ITU Journal
ICT Discoveries

First special issue on

The impact of Artificial Intelligence on communication networks and services

ITU Journal
ICT Discoveries

First special issue on

The impact of Artificial Intelligence on communication networks and services

ITU Journal
ICT Discoveries

The ITU Journal: *ICT Discoveries* publishes original research on telecommunication/ICT technical developments and their policy and regulatory, economic, social and legal dimensions. It builds bridges between disciplines, connects theory with application, and stimulates international dialogue. This interdisciplinary approach reflects ITU's comprehensive field of interest and explores the convergence of telecommunication/ICT with other disciplines. It also features review articles, best practice implementation tutorials and case studies.

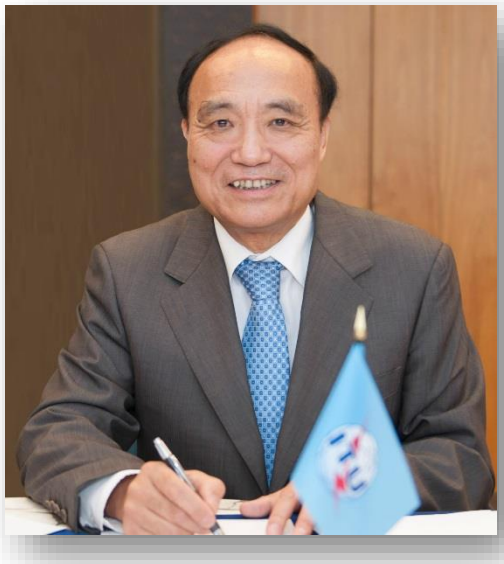
Publication rights

SUGGESTED CITATION: ITU Journal: *ICT Discoveries*, vol. 1(1), Mar. 2018; License: CC BY-NC-ND 3.0 IGO

ADDITIONAL INFORMATION
Please visit the ITU Journal website at:
https://www.itu.int/en/journal/Pages/default.aspx

Inquiries should be addressed to
Alessia Magliarditi at: *journal@itu.int*

Foreword



The International Telecommunication Union is proud to have launched the new, scholarly and professional ITU Journal: *ICT Discoveries*, and published its first special issue.

This Journal was established to encourage the participation of universities and research institutions from different fields in the work of ITU. It identifies emerging technical developments, including their policy, regulatory, economic, social and legal dimensions, in information and communication technologies (ICTs).

The ITU Journal builds bridges between disciplines, connects theory with application, and promotes the critical role that ICTs will play in the pursuit of United Nations' Sustainable Development Goals.

The interdisciplinary scope of the ITU Journal reflects the inclusive character of ITU.

ITU is the United Nations specialized agency for ICTs. We are a public-private partnership, given life by a membership of 193 Member States, over 800 members from industry, international and regional organizations and academia. Researchers participate alongside industry-leading engineers and policymakers in ITU expert groups responsible for radiocommunication, standardization and development.

The long-term vision of academia helps ITU to prepare for the future. The ITU Journal matches research on technical innovation in ICT with analysis of associated transformations in business and society and the related complexities of governance in the digital era. It will help industry players and policymakers to prepare for the impacts of major breakthroughs in research.

The ITU Journal is a peer-reviewed, open access, freely available digital publication.

This first special issue of the ITU Journal forecasts how Artificial Intelligence (AI) will improve user experience by enhancing the performance and efficiency of communication networks. It will assist ITU members in their preparations for AI's expected influence on their work, providing the context for this influence from the perspectives of technology as well as law, ethics, society and regulation.

I express my gratitude to all contributors to the ITU Journal and I would especially like to thank the ITU Journal's Editor-in-Chief, Professor Jian Song of Tsinghua University, for the great dedication with which he has led the curation of this first special issue.

Houlin Zhao
Secretary-General
International Telecommunication Union

Foreword

Artificial Intelligence (AI) and Machine Learning (ML) are certain to influence a broad scope of ITU's technical work. This first special issue of the ITU Journal provides the context for this influence from the perspectives of technology as well as business, law and ethics.

Standardization experts contributing to ITU-T Study Groups are accelerating their studies of AI's potential to assist their preparations for the 5G era.

These studies are looking to AI to bring more automation and intelligence to network design, operation and maintenance, introducing greater efficiency with network self-optimization. Automated virtual assistants are expected to support the customization of multimedia services, and learning algorithms are playing an increasing role in the development of video compression algorithms and algorithmic tools to monitor quality of service and user experience. Cities of the future will be built on the smart use of data, with AI and machine learning delivering data-driven insight to assist cyber-physical systems in adapting their behavior autonomously in the interests of efficiency.

A new ITU Focus Group on ML for 5G is offering essential support to these studies by proposing technical frameworks to assist machine learning in contributing to the efficiency of emerging 5G systems.

Machine Learning is expected to assist the ICT industry in meeting the challenges brought on by 5G and the Internet of Things, shifts representative of considerable increases in network complexity and the diversity of device requirements. The new ITU Focus Group on ML for 5G will define the requirements of machine learning as they relate to technology, network architectures and data formats. Key to this work will be the definition of required data formats and associated mechanisms to safeguard security and privacy.

The AI for Good Global Summit is the leading United Nations platform for dialogue on AI, aiming to ensure that AI accelerates progress towards the United Nations' Sustainable Development Goals.

The 2nd AI for Good Global Summit at ITU Headquarters in Geneva, 15-17 May 2018, will continue to formulate strategies to ensure trusted, safe and inclusive development of AI technologies and equitable access to their benefits. The action-oriented 2018 summit will identify practical applications of AI and supporting strategies to improve the quality and sustainability of life on our planet. It builds on the success of the ground-breaking AI for Good Global Summit in June 2017, the first event to launch inclusive global dialogue on the actions necessary to ensure that AI benefits humanity.

ITU's standardization sector will continue to play an important role in ITU's expanding set of activities on AI.

This first special issue of the ITU Journal, the AI and Machine Learning studies of ITU-T Study Groups and our new Focus Group – alongside the AI for Good series – will be of great value to ITU standardization experts in their efforts to determine how their work could be of most value to AI innovation.

Chaesub Lee
Director
ITU Telecommunication Standardization Bureau

Editor-in-Chief Message

The research community works in service of the public interest. We hope to ensure that our work supports social and economic development on a global scale. We share these ideals with the United Nations.

The ITU Journal: *ICT Discoveries* is a prime example of how ITU and academia are enhancing collaboration to our mutual benefit. In the ITU Journal, the research community sees a valuable new opportunity to serve the public interest. We see a new opportunity to make our research known to public and private-sector decision-makers worldwide.

The ITU Journal looks at both technical and social aspects of ICTs' influence on business, governance and society. It will find new connections between technical and social sciences, as well as new connections between various industry sectors and public-sector bodies. We are certain to uncover new opportunities to work together.

This first special issue of the ITU Journal demonstrates the interdisciplinary scope of this publication.

The issue highlights the potential of Artificial Intelligence to support communication networks and services in fields including cognitive radio, automated driving and the monitoring of our environment. It also explores design principles for AI systems sensitive to human values as well as the ethical implications of advancing AI capabilities as they relate to data security.

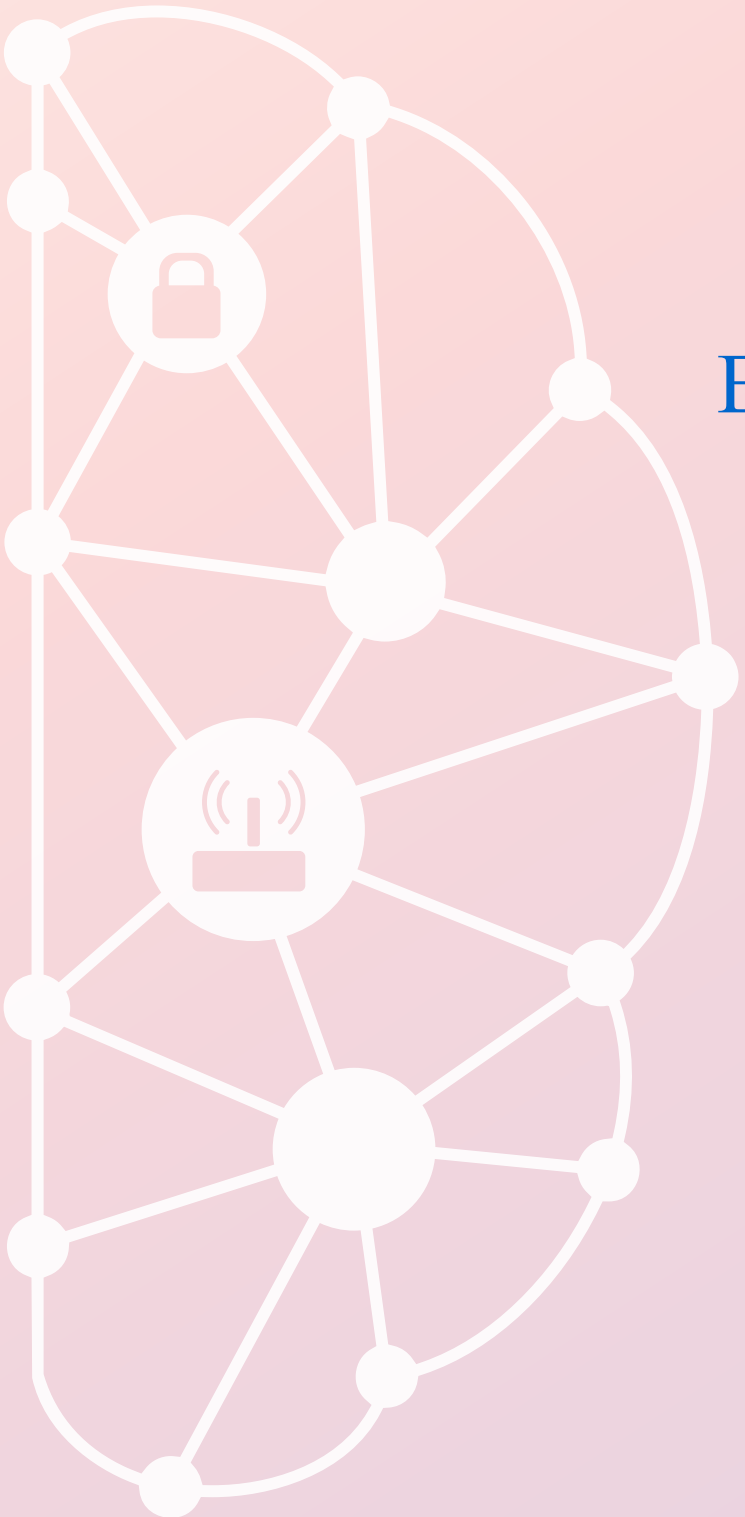
I would like to thank the ITU Secretary-General, Houlin Zhao, for entrusting me with the role of ITU Journal Editor-in-Chief. I also thank the Director of the ITU Telecommunication Standardization Bureau, Chaesub Lee, for the outstanding support that I have received from his bureau.

I would like to express my gratitude to all contributors to this first special issue of the ITU Journal as well as the Outreach Chairman, Stephen Ibaraki and my Associate Editors-in-Chief: Rajkumar Buyya, University of Melbourne, Australia; Jun Kyun Choi, Korea Advanced Institute of Science and Technology; Xiaolan Fu, University of Oxford, UK; Urs Gasser, Harvard University, USA; Alison Gillwald, Research ICT Africa, South Africa; Terry Kramer, University of California, Los Angeles; and Mostafa Hashem Sherif, AT&T, USA.

For their significant contribution to the review process, my special thanks go to all the reviewers and Guest Editors: Antoine Bigomokero Bagula, University of Western Cape, South Africa; Loreto Bravo, Universidad del Desarollo, Chile; Urs Gasser, Harvard University; Larry Holder, Washington State University, USA; Deyi Li, Chinese Academy of Engineering, China; Kazuo Sugiyama, NTT DOCOMO, Japan; Daniel Zeng, University of Arizona, USA; Jun Zhu, Tsinghua University, China.

I look forward to feedback from our readers and our continued work together to stimulate this exciting new direction in ITU-Academia collaboration.

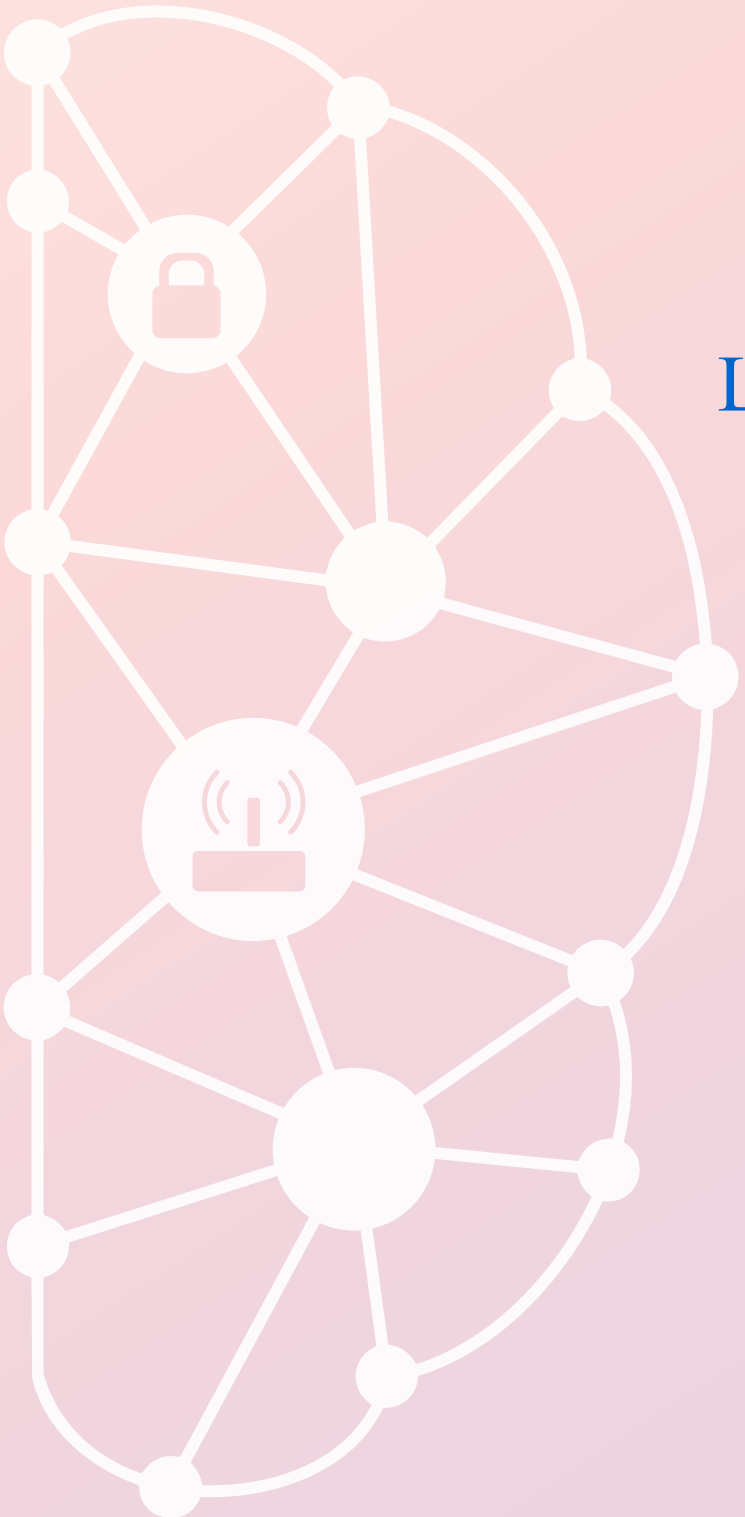
Jian Song
Tsinghua University
China



EDITORIAL BOARD

ITU Editorial Team

TABLE OF CONTENTS



LIST OF ABSTRACTS

Invited Papers

Inclusion of Artificial Intelligence in Communication Networks and Services

Pages 33-38

Xu Guibao, Mu Yubo, Liu Jialiang

AI with learning abilities is a revolutionary technology which the communication industry is exploring, with the aim of introducing it into communication networks and to provide new services, and to improve network efficiency and user experience. At this time there is no total solution or complete framework to do so. One contender in the steps towards a solution is a FINE framework, which can be illustrated by the example of an SDN/NFV collaboratively-deployed network.

View Article

Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models

Pages 39-48

Wojciech Samek, Thomas Wiegand, Klaus-Robert Müller

With the availability of large databases and recent improvements in deep learning methodology, the performance of AI systems is reaching, or even exceeding, the human level on an increasing number of complex tasks. Impressive examples of this development can be found in domains such as image classification, sentiment analysis, speech understanding or strategic game playing. However, because of their nested non-linear structure, these highly successful machine learning and artificial intelligence models are usually applied in a black-box manner, i.e. no information is provided about what exactly makes them arrive at their predictions. Since this lack of transparency can be a major drawback, e.g. in medical applications, the development of methods for visualizing, explaining and interpreting deep learning models has recently attracted increasing attention. This paper summarizes recent developments in this field and makes a plea for more interpretability in artificial intelligence. Furthermore, it presents two approaches to explaining predictions of deep learning models, one method which computes the sensitivity of the prediction with respect to changes in the input and one approach which meaningfully decomposes the decision in terms of the input variables. These methods are evaluated on three classification tasks.

View Article

The Convergence of Machine Learning and Communications

Pages 49-58

Wojciech Samek, Slawomir Stanczak, Thomas Wiegand

The areas of machine learning and communication technology are converging. Today's communication systems generate a large amount of traffic data, which can help to significantly enhance the design and management of networks and communication components when combined with advanced machine learning methods. Furthermore, recently developed end-to-end training procedures offer new ways to jointly optimize the components of a communication system. Also, in many emerging application fields of communication technology, e.g. smart cities or Internet of things, machine learning methods are of central importance. This paper gives an overview of the use of machine learning in different areas of communications and discusses two exemplar applications in wireless networking. Furthermore, it identifies promising future research topics and discusses their potential impact.

View Article

Application of AI to Mobile Network Operations

Pages 59-65

Tomoyuki Otani, Hideki Toube, Tatsuya Kimura, Masanori Furutani

With the introduction of network virtualization and the implementation of 5G/IoT, mobile networks will offer more diversified services and be more complex. This raises a concern about a significant rise in network operation workload. Meanwhile, artificial intelligence (AI) technology is making remarkable progress and is expected to solve human resource shortages in various fields. Likewise, the mobile industry is gaining momentum toward the application of AI to network operation to improve the efficiency of mobile network operation. This paper will discuss the possibility of applying AI technology to network operation and presents some use cases to show good prospects for AI-driven network operation.

View Article

Selected Papers

On Adaptive Neruo-Fuzzy Model for Path Loss Prediction in the VHF Band

Pages 67-75

Nazmat T. Surajudeen-Bakinde, Nasir Faruk, Muhammed Salman, Segun Popoola, Abdulkarim Oloyede, Lukman A. Olawoyin

Path loss prediction models are essential in the planning of wireless systems, particularly in built-up environments. However, the efficacies of the empirical models depend on the local ambient characteristics of the propagation environments. This paper introduces artificial intelligence in path loss prediction in the VHF band by proposing an adaptive neuro-fuzzy (NF) model. The model uses five-layer optimized NF network based on back propagation gradient descent algorithm and least square errors estimate. Electromagnetic field strengths from the transmitter of the NTA Ilorin, which operates at a frequency of 203.25 MHz, were measured along four routes. The prediction results of the proposed model were compared to those obtained via the widely used empirical models. The performances of the models were evaluated using the Root Mean Square Error (RMSE), Spread Corrected RMSE (SC-RMSE), Mean Error (ME), and Standard Deviation Error (SDE), relative to the measured data. Across all the routes covered in this study, the proposed NF model produced the lowest RMSE and ME, while the SDE and the SC-RMSE were dependent on the terrain and clutter covers of the routes. Thus, the efficacy of the adaptive NF model was validated and can be used for effective coverage and interference planning.

View Article

Beyond MAD?: The Race for Artificial General Intelligence

Pages 77-84

Anand Ramamoorthy, Roman Yampolskiy

Artificial intelligence research is a source of great technological advancement as well as ethical concern, as applied AI invades diverse aspects of human life. Yet true artificial general intelligence remains out of reach. Based on the history of deeply transformative technologies developed by multiple actors on the global stage and their consequences for global stability, we consider the possibility of artificial general intelligence arms races and propose solutions aimed at managing the development of such an intelligence without increasing the risks to global stability and humanity.

View Article

Artificial Intelligence for Place-Time Convolved Wireless Communication Networks

Pages 85-93

Ambuj Kumar

Previous works have brought to light that the dynamic and variable motions of potential network users, and other environmental factors, are the eternal threat to present and future wireless communications. One of my earlier works discusses this perennial and perpetual challenge as the place-time dependent functions. The phenomena was coined in my work as place time capacity (PTC), and place time coverage (PTCo), with both collectively known as place time coverage and capacity (PTC2), are derived as the outcomes of dynamics that can be expressed as the functions of place and time. These phenomena degrade the efficiency of any wireless communication network (WCN) to that lowest point, from where, a network service provider (NSP) may not have any choice but to revamp the network. Artificial intelligence (AI), on the other hand, has been striding profoundly for the past several decades, fanning out its influence in various sectors of scientific and technological developments. However, AI is almost absent in the area of WCN dimensioning and optimization, especially for place–time events. This paper revisits the two place-time functions as WCN phenomena and, with the backdrop of these aspects, shall investigate the inevitable need for AI in WCNs, as well as demonstrating how AI can be part of present and future wireless communications.

View Article

Bayesian Online Learning-Based Spectrum Occupancy Prediction in Cognitive Radio Networks

Ahmed Mohammed Mikaeil

Predicting the near future of primary user (PU) channel state availability (i.e., spectrum occupancy) is quite important in cognitive radio networks in order to avoid interfering its transmission by a cognitive spectrum user (i.e., secondary user (SU)). This paper introduces a new simple method for predicting PU channel state based on energy detection. In this method, we model the PU channel state detection sequence (i.e., "PU channel idle" and "PU channel occupied") as a time series represented by two different random variable distributions. We then introduce Bayesian online learning (BOL) to predict in advance the changes in time series (i.e., PU channel state.), so that the secondary user can adjust its transmission strategies accordingly. A simulation result proves the efficiency of the new approach in predicting PU channel state availability.

View Article

The Evolution of Fraud: Ethical Implications in the Age of Large-Scale Data Breaches and Widespread Artificial Intelligence Solutions Deployment

Abhishek Gupta

Artificial intelligence is being rapidly deployed in all contexts of our lives, often in subtle yet behavior-nudging ways. At the same time, the pace of development of new techniques and research advancements is only quickening as research and industry labs across the world leverage the emerging talent and interest of communities across the globe. With the inevitable digitization of our lives, increasingly sophisticated and ever larger data security breaches in the past few years, we are in an era where privacy and identity ownership is becoming a relic of the past. In this paper, we will explore how large-scale data breaches coupled with sophisticated deep learning techniques will create a new class of fraud mechanisms allowing perpetrators to deploy "Identity Theft 2.0".

View Article

Machine Intelligence Techniques for Next-Generation Context-Aware Wireless Networks

Tadilo Endeshaw Bogale, Xianbin Wang, Long Bao Le

Next generation wireless networks (i.e., 5G and beyond), which will be extremely dynamic and complex due to the ultra-dense deployment of heterogeneous networks (HetNets), pose many critical challenges for network planning, operation, management and troubleshooting. At the same time, the generation and consumption of wireless data are becoming increasingly distributed with an ongoing paradigm shift from people-centric to machine-oriented communications, making the operation of future wireless networks even more complex. In mitigating the complexity of future network operation, new approaches of intelligently utilizing distributed computational resources with improved context awareness becomes extremely important. In this regard, the emerging fog (edge) computing architecture aiming to distribute computing, storage, control, communication, and networking functions closer to end users, has a great potential for enabling efficient operation of future wireless networks. These promising architectures make the adoption of artificial intelligence (AI) principles, which incorporate learning, reasoning and decision-making mechanisms, natural choices for designing a tightly integrated network. To this end, this article provides a comprehensive survey on the utilization of AI integrating machine learning, data analytics and natural language processing (NLP) techniques for enhancing the efficiency of wireless network operation. In particular, we provide comprehensive discussion on the utilization of these techniques for efficient data acquisition, knowledge discovery, network planning, operation and management of next generation wireless networks. A brief case study utilizing the AI techniques for this network has also been provided.

View Article

New Technology Brings New Opportunity for Telecommunication Carriers: Artificial Intelligent Applications and Practices in Telecom Operators

Pages 121-127

Wei Liang, Mingjun Sun, Baohong He, Mingchuan Yang, Xiaoou Liu, Bohuan Zhang, Yuntao Wang

In the era of "computational intelligence, perceptional intelligence and cognitive intelligence" as the main direction for the future, telecom operators are on their way to building their own artificial intelligence (AI) ecosystem. In terms of developing AI technology, telecom operators have unique resources and technology advantages: big data resources, superior computing power, lots of investment in AI algorithmic research, broad government and enterprise customer resources. By making full use of these strengths, they have carried out a series of effective practices in the various field and achieved constructive results. This report will be arranged as follows. In the first part the history and the development status of AI has been introduced, as well as the Chinese powerful policy which was released to support its development. In the second part, the unique advantages for operators to develop AI have been introduced, whilst in the meantime, the AI development idea for telecom operators has been provided. In the third part, in order to be more persuasive, the practice of AI by telecom operators in multiple fields to satisfy internal requirements and meet customer needs, has been described. Finally, based on the current development trends of AI, its future prospects are made by this report. Undoubtedly, in the future, operators will further use their advantages to explore more AI development opportunities.

View Article

Correlation and Dependence Analysis on Cyberthreat Alerts

Pages 129-135

John M.A. Bothos, Konstantinos-Georgios Thanos, Dimitris M. Kyriazanos, George Vardoulias, Andreas Zalonis, Eirini Papadopoulou, Yannis Corovesis, Stelios C.A. Thomopoulos

In this paper a methodology for the enhancement of computer networks' cyber-defense is presented. Using a time-series dataset, drawn for a 60-day period and for 12 hours per day and depicting the occurrences of cyberthreat alerts at hourly intervals, the correlation and dependency coefficients that occur in an organization's network between different types of cyberthreat alerts are determined. Certain mathematical methods like the Spearman correlation coefficient and the Poisson regression stochastic model are used. For certain types of cyberthreat alerts, results show a significant positive correlation and dependence between them. The analysis methodology presented could help the administrative and IT managers of an organization to implement organizational policies for cybersecurity.

View Article

RESPONSIBLE ARTIFICIAL INTELLIGENCE: DESIGNING AI FOR HUMAN VALUES

Virginia Dignum
Delft University of Technology, The Netherlands

Abstract – Artificial intelligence (AI) is increasingly affecting our lives in smaller or greater ways. In order to ensure that systems will uphold human values, design methods are needed that incorporate ethical principles and address societal concerns. In this paper, we explore the impact of AI in the case of the expected effects on the European labor market, and propose the accountability, responsibility and transparency (ART) design principles for the development of AI systems that are sensitive to human values.

Keywords – Artificial intelligence, design for values, ethics, societal impact

1. INTRODUCTION

Artificial intelligence (AI) is becoming rapidly present in all aspects of everyday life. It is everywhere, it affects everyone, and its capabilities are evolving extremely rapidly. AI can help us in many ways: it can perform hard, dangerous or boring work for us; it can help us to save lives and cope with disasters; and, it can entertain us and make our daily life more comfortable. AI systems manage complex, data-intensive tasks, e.g. monitoring credit card systems for fraudulent behavior, enabling high-frequency stock trading, supporting medical diagnoses and detecting cybersecurity threats. Embodied as robots, AI is soon to move and work among us, in the form of service, transportation, medical and military robots. Nevertheless, current perceptions and expectations regarding the capabilities of AI vary widely and consensus on the societal impact of AI is hard to find. In the first part of this paper, we analyze this situation by means of a study on the expected effect of AI on the European job market.

The second part of the paper explores the social, economic, political, technological, legal, ethical and philosophical questions raised by AI and how design methods can deal with these. Currently, there is an increasing awareness that a responsible approach to AI is needed to ensure the safe, beneficial and fair use of AI technologies. This also includes the need to consider the ethical implications of decisions made by machines, and to define the legal status of AI. However, concrete approaches to the responsible design of AI are mostly non-existent. The responsible design, development and use of AI systems is of the utmost relevance to AI applications such as self-driving vehicles, companion, healthcare robots, and ranking and profiling algorithms, which are already affecting society or will be in a few years. In all these applications, AI reasoning should be able to take into account societal values, moral and ethical considerations, weigh up the respective priorities of values held by stakeholders and in different multicultural contexts, explain its reasoning and guarantee transparency.

Answering these and related questions requires a whole new understanding of ethics and to rethink the concept of agency in the changing socio-technical reality. Moreover, implementing ethical actions in machines will help us better understand ethics overall.

To enable the required technological developments and responses, AI researchers and practitioners will need to be able to take moral, societal and legal values into account in the design of AI systems. Developing AI responsibly requires the means to elicit and represent human values, translate these values into technical requirements, develop the means to deal with moral dilemmas and values preferences, and to evaluate systems in terms of their contribution to human wellbeing.

Developments in autonomy and machine learning are rapidly enabling AI systems to decide and act without direct human control. Greater autonomy must come with greater responsibility, even when these notions are necessarily different when applied to machines than to people. Ensuring that systems are designed responsibly contributes to our trust of their behavior, and requires both accountability, i.e. being able to explain and justify decisions, and

transparency, i.e. understanding the ways systems make decisions and how the data is being used, collected and governed. To this effect, we have proposed the principles of accountability, responsibility and transparency (ART) [7]. ART implements a design for values approach [26, 10], to ensure that human values and ethical principles, and their priorities and choices are explicitly included in the design processes in a transparent and systematic manner.

2. EXPECTATIONS ON THE IMPACT OF AI

In the past technical innovation has always created more jobs and led to a higher average standard of living; however, this does not mean that the implementation of new technologies has ever gone without opposition [4]. As shown by the Luddite movement in the 18th century and superbly demonstrated in Charlie Chaplin's influential movie "Modern Times", technological change and the subsequent displacement or change in the nature of jobs has led to great social unrest in the past [27].

The current wave of AI development has already incited wide public discussion on its effects on jobs and standards of living. An increasing number of people and organizations are warning about the possible negative impact of AI implementation on jobs and society, and several expect AI to cause more extreme effects than previous technological revolutions [4].

Boasting one of the world's largest economies and a highly educated workforce, this problem is very relevant to the European Union. Already in 2014, European Commissioner Kroes indicated that up to 70% of EU citizens believe that robots will steal people's jobs"[1].

Somber predictions on future AI capabilities put an increasing pressure on policy makers to protect the European economy and workforce. However, comparing possible policies proves to be hard given the uncertainty of future effects of AI. In fact, current studies on the influence of AI on the jobs market vary from a Utopian society in which nobody has to work, to the ending of economic growth in the western world [13, 23].

In order to provide European policy makers with a clear forecast of the future effects of AI on the European labor market and a recommendation on future policy directions combating potential harmful effects to this market we have performed a qualitative study on the expectations on AI. This forecast will be constructed by means of an adapted Delphi method study, facilitating discussion among European AI experts to create a consensus-based forecast of future AI effects on the European labor market.

2.1 Literature analysis

Existing reports on the number of jobs that can theoretically be replaced by AI in the long term, indicate figures as high as 47% of job losses in the US [9], and 35% in the UK [6]. Nevertheless, policy discussions on the effects of AI on jobs are still scarce in Europe. The topic seems to be of importance to some national governments [27, 6] but there is no clear European policy vision on potential harmful effects on the jobs market. In other countries, namely the USA, protecting workers from technological change is a more regular policy topic [20]. Moreover, few studies have provided a clear estimate of the amount of jobs that will be created or on the nature of future jobs. Some researchers looking at historical data expect that created jobs will outnumber those lost [18]. On the other hand, [25] states: "Experts envision automation and intelligent digital agents permeating vast areas of our work […], but they are divided on whether these advances will displace more jobs than they create".

Consensus does exist on the necessity of re-education of employees as preparation for future changes [9]. A panel of experts, hosted by McKinsey in 2014, expected that the number of US manufacturing jobs is rising and will continue to be in the coming years but it is very important to educate these people to work with machines otherwise they will not be needed in the future [16].

With regard to how AI contributes to this changing market, existing studies show an almost even divide among researchers between a positive and negative impact on the European economy [25]. Existing literature on this topic shows a very theoretical, sometimes philosophical, future view on labor markets. Testing these theories is hard, as they reflect the researchers' interpretation of existing data. Polling studies also show little consensus between researchers [25].

We use the four scenarios proposed by [27] as a means to classify the different studies:

A. **Business-as-usual:** According to this view, technological innovation always leads to higher productivity and the effect of AI will not be different. This productivity can in turn lead to either a larger or a smaller labor market, but, at

[1] http://europa.eu/rapid/press-release_SPEECH-14-421_en.htm

least in the long term, technological innovation has always had a positive effect on the number of European jobs [28]. The business-as-usual scenario therefore predicts a growth in the amount of European jobs market and economy in the long-term, coupled with a change in the nature of jobs and possibly short-term unrest. This unrest can be prevented with timely re-education of employees. Large wealth redistribution programs like the introduction of a basic income are not expected to be necessary.

B. Techno-revolutionists: According to this scenario, AI applications will in time compete with and take over an increasing number of human jobs. The deployment of autonomous systems will cause high levels of unemployment and create a growing gap between income from labor and income from assets, leading to an increasing divide in wealth. Major re-education policies are a necessity to make sure that humans will work with machines rather than compete with them for jobs, in a world where machines will outperform a majority of humans. Increasing wealth inequality is a result of big technological revolutions [21], leading to the need for a more balanced distribution of wealth [5]. This could lead to great societal challenges which require major (public) policy changes, such as the introduction of a basic income or a negative income tax. Tax incentives like subsidies for companies that keep humans on the payroll are also mentioned as policy options.

C. Techno-utopists: A small group of researchers expects that the exponential growth of technological developments will lead to negligible costs of information and energy, through which many physical goods and services will become (almost) free. Technological innovation, in this scenario, will eventually create a society of abundance rather than one of scarcity. Ownership and marginal costs will disappear, leading to the end of capitalism. According to [23], AI will be one of the enabling technologies for this scenario. As humans will spend less time on their jobs, and robots and computer programs will not have a salary, new forms of wealth distribution, such as a universal basic income will be needed to maintain the future economy [14].

D. Techno-pessimists: In contrary to the techno-revolutionists and the techno-utopists, techno-pessimists expect future economic growth to be lower than it is today. In fact [13] indicates that many innovations that can lead to strong economic growth are already implemented and cannot be repeated, whereas at the same time novel technological improvements fail to deliver strong economic effects. [13] Therefore expects future AI to have only a very limited impact on the European labor market. Techno-pessimists argue for increased policy to tackle existing economic headwinds rather than investment in AI.

Table 1 gives an overview of the expected impact of the different scenarios on the economic growth, the role of AI and the effect of different policies, based on the qualitative analysis of literature.

Table 1. An overview of the expected impact on economic growth and effects of different policy directions for the four scenarios: negative (-), neutral (0), positive (+) or very positive (++)

	A	**B**	**C**	**D**
Expected economic growth	Medium	Medium	High	High
Impact AI on economic growth	Medium	High	High	Low
Re-education	+	++	0	0
Wealth redistribution programs	+	+	++	0
Investment in AI	+	0	+	-
Subsidized human workforce	-	+	+	0

2.2 The views of AI experts

The scenarios presented in the previous clause highlight a fundamental disagreement on the impact of AI on the labor market and on the policies that are needed to regulate this impact. In this clause, we describe research performed at the Delft University of Technology in the Netherlands, using an adapted Delphi method [19] to facilitate an open discussion among researchers across Europe, This method combines the benefits of survey research, interview sessions and group discussions, and aims to identify the reasoning and rationale behind differences in opinions among AI experts while guarding against the occurrence of group think.

Delphi studies take an iterative approach to ensure that the strongest possible consensus among participants is reached by asking experts for their opinions on the combined results of previous rounds. The Delphi method does not state a fixed boundary on the amount of participants to form an adequate sample size. Finding motivated and knowledgeable respondents is more important than creating a statistically significant sample size. Delphi studies

are often conducted with small sample sizes [22] and participants are not selected at random but because of their particular expertise. The most significant features of the Delphi method are its recursion and the possibility to get feedback and evaluate one's own answers. These characteristics of the Delphi method have been proven to guarantee the validity and reliability in case of studies aiming at predicting or understanding possible future scenarios [15, 11]. This method is therefore suitable to study the impact of AI on the European labor market.

In the above-mentioned study, that took place mid-2016, experts were invited to participate by email; emails were sent to relevant mailing lists and through the European Association for Artificial Intelligence (EurAI). All respondents were screened on their experience by the researchers. This approach led to a total of five respondents. Five additional experts, prominent European AI researchers from three different European countries, indicated their willingness to participate in an interview session for the validation of the study. The first questionnaire aimed at establishing an initial understanding of respondents' views on future AI capabilities, effects on the nature and size of the European labor markets, the factors influencing these effects and possible governmental roles and policies. The second questionnaire provided more detailed predictions through the identification of the timeline and specific factors expected to influence the effect of AI. As such, respondents were asked to comment on short to mid-term (0-10 years) and long-term (>10 years) effects. Respondents were further asked to reflect on specific policies, along the aspects identified in the literature study. The second questionnaire also included a section where respondents could rate the influence of a variety of factors on the effects AI will have on the amount of European jobs and on the nature of European jobs (on a scale of one to five). These results are depicted in Fig. 1.

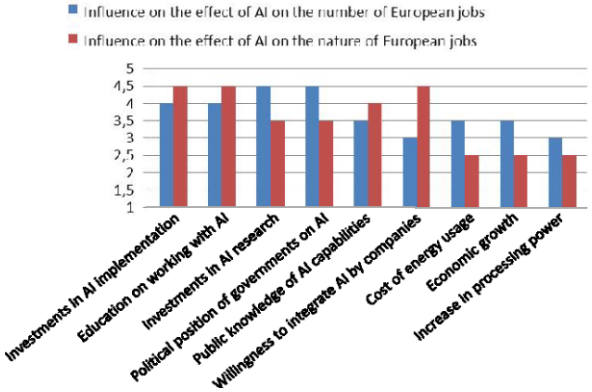


Fig. 1. Influence of different factors on the effects of AI on the nature and number of European jobs. Scores range from 1, no influence to 5, huge influence

After these two rounds of questionnaires, further validated by means of an interview session, respondents reached consensus on the following points: (i) Future AI will decrease the number of mechanical/non-knowledge intensive jobs in the short term; (ii) it will create new, most likely very specialized jobs; (iii) and will have a large impact on the nature of European jobs; (iv) governments will need to revise their education system to make sure their future workforce can work with AI. However, no consensus has been reached on the net result of the influence of AI on the number of European jobs in the long term and on the factors influencing the impact of AI on the number and nature of those jobs. In combination with the results of existing literature, described in the previous clause, this analysis of the views of leading AI researchers provides a useful forecast on the future effects of artificial intelligence on the European labor market to aid policy makers in preparing Europe for a smart future. The results of the Delphi study bring a somewhat moderate view on the effects of AI on the European labor market, which complement and extend current scientific literature. In their short term prediction, the views of the experts consulted are fairly consensual, and mostly aligned with the outcomes predicted by the business-as-usual scenario. Consensus on long term effects is narrower and includes elements from the business-as-usual, techno-optimist and techno-utopist scenarios. Nevertheless, it is important to note that Delphi style research leads to findings that are not necessarily statistically supported, but that can be used to inform further research on the expectations on the social impact of AI at a larger scale. The main contribution of this study is that it tempers the current hype on the impact of AI, by bringing in the views of AI experts with a long experience in the field. This can support policy makers in tempering their expectations.

3. RESPONSIBILITY IN AI

In this clause, we discuss how to approach the design of AI systems that are sensitive to moral principles and human value. Responsible AI is more than the ticking of some ethical 'boxes' or the development of some add-on features in AI systems. Rather, responsibility is fundamental to intelligence and no system can be truly intelligent if it cannot understand responsibility.

Responsible AI rests in three pillars of equal importance. Firstly, society in general must be prepared to take responsibility for the impact of AI. This means that researchers and developers should

be trained to be aware of their own responsibility where it concerns the development of AI systems with direct impact in society. This requires efforts in education and training and the development of codes of conduct. Moreover, responsible AI is an issue of regulation and legislation. It is up to governments and citizens to determine how issues of liability should be regulated. For example, who will be to blame if a self-driving car harms a pedestrian? Is it the builder of the hardware (e.g. of the sensors used by the car to perceive the environment)?; the builder of the software that enables the car to decide on a path?; the authorities that allow the car on the road?; the owner that personalized the car decision-making settings to meet her preferences?; or, the car itself because its behavior is based on its own learning? All these, and more questions must be informing the regulations that societies put in place towards responsible use of AI systems.

Secondly, responsible AI implies the need for mechanisms that enable AI systems themselves to reason about, and act according to, ethics and human values. This requires models and algorithms to represent and reason about, and take decisions based on, human values, and to justify their decisions according to their effect on those values. Current (deep-learning) mechanisms are unable to meaningfully link decisions to inputs, and therefore cannot explain their acts in ways that we can understand.

Thirdly, participation; it is necessary to understand how different people work with and live with AI technologies across cultures in order to develop frameworks for responsible AI. In fact, AI does not stand in itself, but must be understood as part of socio-technical relations. Here again education plays an important role, both to ensure that knowledge of the potential AI is widespread, as well as to make people aware that they can participate in shaping the societal development. A new and more ambitious form of governance is one of the most pressing needs in order to ensure that inevitable AI advances will serve societal good.

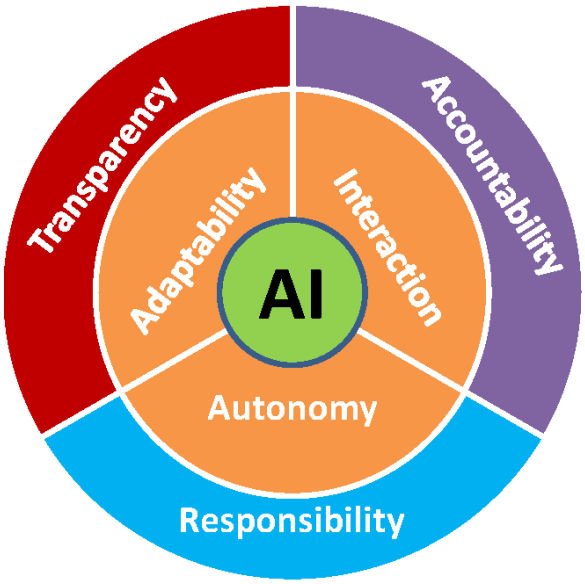


Fig. 2: The ART principles: accountability, responsibility, transparency

AI systems are often characterized by their autonomy, interactivity and adaptability [8, 24]. To reflect societal concerns about the ethics of AI, and ensure that AI systems are developed responsibly, incorporating social and ethical values, we propose to complement these properties with the principles of accountability, responsibility and transparency (ART) [7], as depicted in Fig. 2.

Accountability refers to the need to explain and justify one's decisions and actions to its partners, users and others with whom the system interacts. To ensure accountability, decisions must be derivable from, and explained by, the decision-making algorithms used. This includes the need for representation of the moral values and societal norms holding in the context of operation, which the agent uses for deliberation. Accountability in AI requires both the function of guiding action (by forming beliefs and making decisions), and the function of explanation (by placing decisions in a broader context and by classifying them along moral values).

Responsibility refers to the role of people themselves, and to the capability of AI systems to answer for one's decision and identify errors or unexpected results. As the chain of responsibility grows means are needed to link the AI system's decisions to the fair use of data and to the actions of stakeholders involved in the system's decision.

Transparency refers to the need to describe, inspect and reproduce the mechanisms through which AI systems make decisions and learn to adapt to their environment, and to the governance of the data used or created. Current AI algorithms are basically black

boxes. However, regulators and users demand explanation and clarity about the data used. Methods are needed to inspect algorithms and their results and to manage data, their provenance and their dynamics.

3.1. Responsible AI challenges

In this clause, we discuss how the general principles described above can direct the development of AI systems. Assuming that the development of AI systems follows a standard engineering cycle of Analysis-Design-Implement-Evaluate, taking a design for values approach basically means that the analysis phase will need to include activities for (i) the identification of societal values, (ii) deciding on a moral deliberation approach (e.g. through algorithms, user control or regulation), and (iii) methods to link values to formal system requirements [1].

Responsibility is associated with the capability of moral deliberation, in particular that which is related to dealing with moral dilemmas for which there is not one optimal solution. Several authors have discussed the trolley problem as an example of such a situation. In this scenario, an AI system, e.g. an autonomous vehicle, must decide between harming pedestrians or its own passengers when an accident cannot be avoided. Approaches to moral deliberation reflect ethical theories, such as utilitarianism (save the most lives) or deontological/Kantian (do no harm deliberately). From an implementation perspective, the different ethical theories differ in terms of computational complexity of the required deliberation algorithms. To implement consequentialist agents, reasoning about the consequences of actions is needed, which can be supported by, e.g. dynamic logics. For deontological agents, higher order reasoning is needed to reason about the actions themselves, i.e. the agent must be aware of its own action capabilities and their relations to institutional norms and the rule of law.

Accountability requires both the function of guiding action (by forming beliefs and making decisions), and the function of explanation (by placing decisions in a broader context and by classifying them along moral values). To this effect, machine learning techniques can be used to classify states or actions as 'right' or 'wrong', basically in the same way as classifiers learn to distinguish between cats and dogs.

Another approach to develop explanation methods is to apply evolutionary ethics [2] and structured argumentation models [17].

This moreover provides a model-agnostic approach potentially able to deal with transparency in stochastic, logic and data-based models in a uniform way. Further research is needed to verify this approach. Yet another approach is proposed in [12] based on pragmatic social heuristics instead of moral rules or maximization principles. This approach takes a learning perspective integrating both the initial ethical deliberation rules with adaptation to the context. Finally, poorly understood behavior by AI systems can have large and lasting consequences, and adaptive systems may arrive at "perverse instantiations" of their programmed goals [3].

4. CONCLUDING REMARKS

Increasingly, AI systems will be taking decisions that affect our lives and our way of living in smaller or greater ways. In all areas of application, AI must be able to take into account societal values, moral and ethical considerations, weigh up the respective priorities of values held by different stakeholders and in multicultural contexts, explain its reasoning, and guarantee transparency. As the capabilities for autonomous decision making grow, perhaps the most important issue to consider is the need to rethink responsibility. Being fundamentally tools, AI systems are fully under the control and responsibility of their owners or users. However, their potential autonomy and capability to learn, require that design considers accountability, responsibility and transparency principles in an explicit and systematic manner. The development of AI algorithms has so far been led by the goal of improving performance, leading to opaque black boxes. Putting human values at the core of AI systems calls for a mind shift of researchers and developers towards the goal of improving transparency rather than performance, which will lead to novel and exciting techniques and applications.

As AI systems replace people in many traditional jobs, it is necessary to rethink the meaning of work. Jobs change but more importantly the character of jobs will change. Meaningful occupations are those that contribute to the welfare of society, the fulfillment of oneself and the advance of mankind. These are not necessarily equated with current 'paid jobs'. AI systems can free us to, and be reward for, care for each other, engage in arts, hobbies and sports, enjoy nature, and, meditate, i.e. those things that give us energy and make us happy.

Increasingly, robots and intelligent agents will be taking decisions that can affect our lives and way of living in smaller or greater ways. Being fundamentally artifacts, AI systems are fully under the control and responsibility of their owners or users. However, developments in autonomy and learning are rapidly enabling AI systems to decide

and act without direct human control. That is, in dynamic environments, their adaptability capabilities can lead to situations in which the consequences of their decisions and actions will not be always possible to direct or predict.
More than being a risk to human values, AI brings in itself enormous potential to improve the lives of many, and to ensure human rights to all. However, how this will be realized, depends on us.

ACKNOWLEDGEMENT

The author is grateful to Thomas van Biemen who performed the study described in clause 2.

REFERENCES

[1] H Aldewereld, V Dignum, and YH Tan. Design for values in software development, 2015.

[2] Ken Binmore. Natural justice. Oxford University Press, 2005.

[3] Nick Bostrom. Superintelligence: Paths, dangers, strategies. OUP Oxford, 2014.

[4] Erik Brynjolfsson and Andres McAfee. The second machine age: Work, progress, and prosperity in a time of brilliant technologies. WW Norton & Company, 2014.

[5] Erik Brynjolfsson, Andrew McAfee, and Michael Spence. Labor, capital, and ideas in the power law economy. Foreign Aff., 93:44, 2014.

[6] Deloitte. From brawn to brains, the impact of technology on jobs in the UK. Technical report, 2015. Accessed: 15-08-2017.

[7] Virginia Dignum. Responsible autonomy. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJ-CAI'2017), pages 4698-4704, 2017.

[8] Luciano Floridi and Jeff W Sanders. On the morality of artificial agents. Minds and machines, 14(3):349-379, 2004.

[9] Carl Benedikt Frey and Michael A Osborne. The future of employment: how susceptible are jobs to computerisation? Technological Forecasting and Social Change, 114:254-280, 2017.

[10] Batya Friedman, Peter Kahn, and Alan Borning. Value sensitive design and information systems. Advances in Management Information Systems, 6:348-372, 2006.

[11] Dolores Gallego and Salvador Bueno. Exploring the application of the Delphi method as a forecasting tool in information systems and technologies research. Technology Analysis & Strategic Management, 26(9):987–999, 2014.

[12] Gerd Gigerenzer. Moral satisficing: Rethinking moral behavior as bounded rationality. Topics in cognitive science, 2(3):528–554, 2010.

[13] R Gorden. The demise of US economic growth: Restatement, rebuttal and reflections. Technical Report NBER Working Paper No. 19895, National Bureau Economic Research, 2014.

[14] James J. Hughes. A strategic opening for a basic income guarantee in the global crisis being created by AI, robots, desktop manufacturing and biomedicine. Journal of Evolution & Technology, 24(1):45–61, 2014.

[15] K. K. Lilja, K. Laakso, and J. Palomäki. Using the delphi method. In 2011 Proceedings of PICMET '11: Technology Management in the Energy Smart World (PICMET), pages 1–10, July 2011.

[16] McKinsey. Manufacturing the future. Technical report, 2014.

[17] Sanjay Modgil and Henry Prakken. A general account of argumentation with preferences. Artificial Intelligence, 195:361–397, 2013.

[18] Int. Federation of Robotics. The impact of robots on productivity, employment and jobs. Technical report, 2017. Accessed: 15-08-2017.

[19] Chitu Okoli and Suzanne D Pawlowski. The Delphi method as a research tool: an example, design considerations and applications. Information & management, 42(1):15–29, 2004.

[20] Committee on Technology National Science, Technology Council, and Penny Hill Press. Preparing for the Future of Artificial Intelligence. CreateSpace Independent Publishing Platform, 2016.

[21] Thomas Piketty. Capital in the twenty-first century. Harvard University Press, 2017.

[22] Adriano Bernardo Renzi and Sydney Freitas. The Delphi method for future scenarios construction. Procedia Manufacturing, 3:5785–5791, 2015.

[23] Jeremy Rifkin. The Zero Marginal Cost Society. The Internet of Things, the Collaborative Commons, and the Eclipse of Capitalism. Palgrave Macmillan, 2014.

[24] SJ Russell and P Norvig. Artificial Intelligence: A Modern Approach. Prentice Hall, 2nd ed, 2002.

[25] A Smith and J Anderson. AI, robotics, and the future of jobs. Pew Research Center, 6, 2014.

[26] Jeroen van den Hoven. Design for values and values for design. Information Age +, Journal of the Australian Computer Society, 7(2):4–7, 2005.

[27] Quirien van Est and Linda Kool. Werken aan de robotsamenleving: visies en inzichten uit de wetenschap over de relatie technologie en werkgelegenheid. 2015. in Dutch.

[28] John Van Reenen. Employment and technological innovation: evidence from UK manufacturing firms. J Labor Economics, 15(2):255–284, 1997.

RECONFIGURABLE PROCESSOR FOR DEEP LEARNING IN AUTONOMOUS VEHICLES

Yu Wang[1, 2], Shuang Liang[3], Song Yao[2], Yi Shan[2], Song Han[2, 4], Jinzhang Peng[2] and Hong Luo[2]

[1]Department of Electronic Engineering, Tsinghua University, Beijing, China
[2]Deephi Tech, Beijing, China
[3]Institute of Microelectronics, Tsinghua University, Beijing, China
[4]Department of Electrical Engineering, Stanford University, Stanford CA, USA

Abstract - The rapid growth of civilian vehicles has stimulated the development of advanced driver assistance systems (ADASs) to be equipped in-car. Real-time autonomous vision (RTAV) is an essential part of the overall system, and the emergence of deep learning methods has greatly improved the system quality, which also requires the processor to offer a computing speed of tera operations per second (TOPS) and a power consumption of no more than 30 W with programmability. This article gives an overview of the trends of RTAV algorithms and different hardware solutions, and proposes a development route for the reconfigurable RTAV accelerator. We propose our field programmable gate array (FPGA) based system Aristotle, together with an all-stack software-hardware co design workflow including compression, compilation, and customized hardware architecture. Evaluation shows that our FPGA system can realize real-time processing on modern RTAV algorithms with a higher efficiency than peer CPU and GPU platforms. Our outlook based on the ASIC-based system design and the ongoing implementation of next generation memory would target a 100 TOPS performance with around 20 W power.

Keywords - Advanced driver assistance system (ADAS), autonomous vehicles, computer vision, deep learning, reconfigurable processor

1. INTRODUCTION

If you have seen the cartoon movie WALL-E, you will remember when WALL-E enters the starliner Axiom following Eve, he sees a completely automated world with obese and feeble human passengers laying in their auto driven chairs, drinking beverages and watching TV. The movie describes a pathetic future of human beings in the year of 2805 and warns people to get up from their chairs and take some exercise. However, the inside laziness has always been motivating geniuses to build auto driven cars or chairs, whatever it takes to get rid of being a bored driver stuck in traffic jams.

At least for now, people find machines genuinely helpful for our driving experience and sometimes they can even save peoples lives. It has been nearly 30 years since the first successful demonstrations of ADAS [1][2][3], and the rapid development of sensors, computing hardware and related algorithms has brought the conceptual system into reality. Modern cars are being equipped with ADAS and the numbers are increasing. According to McKinseys estimation [4], auto-driven cars will form a 1.9 trillion dollars market in 2025. Many governments like those in the USA [5], Japan [6] and Europe [7][8][9] have proposed their intelligent transportation system (ITS) strategic plans, which have drawn up timetables for the commercialization of related technologies.

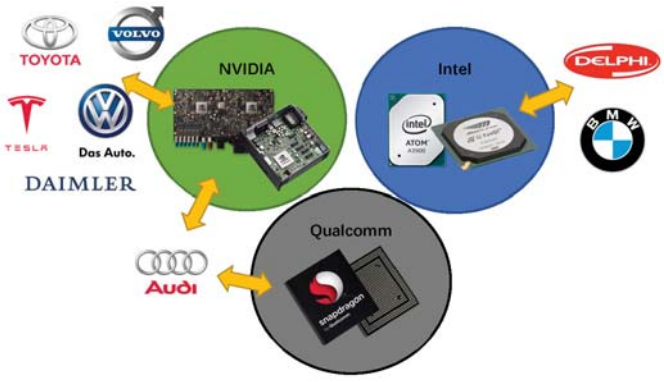


Figure 1. The market pattern of automotive cars.

In current ADASs, machine vision is an essential part; it is also called autonomous vision [10]. Since the conditions of weather, roads and the shapes of captured objects are complex and variable with little concern for safety, the anticipation for high recognition accuracy and rapid system reaction to these is urgent. For state-of-the-art algorithms, the number of operations has already increased to tens and hundreds of giga-operations (GOPs). This has set a great challenge for real time processing, and correspondingly we need to find a powerful processing platform to deal with it.

Fig. 1 shows a pattern of the current market of automotive cars. NVIDIA is leading the market with its Drive series GPU platforms, and has already built cooperation with car manufacturers like Audi, Tesla, Daimler, etc. Intel is also focusing on this area. It has acquired many relevant companies such as Mobileye, Nervana, Movidius and Altera, and has collaborated with BMW and Delphi to build its ecosystem circle. It has also released products such as Atom A3900 for the automotive scene[11]. Another chip giant Qualcomm is also trying to make inroads in this market. It has release dedicated processors like Snapdragon 602A and 820A chips [12], and it has bought NXP to strengthen its impact in the ADAS market.

Many ADAS solutions have chosen graphic processing unit (GPU)-based systems to carry their autonomous vision algorithms, not only because of their powerful computational ability since GPU-based systems can offer massive parallelisms in datapaths and the latest GPU processors can offer a throughput of several TOPS such as the NVIDIA Drive PX2 system [13] with Xavier chips, but also because of the robust and efficient developing environment support such as CUDA [14] and cuDNN [15].

While GPU can offer a computing speed of TOPS, the power consumption can often be the bottleneck for in-car system implementation as some modern GPUs can cost 200-300 W. One solution is to improve the power efficiency, and this can be achieved through the dedicate logic customization, and reconfigurable processors can be a suitable choice. One representative reconfigurable processor is FPGA. FPGA suppliers Xilinx and Altera have already introduced their FPGA products into ADAS scenarios such as Zynq-7000 [16] and Cyclone-V [17] series SoC. While the power is around 10 W, FPGA can also get a peak performance of around 100 GOPS. Together with the features of multi-threading, parallel processing and low latency, FPGA could be expected to be a favorable choice for autonomous vision systems.

Naturally, we can convert an FPGA design into an application-specific integrated circuit (ASIC), and the circuit system can further improve its efficiency by at least one order of magnitude with its reconfigurability maintained, which makes ASIC another mainstream ADAS solution. Suppliers including Qualcomm, Intel, Infineon, and Texas Instruments have released their ASIC-based SoC products for ADAS. One representative product is Intel Mobileyes EyeQ4 chip [18], which will be released in 2018 and can get a 2.5 TOPS performance drawing only 3-5 W. The low power feature makes it quite suitable for in-car supercomputing.

Both the chances and challenges for reconfigurable in-car systems lie ahead. This article will firstly analyze the development of modern RTAV algorithms, then evaluate the performance of each hardware platform, and finally discuss how we can build a more efficient reconfigurable system for RTAV.

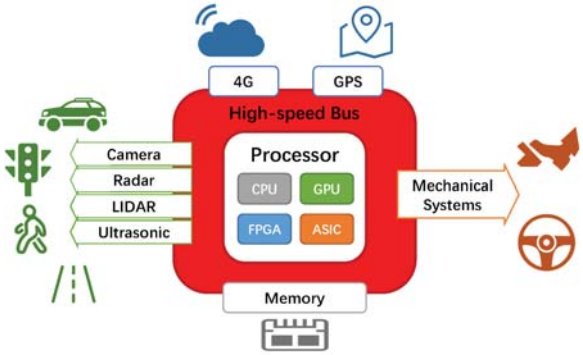


Figure 2. A block diagram for ADAS system description.

2. TRENDS IN AUTONOMOUS VISION

2.1. An overview of an ADAS system

An ADAS system collects data of the surrounding environment from sensors and remote terminals such as cloud servers and satellites, and makes real-time recognition of surrounding objects to assist drivers or automatically make judgements for a better driving experience and avoid potential accidents. A typical system is depicted in Fig. 2. As we can see, there could be a series of sensors on vehicles such as cameras, radars, LIDAR and ultrasonics to get input for a real-time surrounding condition description, and processors will react to give driver warnings or control the mechanical system of the vehicle in some certain circumstances with the trained algorithm models stored in the memory. Communication interfaces can help to locate cars with map data, and can obtain traffic information from datacenters and even offload some compute-intensive tasks to cloud servers, and this can be even more powerful in the future as much faster communication protocols like 5G is already on the way.

Various functions can be achieved with an equipped ADAS system, and autonomous vision has taken up a great portion of this. As we can see from Fig. 3, functions such as vehicle detection (VD), lane departure warning (LDW), forward collision warning (FCW), pedestrian detection (PED), traffic sign recognition (TSR), etc. are achieved by the autonomous vision system itself or together with audio and radar systems. Hence, it is important to find an efficient solution for autonomous vision processing. Next, we will take an overview of the vision algorithms, and present an analysis of potential hardware carriers.

2.2. Traditional algorithms of autonomous vision

For most autonomous vision functions such as PED, VD, LDW, TSR, etc., the kernel algorithm can be generalized into a 2D object detection question. As shown in Fig. 4, a traditional detection process consists of the following stages: image preprocessing, region of interest (ROI) selection, feature extraction and classification.

For traditional algorithms, usually steps like gain and expo-

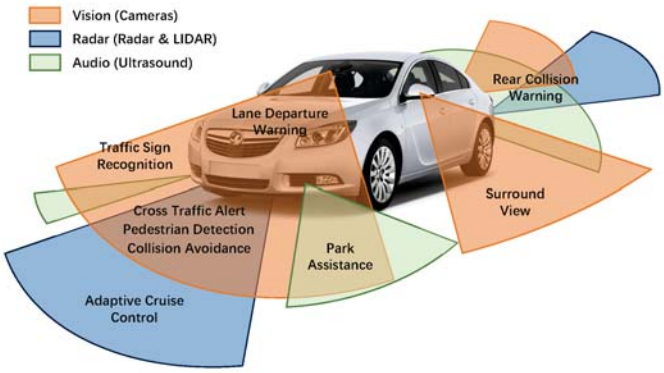


Figure 3. Common functions in ADAS system.

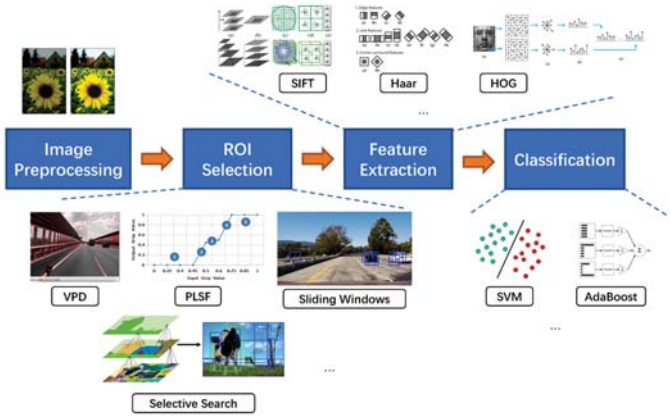


Figure 4. Workflow of traditional detection algorithms.

sure adjustment and image rectification would be performed to preprocess the collected images. ROI selection methods depend on the type of task, such as vanishing point detection (VPD) [19] and piecewise linear stretching function (PLSF) [20] are used in LDW, and sliding window methods are taken in PED, VD and TSR. It would be time consuming to execute an exhaustive ROI search, so various optimizations are also taken for ROI selection. Broggi et al. [21] use morphological characteristics of objects and distance information. Uijlings et al. [22] propose a selective search approach to efficiently generate ROIs. For feature extraction, various manually designed features such as Scale-Invariant-Feature-Transform (SIFT) [23], Histogram-of-Oriented-Gradients (HOG) [24], Haar [25], etc. have been widely used in detection tasks. For classification, combined simple classifiers like AdaBoost [26] and support vector machines (SVMs) [27] are popular to work with traditional features. Some part based methodologies also appear to reduce the complexity of the overall task, such as Felzenszwalb et al. [28] proposes a deformable part model (DPM) to break down the objects into simple parts.

2.3. The rise of convolutional neural network (CNN)

In recent years, the rise of CNN has set off a revolution in the area of object detection. A typical CNN consists of a

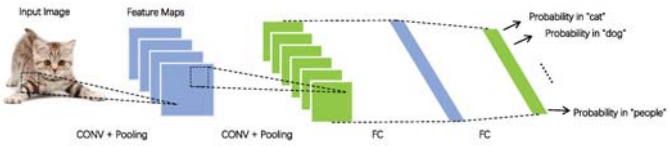


Figure 5. A typical CNN architecture.

number of layers that run in sequence as shown in Figure 5. Convolutional layer (CONV layer) and fully-connected layer (FC layer) are two essential types of layer in CNN, followed by optional layers such as pooling layers for down-sampling and normalization layers. The first CONV layer takes an input image and outputs a series of feature maps, and the following CONV layers will extract features to higher levels layer by layer through convolving the input feature maps with filters. After CONV layers, FC layers will classify the extracted features and output the probability of each category that the input image might belong to.

State-of-the-art CNN models have achieved outstanding performance in computer vision areas. Take image classification as example, in 2012 Krizhevsky et al. announced an 8-layer CNN model AlexNet [29] which achieved 84.7% top-5 accuracy on ImageNet [30], which was far beyond the performance of conventional algorithms. Five years have passed, many organizations such as Google [31][32][33][34], Oxford [35], Microsoft [36] have been focusing on novel CNN model designs with more complex computing patterns, and the accuracies of the top models have already surpassed the human vision level [37].

The excellent performance of CNN is because the generic descriptor extracted from CNN that trained on large scale datasets is much richer than the traditional manually designed features, and can be used for various tasks with some fine tuning [38]. Hence for object detection problems, CNN-based algorithms can get a much better performance than the traditional ones.

The workflows of different detection algorithms are shown in Fig. 6. R-CNN was first proposed [39]. It generates a set of region proposals with selective search, warp/crop each region into a static size, then extracts the feature maps with CONV layers, and finally completes the classification with FC and SVM layers. Since R-CNN needs to run CONV layers for every region proposal which is very expensive in computations, SPP-net has appeared [40]. It merely needs to compute CONV layers only once with spatial pyramid pooling to transfer feature maps into fixed length vectors for FC layers. Based on SPP-net, Fast R-CNN was designed by Girshick et al. [41] which used multi-task loss to train the classifier and bounding-box (BB) localizers jointly, with single-sized ROI pooling to the feature maps of the last CONV layer which are projected with region proposals. Then Ren et al. [42] proposed Faster R-CNN, using the region proposal network (RPN), which was actually a Fast R-CNN network, to generate region proposals and to get rid of the large computations of traditional region proposal methods, and reused the Fast

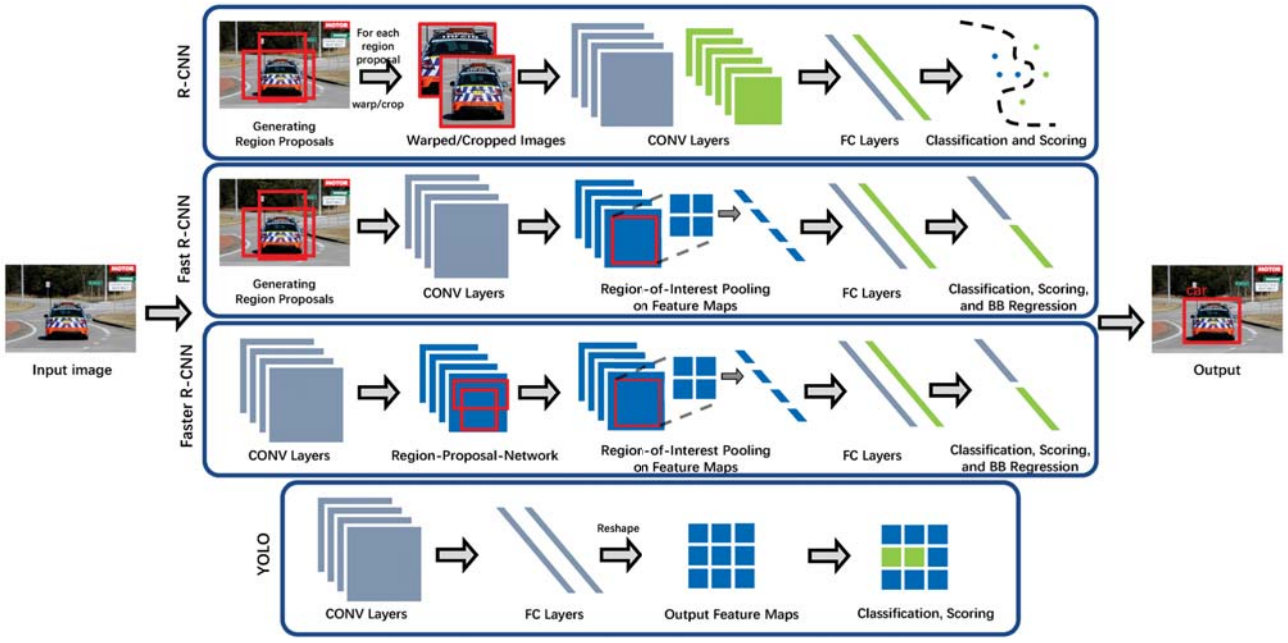


Figure 6. The processing flow of typical CNN-based detection methods.

R-CNN model to train the classifier and BB localizers. Unlike the former algorithms which could only get satisfying mean Average Precision (mAP) performance with the weakness of slow speed, Faster R-CNN can achieve real-time processing since it benefits from RPN and can get a 5fps speed with one NVIDIA K40 GPU. Redmon et al. designed YOLO [43] which directly took the whole input images to train the model, and classifies each pixel in the output feature maps. This equals to dividing the input image into several cells and doing the classification inside each cell, which avoids the expensive process for proposals and can be around seven times faster than Faster R-CNN to realize a more feasible real-time detection with acceptable accuracy drop.

These detection algorithms have shown outstanding performance on a PASCAL VOC dataset [44]. However, for the autonomous vision scene, the detection mission would be much tougher since the objects will be presented in much worse quality for the big variance of object scale and the incomplete captured object shape. Therefore, we need to optimize the way we obtain proposals during our detection algorithms. The corresponding representative benchmark for autonomous vision is KITTI [45], and various algorithms have been proposed for the dataset. We have selected some top ranked detection algorithms and have listed them in Table. 1. Actually, most of these algorithms have taken CONV layers to extract the features based on the classic CNN models with small revisions followed by application dependent FC layers. We compare the CONV layers of classic CNN models in Table. 2. As we can see, giga MACs need to be solved for each input frame. Together with FC layers and considering the number of region proposals, in order to realize real-time processing, the hardware needs to provide a throughput speed of over 100-1000 GOPS. With the growing number of image data collected from cameras, future requirement of computing speed could reach 10-100 TOPS. To build such a powerful processor with programmability and a power consumption of less than 30 W is a challenging task, and we will discuss the contenders in the next section.

Table 1. Top-ranked detection algorithms on KITTI.

Algorithm	Target object (Moderate level)		
	Car	Pedestrian	Cyclist
MS-CNN [46]	89.02%	73.70%	75.46%
SubCNN [47]	89.04%	71.33%	71.06%
SDP+RPN [48]	88.85%	70.16%	73.74%
3DOP [49]	88.64%	67.47%	68.94%
Mono3D [50]	88.66%	66.68%	66.36%
SDP+CRC [48]	83.53%	64.19%	61.31%
Faster R-CNN [42]	81.84%	65.90%	63.35%

Table 2. Comparison of CONV layers in classic CNN models.

Model	AlexNet [29]	VGG-16 [35]	Inception v1 [31]	ResNet-50 [36]
Top-5 Error	19.8%	8.8%	10.7%	7.0%
# of Weights	2.3M	14.7M	6.0M	23.5M
# of MACs	666M	15.3G	1.43G	3.86G

3. PROCESSORS FOR REAL-TIME AUTONOMOUS VISION

3.1. Heterogeneous platforms for CNN acceleration

As the CNN algorithm rapidly develops, so have the related hardware accelerator designs, in recent years. The work of

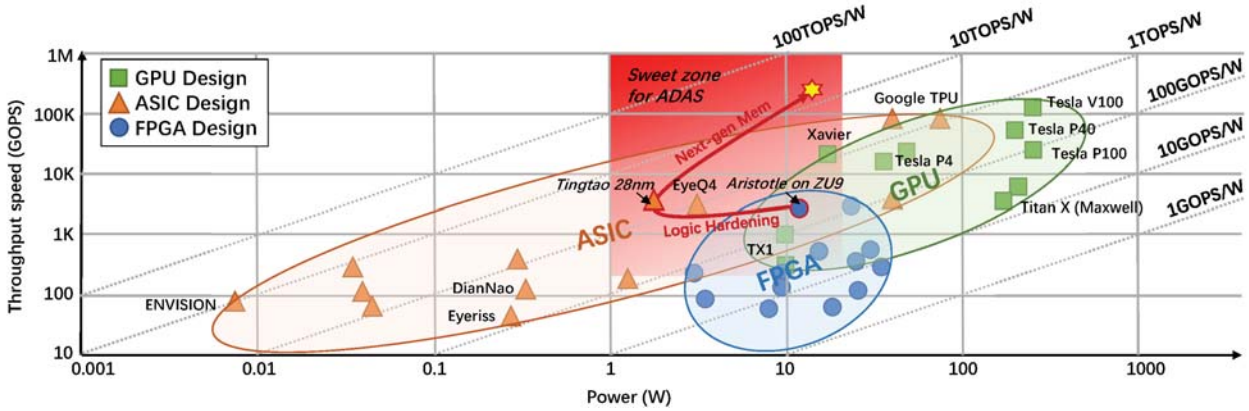


Figure 7. Hardware designs of CNN accelerators on different platforms and development route for RTAV accelerator in ADAS.

(Source by:https://nicsefc.ee.tsinghua.edu.cn/projects/neural-network-accelerator/)

[51] shows the comparison between neural network accelerators, as depicted in Fig. 7. We can see from the image that GPUs are among the top tier of computing speeds, but the power consumption is also very high. The freshly released NVIDIA Tesla V100 can get an astounding computing speed of 120 TOPS [52], with a power consumption of 300 W. This can be useful in datacenter scenarios for cases like model training where power is not the main concern. There are also some GPUs designed for low-power embedded environments, like NVIDIA Jetson TX1 mobile GPU, which brings a 300 GOPS speed on VGG-16 and a peak performance of 1 TOPS with only a 10 W cost [53]. The large general purpose stream processors on chip might bring a considerable parallelism, but the efficiency remains a question. With the technology of 28nm, the NVIDIA TITAN-X and TX1 GPU can only get an efficiency of 20-100 GOPS/W.

To improve the efficiency, we need to customize the inside logic of processing elements (PEs) to enhance processing parallelism and optimize memory access patterns. FPGA could be a suitable initial selection, since it can provide a large amount of computing and memory resources and enough reconfigurability with programmable interconnection to map common algorithms on. In Fig. 7 we can see that there have been many FPGA designs. The top designs, including our Aristotle system on the Xilinx ZU9 FPGA platform, can get a throughput speed at around 2 TOPS, which is quite close to the same technology generation NVIDIA TITAN-X GPUs, but of almost 10 times better efficiency. This proves the capability of FPGA of being a strong competitor.

As we can see, most CNN layers consist of MAC operations and have similar computing patterns which could be possibly generalized and parameterized. Therefore, with mature hardware architecture and processing flow, it is feasible to harden the original FPGA accelerator design into an ASIC chip with a programmable interface for reconfigurability, which can further improve performance. Kuon et al. [54] have measured the performance gap between FPGA and ASIC. It is said that the critical-path delay of ASIC is three to four times less than FPGAs, and the dynamic power consumption ratio is approximately 14 for FPGA to ASIC, while the average chip area of ASIC is also 18 times smaller than FPGA. This means we can realize a much better performance with ASIC within a given hardware area. ASIC designs have the relatively better energy efficiency, mostly between 100 GOPS/W to 10 TOPS/W. They have shown excellent performance in low-power area, and as we can see from Fig. 7 some representative designs such as DianNao [55], Eyeriss [56] and Envision [57] are showing a performance of around 100 GOPS with only milli-watt level power consumption. The efficiency can even reach 10 TOPS/W at extreme low voltage status. To the other side, those ASICs with larger chip sizes are capable of offering more abundant PEs and memory bandwidth, which can lead to a faster throughput speed, such as Googles TPU [58] which can get a peak performance of 86 TOPS. From the business aspect, a large quantity production of ASIC could also reduce the overall cost. However, note that the deep-learning algorithms for RTAV have a quite short evolving cycle, usually within six to nine months. Moreover, the benchmarks for RTAV are also far from perfect and new tasks appear nearly every year. While ASICs time to market is no less than one year, there is a potential risk of incompatibility between hardware processors and fresh algorithms and application scenes. Solution providers need to make a risk-return analysis.

Recently, some breakthroughs have taken place in the area of near-memory and in-memory computing. The 3-D memory can offer an order of magnitude higher bandwidth and several times power consumption than 2-D memory, such as Hyper Memory Cube (HMC) proposed by Micron [59], which uses through silicon vias (TSV) to stack the dynamic random-access memory (DRAM) on top of the logic circuit. Through this method, the memory bandwidth can be increased by an order of magnitude from 2-D memory, and the power consumption can be five times less. There have already been some designs combining the CNN accelerator architecture with HMC [60][61]. Another technology is to embed the computation inside memory, such as memristor

[62]. It can realize a MAC operation through the summation of currents from different memristor branches. This avoids the data movement and can save energy. Recent simulation works such as ISAAC [63] and PRIME [64] have evaluated the efficiency of memristors in CNN acceleration.

An ideal ADAS system should be able to offer a computing speed of over 200 GOPS with no more than 40 W, and hence we can mark the sweet zone for ADAS systems as the red painted area in Fig. 7. Inside this sweet zone, we can sketch a development route for the reconfigurable processors for RTAV acceleration, shown as the dark red curve. Starting from the FPGA design, we can climb up through logic hardening for an efficiency of above 1 TOPS/W, and with the help of the implementation of next generation memory technology, the bandwidth can be broaden and the memory access cost could be reduced, which can lead to an even higher efficiency, to more than 10 TOPS/W. We use the yellow star to indicate our target in Fig. 7. With a larger die size, a throughput speed of over 100 TOPS could be expected, which can be a suitable choice for an ideal RTAV system.

3.2. Chances and challenges for reconfigurable processors

In the area of RTAV, chances and challenges coexist for a wide application of reconfigurable processors. The following features of reconfigurable processors will bring them opportunities:

1) **Programmability**. Reconfigurable processors can offer a pool of logic and memory resources on-chip. Considering the fast evolving RTAV algorithms, it is not hard for users to update the on-chip functions after they bought it from the supplier.

2) **Reliability**. For example, the industrial grade FPGAs can stably work in a temperature range between $-40°C \sim 100°C$. This makes them able to satisfy the requirement of standards AEC-Q100 and ISO 26262.

3) **Low-power**. The power consumption for reconfigurable processors is no more than 30 W. Low-power consumption is suitable for the in-car environment.

4) **Low-latency**. Since algorithms mapped onto reconfigurable processors provide deterministic timing, they can offer a latency of several nanoseconds, which is one order of magnitude faster than GPUs. A quick reaction of ADAS systems is essential to dealing with sudden changes on the road.

5) **Interfaces**. Unlike GPU which can only make communication through the PCI Express protocol, both ASIC and FPGA designs can provide huge interface flexibility, which can be very helpful for ADAS system integration.

6) **Customizable logic**. Recently there has been great progress in the area of model compression, including data quantization and sparsity exploration. For general purpose processors like CPU and GPU, only fixed data types could be supported and the memory access pattern would be regular. Reconfigurable processors can offer fine-grained customizability which can support data type as low as to 1 bit, and specialized controllers could be introduced to deal with irregular sparsity inside the models.

7) **Multi-thread processing**. For ADAS systems, it would be best for different algorithms to be processed simultaneously, such as LDW would work on grayscale images while PD would process RGB images. Reconfigurable processors can provide vast spatial parallelism for algorithms to work in individual channels.

However, challenges remain for the wide use of reconfigurable processors such as:

1) **Programming language gap**: Most developers use high-level programming languages to build their project, while for reconfigurable processors they need to start from the bottom-level hardware and describe the logic with register-transfer level (RTL) hardware description language (HDL) such as Verilog and VHDL.

2) **Limited on-chip resource**: There is limited area for on-chip arithmetic and memory resource to map the tiled algorithm on spatially. This might form a bottleneck for some large-scale algorithms.

3) **Limited off-chip bandwidth**: To communicate reconfigurable processors with off-chip memories like DDR, the bandwidth is often limited by the clock frequency of the controller and the width of data wires.

3.3. Related reconfigurable processors

There have been many excellent reconfigurable processor designs for deep learning models. Initial designs are mostly based on FPGAs. Chakaradhar et al. [65] proposed a runtime reconfigurable architecture for CNN on FPGA with dedicated switches to deal with different CNN layers. Zhang et al. [66] used a nested loop model to describe CNN and designed the on-chip architecture based on high-level synthesis optimizations. Suda et al. [67] presented an OpenCL-based FPGA accelerator with fully-connected layers also implemented on-chip.

ASIC-based reconfigurable processors have appeared in recent years. The representative work is Diannao [55] and its subsequent series [68][69][70], which focused great efforts on memory system optimization. Eyeriss [56] focused on the dataflow optimization and used smaller PEs to form a coarse-grained computing array. ENVISION [57] utilized a dynamic-voltage-accuracy-frequency-scaling (DVAFS) method to enhance its efficiency and reached 10 TOPS/W with low voltage supply. Googles TPU [58] has been the recent star with large on-chip memories and has reached a similar throughput speed to peer GPUs withdrawing much less energy.

Most of these precedent reconfigurable processors have their own features with partial optimization of the entire flow, but few consider the entire flow of the neural network accelerator system. Therefore, the on-chip utilization rate of different CNN layers will eventually fluctuate [58] which may drag down the overall efficiency of the system, and there has been a large space left for improvement from the aspect of

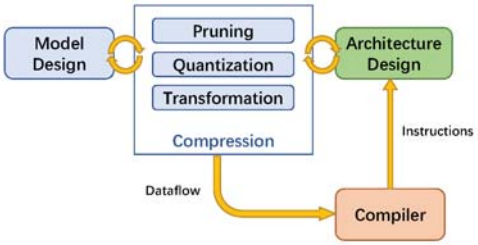


Figure 8. The software-hardware co-design workflow of our system.

software. With this motivation, we will introduce our system design in the following section.

4. SOFTWARE-HARDWARE CO-DESIGN FOR A RECONFIGURABLE AUTONOMOUS VISION SYSTEM

4.1. The overall system workflow

What we have already achieved is an FPGA-based system called *Aristotle* to target CNN acceleration, which can deal with various CNN-based applications and can be conveniently mapped onto different FPGA platforms. For a better processing performance, we should reduce the software workload and improve the hardware utilization rate. Accordingly, we design the software hardware co-design workflow of our Aristotle system depicted in Fig. 8. To reduce the workload, we compress the models using software methods like quantization, pruning and matrix transformation. To improve the utilization rate, the compiler will take the compressed model and hardware parameters of different FPGA platforms as inputs, and execute a task tiling with dataflow optimizations to generate instructions for the hardware. The hardware architecture will exploit the parallelism on-chip for higher throughput with proper granularity choice and datapath reuse. The details will be introduced as follows.

4.2. Compression methods

Usually, an algorithm model is trained in floating-point form, but there exists redundancy. Previous work has shown that it is not necessary to represent every datum with 32-bit, and an appropriate data quantization would not hurt the overall accuracy of the model. In Fig. 9 we have made an experiment of quantization on state-of-the-art CNN models, and as we can see from an 8-bit quantization brings little loss to the accuracy. A lower bit-width can directly compress the size of memory footprint, and can bring chance to share datapath consists of integrated DSP blocks. We can implement two multipliers for 8-bit inputs with one 25×18 DSP block on Xilinx platform.

Another method is to implement a pruning process to the pre-trained model, in order to decrease the number of connections inside a model [71]. It has been proved that some of the connections that have weights close to zero will make a small impact on the output pixel, and can be pruned without much

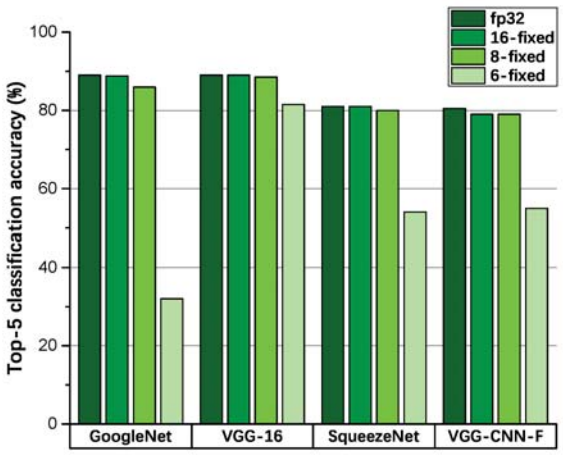


Figure 9. Quantization results for different CNN models.

Table 3. Comparison of compression ratio between quantization, pruning and matrix transformation methods at different accuracy loss levels (baseline 32-bit floating-point).

Accuracy Loss	SVD	Quantization Only	Pruning Only	Quantization and Pruning
0%	-	5.8x	10.3x	27.0x
1%	5.4x	14.1x	15.6x	35.7x
2%	6.5x	14.9x	19.1x	37.0x
4%	6.9x	15.4x	22.9x	37.7x

loss and the loss can be further healed by retraining. Table 3 has shown that if we combine pruning and quantization together, the compressed model size would be the smallest with negligible accuracy loss. Together with Huffman coding, the model size of AlexNet can be reduced by 35 times, and that of VGG-16 can be reduced by 49 times. We should notice the randomness of sparsity from pruning, which is tough to be efficiently used for hardware execution. To deal with this case, we add some constraints to limit the pruned connections in regular patterns, and this can increase the number of all zero channels for more skips during the acceleration process.

Moreover, we can see that inside the basic MAC operations of CNN, multiplication is always the most resource consuming operation, so reducing the number of multiplications can also enhance the hardware performance. Matrix transformation like Winograd [72] and FFT [73] can achieve this goal by targeting different sizes of filters. Take Winograd transformation as example, if we tile the input feature maps into 6×6 blocks and convolve it with 3×3 filters, through transformation we can reduce the number of multiplications by 2.25 times and replace them with cheap add and shifting operations.

With all these compression methods above, we can reduce the workload of the original model, which will benefit the on-chip memory and arithmetic resources and system throughput speed.

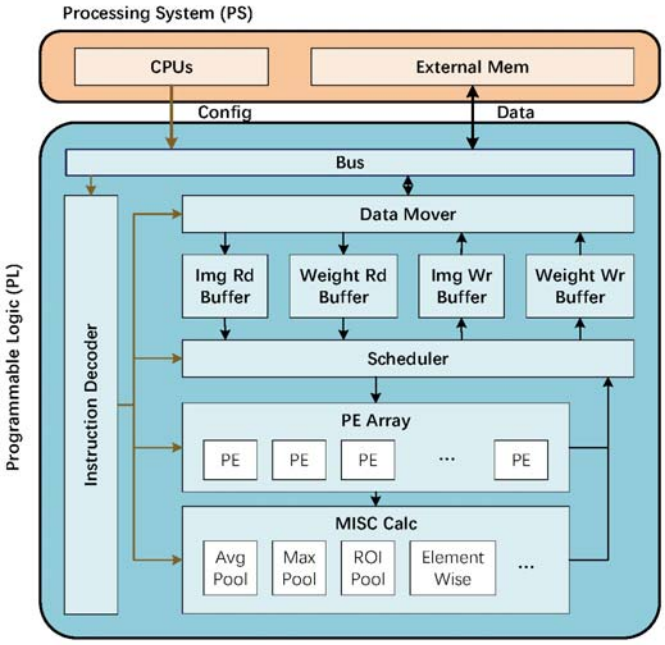


Figure 10. Our CPU+FPGA system architecture.

4.3. Hardware architecture design

Our Aristotle hardware architecture design [74] is given in Fig. 10. A CPU+FPGA accelerator design is adopted, which consists of two parts: the processing system (PS) and the programmable logic (PL). PS contains the low-power CPU processors and the external memory, which offers programmability and data capacity. Instructions will be transferred into PL and decoded to implement the control of PL. PL is the on-chip design where the majority of the CNN accelerator logic is located, and can be scalable due to the chosen FPGA platform. PEs are placed inside PL for parallel MAC operations, which can complete the convolving process through multiple iterations. Some functions that cannot be efficiently accelerated with PE, such as several kinds of pooling and an element-wise dot product, will be contained inside a MISC calculation pool for optional use. On-chip buffers will be provided to offer data reuse opportunities controlled by a scheduler, and communicate with external memories using a data mover such as a direct memory access controller (DMAC). Such hardware architecture design can be easily shared between layers which are friendly to instruction generation and high-level programming.

Instead of combining every multiplication of one filter window together, we split the computing kernel into smaller granularity, which can avoid the waste of arithmetic resource while dealing with a large filter size or window stride, and can ensure a regular data access pattern for easier control. Furthermore, a smaller granularity of PE can increase the chance of skipping for sparsity, which can save the overall execution time of the system.

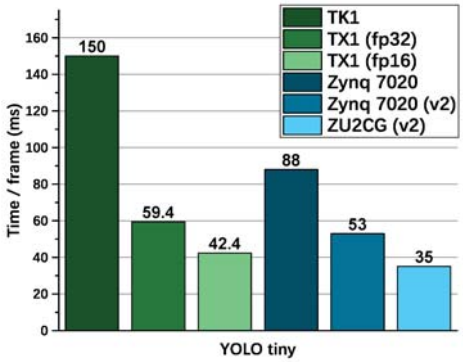


Figure 11. Evaluation results of YOLO-tiny on mobile GPUs and different FPGA platforms.

Table 4. Evaluation results of SSD on CPU, GPU and FPGA platforms.

Platform	Intel Xeon E5-2640 v4	NVIDIA GTX 1080TI GPU	Xilinx ZU9 FPGA	
Task	SSD (YOLO)			SSD (YOLO) Pruned
Operations (GOPs)	16.6			7.4
fps	4.88	183.48	9.09	20.00
Power (W)	90	250	14	
Efficiency (fps/W)	0.054	0.734	0.649	1.429

4.4. Performance evaluation

We use the YOLO algorithm to evaluate our Aristotle system, which is the most popular real-time detection algorithm in the RTAV area. Fig. 11 shows the comparison of performance on different platforms. We can see that compared with the same level mobile GPU platforms our system can reach a similar performance. However, the power consumption of our Zynq-7020 and ZU2 based systems are around 3 W, while the power of GPU is 15 W. Moreover, the peak performance of TK1 is 326 GOPS and that of TX1 is 1 TOPS, while the peak performance of our FPGA platforms is only around 100 GOPS. These can prove a much better efficiency of our system design.

We also use the YOLO version SSD [75] algorithm to compare our larger FPGA systems with CPUs and GPUs. SSD is an optimized algorithm based on YOLO with multi-scale feature extractions which can improve the ability to capture small objects. Table. 4 lists the results on different platforms. We can see that both GPU and FPGA solutions can reach a faster performance than the Intel Xeon CPU. The power consumption of the NVIDIA GTX 1080TI GPU can get up to 250 W, while the value of FPGA is only 14 W. From the perspective of efficiency, with the pruning method implemented, our design can get an efficiency almost twice that of 1080TI GPU.

Furthermore, we have tested a Densebox [76] model on our

Table 5. Evaluation results of Densebox on GPU and FPGA platforms.

Platform	NVIDIA GTX 1080TI GPU		Xilinx ZU9 FPGA
Input Size	640x360		
Task	Densebox	Densebox Pruned	
Operations (GOPs)	28	1.2	
fps	150	330	300
Power (W)	250		14
Efficiency (fps/W)	0.60	1.32	21.43
Recall	0.875		

platform and a peer GPU. Densebox is an end-to-end fully convolutional network (FCN) which has been widely used in face detection applications, and face detection is an essential part of the in-vehicle driver status recognition, such as drowsiness detection. We have pruned the model with the method mentioned in clause 4.2 from 28 GOPs to 1.2 GOPs, with the recall rate staying the same. Table. 5 shows that with the help of pruning, our ZU9-based platform can reach twice the speed of the 1080TI GPU. The GPU can also get a 330 fps with the pruned model, but the utilization rate of model sparsity is quite low considering the peak performance of 1080TI is almost 10.6 TOPS, which results in an efficiency which is 16 times worse than our ZU9 FPGA, reflecting the fit between our compression methods and our hardware system.

4.5. Tingtao: an ASIC-based reconfigurable accelerator

Our ASIC-based reconfigurable accelerator Tingtao is already on schedule. The PS of Tingtao is an ARM Cortex-A5 processor, and the PL includes two deep-learning processing units (DPUs), each containing 2048 MAC PEs and works at 500MHz. Some necessary interfaces for RTAV application are also integrated. Tingtao has taken a 28nm CMOS technology and is projected to provide a peak performance of 4 TOPS at a power of 3 W, which is slightly better than the EyeQ4 product. With the compression method and compiling optimization introduced, the performance could get even better. As shown in Fig. 7, Tingtao has filled the sparse area of 1 to 10 W of power and TOPS level throughput. We are also planning to try a larger design for our next version, and we will pay efforts in the ongoing research of the implementation of emerging memory technology based on our precedent work [64] for the target of our development route.

5. CONCLUSION

This article has reviewed the algorithms for RTAV applications of ADAS, a comparative analysis has been done over different types of platforms, and an enumeration of chances and challenges for reconfigurable RTAV platforms. We have introduced the software-hardware co-design workflow for our reconfigurable RTAV system, with detailed hardware architecture design and implemented compression methods, which ensure an efficient execution with programmability. Evaluation shows that our system can get the best efficiency among peer processors with a satisfying real-time processing performance. An ASIC-based solution can further exploit the efficiency, which means a similar throughput speed with the FPGA-based Aristotle system and an energy cost of one order of magnitude less.

There are some other deep learning models utilized in RTAV applications. Recurrent neural network (RNN) is one of them, and the long-short term memory (LSTM) model [77] shows excellent performance in classifying, processing and predicting time series. This feature can be helpful for object tracking and action predicting functions in ADAS systems. We have not expanded on this topic in this article, but we have already released a similar design based on our Aristotle system framework [78], which has proved the capability of processing various deep learning models.

Future RTAV processors need to offer a 10-100 TOPS throughput speed with less than 30 W power, and to realize this we could count on the rapid development of workload compression such as extreme low-bitwidth CNNs [79][80][81][82] and novel pruning ideas [83][84], hardware design such as dataflow optimization [85][86] and sparsity supported architecture [87][88], and emerging memory technology implementation [60][89]. We are confident that with all those mentioned above, the reconfigurable products will thrive in the ADAS market.

REFERENCES

[1] E. D. Dickmanns and V. Graefe, “Dynamic monocular machine vision,” *Machine vision and applications*, vol. 1, no. 4, pp. 223–240, 1988.

[2] C. Thorpe, M. H. Hebert, T. Kanade, and S. A. Shafer, “Vision and navigation for the carnegie-mellon navlab,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 10, no. 3, pp. 362–373, 1988.

[3] E. D. Dickmanns and B. D. Mysliwetz, “Recursive 3-d road and relative ego-state recognition,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 14, no. 2, pp. 199–213, 1992.

[4] J. Manyika, M. Chui, J. Bughin, R. Dobbs, P. Bisson, and A. Marrs, *Disruptive technologies: Advances that will transform life, business, and the global economy*. McKinsey Global Institute San Francisco, CA, 2013, vol. 180.

[5] J. Barbaresso, G. Cordahi, D. Garcia, C. Hill, A. Jendzejec, and K. Wright, “USDOT’s intelligent transportation systems (ITS) ITS strategic plan 2015-2019,” Tech. Rep., 2014.

[6] Cabinet of Japan, “Statement on ”forging the world-leading it nation”,” Tech. Rep., 2013.

[7] Directive 2010/40/EU of the European Parliament and of the council, "Directives on the framework for the deployment of intelligent transport systems in the field of road transport and for interfaces with other modes of transport," Tech. Rep., 2010.

[8] European Comission. Directorate-General for Mobility and Transport, *White Paper on Transport: Roadmap to a Single European Transport Area: Towards a Competitive and Resource-efficient Transport System*. Publications Office of the European Union, 2011.

[9] European Comission, "Preliminary descriptions of research and innovation areas and fields, research and innovation for europe's future mobility," Tech. Rep., 2012.

[10] J. Janai, F. Güney, A. Behl, and A. Geiger, "Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art," *arXiv preprint arXiv:1704.05519*, 2017.

[11] Intel, "Intel atom processor e3900 series." [Online]. Available: https://www.qualcomm.com/solutions/automotive/drive-data-platform

[12] Qualcomm, "Drive data platform." [Online]. Available: https://www.qualcomm.com/solutions/automotive/drive-data-platform

[13] N. Corp., "NVIDIA drive PX - the AI car computer for autonomous driving," 2017. [Online]. Available: http://www.nvidia.com/object/drive-px.html

[14] NVIDIA CUDA, "NVIDIA CUDA C programming guide," *Nvidia Corporation*, vol. 120, no. 18, p. 8, 2011.

[15] S. Chetlur, C. Woolley, P. Vandermersch, J. Cohen, J. Tran, B. Catanzaro, and E. Shelhamer, "cuDNN: Efficient primitives for deep learning," *arXiv preprint arXiv:1410.0759*, 2014.

[16] Xilinx, "Xilinx automotive Zynq-7000." [Online]. Available: https://www.xilinx.com/publications/prod_mktg/ZynqAuto_ProdBrf.pdf

[17] Altera, "A safety methodology for ADAS designs in FPGAs." [Online]. Available: https://www.altera.com/en_US/pdfs/literature/wp/wp-01204-automotive-functional-safety.pdf

[18] Mobileye, "The evolution of EyeQ," 2017. [Online]. Available: https://www.mobileye.com/our-technology/evolution-eyeq-chip/

[19] J. Son, H. Yoo, S. Kim, and K. Sohn, "Real-time illumination invariant lane detection for lane departure warning system," *Expert Systems with Applications*, vol. 42, no. 4, pp. 1816–1824, 2015.

[20] V. Gaikwad and S. Lokhande, "Lane departure identification for advanced driver assistance," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 910–918, 2015.

[21] A. Broggi, M. Bertozzi, A. Fascioli, and M. Sechi, "Shape-based pedestrian detection," in *Intelligent Vehicles Symposium, 2000. IV 2000. Proceedings of the IEEE*. IEEE, 2000, pp. 215–220.

[22] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[23] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[24] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[25] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–I.

[26] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *European conference on computational learning theory*. Springer, 1995, pp. 23–37.

[27] C. Cortes and V. Vapnik, "Support vector machine," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[28] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[32] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

[33] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[34] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning." in *AAAI*, 2017, pp. 4278–4284.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[37] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[38] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2014, pp. 806–813.

[39] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European Conference on Computer Vision*. Springer, 2014, pp. 346–361.

[41] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[42] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[43] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[44] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[45] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3354–3361.

[46] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *European Conference on Computer Vision*. Springer, 2016, pp. 354–370.

[47] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, "Subcategory-aware convolutional neural networks for object proposals and detection," in *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*. IEEE, 2017, pp. 924–933.

[48] F. Yang, W. Choi, and Y. Lin, "Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2129–2137.

[49] X. Chen, K. Kundu, Y. Zhu, A. G. Berneshawi, H. Ma, S. Fidler, and R. Urtasun, "3d object proposals for accurate object class detection," in *Advances in Neural Information Processing Systems*, 2015, pp. 424–432.

[50] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3d object detection for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2147–2156.

[51] NICS Lab of Tsinghua University, "Neural network accelerator inference," 2017. [Online]. Available: https://nicsefc.ee.tsinghua.edu.cn/projects/neural-network-accelerator/

[52] NVIDIA, "NVIDIA Tesla V100," 2017. [Online]. Available: https://www.nvidia.com/en-us/data-center/tesla-v100/

[53] ——, "NVIDIA Jetson - the embedded platform for autonomous everything," 2017. [Online]. Available: http://www.nvidia.com/object/embedded-systems-dev-kits-modules.html

[54] I. Kuon and J. Rose, "Measuring the gap between fpgas and asics," *IEEE Transactions on computer-aided design of integrated circuits and systems*, vol. 26, no. 2, pp. 203–215, 2007.

[55] T. Chen, Z. Du, N. Sun, J. Wang, C. Wu, Y. Chen, and O. Temam, "Diannao: A small-footprint high-throughput accelerator for ubiquitous machine-learning," in *ACM Sigplan Notices*, vol. 49, no. 4. ACM, 2014, pp. 269–284.

[56] Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, 2017.

[57] B. Moons, R. Uytterhoeven, W. Dehaene, and M. Verhelst, "Envision: A 0.26-to-10 tops/w subword-parallel dynamic-voltage-accuracy-frequency-scalable convolutional neural network processor in 28nm fdsoi," in *IEEE International Solid-State Circuits Conference (ISSCC)*, 2017, pp. 246–257.

[58] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers *et al.*, "In-datacenter performance analysis of a tensor processing unit," *arXiv preprint arXiv:1704.04760*, 2017.

[59] J. Jeddeloh and B. Keeth, "Hybrid memory cube new dram architecture increases density and performance," in *VLSI Technology (VLSIT), 2012 Symposium on*. IEEE, 2012, pp. 87–88.

[60] M. Gao, J. Pu, X. Yang, M. Horowitz, and C. Kozyrakis, "Tetris: Scalable and efficient neural network acceleration with 3d memory," in *Proceedings of the Twenty-Second International Conference on Architectural Support for Programming Languages and Operating Systems*. ACM, 2017, pp. 751–764.

[61] D. Kim, J. Kung, S. Chai, S. Yalamanchili, and S. Mukhopadhyay, "Neurocube: A programmable digital neuromorphic architecture with high-density 3d memory," in *Computer Architecture (ISCA), 2016 ACM/IEEE 43rd Annual International Symposium on*. IEEE, 2016, pp. 380–392.

[62] L. Chua, "Memristor-the missing circuit element," *IEEE Transactions on circuit theory*, vol. 18, no. 5, pp. 507–519, 1971.

[63] A. Shafiee, A. Nag, N. Muralimanohar, R. Balasubramonian, J. P. Strachan, M. Hu, R. S. Williams, and V. Srikumar, "Isaac: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars," in *Proceedings of the 43rd International Symposium on Computer Architecture*. IEEE Press, 2016, pp. 14–26.

[64] P. Chi, S. Li, C. Xu, T. Zhang, J. Zhao, Y. Liu, Y. Wang, and Y. Xie, "Prime: A novel processing-in-memory architecture for neural network computation in reram-based main memory," in *Proceedings of the 43rd International Symposium on Computer Architecture*. IEEE Press, 2016, pp. 27–39.

[65] S. Chakradhar, M. Sankaradas, V. Jakkula, and S. Cadambi, "A dynamically configurable coprocessor for convolutional neural networks," in *ACM SIGARCH Computer Architecture News*, vol. 38, no. 3. ACM, 2010, pp. 247–257.

[66] C. Zhang, P. Li, G. Sun, Y. Guan, B. Xiao, and J. Cong, "Optimizing fpga-based accelerator design for deep convolutional neural networks," in *Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*. ACM, 2015, pp. 161–170.

[67] N. Suda, V. Chandra, G. Dasika, A. Mohanty, Y. Ma, S. Vrudhula, J.-s. Seo, and Y. Cao, "Throughput-optimized OpenCL-based FPGA accelerator for large-scale convolutional neural networks," in *Proceedings of the 2016 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*. ACM, 2016, pp. 16–25.

[68] Y. Chen, T. Luo, S. Liu, S. Zhang, L. He, J. Wang, L. Li, T. Chen, Z. Xu, N. Sun *et al.*, "Dadiannao: A machine-learning supercomputer," in *Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture*. IEEE Computer Society, 2014, pp. 609–622.

[69] Z. Du, R. Fasthuber, T. Chen, P. Ienne, L. Li, T. Luo, X. Feng, Y. Chen, and O. Temam, "Shidiannao: Shifting vision processing closer to the sensor," in *ACM SIGARCH Computer Architecture News*, vol. 43, no. 3. ACM, 2015, pp. 92–104.

[70] D. Liu, T. Chen, S. Liu, J. Zhou, S. Zhou, O. Teman, X. Feng, X. Zhou, and Y. Chen, "Pudiannao: A polyvalent machine learning accelerator," in *ACM SIGARCH Computer Architecture News*, vol. 43, no. 1. ACM, 2015, pp. 369–381.

[71] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," *arXiv preprint arXiv:1510.00149*, 2015.

[72] A. Lavin and S. Gray, "Fast algorithms for convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4013–4021.

[73] M. Mathieu, M. Henaff, and Y. LeCun, "Fast training of convolutional networks through ffts," *arXiv preprint arXiv:1312.5851*, 2013.

[74] J. Qiu, J. Wang, S. Yao, K. Guo, B. Li, E. Zhou, J. Yu, T. Tang, N. Xu, S. Song *et al.*, "Going deeper with embedded fpga platform for convolutional neural network," in *Proceedings of the 2016 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*. ACM, 2016, pp. 26–35.

[75] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[76] L. Huang, Y. Yang, Y. Deng, and Y. Yu, "Densebox: Unifying landmark localization with end to end object detection," *arXiv preprint arXiv:1509.04874*, 2015.

[77] J. Schmidhuber and S. Hochreiter, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[78] S. Han, J. Kang, H. Mao, Y. Hu, X. Li, Y. Li, D. Xie, H. Luo, S. Yao, Y. Wang *et al.*, "ESE: Efficient speech recognition engine with sparse LSTM on FPGA." in *FPGA*, 2017, pp. 75–84.

[79] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks," in *Advances in neural information processing systems*, 2016, pp. 4107–4115.

[80] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "Xnor-net: Imagenet classification using binary convolutional neural networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 525–542.

[81] F. Li, B. Zhang, and B. Liu, "Ternary weight networks," *arXiv preprint arXiv:1605.04711*, 2016.

[82] C. Zhu, S. Han, H. Mao, and W. J. Dally, "Trained ternary quantization," *arXiv preprint arXiv:1612.01064*, 2016.

[83] A. Zhou, A. Yao, Y. Guo, L. Xu, and Y. Chen, "Incremental network quantization: Towards lossless cnns with low-precision weights," *arXiv preprint arXiv:1702.03044*, 2017.

[84] V. Sze, T.-J. Yang, and Y.-H. Chen, "Designing energy-efficient convolutional neural networks using energy-aware pruning," 2017.

[85] M. Alwani, H. Chen, M. Ferdman, and P. Milder, "Fused-layer cnn accelerators," in *Microarchitecture (MICRO), 2016 49th Annual IEEE/ACM International Symposium on*. IEEE, 2016, pp. 1–12.

[86] F. Tu, S. Yin, P. Ouyang, S. Tang, L. Liu, and S. Wei, "Deep convolutional neural network architecture with reconfigurable computation patterns," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2017.

[87] S. Zhang, Z. Du, L. Zhang, H. Lan, S. Liu, L. Li, Q. Guo, T. Chen, and Y. Chen, "Cambricon-x: An accelerator for sparse neural networks," in *Microarchitecture (MICRO), 2016 49th Annual IEEE/ACM International Symposium on*. IEEE, 2016, pp. 1–12.

[88] A. Parashar, M. Rhu, A. Mukkara, A. Puglielli, R. Venkatesan, B. Khailany, J. Emer, S. W. Keckler, and W. J. Dally, "SCNN: An accelerator for compressed-sparse convolutional neural networks," in *Proceedings of the 44th Annual International Symposium on Computer Architecture*. ACM, 2017, pp. 27–40.

[89] L. Song, X. Qian, H. Li, and Y. Chen, "Pipelayer: A pipelined reram-based accelerator for deep learning," in *High Performance Computer Architecture (HPCA), 2017 IEEE International Symposium on*. IEEE, 2017, pp. 541–552.

REAL-TIME MONITORING OF THE GREAT BARRIER REEF USING INTERNET OF THINGS WITH BIG DATA ANALYTICS

Marimuthu Palaniswami[1], Aravinda S. Rao[1], Scott Bainbridge[2]
[1]The University of Melbourne, Dept. of Electrical and Electronic Engineering Parkville, VIC - 3010, Australia
[2]Australian Institute of Marine Science PMB #3, Townsville, QLD - 4810, Australia

Abstract –The Great Barrier Reef (GBR) of Australia is the largest size of coral reef system on the planet stretching over 2300 kilometers. Coral reefs are experiencing a range of stresses including climate change, which has resulted in episodes of coral bleaching and ocean acidification where increased levels of carbon dioxide from the burning of fossil fuels are reducing the calcification mechanism of corals. In this article, we present a successful application of big data analytics with Internet of Things (IoT)/wireless sensor networks (WSNs) technology to monitor complex marine environments of the GBR. The paper presents a two-tiered IoT/WSN network architecture used to monitor the GBR and the role of artificial intelligence (AI) algorithms with big data analytics to detect events of interest. The case study presents the deployment of a WSN at Heron Island in the southern GBR in 2009. It is shown that we are able to detect Cyclone Hamish patterns as an anomaly using the sensor time series of temperature, pressure and humidity data. The article also gives a perspective of AI algorithms from the viewpoint to monitor, manage and understand complex marine ecosystems. The knowledge obtained from the large-scale implementation of IoT with big data analytics will continue to act as a feedback mechanism for managing a complex system of systems (SoS) in our marine ecosystem.

Keywords – Artificial intelligence, big data analytics, coral bleaching, Internet of Things, wireless sensor networks, real-time monitoring, event detection

1. INTRODUCTION

The Great Barrier Reef (GBR) of Australia consists of 3200 coral reefs extended over 280 000 square km [1]. The GBR has about 900 islands covering 2600 km that include mangrove forests, coastal wetlands and estuaries, deep shoals, seagrass meadows, continental shelf margin and slope [2]. Both economically and ecologically, Australia significantly gains benefits from this geographically-important marine ecosystem. However, the burning of fossil fuels releases carbon dioxide (CO_2), which in turn is absorbed by oceans, resulting in acidification. This process inhibits corals from secreting calcium carbonate exoskeletons [3], reducing (calcification) the reef-building mechanism and associated organisms. Rise in global temperature is also putting more stress on the marine species. Coral bleaching is the process where the relationship between the coral and its symbiotic algae breaks down during rapid changes in sea-water temperature (hot or cold), making corals vulnerable [4].

Anthropogenic activities are attributed to increased stresses on coral reefs as the prominent reason for coral bleaching. Episodes of bleaching at regional scales have been occurring for many decades (prior to the 1980s), but due to a lack of reporting, documentation and understanding, it is difficult to measure the extent of the bleaching effect prior to the 1980s [5]. In 1911, the first thermal bleaching incident was reported at Bird Key Reef in the Florida Keys, where large numbers of corals were injured during abnormally hot and calm weather conditions, killing many fish, Diadema and molluscs [6]. In 1929, a similar bleaching incident was reported at Low Isles on the GBR, killing many corals [7]. The reports of bleaching incidents have grown significantly since 1971, and this has been linked to climate change [8].

The Australian Institute of Marine Science (AIMS) collects environmental data to analyze and address these challenging questions. It is understood that the catastrophic thermal stress might seriously impact the GBR over the next century [8]. As a result, it is imperative that we understand the temperature

patterns and ecological response to mitigate the human-activity-induced stresses [5]. Given the lack of evidence, complex environmental simulation models with detailed characterization are more likely to lead to more uncertainty [9]. The only way to approach this problem is to collect information on the tropical marine environment, assisting to develop more robust models with evidence. However, the challenge here is to collect data at the appropriate spatial and temporal scales [10]. The sensitive environmental dynamics on the GBR necessitates real-time monitoring as a way of managing and understanding anthropogenic stresses effectively.

Internet of Things (IoT)/wireless sensor networks (WSNs) enable real-time, remote sensing at fine spatial and temporal scales of large areas (such as the GBR) [11]. WSNs consist of a network of sensor nodes deployed at multiple, statistically important locations. Sensor nodes are equipped with relevant sensing elements, data processor units, transceiver with antennas, power systems and protective housings [12]. The network of sensor nodes is formed by directing the nodes to communicate with specific nodes in the network. Sensor networks promise to allow data collection at a higher sampling frequency (including finer spatial and temporal scales) while able to keep the cost to a minimum, and provide real-time access to a range of parameters [13]. In addition, visualization of sensor data on a web portal in real time with modeling and simulation results, have clearly changed the approaches to monitoring the GBR.

Artificial intelligence (AI) plays a critical role in analyzing real-time streaming sensor data from such large-scale environments. Given the volume of data received from the sensor nodes, the data needs to be modeled to make a meaningful sense of the data. To extract useful information from the marine system, WSNs/IoT need appropriate network architecture, protocols, communication with AI-based analytics helping to inform end users [14]. Designing such networks requires categorizing sensor networks into different communication models, data delivery models, and network dynamic models. However, technical challenges in implementing such networks include network discovery, control and routing, collaborative signal and information processing, tasking and querying, and security. The role of AI and data analytics, is vital in such situations. AI incorporates several elements of learning, adaptation, evolution and fuzzy logic to intelligently analyze data and create intelligent machines to extract and represent information in a meaningful way [15].

In this article, we present our previous experiences in implementing real-time WSN/IoT for monitoring the GBR. The article focuses on implementation challenges and how AI was used to detect interesting events from the deployed WSN. Clause 2 describes the measures taken to monitor and understand GBR. It also provides the challenges faced in deploying WSN on the GBR. Clause 3 provides the proposed network architecture used in monitoring the GBR using WSN. Clause 4 provides a case study of detecting Cyclone Hamish (that passed through the GBR during March 2009) using a suite of AI algorithms and some of the open challenges in system of systems (SoS) integration with AI. The conclusion of this article is provided in clause 5.

2. THE GREAT BARRIER REEF MONITORING

The GBR is the largest living structure that stretches over 2300 kilometers. It includes 600 types of coral, over 100 jellyfish species, more than 3000 varieties of molluscs, 1625 kinds of fish, 133 types of sharks and rays, and over 30 different types of whales and dolphins. It is also unique as the GBR extends 14 degrees of latitude, including 600 continental islands and about 150 inland mangrove islands [16].

The Great Barrier Reef Ocean Observing System (GBROOS) Project, which is part of the Australian Integrated Marine Observing System (IMOS), has been supported by a special National Collaborative Research Infrastructure Strategy (NCRIS) grant from the Australian Government. GBROOS is an observation system that looks to record the impact of the Coral Sea on the GBR. Specifically, GBROOS aims to provide the observational data to understand the long-term change and impact on the GBR. The GBROOS has five components of monitoring [17]:

1) nine long-term moorings (temperature and salinity profiles, waves and currents)
2) two reference moorings (basic oceanographic parameters)

3) IoT/WSNs on seven islands (reefs) (temperature profiles and weather data)
4) remote sensing using satellites (surface temperature and ocean color data)
5) underway sampling (temperature, salinity, chlorophyll).

WSN/IoT provides real-time sensing of spatially and temporally dense measurements of a range of bio-physical parameters [18]. Without WSN/IoT technology, it is difficult to get such spatially and temporally dense record of bio-physical events, which is what makes the WSN deployment so important. In addition, WSN/IoT significantly improves the access to real-time data covering long time and large-scale geographical areas. WSN/IoT find its use highly important in benthic zones [19], as well as to understand the effect of heat and light on coral bleaching [20]. Furthermore, WSN/IoT data allows us to understand complex ocean processes impacting reefs, and providing detailed environmental information up to the coral bommie (outcrop of coral reef) level [20]. The WSN/IoT data from Lizard Island, Orpheus Island, Rib Reef, Myrmidon Reef, Davies Reef, Heron Island and One Tree Island, coupled with four other complementary sensing components (as listed above), provide a dense environmental information source. The real-time data from the integrated system enables the detection of interesting events and for managers to take immediate action.

Figure 1 shows the IoT/WSN deployment sites at seven locations on the GBR. The harsh marine environments of the GBR poses several challenges in implementing large-scale IoT/WSN and observing the data in real time. The sensor nodes are aware of their spatial locations, providing three dimensional data [spatial position (x,y) and depth (z)] [10]. In the event of unsuccessful transmission by a sensor node, a node could be reconfigured to transmit data to other nodes. The data could then be rerouted to the base station without loss. The implementation challenges include network design, sensor node design with protective casings, floating buoys to house sensor nodes, reliable moorings that can withstand tides, water currents and heavy storms. There is always a chance of sea creatures dismantling the setup either due to curiosity or accidently. The following subclause highlights some of the challenges:

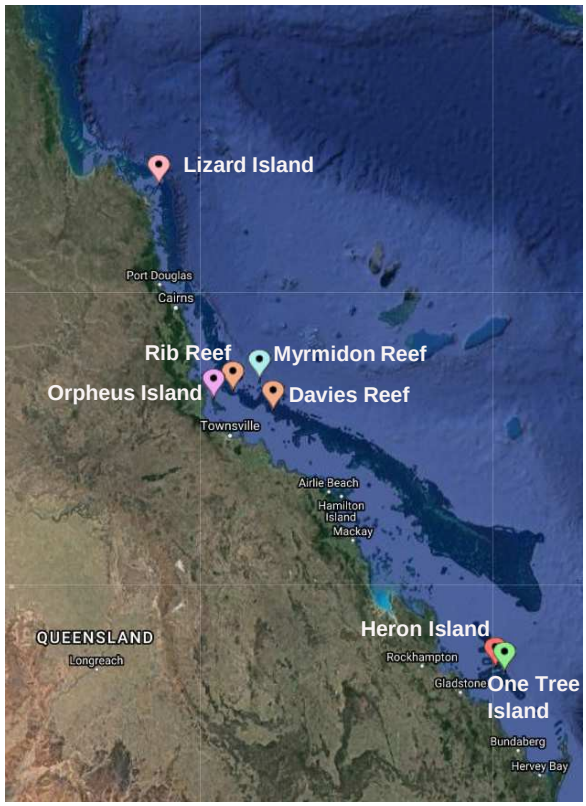


Fig. 1. Map of seven IoT/WSN deployments sites on the GBR. Monitoring of Rib Reef, Myrmidon Reef and Davies Reef have been decommissioned in 2014. Image source: Map Data © 2017 Google Images

2.1. Sensor network and sensing elements

Sensor nodes are resource constrained i.e., they have limited processing power, battery, memory to store and process data. Therefore, the design of sensor networks is application-specific. The architecture design of sensor networks is aimed at maximizing the lifetime of the network at the cost of expending limited resources. These constraints also influence the data sampling times and spatial distribution of sensor nodes. Marine environment is relatively aggressive compared with other environments, requiring specialized sensing elements for continuous monitoring [21]. Marine environment monitoring requires the integration of sensor nodes, such as WSN based iMote2 [22] and IoT-enabled Waspmote [23]. The sensors also need to be calibrated prior to deployment and corrected for drifts in readings from true value over time, as a result of gradually degrading calibration.

2.2. Securing buoys and casing

Sensor nodes need floating buoys to hold the electronics in a secure casing, protection from surrounding environments to avoid water, humidity build-up or condensation. Experience from previous deployments have indicated equipment will foul and corrode [24]. Therefore, utmost care must be taken when deploying sensor nodes in marine environments, given there is a high chance of contact with sea water. The floating buoys are the preferred protective casings for marine environments. However, the buoys introduce several challenges. First, the buoys consisting of sensor nodes need moorings fixed usually to the sea floor using cables. Second, depending on the type of sensing element we may also have to run a long cable with a sensing element, causing deterioration of the measured signal from the sensing element. Third, the buoys drift due to ocean currents and tides, causing radial (vertical) and tangential (horizontal) displacements of sensors nodes. Therefore, these displacements are likely to cause issues with wireless data transmission as the nodes move in a manner that they are unable to be in the communication range of the deployed sensor network.

2.3. Communication and scheduling constraints

Communication in a sensor network can be categorized into two kinds: local coordination and sensor-base communication [25]. Local coordination involves aggregating data among a group of nodes. The sensor-base communication is concerned with communicating the aggregated data to a base station. Both these types of data aggregation could utilize single-hop or multi-hop communication [26]. The data could be logged to a data logger, or collected manually using regular site visits, or using a microwave communication link (such as, the setup at Davies Reef) [27]. Radiocommunication draws most of the battery power. Hence, scheduling of sleep and wake-up cycles to sense environment, store data, and transmit data are critical to prolong the operational time of the network.

2.4. Scalable networking architecture

Marine environments need sensor networks that have flexible architecture to cover spatially small (few meters) as well as large (few kilometer) areas. Therefore, the sensor network architectures need to be scalable. Homogenous sensor networks consist of sensor nodes of the same processing power, radio range, data storage capacity, and networking abilities. These attributes of a homogenous network limit the topology of the sensor network to be flat i.e., it is not scalable in the event of dynamic rerouting or reconfiguration of network nodes. On the other hand, heterogeneous network includes sensor nodes having varied processing, communication, and storage abilities. These attractive features of heterogeneous sensor nodes provide scalable networking solutions through hierarchical architectures. Tenet [28], Tsar [29], SensEye [30], Asap [31], and Citric [32] are some of the examples of heterogeneous networks.

2.5. Detecting interesting events using AI

The network topology of IoT/WSN can be a star, mesh, tree or a combination of all [33]. The way that data moves across the network from one location to another depends on the topology. Thus, event detection can be local or global relative to the network. Local events are specific to the sensor nodes, whereas global events require the sensor data to be gathered at a centralized location. Local event detection requires processing power at the sensor nodes. Often, the computational power, memory and resources are limited in individual nodes. Therefore, simple event detections locally are often relatively easy and will be based on one-dimensional sensor data. For example, detecting changes in temperature level above a certain threshold is an example of a local event. On the other hand, global event detection involves high computational power, memory, sophisticated AI algorithms and network resources. Global events also consider multidimensional sensor data from multiple sensors. For example, detecting an island-wide temperature change is an example of a global event. AI algorithms face challenges due to error in data, data loss, non-generative data models, temporally and spatially incomplete records of physical phenomenon and certain marine artefacts affecting the sensor data. The challenge in IoT/WSN to detect events is to identify anomalous events in a resource-constrained setup.

3. CLOUD-CENTRIC NETWORK ARCHITECTURE FOR REAL-TIME MONITORING

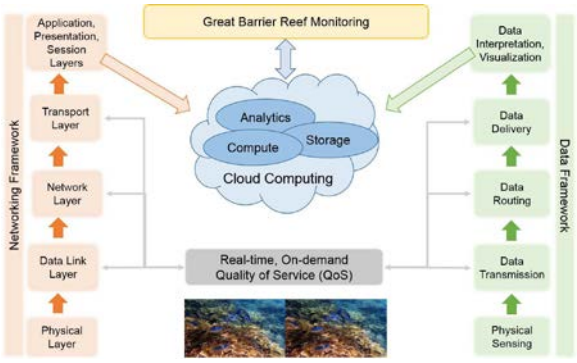


Fig. 2. Proposed IoT/WSN architecture for real-time monitoring, managing and understanding of environment

Figure 2 shows the proposed cloud-centric network architecture for real-time sensing, monitoring and decision-making. The data framework in Fig. 2 illustrates how data is converted when transferring from lower layers of the network architecture to layers above. The data from the end-user applications are managed through a cloud platform. The platform contains computing hardware and software, data storage capacity and AI analytics to service end users on a real-time basis.

3.1. Networking framework

The networking framework shows the networking protocols from physical layer up to the application layer. The layers are conceptually similar to the Open Systems Interconnection (OSI) model. The layers ensure that the control is passed from the layer below to the layer above. The physical layer allows physical transmission of data bits, whereas the data link layer allows transfer of data from one sensor node to another. The network layer determines the path on the network to the correct physical node by managing Internet Protocol (IP) addresses. The transport layer manages end-to-end connections and reliability of network. This is achieved by transferring data across network connections. Transfer Control Protocol (TCP) is an example of a transfer layer protocol.

Figure 2 also shows that the data link layer, network layer and transport layers together are responsible for controlling the quality of service (QoS) demands requested by the users as well as for prioritizing data for real-time applications, such as video [34]. The session layer manages different types of communications between hosts. This operation includes the opening and closing of sessions. The presentation layer manages contexts between applications by handling format conversions, encryption/decryption, independent of application – web content is an example. The application layer is the topmost layer and is used by the end users. It provides services to end-user applications through appropriate networking protocols, such as Hypertext Transfer Protocol (HTTP) and Hypertext Transfer Protocol Secure (HTTPS). Web browser, is an example of application layer that uses HTTP or HTTPS networking protocol for a wide variety of services.

3.2. Data framework

The data framework provides a model of the data flow from physical sensing to data transmission among sensor nodes, routing through different networks, delivery with appropriate encapsulations, interpretation and visualization based on cloud computing and analytics. Sensors attached to the nodes measure physical phenomena and processing boards are programmed to sample at specific sampling rates. Sensor nodes are spatially distributed based on the project plan, cost, application and scientific objectives. The data from the sensors are stored temporarily on board the nodes, before they are transmitted to gateway nodes. The data link layer handles data transmission by sensing the physical medium and channel availability. Once the data is transferred from nodes to gateways, the data will be directed to high-level nodes, such as the cloud servers. Gateways are programmed to direct the data from sensor nodes to cloud databases through IP addresses. The network layer manages the IP addresses and data routing. The delivery of data to a particular database is ensured by the transport layer. Cloud servers manage the storage of received data from sensor nodes and also have high computation compared to sensor nodes. The data is made available to end users through applications from the cloud servers. Servers also manage application sessions among different hosts, as well as allow multiple sessions for the same application from multiple users.

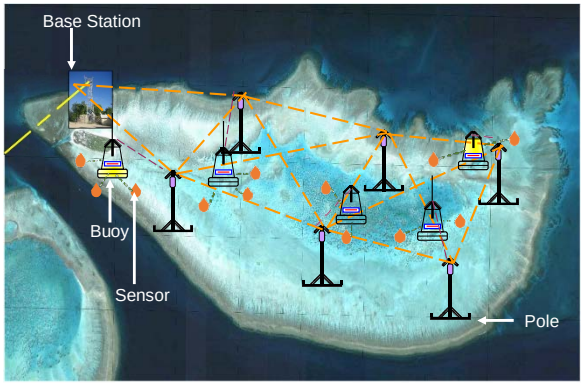


Fig. 3. IoT/WSN deployed at Heron Island of the Great Barrier Reef. Figure shows sensor nodes networked to communicate with buoys, which in turn communicate with poles. Finally, the data from the poles are transmitted to a base station. The base station transmits the data to the mainland that is 75 km away

4. CASE STUDY: DETECTING CYCLONE HAMISH ON HERON ISLAND OF GBR USING AI

In this clause, we provide a case study from the IoT/WSN deployed on Heron Island for real-time monitoring of the GBR and in particular the passage of Tropical Cyclone Hamish in 2009. The Australian Federal Government and Queensland State Government provided funding for the Australian Integrated Marine Observing System (IMOS) [20], which included five components to observe ocean parameters with the IoT/WSN being one of them. By mid-2008 the IoT/WSN had been installed at Heron Island with the other sites completed by 2010.
Figure 3 shows the deployment of heterogeneous sensor nodes with hierarchical network architecture. The network consists of 5 buoys and 6 poles in the lagoon area of the GBR, with a spatial resolution of 2 km. The first tier (top level) consists of poles, followed by floating buoys (as second tier). Sensor nodes are connected to buoys via appropriate cabling. Further, sensor nodes are equipped with temperature probes that measure sea temperatures below the surface. For this case study, we have considered one month of data (collected from 21 February 2009 to 22 March 2009, 9:00am to 3:00pm with 10 minutes sampling frequency) [15, 21, 35].

4.1. WSN network architecture

Figure 4 shows the network architecture of the deployed IoT/WSN on Heron Island. The buoys use single-hop communication to send data to poles, and poles use multi-hop communication to send data to the base station. One of the poles is housed with a weather station, measuring air temperature, pressure, humidity, rain, wind speed and direction. The data from this weather station is collected every 10 minutes. The data received by the main base station is then transmitted to a database that is 75 km away on the Australian mainland using the Telstra 3G network.

4.2. Cyclone Hamish detection using AI

Event detection in the case of Heron Island involved detecting anomalous patterns from the sensor data. The key challenge here is to identify anomalous events in the resource-constrained IoT/WSN setup while achieving high detection accuracy [36]. We approached the problem using our previously established method of detecting elliptical anomalies or Elliptical Summaries Anomaly Detection (ESAD). This is achieved by first modeling the collected data at individual sensor nodes by sample-based ellipsoids and numerically clustering the sets of ellipsoids [36, 37]. Next, a dissimilarity measure of the data is constructed using improved visual assessment of cluster tendency (iVAT) [38]. This step provides us a visual tendency of assessment (VAT) to seek the presence of the number of clusters of ellipsoids in the data. The block within the dissimilarity matrix is reordered using a recursive iVAT algorithm. As a final step, a single linkage algorithm is employed to extract anomalous clusters from the dissimilarity data. Using this AI approach, we were able to clearly identify the pattern of the passage of Tropical Cyclone Hamish appearing as an anomaly before and after the cyclone passed through Heron Island in 2009. The algorithm has since been transformed for real-time applications. Recently, we were able to demonstrate that the Cyclone Hamish event from the WSN data can be detected in real-time IoT settings using our suite of AI algorithms.

4.3. System of systems (SoS) view of integrated AI

Our ecosystem not only consists of environmental applications, but also agriculture, smart city, healthcare, transport, energy and many others. From a global technological ecosystem, we need to have a holistic view of the entire ecosystem to understand and solve the emerging issues. In other words, we need to have an integrated view of the SoS. Figure 5 provides a common operating picture (COP) of a system of systems using cloud-centric AI analytics.

AI algorithms are used to provide application-specific analytics to end users. These are generally provided as a service to end users. The SoS concept goes beyond traditional analytics to provide a complete understanding of the issues to governments, policymakers and decision managers. In this respect, systems are to be integrated in a seamless manner and solve technological issues within and outside the realm of specific solutions.

4.4. Open research challenges

- An estimated 8.4 billion IoT devices are used in 2017 for numerous applications [39]. The challenge lies in SoS integration with AI to process, analyze and generate actionable knowledge. For data originating from billions of devices, it will be nearly impossible to analyze manually. This challenge is attributed to the **big data** challenge. The analysis should include the growth of multidimensional data attributes (i.e. volume, velocity, variety, veracity, variability, and value) to produce actionable knowledge using AI algorithms.

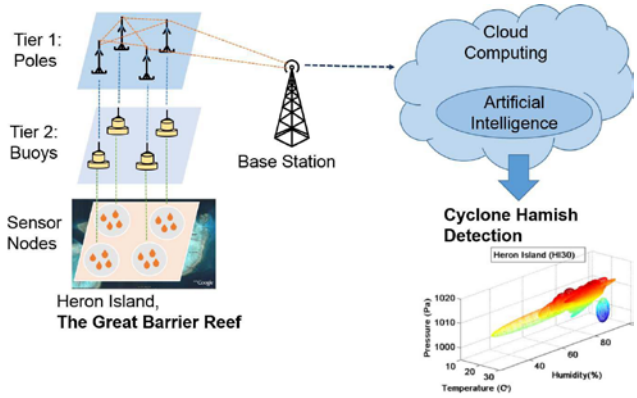


Fig. 4. WSN/IoT system architecture used to monitor Heron Island. The data is collected using a two-tiered (tree) hierarchical network architecture. The data is then transmitted to a central base station, which then transmits it to a mainland server. The cloud server consists of AI analytics that can monitor streaming real-time data and detect events

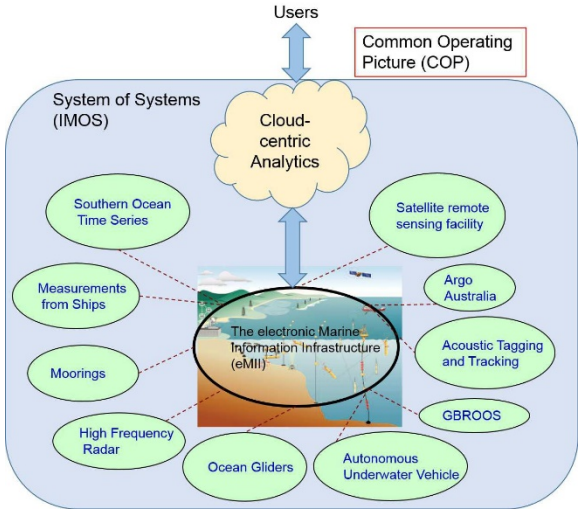


Fig. 5. The illustration depicts the idea of achieving a common operating picture (COP) using system of systems approach. The eMarine Information Infrastructure (eMII) provides a single integrative framework for data and information management that will allow discovery and access of the data by scientists, managers and the public. The data from real-world are then fed to AI-based cloud-centric analytics. The output from the analytics is used to make decisions as well as to provide feedback to the existing systems

- Clustering of data from sensors and IoT devices is an unsupervised task to extract hidden patterns without a priori information. However, with the big data challenges (pointed above), the clustering algorithms need to be **scalable** (algorithms could be used on large volumes of data), **self-tuning** (the AI algorithms should work without any input parameters from end users), **immune to outliers** (eliminate outliers, missing data points, and error points from the sensed data), and **adaptive** (the AI models must adapt the models to newly arrived data points without retraining using all data points).
- The AI algorithms must provide a holistic knowledge of SoS in **real-time**. This requires AI algorithms to be aware of not only the current system where the algorithm resides, but also about dependent and interconnected systems. Currently, most of the AI algorithms are designed to be performed for a specific system, ignoring the implications of data flow and connectedness of dependent systems.
- One of the important challenges in today's IoT/WSN paradigm is the **data security** issue. The algorithms, services and infrastructure face tremendous challenge in maintaining security,

authenticity, trust, privacy and transparency of the data. This also becomes further complicated when citizen-centric data is allowed (such as, the Open Data initiatives from governments) and through crowdsourcing. New methods, such as, block-chain is a possible future solution.

CONCLUSION

The Great Barrier Reef (GBR) of Australia is the largest living structure (coral reef) on the planet and stretches over 2300 kilometers. Anthropogenic stresses on coral reefs are causing coral bleaching. The burning of fossil fuels releases carbon dioxide which in turn is absorbed by the oceans, reducing the efforts of the reef-building mechanism by corals. Therefore, it is necessary to monitor and manage our marine environment as well as to prevent ecosystem collapse. In this article, we presented an overview and a use case of the WSNs/IoT to monitor the complex marine environments, including the GBR. The article presented an architecture used to monitor the GBR as well as the role of AI algorithms to detect events. With a suite of AI algorithms, we were able to detect Cyclone Hamish (which occurred in 2009) patterns using temperature, pressure and humidity sensors using two-tiered IoT/WSN network architecture. The article highlights the role of AI algorithms that could be used to monitor, manage and understand complex marine ecosystems.

ACKNOWLEDGEMENT

Authors acknowledge the support from ARC Research Network on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP) and SensorMap - Microsoft Research RFP Project grant, the Queensland State Government via the Department of Science, Information Technology and Innovation, the National Collaborative Research Infrastructure Strategy (NCRIS) and the Wireless Sensor Network Facility (FAIMMS) of the Integrated Marine Observing System (IMOS). The project is a collaborative effort between The University of Melbourne (ISSNIP research group) and Australian Institute of Marine Science (AIMS).

REFERENCES

[1] Miles MJ Furnas et al., Catchments and corals: terrestrial runoff to the Great Barrier Reef, Australian Institute of Marine Science & CRC Reef Research Centre, 2003.

[2] Robert C Ferrier and Alan Jenkins, Handbook of catchment management, John Wiley & Sons, 2009.

[3] Marc Lallanilla and Staff Writer, "What are coral reefs?," https://www.livescience.com/ 40276-coral-reefs.html, October 2013.

[4] National Ocean Service, National Oceanic and Atmospheric Administration, "What is coral bleaching?," https://oceanservice.noaa.gov/facts/ coral_bleach.html, July 2017.

[5] Ray Berkelmans, Glenn Death, Stuart Kininmonth, and William J Skirving, "A comparison of the 1998 and 2002 coral bleaching events on the great barrier reef: spatial correlation, patterns, and predictions," Coral reefs, vol. 23, no. 1, pp. 74–83, 2004.

[6] Alfred Goldsborough Mayor, The effects of temperature upon tropical marine animals, Carnegie Institution of Washington, 1914.

[7] Charles Maurice Yonge, A Gr Nicholls, and Martha Jane Yonge, Studies on the physiology of corals, vol. 1, British Museum, 1931.

[8] Ove Hoegh-Guldberg, "Climate change, coral bleaching and the future of the world's coral reefs," Marine and freshwater research, vol. 50, no. 8, pp. 839–866, 1999.

[9] Scott Wooldridge and Terry Done, "The use of bayesian belief networks to aid in the understanding and management of large-scale coral bleaching," MODSIM. Modsim, Townsville, pp. 614–620, 2003.

[10] Stuart Kininmonth, Scott Bainbridge, Ian Atkinson, Eric Gill, Laure Barral, and Romain Vidaud, "Sensor networking the great barrier reef," Spatial Sciences Qld Journal, pp. 34–38, 2004.

[11] Ron Johnstone, Davide Caputo, Umberto Cella, Alessandro Gandelli, Cesare Alippi, Francesco Grimaccia, N Haritos, and Riccardo E Zich, "Smart environmental measurement & analysis technologies (se-mat): Wireless sensor networks in the marine environment," Stockholm, Sweden, June 2008.

[12] Jiong Jin, Jayavardhana Gubbi, Slaven Marusic, and Marimuthu Palaniswami, "An information framework for creating a smart city through internet of things," IEEE Internet of Things Journal, vol. 1, no. 2, pp. 112– 121, 2014.

[13] Stuart Kininmonth, "Considerations in establishing environmental sensor networks," in 3rd International Conference on Intelligent Sensors, Sensor Networks and Information. IEEE, 2007, pp. 687–691.

[14] Jayavardhana Gubbi, Rajkumar Buyya, Slaven Marusic, and Marimuthu Palaniswami, "Internet of things (iot): A vision, architectural elements, and future directions," Future generation computer systems, vol. 29, no. 7, pp. 1645–1660, 2013.

[15] James C Bezdek, Sutharshan Rajasegarar, Masud Moshtaghi, Chris Leckie, Marimuthu Palaniswami, and Timothy C Havens, "Anomaly detection in environmental monitoring networks [application notes]," IEEE Computational Intelligence Magazine, vol. 6, no. 2, pp. 52–58, 2011.

[16] Australian Government, Great Barrier Reef Marine Park Authority, "Facts about the great barrier reef," http: //www.gbrmpa.gov.au/about-the-reef/ facts-about-the-great-barrier-reef, 2017.

[17] Australian Institute of Marine Science (AIMS), "GBROOS Data - Observing the Great Barrier Reef," http://data.aims.gov.au/gbroos/, 2017.

[18] Aravinda S Rao, Stephen Marshall, Jayavardhana Gubbi, Marimuthu Palaniswami, Richard Sinnott, and Vincent Pettigrovet, "Design of low-cost autonomous water quality monitoring system," in 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, 2013, pp. 14–19.

[19] Philip W Rundel, Eric A Graham, Michael F Allen, Jason C Fisher, and Thomas C Harmon, "Environmental sensor networks in ecological research," New Phytologist, vol. 182, no. 3, pp. 589–607, 2009.Integrated Marine Observing System (IMOS), "Wireless Sensor Networks," http://imos.org.au/ facilities/wirelesssensornetworks/, 2017.

[20] Sutharshan Rajasegarar, Jayavardhana Gubbi, Olga Bondarenko, Stuart Kininmonth, Slaven Marusic, Scott Bainbridge, Ian Atkinson, and Marimuthu Palaniswami, "Sensor network implementation challenges in the great barrier reef marine environment," in Conference Proceedings of the ICT-MobileSummit. International Information Management Corporation (IIMC) Ltd, 2008, pp. 1–9.

[21] MEMSIC, Inc, "Wireless sensor networks,"http://www.memsic.com/ wireless-sensor-networks/, 2017, [Online].

[22] Libelium, "Waspmote," http://www.libelium. com/products/waspmote/, 2017, [Online].

[23] Olga Bondarenko, Stuart Kininmonth, and Michael Kingsford, "Underwater sensor networks, oceanography and plankton assemblages," in 3rd International Conference on Intelligent Sensors, Sensor Networks and Information. IEEE, 2007, pp. 657–662.

[24] Chenyang Lu, Brian M Blum, Tarek F Abdelzaher, John A Stankovic, and Tian He, "Rap: A real-time communication architecture for large-scale wireless sensor networks," in Eighth IEEE Real-Time and Embedded Technology and Applications Symposium. IEEE, 2002, pp. 55–66.

[25] Ian F Akyildiz, Weilian Su, Yogesh Sankarasubramaniam, and Erdal Cayirci, "A survey on sensor networks," IEEE Communications magazine, vol. 40, no. 8, pp. 102–114, 2002.

[26] Cameron Huddlestone-Holmes, Gilles Gigan, Graham Woods, Adam Ruxton, Ian Atkinson, and Stuart Kininmonth, "Infrastructure for a sensor network on davies reef, great barrier reef," in 3rd International Conference on Intelligent Sensors, Sensor Networks and Information. IEEE, 2007, pp. 675–679.

[27] Omprakash Gnawali, Ki-Young Jang, Jeongyeup Paek, Marcos Vieira, Ramesh Govindan, Ben Greenstein, August Joki, Deborah Estrin, and Eddie Kohler, "The tenet architecture for tiered sensor networks," in Proceedings of the 4th international conference on Embedded networked sensor systems. ACM, 2006, pp. 153–166.

[28] Peter Desnoyers, Deepak Ganesan, and Prashant Shenoy, "Tsar: a two tier sensor storage architecture using interval skip graphs," in Proceedings of the 3rd international conference on Embedded networked sensor systems. ACM, 2005, pp. 39–50.

[29] Purushottam Kulkarni, Deepak Ganesan, Prashant Shenoy, and Qifeng Lu, "Senseye: a multi-tier camera sensor network," in Proceedings of the 13th annual ACM international conference on Multimedia. ACM, 2005, pp. 229–238.

[30] Junsuk Shin, Rajnish Kumar, Dushmanta Mohapatra, Umakishore Ramachandran, and Mostafa Ammar, "Asap: A camera sensor network for situation awareness," Principles of Distributed Systems, pp. 31–47, 2007.

[31] Phoebus Chen, Parvez Ahammad, Colby Boyer, ShihI Huang, Leon Lin, Edgar Lobaton, Marci Meingast, Songhwai Oh, Simon Wang, Posu Yan, et al., "Citric: A low-bandwidth wireless camera network platform," in Distributed smart cameras, 2008. ICDSC 2008. Second ACM/IEEE international conference on. IEEE, 2008, pp. 1–10.

[32] Chris Townsend and Steven Arms, "Wireless sensor networks: Principles and applications," in Sensor Technology Handbook, Jon S. Wilson, Ed., chapter 22, pp. 575–589. Newnes, 2005.

[33] Jesu´s Pe´rez, Victor Za´rate, and Christian Cabrera, "A network and data link layer qos model to improve traffic performance," Emerging Directions in Embedded and Ubiquitous Computing, pp. 224–233, 2006.

[34] Scott Bainbridge, Damien Eggeling, and Geoff Page, "Lessons from the fieldtwo years of deploying operational wireless sensor networks on the great barrier reef," Sensors, vol. 11, no. 7, pp. 6842–6855, 2011.

[35] Sutharshan Rajasegarar, James C Bezdek, Christopher Leckie, and Marimuthu Palaniswami, "Elliptical anomalies in wireless sensor networks," ACM Transactions on Sensor Networks (TOSN), vol. 6, no. 1, pp. 7, 2009.

[36] Masud Moshtaghi, Sutharshan Rajasegarar, Christopher Leckie, and Shanika Karunasekera, "Anomaly detection by clustering ellipsoids in wireless sensor networks," in Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2009 5th International Conference on. IEEE, 2009, pp. 331–336.

[37] Timothy C Havens and James C Bezdek, "An efficient formulation of the improved visual assessment of cluster tendency (ivat) algorithm," IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 5, pp. 813–822, 2012.

[38] Gartner, Inc, "Gartner says 8.4 billion connected "things" will be in use in 2017, up 31 per- cent from 2016," http://www.gartner.com/ newsroom/id/3598917, 2017, [Online].

INCLUSION OF ARTIFICIAL INTELLIGENCE IN COMMUNICATION NETWORKS AND SERVICES

XU Guibao, MU Yubo, LIU Jialiang
Institute of Technology and Standards Research, China Academy of Information and Communication Technology, Beijing, China PR

Abstract – AI with learning abilities is a revolutionary technology which the communication industry is exploring, with the aim of introducing it into communication networks and to provide new services, and to improve network efficiency and user experience. At this time there is no total solution or complete framework to do so. One contender in the steps towards a solution is a FINE framework, which can be illustrated by the example of an SDN/NFV collaboratively- deployed network.

Keywords – Artificial intelligence, communication network, network functions virtualization (NFV), software-defined network (SDN)

1. INTRODUCTION

In recent years, with the development and maturation of such technologies as cloud computing, big data and deep learning, the industrialization of artificial intelligence (AI) has been developed accordingly. Since AlphaGo won the Go match against Lee Se-dol in 2016, AI has attracted more and more attention. AI technology has been introduced into a number of areas. As a revolutionary force, AI has been making great progress and realized many achievements in these areas.

Communications is a sector with heavy ICT use, dealing with a variety of consumer demands on individualization requirements, multimedia services and precision management, which has made network security become more and more important. With AI's advantages in learning, understanding, reasoning, and cooperating gradually being discovered, software-defined networks (SDN) and network functions virtualization (NFV) have appeared, technologies of deep packet inspection and service aware networks are almost in maturity, and the intellectualization of communication networks and services are becoming possible. Furthermore, operators have a keen interest in AI which may decrease capital expenditure (CAPEX) and operating expense (OPEX).

2. TRENDS IN COMMUNICATION NETWORKS AND SERVICES

2.1. Characterized requirements

With an increasing number of users and the growing size of the communication network, differences of preferences, habits and the information needs of enterprises and individual users are gradually exposed. The demand for specialized businesses is becoming stronger with customized networks and services now being provided for enterprise users. In the future there will be a special service package for each user, and even a special network. Such complex requirements would be unimaginable without an intellectual tool.

2.2. Multimedia services

With the arrival of the Web 2.0 era, Internet users have become information producers, as well as information consumers, and are producing more and more information in multimedia. User-generated content increases Internet traffic at an unbelievable speed. Under these circumstances, both storage and transmission are a great challenge. The inclusion of AI bolsters our abilities to handle this challenge.

2.3. Precision management

The use of smartphones makes it inevitable that the various dimensions and granularities in today's wireless traffic models should be considered in the networks.

With the development of the technologies of network function virtualization and software-defined networks, the management of the network has become more precise. Virtualization is not only at the level of network elements, but also at the level of components such as the CPU, memory, port, bandwidth, etc.
AI-based technologies allow operators to set up an on-demand networks for special users. Operators can also attain their energy-saving goals, as well as other goals with AI.

2.4. Predictable future

The expansion of business requirements and increasing numbers of users has meant that the gap between the peaks and troughs of network usage is becoming greater. In this case, operators are requested to predict the future status of networks more accurately to satisfy users' demand and improve their experience.

2.5. Intellectualization

Networks are becoming more heterogeneous. Users often use a variety of equipment with different wireless access technologies such as 2G, 3G, 4G, Wi-Fi and Internet of Things (IoT), and the adoption of 5G will reshape telecommunication networks in the near future.
The increase in network equipment and user terminals, the expansion of network size, the increase in the number of users, and the increasing complexity of the network has resulted in the network management becoming more difficult to maintain with an acceptable quality of service (QoS). As well as expanding capacity by introducing more equipment, operators are expected to raise their network performance with smart tools and intelligence technologies. This includes introducing more intelligence into networks and management to meet customer needs, make more profits, reduce operating costs, and improve network performance.

2.6. More attention to security and safety

Security incidents are growing and becoming more severe. These events have resulted in significant commercial consequences, including broken networks, economic losses, etc. AI can be introduced into several network layers to establish strong security protection and behavioral analysis based on machine learning will significantly improve the ability of network detection attacks, automatic analysis of data, and the identification of relationships between isolated behaviors.

3. ADVANTAGES OF AI

AI continues to develop rapidly. In the communications industry, whether it's network operators, equipment manufacturers or solution providers, etc., the industry hopes to take advantage of AI to assist in areas in which they are currently struggling, such as in designing, operating, maintaining and managing communication networks and services. The next few subclauses describe some of the advantages of AI.

3.1. Abilities of learning

Operators need intelligent decisions to manage complex resources and dynamic traffic. But so far no one single model has the ability to accurately describe the network traffic characteristics. Fortunately, AI has entered into the cognitive age, and deep learning can be used (confirmed by Hinton in 2006 [1]). Through deep learning, the machine system can use the existing training data to process large amounts of data through data mining. AI can also learn the characteristics of data traffic, management, controls and other characteristics automatically and master expert experience of operating, managing and maintaining networks. By these efforts, the accuracy of analysis can be enhanced, and the intelligent management and services of communication networks can be realized.

3.2. Abilities of understanding and reasoning

Due to the dynamics of the network system, the state information of a resource may have changed when it is transmitted to the network management system. Therefore, the network management can only know the local state information without the knowledge about the system internal state. Machine learning happens to have the strength to deal with this kind of fuzzy logic and uncertainty reasoning. In order to make the classification or prediction easier, deep learning constructs a multi-hidden layer model and uses the hierarchical network structure to transform the feature representation of the sample into a new feature space layer by layer. In particular, AI does not need to describe the mathematical model of the system accurately, and therefore has the ability to deal with uncertainty or even 'unknowability.

3.3. Ability of collaborating

Due to the expansion of the network both in scale and size, the structure complexity of communication networks are increasing quickly. Concepts such as distribution and hierarchy are often talked about in the network management.

Management tasks and controls are distributed to the entire network. As a result, we have to deal with issues such as tasks' distribution, communication and collaboration between management nodes. If we introduce the multi-agent collaboration of distributed AI into the network management, we can expect the ability to collaborate between network managers distributed in every layer.

4. POSSIBILITY TO USE AI IN COMMUNICATIONS

From the TDM automatic switch, it has been the pursuit of the communications industry to introduce intelligence into network operations, management and maintenance management. N. Kojic, et al. [2] suggested a neural network algorithm for the optimization of routing in communication networks. SUI Dan and JIN Xian-hua [3] suggested a network congestion control method based on AI. Sandra Sendra et al. [4] introduced AI into a routing protocol using SDN. Sahebu, K.M. [5] suggested an AI approach to planning and managing communication networks. Undoubtedly, they achieved very good results in their research. But under the condition that networks, equipment and systems must be treated as undivided management objects, their research mainly focused on the theoretical analysis and simulation to give a certain kind of solution and this could not be used in real communication networks or services.

However, SDN, NFV, network slicing and other technologies, coupled with integrated network management systems have been able to directly issue orders which can be executed by network equipment, and DPI systems can be deployed on network equipment, and it is possible to realize real-time monitoring of networks and services and intelligent management.

4.1. AI in SDN

Through the separation of control and forwarding, SDN provides network operators with a logical centralized control and flexible programming interfaces which greatly promote the capabilities of network automated management and control; more than was previously possible.

A typical SDN framework [6] is composed of three layers: infrastructure layer, control layer and application layer. The infrastructure layer includes some network elements which can provide network traffic, acting as the object controlled by the SDN controller, as well as a data source of the network resource. The control layer has the SDN controller, which is the core component of the SDN network carrying out important tasks of controlling network traffic. The application layer includes various applications. The southbound interface D-CPI (Data-Controller Plane Interface) is responsible for exchanging data between the SDN controller and the network element. The northbound interface A-CPI (Application-Controller Plane Interface) is responsible for providing the upper-level application with the channel exchange to obtain the underlying network resource information and send data to the lower-level network. SDN provides a good interface with its programmability to introduce AI into the communication networks. This is SDN's biggest advantage. SDN uses the application programming interface to send powerful programming instructions to the network device. With AI, network managers cannot only schedule an automated intelligent business orchestrator, but also program the AI-optimized network strategy and automatically compile them into the task script, then assign them into the network allocation tasks with the application programming interface (API). Network managers can also automatically collect network statistics information to lay a solid foundation for continuous network optimization. If necessary, some new functionalities can also be added intelligently through the SDN application for the network environment.

4.2. AI in NFV

With virtualization technology, network functions virtualization (NFV) [7] can divide network-level functions and applications, such as routing, customer premises equipment (CPE), mobile core, IP multimedia subsystems (IMS), content delivery networks (CDN), switching elements, mobile network nodes, home routing operations, set-top box business, tunnel gateway elements, traffic analysis, service assurance, service level agreement (SLA) monitoring, testing and diagnosis, next generation network (NGN) signal, aggregation and network range functions, application optimization, security policy, etc., into several functional blocks, and run them in software mode respectively. This means that they are no longer limited to the hardware architecture.

The typical NFV reference architecture includes three layers of the complete infrastructure layer, the resource management layer, and the business flow orchestrator layer. NFV helps ISV and telecommunication operators to achieve virtual network functions by deploying hypervisor at the infrastructure layer to virtualize infrastructure resources such as commercial general computing, storage and network resources and others. The resource management layer is in charge of the NFV infrastructure's management, configuration and collaboration. The business flow orchestrator layer is a key part of the NFV network function for network operating; it is used to organize and orchestrate the functions of the NFV network. It is also in charge of managing and monitoring the global resources across the data center or the resource pool.

With the virtualization of network functions NFV can realize an on-demand dynamic network configuration separated from the underlying architecture. As key issues have been solved, AI can play its full role in critical network management.

4.3. Network monitor and control

To master the real-time information of the communication network, the network must have the function of initiative uploading. Currently there are many DPI systems. With inspectors, the deep packet inspection (DPI) [8] system can collect the information such as the running state of network equipment, the usage of resources and the quality of services.

With the big data obtained from the DPI system, the AI system can rapidly analyse and find if there are or will be abnormity within the information. For example, if the AI system finds a burst a continuous traffic, it can doubt a distributed denial of service (DDoS) attack in the network and analyse the package characteristics immediately, then orchestrate an inspector collaboration task to drop all packages with the characteristics to avoid the damage. It could write a new record in the security database in case of the appearance of unknown hack attacks or new virus flooding.

5. AN AI-BASED NETWORK FRAMEWORK

On the basis of the above analysis and Xu's work [9] the establishment of an intelligent communication network, called future intelligent network (FINE) can be considered in the near future. The system architecture of FINE is shown in Fig. 1.

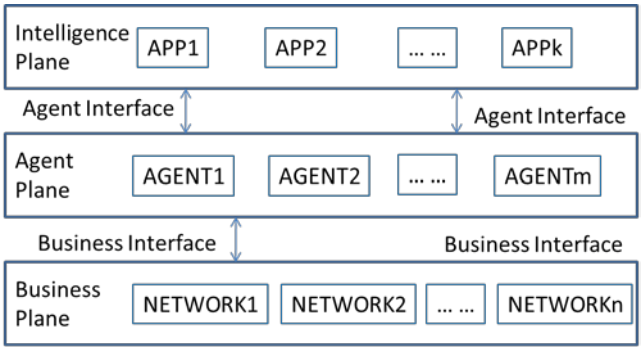


Fig.1. A system architecture of FINE

5.1. Intelligence plane

The intelligence plane, see Fig. 2, is in charge of providing intelligence for the entire FINE, and it acts as the brain of the FINE system. Therefore, FINE is an intelligent network with an AI core. The intelligence plane can be composed of the basic layer, the core layer, the platform layer, the application and terminal layer and the scheme layer.

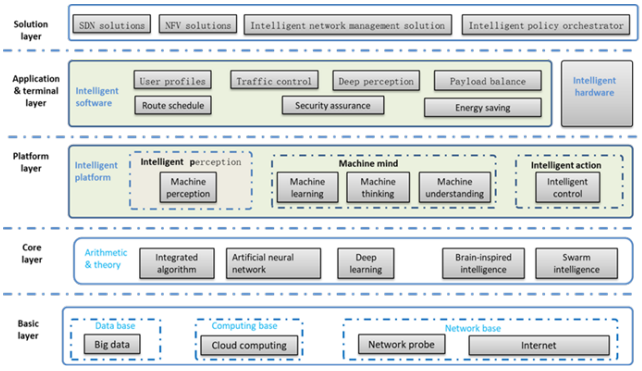


Fig. 2. The intelligence plane

The base layer provides support in data, calculation and the network for the intelligent plane. The data here is big data, not only including static data such as expert knowledge data, network infrastructure data, user profile data and others, but also including dynamic original data collected by the network probes from the business layer, such as status data of various types of equipment, applications and services.

The core layer is the provider of intelligent algorithms in the intelligent plane, such as integrated algorithms, an artificial neural network, depth learning, brain-inspired intelligence and swarm intelligence. It is the kernel of the FINE core.

The platform layer provides intelligent planes for the realization of the intelligent logic of AI ability and behavior, such as intelligent perception, machine mind, intelligent action etc. The intelligent

perception function can make use of theories and algorithms of the core layer, and deal with the big data of the basic layer supported by the computing resources, so as to perceive the development trends of networks and services. The machine mind function includes machine learning, machine thinking, machine understanding, etc. The machine learning consists of machine learning abilities generated by algorithms such as deep learning, brain-inspired intelligence and swarm intelligence. The machine thinking function provides the ability of knowledge mapping and knowledge reasoning. The machine understanding function provides the abilities of understanding based on the existing knowledge and the phenomenon, solving the ambiguity problem in reasoning, etc.

The application and terminal layer provides abilities of modular realization of functions needed by the solution layer. The functions here may include the user portrait, the flow control, the load balancing, the depth perception, the routing, the security, the energy saving, etc. These realizations may be in software or hardware using the abilities of perception, thinking and action provided by the platform layer.

The solution layer is in charge of designing flexible policies and related activities related to satisfy the requirements to operate or manage the network, the network element, the network management system, etc.

5.2. Agent plane

The agent plane consists of a series of agents with characteristics such as autonomy, sociality, responsiveness, initiative, rationality, learning and adaptability, reasoning ability etc. An agent is usually composed of a user interface module, a learning module, a task technology module, an operating system interface module, an execution module, a knowledge base and a central control module. Among them, the central control module is the core of the agent and it controls all other modules. These agents independently play their roles of communicators. It sends the information obtained from network probes to the data module of the intelligence plane, and sends intelligent control instructions from the intelligence plane to the business plane. At the same time, these agents can also communicate with each other and form several linkages for cooperative operation.

The technology of intelligent agents has been deeply researched and widely applied since J. Holland put it forward in 1995 .[10] . Therefore, there are no further details to describe it here.

5.3. Business plane

The business plane, see Fig. 3, mainly includes networks, services, systems intelligently serving the network manager and the end user. It is in charge of executing services orchestrated by the intelligence plane with its components such as the communication network and its operation support system, service support system, etc. Every component in this plane is accompanied by a DPI probe.

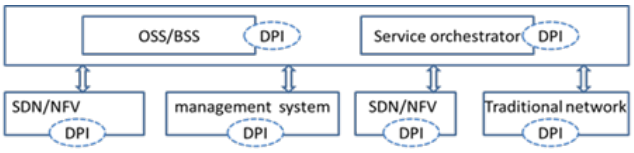


Fig. 3. The business plane

The DPI probe is the sensor and the executor of FINE. Firstly, it perceives the status of its accompanying object. It collects any useful information of its accompanying object, and sends the information to the AI plane through the agent. Secondly, it executes instructions from the intelligence plane and necessary actions reasoned by itself.

Service components include management systems, communication networks, user service systems and infrastructures, etc. Management systems serving the manager could include OSS, BSS, NMS, EMS, etc. Communication networks could include SDNs/NFVs, traditional networks, etc. User service systems include cloud computing, mobile Internet, data communications, 5G, etc. Infrastructures include data centers, equipment, etc.

6. A FINE EXAMPLE

Figure 4 shows an SDN/NFV collaboratively-deployed network. Nodes in this network may be real network elements or virtual ones.

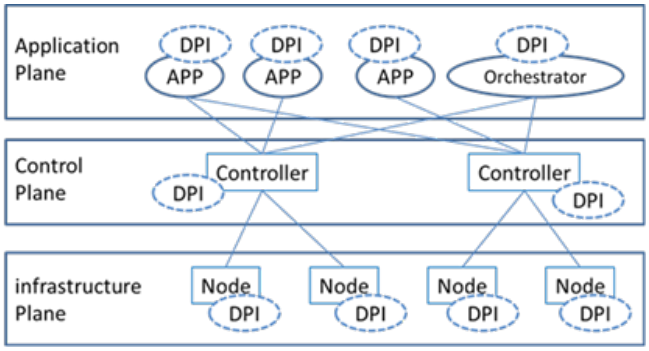


Fig. 4. An SDN/NFV collaboratively-deployed network

To include AI into this network, we can deploy DPIs for every component in it. All the information collected by DPIs will be sent to the big data module in the basic layer of the intelligence plane. The intelligent perception module mines the data to find the characteristics of the changed data. Then the Machine mind module actions a 'reasoning and understanding' supported by algorithms at the core layer and the intelligent policy orchestrator module at the solution layer, and gives its judgment. After that, the intelligent control module makes a decision and provides instructions to the control plane through the agent plane. When controllers at the control plane receive instructions sent to them, they carry them out at related nodes.

7. CONCLUSION

In this paper, we highlighted an AI-based network framework, FINE, to give a total solution to introducing AI in communication networks and services. This was then illustrated with an SDN/NFV collaboratively-deployed network.

It has been proven that the FINE framework is feasible to be used in real communication networks and services. Not only that, we can depend on this framework to set up a standard system for AI-based communication networks and services by defining detailed functions of nodes, layers, planes, related interfaces, etc. And finally, questions around privacy and security should be considered.

REFERENCES

[1] G. E. Hinton, R. R. Salakhutdinov, Reducing the Dimensionality of Data with Neural Networks, Science, Jul. 2006, vol. 313:504-507.

[2] N. Kojic, et al. Neural Network for Optimization of Routing in Communication Networks, Facta Universitatis(NIS) Ser: Elec. Energ. Aug. 2006, vol. 19, no.2:317-329.

[3] SUI Dan, JIN Xian-hua, Network Cognition Control Method Based on Artificial Intelligence, Computer Simulation, Sept. 2011, vol. 28, No.9: 102-105.

[4] Sandra Sendra, et al, Including Artificial Intelligence in a Routing Protocol Using Software Defined Network, ICC2017: wt04-5th IEEE International Workshop on Smart Communication Protocols and Algorithms (SCPA 2017).

[5] Sahebu, K.M. Artificial intelligence approach to planning and managing communication networks, International Conference on Electromagnetic Interference & Compatibility, 2002 :193-202.

[6] Open Networking Foundation (ONF), Software-defined Networking: The New Norm of Networks [EB/OL], http://www.opennetworking.org/images/stories/downloads/sdn-resources/white-papers/wp-sdn-nownorm.pdf .

[7] Chiosi M. et al. Network Functions Virtualization: An Introduction, Benefits, Enablers, Challenges and Call for Action [EB/OL], http://www.etsi.org .

[8] M. Al-Hisnawi, M. Ahmadi, Deep Packet Inspection Using Quotient Filter, IEEE Communications Letters, 2016, 20 (11) :2217-2220.

[9] XU Guibao, A technological architecture of artificial intelligence, Telecommunication Network Technology, Dec. 2016, no.12:1-6.

[10] J. Holland, Can there be a unified theory of complex adaptive systems?, in: The Mind, the Brain, and the Complex Adaptive Systems, H. Morowitz and J. L. Singer, Eds., Addison-Wesley, 1995, pp. 45–50.

EXPLAINABLE ARTIFICIAL INTELLIGENCE: UNDERSTANDING, VISUALIZING AND INTERPRETING DEEP LEARNING MODELS

Wojciech Samek[1], Thomas Wiegand[1,2], Klaus-Robert Müller[2,3,4]

[1]Dept. of Video Coding & Analytics, Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany
[2]Dept. of Electrical Engineering & Computer Science, Technische Universität Berlin, 10587 Berlin, Germany
[3]Dept. of Brain & Cognitive Engineering, Korea University, Seoul 136-713, South Korea
[4]Max Planck Institute for Informatics, Saarbrücken 66123, Germany

Abstract – With the availability of large databases and recent improvements in deep learning methodology, the performance of AI systems is reaching, or even exceeding, the human level on an increasing number of complex tasks. Impressive examples of this development can be found in domains such as image classification, sentiment analysis, speech understanding or strategic game playing. However, because of their nested non-linear structure, these highly successful machine learning and artificial intelligence models are usually applied in a black-box manner, i.e. no information is provided about what exactly makes them arrive at their predictions. Since this lack of transparency can be a major drawback, e.g. in medical applications, the development of methods for visualizing, explaining and interpreting deep learning models has recently attracted increasing attention. This paper summarizes recent developments in this field and makes a plea for more interpretability in artificial intelligence. Furthermore, it presents two approaches to explaining predictions of deep learning models, one method which computes the sensitivity of the prediction with respect to changes in the input and one approach which meaningfully decomposes the decision in terms of the input variables. These methods are evaluated on three classification tasks.

Keywords – Artificial intelligence, black-box models, deep neural networks, interpretability, layer-wise relevance propagation, sensitivity analysis

1. INTRODUCTION

The field of machine learning and artificial intelligence has progressed over the last few decades. A driving force for this development was the earlier improvements in support vector machines and more recent improvements in deep learning methodology [22]. Also the availability of large databases such as ImageNet [9] or Sports1M [17], the speed-up gains obtained with powerful GPU cards and the high flexibility of software frameworks such as Caffe [15] or TensorFlow [1] were crucial factors in the success. Today's machine learning based AI systems excel in a number of complex tasks ranging from the detection of objects in images [14] and the understanding of natural languages [8] to the processing of speech signals [10]. On top of this, recent AI[1] systems can even outplay professional human players in difficult strategic games such as Go [34] and Texas hold'em poker [28]. These immense successes of AI systems, especially deep learning models, show the revolutionary character of this technology, which will have a large impact beyond the academic world and will also give rise to disruptive changes in industries and societies.

However, although these models reach impressive prediction accuracies, their nested non-linear structure makes them highly non-transparent, i.e. it is not clear what information in the input data makes them actually arrive at their decisions. Therefore, these models are typically regarded as black boxes. The 37th move in the second game of the historic Go match between Lee Sedol, a top Go player, and AlphaGo, an artificial intelligence system built by DeepMind, demonstrates the non-transparency of the AI system. AlphaGo played a move which was totally unexpected and which was commented on by a Go expert in the following way:

This work was supported by the German Ministry for Education and Research as Berlin Big Data Center (BBDC) (011S14013A). We thank Grégore Montavon for his valuable comments on the paper.

[1] The terms artificial intelligence and machine learning are used synonymously.

"It's not a human move. I've never seen a human play this move." (Fan Hui, 2016).
Although during the match it was unclear why the system played this move, it was the deciding move for AlphaGo to win the game. In this case the black-box character of the AlphaGo did not matter, but in many applications the impossibility of understanding and validating the decision process of an AI system is a clear drawback. For instance, in medical diagnosis, it would be irresponsible to trust predictions of a black-box system by default. Instead, every far-reaching decision should be made accessible for appropriate validation by a human expert. Also in self-driving cars, where a single incorrect prediction can be very costly, the reliance of the model on the right features must be guaranteed. The use of explainable and human interpretable AI models is a prerequisite for providing such a guarantee. More discussion on the necessity of explainable AI can be found in clause 2.
Unsurprisingly, the development of techniques for "opening" black-box models has recently received a lot of attention in the community [6, 35, 39, 5, 33, 25, 23, 30, 40, 11, 27]. This includes the development of methods which help to better understand what the model has learned (i.e. its representation) [12, 24, 29], as well as techniques for explaining individual predictions [19, 35, 39, 5, 26]. A tutorial on methods from these two categories can be found in [27]. Note that explainability is also important for support vector machines and other advanced machine learning techniques beyond neural networks [20].
The main goal of this paper is to foster awareness for the necessity of explainability in machine learning and artificial intelligence. This is done in clause 2. After that in clause 3, we present two recent techniques, namely sensitivity analysis (SA) [6, 35] and layer-wise relevance propagation (LRP) [5], for explaining the individual predictions of an AI model in terms of input variables. The question of how to objectively evaluate the quality of explanations is addressed in clause 4 and results from image, text and video classification experiments are presented in clause 5. The paper concludes with an outlook on future work in clause 6.

2. WHY DO WE NEED EXPLAINABLE AI?

The ability to explain the rationale behind one's decisions to other people is an important aspect of human intelligence. It is not only important in social interactions, e.g. a person who never reveals one's intentions and thoughts will be most probably regarded as a "strange fellow", but it is also crucial in an educational context, where students aim to comprehend the reasoning of their teachers. Furthermore, the explanation of one's decisions is often a prerequisite for establishing a trust relationship between people, e.g. when a medical doctor explains the therapy decision to his patient.
Although these social aspects may be of less importance for technical AI systems, there are many arguments in favor of explainability in artificial intelligence. Here are the most important ones:

- **Verification of the system**: As mentioned before, in many applications one must not trust a black-box system by default. For instance, in healthcare the use of models which can be interpreted and verified by medical experts is an absolute necessity. The authors of [7] show an example from this domain, where an AI system which was trained to predict the pneumonia risk of a person arrives at totally wrong conclusions. The application of this model in a black-box manner would not reduce but rather increase the number of pneumonia-related deaths. In short, the model learns that asthmatic patients with heart problems have a much lower risk of dying of pneumonia than healthy persons. A medical doctor would immediately recognize that this cannot be true as asthma and heart problems are factors which negatively affect the prognosis for recovery. However, the AI model does not know anything about asthma or pneumonia, it just infers from data. In this example, the data was systematically biased, because in contrast to healthy persons the majority of asthma and heart patients were under strict medical supervision. Because of that supervision and the increased sensitivity of these patients, this group has a significant lower risk of dying of pneumonia. However, this correlation does not have causal character and therefore should not be taken as a basis for the decision on pneumonia therapy.
- **Improvement of the system**: The first step towards improving an AI system is to understand its weaknesses. Obviously, it is more difficult to perform such weakness analysis on black-box models than on models which are interpretable. Also detecting biases in the model or the dataset (as in the pneumonia example) is easier if one understands what the model is doing and why it arrives at its predictions. Furthermore, model interpretability can be helpful when comparing different models or architectures. For instance, the authors of [20, 2, 3] observed that models may have the same classification performance,

but largely differ in terms of what features they use as the basis for their decisions. These works demonstrate that the identification of the most "appropriate" model requires explainability. One can even claim that the better we understand what our models are doing (and why they sometimes fail), the easier it becomes to improve them.

- **Learning from the system**: Because today's AI systems are trained with millions of examples, they may observe patterns in the data which are not accessible to humans, who are only capable of learning with a limited number of examples. When using explainable AI systems, we can try to extract this distilled knowledge from the AI system in order to acquire new insights. One example of such knowledge transfer from AI system to human was mentioned by Fan Hui in the quote above. The AI system identifies new strategies to play Go, which certainly now have also been adapted by professional human players. Another domain where information extraction from the model can be crucial are the sciences. To put it simply, physicists, chemists and biologists are interested in identifying the hidden laws of nature rather than just predicting a quantity with black-box models. Thus, only models which are explainable are useful in this domain (c.f., [37, 32]).
- **Compliance to legislation**: AI systems are affecting more and more areas of our daily life. Along with this related legal aspects such as the assignment of responsibility when the systems make a wrong decision, have also recently received increased attention. Since it may be impossible to find satisfactory answers for these legal questions when relying on black-box models, future AI systems will necessarily have to become more explainable. Another example where regulations may become a driving force for more explainability in artificial intelligence are individual rights. People who are immediately affected by decisions of an AI system (e.g. those who are rejected for loans by the bank) may want to know why the systems have decided in this way. Only explainable AI systems will provide this information. These concerns brought the European Union to adapt new regulations which implement a "right to explanation" whereby a user can ask for an explanation of an algorithmic decision that was made about her or him [13].

These examples demonstrate that explainability is not only of important and topical academic interest, but it will play a pivotal role in future AI systems.

3. METHODS FOR VISUALIZING, INTERPRETING AND EXPLAINING DEEP LEARNING MODELS

This clause introduces two popular techniques for explaining predictions of deep learning models. The process of explanation is summarized in Fig. 1. First, the system correctly classifies the input image as "rooster". Then, an explanation method is applied to explain the prediction in terms of input variables. The result of this explanation process is a heatmap visualizing the importance of each pixel for the prediction. In this example the rooster's red comb and wattle are the basis for the AI system's decision.

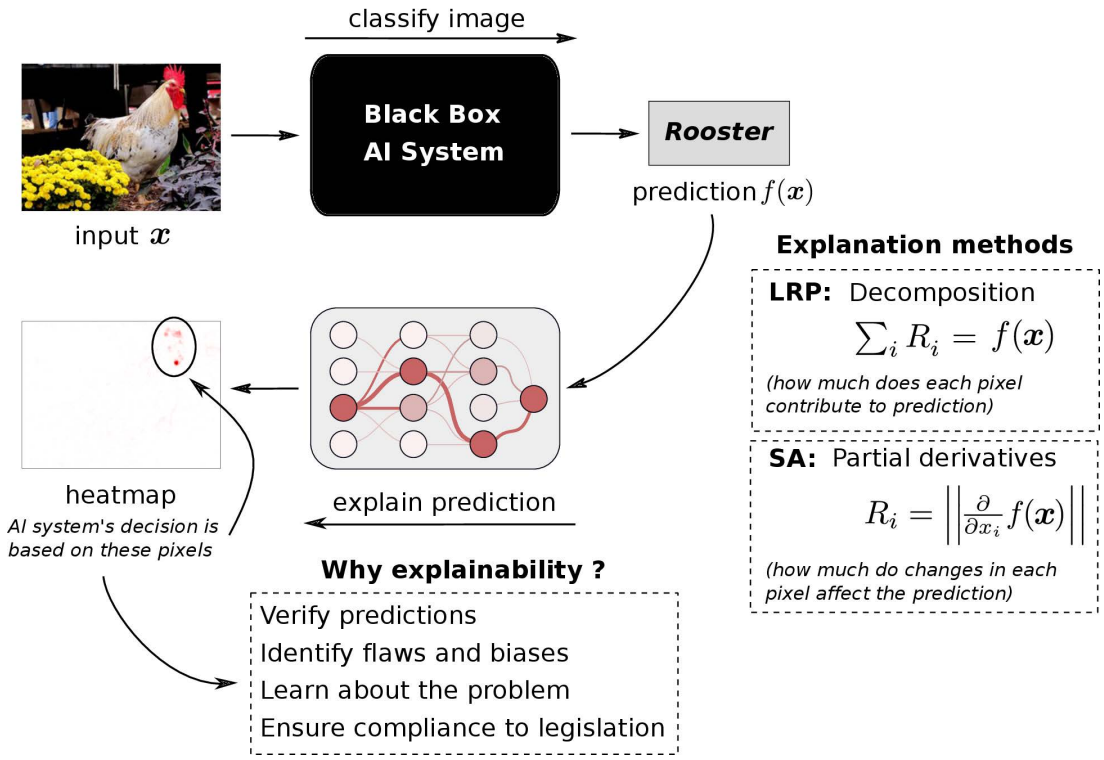


Fig. 1. Explaining predictions of an AI system. The input image is correctly classified as "rooster". In order to understand why the system has arrived at this decision, explanation methods such as SA or LRP are applied. The result of this explanation is an image, the heatmap, which visualizes the importance of each pixel for the prediction. In this example the rooster's red comb and wattle are the basis for the AI system's decision. With the heatmap one can verify that the AI system works as intended

3.1. Sensitivity analysis

The first method is known as sensitivity analysis (SA) [6, 35] and explains a prediction based on the model's locally evaluated gradient (partial derivative). Mathematically, sensitivity analysis quantifies the importance of each input variable *i* (e.g. image pixel) as

$$R_i = \left|\left|\frac{\partial}{\partial x_i} f(\mathbf{x})\right|\right|$$

This measure assumes that the most relevant input features are those to which the output is most sensitive. In contrast to the approach presented in the next subclause, sensitivity analysis does not explain the function value $f(\mathrm{x})$ itself, but rather a variation of it. The following example illustrates why measuring the sensitivity of the function may be suboptimal for explaining predictions of AI systems. A heatmap computed with sensitivity analysis indicates which pixels need to be changed to make the image look (from the AI system's perspective) more/less like the predicted class. For instance, in the example shown in Fig. 1 these pixels would be the yellow flowers which occlude part of the rooster. Changing these pixels in a specific way would reconstruct the occluded parts of the rooster, which most probably would also increase the classification score, because more of the rooster would be visible in the image.

Note that such a heatmap would not indicate which pixels are actually pivotal for the prediction "rooster". The presence of yellow flowers is certainly not indicative of the presence of a rooster in the image. Because of this property SA does not perform well in the quantitative evaluation experiments presented in clause 5. More discussion on the drawbacks of sensitivity analysis can be found in [27].

3.2. Layer-wise relevance propogation

In the following we provide a general framework for decomposing predictions of modern AI systems, e.g. feed-forward neural networks and bag-of-words models [5], long-short term memory (LSTM) networks [4] and Fisher Vector classifiers [20], in terms of input variables. In contrast to sensitivity analysis, this method explains predictions relative to

the state of maximum uncertainty, i.e. it identifies pixels which are pivotal for the prediction "rooster". Recent work [26] also shows close relations to Taylor decomposition, which is a general function analysis tool in mathematics.

A recent technique called layer-wise relevance propagation (LRP) [5] explains the classifier's decisions by decomposition. Mathematically, it redistributes the prediction $f(x)$ backwards using local redistribution rules until it assigns a relevance score R_i to each input variable (e.g. image pixel). The key property of this redistribution process is referred to as relevance conservation and can be summarized as

$$\sum_i R_i = \ldots = \sum_j R_j = \sum_k R_k = \ldots = f(\mathbf{x}) \quad (1)$$

This property says that at every step of the redistribution process (e.g. at every layer of a deep neural network), the total amount of relevance i.e. the prediction $f(x)$ is conserved. No relevance is artificially added or removed during redistribution. The relevance score R_i of each input variable determines how much this variable has contributed to the prediction. Thus, in contrast to sensitivity analysis, LRP truly decomposes the function value $f(x)$.

In the following we describe the LRP redistribution process for feed-forward neural networks; redistribution procedures have also been proposed for other popular models [5, 4, 20]. Let x_j be the neuron activations at layer l, R_k be the relevance scores associated to the neurons at layer $l+1$ and w_{jk} be the weight connecting neuron j to neuron k. The simple LRP rule redistributes relevance from layer $l+1$ to layer l in the following way:

$$R_j = \sum_k \frac{x_j w_{jk}}{\sum_j x_j w_{jk} + \epsilon} R_k \quad (2)$$

where a small stabilization term ϵ is added to prevent division by zero. Intuitively, this rule redistributes relevance proportionally from layer $l+1$ to each neuron in layer l based on two criteria, namely (i) the neuron activation x_j, i.e. more activated neurons receive a larger share of relevance, and (ii) the strength of the connection w_{jk}, i.e. more relevance flows through more prominent connections. Note that relevance conservation holds for $\epsilon = 0$.

The "alpha-beta" rule is an alternative redistribution rule introduced in [5]:

$$R_j = \sum_k \Big(\alpha \cdot \frac{(x_j w_{jk})^+}{\sum_j (x_j w_{jk})^+} - \beta \cdot \frac{(x_j w_{jk})^-}{\sum_j (x_j w_{jk})^-} \Big) R_k \quad (3)$$

where $()^+$ and $()^-$ denote the positive and negative parts, respectively. The conservation of relevance is enforced by an additional constraint $\alpha - \beta = 1$. For the special case $\alpha = 1$, the authors of [26] showed that this redistribution rule coincides with a "deep Taylor decomposition" of the neural network function when the neural network is composed of ReLU neurons.

3.3. Software

The LRP toolbox [21] provides a python and matlab implementation of the method, as well as an integration into popular frameworks such as Caffe and TensorFlow. With this toolbox one can directly apply LRP to other people's models. The toolbox code, online demonstrators and further information can be found on www.explain-ai.org.

4. EVALUATING THE QUALITY OF EXPLANATIONS

In order to compare heatmaps produced by different explanation methods, e.g. SA and LRP, one needs an objective measure of the quality of explanations. The authors of [31] proposed such a quality measure based on perturbation analysis. The method is based on the following three ideas:

- The perturbation of input variables which are highly important for the prediction leads to a steeper decline of the prediction score than the perturbation of input dimensions which are of lesser importance.
- Explanation methods such as SA and LRP provide a score for every input variable. Thus, the input variables can be sorted according to this relevance score.
- One can iteratively perturb input variables (starting from the most relevant ones) and track the prediction score after every perturbation step. The average decline of the prediction score (or the decline of the prediction accuracy) can be used as an objective measure of explanation quality, because a large decline indicates that the explanation method was successful in identifying the truly relevant input variables.

In the following evaluation we use model-independent perturbations (e.g. replacing the input values by random sample from uniform distribution), in order to avoid biases.

5. EXPERIMENTAL EVALUATION

This clause evaluates SA and LRP on three different problems, namely the annotation of images, the classification of text documents and the recognition of human actions in videos.

5.1. Image classification

In the first experiment we use the GoogleNet model [38], a state-of-the art deep neural network, to classify general objects from the ILSVRC2012 [9] dataset.

Figure 2 (A) shows two images from this dataset, which have been correctly classified as "volcano" and "coffee cup", respectively. The heatmaps visualize the explanations obtained with SA and LRP. The LRP heatmap of the coffee cup image shows that the model has identified the ellipsoidal shape of the cup to be a relevant feature for this image category. In the other example, the particular shape of the mountain is regarded as evidence for the presence of a volcano in the image. The SA heatmaps are much noisier than the ones computed with LRP and large values R_i are assigned to regions consisting of pure background, e.g. the sky, although these pixels are not really indicative for image category "volcano". In contrast to LRP, SA does not indicate how much every pixel contributes to the prediction, but it rather measures the sensitivity of the classifier to changes in the input. Therefore, LRP produces subjectively better explanations of the model's predictions than SA.

The lower part of Fig. 2 (A) displays the results of the perturbation analysis introduced in clause 4. The y axis shows the relative decrease of the prediction score average over the first 5040 images of the ILSVRC2012 dataset, i.e. a value of 0.8 means that the original scores decreased on average by 20%. At every perturbation step a 9x9 patch of the image (selected according to SA or LRP scores) is replaced by random values sampled from a uniform distribution. Since the prediction score decrease is much faster when perturbing the images using LRP heatmaps than when using SA heatmaps, LRP also objectively provides better explanations than SA.

More discussion on this image classification experiment can be found in [31].

5.2. Text document classification

In this experiment, a word-embedding based convolutional neural network was trained to classify text documents from the 20Newsgroup dataset[2]. Fig. 2 (B) shows SA and LRP heatmaps (e.g. a relevance score R_i is assigned to every word) overlaid on top of a document, which was classified as topic "sci.med", i.e. the text is assumed to be about a medical topic. Both explanation methods, SA and LRP, indicate that words such as "sickness", "body" or "discomfort" are the basis for this classification decision. In contrast to sensitivity analysis, LRP distinguishes between positive (red) and negative (blue) words, i.e. words which support the classification decision "sci.med" and words which are in contradiction, i.e. speak for another category (e.g."sci.space"). Obviously, words such as "ride", "astronaut" and "Shuttle" strongly speak for the topic space, but not necessarily for the topic medicine. With the LRP heatmap, we can see that although the classifier decides for the correct "sci.med" class, there is evidence in the text which contradicts this decision. The SA method does not distinguish between positive and negative evidence. The lower part of the figure shows the result of the quantitative evaluation. The y axis displays the relative decrease of the prediction accuracy over 4154 documents of the 20Newsgroup dataset. At every perturbation step, the most important words (according to SA or LRP score) are deleted by setting the corresponding input values to 0. Also this result confirms quantitatively that LRP provides more informative heatmaps than SA, because these heatmaps lead to a larger decrease in classification accuracy compared to SA heatmaps. More discussion on this text document classification experiment can be found in [3].

5.3. Human action recognition in videos

The last example demonstrates the explanation of a Fisher Vector/SVM classifier [16], which was trained for predicting human actions from compressed videos. In order to reduce computational costs, the classifier was trained on block-wise motion vectors (not individual pixels). The evaluation is performed on the HMDB51 dataset [18].

[2] http://qwone.com/~jason/20Newsgroups

Fig. 2 (C) shows LRP heatmaps overlaid onto five exemplar frames of a video sample. The video was correctly classified as showing the action "sit-up". One can see that the model mainly focuses on the blocks surrounding the upper body of the person. This makes perfect sense, as this part of the video frame shows motion which is indicative of the action "sit-up", namely upward and downward movements of the body.

The curve at the bottom of Fig. 2 (C) displays the distribution of relevance over (four consecutive) frames. One can see that the relevance scores are larger for frames in which the person is performing an upwards and downwards movement. Thus, LRP heatmaps not only visualize the relevant locations of the action within a video frame (i.e. where the relevant action happens), but they also identify the most relevant time points within a video sequence (i.e. when the relevant action happens).

More discussion on this experiment can be found in [36].

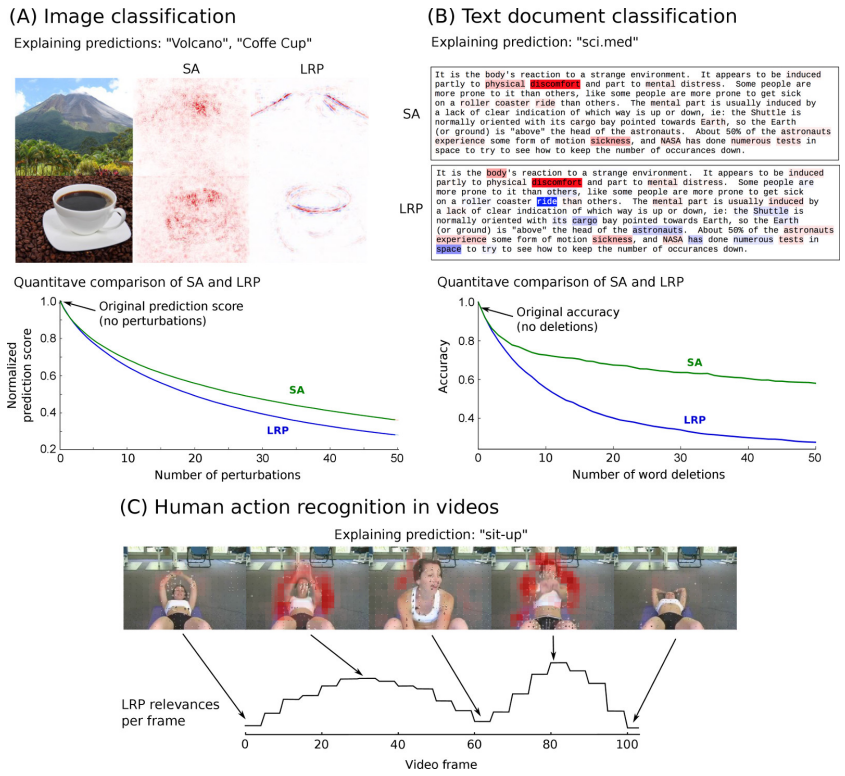


Fig. 2. Explaining predictions of AI systems. (A) shows the application of explainable methods to image classification. The SA heatmaps are noisy and difficult to interpret, whereas LRP heatmaps match human intuition. (B) shows the application of explainable methods to text document classification. The SA and LRP heatmaps identify words such as "discomfort", "body" and "sickness" as the relevant ones for explaining the prediction "sci.med". In contrast to sensitivity analysis, LRP distinguishes between positive (red) and negative (blue) relevances. (C) shows explanations for a human action recognition classifier based on motion vector features. The LRP heatmaps of a video which was classified as "sit-up" show increased relevance on frames in which the person is performing an upwards and downwards movement

6. CONCLUSION

This paper approached the problem of explainability in artificial intelligence. It was discussed why black-box models are not acceptable for certain applications, e.g. in the medical domain where wrong decisions of the system can be very harmful. Furthermore, explainability was presented as a prerequisite for solving legal questions which are arising with the increased usage of AI systems, e.g. how to assign responsibility in case of system failure. Since the "right to explanation" has become part of European law, it can be expected that it will also greatly foster explainability in AI systems.

Besides being a gateway between AI and society, explainability is also a powerful tool for detecting flaws in the model and biases in the data, for verifying predictions, for improving models, and finally for gaining new insights into the problem at hand (e.g. in the sciences).

In future work we will investigate the theoretical foundations of explainability, in particular the connection between post-hoc explainability, i.e. a trained model is given and the goal is to explain its predictions, and explainability which is incorporated directly into the structure of the model. Furthermore, we will study new ways to better understand the learned representation, especially the relation between generalizability, compactness and explainability. Finally, we will apply explaining methods such as LRP to new domains, e.g. communications, and search for applications of these methods beyond the ones described in this paper.

REFERENCES

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467, 2016.

[2] L. Arras, F. Horn, G. Montavon, K.-R. Müller, and W. Samek. Explaining predictions of non-linear classifiers in nlp. In Proceedings of the 1st Workshop on Representation Learning for NLP, pages 1–7. ACL, 2016.

[3] L. Arras, F. Horn, G. Montavon, K.-R. Müller, and W. Samek. ”What is relevant in a text document?”: An interpretable machine learning approach. PLoS ONE, 12(8):e0181142, 2017.

[4] L. Arras, G. Montavon, K.-R. Müller, and W. Samek. Explaining recurrent neural network predictions in sentiment analysis. In Proceedings of the EMNLP’17, Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA), pages 159–168, 2017.

[5] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE, 10(7):e0130140, 2015.

[6] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller. How to explain individual classification decisions. Journal of Machine Learning Research, 11:1803–1831, 2010.

[7] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1721–1730, 2015.

[8] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078, 2014.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 248–255, 2009.

[10] L. Deng, G. Hinton, and B. Kingsbury. New types of deep neural network learning for speech recognition and related applications: An overview. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 8599–8603, 2013.

[11] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608, 2017.

[12] D. Erhan, Y. Bengio, A. Courville, and P. Vincent. Visualizing higher-layer features of a deep network. Technical Report 1341, University of Montreal, 2009.

[13] B. Goodman and S. Flaxman. European union regulations on algorithmic decision-making and a ”right to explanation”. arXiv preprint arXiv:1606.08813, 2016.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016.

[15] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM international conference on Multimedia, pages 675–678, 2014.

[16] V. Kantorov and I. Laptev. Efficient feature extraction, encoding and classification for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2593–2600, 2014.

[17] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), pages 1725–1732, 2014.

[18] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 2556–2563. IEEE, 2011.

[19] W. Landecker, M. D. Thomure, L. M. A. Bettencourt, M. Mitchell, G. T. Kenyon, and S. P. Brumby. Interpreting individual classifications of hierarchical networks. In Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining (CIDM), pages 32–38, 2013.

[20] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Müller, and W. Samek. Analyzing classifiers: Fisher vectors and deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2912–2920, 2016.

[21] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Müller, and W. Samek. The layer-wise relevance propagation toolbox for artificial neural networks. Journal of Machine Learning Research, 17(114):1–5, 2016.

[22] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller. Efficient backprop. In Neural networks: Tricks of the trade, pages 9–48. Springer, 2012.

[23] Z. C. Lipton. The mythos of model interpretability. arXiv preprint arXiv:1606.03490, 2016.

[24] A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5188–5196, 2015.

[25] A. Mahendran and A. Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. International Journal of Computer Vision, 120(3):233–255, 2016.

[26] G. Montavon, S. Bach, A. Binder, W. Samek, and K.R. Müller. Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognition,65:211–222, 2017.

[27] G. Montavon, W. Samek, and K.-R. Müller. Methods for interpreting and understanding deep neural networks. arXiv preprint arXiv:1706.07979, 2017.

[28] M. Moravčík, M. Schmid, N. Burch, V. Lisy´, D. Morrill, N. Bard, et al. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. Science, 356(6337):508–513, 2017.

[29] A. Nguyen, J. Yosinski, and J. Clune. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. arXiv preprint arXiv:1602.03616, 2016.

[30] M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1135–1144. ACM, 2016.

[31] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller. Evaluating the visualization of what a deep neural network has learned. IEEE Transactions on Neural Networks and Learning Systems, 2017, in press.

[32] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko. Quantum-chemical insights from deep tensor neural networks. Nature communications, 8:13890, 2017.

[33] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje. Not just a black box: Learning important features through propagating activation differences. arXiv preprint arXiv:1605.01713, 2016.

[34] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, et al. Mastering the game of go with deep neural networks and tree search. Nature, 529(7587):484–489, 2016.

[35] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034, 2013.

[36] V. Srinivasan, S. Lapuschkin, C. Hellge, K.-R. Müller, and W. Samek. Interpretable human action recognition in compressed domain. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1692–1696, 2017.

[37] I. Sturm, S. Lapuschkin, W. Samek, and K.-R. Müller. Interpretable deep neural networks for single-trial eeg classification. Journal of Neuroscience Methods, 274:141–145, 2016.

[38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1–9, 2015.

[39] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In European Conference Computer Vision - ECCV 2014, pages 818–833, 2014.

[40] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling. Visualizing deep neural network decisions: Prediction difference analysis. arXiv preprint arXiv:1702.04595, 2017.

THE CONVERGENCE OF MACHINE LEARNING AND COMMUNICATIONS

Wojciech Samek[1], Slawomir Stanczak[1,2], Thomas Wiegand[1,2]
[1]Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany
[2]Dept. of Electrical Engineering & Computer Science, Technische Universität Berlin, 10587 Berlin, Germany

Abstract - The areas of machine learning and communication technology are converging. Today's communication systems generate a large amount of traffic data, which can help to significantly enhance the design and management of networks and communication components when combined with advanced machine learning methods. Furthermore, recently developed end-to-end training procedures offer new ways to jointly optimize the components of a communication system. Also, in many emerging application fields of communication technology, e.g. smart cities or Internet of things, machine learning methods are of central importance. This paper gives an overview of the use of machine learning in different areas of communications and discusses two exemplar applications in wireless networking. Furthermore, it identifies promising future research topics and discusses their potential impact.

Keywords – Artificial intelligence, communications, machine learning, wireless networks

1. INTRODUCTION

We are currently observing a paradigm shift towards "smart" communication networks that take advantage of network data. In fact, modern communication networks, and in particular mobile networks, generate a large amount of data at the network infrastructure level and at the user/customer level. The data in the network contains a wealth of useful information such as location information, mobility and call patterns. The vision of network operators is to either enable new businesses through the provisioning of this data (or the information contained within it) to external service providers and customers or to exploit the network data for in-house services such as network optimization and management.

In order to make the vision reality, there is a strong need for the development and implementation of new machine learning methods for big data analytics in communication networks. The objective of these methods is to extract useful information from the network data while taking into account limited communication resources, and then to leverage this information for external or in-house services.

Moreover, machine learning methods are a core part in many emerging applications of communication technology, e.g. smart cities [47] or the Internet of things [46]. Here, topics such as monitoring, fault prediction and scheduling are addressed with modern learning algorithms. The use of machine learning methods in communications may provide information about individuals that affect their privacy. Therefore, various privacy-preserving approaches to data analysis have been recently proposed (e.g. [1]). Machine learning methods are also widely applied to tackle security-related problems in communications, e.g. as part of defense mechanisms against spam attacks and viruses [21].

The increasing convergence can be also observed in specific domains of communications such as image and video communication. While the direct approach to designing compression algorithms using autoencoders has provided very limited results compared to the state-of-the-art approach, the use of machine learning as an enhancing component for aspects like video encoding, bit allocation or other parts became a promising research direction [55]. As most video signals are stored as compressed data, the topic of object recognition and tracking in the compressed domain is also of high relevance [41]. Video streaming is another application which benefits from the use of learning algorithms [28].

Despite the successful use of machine learning methods in various communication applications, there are still many challenges and questions that need to be addressed. For instance, the large size and high computational demands of modern machine learning algorithms prevent the large-scale use of these models in embedded devices. Also 5G networks call for novel machine learning-based approaches to radio resource management and

network management approaches that can cope with uncertainties and incomplete channel and network state information. Other problems concern reliability, privacy and security aspects of machine learning models.

The following clauses review the literature (clause 2), present two applications (clause 3) and discuss future research topics in machine learning and communications (clause 4). The paper concludes with a summary (clause 5).

2. MACHINE LEARNING IN COMMUNICATIONS

This clause reviews the use of machine learning algorithms in different application fields of communications (see Figure 1).

2.1. Communication networks

Routing has a significant impact on the network's performance, and it is a well-studied topic in communications. Machine learning methods have been used to tackle different types of routing problems in the past, including shortest path routing, adaptive routing and multicasting routing. The authors of [7] proposed an algorithm for package routing in dynamically changing networks based on reinforcement learning. This algorithm learns a routing policy which balances between the route length and the possibility of congestion along the popular routes. Extensions on this idea have been proposed in [27]. Other researchers approached the routing problem with genetic algorithms [34]. Here alternative routes are created by crossover and mutation of the existing routes. Genetic algorithms have been also used for tackling the multicasting routing problem which emerges when data is sent to multiple receivers through a communication network [54]. Also, in mobile ad-hoc networks the construction of multicast trees has been addressed using genetic algorithms. Here additional objectives such as bounded end-to-end delay and energy efficiency are added to the optimization [30].

Several works (e.g. [17]) have also used machine learning techniques for throughput or traffic prediction in communication networks. This is an important topic as with a dynamic throughput control and allocation one can fulfill the quality of service (QoS) requirements while efficiently utilizing the network resources. For instance, the authors of [28] applied neural networks for variable-bit-rate video traffic prediction in order to dynamically allocate throughput for real-time video applications. Traffic identification is another important topic for network operators as it helps them to manage their networks, to assure the QoS and to deploy security measures.

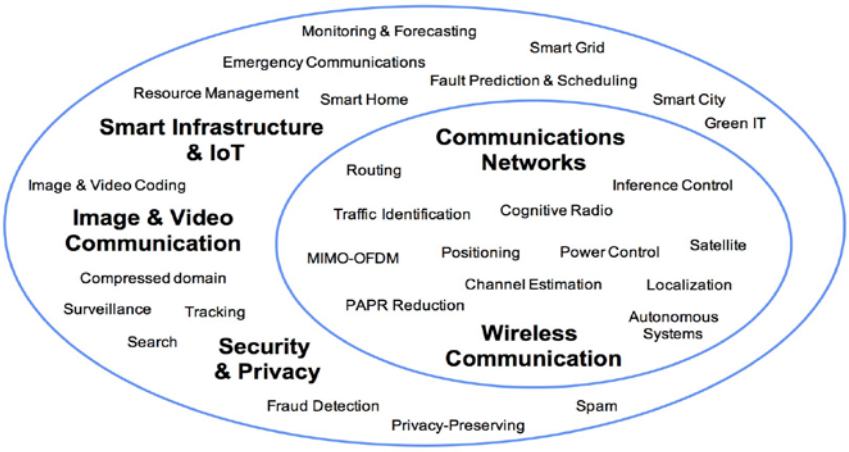


Fig. 1. Applications on machine learning in different areas of communications.

Here, machine learning methods recognize statistical patterns in the traffic data by analyzing packet-header and flow-level information. An excellent review of traffic classification with machine learning methods is [35].

2.2. Wireless communications

To achieve a high efficiency at the desired QoS, it is essential in wireless systems to continuously adapt different parameters of MIMO-OFDM systems, in particular the link parameters, to the variations in the communication environment. Various work (e.g. [52]) tackle this parameter selection problem using machine learning. Due to the dynamic nature of the

wireless communication environment, there is also a strong need for adapting hardware parameters, e.g. to select a suitable set of transmit and receive antennas [25].

The problem of reducing the peak-to-average power ratio (PAPR) is one of the key aspects in the design of OFDM-based wireless systems. This problem has attracted much attention for many years. Application examples of machine learning to the PAPR reduction problem include neural networks [23] and set-theoretic approaches [10] that are particularly suitable for online learning. Methods of machine learning and compressive sensing can also provide a key ingredient in enhancing the efficiency of OFDM channel estimation. For instance, the authors of [13] address the problem by considering a neural network with known pilot signals at its input and the corresponding channel response at its output. Other works (e.g. [39]) turn their attention towards the problem of channel estimation in MIMO systems in the presence of nonlinearities. Learning-based approaches have been also applied for the estimation of mmWave channels [32].

In order to enable an efficient and reliable opportunistic spectrum access, several approaches based on supervised, unsupervised, or reinforcement learning have been proposed in the literature. For instance, the study [44] considers a cognitive radio system with cooperative spectrum sensing where multiple secondary users cooperate to obtain robust spectrum-sensing results. Other approaches [8] apply distributed adaptive learning to tackle this problem.

Power control is a key mechanism for resource allocation in wireless systems. Machine learning has attracted some attention in the context of MIMO power control (e.g. [31]). Various learning-based approaches (e.g. [19]) have also been proposed to tackle the inter-cell interference problem, which may have a detrimental impact on the performance of wireless users in mobile networks. Furthermore, human supervision is still an indispensable element of current network management tools that are used to operate and manage mobile networks. Much research effort has been spent in the last decade to fully automate the network management process and with it to realize the vision of self-organizing networks that operate without human intervention (see [2]).

Information on the position of wireless devices is a key prerequisite for many applications. Machine learning methods have been used for localization [53] as well as navigation and positioning in car-to-car communication systems [40].

2.3. Security, privacy and communications

Machine learning methods play a pivotal role in tackling privacy and security-related problems in communications. For instance, they monitor various network activities and detect anomalies, i.e. events that deviate from the normal network behavior. Various machine learning methods have been applied for network anomaly detection in the past (see [45]). Other security applications are automatic spam filtering [21] and phishing attack detection [4].

Preserving data privacy is an important security aspect in communications, especially when sensitive data is involved. The design of machine learning algorithms that respect data privacy has recently gained increased attention. The authors of [1] demonstrated that it is possible to build a decision-tree classifier from corrupted data without significant loss in accuracy compared to the classifiers built with the original data, while at the same time it is not possible to accurately estimate the original values in the corrupted data records. This way one can hide private information from the algorithm, but still obtain accurate classification results.

2.4. Smart services, smart infrastructure and IoT

With the recent advances in communication technology the new field of "smart" applications has gained increased attention (e.g. smart homes, smart cities, smart grids, Internet of things). Machine learning algorithms are often the core part of such applications. For instance, the authors of [14] used a neural network based prediction algorithm to forecast and manage the power production of a photovoltaic plant. Other researchers have applied similar techniques to traffic light control [48] in smart cities or context aware computing in IoT [37].

Machine learning can also help detecting malicious events before they occur, e.g. in smart-grid networks [18]. Tasks such as the prediction of a resource usage, estimation of task response times, data traffic monitoring and optimal scheduling have also been tackled with learning algorithms [49].

2.5. Image and video communications

Machine learning methods have been used for various tasks in multimedia communication and processing (e.g. more than 200 applications of neural networks for images are summarized in [16]). Signal compression is one important field of application of these methods as it is part of almost every

multimedia communication system. A survey on image compression methods with neural networks can be found in [24]. Tracking is another well-studied topic in machine learning which is also relevant in multimedia communication. A new generation of object tracking methods based on deep neural networks has been described in [12]. Tracking algorithms which make use of the compressed video representation have also gained attention recently [41]. In multimedia applications such as video streaming the quality of the displayed video is of crucial importance. Different machine learning methods have been proposed to estimate the subjective quality of images perceived by a human [5, 6].

3. EXEMPLAR APPLICATIONS IN WIRELESS NETWORKING

The design and operation of wireless networks is a highly challenging task. On the road to the fifth generation of mobile networks (5G), researches and practitioners are, in particular, challenged by a multitude of conflicting requirements and promises as ever higher data-rates, lower-latency and lower energy consumption.

The main cause of the problems and limitations in the context of 5G is the radio propagation channel. This so-called wireless channel can strongly distort transmission signals in a manner that varies with frequency, time, space and other system parameters. The channel distortions are therefore of random nature and are notoriously difficult to estimate and predict. In addition, the wireless channel is a shared communication medium so that different wireless (communication) links must share the available communication resources. In modern mobile networks, this leads to interference between different mobile users, which in turn may have a detrimental impact on network operation. As a result, the capacity of wireless links is of an ephemeral and highly dynamic nature, and it depends on global channel parameters such as path loss, path delay and carrier phase shifts, all of which vary with time, frequency and space.

Against this background, it is not surprising that the problem of reconstructing, tracking and predicting channel parameters play a prominent role in the design and operation of modern wireless networks such as 5G. Traditional approaches to this problem are usually based on the assumptions that 1) the wireless channel can be modeled with a sufficient accuracy and 2) a sufficient number of pilot-based channel measurements can be performed in real time. However, the continuously increasing need for high-spectral efficiency and the utilization of extremely high frequencies (above 6 GHz) makes these assumptions untenable in future networks. A potential solution will not be an adaptation or extension within an existing framework, but rather a paradigm shift is necessary to meet the requirements of 5G networks. This in turn requires large strides both with respect to theoretical foundations and practical implementations.

Modern wireless networks collect and process a large amount of data and this data (including measurement data) can be used for tackling the previously mentioned problem of channel reconstruction, tracking and prediction. Therefore, in this context, special attention has been given to the development of new machine learning algorithms that are able to process spatially distributed data in real time while efficiently using scarce wireless communication resources. This calls for the development of distributed algorithms that in addition must provide robust results, have good tracking (online) capabilities, and exhibit a relatively low complexity. Finally, they need to exploit context and side information such as spatial and temporal sparsity in the wireless channel.

In the following subclause, we present one promising machine learning approach to the problem of reconstructing and tracking path loss maps in cellular networks. Clause 3.2 exemplifies the possibility of designing deep neural networks that exploit sparsity in the input data and which are amenable to real-time implementation.

3.1. Reconstruction of radio maps

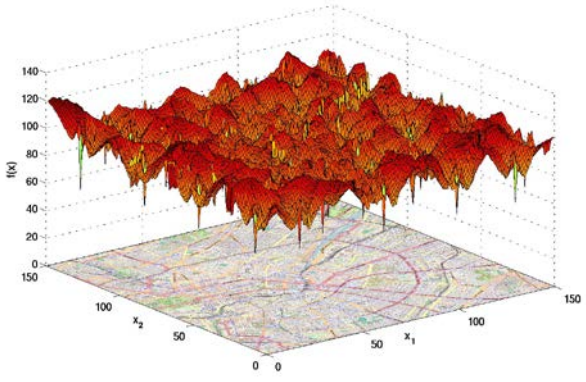


Fig. 2. An example of a path loss map for the downlink scenario with multiple base stations. The path loss map is a 2-dimensional function that assigns to a geographical position its path loss to the strongest base station.

We consider the downlink of a cellular network in which a number of base stations (transmitters) send the data to mobile users. While the users move over a geographical area covered by the network and send periodically their path loss measurements to the base stations, the problem is to reconstruct and track a (complete) path loss map in an online fashion from these measurements. The path loss determines the reduction in the power density of a transmission signal as it propagates from a base station to a geographical position. Note that for every geographical position, its path loss is defined with respect to the strongest base station, which is the base station with the smallest path loss. A radio map is then a function $f: \mathbb{R}^2 \to \mathbb{R}_{\geq 0}$ that assigns to every geographical position in a given area its path loss associated with the strongest base station. Figure 2 shows an example of a path loss map (the 2-dimensional function over the geographical area) for the downlink of a cellular network.

The general setting is as follows: Each base station collects path loss measurements sent by a subset of mobile users and it updates its estimate of the path loss map in an online manner whenever a new measurement arrives. Measurements may contain errors since geographical location cannot be determined with arbitrary precision and measured path loss values can be erroneous. Finally, measurements are not uniformly distributed over a given geographical area so that more measurements may be available for some sub-areas than for others. The challenge is to reliably reconstruct the path loss map, including the path loss values for geographical positions for which no path loss measurements are available.

The problem was considered in [26] where the authors propose using a multi-kernel approach based on adaptive projection methods. To be more precise, consider an arbitrary base station and let $(x_n, y_n) \in \mathbb{R}^2 \times \mathbb{R}$ be its measurement at time $n \in \mathbb{N}$, where $x_n \in \mathbb{R}^2$ is a sample (position measurement) at time n and $y_n \in \mathbb{R}_{\geq 0}$ is the corresponding response (a noisy path loss measurement). An estimate $\hat{f} : \mathbb{R}^2 \mapsto \mathbb{R}_{\geq 0}$ of the path loss map must be consistent with the available measurements. To this end, we require that $\forall_{n\in\mathbb{N}} |y_n - \hat{f}(x_n)| \leq \epsilon$ for some suitably chosen small $\epsilon > 0$. In [26], this requirement is met by projecting the estimate $\hat{f}$ on the hyperslabs given by $S_n = \{f \in \mathcal{H} : |y_n - \langle f, \kappa(x_n, \cdot)\rangle| \leq \epsilon\}, n \in \mathbb{N}$ where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) and $\kappa : \mathbb{R}^2 \times \mathbb{R}^2 \mapsto \mathbb{R}$ is the reproducing kernel for $\mathcal{H}$ so that $\langle f, \kappa(x_n, \cdot)\rangle = f(x_n)$ (the reproducing property). Due to a lack of space, we refer the reader to [26] for a rigorous definition of the concept of RKHS.

Since S_n is a closed convex set, the method of projection on convex sets (POCS) [20] provides the basis for the development of an iterative algorithm. However, the POCS framework cannot be directly applied to our problem at hand because the number of measurements grows without bound as time evolves. Therefore, the authors of [26] considered a different algorithmic approach that is a special case of the adaptive projected sub-gradient methods (APSM) developed in [50, 11, 9]. These methods open the door to distributed implementation and real-time online processing via adaptive parallel projections on closed convex sets such as the hyperslabs. Moreover, they allow for incorporating context information in a systematic manner, while exhibiting relatively low-complexity and robustness against errors. For more details the reader is referred to [50, 11, 9, 26].

The main disadvantage of the APSM-based approach is the need for choosing appropriate kernel functions. In fact, in practical scenarios, different geographical positions require different kernel functions that in addition need to be adapted over time due to the dynamic nature of the wireless environment. Since a real-time optimization of the kernel functions is an intricate task, inspired by the work [51], the authors of [26] developed a multi-kernel approach that adapts kernel functions over time and space by choosing them from a large set of predefined kernel functions, while maintaining low-complexity and real-time capabilities. In the following, we briefly explain this approach.

To this end, let $\{\kappa_m\}_{m=1}^{M}$ with $\kappa_m : \mathbb{R}^2 \times \mathbb{R}^2 \mapsto \mathbb{R}$ be a given set of some predefined kernel functions, where $M \gg 1$ is sufficiently large to include all relevant kernel functions.

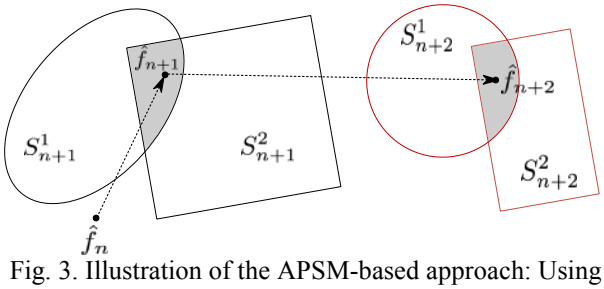


Fig. 3. Illustration of the APSM-based approach: Using parallel projection methods, the estimate $\hat{f}$n follows the intersections of the hyperslabs at times n+1 and n+2 to ensure consistency with new measurements and good tracking capabilities with online processing. At each time, there are two hyperslabs corresponding to two measurements.

Since the number of measurements grows linearly with n, we take into account only the most relevant data which is contained in the dictionary $\{(x_i,y_i)\}_{n\in In}$ where $I_n \subseteq \{n,n-1,\ldots,1\}$ is the dictionary index set at time n. The cardinality $I_n = |I_n|$ of the dictionary must be sufficiently small to satisfy the hardware limitations on memory size and processor speed. With these definitions, for an arbitrary time n, the estimate $\hat{f}_n(x)$ of the path loss at position x can be written as a weighted sum of kernel functions: $\hat{f}_n(x) = \langle A_n, K_n\rangle = \mathrm{trace}(A^{\mathrm{T}}{}_n, K_n)$. Here $K_n = K_n(x) \in \mathbb{R}^{M\times In}$ is a given kernel matrix (evaluated at x) with $[K_n]_{i,m} = \kappa_m(x,x_i)$, and $A_n \in \mathbb{R}^{M\times In}$ is a parameter matrix that needs to be optimized. We point out that since the kernel matrix depends on the position $x \in \mathbb{R}^2$, the parameter matrix should be optimized for different geographical positions.

The most obvious requirement on the parameter matrix A is that it must be chosen to fit the estimate to the measurements. This can be achieved by minimizing the distance (with some suitably chosen metric) of A from the set $S_n = \{A \in \mathbb{R}^{M\times In} : |\langle A_n, K_n\rangle - y_{\mathrm{n}}| \leq \epsilon\}$ for some sufficiently small $\epsilon > 0$. Since M is large, the problem is however computationally prohibitive for many applications in wireless networks. Therefore, the authors of [26] extended the objective function by adding to the distance metric two regularization terms that impose some sparsity in A when the new regularized objective function is minimized. As a result, the approach not only fits the estimate function to the measurements but also discards irrelevant data in the dictionary and reduces the impact of unsuitable kernels.

The regularized objective function provides a basis for the development of new iterative algorithms in [26] based on the forward-backward splitting methods and sparsity-based iterative weighting methods. The algorithms provide good tracking capabilities for the problem of reconstructing and tracking time-varying path loss maps. For more details, we refer the reader to [26].

3.2. Deep neural networks for sparse recovery

Recently, compressed sensing and deep learning have emerged as theoretical and practical toolsets to unleash full potential and approach fundamental theoretical bounds, whether it be for pilot decontamination in channel estimation, user identification, activity detection or PAPR reduction. While in many cases researchers are well aware of optimal solutions, e.g. in terms of optimization problems for channel estimation using a minimal number of pilots, implementing these solutions in embedded devices is considered unfeasible due to unpredictable termination times and incalculable loss of early stopping. In this regard, a provisional solution aimed at large-scale measurement campaigns and utilizing black-box data-driven machine learning techniques. While this approach fits well with many imaging problems, it was soon stripped of its enchantment for communication systems due to the necessity of measuring and preprocessing RF signals under diverse sets of environmental conditions resulting in extremely large training times and disappointing performance gains. In addition, there is still no commonly accepted neural network de-facto standard or baseline architecture for particular communication problems akin to AlexNet or GoogleNet in the imaging domains. One step to close this important gap was made in [29] by using multidimensional Laplace transform techniques to design optimal neural networks for particular sparse recovery problems revealing a very intriguing connection between commonly employed neural networks comprising weights, threshold functions, rectified linear (ReLU) and rectified polynomial (ReP) activation functions and volume and centroid computation problems over sparsity inducing sets. We refer the reader to Figure 4 for a geometric illustration of a small sparse recovery problem. Here, x is to be recovered from dimensionality reduced measurement y = Ax given that x belongs to a particular sparsity inducing set (blue). Then, the neural network of [29] outputs the estimate $\hat{x}$ that minimizes the expected error over the uncertainty set (intersection between red and blue). Using such geometric ideas in the design of neural networks allows for bypassing costly search over exponential candidate networks that consume large portions of available computing resources. Indeed, practitioners can still apply fine-tuning to reduce a possible model mismatch and reduce reconstruction errors even further.

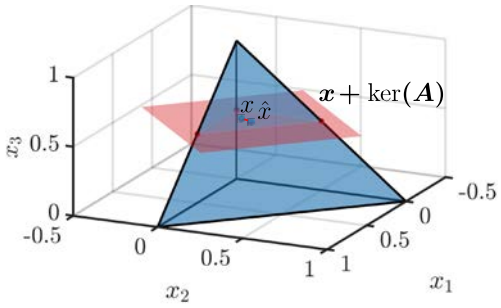


Fig. 4. Geometry of sparse recovery.

4. FUTURE RESEARCH TOPICS

This clause discusses four future research topics in machine learning and communications.

4.1. Low complexity models

State-of-the-art machine learning models such as deep neural networks are known to work excellently in practice. However, since the training and execution of these models require extensive computational resources, they may not be applicable in communication systems with limited storage capabilities, computational power and energy resources, e.g. smartphones, embedded systems or IoT devices. Recent work addressed this problem and proposed techniques for reducing the complexity of deep neural networks. For instance, the authors of [22] demonstrated that the size of VGG-16, a popular deep neural network for image classification, can be reduced by over 95% with no loss of accuracy. In communication applications such compression techniques can be used to store and transmit models efficiently. Other authors (e.g. [15]) targeted the problem of weight binarization in deep neural networks. This type of discretization can be useful, e.g. when adapting models to processor architectures which do not allow floating point operations.
Further research on these topics is of high importance as it can be expected that a large number of new applications would emerge, if the complexity of state-of-the-art models can be reduced to a level, which allows their use in computationally limited environments at minimal performance loss.

4.2. Standardized formats for machine learning

The standardization of algorithms and data formats is of high importance in communications, because it increases the reliability, interoperability and modularity of a system and its respective components. With the increasing use of learning algorithms in communication applications, the need for standardized formats for machine learning is also rising.
For instance, standardized formats could be used to specify how to train, adapt, compress and exchange machine learning models in communication applications. Furthermore, there could be standardized formats for the data and standards which determine how multiple machine learning models interact with each other. Other formats could be specifically designed for ensuring that a model fulfills certain security or privacy requirements.

4.3. Security and privacy mechanisms

Machine learning models are often used in a black-box manner in today's applications. This prevents the human expert from comprehending the reasoning of the algorithm and from validating its predictions. Although recent works [3, 33] proposed techniques for explaining the predictions of a machine learning model, further research on this topic is of high importance as the lack of transparency can be a large disadvantage in communication applications.
Moreover, it is well-known that deep neural networks can be easily fooled or may behave in an unexpected way when being confronted with data with different properties to the data used for training the model [43]. Thus, the establishment of mechanisms which increase the reliability of the model is a prerequisite for large-scale use in communication applications. Such mechanisms can be implemented on different levels, e.g. be an integral part of the model, be integrated into the communication protocol or be part of a separate inspection process.
Besides interpretability and security aspects, future research also needs to investigate how to effectively encrypt machine learning models and how to ensure data privacy during and after learning.

4.4. Radio resource and network management

The end-to-end performance of mobile networks is strongly influenced by the choice of radio resource (e.g. beamforming and medium access control parameters) and network management (e.g. handover parameters, neighborhood lists, loads and power budgets) parameters. Moreover, some of the parameters must be continuously adapted on a relatively short time scale to time-varying radio propagation conditions and changing network topologies [42].
Current approaches are inadequate to cope with the growth of autonomous network elements in 5G small cell deployments based on mobile cloud RAN architectures. Therefore, 5G networks call for new model and data-driven radio resource management and network management methods that are augmented by machine learning techniques for extracting knowledge from the system and gradual learning in the presence of inherent uncertainties and the lack of complete channel and network state information [38]. The realization of these ideas in the context of 5G will require modifications of existing protocols and the development of new ones.

5. CONCLUSION

This paper discussed the increasing mutual influence of machine learning and communication technology. Learning algorithms were not only shown to excel in traditional network management tasks such as routing, channel estimation or PAPR reduction, but also to be a core part of many emerging application fields of communication technology, e.g. smart cities or Internet of things. The availability of large amounts of data and recent improvements in deep learning methodology will further foster the convergence of these two fields and will offer new ways to optimize the whole communication pipeline in an end-to-end manner [36].
However, before resource-intensive models such as deep neural networks can be applied on a large scale in communication applications, several practical challenges (e.g. complexity, security, privacy) need to be solved. Furthermore, more research is required on theoretical topics at the intersection of communications and machine learning, e.g. incremental learning, learning in non-stationary environments or learning with side information.

REFERENCES

[1] R. Agrawal and R. Srikant. Privacy-preserving data mining. In ACM SIGMOD Manag. Data, pages 439– 450, 2000.

[2] O. G. Aliu, A. Imran, M. A. Imran, and B. Evans. A survey of self organisation in future cellular networks. IEEE Commun. Surveys Tuts., 15(1):336–361, 2013.

[3] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS ONE, 10(7):e0130140, 2015.

[4] R. B. Basnet, S. Mukkamala, and A. H. Sung. Detection of phishing attacks: A machine learning approach. Soft Comp. Appl. Indust., 226:373–383, 2008.

[5] S. Bosse, D. Maniry, T. Wiegand, and W. Samek. A deep neural network for image quality assessment. In IEEE ICIP, pages 3773–3777, 2016.

[6] S. Bosse, K.-R. Müller, T. Wiegand, and W. Samek. Brain-computer interfacing for multimedia quality assessment. In IEEE Syst. Man Cyber., pages 002834– 002839, 2016.

[7] J. A. Boyan and M. L. Littman. Packet routing in dynamically changing networks: A reinforcement learning approach. NIPS, pages 671–671, 1994.

[8] R. L. Cavalcante and S. Stanczak. A distributed subgradient method for dynamic convex optimization problems under noisy information exchange. IEEE J. Sel. Topics Signal Process., 7(2):243–256, 2013.

[9] R. L. G. Cavalcante and S. Stanczak. Robust settheoretic distributed detection in diffusion networks. In IEEE ICASSP, 2012.

[10] R. L. G. Cavalcante and I. Yamada. A flexible peak-toaverage power ratio reduction scheme for ofdm systems by the adaptive projected subgradient method. IEEE Trans. Signal Process., 57(4):1456–1468, 2009.

[11] R. L. G. Cavalcante, I. Yamada, and B. Mulgrew. An adaptive projected subgradient approach to learning in diffusion networks. IEEE Trans. Signal Process., 57(7):2762–2774, 2009.

[12] Y. Chen, X. Yang, B. Zhong, S. Pan, D. Chen, and H. Zhang. Cnntracker: online discriminative object tracking via deep convolutional neural network. Appl. Soft Comput., 38:1088–1098, 2016.

[13] C.-H. Cheng, Y.-H. Huang, and H.-C. Chen. Channel estimation in ofdm systems using neural network technology combined with a genetic algorithm. Soft Comput., 20(10):4139–4148, 2016.

[14] L. Ciabattoni, G. Ippoliti, A. Benini, S. Longhi, and M. Pirro. Design of a home energy management system by online neural networks. IFAC Proc. Volumes, 46(11):677–682, 2013.

[15] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio. Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or -1. arXiv:1602.02830, 2016.

[16] M. Egmont-Petersen, D. de Ridder, and H. Handels. Image processing with neural networksa review. Pattern Recognit., 35(10):2279–2301, 2002.

[17] A. Eswaradass, X.-H. Sun, and M. Wu. A neural network based predictive mechanism for available bandwidth. In IPDPS, pages 33a–33a, 2005.

[18] Z. M. Fadlullah, M. M. Fouda, N. Kato, X. Shen, and Y. Nozaki. An early warning system against malicious activities for smart grid communications. IEEE Netw., 25(5), 2011.

[19] A. Galindo-Serrano and L. Giupponi. Distributed qlearning for aggregated interference control in cognitive radio networks. IEEE Trans. Veh. Technol., 59(4):1823–1834, 2010.

[20] L. Gubin, B. Polyak, and E. Raik. The method of projections for finding the common point of convex sets. USSR Comput. Math. Math. Phys., 7(6):1–24, 1967.

[21] T. S. Guzella and W. M. Caminhas. A review of machine learning approaches to spam filtering. Expert Syst. Appl., 36(7):10206–10222, 2009.

[22] S. Han, H. Mao, and W. J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv:1510.00149, 2015.

[23] Y. Jabrane, V. P. G. Jimenez, A. G. Armada, B. A. E.´ Said, and A. A. Ouahman. Reduction of power envelope fluctuations in ofdm signals by using neural networks. IEEE Commun. Lett., 14(7):599–601, 2010.

[24] J. Jiang. Image compression with neural networks–a survey. Signal Process. Image, 14(9):737–760, 1999.

[25] J. Joung. Machine learning-based antenna selection in wireless communications. IEEE Commun. Lett., 20(11):2241–2244, 2016.

[26] M. Kasparick, R. L. G. Cavalcante, S. Valentin, S. Stanczak, and M. Yukawa. Kernel-based adaptive online reconstruction of coverage maps with side information. IEEE Trans. Veh. Technol., 65(7):5461–5473, 2016.

[27] S. Kumar and R. Miikkulainen. Dual reinforcement q-routing: An on-line adaptive routing algorithm. In ICANNE, 1997.

[28] Y. Liang. Real-time vbr video traffic prediction for dynamic bandwidth allocation. IEEE Trans. Syst., Man, Cybern. C, 34(1):32–47, 2004.

[29] S. Limmer and S. Stanczak. Optimal deep neural networks for sparse recovery via laplace techniques. arXiv, 2017.

[30] T. Lu and J. Zhu. Genetic algorithm for energy-efficient qos multicast routing. IEEE Commun. Lett., 17(1):31– 34, 2013.

[31] P. Mertikopoulos and A. L. Moustakas. Learning in an uncertain world: Mimo covariance matrix optimization with imperfect feedback. IEEE Trans. Signal Proc., 64(1):5–18, 2016.

[32] J. Mo, P. Schniter, N. G. Prelcic, and R. W. Heath. Channel estimation in millimeter wave mimo systems with one-bit quantization. In Asilomar Signals Syst. Comput., pages 957–961, 2014.

[33] G. Montavon, W. Samek, and K.-R. Müller. Methods for interpreting and understanding deep neural networks. arXiv:1706.07979, 2017.

[34] M. Munetomo, Y. Takai, and Y. Sato. A migration scheme for the genetic adaptive routing algorithm. In IEEE Sys. Man Cyber., volume 3, pages 2774–2779, 1998.

[35] T. T. Nguyen and G. Armitage. A survey of techniques for internet traffic classification using machine learning. IEEE Commun. Surveys Tuts., 10(4):56–76, 2008.

[36] T. J. O'Shea, T. Erpek, and T. C. Clancy. Deep learning based mimo communications. arXiv:1707.07980, 2017.

[37] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos. Context aware computing for the internet of things: A survey. IEEE Commun. Surveys Tuts., 16(1):414–454, 2014.

[38] T. W. Rondeau. Application of Artificial Intelligence to Wireless Communications. PhD thesis, Virginia Tech, 2007.

[39] M. Sanchez-Fernandez, M. de Prado-Cumplido, J. Arenas-Garc´ıa, and F. Perez-Cruz. Svm multiregres-´ sion for nonlinear channel estimation in multiple-input multiple-output systems. IEEE Trans. Signal Process., 52(8):2298–2307, 2004.

[40] Skog and P. Handel. In-car positioning and navigation technologiesa survey. IEEE Trans. Intell. Transp. Syst., 10(1):4–21, 2009.

[41] V. Srinivasan, S. Lapuschkin, C. Hellge, K.-R. Müller, and W. Samek. Interpretable human action recognition in compressed domain. In IEEE ICASSP, pages 1692– 1696, 2017.

[42] S. Stanczak, M. Wiczanowski, and H. Boche. Fundamentals of resource allocation in wireless networks: theory and algorithms, volume 3. Springer, 2009.

[43] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. arXiv:1312.6199, 2013.

[44] M. Thilina, K. W. Choi, N. Saquib, and E. Hossain. Machine learning techniques for cooperative spectrum sensing in cognitive radio networks. IEEE J. Sel. Areas Commun., 31(11):2209–2221, 2013.

[45] C.-F. Tsai, Y.-F. Hsu, C.-Y. Lin, and W.-Y. Lin. Intrusion detection by machine learning: A review. Expert Syst. Appl., 36(10):11994–12000, 2009.

[46] C.-W. Tsai, C.-F. Lai, M.-C. Chiang, L. T. Yang, et al. Data mining for internet of things: A survey. IEEE Commun. Surveys Tuts., 16(1):77–97, 2014.

[47] P. J. Werbos. Computational intelligence for the smart grid-history, challenges, and opportunities. IEEE Comput. Intell. Mag., 6(3):14–21, 2011.

[48] M. Wiering. Multi-agent reinforcement learning for traffic light control. In ICML, pages 1151–1158, 2000.

[49] Xu, M. Zhao, J. Fortes, R. Carpenter, and M. Yousif. Autonomic resource management in virtualized data centers using fuzzy logic-based approaches. Cluster Comput., 11(3):213–227, 2008.

[50] Yamada and N. Ogura. Adaptive projected subgradient method for asymptotic minimization of sequence of nonnegative convex functions. Num. Func. Analys. Optim., 25(7/8):593–617, 2005.

[51] Yukawa. Multikernel adaptive filtering. IEEE Trans. Signal Process., 60(9):4672–4682, 2012.

[52] S. Yun and C. Caramanis. Reinforcement learning for link adaptation in mimo-ofdm wireless systems. In GLOBECOM, pages 1–5, 2010.

[53] Zhang, Y. Zhuo, and Y. Zhao. Mobile location based on svm in mimo communication systems. In ICINA, volume 2, pages V2–360, 2010.

[54] Q. Zhang and Y.-W. Leung. An orthogonal genetic algorithm for multimedia multicast routing. IEEE Trans. Evol. Comput., 3(1):53–62, 1999.

[55] Y. Zhang, S. Kwong, X. Wang, H. Yuan, Z. Pan, and L. Xu. Machine learning-based coding unit depth decisions for flexible complexity allocation in high efficiency video coding. IEEE Trans. Image Process., 24(7):2225–2238, 2015.

APPLICATION OF AI TO MOBILE NETWORK OPERATION

Tomoyuki Otani[1], Hideki Toube[2], Tatsuya Kimura[1], Masanori Furutani[1]
[1]DOCOMO Technology INC., [2]NTT DOCOMO INC., Japan

Abstract – With the introduction of network virtualization and the implementation of 5G/IoT, mobile networks will offer more diversified services and be more complex. This raises a concern about a significant rise in network operation workload. Meanwhile, artificial intelligence (AI) technology is making remarkable progress and is expected to solve human resource shortages in various fields. Likewise, the mobile industry is gaining momentum toward the application of AI to network operation to improve the efficiency of mobile network operation [1][2].This paper will discuss the possibility of applying AI technology to network operation and presents some use cases to show good prospects for AI-driven network operation.

Keywords – Artificial intelligence (AI), mobile network, network operation

1. INTRODUCTION

1.1. Characteristics of artificial intelligence (AI)

Businesses today are faced with a great deal of information and a shortage of human resources. To solve these issues, they are introducing AI/big data technologies at an accelerated pace to improve the efficiency of business operation.
As the world today is experiencing the third AI boom, the practical use of "machine learning" helps us to automatically identify and learn the patterns and rules from large amounts of data (the so-called "big data") based on specific criteria. This technology will enable us to set more appropriate rules based on learning from data and make more accurate judgments once evaluation criteria are given. Following this, “deep learning” has emerged, in which AI learns by itself and accumulates the knowledge of patterns and rules with no specific criteria given.
According to the 2016 White Paper on Information and Communications in Japan published by the Ministry of Internal Affairs and Communications, there are three major types of functions that AI plays in actual services: "identification", "prediction" and "execution" [3]. These functions can be utilized and applied across all industries. Standard usages of each function are as follows:

- "Identify" the current situation (characteristics) from a large amount of data (big data).
- Analyze the time characteristics of the data and "predict" future tendencies.
- Make and "execute" an optimum plan which is based on the "identified"/"predicted" data.

As shown in the white paper, AI is expected to be utilized in various business fields and perform highly advanced analysis (for improving operational efficiency), using big data in a short period of time without human manual operation

1.2. Trends of mobile network

Today, the application of virtualized network functions (NFV) to mobile core networks is in progress [4]. For years, various network functions, such as the conventional Evolved Packet Core (EPC), have been provided in their dedicated hardware (HW) such as Advanced Telecom Computing Architecture (ATCA) hardware. With the introduction of NFV, software will be able to run on a virtualized operating system (OS) of generic Intel architecture (IA) servers and be provided separately from hardware [3]. Furthermore, the NFV architecture enables integrated management and control (orchestration) of network services and resources, interworking with Management and Orchestration (MANO) and a software-defined network (SDN).
These technologies will enable most of the construction and increase/decrease work of network elements (NEs) to be performed remotely by software control without manual operation. In

addition, once software and hardware are separated, they can be constructed and increased or reduced independently at different timings. For instance, once the hardware, e.g. the IA server, is prepared, software can be readily increased/reduced on the server. Resources can be allocated whenever they are needed to accommodate the traffic, reducing the surplus network facilities (Fig. 1.1). In the conventional method, network facilities (blue line) are established in advance based on the specific network traffic volume predicted (red line). In contrast, in the virtualized network, lots of equipment units can be flexibly arranged as virtual machines (VMs) on the servers/cloud and such VMs can be allocated dynamically to allow dynamic changes in the network resources allocated.

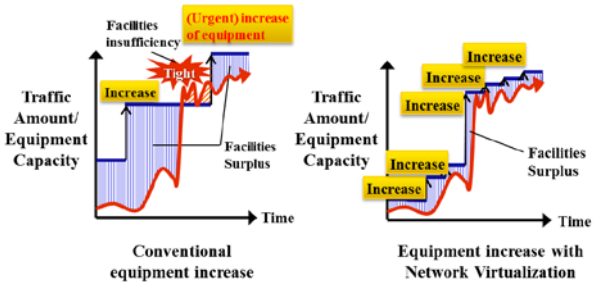


Fig. 1.1. Network resource allocation with network virtualization

Moreover, the introduction of 5G scheduled for 2020 will bring with it a variety of network services (such as IoT) with traffic characteristics and network requirements different from those of existing smartphones [5].

For example, a factory equipment monitoring system, in which mass data from sensors are sent to the administrator periodically, will require a highly secure and reliable network (no data loss). On the other hand, a traffic system, in which inter-vehicle communication prevents traffic congestion, will require a low-latency and highly secure network. Although the networks with different requirements need to be separated, it is inefficient to do so physically. Instead, we need to utilize the network slicing technology to virtually construct separate networks with different requirements on the same physical network.

With new technologies such as NFV and the network slicing technology mentioned above, the physical hardware configuration will remain unchanged or become simpler. However, this will bring about more issues for manual network operation, as the logical network configuration used for service provision will become complicated due to the use of multiple virtualized logical resources.

2. ISSUES OF MOBILE NETWORK OPERATION

Network operation in general can be illustrated as a cycle of activities that consist of planning, construction and maintenance, as shown in Fig. 2.1.

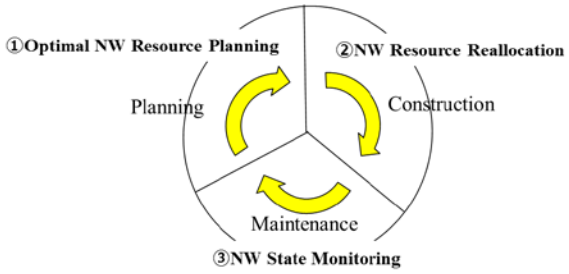


Fig. 2.1. Network operation cycle

In light of conventional mobile network trends, whether to achieve the following objectives can have an impact on the network operation:

1) accelerate the network operation cycle to provide services quickly; and
2) establish an analysis and operation method for advanced and complicated networks.

The issues towards achieving the above objectives are described below.

Planning is the step to analyze the network traffic and formulate future plans for network facility resources. The main activity of the conventional planning process is the analytic work comparable "prediction," which is performed to analyze the past network traffic and calculate the amounts of long-term (yearly) network resources (such as bandwidths). However, when network virtualization and various 5G/IoT services come into play and different services are superimposed on the same physical network/ hardware, it will be necessary to factor in the traffic of all services in traffic volume prediction for a single network/hardware unit. This will make the analysis very difficult and time consuming if it is done manually by humans or based on the expertise of specialists as before. Specific issues for accelerating the planning process are described in clause 2.1.

Generally speaking, the construction process involves designing, building, and testing the network equipment. When network virtualization is applied, it will be possible to increase or reduce the number of VMs used as NE equipment at any time if IA servers are in place as hardware. The network resources can be created by software operation only.

As construction can be completed simply by setting VM servers, the time for the construction will be shortened remarkably, accelerating the construction process.
The maintenance process is the step taken to analyze the impact of problems on customer services based on alarm information from the network, recover services in a way appropriate to the details of impact, and identify the faulty equipment and fix the problem using the alarm information. With the progress of NFV and network slicing technology, the network configuration for customer service provision is divided into two: the logical network configuration and the hardware configuration that configures the logical network. These network configurations will make it difficult to quickly and accurately analyze/identify the impact on customer services at the time of hardware failure.
The application of network virtualization to the construction process will accelerate the cycle of network operation. On the other hand, the planning and maintenance processes are facing some issues, which will be discussed in more specific terms in the following subclause.

2.1. Issues of planning process

For conventional traffic prediction, it has been common to predict the volume of traffic based on the time-series analysis of the measured traffic volume. However, predicting the traffic volume accurately through the time-series analysis is becoming difficult due to a dynamic variation of factors contributing to the generation of traffic, such as the emergence of new applications/contents and temporary concentration of user population associated with events.
For example, traffic prediction based on time-series analysis cannot cope with such disturbing factors as temporary traffic surges associated with events. Therefore, we need traffic tendency analysis that has eliminated those disturbing factors and other anomalous aspects responsible for the significant deviation between prediction and reality. In addition, we also need to analyze short-term traffic tendencies such as temporary traffic spikes in order to realize the network resource allocation in line with the volume of communication traffic as shown in Fig.1.1. However, accurate analysis as mentioned above requires not only an enormous analysis workload but also different sets of special skills for developing long/short-term models. Furthermore, an assumption of various services such as those of 5G/IoT presents the limitation of human analysis regarding anomalous traffic tendencies for all services.

2.2. Issues of maintenance process

The conventional network has a fixed mapping relation between the logical network configuration for service provision and the hardware that configures the logical network. When hardware fails, this fixed relation has allowed maintenance staff to analyze/identify the affected services by looking at the service topology information consisting of hardware and network configuration information as well as the equipment alarm information sent from the hardware.
With the progress of NFV and network slicing technology, the logical network configuration for service provision will include multiple virtualized logical resources. There will be a various way of service provision. In one case, the same service will be configured on different hardware every day. In another case, the same hardware may provide a different service every day. As a result, it will be difficult to achieve quick and accurate analysis with the conventional method, which analyzes/identifies the impact on customer services manually by humans based on the alarms from various equipment units.

3. NETWORK OPERATION WITH AI

The application of AI will enable us to respond to the above-mentioned problems in the planning and maintenance processes quickly and efficiently even when sufficient human resources, experience and special skills are not available. So, we are aiming to apply AI to the planning/maintenance processes in order to conduct more efficient and advanced analysis work for the planning/ maintenance tasks.
This clause qualitatively explains how the application of AI will make network operation effective.

3.1. Approach to applying AI to planning process

This section explains how AI is applied to traffic demand prediction during the planning process.
We use AI to predict and analyze traffic demand. Basically, traffic tendencies can be divided into two types: short-term tendencies, such as temporary traffic increases during events; and long-term tendencies, from which anomalous tendencies such as temporary traffic increases during events have been removed. In AI-driven traffic prediction, we make AI learn the short-/long-term traffic tendencies that have respectively totally different factors causing traffic fluctuations and mechanisms in order

to predict both short-/long-term traffic demands. An example of AI-driven traffic demand prediction is shown in Fig. 3.1.

1) Input traffic information and date and time information of events in each area as the data for AI learning.
2) Have AI learn long-term traffic increase tendencies excluding temporary traffic surges such as those during events (for generation of long-term traffic demand prediction model). Furthermore, have AI learn the correlation between the time of events and temporary traffic surges mentioned above to learn the occurrence tendencies (timing, increment) of short-term traffic increases (due to events, etc.) (for generation of temporary traffic demand prediction model).
3) Have AI output long-/short-term traffic demand prediction models for each area.

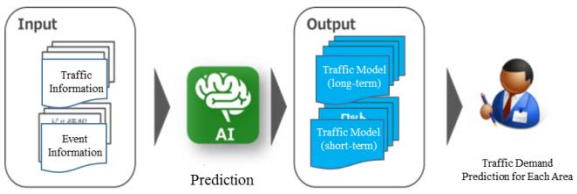


Fig. 3.1. Process of traffic demand prediction with AI

3.2. Approach to applying AI to maintenance process

This section explains the application of AI to network monitoring (network abnormality detection).

3.2.1. Application of AI to network monitoring

3.2.1.1. Necessity of service monitoring

As described in clause 2.2, the network maintenance process will become complicated. Above all, it will be much more difficult to analyze and grasp the impact on customer services accurately in the network monitoring process, in which speed is the key.

With the conventional method of identifying the impact on customer services from equipment alarm information, it is difficult to identify the impact when no alarm is given and information is insufficient, especially in the case of silent alarms.

In addition, as the conventional method often relies on the skill and expertise of maintenance staff, sometimes the impact on service cannot be reasonably grasped from the standpoint of customer experience.

Under such circumstances, it is necessary to realize a service monitoring system that is able to detect the quality close to what customers are actually experiencing, estimating it from the network data, independent of alarm information. Image of Service monitoring is shown in Fig. 3.2.

This can be achieved by collecting various data from the network, including equipment alarm information and integrating them into the big data for multi-angle analysis (AI analysis).

If service monitoring is realized, the maintenance process will be innovated from facility maintenance centered on device alarm to network quality maintenance based on customer experience. Taking the initial reaction amount, for example in the conventional method, it corresponds to the number of alarms notified from the network. While in the service monitoring method, it corresponds to the number of influences on customer service caused by the equipment failure, which is expected to be reduced drastically.

Regarding the workload of fixing the failure from the maintenance viewpoint, since the necessity of repairing the equipment is irrelevant to the service effect occurrence, there will not be a great difference in the maintenance workload. In the maintenance process, there is also a possibility to arrange the maintenance automatically by converting the device alarm information and work log into big data for machine learning.

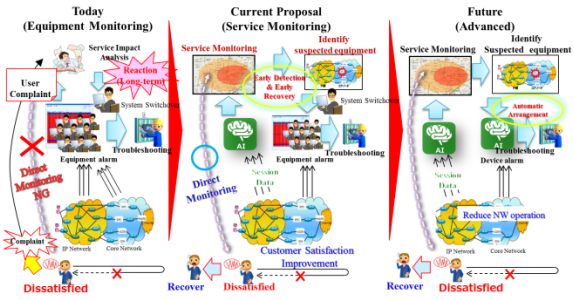


Fig. 3.2. Image of Service monitoring method

3.2.1.2. Application of AI to service monitoring

In order to realize the network service monitoring based on customer experience independent of equipment alarm information, a method of analysis that consists of the following two stages can be used:

(A) Have AI analyze the network data and estimate the index which can express the quality of customer experience (QoE) for each service [6];
(B) Collect the estimated QoE for each service provision area (cell/eNB area) and have AI

learn the feature value of distribution. Have AI detect any unusual state (deterioration of network quality) based on the feature value and make a judgement on the distribution state to decide whether it is "different from the usual state or not" [1], [7].

The illustration of the proposed method is shown in Fig. 3.3.

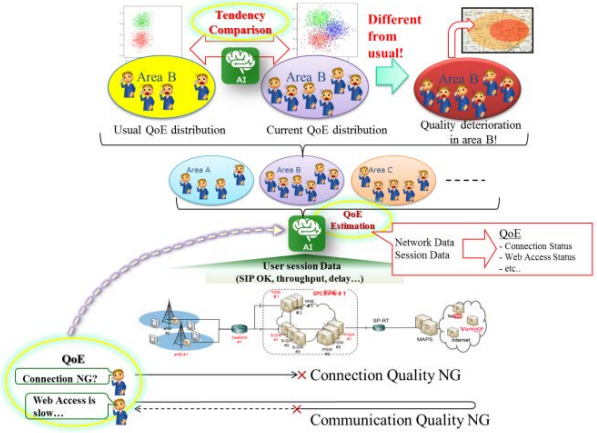


Fig. 3.3. Apply AI to service monitoring method

The quality of customer experience (QoE) estimation and service abnormality determination methods are described below.

(A) (QoE) estimation

QoE is an index of customers' subjective feeling (comfortable, clean, etc.), which is difficult to measure directly on the network.

Therefore, we use another quality index, for example an application quality of service (QoS) such as webpage loading time as an alternative for QoE. As it is also difficult to measure the application QoS on the network directly, we estimate the application QoS from the network QoS (TCP throughput, etc.), which can be measured on the network, and use the estimated application QoS as a substitute of QoE (Fig. 3.4).

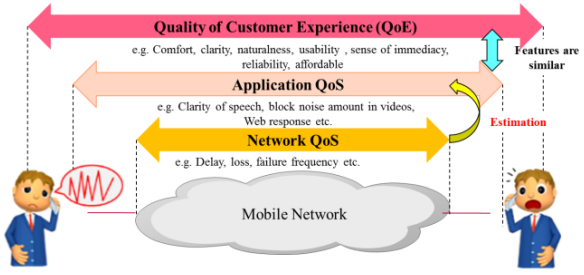


Fig. 3.4. Relation between QoS and QoE

The following steps explain the method of using AI analysis to estimate the application QoS from the network QoS obtainable from the network, using the loading time of web access as an example. (Hereinafter, QoE means application QoS.)

(1) Measure and acquire information on the actual webpage loading time by the test terminal, etc.as training data.
(2) Collect communication quality information at network level observable on the network (throughput, etc.) or equipment traffic information (statistical values such as the number of established connections for each equipment unit). Have AI learn the information's correlation with the webpage loading time mentioned above in (1) and detect highly-correlated information from the web access session information and equipment traffic information. Then, have AI generate an analysis model (AI model) for calculating an estimated value of webpage loading time, using high-correlated information (in other words, information which can be acquired on the network side).
(3) Using the generated AI model, estimate the webpage loading time from the communication quality information at a network level (throughput, etc.) or equipment traffic information.

The explanation above uses an example of communication quality in which QoE is difficult to measure. In contrast, when the connection quality is determined simply by whether the connection is OK (successful) or NG (unsuccessful), such network information (disconnection information) can be directly applied as the QoE index. In this case, AI analysis is not necessarily needed for estimating QoE.

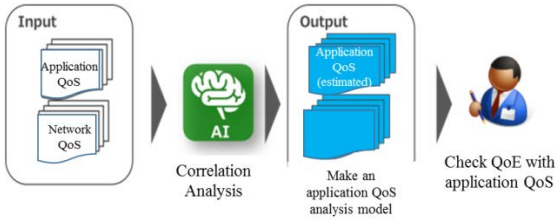


Fig. 3.5. Process of QoE estimation with AI

(B) QoE anomaly analysis

Using the following steps, we explain service quality abnormality analysis in which quality degradation caused by the network is detected based on the QoE index output in (A).

In a mobile network, where resource utilization changes greatly according to the mobility of user population, it is highly likely that QoE will also fluctuate by the time and place. Taking this into account, we are studying an analysis method which consists of the steps as described below.

(1) Classify the collected QoE data into macroscopic observation units, which are in this case communication areas. Have the AI learn the distribution of feature amounts in the QoE data observed in each area in time series for both cases in which the network is in normal and abnormal states. Generate the AI analysis model with the QoE distribution (steady state) for each area when the network is normal. (In the distribution learning of QoE, there will be a large amount of analysis if every individual QoE value is used. Therefore, it will be more effective to learn and determine with the statistical values generated in advance, such as the maximum, minimum, average and dispersion of QoE for each area, which represents a macroscopically aggregated unit.)

(2) With the AI analysis model, create a clustering of normal and abnormal states from the QoE information collected regularly. Determine the service quality as abnormal when detecting a deviation tendency from the steady state.

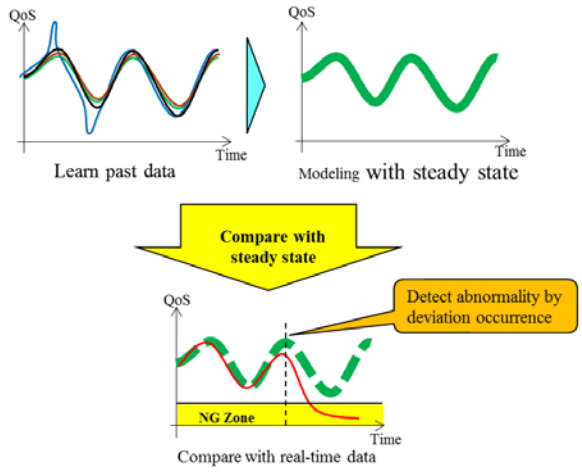


Fig. 3.6. Abnormality determination in comparison with steady state

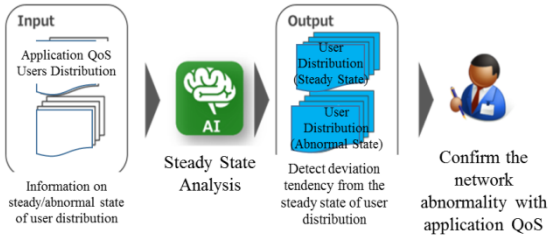


Fig. 3.7. Process of service quality abnormality determination with AI

Moreover, in this method, in which the presence of users at the edge of a service area is also taken into consideration; any case that deviates from the normal network state is considered to be under the influence of network abnormality (abnormality of the network itself or another network) and only service quality degradation caused by network abnormality can be detected.

4. CONCLUSION

In this article, we described the possibility of applying AI technology to mobile network operation and presented some use cases to show significant benefit from AI-driven network operation.

The study on the application of AI to mobile network operation in the telecommunications field is still in its infancy and there has been no report of a commercial network which has actually introduced AI to its operation. Based on the concept described in clause 3, we plan to verify technical aspects of AI application in a trial environment in which commercial network data is available and aim to realize network operation taking advantage of AI.

REFERENCES

[1] "Development of NTT Group AI technology "corevo™" for network operation, the AI technology able to autonomously derive the correlation between the "causes of faults and alarms" for highly accurate detection of network faults. http://www.ntt.co.jp/news2017/1701/170130a.html

[2] "KDDI R&D Laboratories completes World's first successful AI-assisted automated network operation system PoC" ., News release on Feb. 22 2016. https://wirelesswire.jp/2016/02/50505/

[3] "Present and Future of Artificial Intelligence (AI)," White Paper on Information and Communications in Japan 2016, Chapter 4, Section 2.

[4] Hiroyuki Oto, "Application of Virtualization Technology to Communications Networks," NTT DOCOMO Technical Journal Vol. 18 No. 1 pp. 4-10.

[5] Takuya Shimojo., "Future Core Network for the 5G Era," NTT DOCOMO Technical Journal Vol. 17 No. 4 pp. 50-59.

[6] Takanori Hayashi, "QoE-centric Operation for Optimizing User Quality of Experience." https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201509fa3.html

[7] "MasterScope: Features and Experimental Applications of System Invariant Analysis Technology" (NEC Technical Journal Vol. 65 No. 2/2012).

ON ADAPTIVE NEURO-FUZZY MODEL FOR PATH LOSS PREDICTION IN THE VHF BAND

Nazmat T. Surajudeen-Bakinde[1], Nasir Faruk[2], Muhammed Salman[1], Segun Popoola[3], Abdulkarim Oloyede[2], Lukman A. Olawoyin[2]
[1]Department of Electrical and Electronics Engineering, University of Ilorin, Nigeria
[2]Department of Telecommunication Science, University of Ilorin, Ilorin, Nigeria
[3]Department of Electrical and Information Engineering, Covenant University, Ota, Nigeria
Email: engrsalman011@gmail.com; segun.popoola@stu.cu.edu.ng; faruk.n, deenmat, oloyede.aa, olawoyin.la{@unilorin.edu.ng}

Abstract – Path loss prediction models are essential in the planning of wireless systems, particularly in built-up environments. However, the efficacies of the empirical models depend on the local ambient characteristics of the propagation environments. This paper introduces artificial intelligence in path loss prediction in the VHF band by proposing an adaptive neuro-fuzzy (NF) model. The model uses five-layer optimized NF network based on back propagation gradient descent algorithm and least square errors estimate. Electromagnetic field strengths from the transmitter of the NTA Ilorin, which operates at a frequency of 203.25 MHz, were measured along four routes. The prediction results of the proposed model were compared to those obtained via the widely used empirical models. The performances of the models were evaluated using the Root Mean Square Error (RMSE), Spread Corrected RMSE (SC-RMSE), Mean Error (ME), and Standard Deviation Error (SDE), relative to the measured data. Across all the routes covered in this study, the proposed NF model produced the lowest RMSE and ME, while the SDE and the SC-RMSE were dependent on the terrain and clutter covers of the routes. Thus, the efficacy of the adaptive NF model was validated and can be used for effective coverage and interference planning

Keywords – COST 231, generalized bell, Hata, neuro-fuzzy, path loss

1. INTRODUCTION

Predicting the propagation of electromagnetic waves is of great significance in the design and planning of wireless communication systems. Propagation models are essential in evaluating the performance of a wireless system and quality of the received signal. Empirical path loss models have been found to be the widely used models due to their simplicity and ease of use, as the implementation of the models do not require much computational efforts, and, are less responsive to the physical and geometrical structures of the environments [1]. These make them attractive, although a major drawback of utilizing these models is the lack of accuracy, especially when deployed in another environment other than the one where the measurement was taken. For example in [2-5], several of these models were tested in a typical urban and rural Nigerian terrain and they were found to be inconsistent in prediction, and have high prediction errors. Although, in [6-8], some of the most performing models, were tuned to minimize errors and improve prediction accuracy. Yet the tuned models were found to be site-specific. On the other hand, the deterministic models seem to have better prediction accuracy because of the availability of detailed information about the propagation environment. However, they are computationally intensive and time consuming [9]. Moreover, despite the inclusion of site-specific information, the deterministic models do not always provide more accurate predictions than the empirical models [9-10]. This therefore raises more questions to which model(s) can provide optimum prediction with minimal complexity, as such, the need to incorporate artificial intelligence (AI) and heuristic algorithms to improve path loss prediction. Different AI techniques have been adopted, as evident in the literature, for path loss prediction. Tamma *et al* [11] developed an artificial neural network (ANN) model for path loss prediction in the UHF (ultra-high frequency) band based on the measurement data collected in Tripoli, Turkey. The accuracy of the proposed model was evaluated and compared to that of the Hata model and it was found that the ANN model provided more accurate prediction. In [12], an adaptive network based fuzzy inference system (ANFIS) was used to predict path loss in the urban settlement of the Habiye region of Istanbul in the 900 MHz band. The ANFIS model

increased the prediction accuracy by 15% relative to the Bertoni-Walfisch model. Joel and Elmer in [13] conducted a comparative analysis of the neural network (NN), free space loss (FSL), and Egli models. The NN model was most efficient for path loss prediction in digital TV macro-cells in the UHF band. Achieving the lowest prediction error using the AI models was not the only benefit over the empirical path loss propagation models as Ozdemir *et al* [14] showed that the ANN model performed better than the theoretical and the empirical propagation models in terms of prediction accuracy, less complexity and time. Furthermore, the authors proved that within the ANN model, the model that employed the Levenberg-Marquardt learning algorithm had minimal prediction error compared to the one that used Epstein-Peterson. In [15] the performance of ANFIS for optimal power control for cognitive radio spectrum distribution was investigated. ANFIS produced the lowest prediction error and was recommended as the most suitable method for power scale control. Vishal and Sharma [16] employed a fuzzy logic (FL) model to predict path loss as a function of path loss exponent in the fringe areas of a suburban region of Clementown and Dehradun in India. It is worth mentioning that some efforts have also been made to compare the performance of some AI models used in path loss prediction. For instance, Vahala et *al.*, [17] investigated the electromagnetic interference pattern caused by portable devices onboard Airbus 319 and 320 with respect to various receivers on the aircraft using NF modelling (NFM). The results obtained were compared to the ANN model and it was found that the NFM performed better.

Although application of heuristic algorithms for path loss prediction in an urban macro-cellular environment [18-21] is gaining momentum, most of the works that focus on investigating the suitability of adaptive NF technique for path loss prediction in the VHF band are very limited. Moreover, due to the peculiar nature of our terrain environment and the wide deployments of wireless systems operating on the VHF bands, there is a need to test the efficacy of the NF model. Therefore, this paper introduces an adaptive neuro-fuzzy (NF) approach to path loss prediction in the VHF band within the Nigerian propagation terrain context. The predictions of the NF model were compared to those of the widely used empirical models such as Hata, COST 231, Egli and ECC-33 models. The performances of the models under investigation were evaluated using the Root Mean Square Error (RMSE), Spread Corrected RMSE (SC-RMSE), Mean Error (ME), and Standard Deviation (SD), relative to the measured data.

2. METHODOLOGY

This section is divided into two parts: the first part describes the measurement procedure and the second part explains the adaptive NF approach to path loss modelling in the VHF band.

2.1. Measurement Campaign Procedure

Measurements were carried out in Ilorin, Kwara State, Nigeria (Long 4°36'25"E, Lat 8°25'55"N). The received signal power was measured from the NTA Ilorin transmitter which operates on VHF band at a frequency of 203.25 MHz. For the receiver, a dedicated Agilent spectrum analyzer, N9342C, was used and this was properly positioned in a vehicle and driven at an average speed of 40 km/hr to minimize Doppler effects. The analyzer has a displayed average noise level (DANL) of −164 dBm/Hz and can detect even very weak signals. A whip retractable antenna (70 MHz – 1 GHz), a global positioning system (GPS) receiver and a dedicated memory stick for data storage were coupled to the analyzer. The external GPS receiver was attached to the roof of the vehicle, while the spectrum analyzer was positioned inside the vehicle. The four measurement routes visited are: Murtala Mohammed way, Old Jebba road, Pipeline road, and Ogbomoso road. These routes are characterized with complex terrain propagation features with the presence of hills, valleys and urban clutters. The terrain elevation varies between 350 m to 403.7 m. The routes i.e. MURTALA, PIPELINE, OLD JEBBA AND OGBOMOSO are dual carriage, single lane road, two-lane road and two-lane respectively. The MURALA route is considered the busiest among the routes. The average buildings along these routes are two storeys. The total length of the routes and total number of data points were 75.5 km and 92,280 respectively. The received signal strength data were filtered to minimize noise and preserve the shadowing effects. This reduced the number of data sets per route to 500.

2.2. Prediction Model

NF modelling is a kind of fuzzy inference system which prepares the mapping of inputs to outputs. It consists of both FL and ANN in the development of mapping the inputs to the output [15]. It consists of five layers as shown in Fig. 1. The nodes in these layers are either fixed or adaptive. The adaptive nodes are symbolized by the square shapes, while the fixed nodes are represented by the circular shapes. To describe the structure, a first order Sugeno model has been used because the output is a crisp value that does not require defuzzification. A Sugeno based NF has a rule of the form [15]:

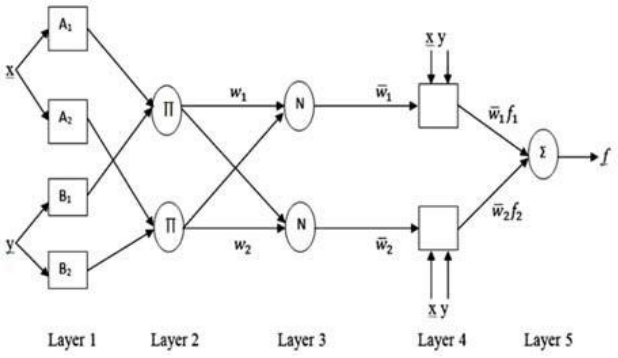


Fig. 1. Neuro-fuzzy structure

Rule 1:

If x is A_1 and y is B_1 then $f_1 = p_1x + q_1y + r_1$ (1)

Rule 2:

If x is A_2 and y is B_2 then $f_2 = p_2 x + q_2y + r_2$ (2)

Layer 1: A node in this layer is adaptable and is given as:

$$L_j^1 = \mu A_i(x) \quad i = 1,2 \tag{3}$$

x is the input to *ith* node, A_i is the alterable language related to this node and the membership function of A_i is $\mu A_i(x)$ and normally taken as:

$$\mu A_i(x) = \frac{1}{1+[(\frac{x-f_i}{d_i})^2]^{e_i}} \tag{4}$$

$\{d_i, e_i, f_i\}$ is the antecedent parameters set. Eqn. (4) represents the generalized bell membership function (MF) which was chosen for this work because it produced the best accuracy when compared to the other membership functions.

Layer 2: This layer comprises of fixed nodes and it solves the firing power w_i of a rule. The multiplication of the incoming signals is the output of each node and is given by:

$$L_i^2 = w_i = \mu A_i(x) \times \mu B_i(y), i = 1,2 \tag{5}$$

$\{p, q \text{ and } r\}$ is the consequent parameters set which are established by the least squares method.

Layer 3: Each node is constant in this layer with the output given by

$$L_i^3 = w_l = \frac{w_i}{\sum w_i}, i = 1,2 \tag{6}$$

Layer 4: The adaptable output of this layer is given by

$$L_i^4 = w_i f_i = w_i(p_i x + q_i y + r_i, i = 1,2 \tag{7}$$

Layer 5: The output of this layer is the summation of all incoming signals and is given by

$$L_i^s = \sum_{i=1}^{2} w_l f_i = \frac{\sum w_i f_i}{\sum w_i} \tag{8}$$

The optimization method used for training the network in this work is the hybrid method which combines both the back propagation gradient descent algorithm and the least square errors estimate used to establish the input and output parameters respectively. The output parameters are adjusted first using the least squares algorithm and those of input parameters by back propagating the faults from the output using the gradient descent method until the training is completed.

3. RESULTS AND DISCUSSION

Figs. 2 to 5 show the graphical depiction of the measured and predicted path losses as a function of distance for each of the four routes considered. It is worth pointing out that from the figures, the prediction by the NF model in all the four routes performed the best among all the considered models as it mimicked the measured data. In Fig 2, the variation of path loss with distance for MURTALA along the route and the predictions of the four empirical models were superimposed on the measured loss. The Egli, COST 231 and Hata models under-predicted the path loss throughout the measurement route, except within 3-4.5 km where the Hata model provided good fitness. The ECC-33 model generally fluctuated between over and under-prediction of the path losses with respect to the measured path losses. Along this route, it can be concluded that the Egli model has the worst prediction with a mean error of −34.5 dB when compared to −1.27 dB for the ECC-33 model and −6.52E-07 dB for the neuro-fuzzy model. The mean error for the NF model is insignificant and this result indicated the good fitness of the model along this route.

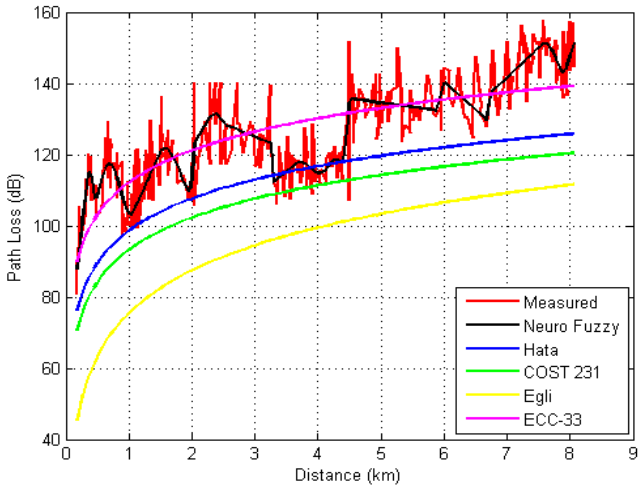


Fig. 2. Comparison of NF model path loss with measured path loss and other empirical models path loss for the MURTALA route

Fig. 3 depicts the result for the PIPELINE route. Again, the ECC-33 model gave optimum prediction up to 3 km; thereafter, it overestimated the path losses while the Hata model provided good fitness at distances above 3 km. Other models underestimated the path losses throughout the measurement distances. In Figures 4 and 5, the ECC-33 model provided good fitness, while all other empirical models underestimated the losses with various offset values.

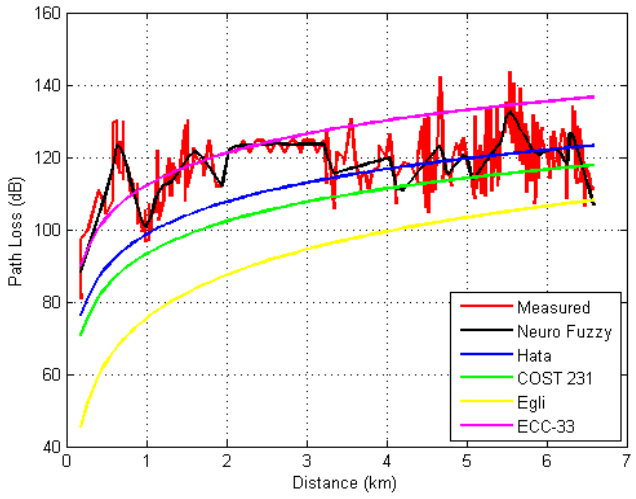


Fig.3. Comparison of NF model path loss with measured path loss and other empirical models path loss for the PIPELINE route

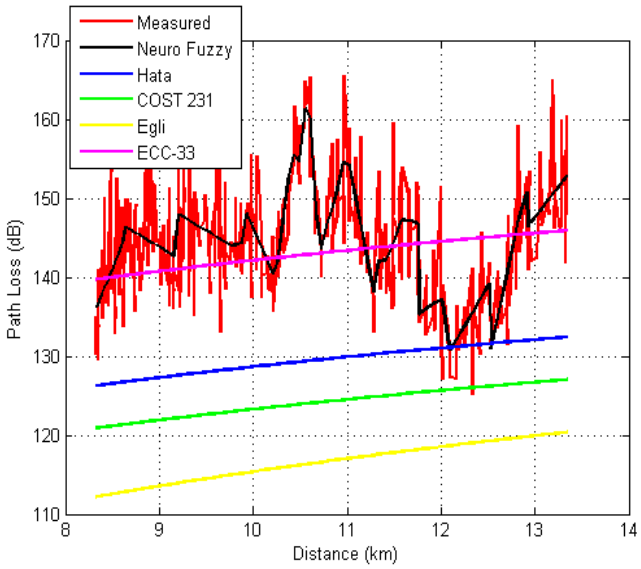


Fig. 4. Comparison of NF model path loss with measured path loss and other empirical models path loss for the OLD JEBBA route

From the figures, it was observed that the NF model provided better prediction as it followed the measured losses. In Table 1, the statistical analysis of the errors for each model across all the routes is provided. Table 1 shows how each of the models performed in terms of their RMSE, SC-RMSE, ME and SDE. RMSE between 0-7 dB is considered acceptable for urban areas [21], although for typical suburban and rural areas up to 10-15 dB [22] can still be acceptable.

For the RMSE and ME, the NF model is the lowest in average across all the routes with 5.2 dB, and −0.00000388 dB respectively, which proved to be the fittest among all the models. These values fell within the acceptable range for urban environments and as such the model did not either underestimate or overestimate the losses as the ME was found to be insignificant. The ECC-33 model gave the least values when compared to other empirical models. The average RMSE and ME were 9.48 dB and 2.27 dB respectively, although the RMSE was found to be a bit higher than the threshold limit for urban environments and the ME clearly indicated that the model underestimated the losses. Surprisingly, this model was developed to suit fixed wireless systems and recommended for European cities but is found to provide optimum predictions when compared to other contending empirical path loss propagation models.

Table 1. Performance metrics for the measurement routes

MODEL		ROUTES				
		OGBOMOSO	MURTALA	PIPELINE	OLD JEBBA	AVERAGE
NF	RMSE (dB)	5.0377	5.3232	5.5736	4.6727	5.1518
	SC-RMSE (dB)	5.8071	11.3713	4.5005	3.6978	6.3442
	ME (dB)	−4.88E-06	−6.52E-07	−3.33E-06	−6.65E-06	−3.88E-06
	SDE (dB)	8.6174	14.7883	7.3182	5.7409	9.1162
COST 231	RMSE (dB)	17.6246	22.0911	13.4102	21.6572	18.696
	SC-RMSE (dB)	11.8382	11.2586	8.0229	19.7363	12.7140
	ME (dB)	−15.8773	−20.1378	−9.9811	−20.4191	−16.6038
	SDE (dB)	6.7797	13.3675	9.2027	2.0496	7.8499
HATA	RMSE (dB)	12.985	17.3235	10.0662	16.6762	14.264
	SC-RMSE (dB)	8.138	8.592	6.2476	14.8439	9.4554
	ME (dB)	−10.4914	−14.752	−4.5952	−15.0332	−11.2180
	SDE (dB)	6.7797	13.3675	9.2027	2.0496	7.8499
EGLI	RMSE (dB)	21.8275	36.0905	24.4704	29.1706	27.890
	SC-RMSE (dB)	13.8329	19.7112	14.5071	26.5375	18.6472
	ME (dB)	−20.2792	−34.5799	−21.9709	−28.2467	−26.2692
	SDE (dB)	9.0246	17.7937	12.2498	2.7283	10.4491
ECC-33	RMSE (dB)	8.2447	9.1536	12.5833	7.3857	9.487
	SC-RMSE (dB)	5.0961	7.8597	7.0676	5.9038	6.4818
	ME (dB)	3.0725	−1.273	8.8542	−1.5622	2.2729
	SDE (dB)	6.8434	13.3076	9.1691	2.0619	7.8455

Despite the fact that the system parameters, such as the operating frequency, height of the transmitter and distance of the measurement routes fell within the validity of the Hata, COST 231 and Egli models, the models performed woefully, with an RMSE and ME of 14.26 dB, and −11.21 dB, 18.69 dB and −16.60 dB, and 27.89 dB and −26.26 dB respectively. The average standard deviation error (SDE) for the NF model is 9.11 dB, while, 7.84 dB, 7.84 dB, 10.44 dB and 7.84 dB for COST 231, Hata, Egli and ECC-33 models respectively. However, the excellent performance of the NF in terms of the mean prediction error and RMSE may not be over-emphasized as the route specific SDEs were 8.61 dB, 14.78 dB, 7.31 dB and 5.74 dB for OGBOMOSO, MURTALA, PIPELINE and OLD JEBBA routes. These are quite high and this is because the model mimicked the measured data and the deviations are noticeable along each route with varying degree of clutter effects. Furthermore, in terms of SC-RMSE, no significant impacts were observed for the NF model when compared to RMSE, apart from the MURTALA route. Ordinarily, SC-RMSE negates the impact of dispersion from the overall errors attributed to a noisy link. For other models, the SC-RMSE was very significant as their respective mean prediction errors were quite high. The prediction errors as a function of radial distance from the transmitter for each route were equally investigated and the results are presented in Figures 6 to 9.

In Fig. 6, it was observed that the shadowing effects on the PE for the NF model along the MURTALA route, as the PE undulated along the 0 dB baseline with a varying degree of impact due to different clutter types along the route. Other models also undulated but with high amplitudes of PE with varying offsets. Interestingly, all the empirical models tend to have high prediction errors between the 0-1 km distance and which is due to their initial offset values for the models. Between 3-4.5 km, the PEs of COST 231 and Hata models were close to the 0 dB baseline, this conformed to the earlier findings reported in Fig 2. The ECC-33 model provided the least errors when compared to other empirical models but with a high prediction error within 3 and 4.5 km.

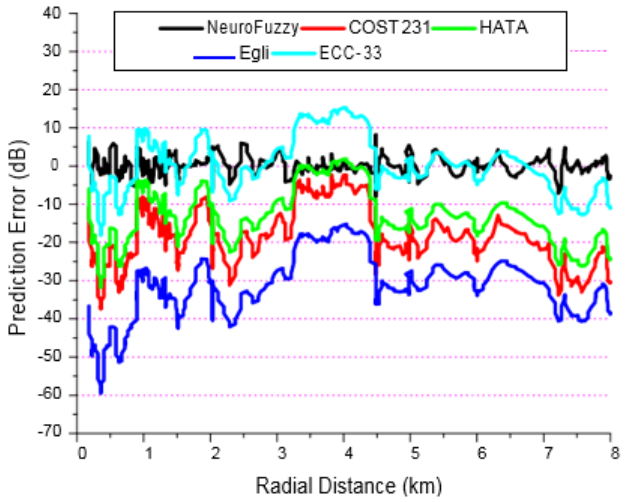


Fig. 6. Comparison of NF model prediction errors with other empirical models prediction errors for the MURTALA route

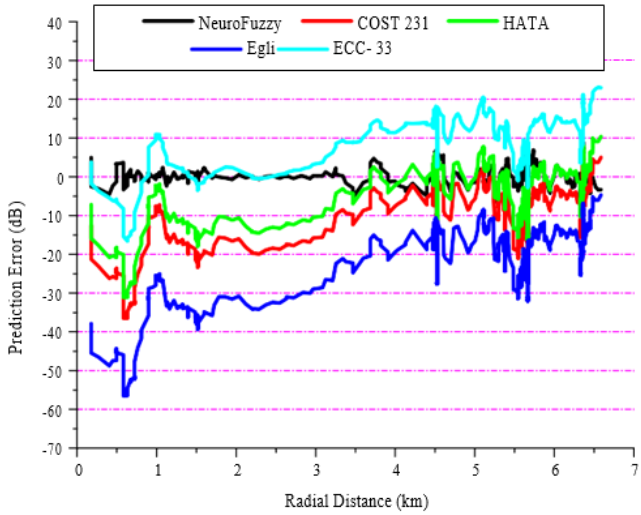


Fig. 7. Comparison of NF model prediction errors with other empirical models prediction errors for the PIPELINE route

In Fig. 7 it was shown that the NF model provided the least PE along the route. However, the ECC-33 model tried to emulate the NF model aside the initial spike between 0-1 km and $d > 3$ km, while all other models converged towards the 0 dB baseline afterwards. The situation in Figs 8 and 9 were quite different, as severe clutter effects on the PE were noticeable with several spikes along the routes.

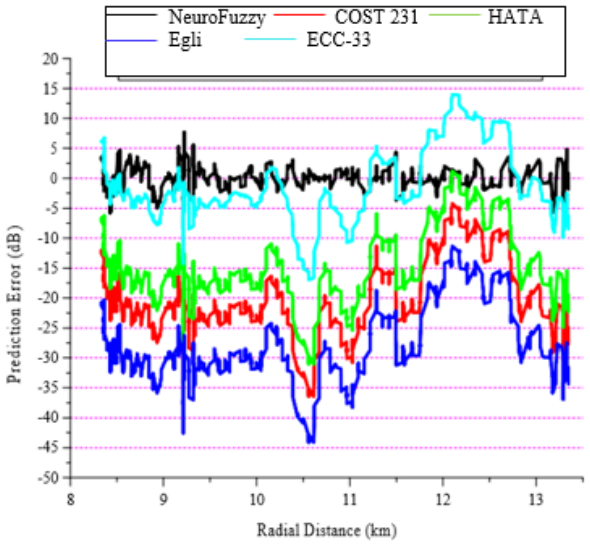


Fig. 8. Comparison of NF model prediction errors with other empirical models prediction errors for the OLD JEBBA route

Generally, the NF model can be said to have the best PE in all the routes among the considered models as it wavered close to the 0 dB line between over and under-prediction in a uniform manner throughout the distances. The result obtained in Fig. 9, for the OGBOMOSO route showed that the PEs for the Egli and COST 231 are almost similar. They largely under-predicted and showed slight over-predictions between 20-25 km and 30-35 km. The NF and the ECC-33 models' predictions are similar as well.

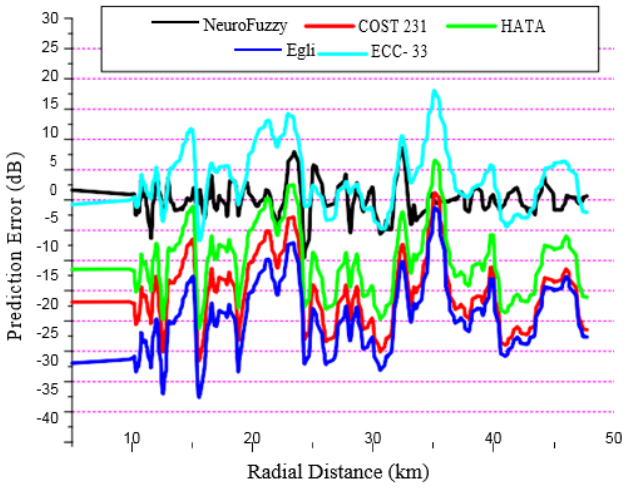


Fig. 9. Comparison of NF model prediction errors with other empirical models prediction errors for the OGBOMOSO route

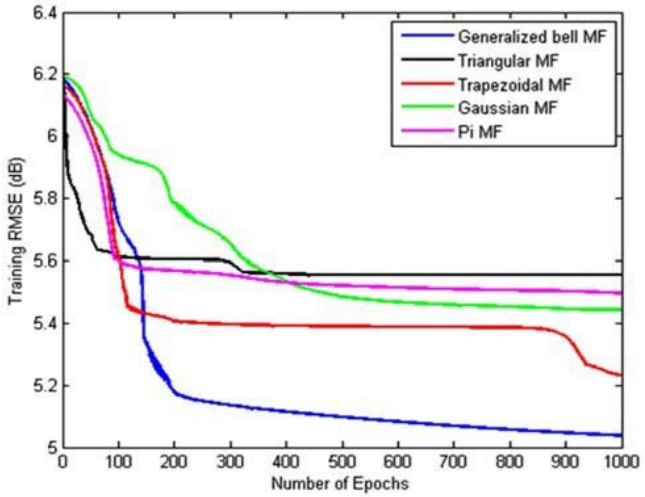


Fig. 10. Effects of membership functions types and the number of epochs on the training RMSE for the OGBOMOSO route

In Fig. 10, the impact of membership function types and epochs size on the RMSE were investigated, and these were done to establish which of the membership function types provided stability and attained fast convergence at the minimum number of epochs (number of iterations). For this exercise the OGBOMOSO route was used. Generally, the figure showed that an increase in the number of epochs for the different types of membership functions translated to a decrease in the RMSE. However, the increment in the number of epochs got to a point where the antecedent and consequent parameter sets could no longer be updated or the updates were infinitesimal and negligible, and therefore attained a steady state output for the RMSE. Also, it could be seen that the generalized bell membership function produced the lowest RMSE among the membership functions considered for this route.

4. CONCLUSION

This work is centered on the incorporation of artificial intelligence in path loss prediction. A neuro-fuzzy model was developed and used to predict path losses in the VHF band. The path loss predictions as well as the prediction errors of the proposed model were compared to that of four widely used empirical models. The proposed NF model provides the lowest errors with an average RMSE and ME of 5.2 dB and −0.00000388 dB respectively, across all the routes. The ECC-33 model gave the least values when compared to other empirical models. The average RMSE and ME were 9.48 dB and 2.27 dB respectively, although the RMSE was found to be a bit higher than the threshold limit for urban areas and the ME clearly indicated the model overestimated the losses. The work showed that the Hata, COST 231 and Egli models performed woefully, with higher errors despite the fact that the system parameters such as the operating frequency, height of the transmitter and distance of the measurement routes fell within the validity of the models. Furthermore, the paper showed that route-specific SDEs of the proposed model are quite high, as the model mimicked the measured data and the clutter effects were noticeable along each route with varying degrees. It was also discovered that the SC-RMSE had no significant impact on the NF model when compared to the RMSE. For other models, the SC-RMSE was very significant as their respective mean prediction errors were quite high. The paper also showed that an increase in the number of epochs for the different types of membership functions translated to a decrease in the RMSE and the generalized bell membership function produced the lowest RMSE among the membership functions considered for this route. In conclusion, the NF model proved to be the fittest for path loss prediction among the other models for this work. However, future work can be extended to other frequency bands, more routes with respect to the same transmitter and other transmitters, consideration of more path loss models and extension to other geographical areas.

ACKNOWLEDGMENT

The authors would like to express their sincere appreciation to the University of Ilorin, Nigeria, for the purchase of the dedicated Agilent spectrum analyzer and for the funds provided to facilitate this study.

REFERENCES

[1] J.D. Parsons (2000), “The mobile radio propagation channel”, Second Edition, John Wiley & Sons Ltd., pp. 15-127.

[2] O.S. Onidare, N. Faruk, W.B. Olayiwola, M.Y. Muhammad, O. Sowande, A.A. Adeseko, “Practical Error Bounds for Empirical Models at VHF/UHF bands” Bayero Journal of Engineering and Technology, vol. 11(1), pp. 28-39, 2016.

[3] N. Faruk, Y.A. Adediran, A.A. Adeseko, “Error Bounds of Empirical Path Loss Models at VHF/UHF Bands in Kwara State, Nigeria” Eurocon, Croatia, July 2013.

[4] A.A. Jimoh, Surajudeen-Bakinde, N.T., Faruk, N, Adeseko, A.A., Obiyemi, O.O. and Bello, O.W. (2015), Performance Analysis of

Empirical Path Loss Models in VHF and UHF Bands. In proc IEEE International Conference on Information and Communication Systems, Jordan, April, 7 – 9, 2015. pp. 194-199.

[5] N. Faruk, A.A. Adeseko, Y.A. Adediran, "On the Study of Empirical Path Loss Models for Accurate Prediction of TV Signal for Secondary Users" Progress in Electromagnetic Research (PIER), B, vol. 49, pp. 155-176, 2013.

[6] S.I. Popoola, A.A. Atayero, N. Faruk, C.T. Calafate, L.A. Olawoyin and V.O. Matthews (2017), "Standard Propagation Model Tuning for Path Loss Predictions in Built-Up Environments", in *International Conference on Computational Science and Its Applications*, 2017, pp. 363-375.

[7] S.I. Popoola, A.A. Atayero, N. Faruk, C.T. Calafate, E. Adetiba, and V.O. Matthews (2017), "Calibrating the Standard Path Loss Model for Urban Environments using Field Measurements and Geospatial Data," in Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2017, 5-7 July, 2017, London, U.K., pp. 513-518, 2017.

[8] N. Faruk, A.A.Ayeni, Y.A. Adediran and N.T Surajudeen-Bakinde (2014), "Improved Path Loss Model for Predicting DTV Coverage for Secondary Access" Int. J. Wireless and Mobile Computing, Vol. 7, No. 6, pp. 565-576, 2014.

[9] D.W. Kifle, L.C. Gimenez, B. Wegmann, I. Viering, and A. Klein (2014), Comparison and Extension of Existing 3D Propagation Models with Real-World Effects Based on Ray-Tracing A Basis for Network Planning and Optimization, Wireless Personal Communication, 78:1719-1738.

[10] E. Greenberg, E. Klodzh (2005), "Comparison of deterministic, empirical and physical propagation models in urban environments", Microwaves, Communications, Antennas and Electronic Systems (COMCAS), 2015 IEEE 2005. PIMRC 2005. IEEE 16th Int. Symp. on, vol. 1, no., pp. 659-663, 11-14 Sept. 2005.

[11] A.B. Tamma, A. Rabbie, K.S. Mustafa, "Neural Network Approach to Model the Propagation Path Loss for Great Tripoli Area at 900, 1800, and 2100 MHz frequency bands" 16th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering-STA, Tunisia, pp. 793-798, December, 2015.

[12] E.D. Turkan, Y.H. Berna, A. Aysen, "Fuzzy Adaptive Neural Network Approach to Path Loss Prediction in Urban Areas at GSM-900 band", vol. 18, No. 6, pp. 1077-1094, 2010.

[13] C.D. Joel, P.D. Elmer, "Neural Network Based Path Loss Prediction for Digital TV Macrocells" 8th IEEE International Conference Humanoid, Nanotechnology, Information Technology Communication and Control, Environment and Management (HNICEM), Phillipines, December, 2015.

[14] A.R. Ozdemir, Mustapha Alkan, Mehmet Kabak, M. Gulsen, M.H. Sazli, "The Prediction of Propagation Loss of FM Radio Station Using Artificial Nueral Network" Journal of Electromagnetic Analysis and Applications vol. 6, pp. 358-365, 2014.

[15] Rupanwita DasMahapatra, "Optimal Power Control for Cognitive Radio in Spectrum Distribution Using ANFIS" IEEE, 2015.

[16] G. Vishal, S.C. Sharma, "Secure Path Loss Prediction in Fringe Areas Using Fuzzy Logic Approach" IEEE International Conference of Advances in Computing, Control and Telecommunication Technologies, pp. 372-375, 2009.

[17] M.J. Jafri, J.J. Ely and L.L. Vahala, "Interference Path Loss Prediction in A319/320 Airplanes Using Modulated Fuzzy Logic and Neural Networks," NAFIPS 2007 - 2007 Annual Meeting of the North American Fuzzy Information Processing Society, San Diego, CA, 2007, pp. 337-342.

[18] S.P. Sotiroudis, S.K. Goudos, K.A. Gotsis, K. Siakavara and J.N. Sahalos, "Modeling by optimal Artificial Neural Networks the prediction of propagation path loss in urban environments," 2013 IEEE-APS Topical Conference on Antennas and Propagation in Wireless Communications (APWC), Torino, 2013, pp. 599-602.

[19] J.C. Dela Cruz and F.S. Caluyo, "Heuristic modelling of outdoor path loss for 9m, 3m and 1.5m antenna at 677 MHz," 2013 IEEE

Conference on Cybernetics and Intelligent Systems (CIS), Manila, 2013, pp. 117-122.

[20] S.P. Sotiroudis, S.K. Goudos, K.A. Gotsis, K. Siakavara and J.N. Sahalos, "Optimal Artificial Neural Network design for propagation path-loss prediction using adaptive evolutionary algorithms," 2013 7th European Conference on Antennas and Propagation (EuCAP), Gothenburg, 2013, pp. 3795-3799.

[21] S.P. Sotiroudis, S.K. Goudos, K.A. Gotsis, K. Siakavara and J.N. Sahalos, "Application of a Composite Differential Evolution Algorithm in Optimal Neural Network Design for Propagation Path - Loss Prediction in Mobile Communication Systems," in IEEE Antennas and Wireless Propagation Letters, vol. 12, pp. 364-367, 2013.

[22] V.S. Abhayawardhana, I.J.Wassell, D. Crosbsy, M.P. Sellars., and M.G. Brown, (2005) 'comparison of empirical propagation path loss models for fixed wireless access systems' IEEE Vehicular Technology Conference, Vol 1, pp. 73-77, 2005.VTC 2005-Spring.

[23] N. Blaunstein, D. Censor, and D Katz, (2003) 'Radio propagation in rural residential areas with vegetation', Progress in Electromagnetics Research, PIER 40, pp. 131-153.

BEYOND MAD?: THE RACE FOR ARTIFICIAL GENERAL INTELLIGENCE

Anand Ramamoorthy[1], Roman Yampolskiy[2]
[1]Dept. of Anaesthesiology, UW-Madison, WI, USA
[2]Computer Engineering and Computer Science, University of Louisville, KY, USA

Abstract – Artificial intelligence research is a source of great technological advancement as well as ethical concern, as applied AI invades diverse aspects of human life. Yet true artificial general intelligence remains out of reach. Based on the history of deeply transformative technologies developed by multiple actors on the global stage and their consequences for global stability, we consider the possibility of artificial general intelligence arms races and propose solutions aimed at managing the development of such an intelligence without increasing the risks to global stability and humanity.

Keywords – AGI, AI safety, AI Treaty, arms races, global coordination, MAD

1. INTRODUCTION

Artificial intelligence (AI) research has a decades-long history. However, AI systems have come to occupy the public imagination now more than ever, thanks to some remarkable recent technical developments with compelling public demonstrations, as well as an unprecedented level of hype and concerns of existential risk.

Applied AI stands poised to revolutionize multiple industries and avenues of human life. AI systems are expected to replace human drivers in autonomous vehicles, human combatants in warfare, and the relevance of the human operator to many forms of professional activity will be called into question as these systems become more proficient and truly autonomous.

A WIRED magazine article in 2014 designated Go as the "ancient game that computers still cannot win", with some experts predicting that computers would only be able to win in a decade or more [1]. Well before the predicted time, Deep Mind's AlphaGo [2], attracted significant attention owing to its success against Lee Sedol, a 9th Dan in the game of Go [3], and the world's leading player Ke Jie [4].

A Genetic Fuzzy Tree method based AI, ALPHA, which is currently a simulation tool, has emerged as a likely candidate for future military applications involving unmanned aerial vehicles as well as mixed squadrons of manned and unmanned vehicles in aerial combat [5].

While such specialist AI systems exist, and are evidently becoming more and more capable, the most exciting (and according to AI safety researchers, the most risk-laden) goal of AI research has always been what can be termed Strong AI, that is, an artificial intelligence that is truly general and not constrained to a task-domain. Such an artificial general intelligence (AGI) does not exist in our world yet, to the best of our knowledge.

The possibility of such an AGI leading to so-called artificial superintelligences (ASI) via intelligence explosion (a scenario where the AGI recursively improves itself to the point of exceeding human-level intelligence to an unpredictable extent), has led scientists, philosophers and technologists to consider the existential risks (to humanity) posed by the development of AGI. Unlike technologists and techno-futurists, AI and machine-learning researchers are more conservative in their estimates as to when such a system (or multiple such systems) can be expected on the scene. It is to be noted that the machine learning/AI research community is mostly unsure as to whether AGI would come to be or result in ASI. There also exist optimistic views on AI development which provide a healthy counterbalance to the apocalyptic visions commonly encountered in the media [6].

That said, 48% of the expert respondents in a recent survey [7] did agree that "society should prioritize work on minimizing the potential risks of AI". A recent report from the U.S Government's Office of Science and Technology Policy examines the state of the art in AI technology and provides an overview of the benefits and risks of AI, concluding with policy recommendations for the administration [8]. AI and AI safety are predicted to be vital to both economic development and national security [9]. It stands to

reason that such considerations cut across borders, and are shared by nation-state stakeholders in the AI sector. AI arms races can happen in the context of "narrow" AI systems as well as AGI development. In the former instance, the most salient manifestation of such a race would be adversarial programs, administered by militaries around the world, to develop intelligent controllers and autonomous weapons systems. This possibility has attracted tremendous attention from AI safety experts as well as ethicists who rightly fear the ramifications of such systems being introduced into human armed conflict. The second type of AI arms race concerns multiparty, adversarial programs aimed at developing the world's first AGI.

Here we examine the prospect of AGI development in terms of escalation and arms races between diverse actors and stakeholders: state actors, non-state (corporate) actors, and rogue actors. We conclude by providing policy recommendations aimed at mitigating the risks identified.

2. ARMS RACES AND AGI: BEYOND MAD?

The success of The Manhattan Project and the deployment of nuclear weapons by the United States military in the Second World War led to a new kind of international conflict, a nuclear arms race, where powerful nations sought to acquire the same destructive capabilities as the U.S.A. This resulted in a world where an unstable peace is kept alive, informed to a significant extent, by the doctrine of mutually assured destruction (MAD) in addition to global non-proliferation efforts. A detailed discussion of the current understanding of MAD and the status of nuclear non-proliferation is beyond the scope of this paper. It suffices to note that examining the case of MAD in its original context provides insights that can be used to understand the role of disruptive technologies in international conflict (for case studies of such technologies see [9]). AGI, if and when it happens, may well be the final disruptive technological development engineered by humans. AGI represents a level of power that remains firmly in the realm of speculative fiction as on date. It stands to reason that if true AI were achievable, state actors would be invested in achieving this and with priority if possible. Such a quest for priority might have disastrous consequences due to corner-cutting when it comes to safety, and has been described as "racing to the precipice" [10]. An AI arms race is often spoken of in the context of the development of autonomous weapons systems which become increasingly sophisticated, changing the face of warfare. Were we to adopt Clausewitz's observation that "war is the continuation of politics and with other means" [11] and examine international conflict, it becomes obvious that the role of AI would extend well beyond, and emerge well before, armed conflict. A nation equipped with a fully general AI, would stand to benefit in the negotiation of conflict and agendas, regardless of means. If said AI were both general AND endowed with the ability to act in the world (i.e., not merely an Oracle-in-a-box as some have proposed, see [12] for an analysis of AI confinement), then, all arguments pertaining to the existential risk posed by AI would apply. Having AI systems autonomously determine the deployment of weapons in armed conflict is one major route to potential catastrophe, but we would like to emphasize that matters are likely to become fraught even before this development.

AGI development would push the global security strategy beyond what is currently in place. In the event of human control over the AGIs (which is a problem worth examining in its own right), MAD would not be sufficient to avert catastrophe. This would be because of the greater complexity associated with AGI and the capabilities such a system would present to human controllers, for instance, the AGI of a particularly belligerent state could calculate the optimal means to mortally destabilize the economy of a rival state (however powerful), or develop weaponized code disseminated globally to control, if not disrupt, vital systems such as power grids and communication networks. In other words, the cyber-warfare capabilities of an AGI-assisted nation-state would pose a serious threat to global stability and humanity. The current forms of narrow AI are capable of interfering with communication services. AI-enabled surveillance across communication networks is likely to become the norm. AI tools with the potential to perturb or alter the content of communications are already in development (see https://lyrebird.ai/ for an interesting example in the context of mimicking human speech; see also: https://lyrebird.ai/ethics/). An AGI with access to the Internet and communication networks in general would be able to, depending on its objectives (or of those who deploy it), selectively impede communication across a certain network/region, or fabricate misinformation to probe human responses if it develops the objective to understand social impact of communication networks. Much as these scenarios remind us of science fiction, it is worth noting that we encounter reports of computational propaganda or technology-assisted disinformation with increasing regularity. On a more optimistic note, an AGI that is constrained to cooperate with humans could help

envision more efficient use of resources for optimizing the communication networks we have available, or design altogether novel and better architectures.

As we discuss below, the potential development of AGI is unlikely to be an exclusively state-funded affair. Current breakthrough AI systems all appear to be products designed by technology companies.
Given the fact that technology giants such as Google, Facebook and others are beginning to open-source their machine-learning tools (e.g. TensorFlow), it is well within the realm of possibility that non-commercial, non-state actors including individuals would be able to develop applied artificial intelligence.

2.1. Actors in the AGI race

The development of AGI (or AGIs) could be due to the actions of many actors, and each type of genesis would likely be linked to the emergence of risks shaped by the intentions of the actors, the corresponding core objectives of the AGI and the context of AGI deployment. Here we examine possible scenarios as a function of actors and likelihood.

2.1.1. State actors

States function as agents/actors in the international environment, and pursue policies that benefit them as well as their allies whilst potentially mitigating the risks posed by states that are not allies or neutral, if not diminishing their influence regionally/ internationally. Global coordination remains a non-trivial challenge and one that is not easily resolved.
Imagine throwing in one state which suddenly had an AGI Oracle that could examine scenarios exhaustively for any given situation, access all relevant information from past situations of the sort being considered, and glean insights from actionable intel available to the highest executive in the state to guide decision making. If the Oracle has no influence on the world physically, and the state concerned is one with more to gain from a better solution to the global coordination problem, their use of the Oracle's counsel would likely benefit humanity.

It is conceivable that this may not be the most likely scenario. Powerful technology has often been a driving force in the pursuit of a world order where the nation-state with the technology in question is in a privileged position to pursue its objectives. If the state in question is less interested in achieving better global coordination or more interested in exercising disproportionate influence globally, the AGI would provide such a state actor with a potentially incalculable advantage.

Contemporary trends in various nations across the globe evidence a resurgence of nationalistic themes in politics, with elections delivering governance to those who promise a stronger nation "above all else". The world is not presently geared to support a single global community, and in a setup where nationalistic impulses influence both intranational politics and geopolitics, it is neither realistic nor prudent to assume that the achievement of AGI by any single state actor would be beneficial to humanity. It stands to reason that multiple state actors would seek unprecedented strategic advantage through AGI. If there is no commensurate development on the global coordination problem which renders the balance of power stable, this scenario would lead to catastrophe.

2.1.2. Corporate actors

Industry remains the face of AI research. With multiple corporate entities vying to develop true artificial intelligence, or artificial general intelligence, there already is a race to harness the power of AI for commercial ends whilst ostensibly impacting the world positively. Recent calls for research on AI safety have been made by researchers as well as leaders in the tech industry. Such a confluence of academic and corporate/industrial camps on existential risk posed by AI bodes well for research and development on this front. That said, the competitive and technological advantage presented by achieving priority in the development of AI is likely to incentivize some corporate actors to compromise on safety, resulting in unregulated, unsafe development of AI/AGI. One is reminded of the reported discrepancy between management and engineering divisions at NASA on the risk associated with continued operation of the space shuttle, as discussed in Appendix F of the report on the Challenger explosion [13]. Also, scientific and ethical consensus may not be sufficient to motivate technologically capable enterprises to focus on safety prior to developing AI across the globe. The problem of global coordination would remain a factor that would increase degrees of freedom in any given scenario where actors interact internationally.

2.1.3. Rogue actors

Unlike other dangerous technological developments in history, AI breakthroughs may not occur exclusively

in academic, governmental or industrial centers of research. In principle, powerful AI systems could be developed by individuals or groups with no national or corporate agendas. Such *homebrewed* AI would be hard to deal with precisely due to a lack of oversight, monitoring or consensus on architectures and objectives. More worrisome is the prospect of such rogue actors developing AI without safety considerations or with malicious intent (for example, see the case of commercially available unmanned aerial vehicles being repurposed for guerrilla warfare by terror groups [14]). It could be argued that the resources required to develop a powerful and truly general artificial intelligence may not be available to rogue actors, but it is far from clear that this would be the case, and it may be unwise to presume that any such obstacles would be insurmountable.

Cyberattacks originating from individuals/small groups are commonplace in our increasingly interconnected world, and it is conceivable that the development of an artificial general intelligence by rogue actors would be similar in terms of execution, but more harmful in terms of impact on human society and life. As a case in point, consider the recent large-scale spread of the WannaCry ransomware, exploiting a vulnerability in the Windows operating system (particularly versions past). Investigations seem to suggest that the architects of the attack were not well organized and the attack not as nightmarish as it could have been. Yet, it precipitated a significant amount of chaos and affected networked computers worldwide. Now replace the ransomware with an AGI that is released into the wild by hackers motivated by political ideology, notoriety or curiosity [15]. Even if the AGI is not inherently dangerous, the consequences in such a scenario would be hard to predict or plan for and could be catastrophic.

3. AGI AND VALUE ALIGNMENT

Researchers working on making any potential AGI "friendly" or compatible with human existence, if not values, speak of the AI value alignment problem. There are technical as well as pragmatic considerations attendant upon AGI research (see [16]), which increase the complexity of any proposed attempt to align the values (or objectives) of a putative AGI with human ones. The most obvious, and most non-trivial consideration (if not constraint), is the fact that humans across the world are not capable of value alignment to the extent that actions which increase existential risk for all humanity would be rendered extremely unlikely. Climate change policy is one example of a failure of global coordination on a matter of utmost importance. The current near-impossibility of universal nuclear disarmament is another. International conflict is never devoid of human cost. It appears to be the case that several developments in the AI sector are likely to exacerbate inequities in the bargaining power of nations on the international scene, as well as their ability to administer coercive force via an AI-augmented military. Given the lack of value alignment within human groups, it would be highly prudent to seek a solution to this problem in parallel with, if not prior to, working on AGI value alignment. It is to be noted that discussions on AGI emergence and global coordination posit the scenario of a singleton (one global governing entity), which appears markedly less likely to happen relative to the birthing of an AGI. In all likelihood, even the most benign AGI would be developed within a world where human groups (nation states) do not see eye-to-eye on several crucial issues. Given this, it would be wise to temper any optimism on the AGI front with a healthy appreciation for risks, safety concerns and the need to respect reality, particularly with regard to international conflict and human factors.

4. SHAPING AGI RESEARCH

[17] discusses the means and measures to shape AI (and potential AGI) research to promote safe and beneficial AI development, and makes the distinction between extrinsic measures such as constraints on design and intrinsic measures such as inducing a normative shift towards wanting to build beneficial AI and creating a stigma around dangerous AI research. [17] makes the compelling argument that hard, extrinsic measures such as outright bans might have effects counter to what was intended, (as a ban would draw attention to the problem in a manner that evokes curiosity and desire for boundary breaking). The intentional shaping of AGI research by targeting the culture in the research communities, both academic and non-academic, to make the wish to build safe and beneficial AI a social norm, with strong normative factors encouraging the avoidance of unsafe designs etc. appears to be an interesting strategy and one that is likelier to have a stronger impact over time, especially in the context of open AI development.

[18] discusses the implications of open AI development and elaborates on the complexities inherent to the pursuit of openness as a policy across multiple dimensions, such as the political, scientific and technological and it appears that the solution proposed here would make concerns regarding long-term costs of openness irrelevant, given that, assuming this solution is workable, a coalition of

states and researchers would work on AI (AGI) development as a public, open enterprise, inspired by ventures such as OpenAI.

[19] presents a formal account of determining a Pareto optimal policy by a machine built and deployed by two actors (individuals, companies or states), reflecting the beliefs and utility functions associated with each actor, and demonstrates that such a policy would sequentially shift the prioritization of one actor's utility function over the other's as a consequence of the accuracy of the actors' beliefs regarding the input to the machine (i.e. the state of the world in which the machine functions). This analysis raises the question whether the fruits of strategic cooperation in AI development could still be distributed unequally if the actors involved do not have access to either the same information regarding the machine's world (and inputs), or if one or more actors actively shape the beliefs of the other actors through disinformation.

5. PERSPECTIVES AND SOLUTIONS

In examining the potential for an AGI "arms race", and the trends observed in the present-day world, we see that there are multiple paths to averting such conflict in the decades to come. They are not mutually exclusive and can be integrated to form a comprehensive strategy. We recommend developing a solution that is layered, with failures at one level compensated for by policies and mechanisms in place at another, for redundancies and multiple defenses would render the solution robust.

5.1. Solution 1: Global collaboration on AGI development and safety

Postwar advancements in space science led to both the Space Age, characterized by the race for space. The erstwhile Soviet Union and the United States of America vied with each other for priority in space exploration and successful manned missions. While spacefaring nation states pursue space programs of their own with national agendas and strategic goals, several have also come together to create and sustain an international space station (ISS), which has become a new benchmark for international cooperation [20]. Taking this as a paradigmatic case of successful international cooperation, we refer to this as the ISS pathway.

If the ISS pathway is chosen with coordination and foresight, an AGI arms race could be avoided altogether by adopting the safe and beneficial development of AGI as a global, non-strategic humanitarian objective, under the aegis of a special agency within the United Nations, established for this exclusive purpose. In this scenario, countries would supply resources and invest equitably in the creation of the world's *first and only* AGI with safety considerations imposed at every level. As a first step towards securing such global cooperation, a comprehensive Benevolent AGI Treaty must be developed and ratified by all member nations of the UN. Recent calls for a ban on autonomous weapons are a step in this direction (see [21]). The 23 principles, enunciated in conjunction with the 2017 Asilomar Conference, provide a foundation for such a treaty (see: [22]). If such a treaty succeeds, any potential beneficial AGI would be treated as a global, public good with equal distribution of (carefully minimized) risks and (carefully maximized) benefits and no room for monopolies, adversarial co-opting of the system's potential, etc. The latter constraints are vital to the success of the treaty as an instrument and even more important to its efficacious implementation, as an AGI would represent a level of power hitherto unprecedented. Additional safeguards need to be researched on account of the fact that an AGI would be much easier to reproduce compared to other disruptive technologies humans have developed thus far. Further research needs to be done on how to convolve AI safety approaches such as "boxing in the AI", with collaborative AGI development, to prevent possible unauthorized, undesirable reproduction of the AGI by actors not sanctioned by the treaty.

The game theoretic analysis presented by [10] appears to support a cooperative approach, as teams working together had positive implications for safety, as opposed to teams racing to succeed. However, they also present an intriguing result that increased knowledge of the work being done by other teams increased risk. It is to be noted that this informational hazard would apply only if nations seek to gain strategic dominance or advantage, and a concerted, transparent effort to build an AGI for global welfare, is expected to improve safety and risk-mitigation efforts.

5.2. Solution 2: Global Task Force on AGI to monitor, delay and enforce safety guidelines

[23] proposes general policy desiderata for a world about to countenance AGI/ASI and recommend with specific reference to the problem of potential global coordination failure, that control of such technology be centralized, or a monitoring regime devoted to identifying harmful applications of AI and

intercepting them pre-deployment. As seen above, creating a global AGI project would address the global coordination issue head-on. Achieving a level of coordination superior to that which currently sustains the International Space Station would be a precondition for an international AGI development program.

What about a world which is not yet ready for such coordination or the advent of AGI; namely, our world? We propose that a global watchdog agency, be created for the express purpose of tracking progress of AGI programs, state-funded as well as corporate (the third variety, rogue AGI development, may be harder to monitor, but not impossible). This agency would have as its operating charter, the treaty on safe AGI development for the benefit of all humanity, as proposed above, with jurisdiction across all nations (a singleton, but within the specific context of AGI development, without authority over other aspects of governance or administration) and the lawful authority to both intercept unlawful attempts at AGI development and unilaterally terminate or freeze such programs.

The creation of such a body with such vast powers (albeit within a specific context), would also be constrained by the efficacy of global coordination and is less likely to happen relative to the space-station analogue proposed earlier. It is perhaps easier to bring people together to do something (build AGI as a global effort), rather than to create a group that tells everyone else not to do specific things (watchdog with the power to stop AGI development that runs afoul of the treaty). The history of the IAEA may be of interest if this path is chosen, to learn from its successes and preempt challenges likely to be faced by an international agency tasked with regulating technological development of a specific sort. Any such agency would benefit from drawing upon the cybersecurity infrastructures established by nation states, as well as the intelligence communities of the participating states. The proposed agency would, however, have a focus that is global, and not constrained by the national interests of participating nations, as this could easily lead to conflicts with undesirable ramifications.

Now, there are reasons to believe that AGI development may occur in the corporate sphere, as opposed to within a state-funded program. OpenAI is a non-profit company formed recently by a number of entrepreneurs concerned about safe AGI development. Several companies such as Google DeepMind, Vicarious, etc. are pushing the boundaries of what AI is capable of, with an increasing rate of progress. Many such companies could form a consortium driven by the need for safe AGI development and public good. Indeed, such an entity exists in the form of Partnership on AI [24] a timely non-profit organization bringing together diverse parties and actors with public safety and benefit as foci.

The problem with this scenario is that the raw capability for AGI development is not strictly limited to one corner of the globe, and in the absence of intergovernmental coordination and a global regulatory authority with real legal power to halt unsafe programs, it is entirely conceivable that AGI development could occur in multiple parts of the world and not all players may accept the rules likely to ensure the safest and most beneficial outcome. [25] proposes the creation of an AI Standards Developing Organization, whose role would be to provide strict guidelines for risk management and AI safety in an industrial context. This is indeed a good non-state analogue to the adoption of a treaty, and perhaps closer to the theater of AI development/deployment. Although it is to be expected that complexities associated with developing regulatory standards would necessitate the involvement of states or a coalition of states.

In addition to the solutions proffered above, powerful impetus could be provided to the creation of a Nanny AI [26] which would be tasked with the monitoring of AGI development worldwide with a clear mandate to delay any and all such programs until the coordination issues and safety considerations can be addressed rigorously. This, however, is not without risks of its own.

6. CONCLUSION

An artificial intelligence arms race most likely cannot be stopped, only managed. While there are many possible scenarios and outcomes, it is in the best interest of humanity that the dangerous ones be given due consideration before we develop AGI. We believe a systematic and tempered, public elucidation of the risks would help the cause of safe AGI development more than an approach characterized by hype and apocalyptic messaging. Technological progress tends to have a life of its own, and given the rate at which AI systems are achieving feats of intelligence and expertise, it is merely a matter of time, perhaps a few decades hence, perhaps more, before a truly general AI comes into existence. In this paper we have examined the prospect of such AGI development being prosecuted as an “arms race”, and have offered

a set of solutions, including the development of a comprehensive treaty on AGI development, international collaboration on a singular AGI program, a regulatory global watchdog designed to enforce the aforementioned treaty and potential recruitment of a Nanny-AI system [26], to delay AGI development until pragmatic considerations and risks can all be addressed with sufficient rigor. With this, we seek to add to the emerging discussion on AI safety within the technology and policy communities, and hope that the ideas presented herein are investigated thoroughly with concrete application in mind.

ACKNOWLEDGEMENT

We thank Beth M.Barnes for valuable comments on this paper.

REFERENCES

[1] A. Levinovitz. “The mystery of Go, the ancient game that computers still can’t win.” Wired Magazine. Retrieved August 27, 2017, from https://www.wired.com/2014/05/the-world-of-computer-go/, 2014.

[2] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Pannershelvam, M. Lanctot, and S. Dieleman, “Mastering the game of Go with deep neural networks and tree search,” Nature, vol. 529 no.7587, pp. 484-489, 2016.

[3] E. Gibney, "What Google’s winning Go algorithm will do next," Nature, vol. 531 no. 7594, pp. 284-285, 2016.

[4] E. Gibney, "Google secretly tested AI bot," Nature, vol. 541, pp. 142-142, 2017.

[5] N. Ernest, D. Carroll, C. Schumacher, M. Clark, K. Cohen, and G. Lee, "Genetic Fuzzy based Artificial Intelligence for Unmanned Combat Aerial Vehicle Control in Simulated Air Combat Missions," J Def Manag, vol. 6 no. 144, pp. 2167-0374, 2016.

[6] G. Booch, "I, for One, Welcome Our New Computer Overlords," IEEE Software, vol. 32 no. 6, pp. 8-10, 2015.

[7] K. Grace, J. Salvatier, A. Dafoe, B. Zhang, and O. Evans, "When Will AI Exceed Human Performance? Evidence from AI Experts," arXiv preprint arXiv:1705.08807, 2017.

[8] "Preparing for the future of artificial intelligence" The White House. Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf, 2016

[9] G. Allen, and T. Chan, "Artificial Intelligence and National Security," Retrieved from http://www.belfercenter.org/sites/default/files/files/publication/AI%20NatSec%20-%20final.pdf, 2017.

[10] S. Armstrong, N. Bostrom, and C. Shulman, "Racing to the precipice: a model of artificial intelligence development," Technical Report #2013-1, Future of Humanity Institute, Oxford University, 2013.

[11] C. Von Clausewitz, "On war: translated from the German by OJ Matthijs Jolles," The Modern Library, 1943.

[12] R. Yampolskiy, "Leakproofing the singularity: artificial intelligence confinement problem," Journal of Consciousness Studies, vol. 19 no. 1-2, pp. 194-214, 2012.

[13] R. Feynman, "Report of the presidential commission on the space shuttle Challenger accident. Appendix F," Retrieved from https://engineering.purdue.edu/~aae519/columbialoss/feynman-rogersrpt-app-f.pdf, 1986

[14] B. Watson, “The Drones of ISIS,” Defense One. Retrieved, August 26, 2017 from http://www.defenseone.com/technology/2017/01/drones-isis/134542/, 2017

[15] F. Pistono, and R. Yampolskiy, “Unethical research: how to create a malevolent artificial intelligence,” arXiv preprint arXiv:1605.02817, Retrieved August 27, 2017, from https://arxiv.org/abs/1605.02817, 2016

[16] S. Russell, D. Dewey, and M. Tegmark, "Research priorities for robust and beneficial artificial intelligence," Ai Magazine, vol. 36 no. 4, pp. 105-114, 2015.

[17] S. D. Baum, "On the promotion of safe and socially beneficial artificial intelligence," AI & Society, pp. 1-9, 2016.

[18] N. Bostrom, "Strategic implications of openness in AI development," Global Policy, vol. 8 no. 2, pp. 135-148, 2017.

[19] A. Critch, "Toward negotiable reinforcement learning: shifting priorities in Pareto optimal sequential decision-making," arXiv preprint arXiv:1701.01302, 2017.

[20] G. H. Kitmacher, W. H. Gerstenmeier, J. Bartoe, and N. Mustachio, "The international space station: A pathway to the future," Acta astronautica, vol. 57 no. 2, pp. 594-603, 2005.

[21] https://futureoflife.org/2017/08/20/leaders-top-robotics-ai-companies-call-ban-killer-robots/

[22] https://futureoflife.org/ai-principles/

[23] N. Bostrom, A. Dafoe, and C. Flynn, "Policy Desiderata in the Development of Machine Superintelligence," Working Paper, Future of Humanity Institute, Oxford University, 2016.

[24] https://www.partnershiponai.org/#

[25] S. Ozlati, and R. Yampolskiy, "The Formalization of AI Risk Management and Safety Standards," 31st AAAI Conference on Artificial Intelligence (AAAI-2017). 3rd International Workshop on AI, Ethics and Society. San Francisco, CA, USA. 2017.

[26] B. Goertzel, "Should Humanity Build a Global Nanny AI to Delay the Singularity Until It's Better Understood?" Journal of Consciousness Studies, vol. 19 no. 1-2, pp. 96-111, 2012.

ARTIFICIAL INTELLIGENCE FOR PLACE-TIME CONVOLVED WIRELESS COMMUNICATION NETWORKS

Ambuj Kumar,
Aarhus University, Herning, Denmark,
ambuj@btech.au.dk

Abstract – Previous works have brought to light that the dynamic and variable motions of potential network users, and other environmental factors, are the eternal threat to present and future wireless communications. One of my earlier works discusses this perennial and perpetual challenge as the place-time dependent functions. The phenomena was coined in my work as place time capacity (PTC), and place time coverage (PTCo), with both collectively known as place time coverage and capacity (PTC^2), are derived as the outcomes of dynamics that can be expressed as the functions of place and time. These phenomena degrade the efficiency of any wireless communication network (WCN) to that lowest point, from where, a network service provider (NSP) may not have any choice but to revamp the network. Artificial intelligence (AI), on the other hand, has been striding profoundly for the past several decades, fanning out its influence in various sectors of scientific and technological developments. However, AI is almost absent in the area of WCN dimensioning and optimization, especially for place–time events. This paper revisits the two place-time functions as WCN phenomena and, with the backdrop of these aspects, shall investigate the inevitable need for AI in WCNs, as well as demonstrating how AI can be part of present and future wireless communications.

Keywords – Artificial intelligence, dynamic path loss, network planning and dimensioning, place time capacity, place time coverage

1. INTRODUCTION

Network planning, dimensioning, and deployment are the processes that assure that any geographical area, where a certain amount of network users is expected, must be catered for in such a way that 'almost' every location in the region has a sufficient level of signal strength with a minimum assured capacity. Such kind of geographical area is defined as the 'Area of Interest' (AoI) in my previous works in [1] and in the PhD dissertation [2]. These processes are accomplished by suitably deciding the proper locations of the network sites[1], and then, installing the necessary equipment with the appropriate configuration (see Fig. 4).

Existing WCNs are very unlike their predecessors. It is expected that there will be more than 5 billion devices [3] by the year 2020, and this estimate greatly differs from what it was 40 years ago. As per ITU's document, about 80% of the youth population in more than 100 countries is online [4]. Modern wireless communication devices fall into the categories of high-power high-data-rate, low-power high-data-rate, and low-power low-data-rate [5]. These mixed characteristics of devices correspond to the type of user that holds them, and the nature of their use. Additionally, these devices can coexist in a single AoI, making things complicated for the NSPs.

It has been observed that, however well planned and deployed the WCNs are, there is always an ongoing demand for new sites, network upgrades, or even network revamps to deal with network users' needs. Many times these network upgrades are not due to any change in communication technology, such as 5th Generation [5] at our doorstep, or to repair equipment, but to accommodate a perpetual and ever-increasing challenge of an "ostentatious nature" [2] of the wireless communication environment, the PTC^2. In section 2, these phenomena will be discussed in detail, where it will be seen why this is most threatening for past, present, and future WCNs. Section 3 relates PTC^2 challenge to the problems in static network planning and dimensioning. Several attempts have been made to cater for this problem [2]. As we know, artificial intelligence (AI) is a technology that involves computational machines in learning a subject and

[1] A physical geographical location where the network base stations, along with necessary equipment, are installed to illuminate the surrounding area with necessary resources.

generating suitable actions for both predicted and unpredicted relevant challenges [6].
This problem has been attempted in this paper, and later, it shall be seen how AI can be an inevitable stairway to attain this seemingly impossible cliff. Section 4 reflects the need (influence) of artificial intelligence (AI) on the place and time events.

2. BACKDROP: OSTENTANEITY OF AN EVENT

Before we get into the broader discussions, let us take a moment to understand the background of why things are not as simple as they seem.

2.1. Unostentatious Events

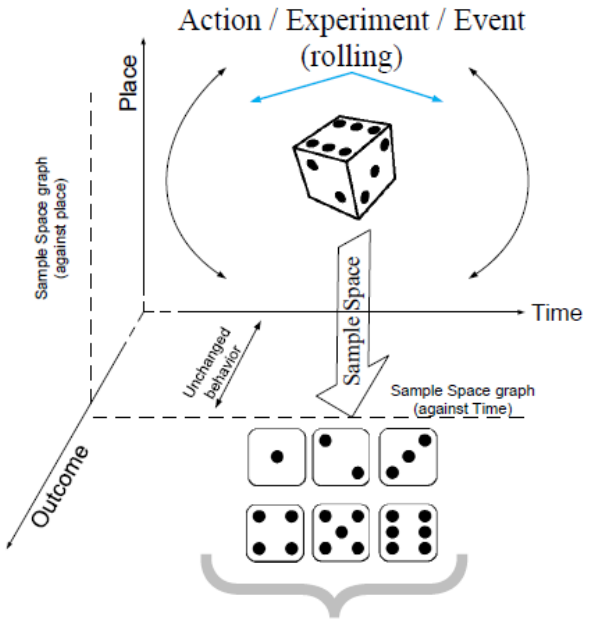


Fig. 1. Unostentatious Event

An 'Event', by definition in mathematics, is an experiment, process, or action that generates some measurable outcome. Fig. 1 shows a conventional die, which produces one of the six results when rolled on a flat surface, and when sample space is plotted against place and time, the graph is linear.
This indicates that no matter how many times we roll a die (on a flat surface), and anywhere across space, we always have a probability of 1/6 for every outcome [7]. Such events, where the behavior remains unchanged with time and place, are defined as an *Unostentatious Event* (UE) in our work in [2]. UEs, as illustrated in [2], are "*Events whose probabilities of all outcomes, or sample space, remain unchanged no matter at what places and at what times the event takes place.*"

Most, but not all, simple events are unostentatious and show the same or similar behavior at different places and/or different times.

2.2. Place-Time Events

In Fig. 2, we revisit the famous Schrodinger's cat example [8], except with a change, where this time the cage is a large walled circular park and nuclear radiation is somewhere within the contour of the park. In this example, as we can imagine, the probability of having the cat alive is not ½ (50%) all the time, and its likelihood is distributed across place and time. The death of the cat is directly proportional to its closeness to the radiation and the time it spends in the park at a particular position.
In our previous work [2], we have defined such events as the *Place-Time Events (PTE)s, whose one or more outcomes depend on at what time and at what position the incidence took place.* This means that an event can have different probabilities of outcomes at various times, while being at the same location. The vice-versa of this, however, is not true, as whenever the position changes, there is always some lapse in time.
By the term 'time', we mean both kinds of time, (i) ***absolute time***, saying the length of time for which the process is ongoing, and (ii) the ***event time***, indicating the time at which the event happened. As an example, I drove a car for 2 hours, is the absolute time, and I parked my car at 10:30 am, is the event time.
Nonetheless, in the present example, the probability of an outcome is definite concerning place and time, meaning that if the position of the cat and the time it spends at that location are known, we can determine the probability of the cat being dead or alive. Hence, this is a *partially unostentatious event*, as we see how the system behaves, and place and time are convolved with each other.

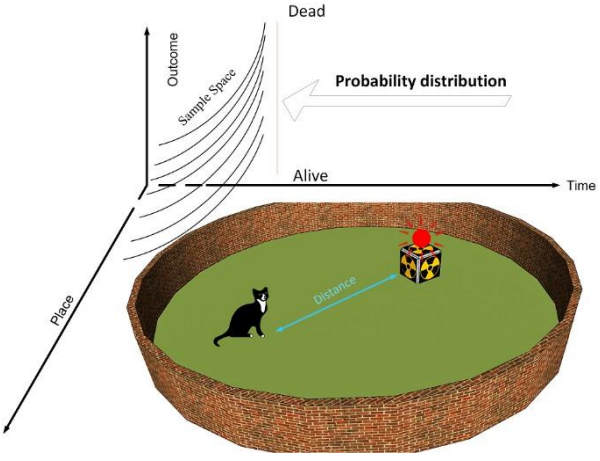


Fig. 2: Example of a place-time event: Schrodinger's thought experiment.

Such probabilities are also defined as the entangled probability in our previous work [2], if the experiment is performed at the same positions, dissipating the same length of time, but at different times, the probabilities of the outcomes are congruent. It is only the *absolute time* that matters and not the *event time*. In subsection 2.3, we discuss a phenomenon, where the probability is almost unpredictable. The discussion shall set a base for further investigations.

2.3. Ostentatious Events

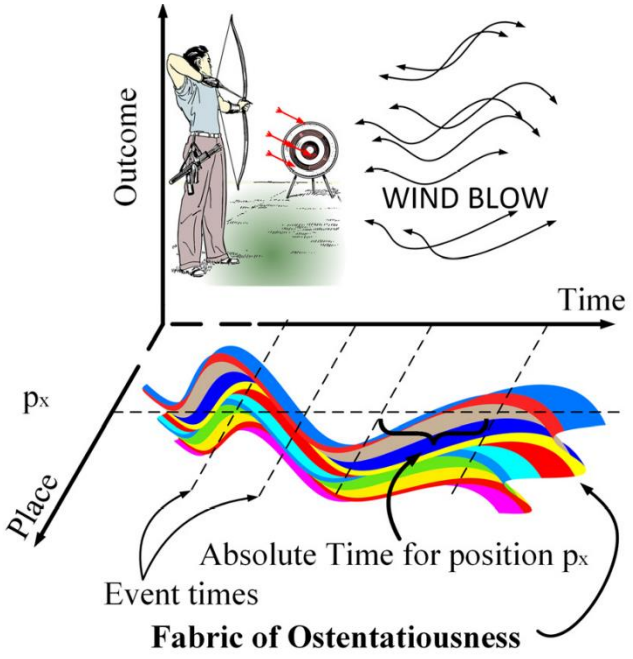


Fig. 3. Ostentatiousness of an event: An archer trying to meet bullseye in a windy field.

Let us examine the situation as illustrated in Fig. 3, where an archer targets a bullseye in an open and windy ground. Since the flow, thrust and direction of the wind are abrupt and unpredictable, the probability of releasing an arrow and the likelihood of hitting the target (bullseye) are not congruent. The situation is purely random and cannot be expressed by any probability distribution function (PDF) and can only be measured when the event occurs. The uncertainty in outcomes is due to varying air pressure hitting an arrow differently at different times, and, at different positions across a 3-dimensional space. This is an example of an **Ostentatious Event**, which we have defined in our work in [2]as "*Those place time events, whose one or more outcomes (sample space) are not interdigitated with either of place or time."* This means that outcomes shall show different probabilities at the same place and at different times, and therefore depend on both *absolute and event time*.

If we plot the outcomes of an ostentatious event against position and time, we have a graph (see, Fig. 3) that is defined as ***Fabric of Ostentatiousness***. Thus, for an ostentatious event, the probability of an outcome changes with position (place) and time. In Fig. 2, the graph is indicative, to show how the probability of 'the cat is alive' (or 'the cat is dead') varies with the position (distance) and time it spends. The ostentatious events here are defined as the ***Absolute Place Time Events*** (APE).

From the arguments above, it can be made out that the outcomes of an ostentatious event can have different probabilities of outcomes that are completely uncorrelated with each other.

2.4. Appropriateness of APE

The discussions so far might show familiarity with other previous works of probabilities [9], or space-time theories [10], etc. However, there is a big difference, and that is why the word '*Ostentatious*' was used to segregate this perspective from others. While in other cases, the probabilities are continuous/ continuum and may be described by the appropriate probability density functions, in this case, the events have entirely random outcomes.

It is essential to mention the three definitions, especially the ostentatiousness, to understand the challenges that are going to be discussed in the next section of this paper.

3. PLACE TIME COVERAGE AND CAPACITY: NSP's DUO ORDEAL

Can a WCN environment be ostentatious? Moreover, what does it mean by being ostentatious for any WCN? This section elaborates to answer these questions.

3.1. Understanding the network environment

In the previous section, the defined ostentatious events were introduced, the phenomena that have outcomes dependent on place and time. Apropos to which, assume that an area that has a complex morphology of dense urban, urban, rural, vegetation, etc., as shown in Fig. 4. The example in this figure, which is an area, also mentioned earlier as AoI, is a typical example of an NSP, which is trying to cater for this AoI with a certain amount of network sites. It can be seen in Fig. 4, that like any topology, this area also has almost all kinds of environments, namely Dense Urban (DU), Urban (U), Sub-Urban (SU), Rural (RU), Vegetation (VE), and water bodies (WB). Such a scenario can be found in any metropolitan to medium cities across the world; Paris is one of such example, which embraces Dense

Urban surroundings of the Eiffel Tower, River Seine, and vegetation spots like Parc Floral, etc.

What is important here is to see that the density and property of network sites are dependent on their location in AoI. The sites in DU and U are tightly packed, may be of less height, low transmit powers, etc., whereas, those located in SU and RU are taller and have higher transmit power. Some are umbrella sites that just have a purpose of catering for highways and dense vegetation, which have lower carriers and highest transmit power.

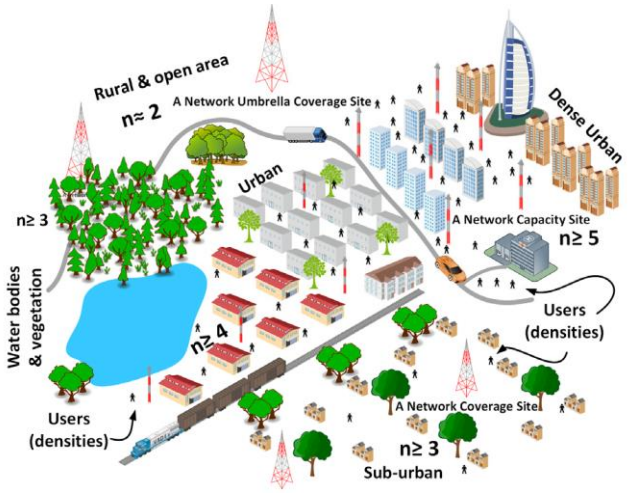


Fig. 4. A complex network environment

Here, 'n' is the **Path Loss Exponent** (PLE), which is the numerical representation of the 'area type' between Tx and Rx. For example, if the area between Tx and Rx is open, then 'n' takes the value of 2 and similarly, n=3 for urban, n=4 for high dense urban, and likewise. Table 1 in [11] describes characteristics of different physical operating environments that determine the values of 'n' corresponding to the different area types. Therefore, while planning the number of Tx sites required to cover an area (also termed as target area), an NSP can have an estimate of the density of sites by identifying the area type. This means that once the area type is known, how quickly a signal dissipates in the target area while moving away from Tx can be calculated, which eventually, evaluates the site density. The argument behind these different configurations is the way network planning and deployment is approached. Usually, a network is planned based on the probability of finding users in a particular location within AoI. The likelihood of more users, means more resources, thereby, causing higher demand for network sites in a unit area. Another reason for the different site density at various locations in AoI is the ease with which the electromagnetic wave (EMW) propagates through the medium. The EMWs from multiple transmitters around nearby areas have chances of finding a material, in forms of concrete, metals, plastics, glass, trees, rock, water, etc. These may be parts of topology or human-made structures, which may absorb, reflect, and/or diffract the EMW more profoundly than when material was absent. The absorption of signal leads to drop in the signal strength in the following path, and the reflections causes EMWs to divert from its designated path and break down into multiple signals in various other directions, which interacts with other EMWs creating interference at the point of intersection, thereby decreasing the signal strength. In DU and U, these two conjunctures are much more significant than other area types. In vegetation, the absorption is higher than reflections. The configuration, position, and the number of network sites immensely relates with the subscribers' distributions and propagation behavior. As mostly the subscriber densities are proportional to the area type, for example, DU has most users (network subscribers) available at most of the time, and likewise, comes DU, SU, RU, etc., the planning and deployment of network site distribution becomes easier.

Keeping the users distribution and area type, an NSP deploys a WCN; nevertheless, how long an appropriate network deployment can withstand, is the question we take up in next subsection.

3.2. The NSP's nightmare: Ostentations network behavior

(a) Aalborg City (Denmark) Carvival[2]

(b) Greek Open air theatre[3]

(c) Traffic Jam[4]

(d) Heavy Sale during Thanksgiving in Burbank, California[5]

Fig. 5. Unprecedented accumulation of people/ users

[2] Source: https://imgur.com/gallery/WMOMa
[3] Source: https://cbsla.files.wordpress.com
[4] Source: https://www.zmescience.com
[5] Source: https://www.nbcnews.com

In reference to the above discussions, let us assume the following:

- There is a network under consideration,
- there are 'u' users (subscribers) in the network,
- there are 's' sites in the network,
- the 'j^{th}' user is served by 'k^{th}' site,
- the users have instantaneous capacity, demand (or, throughput demand),
- the capacity demand function, in terms of bits per second of j^{th} user, is defined as bj(t),
- all users are dynamic in position and time,
- the velocity and position of jth user at any time 't' is defined as $v_j(t)$ and $p_j(t)$, which is the velocity and position function of user 'j'.

Let it be assumed that the j^{th} user, was the one staying at point 'P', decides to move within the network to reach point 'Q'. In this case, wherever the user j is moving, it is raising an additional demand at the location. Then, all the problems that were once associated with point P are now progressing to point Q. Each time a person or group of people move to a new place, an additional resource demand is created, which keeps on shifting from point to point, as the person or group moves. Thus, the locus of the path that a person or group moves, creates a demand curve in the network area.

Pictures in Fig. 5 show that there can be '**triggers**', such as music concerts, carnivals, market sale, etc., that compel or motivate users to accumulate and move in groups at unprecedented locations and at unforeseen times. The more populated the AoI is the more chances there are of such random wobbles of accumulations. These uncertainties give rise to dubiousness in capacity and coverage of a network, discussed below.

3.3. Place Time Coverage

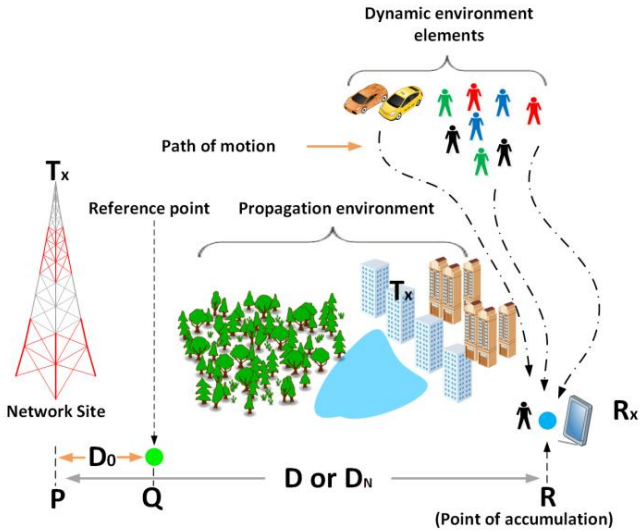


Fig. 6. Place time coverage; dynamic path loss model

As shown in Fig. 6, let us consider a scenario in which, a Transmitter (Tx) at position P and a Receiver (Rx) at position R are separated by a distance 'D', and, there exists a point Q between P and R, such that PQ is a free space. Then, to the Friis Model, the transmit and receive signal power relation can be expressed by [12]:

$$\frac{P_r}{P_t} = G_t G_r \left(\frac{\lambda}{4\pi D_0}\right)^2 \quad (1)$$

Where,

- P_t is the transmit power of the network site,
- P_r is the receive signal power at Q by the receiver,
- G_t is the transmit antenna gain,
- G_r is the receiver antenna gain,
- λ is the wavelength of the carrier in use,
- D_0 is the distance between P and Q,

Considering the antenna gains as unity, that is, $G_t = G_r = 1$, equation 1 yields free space Path Loss Model (PLM), given as [12]:

$$PL_{fs}(D_0) = \frac{P_r}{P_t} = \left(\frac{\lambda}{4\pi D_0}\right)^2 \quad (2)$$

The decibel format of equation 2 can be obtained as:

$$PL_{fsdB}(D_0) = 10\log_{10}(PL_{fs}(D_0)), \text{ or}$$
$$PL_{fsdB}(D_0) = 10log_{10}\left\{\left(\frac{\lambda}{4\pi D_0}\right)^2\right\} \quad (3)$$

The path loss experienced at point R, at a distance D from P, can be obtained in relation to the free space path loss at D_0, using the well-known path loss equation [12], mentioned below as:

$$PL(D) = PL(D_0) + 10log_{10}\left(\frac{D}{D_0}\right)^n. \quad (4)$$

As we can see, from equation 1, the received signal strength at a certain distance 'd' from Tx is primarily influenced by the path loss exponent 'n'.

Now, let us assume that a person from somewhere reaches point P and stands between Tx and Rx path (see, Fig. 6). In such a case, the path loss will not be the same as what is given in equation 1. Likewise, if more people surround Rx, there will be subsequent drops in the receive signal value. If an NSP decides to deploy a mobile WCN using standard path loss models, then the deployed network ought to fail in such conditions.

The new value of PLE can be expressed as:

$$n_{var} = n + N_{AR} \quad (5)$$

Where 'n_{var}' is the varying path loss, 'n' is the original PLE, and N_{AR} is the ostentatious PLE, which is the place-time factor of the path loss. The PLM in equation 4 can now be modified as:

$$PL_{dyn}(D) = PL(D_0) + 10log_{10}\left(\frac{D}{D_0}\right)^{n_{var}} \quad (6)$$

putting n_{var} in equation (5), we have,

$$PL_{dyn}(D) = PL(D_0) + 10log_{10}\left(\frac{D}{D_0}\right)^{n+N_{AR}} \quad (7)$$

expanding (7) we have,

$$PL_{dyn}(D) = PL(D_0) + 10log_{10}\left(\frac{D}{D_0}\right)^{n} + 10log_{10}\left(\frac{D}{D_0}\right)^{N_{AR}} \quad (8)$$

from (4) and (8), we have,

$$PL_{dyn}(D) = PL(D) + 10log_{10}\left(\frac{D}{D_0}\right)^{N_{AR}} \quad (9)$$

Differentiating both sides with time, we have

$$PL'_{dyn}(D) = PL'(D) + (N'_{AR})10log_{10}\left(\frac{D}{D_0}\right) \quad (10)$$

From previous discussions, we understand that $PL(D)$, and $10log_{10}\left(\frac{D}{D_0}\right)$ are constants; therefore,

$$PL'_{dyn}(D) = (N'_{AR})10log_{10}\left(\frac{D}{D_0}\right) \quad (11)$$

This indicates that with the change in the number of people, the environment's characteristics have also changed. An environment, which might have been 'rural' has now become pseudo suburban, or urban. Hence, the rate of change of PLM is now no more constant and depends on how N_{AR} incorporates user dynamics with time. The same equations can be expressed with the position. This dynamic path loss model is presented as PL_{dyn} which, apart from all previous parameters, considers another factor, N_{AR}, which is defined as ***Augmented Repercussive Exponent*** in [2]. The PL_{dyn} can accommodate any factor that may affect EMW propagation.

From the above discussion, it can be said that Pl_{dyn} can have variable values at the same or different locations. As the PLMs decide the number of sites needed to illuminate an area, the PL_{dyn} changes the paradigm by creating a perpetual variation in demand at different places, at different times. This kind of variable demand in network coverage is defined as the ***Place Time Coverage*** (PTCo) [2].

3.4. Place Time Capacity

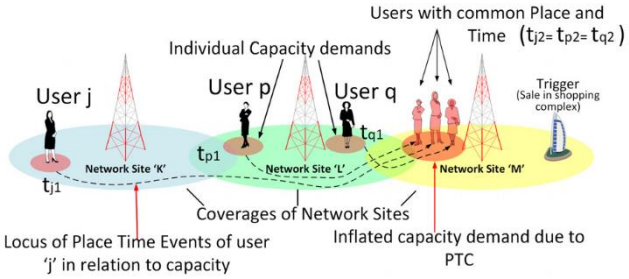


Fig. 7. Place time capacity

Considering Fig. 7, let us assume that j^{th} user, with a throughput demand of b_j, is standing at a position p_j. Then, accordingly, we can define a function that maps throughput with a position at any instant of time as:

$$iPTC(t) = p_j(t)b_j(t) \quad (12)$$

Where iPTC is termed as the "*Instantaneous Place Time Capacity (iPTC)*" in [2].

If we assume that user 'j' has traversed a path during the time interval (t_{j1}, t_{j2}), then, integrating (12) with respect to time for the range (t_{j1}, t_{j2}), we have

$$\int_{tj1}^{tj2} iPTC\ dt = \int_{tj1}^{tj2} p_j(t)b_j(t)\ dt \quad (13)$$

Applying integration by parts on RHS, we have,

$$RHS = p_j \int_{tj1}^{tj2} b_j dt - \int_{tj1}^{tj2} \left(\frac{dp_j}{dt} \int b_j dt\right) dt \quad (14)$$

Since, $p_j(t)$ is position function, differentiating it with time will give velocity function for user j, which, as mentioned earlier, is $V_j(t)$. Further, as we know that $b_j(t)$ is data rate in bits per second (bps), integrating it with time will give us the data volume, $D_j(t)$, which is generated during the time interval (t_{j1}, t_{j2}). Therefore, accommodating these points in (14), and equating LHS=RHS in (13) and (14), we have,

$$\int_{tj1}^{tj2} iPTC\ dt = p_j D_j(t)|_{tj1}^{tj2} - \int_{tj1}^{tj2} v_j\, D_j dt \quad (15)$$

iPTC, when integrates over a time interval, yields the place time capacity (PTC) demand generated during the given time interval. Therefore, equation (15) can be rewritten as,

$$PTC|_{tj1}^{tj2} = p_j D_j(t)|_{tj1}^{tj2} - \int_{tj1}^{tj2} v_j\, D_j dt \quad (16)$$

Equation 16 clearly indicates the ostentatiousness of a network with user dynamics. The first term on the RHS

of equation 16 shows that user j has raised a demand to cater a volume of data Dj, while being at position p_j and, the second term shows the aggregation of data volumes while in motion. We can extend equation (16) for other individual users and groups (see Fig. 7). We can also see that in Fig. 7 that each user creates its own locus of PTC demand, which is an equivalent to the requirement of additional resources.

3.5. PTC^2: Need of unorthodox approach

There are a variety of ways and complexities to identify and define PTC and PTCo, and the discussions in subsections 3.3 and 3.4 are confined to fewer parameters. Nonetheless, as both PTC and PTCo depend on dynamics within the network area, they usually complement each other. We have combined them with the common definition of place time coverage and capacity, or PTC^2 [2] to indicate their correlation.

From our previous discussions, we can conclude that the PTC^2, being an aggregation of the dynamic behaviors of all users in a network, cannot be handled by conventional network planning and deployment. We need 'intelligence' to understand its behavior. In the next and final section of this paper, how artificial intelligence can handle this challenge is elaborated.

4. DEALING WITH PTC^2 'ARTIFICIAL INTELLIGENTLY'

Despite efficient network planning and volumes of investments in optimization, and revamping of networks, WCNs are not able to get rid of PTC^2, and it is an eternal and perpetual phenomenon. With future demand on more throughput, humongous user mobilities, and many more devices, with the latest being the Internet of things (IoT), the PTC^2 is undoubtedly a significant worry. The increasing user densities and convenient but powerful handheld devices increase the chances of frequent and severe wobbles of accumulations and movements in groups, which is undoubtedly going to hurt more. As WCNs are planned based on some specific user distributions, the cushion for users roaming into a site is very narrow, which often leads to network congestion. If the problem is itinerant and clandestine, the only way out is to handle it 'intelligently'. Involvement of machines to perform intelligent tasks is even better; as then, the system is more robust and accurate. This leads us to propose the percolation of artificial intelligence (AI) in the WCN system. Later sections shall elaborate this scope.

4.1. AI-Assisted Architecture

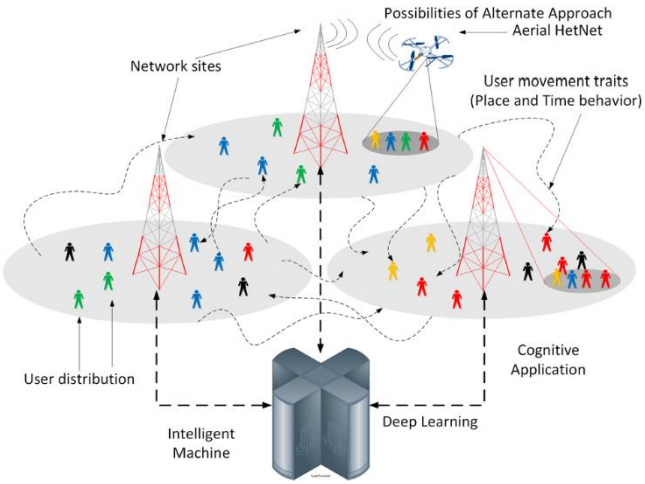


Fig. 8. AI-Assisted Architecture

Fig. 8 shows an AI-Assisted Architecture (AAA) that is found potent to handle the ostentatious network behavior. AAA is expected to have the following components:

- an efficient machine that has sufficient capacity to control large data and functions;
- the set of deep learning and cognitive response algorithms to iteratively configure and drive the network according to the contentiousness of the system;
- an AI-assisted network, means all the network equipment are controlled by the AI machine.

4.2. How should AAA respond to the PTC^2?

The functioning of AAA is discussed below.

4.2.1. Information aggregation

It is expected that both the base stations network and core network be connected with the common AI platform handled by the AI machine. In this way, the change in the environmental characteristics and user wobbles that are sensed by the on-site base stations can be sent to the central computer to learn the environment user behaviors. This is to be noted that environment and user behavior are distinct functions. An environment can be varied by a change in the material composition, which changes the dielectric property of the environment, and therefore, is purely a coverage issue (PTCo). Whereas the accumulation of users is both a coverage (change in material dielectric) and capacity (change in the count of user devices) issue. Both these changes can be sensed by an intelligent antenna system and advanced base stations. We have discussed once such kind of smart network system in our work [2].

4.2.2. Deep Learning

Although the PTC^2 wobbles may seem random, they are not entirely so. If it is observed, most of the severe wobbles are triggered by some periodic events, like carnivals and festivals, and so on. It is proposed that this part is to be done by the deep learning algorithms. Such algorithms can precisely identify the locations and many of wobbles, if not all of them, then at least many of them. Also, with efficient deep learning algorithms, the system can identify where the groups of potential users are heading to. Therefore, we have to involve capable machines that can aggregate all the information and process it in a single time to identify the possibilities of accumulations and the additional resource requirement. The predictions, however, can be modified or regenerated iteratively by continuously processing the network information and learning the network.

4.2.3. Disseminating actions

Once the system knows the probable problematic locations, the AI machine can allocate the new configuration of the network, which can be observed by the connected smart equipment, such as smart antennas, and intelligent BSs.

It is to be noted that a WCN is a closed system. This means that even though things are dynamic in the system, there is a marginal change in the number of users. Hence, the reconfiguration of the network may require the shifting of carrier channel from one base station to another and down-tilting the antennas to confine resources (see Fig. 8). However, many times the accumulation is so severe that even the shifting of resources does not solve the problem. In such cases an AAA system can also use the on-demand basis of spectrum allocation to have the time-based additional spectrum to cater enormous accumulations, as shown in Fig. 5.

4.2.4. Integrating Alternate Solutions

Several research works are being done to cater users through alternative ways. The most promising and looked upon solution is using unmanned aerial vehicles, aka drones, instead of using ground-fixed structures for mounting base stations (BS) [13], Purnima et al. proposed Aerial Hetnet that follows moving crowds.

AI can also play a primary role in integrating such alternate solutions to work in coordination with the prime network to offload the additional resource demand.

A lot of work is required to be done in this regard. Active and smart networks, deep learning, cognitive algorithms, etc., are some of the many areas to be worked on. Nonetheless, it must be said that AI is an inevitable paradigm that is yet, but very soon to, come full-fledged.

REFERENCES

[1] Kumar Ambuj, P.L. Mehta, and R. Prasad, "Place Time Capacity – A novel concept for defining challenges in 5G networks and beyond in India," 2014 IEEE Global Conference on Wireless Computing and Networking (GCWCN), 2014, pp. 278–282.

[2] Kumar, Ambuj. "Active Probing Feedback based Self Configurable Intelligent Distributed Antenna System: For Relative and Intuitive Coverage and Capacity Predictions for Proactive Spectrum Sensing and Management", Aalborg University, 2016.

[3] Cisco Visual Networking Index: "Global Mobile Data Traffic Forecast Update, 2016-2021 White Paper", https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html.

[4] http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2017.pdf.

[5] ITU RECOMMENDATION: "ITU-R M.2083-0, IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond", 2015.

[6] W. Ertel, Introduction to Artificial Intelligence. Springer Science & Business Media, 2011.

[7] "Mathematics Stack Exchange." [Online]. Available: http://math.stackexchange.com/questions/598838/.

[8] Gribbin, John, "In Search of Schrodinger's Cat: Quantum Physics and Reality", Random House Publishing Group, 2011, p. 234. ISBN 0307790444.

[9] C. Ash, "The Probability Tutoring Book: An Intuitive Course for Engineers and Scientists", 1 Edition. New York: Wiley-IEEE Press, 1996.

[10] Will, Clifford M, "Space-Time Continuum". Grolier Multimedia Encyclopedia. August 2010.

[11] RECOMMENDATION "ITU-R P.1411-6, Propagation data and prediction methods for the planning of short-range outdoor radiocommunication systems and radio local area networks in the frequency range 300 MHz to 100 GHz", 2012.

[12] RECOMMENDATION "ITU-R P.525-3, Calculation of free-space attenuation", 2016.

[13] Mehta, Purnima Lala; Prasad, Ramjee., "Aerial-Heterogeneous Network: A Case Study Analysis on the Network Performance Under Heavy User Accumulations", Wireless Personal Communications (Springer), Volume 96, Issue 3, October 2017, pp 3765-3784.

BAYESIAN ONLINE LEARNING-BASED SPECTRUM OCCUPANCY PREDICTION IN COGNITIVE RADIO NETWORKS

Ahmed Mohammed Mikaeil
Department of Electronic Engineering
Shanghai Jiao Tong University, Shanghai, China
ahmed_mikaeil @sjtu.edu.cn, ahmed_mikaeil@yahoo.co.uk

Abstract – Predicting the near future of primary user (PU) channel state availability (i.e., spectrum occupancy) is quite important in cognitive radio networks in order to avoid interfering its transmission by a cognitive spectrum user (i.e., secondary user (SU)). This paper introduces a new simple method for predicting PU channel state based on energy detection. In this method, we model the PU channel state detection sequence (i.e., "PU channel idle" and "PU channel occupied") as a time series represented by two different random variable distributions. We then introduce Bayesian online learning (BOL) to predict in advance the changes in time series (i.e., PU channel state.), so that the secondary user can adjust its transmission strategies accordingly. A simulation result proves the efficiency of the new approach in predicting PU channel state availability.

Keywords – Bayesian online learning, cognitive radio, primary user, spectrum occupancy prediction.

1. INTRODUCTION

In cognitive radio networks a secondary user is allowed to opportunistically utilize the vacant spectrum channels left by the primary user without interfering with their transmission. One of the key challenges for secondary users in cognitive radio networks is how to know when to occupy or leave the spectrum (i.e., the channels) for primary users' transmission. To tackle this problem, the secondary user must be capable of predicting in advance the channel availability of the primary user (i.e., whether the PU channels' status are "idle" or "busy") so that it can occupy or leave the channels for PU transmission.

The spectrum occupancy prediction problem has been widely investigated, for example, the idea of predictive spectrum access was first introduced in [1], in which the authors utilize Hidden Markov Model (HMM) to solve the spectrum occupancy prediction problem. Later on, the HMM-based spectrum prediction model received great attention in the literature [2-4]. And, due to the fact that HMM-based approaches require a priori knowledge of the PUs' traffic pattern, other machine learning approaches such as neural network [5], Bayesian inference [6] and online support vector regression (SVR) [7] have been adopted for the prediction of PU channel availability. However, these prediction techniques consider only time-invariant PU model behaviors. While in real-world cognitive radio systems, PU traffic patterns can also exhibit time-variant traffic patterns, which is hard to characterize using the above-mentioned machine learning algorithms. On the other hand, the Bayesian online learning algorithm (BOL) [8] has a capability to track both time-variant and time-invariant dual-states switching time series behaviors. Motivated by the fact that the nature of the PUs channel state availability can be also modeled as dual-states switching time series, we propose a new spectrum occupancy (PUs channel state) prediction technique that utilizes BOL to perform PU channel availability prediction in cognitive radio network. In more details, we captured the PU channel state energy detection sequence using a time series that switches over the time between two different random distributions representing the PU channel state (i.e., PU idle or PU occupied). We then fed this time series as an observations sequence into a BOL prediction algorithm to estimate or predict in advance the point of the time when the change will occur between the two states of the time series so that SUs can adjust their transmission strategies accordingly. The experimental results show the effectiveness of the BOL algorithm in predicting the changing points of the time series that were generated to capture PU channel availability.

The paper is organized as follows: first, we introduce the system model for energy detection and time series generation. Second, we present the BOL method to predict the time series that is generated to capture the PU channel state, followed by the simulation results and the overall work conclusion.

2. SYSTEM MODEL

The system model for predicting the PU channel state is illustrated by the block diagram Fig. 1. This diagram contains the energy detection block model that detects the PU channel state (PU signal present or absent), the time series generation block model to capture the PU channel state detection sequence followed by the Bayesian online learning algorithm block to predict the near future of the PU channel state (i.e., to detection changing points or the switching point of the time series) by utilizing the previously detected channel state information.

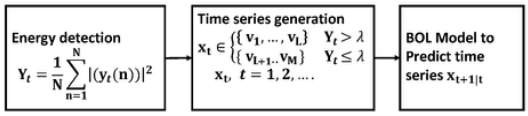


Fig. 1. Block diagram of PU channel state perdition based on BOL algorithm

In the diagram depicted in Fig.1, Y_t : is the instantaneous energy statistic, λ : is the detection threshold, x_t : is the time series that is generated to capture the detection sequence or PU channel states (i.e., PU present or PU absent) over the time. The time series x_t is represented using two different random distributions for each state (i.e., $x_t \in \{v_1, v_2 \dots v_L\}$ for PU signal absent state, and $x_t \in \{v_{L+1} \dots . v_M\}$ for PU signal present state).

3. ENERGY DETECTION MODEL

In energy detection based spectrum sensing the SU sensor observes the licensed spectrum to determine whether the primary user signal is present or absent by doing the binary hypothesis test on the received signal over the time $1 \le t \le T$, as follows:

$$y_t(n) = \begin{cases} w_t(n) & H0 \text{ (absent)} \\ \sqrt{\gamma_t}\, s_t(n) + w_t(n) & H1 \text{(present)} \end{cases} \quad (1)$$

where $y_t(n)$ the observed received signal $s_t(n)$ the primary user's signal, and it is assumed to be Gaussian i.i.d random process with zero mean and variance ${\sigma_t}^2$, $w_i(n)$ is the noise, and assumed to be Gaussian i.i.d random process with zero mean and variance ${\sigma_t}^2$, γ_t: the SNR at time t. The instantaneous energy statistic at the secondary user sensing node Y_t can be represented by

$$Y_t = \frac{1}{N}\sum_{n=1}^{N}|(y_t(n))|^2 \quad , 1 \le t \le T \quad (2)$$

where, N: is the number of samples used by the sensing node SU for energy detection, T: the overall system simulation time and Y_t is a random variable whose probability density function (PDF) is chi-square distribution with 2N degrees of freedom for the complexed value ($y_t(n)$,and N degrees of freedom for the real value case. For the value of $N \ge 200$, Y_t can be approximated using the Gaussian distribution. Therefore, the distribution of the power test Y_t for wide band signal follows:

$$Y_t \sim \begin{cases} (\sigma_t{}^2, 2\sigma_t{}^4/N) & H0 \\ (\sigma_t{}^2(1+\gamma_t), 2\sigma_t{}^4(1+\gamma_t)^2/N & H1 \end{cases} \quad (3)$$

If we assume that the noise variance and SNR at every sensing node remains unchanged during the observation time t, then ${\sigma_t}^2$ and can be written as γ_t ${\sigma_t}^2 = {\sigma_u}^2, \gamma_t = \gamma_u$. Thus, for a chosen threshold $\lambda_t = \lambda$ the probability of false alarm P_f can be written as:

$$\begin{aligned} P_f(\lambda) &= \Pr(Y_t > \lambda | H0) \\ &= \frac{1}{\sqrt{2\pi}\sigma_u} \int_{\lambda}^{\infty} e^{-(\lambda-\sigma_u)^2/\sqrt{2}{\sigma_u}^2} \\ &= Q\left(\frac{\lambda}{{\sigma_u}^2} - 1\right) \end{aligned} \quad (4)$$

where $Q(.)$ is the complementary distribution function of Gaussian distribution with zero mean and unit variance. From equation (4) and for a given probability of false alarm P_f, the single user decision threshold can be written as:

$$\lambda = \left(\sqrt{\frac{2}{N}} Q^{-1} + 1\right) {\sigma_u}^2 \quad (5)$$

where $Q^{-1}(.)$ is the inverse of the $Q(.)$ function. And the instantaneous primary user channel state detection sequences the sensing results over the time can be written as a function of the decision threshold as follows:

$$D_t = \begin{cases} 0 & \text{PU signal absent} \quad Y_t < \lambda \\ 1 & \text{PU signal present} \quad Y_t \ge \lambda \end{cases} \quad ,1 \le t \le T. \quad (6)$$

4. TIME-SERIES GENERATION BASED ON ENERGY PRIMARY USER DETECTION SEQUENCE

After the detection of the instantaneous primary user channel state D_t (i.e., PU signal present or PU absent) based on the energy detector as explained in the previous section (other spectrum-sensing algorithms such as data fusion based cooperative sensing approach could be also used here). Our next goal is to predict the near future of the primary user channel state. In order to do so, we denote the period that the primary user signal is absent as "idle state" and the period that the primary user signal present as "occupied state". And to estimate the time when the channel state will switch from idle to occupied or vice versa (i.e., the change-points time) we generate a time series x_t to capture the instantaneous detection sequence signal D_t .Then we transfer the two states of the time series ("PU present" or "absent",) into observations using two different random distributions for each state (i.e., $x_t \in \{v_1, v_2 \dots v_L\}$ for idle state and $x_t \in \{v_{L+1} \dots . v_M\}$ for occupied state. The equation that capture PU detection sequence x_t can be written as follows:

$$x_t \in \begin{cases} \{v_1, v_2, \dots, v_L\} & Y_t < \lambda \\ \{v_{L+1}, \dots, v_M\} & Y_t \geq \lambda \\ 1 \leq t \leq T \end{cases}, \quad (7)$$

We can note that x_t formulates a time series of two non-overlapping states over the time T. The effectiveness of the Bayesian online learning algorithm in predicting the change-points in such a time series has been shown in many applications such as finance, biometrics, and robotics [8].

5. TIME-SERIES PREDICTION BASED ON BAYESIAN ONLINE LEARNING ALGORITHM

We assume a sequence of observations $x_t = x_1, x_2, x_3, \dots$ with two non-overlapping states, as denoted by equation (7). The delineations between the two non-overlapping states are called the change-points. To determine these change-points we need to estimate the posterior distribution over the current "run length=: r_t"or the time since the last change-point, given the data so far observed. Under the assumption that change-points occur as a stochastic process, the data between change-points are i.i.d distributed, the parameters are independent across the change-points, and when the change-point has occurred the run length r_t will drop to zero; otherwise, the run length is increased by one. We use the BOL algorithm to calculate the posterior run length at time t, i.e. $P(r_t|x_{1:t})$, sequentially. This posterior is used to make online predictions robust to underlying regime changes through marginalization of the run length variable as follows:

$$P(x_{t+1}|x_{1:t}) = \sum_{r_t} P(x_{t+1}|x_{1:t}, r_t) P(r_t|x_{1:t}) = \sum_{r_t} P(x_{t+1}|x_r) P(r_t|x_{1:t}) \quad (8)$$

where x_r refers to the set of observations $x_t$associated with run length r_t, and $P(x_{t+1}|x_r)$ is computed using the underlying predictive model UPM (the training model set), to find $P(r_t, x_{1:t})$, we estimate the run length distribution $P(r_{ti}, x_{1:t})$, for $i = 1, 2, \dots,$ tof run length r_t . For each time step t, the run length distribution contains i-elements of probabilities such that $\sum_{i=1}^{t} r_{ti} = 1$. The run length posterior is found by normalizing the joint likelihood as:

$$P(r_t|x_{1:t}) = \frac{P(r_{ti}, x_{1:t})}{\sum_{r_{ti}} P(r_{ti}, x_{1:t})} \quad (9)$$

If we denote the joint likelihood distribution of the run length r_t at time t for the observed data $x_{1:t}$, $P(r_{ti}, x_{1:t})$ as ϕ_t: =$P(r_{ti}, x_{1:t})$ we can update the joint likelihood online recursively using:

$$\phi_t = P(r_{ti}, x_{1:t}) = P(x|r)P(r) = \sum_{(r_t-1)i} P(r_{ti}, x_{1:t} | r_{(t-1)i}, x_{1:t-1}) P(r_{t-1}, x_{1:t-1}) = \sum_{(r_t-1)i} P(r_{ti}|, r_{(t-1)i}) P(x_t|r_{(t-1)i}, x_r) \quad (10)$$

where $P(r_{ti}|, r_{(t-1)i})$ is the change-point prior or hazard function and $P(x_t|r_{(t-1)i}, x_r)$is the likelihood or the underlying predictive model UPM (the model training data set). All the distributions mentioned so far are implicitly conditioned on the set of hyper-parameters θ.

Assuming a simple BOL model represented using a constant hazard function $H(r|\theta_h) := \theta_h$, this means $P(r_t = 0|r_{t-1}, \theta_h)$ is independent of r_{t-1} and is constant. We can represent the underlying predictive model with a basic predictive model that model a scalar ($x_t \in \mathcal{R}$) by placing a normal-inverse-gamma prior on i.i.d Gaussian observations [9]:

$$x_t \sim (\mu_t, {\sigma_t}^2) \quad (11)$$

$$\mu_t \sim (\mu_0, {\sigma_t}^2/\kappa_t), {\sigma_t}^{-2} \sim \text{Gamma}(\alpha_t, \beta_t) \quad (12)$$

For this model the parameters are $\eta := \{\mu_t, {\sigma_t}^2\}$ and the model hyper-parameters are $\theta = \{\mu_t, {\sigma_t}^2, \kappa_t, \alpha_t, \beta_t\}$. A new value for $\mu_t$and ${\sigma_t}^2$ are sampled at each change-point. The posterior on η is updated at every new data point for each run length, for example for a model with a training set $\{x_1, x_3, \ldots x_L\}$ or $x_{1:L}$and initial mean and variance $\mu_0, {\sigma_0}^2$ respectively the first update of η (the corrected prior mean and variance) is: $\mu_1 = (\mu_0 - E(x_{1:L}))/\text{std}(x_{1:L})$ and ${\sigma_1}^2 = {\sigma_0}^2/\text{std}(x_{1:L})$ where std := standard deviation.

The BOL algorithm after the training is written as given in [10] is as follows:

1. Initialize or calculate from the training data set the corrected prior mean μ_1, the corrected prior variance ${\sigma_1}^2$, the degree of freedom β_1, the run length distribution $P(r_{(1)i}) = 1$, α_1 and κ_1.

2. While (new data x_t is available) do:

3. Compute the Gaussian prediction function by the student's t-distribution (which gives a posterior predictive distribution on x_t of)

$$\pi_t^{(r)} = P(x_t|x_r, \theta) = P(x_t|\mu_t, {\sigma_t}^2, \beta_t, \kappa_t) = St_{2\alpha_t}\left(\mu_t, \frac{\beta_t}{{\sigma_t}^2}\frac{\kappa_t}{\kappa_t+1}\right) \quad (13)$$

St =: Student's t-distribution probability density function

4. For $i = 1 = 1$ to $t-1$, compute growth probabilities

$$P(r_{ti}, x_{1:t}) = P(r_{t-1}, x_{1:t-1})\,\pi_t^{(r)}\,(1 - H(r_t)) \quad (14)$$

where we assume that the hazard function $H(r_t) = \lambda^{-1}$,where λ is a timescale parameter.

5. Compute change-point probabilities

$$P(r_{ti}, x_{1:t}) = \sum_{i=1}^{t-1}\frac{P(r_{ti}, x_{1:t})}{\lambda - 1} \quad (15)$$

6. Compute run length distribution

$$P(r_{ti}|x_{1:t}) = \frac{P(r_{ti}, x_{1:t})}{P(x_{1:t})} = \frac{P(r_{ti}, x_{1:t})}{\sum_{i=1}^{t} P(r_{ti}, x_{1:t})} \quad (16)$$

7. Update sufficient statistics

$$\mu_{t+1} = \frac{\kappa_t\mu_t + x_t}{\kappa_t + 1} \quad (17)$$

$$\kappa_{t+1} = \kappa_t + 1 \quad (18)$$

$${\sigma_{t+1}}^2 = \frac{1}{\beta_t}\left[(\kappa_t + 1) + \frac{1}{2}(x_t - \mu_t)^2\right] \quad (19)$$

$$\beta_{t+1} = \beta_t + \frac{\kappa_t (x_t - \mu_t)^2}{2(\kappa_t + 1)} \quad (20)$$

$$\alpha_{t+1} = \alpha_t + \frac{1}{2} \quad (21)$$

8. If

$$\underset{i}{\text{argmax}} \quad P(r_{ti}|x_{1:t}) \quad (22)$$

Then the change-point has occurred, reset run length to zero $r_t = 0$. If not, increment $r_t = r_{t-1} + 1$.

9. Perform prediction

$$P(x_{t+1}|x_t) = \sum_{i=1}^{t-1} P(x_{t+1}|x_t, r_t)\,P(r_{ti}, x_{1:t}) \quad (23)$$

10. Go to step 2.

6. SIMULATION RESULTS

To evaluate the performance of the BOL algorithm in predicting PU channel state availability we generate a simulated detection sequence considering single user cognitive radio system with noise variance ${\sigma_u}^2$=1 and SNR γ_u= −22dB. The local energy detection decision is made after observing N =1000 samples. We employ equation (6) to generate the time series that captures the simulated detection sequence.

Figure 2 shows the performance of the BOL algorithm in predicting the time series that is generated to capture the primary user channel state detection sequence for single user "sensing node" cognitive radio network with a randomly distributed channel occupancy over $T = 350$ time points after training the algorithm offline over a set of training observations $L = 150$ time points (we use millisecond as time point here). For training the algorithm we use $\mu_0 = 2$ and ${\sigma_0}^2 = 1$, the timescale parameter $\lambda = 2000, \alpha_1 = 1$, and $\kappa_1 = \beta_1 = 1/{\sigma_1}^2$. In Fig.2 the top plot shows the simulated detection sequence signal, where 150 ms of the simulated detection sequence signal used for training the hyper-parameters θ, and the remaining 350 ms is used for testing the algorithm. The middle plot shows the generated time series to capture the primary user channel state detection sequence with the random distribution for idle states represented by $x_t \in \{1,2,3\}$ and $x_t \in \{4,5,6\}$ for occupied states.

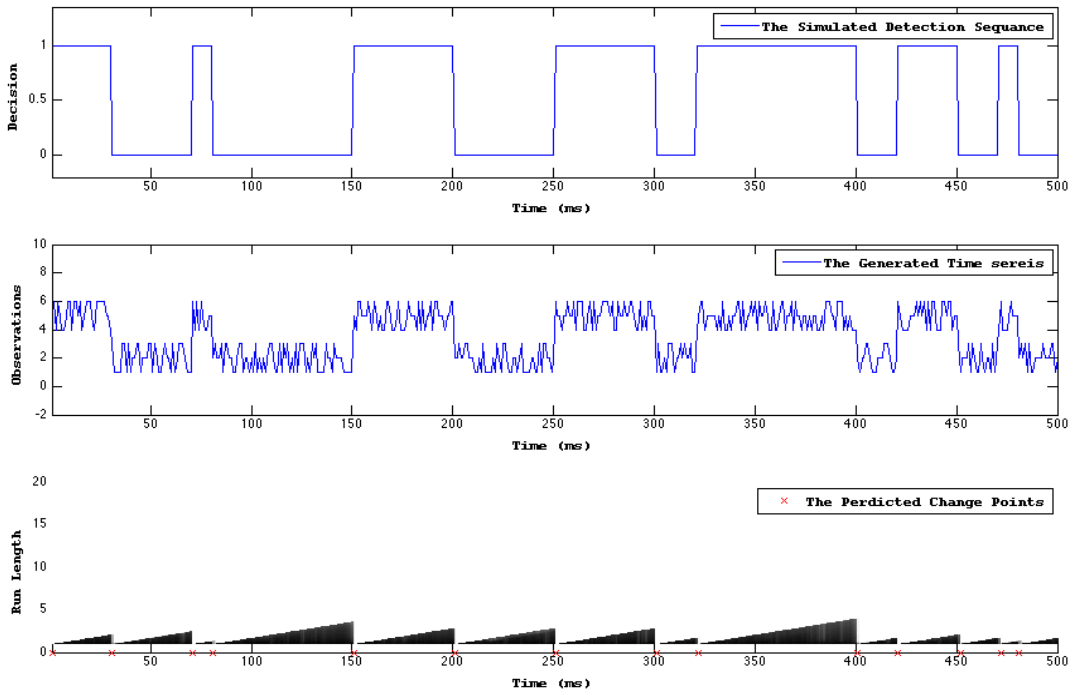


Fig. 2. The simulated detection sequence signal (150 time points used to train the hyper-parameters θ), the generated time series to capture the detection sequence and the performance of the BOL algorithm in predicting the change-points for 350 time points testing time series, (assuming the time point is in millisecond: = ms)

The bottom plot shows the posterior probability of the current run length $P(r_t|x_{1:t})$ at each time step, using a logarithmic color scale; the lighter pixels indicate higher probability, the red crosses are the change-points and the darker pixels represent the current run length. As we can see run-length is dropped to zero immediately after the change-point. The time consumed for training the model is 5.223582 milliseconds while the testing time is only 0.204281 milliseconds.

7. CONCLUSION

In this paper, we have studied the problem spectrum occupancy prediction for a single user cognitive radio network based on the Bayesian online learning model. We modeled the detection sequence of primary user channel state availability as a time series changing over the time between two states (PU idle and PU occupied). We introduced Bayesian online learning to predict in advance the changes in the states of the time series. Finally, we evaluated the performance of our algorithm using a simulated PU detection sequence. The simulation results have verified the effectiveness of the BOL model in predicting PU channel state availability.

ACKNOWLEDGEMENT

I gratefully acknowledge the funding received from Shanghai Jiao Tong University to undertake my PhD. I also thank Prof. Bin Guo for his encouragement and help on the topic.

REFERENCES

[1] I.A. Akbar and W.H. Tranter, "Dynamic spectrum allocation in cognitive radio using hidden Markov models: Poisson distributed case," 1in Proc. 2007 IEEE Southeast Con., pp. 196-201.

[2] Z. Chen and R.C. Qiu, "Prediction of channel state for cognitive radio using higher-order hidden markov model," in IEEE Proceedings of the Southeast Con (Southeast Con), 2010, pp. 276-282.

[3] E. Chatziantoniou, B. Allen, and V. Velisavljevic, "An hmm-based spectrum occupancy predictor for energy efficient cognitive radio," in IEEE 24th International Symposium onPersonal Indoor and Mobile Radio Communications (PIMRC), 2013, pp. 601-605.

[4] Zhao, Yanxiao, Zhiming Hong, Guodong Wang, and Jun Huang. "High-Order Hidden Bivariate Markov Model: A Novel Approach on Spectrum Prediction." In Computer Communication and Networks (ICCCN), 2016 25th International Conference on, pp. 1-7. IEEE, 2016.

[5] V. K. Tumuluru, P. Wang, and D. Niyato, "A neural network based spectrum prediction scheme for cognitive radio," in 2012 IEEE ICC con., pp. 32-57.

[6] X. Xing, T. Jing, Y. Huo, H. Li and X. Cheng, "Channel quality prediction based on Bayesian inference in cognitive radio networks," in Proc. 2013 IEEE INFOCOM, pp. 1465-1473, 2013.

[7] Zhang, Zhao, Kaiqing Zhang, Feifei Gao, and Shun Zhang. "Spectrum prediction and channel selection for sensing-based spectrum sharing scheme using online learning techniques." In Personal, Indoor, and Mobile Radio Communications (PIMRC), 2015 IEEE 26th Annual International Symposium on, pp. 355-359. IEEE, 2015.

[8] Adams, Ryan Prescott, and David JC MacKay. "Bayesian online change-point detection." arXiv preprint arXiv:0710.3742 (2007).

[9] Turner, Ryan Darby. Gaussian processes for state space models and change point detection. Diss. University of Cambridge, 2012.

[10] Lau, Hon Fai, and Shigeru Yamamoto. "Bayesian online change-point detection to improve transparency in human-machine interaction systems." Decision and Control (CDC), 2010 49th IEEE Conference on.IEEE, 2010.

THE EVOLUTION OF FRAUD: ETHICAL IMPLICATIONS IN THE AGE OF LARGE-SCALE DATA BREACHES AND WIDESPREAD ARTIFICIAL INTELLIGENCE SOLUTIONS DEPLOYMENT

Abhishek Gupta
District 3, Concordia University, Montreal, Canada

Abstract – Artificial intelligence is being rapidly deployed in all contexts of our lives, often in subtle yet behavior-nudging ways. At the same time, the pace of development of new techniques and research advancements is only quickening as research and industry labs across the world leverage the emerging talent and interest of communities across the globe. With the inevitable digitization of our lives, increasingly sophisticated and ever larger data security breaches in the past few years, we are in an era where privacy and identity ownership is becoming a relic of the past. In this paper, we will explore how large-scale data breaches coupled with sophisticated deep learning techniques will create a new class of fraud mechanisms allowing perpetrators to deploy "Identity Theft 2.0".

Keywords – Artificial intelligence, data privacy, data security, ethics, fraud

1. INTRODUCTION

With the inevitable deployment of artificial intelligence (AI) at an unprecedented pace touching every conceivable aspect of our lives, it is of paramount importance that we begin considering some of the ethical implications of the development work and deployments being done in AI and how we can work towards mitigating adverse outcomes that might arise from the use of these systems.

This paper concerns itself with analyzing the long-run implications of large-scale data breaches and how, beyond the evident, immediate impacts, this will lead to the emergence of novel fraud techniques challenging societal balance. Specifically, the paper will focus on these negative outcomes being aggravated because of advances in deep learning (a subfield of AI).

Given how many diverse parts of our lives rely on products and services that use some sort of pattern recognition (one of the primary outcomes of applying AI techniques), the consequences of unlocking new hitherto anonymized datasets via the mosaic effect [1] means we will see the rapid rise of sophisticated fraud techniques that will span online, physical and biological characteristics of the target individual.

The idea of an individual's data exhaust [2] has gained traction in the security and privacy industry. We "emit" bits of personal data in numerous interactions, both online and offline. Government services [3], social media [4], grocery purchases [5], credit services [6], the list of places where we leave a digital footprint extends almost indefinitely. As the usage of mobile data increases and with the imminent arrival of 5G technologies, there is an increased risk of even more data flowing through these pipes. Perhaps this will also be supplemented by Internet of Things (IoT) devices, both in residential and industrial settings, which will collect even more information about consumer behavior and transmit them via protocols leveraging faster speeds and increased coverage due to the emergence of 5G technologies [7].

The pattern recognition outcomes as mentioned above are significant for two reasons; firstly, they discover existing correlations between different factors without being explicitly programmed to do so [8]. The second reason is more important because they can discover latent patterns to make decisions that are at once unknown to the developers of the system and opaque given current advances in being able to "explain" the reasoning behind the system taking particular decisions [9].

This paper will begin by explaining some of the key ideas that will be used recurrently to connect seemingly disparate developments in the field of deep learning, product development happening in different parts of the world and most importantly the data breaches that will enable the creation of novel

fraud techniques. Then we will go on to explore briefly some recent data breaches and their current impacts, as assessed by the cybersecurity community. We will tie all of this together to explain fraud scenarios that have the potential to arise as a consequence of piecing together the above and subsequently analyze how advances in deep learning techniques will turbocharge this process. Finally, we will look at some challenges that remain to be solved from an ethical, cybersecurity, regulatory and societal perspective. The paper makes preliminary recommendation to set the stage for future developments in this domain, especially as the pace of advances in AI quickens [10].

2. KEY IDEAS

2.1. Data brokers

Data brokers are companies that shy away from media exposure but are responsible for collecting a vast array of information [11] about individuals via different channels. Often they aggregate information from partners like social media companies, telecommunication firms, retail chains, e-commerce firms, etc. At the same time, these above-mentioned partners are also clients that look to augment the profiles that they have on their customers. Key concerns with the operations of the data brokers are that they collect in-depth information about consumers and their behaviors, often without their consent. The lack of consent is almost less shocking considering that most people are unaware [12] of the existence of these firms and the degree of profiling that they can do on a targeted individual.

In the context of this paper, the data brokers will be important because they often accumulate disparate datasets containing non-overlapping bits of information on an individual which if linked together can form an even richer representation of the individual for targeting purposes.

2.2 Mosaic effect

The mosaic effect results in a deeper analysis by combining several large datasets and making cross-references between them to coalesce previously de-linked information into a single, richer profile [1].

Edward Felten from Princeton University succinctly sums up the process of data collection and analysis today [1] as collecting, merging and analyzing data to infer facts about people. It is the merging process that gets a lot more effective as a consequence of the mosaic effect. During the first cycle of merging data, one is able to create more accurate portraits of users but follow-on cycles can do this in a more refined manner as they know more about the user behavior and identity which allows further merging.

This also has the added benefit of being able to generate smaller, more homogenous groups, i.e., increasing the ability to micro-target individuals while creating fine-grained analysis to inform policy-making (advertising, credit decisions, etc. by the acquiring firms) at a macro-level. On occasion, some of these insights arise without prior knowledge that such an objective could be achieved when data brokers are aggregating and analyzing this information.

Two examples illustrate the power of de-anonymization (the ability to identify an individual in an anonymized dataset) by combining different data points: Netflix movie ratings and AOL data search combination.

In the Netflix example [13], essentially after an anonymized dataset was released as part of a competition by Netflix to incentivize researchers to come up with more efficient recommendation systems, some researchers were able to cross-reference rating scores on rare movies with those on Internet Movie Database (IMDB) which has public ratings to identify the user and consequently reveal inferred information about the users, such as sexual orientation and potential political affiliations.

In the case of AOL [14], in an anonymized dataset that was released, a closer analysis of the actual search queries helped to localize the user to a specific region and then combining that with some easily accessible public information repositories, it became possible to identify users within that anonymized dataset.

Data brokers already leverage the mosaic effect to create rich consumer profiles that they can then sell to clients but following the large-scale data breaches, it will now be possible to “unlock” previously disjointed datasets via attribute groups that act as unique keys to link them together.

3. RECENT DATA BREACHES

This section will briefly look at the largest data breaches that have happened in recent history, looking at the specific bits of information leaked and a high-level view of potential consequences.

3.1. Yahoo

The most recent statement from Yahoo confirmed that all 3 billion accounts [15] with them were compromised, with information such as names, email addresses, dates of birth, encrypted passwords (for some fraction of the accounts the encryption was not state of the art) being leaked, as well as security questions and answers. What is troubling here is that, for example, weakly encrypted passwords and security question-answer pairs are usually similar for users across different web services, and this opens up the risk of users being compromised elsewhere.

Furthermore, this only helps to create more robust rainbow tables [16] which can be used as an easy mechanism to break through other online services that have less than ideal cybersecurity practices.

3.2. Adult Friend Finder

Approximately 412 million accounts were compromised in this leak which represented about 20 years' worth of information cutting across 6 databases [17]. The long temporal nature of this leak provides a unique window to the hacker community to find tidbits of information in terms of both old and new perspectives on the online identities of the users.

This also comes with the additional burden of revealing unsavory details of someone's online life with very tangible real-world consequences. This represents an almost 13 times larger breach compared to the Ashley Madison breach [18] which led to some suicides [19].

3.3. eBay

In this case, in 2014 all 145 million accounts [20] were compromised leading to the release of names, dates of birth, email addresses, physical addresses and salted-hashed passwords (though eBay declined to comment on the algorithm used to do so). What was of particular concern here is the leak of physical address information which adds another crucial link to the databases maintained by data brokers. Even though financial information was not leaked, being able to link the physical address, specifically the zip code gives away information of income ranges among other things via publicly available datasets from the national statistics databases.

3.4. Equifax

Equifax is one of the largest credit bureaus and the most recent large-scale compromise, more than 140 million accounts were hacked and birth dates, addresses, social security numbers (SSN) and in some cases driver license information was leaked. In about 200,000 cases consumers also had their credit card information exposed. Data brokers and other malevolent entities can pick up this information along with some of the ones mentioned before to create a pretty comprehensive profile of individuals who might have been present in all of the above-mentioned compromises. More importantly, the linking of SSN to other bits of personally identifiable information (PII) prepares unique keys that are invaluable in de-anonymizing large existing datasets.

4. EVOLUTION OF FRAUD

If we start to collect all the disparate pieces of information spread across the sample of large-scale compromises mentioned here, there are tremendous "unlocked" datasets that can now be fed into different pattern recognition systems to create the next generation of fraud techniques, i.e. Identity Theft 2.0.

Some of the recent research projects coming out of AI research labs across the world act as prescient reminders of a somewhat dystopic future if combined with the information leaks that we discussed above. As an example, Lyrebird [21] from Montreal, Canada uses "less than one minute" voice samples from an individual to generate voice output of them saying any statement that you input. Lyrebird is not perfect yet because it requires high-quality input samples with an array of sentences but as the technique improves, it would be possible to use voice samples over a telephone connection to collect enough data.

A research paper from the University of Washington [22] highlighted how they could take an audio sample and then synthesize video output of the person saying, with fairly realistic facial expressions and head movements, anything said in the audio. In its current iteration, this is limited in its application because of the need for high-quality video input for training purposes but again as techniques improve this could be applied more broadly.

For both of the above examples, it is already possible to synthesize fake audio and video output for public figures because it is easy to find high-quality training samples for them. In the case of synthesizing audio, anti-spoofing measures are not adequate. These synthesized samples can be used to impersonate users and trick authentication systems, potentially giving attackers access to unauthorized systems [23].

With the rising popularity and easy accessibility of genetic testing from firms like 23&Me [24], data compromises could be disastrous in terms of adding richer information to the open pool. [25] More so, this creates opportunities for targeted attacks on individuals, particularly those who are in the public domain. "Cloudhopping" is a technique where hackers compromise a low-hanging fruit on a particular cloud provider and are then able to "hop" onto the cloud infrastructure of a more valuable target. These techniques further increase the chances of information leaks, especially in targeted attack scenarios.

In a particularly far-reaching hack, Cloudflare leaked important tokens via a vulnerability dubbed "Cloudbleed" [26] which affected services from Uber, Fitbit and 1Password among others. These are all services that store highly personalized information about an individual which if combined with pieces of information from other data breaches mentioned in the previous section, via the mosaic effect, essentially enable the creation of a virtual avatar of an individual, replete with genetic and physical world information that can now be deployed to commit fraud using very complete real-world identities.

The most concerning thing here is that some of the techniques that are being used to create a synthetic identity and output for an individual can be automated; telephony systems can be connected via Twilio [27] to collect legitimate audio samples, video samples can be collected via camera feeds from all the places that an individual visits, their smartphone, their laptop, etc. and upon finding hits within the datasets sourced from data brokers, this can be combined to create a rich representation of an individual with sufficient data points to access even government services on their behalf.

5. OTHER ETHICAL CONSEQUENCES

There will be several long-run implications of these large-scale data breaches that are not immediately obvious.

Credit score ratings will be affected, negatively in some cases where disputed transactions and other historical data are tied to the identity of the user. Advertising firms will be able to further micro-target individuals based on these richer, "unlocked" datasets. Partnerships between large firms, public and private, can enable larger data sharing pools and access to each other's rapid AI advances [10].

Where this becomes especially problematic is in countries where one can be persecuted for particular political leanings, sexual orientation, group membership, etc. and this sort of access to data and advanced pattern recognition only enables that.

Even in progressive countries, hate groups can use this type of information to target individuals and engage in doxing behavior (publishing personal information of an individual into the public domain with malicious intent) that can inflict tremendous damage to the individual. [28] Easier access to AI tools via standard libraries and the falling learning curve to deploy basic techniques heralds a certain "democratization" of these methods.

6. AGGRAVATION BY AI ADVANCES

The emergence of Generative Adversarial Networks (GANs) as a way to produce synthetic data [29] poses an increased level of threat when combined with large-scale data breaches that provide a high amount of training data for these systems. Larger and more representative input into the system leads to more realistic synthetic data coming out of the GANs. They feed a cycle of being able to train AI systems into making more effective predictions and outputs. GANs can also be folded into the techniques mentioned in section 4 to further enhance the virtual avatar of an individual.

Cambridge Analytica [30] is known [31] to have been involved in creating advantages for candidates in senatorial races in the USA, the most recent US Presidential election, elections in the UK and South Africa. They use big data and advanced machine learning techniques to provide a full suite of services to enable highly targeted marketing and campaigning. They also have a commercial offering that is geared towards marketers helping them use "data to change user behavior".

The pace of development of deep learning is only hastening [10] and work is coming out of different research labs across the world (as an example, audio generation and video generation) that can potentially

be combined together to create more powerful tools that have unseen consequences. With strong incentives to monetize such advances, there will be an emergence of many novel use cases that have the potential to dramatically alter our notion of ownership of our identity and how we project ourselves online. In a world where we live increasingly larger portions of our lives online, it is crucial that we be able to manage adverse outcomes.

7. CHALLENGES TO BE ADDRESSED

Given what has been discussed so far, there are quite a few challenges that the research community, industry, policy makers and concerned citizens need to think about to come up with solutions to mitigate adverse outcomes. At the highest level, we need to orient the development and deployment of AI-powered solutions towards enhancing human welfare. But more specifically, to be able to safely unlock the benefits from AI development and to have them be distributed in an equitable, transparent and fair manner will be one of the key challenges facing the community.

There are questions around our ability to "regulate" the development of AI; this idea while noble in its intentions greatly underestimates not only the vastness and distributed nature of research at the moment but also the great difficulty in being able to find such a body that can not only assimilate ongoing research from across the world but also have the right jurisdictional powers that have been agreed upon by such a community to act as their regulators.

The European Union (EU) General Data Protection Regulation (GDPR) [32] serves as a great precedent in the field of data privacy and security on how we can elevate the rights of the users and strengthen the ownership that they can have on their PII. Yet we need to wait till May 2018 when it comes into effect to see how organizations react and how user rights actually get strengthened.

In the context of widely differing standards and approaches used in different industries, one approach to mitigate adverse outcomes is to fold decision-making around best practices into existing regulatory entities within that industry. More so, sharing challenges across research communities, different industries and across geographic borders can hasten the pace at which we discover and develop the right mix of policies that preserve the rights of consumers while not stifling the development of AI. Practitioners can self-impose a "Hippocratic Oath" for AI development as a starting point which while non-binding can at least trigger the right conversations around the distribution of responsibility and share the onus in terms of arriving at solutions to help address some of the issues raised in this paper.

On the data side where organizations today cannot only collect an increasing amount of data on any individual via their online, and slowly their physical, activities as well, they start creating an ability to infer a lot of different facts about an individual. Open datasets that have been de-anonymized fall prey to sophisticated statistical techniques as shown in the case of AOL and Netflix that led to the re-identification of individuals which will now remain in the public domain forever. While it is great to have research efforts that help us identify the shortcomings of de-anonymization and especially public demonstrations of the mosaic effect which help to put the risks in perspective, we need to focus our efforts on being able to create safe ways to advance research efforts, particularly those in the AI community (because they rely heavily on access to large datasets for training purposes).

Some techniques like homomorphic encryption [33] help to address this to a certain degree by leveraging cryptographic and statistical principles to allow computations on encrypted datasets without having to release the underlying data, yet limitations remain around the high cost of computations required to process data in this form. Keeping this in light, there definitely remains dangers of releasing particular snippets of information because it is hard to predict what this can be combined with to unlock other previously existing datasets.

Other challenges remain like the collection of encrypted traffic for future decryption when novel methods might be invented (including the use of quantum computers) further demonstrating the threat from leaked data even if it is in an encrypted format.

8. CONCLUDING REMARKS

The role of the technical community in providing accessible discourse and guiding popular coverage (specifically keeping it within the bounds of practicality) will be crucial as this serves the stakeholders in taking appropriate decisions rather than being swayed by uninformed opinions, especially from a technical feasibility perspective.

It is of paramount importance to educate the public on the existence of such techniques and their effects as highlighted in this paper and the consequences that arise when they are combined together for malevolent purposes. A strong demand from the user base for stricter and more privacy-enhancing regulations will only serve to encourage policy makers and law-making bodies to put together committees to assess the potential impacts and put in place appropriate measures to ensure the wellbeing of their citizens.

ACKNOWLEDGEMENT

The author is grateful to Sydney Swaine-Simon for discussions on the work presented in this paper.

REFERENCES

[1] "'Mosaic Effect' Paints Vivid Pictures of Tech Users' Lives, Felten Tells Privacy Board." Princeton University, The Trustees of Princeton University, www.princeton.edu/news-and-events/news/item/mosaic-effect-paints-vivid-pictures-tech-users-lives-felten-tells-privacy

[2] Williams, Alex. "The Power of Data Exhaust." TechCrunch, TechCrunch, 26 May 2013, https://techcrunch.com/2013/05/26/the-power-of-data-exhaust/

[3] Angela Mitchell, "Artificial Intelligence in Government.", Deloitte UK, 30 June 2017, https://www2.deloitte.com/uk/en/pages/public-sector/articles/artificial-intelligence-in-government.html

[4] Greene, Tristan. "AI Is Analyzing You on Social Media for Market Research." The Next Web, 7 Aug. 2017, https://thenextweb.com/artificial-intelligence/2017/08/03/this-ai-company-leverages-social-media-speech-for-your-marketing-needs/

[5] "How Artificial Intelligence Is Changing Online Shopping." Time, Time, https://time.com/4685420/artificial-intelligence-online-shopping-retail-ai/

[6] Scott Zoldi, Analytics & Optimization, Risk & Compliance. "How to Build Credit Risk Models Using AI and Machine Learning." FICO, 22 May 2017, www.fico.com/en/blogs/analytics-optimization/how-to-build-credit-risk-models-using-ai-and-machine-learning/

[7] Alexander Hellemans, "Why IoT Needs 5G." IEEE Spectrum: Technology, Engineering, and Science News, 20 May 2015, https://spectrum.ieee.org/tech-talk/computing/networks/5g-taking-stock

[8] Unsupervised Feature Learning and Deep Learning Tutorial, https://ufldl.stanford.edu/tutorial/unsupervised/Autoencoders/

[9] "Explainable Artificial Intelligence (XAI)." Defense Advanced Research Projects Agency, www.darpa.mil/program/explainable-artificial-intelligence

[10] "Why AI Development Is Going to Get Even Faster. (Yes, Really!)." Import AI, 4 Apr. 2016, https://jack-clark.net/2016/04/03/why-ai-development-is-going-to-get-even-faster-yes-really/

[11] Edith Ramirez et al, "Data Brokers: A Call for Transparency and Accountability", Federal Trade Commission, pp. 97-100, May 2014.

[12] Mirani, Leo, and Max Nisen. "The Nine Companies That Know More about You than Google or Facebook." Quartz, Quartz, 27 May 2014, https://qz.com/213900/the-nine-companies-that-know-more-about-you-than-google-or-facebook/

[13] Singel, Ryan. "Netflix Spilled Your Brokeback Mountain Secret, Lawsuit Claims." Wired, Conde Nast, 17 Dec. 2009, www.wired.com/2009/12/netflix-privacy-lawsuit/

[14] Michael Barbaro And Tom Zeller Jr. "A Face Is Exposed for AOL Searcher No. 4417749." The New York Times, 8 Aug. 2006, www.nytimes.com/2006/08/09/technology/09aol.html

[15] Stempel, Jonathan, and Jim Finkle. "Yahoo Says All Three Billion Accounts Hacked in 2013 Data Theft." Reuters, Thomson Reuters, 4 Oct. 2017, www.reuters.com/article/us-yahoo-cyber/yahoo-says-all-three-billion-accounts-hacked-in-2013-data-theft-idUSKCN1C82O1

[16] Tzink. "How Rainbow Tables Work." Terry Zink: Security Talk, https://blogs.msdn.microsoft.com/tzink/2012/08/29/how-rainbow-tables-work/

[17] Whittaker, Zack. "AdultFriendFinder Network Hack Exposes 412 Million Accounts." ZDNet, ZDNet, 13 Nov. 2016, www.zdnet.com/article/adultfriendfinder-network-hack-exposes-secrets-of-412-million-users/

[18] Zetter, Kim. "Hackers Finally Post Stolen Ashley Madison Data." Wired, Conde Nast, 18 Aug. 2015, www.wired.com/2015/08/happened-hackers-posted-stolen-ashley-madison-data/

[19] "Pastor Outed on Ashley Madison Commits Suicide." CNNMoney, Cable News Network, https://money.cnn.com/2015/09/08/technology/ashley-madison-suicide/

[20] Ragan, Steve "Raising Awareness Quickly: The EBay Data Breach." CSO Online, 21 May 2014, www.csoonline.com/article/2157782/security-awareness/raising-awareness-quickly-the-ebay-database-compromise.html

[21] "Lyrebird – Create a Digital Copy of Voice." Lyrebird – Create a Digital Copy of Voice, https://lyrebird.ai/

[22] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman, 2017, Synthesizing Obama: Learning Lip Sync from Audio. ACM Trans. Graph. 36, 4, Article 95 (July 2017), 13 pages. DOI: http://dx.doi.org/10.1145/3072959.3073640

[23] Smith, Ms. "Voice Hackers Can Record Your Voice Then Use Morpher to Trick Authentication Systems." CSO Online, CSO, 30 Sept. 2015, www.csoonline.com/article/2988133/security/voice-hackers-can-record-your-voice-then-use-morpher-to-trick-authentication- systems.html

[24] "23andMe Genotypes One Millionth Customer." 23andMe Media Center, https://mediacenter.23andme.com/press-releases/23andme-1million/

[25] "What Can a Hacker Do with Your Genetic Information?" Motherboard, 26 July 2016, https://motherboard.vice.com/en_us/article/gv5w7j/what-can-a-hacker-do-with-your-genetic-information

[26] 1139 – Cloudflare: Cloudflare Reverse Proxies Are Dumping Uninitialized Memory-Project-Zero-Monorail, https://bugs.chromium.org/p/project- zero/issues/detail?id=1139

[27] "Check out All Twilio APIs." Communication APIs for SMS, Voice, Video and Authentication, www.twilio.com/

[28] Ellis, Emma Grey. "Doxing Is a Perilous Form of Justice-Even When It's Outing Nazis." Wired, Conde Nast, 17 Aug. 2017, www.wired.com/story/doxing-charlottesville/

[29] Sanchez, Cassie. "At a Glance: Generative Models & Synthetic Data." Mighty AI, 22 Mar. 2017, https://mty.ai/blog/at-a-glance-generative-models-synthetic-data/

[30] "Cambridge Analytica – Data Drives All That We Do." Cambridge Analytica – Data Drives All That We Do., https://cambridgeanalytica.org/

[31] "How Trump's Campaign Used the New Data-Industrial Complex to Win the Election." USAPP, 29 Nov.2016, https://blogs.lse.ac.uk/usappblog/2016/11/26/how-trumps-campaign-used-the-new-data-industrial-complex-to-win-the-election/

[32] "Home Page of EU GDPR." EU GDPR Portal, www.eugdpr.org/

[33] Gentry, Craig "A Fully Homomorphic Encryption Scheme", Stanford University, 2009.

MACHINE INTELLIGENCE TECHNIQUES FOR NEXT-GENERATION CONTEXT-AWARE WIRELESS NETWORKS

Tadilo Endeshaw Bogale[1,2], Xianbin Wang[1] and Long Bao Le[2]
Western University, London, Canada[1]
Institute National de la Recherche Scientifique (INRS),
Université du Québec, Montréal, Canada[2]
Email: tadilo.bogale@emt.inrs.ca, xianbin.wang@uwo.ca,
and long.le@emt.inrs.ca

Abstract – Next generation wireless networks (i.e., 5G and beyond), which will be extremely dynamic and complex due to the ultra-dense deployment of heterogeneous networks (HetNets), pose many critical challenges for network planning, operation, management and troubleshooting. At the same time, the generation and consumption of wireless data are becoming increasingly distributed with an ongoing paradigm shift from people-centric to machine-oriented communications, making the operation of future wireless networks even more complex. In mitigating the complexity of future network operation, new approaches of intelligently utilizing distributed computational resources with improved context awareness becomes extremely important. In this regard, the emerging fog (edge) computing architecture aiming to distribute computing, storage, control, communication, and networking functions closer to end users, has a great potential for enabling efficient operation of future wireless networks. These promising architectures make the adoption of artificial intelligence (AI) principles, which incorporate learning, reasoning and decision-making mechanisms, natural choices for designing a tightly integrated network. To this end, this article provides a comprehensive survey on the utilization of AI integrating machine learning, data analytics and natural language processing (NLP) techniques for enhancing the efficiency of wireless network operation. In particular, we provide comprehensive discussion on the utilization of these techniques for efficient data acquisition, knowledge discovery, network planning, operation and management of next generation wireless networks. A brief case study utilizing the AI techniques for this network has also been provided.

Keywords – Artificial (machine) intelligence, context-aware wireless, machine learning, ontology

1. INTRODUCTION

The advent of the fifth generation (5G) wireless network and its convergence with vertical applications constitute the foundation of a future connected society which is expected to support 125 billion devices by 2030 (IHS Markit). As these applications and devices are featured by ubiquitous connectivity requirements, future 5G and beyond networks are becoming more complex. Aside from the complexity increase of base stations (BSs) and user equipment (UE), significant challenges arise from the initial network planning to the deployment and situation-dependent operation and management stages.

The network architecture of 5G and beyond will be inevitably heterogeneous and multi-tier with ultra-dense deployment of small cells to achieve the anticipated 1000-fold capacity increase cost-effectively. For instance, the mixed use of planned and centrally controlled macro-BSs and randomly deployed wireless fidelity (WiFi) access points or femto-BSs in the ultra-dense heterogeneous network (HetNet) raises several unexpected operation scenarios, which are not possible to envision at the network design stage. This requires future wireless networks to have self-organizing, configuring and healing capabilities based on the operational condition through the tight coordination among different nodes, tiers and communication layers. These challenges highlight that existing network design strategies, which utilize a fairly simple statistics experience, delivers unacceptable performance (for example, in terms of spectrum and energy efficiency, coverage, delay and cost) [1], [2].

The rapidly growing number of machine-type communication (MTC) devices contributes a considerable portion of the complexity of this ultra-dense network. Many of the future MTC applications supported by 5G and beyond will require the underlying wireless networks to achieve high availability, reliability and security, very short transit times and low latency [3]. Furthermore, in such use cases, uninterrupted and safe operation is often the top priority (for instance, connected vehicles). Taking an MTC application offline for any reason can cause significant business loss or non-tolerable customer experience, and many of the MTC devices are resource-constrained and will not be able to rely solely on their own limited resources to fulfill their processing demands [4].

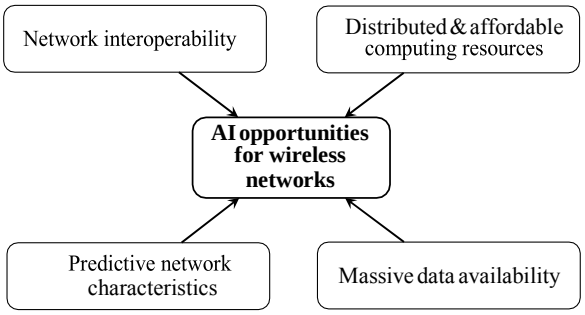


Fig. 1. Favorable conditions for the adoption of machine intelligence techniques in the next generation wireless networks.

As a result, these latency critical applications cannot be moved to the network controller or cloud due to delay, bandwidth, or other constraints. Moreover, the data sets generated from these devices will be extremely diverse and may have large-scale missing (inaccurate) values [5]. In addition, a number of new data hungry MTC immersive use-cases will arise including wearables, virtual realities, intelligent product and support-systems where most of them will use built-in back-end data infrastructure and analytics engine to provide context-aware services. All these necessitate the next generation network (i.e., 5G and beyond) to adopt an intelligent and context-aware approach for network planning, design, analysis, and optimization.

We are in the beginning phase of an intelligent era that has been driven by the rapid evolution of semiconductor industries, computing technologies, and diverse use cases. This is witnessed by the tight integration of networked information systems, sensing and communication devices, data sources, decision making, and cyber-physical infrastructures. The proliferation of tiny wireless sensors and MTC devices, and smart phones also show clear evidence of exceptional processing capability and cost-effectiveness of semiconductor devices. These promising developments facilitate distributed computing resources not only in the cloud but also in the fog and edge nodes. Both fog and edge computing attempt to push the intelligence and processing capabilities down closer to where the data originates.

The edge computing aims to integrate intelligence and processing power capabilities closest to the original data source. The edge node, for example, intelligent programmable automation controllers (PACs), determines which raw input data should be stored locally or sent to the fog (cloud) for further analysis. On the other hand, in the fog computing, all the raw input data will first be converted to the appropriate Internet protocol (such as HTTP) before being sent to the fog nodes. Thus, higher-level data content is processed, stored and sent to the cloud for further analysis in the fog devices (for example, intelligent routers, access points, Internet of things (IoT) gateways). Thus, the edge and fog enabled network allows distributed computing, storage, control, communication and networking functions by reducing the data transmitted and workload of the cloud, latency and system response time especially for applications demanding localized and location-dependent information [6]. Moreover, the node, user, sensor, or MTC device is potentially capable of generating raw data and processed data at different granularity levels, which ultimately helps the network to have a massive amount of data exhibiting a pattern. This will help different nodes to leverage data mining and analytics techniques to predict relevant network metrics such as user mobility, traffic behavior, network load fluctuation, channel variations, and interference levels.

All these opportunities enable efficient and flexible resource allocation and management, protocol stack configuration, and signaling procedure and physical layer optimization, and facilitate existing devices to harness the powers of sensors, edge, fog and cloud-based computing platforms, and data analytics engines [7]-[9]. These also create favorable conditions to engineer a tightly integrated wireless network by adopting the AI principles (see Fig. 1) incorporating learning, reasoning and decision-making mechanisms which are crucial to realize the context-awareness capability. A typical next generation network utilizing the AI principles at different nodes is shown in Fig. 2. To this end, the current paper provides a comprehensive survey on the utilization of AI integrating machine learning, data analytics and natural language processing (NLP) techniques for enhancing the

efficiency of wireless systems. We particularly focus on the utilization of these techniques for efficient wireless data acquisition and knowledge discovery, planning, and operation and management of next generation wireless networks. A brief case study showing the utilization of AI techniques for this network has also been provided.

The paper is organized as follows. In section 2, we discuss data acquisition and knowledge discovery approaches used in AI-enabled wireless networks. Then, a comprehensive discussion on how this knowledge can be used in network planning, operation and management of the next generation wireless network is given in sections 3 and 4. A case study discussing the applications of AI techniques for channel impulse response (CIR) prediction and context-aware data transmission is then provided in section 5. Finally, conclusions are drawn in section 6.

2. DATA ACQUISITION AND KNOWLEDGE DISCOVERY

Efficient data acquisition and knowledge discovery is one of the requirements of future wireless networks as it helps to realize situation aware and optimized decisions as shown in Fig. 3. The gathered data may need to be processed efficiently to extract relevant knowledge. Furthermore, as the available data may contain a large amount of erroneous (missing) values, a robust knowledge discovery may need to be devised [5].

A. Data acquisition

AI-based tools relying on machine learning for the input data mining and knowledge model extraction at different levels could be applied [10]. This includes the cell level, cell cluster level and user level. In general, one can collect data from three main sources; network, user, and external devices. The network data characterizes the network behavior including outage and usage statistics of services or nodes, and the load of a cell. The user data could comprise user subscription information and user device type. And, the external data contains user specific information obtained from different sources such as sensors and channel measurements [11].

One way of collecting wireless data is by employing content caching where the idea is to store popular content at the network edge (at BSs, devices, or other intermediate locations). In this regard, one can enable the proactive cache type if the traffic learning algorithm predicts that the same content will be requested in the near future [12], [13].

Moreover, since different users may request the same content with different qualities, each edge node may need to cache the same content in different granularity (for example, caching video data with different resolutions). This further requires the edge device to apply coded (adaptive) caching techniques based on the quality of service (QoS) requirement of the requester [14]. Coded caching also enables devices to create multicasting opportunities for certain content via coded multicast transmissions [15].

In some cases, a given edge (fog) may collect date from more than one source with different connectivity criteria [16]. In a fog-enabled wireless network this is facilitated by employing IoT devices which leverage a multitude of radio-access technologies such as wireless local area networks (WLAN) and cellular networks. In this regard, context-aware data collection from multiple sources probably in a compressed format by employing appropriate dimensionality reduction techniques under imperfect statistical knowledge of the data while simultaneously optimizing multiple objective functions such as delay and transmission power can be enabled [13].

B. Knowledge discovery

Efficient knowledge discovery is critical for optimized operation and management of the network. The network may need to use a novel learning technique such as deep learning to extract the hidden contextual information of the network which is crucial for knowledge base (KB) creation. In general, context is related to any information used to characterize the situation of an entity, including surrounding location, identity, preferences, and activities. Context may affect the operation and management procedures of complex systems at various levels, from the physical device level to the communication level, up to the application level [17]. For instance, uncovering the relation between the device and network information (user location, velocity, battery level, and other medium access control (MAC) and higher layer aspects) would permit adaptive communication and processing capabilities based on the changes in the environment and application [2], [17].

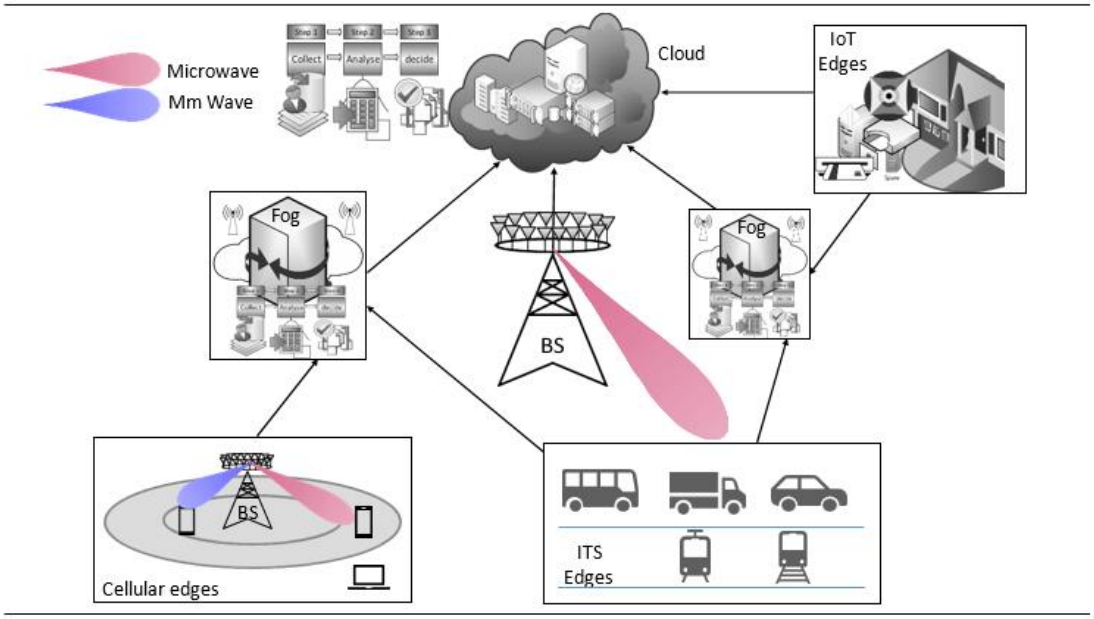


Fig. 2. Typical next generation network adopting AI principles with learning, reasoning and decision making.

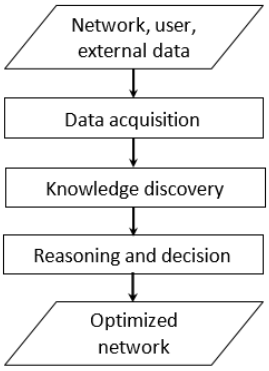


Fig. 3. Optimized network design with AI techniques.

Analyzing wireless data contextually (semantically) facilitates wireless operators to optimize their network traffic. To realize semantic-aware traffic optimization, however, the network may need to understand the content of the signal. One way of understanding the information content of a data is by creating semantic-aware ontology using predefined vocabulary of terms and concepts [18]. The ontology specification can provide an expressive language with many logical constructs to define classes, properties and their relationships. In this regard, the authors of [18] propose a semantic open data model for sensor data called *MyOntoSens* and write using ontology web language 2 description logic language. The prooposed KB has been implemented using *protégé* and pre-validated with pellet reasoner. In a similar context, an ontology for wireless sensor networks (WSNs) dedicated to the description of sensor features and observation has been presented in [19], [20].

Understanding the context also helps to produce context-aware compressed (summary) information which will utilize less radio resources for transmission. For instance, if a BS would like to transmit text information to a user, the BS can transmit only its contextual coded data. The user will then extract the desired content just from the context by utilizing appropriate decoder and big-data analytics techniques such as NLP. As context differs from the user's world knowledge about the content, the coding and decoding technique may vary among users [17]. In general two types of content summarizing approaches are commonly adopted; abstractive and extractive. The extractive approach uses only the relevant content from the original information, whereas the abstractive approach may use new words (expressions or contents) as part of the summary information [21], [22]. Although most of the existing methods can extract useful information for the summary, they are very far from generating a human understandable summary. One of the main reasons is the loose associations and unordered information distribution which make it hard to extract syntactically correct and semantically coherent information from the summary. In fact, modeling the coherence of information summary is one of the active areas of research [21].

3. NETWORK PLANNING

One of the most critical aspects determining the performance of a wireless network is the initial planning. This includes infrastructure (node deployment), frequency, number of parameters and

their configuration setting procedures, and energy consumption to run the network in both idle (no data communication takes place between the network and a user) and active (data communication takes place between the network and a user) conditions. A well-planned network may need to guarantee satisfactory user QoS (in terms of data rate, reliability and latency) and the network operator requirements (in terms of cost). The machine learning technique can be used for planning different parts of the network by utilizing the network and user data.

A. Node deployment, Energy consumption and RF planning

The future generation wireless networks will be extremely dense and heterogeneous, likely equipped with moving and flying BSs, and featured by continually varying network conditions [23]. This fact makes the existing network planning techniques, which are mainly static and designed from expensive field tests, not suitable for future wireless networks [2], [24]. The utilization of AI techniques for network planning has recently received an interest in the research community. For instance, a machine learning technique is suggested for content popularity and request distribution predictions of a cache where its design in the network considers several factors including cache placement and update strategy which are determined by utilizing the users' content request distribution and frequency, and mobility pattern [25].

In [26] an AI-based system which leverages graph theory based problem formulations for the fiber to home network is proposed to automate the planning process. To solve the problems, mixed integer linear programming (MILP), ant colony optimization (ACO) and genetic algorithm (GA) have been applied. The authors of [10] employ the principles of AI for radio access network (RAN) planning which includes new cell, radio frequency (RF) and spectrum of the 5G wireless network. The analysis is performed by processing input data from multiple sources, through learning based classification, prediction and clustering models, and extracting relevant knowledge to drive the decisions made by the 5G network.

Wireless networks contribute an increasing share in the energy consumption of the information communication technology (ICT) infrastructure. Over 80% of a wireless network power consumption is used by the RANs, especially at the BS, since the present BS deployment is designed on the basis of peak traffic loads and generally stays active irrespective of the huge variations in traffic load [27]. This makes the current energy planning inefficient for the future smart cities aiming to realize green communication. To enable energy-efficient wireless networks, different AI techniques have been suggested. For instance, the authors of [28] propose a method to realize the efficient use of electricity by autonomously controlling network equipment such as servers, air-conditioners in an integrated manner. Furthermore, the authors of [29] suggest predictive models, including neural network and Markov decisions, for the energy consumption of IoT in smart cities. Along this line, a BS-switching solution for traffic aware greener cellular networks using AI techniques has also been discussed in [27].

B. Configuration parameter and service planning

The number of configurable parameters in cellular networks fairly increases when moving from one generation to the next. For instance, in typical 3G and 4G nodes these parameters are around 1000 and 1500, respectively. It is anticipated that this trend will continue and the recently suggested 5G network node will likely have 2000 or more parameters. In addition, unlike the current and previous generation networks which provide static services, the next generation network may need to support continuously evolving new services and use cases, and establish sufficient network resource and provisioning mechanisms while ensuring agility and robustness. These necessitate the next generation network to understand parameter variations, learn uncertainties, configure network parameters, forecast immediate and future challenges, and provide timely solutions by interacting with the environment [27]. In this direction, the utilization of big data analytics has been discussed for protocol stack configuration, signaling procedure and physical layer procedure optimizations in [9].

Future smart cities require well-planned wired and wireless networks with ubiquitous broadband connectivity, and flexible, real-time and distributed data processing capability. Although most modern cities have multiple cellular networks that provide adequate coverage and data processing capability, these networks often have limited capacity and peak bandwidths and fail to meet the real-time constraint of different emerging tactile applications. These make the realization of advanced delay critical municipal services envisioned in a smart city (e.g., real-time surveillance, public safety, on-time advisories, and smart buildings) challenging [1]. One way of addressing this challenge would be by deploying an AI integrated fog-based wireless architecture which allows data processing of the network using a number of distributed nodes. This will help analyze network

status, detect anticipated faults and plan new node deployment using AI techniques [1].

4. NETWORK OPERATION AND MANAGEMENT

Energy and spectrum efficiency, latency, reliability, and security are the key parameters that are taken into account during the network operation stage. And properly optimizing these parameters usually yields satisfactory performance for both the service providers and end users. In addition, these optimization parameters usually require simple and real-time learning and decision-making algorithms.

A. Resource allocation and management

Different AI techniques have been proposed for resource allocation, management and optimization of wireless networks such as cellular, wearable, WSN, and body area network (BAN) [24]. In [30], the potential of AI in solving the channel allocation problem in wireless communication is considered. It is demonstrated that the AI-based approach has shown better performance than those of randomized-based heuristic and genetic algorithms (GAs). In [31], radio access technology (RAT) selection utilizing the Hopfield neural networks as a decision-making tool while leveraging the ability of AI reasoning and the use of multi-parameter decisions by exploiting the options of IEEE 802.21 protocol is proposed. Machine learning based techniques including supervised, unsupervised, and reinforcement learning techniques, have been exploited to manage the packet routing in many different network scenarios [32]. Specifically, in [33], [34], a deep-learning approach for shortest traffic route identification to reduce network congestion is presented. A deep-learning technique aiming to shift the computing needs from rule-based route computation to machine learning based route estimation for high throughput packet processing is proposed in [32]. Along this line, a fog computing based radio-access network which exploits the advantage of local radio signal processing, cooperative radio resource management, and distributed storage capability of fog has been suggested to decrease the load on the front haul and avoid large-scale radio signal processing in the centralized baseband controllers [1].

The utilization of unlicensed spectrums as a complement to licensed ones receives an interest to offload network traffic through the carrier aggregation framework, while critical control signaling, mobility, voice and control data will always be transmitted on the licensed bands. In this respect, the authors of [35] propose a hopfield neural network scheme for multi-radio packet scheduling. The problem of resource allocation with uplink-downlink decoupling in a long term evolution-unlicensed (LTE-U) system has been investigated in [36] for which the authors propose a decentralized scheme based on neural networks. The authors in [37] propose a distributed approach based on Q-learning for the problem of channel selection in an LTE-U system. Furthermore, in a multi-RAT scenario, machine learning techniques can allow the smart use of different RATs wherein a BS can learn when to transmit on each type of frequency band based on the network conditions. For instance, one can apply machine learning to predict the availability of a line of sight (LoS) link, by considering the users' mobility pattern and antenna tilt, thus allowing the transmission over the millimeter wave band.

B. Security and privacy protection

The inherent shared nature of radio propagation environment makes wireless transmissions vulnerable to malicious attacks, including eavesdropping and jamming. For this reason, security and privacy protection are fundamental concerns of today's wireless communication systems. Wireless networks generally adopt separate security levels at different layers of the communication protocol stack. Furthermore, different applications usually require different encryption methods [42]. The utilization of AI techniques for wireless security has received significant interest.

TABLE 1. Main issues in AI-enabled wireless network

Data acquisition and knowledge discovery	Ref.
• Context-aware data acquisition from single/multiple sources	[13]
• Coded (adaptive) caching	[14], [15]
• Semantic-aware Ontology (KB) creation from network data	[18]-[20]
• Robust knowledge discovery from erroneous (missing) data	[5]
Network planning	
• Node deployment and radio frequency allocation	[10], [26]
• Caching and computing placement and content update	[38]
• Energy consumption modeling and prediction (idle/active)	[27], [28]
• Parameter and service configuration procedure	[1], [9]
Network operation and management	
• Resource allocation: RAT and channel selection, packet routing, distributed storage and processing, multi-RAT packet scheduling	[30], [31] [1], [32] [35]
• Security: Spoofing attack and intrusion detection	[39]-[41]
• Latency: Context-aware edge computing and scheduling	[5], [38]

In [39], a spoofing attack detection scheme using a random key distribution based artificial immune system (AIS) has been proposed. In a similar way, an approach based on GA and AIS, called GAAIS, for dynamic intrusion detection in mobile ad-hoc networks (MANETs) is suggested in [41]. In [40], advanced detection of intrusions on sensor networks (ADIOS) based intrusion detection and prevention system is

developed. The ADIOS is designed to mitigate denial-of-service attacks in wireless sensor networks by capturing and analyzing network events using AI and an expert system developed using the C language integrated production system tool. In a similar work, the authors of [43] propose an AI-based scheme to secure the communication protocol of connected vehicles.

C. Latency optimization for tactile applications

Next generation wireless networks are featured by several mission critical (tactile) applications such as lane switching in automated vehicles. For vehicular networks, different levels of automation has been defined by the USA department of transportation (DOT) ranging from a simple driver assistance (level 1) to the full automation mode (level 5). For this application, one can apply different message representations including warning alarm, picture and audio information to request intervening. In fact, recently it has been demonstrated via an experiment that the use of natural language generation techniques from imprecise data improves human decision-making accuracy. Such a linguistic description of data could be designed by modeling vague expressions such as *small* and *large*, which are norms in daily life conversation, using fuzzy logic theory [44]. All these facilitate the utilization of predictive machine learning as in [45], [46].

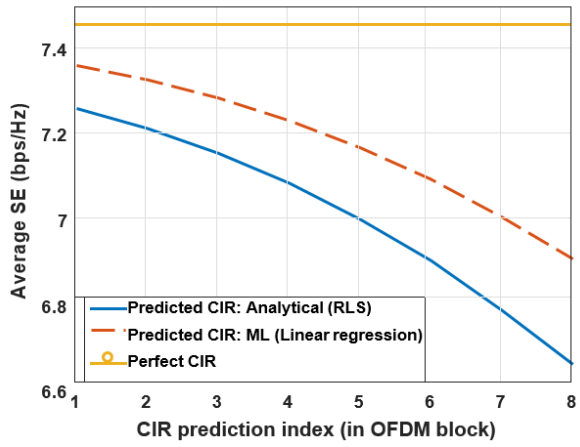


Fig. 4. Comparison of analytical and machine learning (ML) approaches in terms of achieved average SE for different future OFDM block index CIR.

From the computing side, edge devices can be used for effective low-latency computations, using the emerging paradigm of mobile edge computing. However, optimizing mobile edge computing faces many challenges such as computing placement, computational resource allocation, computing task assignment, end-to-end latency, and energy consumption. In this respect, a machine learning technique can be used to address these challenges by utilizing historical data. Predicting computational requirements enable the network devices to schedule the computational resources in advance to minimize the global latency. In this respect, the authors of [38] propose a cross-system learning framework in order to optimize the long-term performance of multi-mode BSs, by steering delay-tolerant traffic towards Wi-Fi. Furthermore, in a fog-enabled wireless system, latency can be addressed by exploiting different levels of awareness at each edge network. In fact, a number of learning techniques can be applied to achieve this awareness including incremental, divide and conquer, parallel and hierarchical [5]. A brief summary of different issues in the AI-enabled wireless network is presented in Table 1.

5. DESIGN CASE STUDIES

This section discusses typical design case studies in which the AI techniques can be applied for the context-aware wireless network.

A. Machine learning for CIR prediction

This study demonstrates the utilization of machine learning tools for optimizing wireless system resources. In this respect, we select the wireless CIR prediction as the design objective. To solve this design objective, the first possibility could be to apply different analytical CIR prediction techniques (for example, the recursive least square (RLS) prediction proposed in [45]). The second possibility could be to predict future CIRs by leveraging the past experience. The former possibility is very expensive particularly when real-time prediction is needed. Furthermore, in most cases, the analytical prediction approach may fail whenever there is a modeling error or uncertainty. The latter possibility, however, is simple as it employs past experience and applies standard vector multiplication and addition operations [47]. This simulation compares the performances of RLS and machine learning prediction approaches. For the machine learning, we employ the well-known multivariate linear regression.

For the comparison, we consider an orthogonal frequency domain multiplexing (OFDM) transmission scheme where a BS equipped with N antennas is serving a single antenna IoT device. The CIR is modeled by considering a typical scenario of the IEEE 802.11 standard with channel correlation both spatially and temporarily. The spatial channel covariance matrix is modeled by considering the uniform linear array

(ULA) structure, and the temporal channel correlation is designed by the well-known Jake's model [48].

The number of multipath taps $L = 4$, fast Fourier transform (FFT) size $M = 64$, $N = 8$, OFDM symbol period $T_s = 166$ µs, RLS window size $S_b = 8$, forward prediction window size $S_f = 8$, carrier frequency 5.6 GHz, and mobility speed of the IoT device 30 km/hr. The signal to noise ratio (SNR) for each subcarrier is set to 10 dB. With these settings, Fig. 4 shows the average spectrum efficiency (SE) obtained by the RLS and machine learning approaches for subcarrier $s = 4$. In both cases, the achieved SE decreases as the future OFDM-block index increases. This is expected since the number of unknown CIR coefficients increase as the future block index increases leading to a degraded CIR prediction quality. However, for a fixed future prediction index, the machine learning approach yields better performance than the RLS one[1]

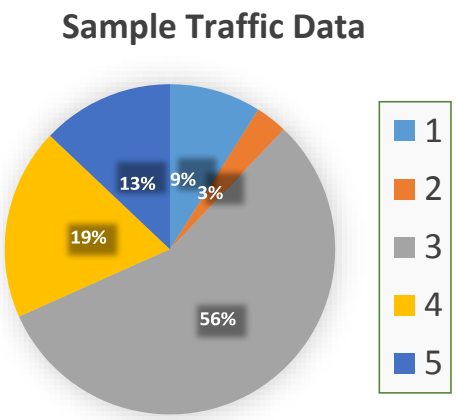


Fig. 5. Data traffic of a sample abstract information: In this figure, (sets) 1 to 5 denote, Background, Objective, Method and Result, Conclusion, and Related and Future Works, respectively.

B. Context-aware data transmission using NLP techniques

Context (semantic) aware information transmission is crucial in the future generation network. To validate this, we employ abstract texts from scientific articles [49]. According to this paper, each scientific abstract text consists of different types of information including research background, methodology, main results etc. Fig. 5 shows the expert annotated data size for different types for 2000 biomedical article abstracts. As can be seen from this figure, different information types use different portions of the overall data set. And for a given user, one can transmit the desired information according to the context. For instance, for a user who is interested in the basics of the article, transmitting the background information could be sufficient which accounts for only 9% of the total traffic. This shows that semantically-enabled data transmission will reduce the network traffic while simultaneously maintaining the desired QoS experience of users.

Such a transmission, however, is realized when sentences of similar types are clustered correctly for each abstract. In scientific papers, the location and part-of-speech voice of a sentence are crucial features to identify its class set [49]. We have employed these features with the commonly used data clustering algorithms (i.e., K-means and Agglomerative) and present the accuracy achieved by these algorithms for each type as shown in Table 2. As can be seen from this table, different clustering algorithms yield different accuracy. One can also notice from this table that significant research work may need to be undertaken to reach the ideal performance.

Table 2. Accuracy of different clustering methods

Clustering Method	Set 1	Set 2	Set 3	Set 4	Set 5
K-Means	0.34	0.17	0.35	0.31	0.16
Agglomerative	0.21	0.18	0.38	0.30	0.15

6. CONCLUSIONS

Next generation wireless networks, which will be more dynamic, complex with dense deployment of BSs of different types and access technologies, pose many design challenges for network planning, management and troubleshooting procedures. Nevertheless, wireless data can be generated from different sources including networked information systems, and sensing and communication devices. Furthermore, the emerging fog computing architecture aiming for distributed computing, storage, control, communication, and networking functions closer to end users contribute to the efficient realization of wireless systems. This article provides a comprehensive survey on the utilization of AI integrating machine learning, data analytics and NLP techniques for enhancing the efficiency of wireless networks. We have given a comprehensive discussion on the utilization of these techniques for efficient wireless data acquisition and knowledge discovery, planning, operation and management of next generation wireless networks. A brief case study showing the utilization of AI techniques has also been provided.

[1] Note that similar average performance is observed for other subcarriers.

REFERENCES

[1] A. Munir, P. Kansakar, and S.U. Khan, "IFCIoT: integrated Fog Cloud IoT: A novel architectural paradigm for the future internet of things," IEEE Consumer Electronics Magazine, vol. 6, no. 3, pp. 74-82, July 2017.

[2] T.E. Bogale and L. Le, "Massive MIMO and mmWave for 5G wireless HetNet: Potential benefits and challenges," IEEE Vehic. Techno. Mag., vol. 11, no. 1, pp. 64-75, Mar. 2016.

[3] M. Weiner, M. Jorgovanovic, A. Sahai, and B. Nikoli, "Design of a low-latency, high-reliability wireless communication system for control applications," in Proc. IEEE Int. Conf. Commun. (ICC), 2014, pp. 3829-3835.

[4] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," IEEE Internet of Things, vol. 3, no. 6, pp. 854-864, Dec. 2016.

[5] X. Wang and Y. He, "Learning from uncertainty for big data: Future analytical challenges and strategies," IEEE Systems, Man, and Cybernetics Magazine, vol. 2, no. 2, pp. 26-31, April 2016.

[6] Ruilong Deng, Rongxing Lu, Chengzhe Lai, Tom H Luan, and Hao Liang, "Optimal workload allocation in fog-cloud computing toward balanced delay and power consumption," IEEE Internet of Things Journal, vol. 3, no. 6, pp. 1171-1181, 2016.

[7] M. Chiang, S. Ha, C. L. I, F. Risso, and T. Zhang, "Clarifying Fog computing and networking: 10 questions and answers," IEEE Communications Magazine, vol. 55, no. 4, pp. 18-20, April 2017.

[8] Y. Sun, H. Song, A. J. Jara, and R. Bie, "Internet of things and big data analytics for smart and connected communities," IEEE Access, vol. 4, pp. 766-773, 2016.

[9] S. Han, C. L. I, G. Li, S. Wang, and Q. Sun, "Big data enabled mobile network design for 5G and beyond," IEEE Communications Magazine, vol. PP, no. 99, pp. 2-9, 2017.

[10] J. Prez-Romero, O. Sallent, R. Ferrs, and R. Agust, "Knowledge-based 5G radio access network planning and optimization," in 2016 International Symposium on Wireless Communication Systems (ISWCS), Sept 2016, pp. 359-365.

[11] J. Prez-Romero, O. Sallent, R. Ferrs, and R. Agust, "Artificial intelligence-based 5G network capacity planning and operation," in 2015 International Symposium on Wireless Communication Systems (ISWCS), Aug 2015, pp. 246-250.

[12] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," IEEE Communications Magazine, vol. 52, no. 8, pp. 82-89, Aug 2014.

[13] E. Bastug, M. Bennis, M. Medard, and M. Debbah, "Toward interconnected virtual reality: Opportunities, challenges, and enablers," IEEE Communications Magazine, vol. 55, no. 6, pp. 110-117, 2017.

[14] G. Paschos, E. Bastug, I. Land, G. Caire, and M. Debbah, "Wireless caching: Technical misconceptions and business barriers," IEEE Communications Magazine, vol. 54, no. 8, pp. 16-22, August 2016.

[15] Y. Fadlallah, A. M. Tulino, D. Barone, G. Vettigli, J. Llorca, and J.M. Gorce, "Coding for caching in 5G networks," IEEE Commun. Magazine, vol. 55, no. 2, pp. 106-113, 2017.

[16] D. Ohmann, A. Awada, I. Viering, M. Simsek, and G.P. Fettweis, "Achieving high availability in wireless networks by inter-frequency multi-connectivity," in 2016 IEEE Inter-national Conference on Communications (ICC), May 2016, pp. 1-7.

[17] F. Chiti, R. Fantacci, M. Loreti, and R. Pugliese, "Context- aware wireless mobile autonomic computing and communications: Research trends and emerging applications," IEEE Wireless Commun., vol. 23, no. 2, pp. 86-92, Apr. 2016.

[18] L. Nachabe, M. Girod-Genet, and B. El Hassan, "Unified data model for wireless sensor network," IEEE Sensors Journal, vol. 15, no. 7, pp. 3657-3667, July 2015.

[19] R. Bendadouche, C. Roussey, G. De Sousa, J. Chanet, and K-M. Hou, "Extension of the semantic sensor network ontology for wireless sensor networks: The stimulus-WSNnode-communication pattern," in Proc. 5th Int. Conf. Semantic Sensor Networks – Vol. 904, 2012, pp. 49-64.

[20] M. Compton, A. A. Henson, L. Lefort, H. Neuhaus, and A. P. Sheth, "A survey of the semantic specification of sensors," in Proc. CEUR Workshop, 2009, pp. 17-32.

[21] E. Cambria and B. White, "Jumping NLP curves: A review of natural language processing research," IEEE Comp. Intelligence Mag., pp. 48-57, May 2014.

[22] T. Hirao, M. Nishino, Y. Yoshida, J. Suzuki, N. Yasuda, and M. Nagata, "Summarizing a document by trimming the discourse tree," IEEE/ACM Trans. Audio, Speech, and Lang. Process, vol. 23, no. 11, pp. 2081-2092, Nov. 2015.

[23] M. Alzenad, A. El-Keyi, F. Lagum, and H. Yanikomeroglu, "3-D placement of an unmanned aerial vehicle base station (UAV-BS) for energy-efficient maximal coverage," IEEE Wireless Communications Letters, vol. 6, no. 4, pp. 434-437, Aug 2017.

[24] Z. Fadlullah, F. Tang, B. Mao, N. Kato, O. Akashi, T. Inoue, and K. Mizutani, "State-of-the-art deep learning: Evolving machine intelligence toward tomorrow's intelligent network traffic control systems," IEEE Communications Surveys Tutorials, vol. PP, no. 99, pp. 1-1, 2017.

[25] M. Chen, M. Mozaffari, W. Saad, C. Yin, M. Debbah, and C. S. Hong, "Caching in the sky: Proactive deployment of cache-enabled unmanned aerial vehicles for optimized quality-of-experience," IEEE JSAC, vol. 35, no. 5, pp. 1046-1061, 2017.

[26] K. F. Poon, A. Chu, and A. Ouali, "An AI-based system for telecommunication network planning," in 2012 IEEE International Conference on Industrial Engineering and Engineering Management, Dec 2012, pp. 874-878.

[27] R. Li, Z. Zhao, X. Zhou, G. Ding, Y. Chen, Z. Wang, and H. Zhang, "Intelligent 5G: When cellular networks meet artificial intelligence," IEEE Wireless Communications, vol. PP, no. 99, pp. 2-10, 2017.

[28] K. Awahara, S. Izumi, T. Abe, and T. Suganuma, "Au- tonomous control method using AI planning for energy- efficient network systems," in 2013 Eighth International Conference on Broadband and Wireless Computing, Communication and Applications, Oct 2013, pp. 628-633.

[29] W. Ejaz, M. Naeem, A. Shahid, A. Anpalagan, and M. Jo, "Efficient energy management for the internet of things in smart cities," IEEE Communications Magazine, vol. 55, no. 1, pp. 84-91, January 2017.

[30] S. I. Suliman, G. Kendall, and I. Musirin, "Artificial immune algorithm in solving the channel assignment task," in 2014 IEEE International Conference on Control System, Computing and Engineering (ICCSCE 2014), Nov 2014, pp. 153-158.

[31] V. Rakovic and L. Gavrilovska, "Novel RAT selection mech- anism based on Hopfield neural networks," in International Congress on Ultra Modern Telecommunications and Control Systems, Oct 2010, pp. 210-217.

[32] B. Mao, Z. M. Fadlullah, F. Tang, N. Kato, O. Akashi, T. Inoue, and K. Mizutani, "Routing or computing? the paradigm shift towards intelligent computer network packet transmission based on deep learning," IEEE Transactions on Computers, vol. PP, no. 99, pp. 1-1, 2017.

[33] N. Kato, Z. M. Fadlullah, B. Mao, F. Tang, O. Akashi, T. Inoue, and K. Mizutani, "The deep learning vision for heterogeneous network traffic control: Proposal, challenges, and future perspective," IEEE wireless commun. Mag, vol. 24, no. 3, pp. 146-153, 2017.

[34] F. Tang, B. Mao, Z. M. Fadlullah, N. Kato, O. Akashi, T. Inoue, and K. Mizutani, "On removing routing protocol from future wireless networks: A real-time deep learning approach for intelligent traffic control," IEEE Wireless Commun. Mag, 2017.

[35] Y. Cui, Y. Xu, R. Xu, and X. Sha, "A multi-radio packet scheduling algorithm for real-time traffic in a heterogeneous wireless network environment," Information Technology Journal, vol. 10, pp. 182-188, Oct. 2010.

[36] M. Chen, W. Saad, C. Yin, and M. Debbah, "Echo state networks for proactive caching in cloud-based radio access networks with mobile users," IEEE Trans. Wireless Commun., vol. 16, no. 6, pp. 3520 – 3535, Jun. 2017.

[37] J. Perez-Romero, O. Sallent, R. Ferrus, and R. Agusti, "A ro- bustness analysis of learning-based coexistence mechanisms for LTE-U operation in non-stationary conditions," in Proc. of IEEE Vehicular Technology Conference (VTC Fall), Sep. 2015.

[38] M. Bennis, M. Simsek, A. Czylwik, W. Saad, W. Valentin, and M. Debbah, "When cellular meets WiFi in wireless small cell networks," IEEE Commun. Magazine, vol. 51, no. 6, pp. 44-50, 2013.

[39] E. S. Kumar, S. M. Kusuma, and B. P. V. Kumar, "A random key distribution based artificial immune system for security in clustered wireless sensor networks," in Electrical, Electronics and Computer Science (SCEECS), 2014 IEEE Students' Conference on, March 2014, pp. 1-7.

[40] V. F. Taylor and D. T. Fokum, "Securing wireless sensor networks from denial-of-service attacks using artificial intelligence and the clips expert system tool," in 2013 Proceedings of IEEE Southeastcon, April 2013, pp. 1-6.

[41] F. Barani, "A hybrid approach for dynamic intrusion detection in ad hoc networks using genetic algorithm and artificial immune system," in 2014 Iranian Conference on Intelligent Systems (ICIS), Feb 2014, pp. 1-6.

[42] Y. Zou, J. Zhu, X. Wang, and L. Hanzo, "A survey on wireless security: Technical challenges, recent advances, and future trends," Proceedings of IEEE, vol. 104, no. 9, pp. 1727-1765, Sept. 2016.

[43] P. Sharma, H. Liu, H. Wang, and S. Zhang, "Securing wireless communications of connected vehicles with artificial intelligence," in 2017 IEEE International Symposium on Technologies for Homeland Security (HST), April 2017, pp. 1-7.

[44] D. Gkatzia, O. Lemon, and V. Rieser, "Data-to-text generation improves decision-making under uncertainty," IEEE Computational Intelligence Magazine, vol. 12, no. 3, pp. 10-17, Aug 2017.

[45] T. E. Bogale, X. Wang, and L. B. Le, "Adaptive channel prediction, beamforming and scheduling design for 5G V2I network," in Proc. IEEE Veh. Technol. Conf (VTC-Fall)., Sep. 2017.

[46] K. Zhang, Y. Mao, S. Leng, Y. He, and Y. ZHANG, "Mobile- edge computing for vehicular networks: A promising network paradigm with predictive off-loading," IEEE Vehicular Technology Magazine, vol. 12, no. 2, pp. 36-44, June 2017.

[47] T. E. Bogale, X. Wang, and L. B. Le, "Adaptive channel prediction, beamforming and feedback design for 5G V2I network," in IEEE (In preparation for submission), 2017.

[48] D. Schafhuber and G. Matz, "MMSE and adaptive prediction of time-varying channels for OFDM systems," IEEE Tran. Wireless Commun., vol. 4, no. 2, pp. 593-602, March 2005.

[49] Y. Guo, A. Korhonen, I. Silins, and U. Stenius, "A weakly-supervised approach to argumentative zoning of scientific documents," in Proc. of Conf. on Empir. Methods in Nat. Lang. Process., 2011, pp. 273-283.

NEW TECHNOLOGY BRINGS NEW OPPORTUNITY FOR TELECOMMUNICATION CARRIERS: ARTIFICIAL INTELLIGENT APPLICATIONS AND PRACTICES IN TELECOM OPERATORS

Wei LIANG[1], Mingjun SUN[2], Baohong HE[2], Mingchuan YANG[1], Xiaoou LIU[1], Bohuan ZHANG[1] and Yuntao WANG[2]
[1]China Telecom, CHINA
[2]CAICT, CHINA

Abstract– In the era of "computational intelligence, perceptional intelligence and cognitive intelligence" as the main direction for the future, telecom operators are on their way to building their own artificial intelligence (AI) ecosystem. In terms of developing AI technology, telecom operators have unique resources and technology advantages: big data resources, superior computing power, lots of investment in AI algorithmic research, broad government and enterprise customer resources. By making full use of these strengths, they have carried out a series of effective practices in the various field and achieved constructive results.
This report will be arranged as follows. In the first part the history and the development status of AI has been introduced, as well as the Chinese powerful policy which was released to support its development. In the second part, the unique advantages for operators to develop AI have been introduced, whilst in the meantime, the AI development idea for telecom operators has been provided. In the third part, in order to be more persuasive, the practice of AI by telecom operators in multiple fields to satisfy internal requirements and meet customer needs, has been described. Finally, based on the current development trends of AI, its future prospects are made by this report. Undoubtedly, in the future, operators will further use their advantages to explore more AI development opportunities.

Keywords –Artificial intelligence, best practice, development idea, telecom operators, unique advantages

1. INTRODUCTION

Artificial intelligence (AI) continues to develop well since it was raised in the US Dartmouth meeting in 1956 [1]. 2017 is its sixtieth anniversary. After more than half a century's development, AI has reached a relatively sophisticated state in terms of both hardware and software, where the potential of industrial application is stimulated. Since 2013, with the rapid development of its related technology such as deep learning and knowledge mapping, the trend of AI technology industrialization has embarked on a fast track. Many giant companies have made remarkable achievements in medical, education as well as Fintech areas. Here are some examples: IBM Watson could diagnose a rare disease within 10 minutes with the help of AI technology, and the accuracy of lung cancer treatments recommendations could be up to 90%. Now AI could make the accuracy of intelligent medical diagnosis comparable to a doctor with years of experience. By proving AI technology can be "trained" to accurately analyze and evaluate test answers, ETS has successfully reviewed the analytical writing test of GRE and SAT, which freed teachers from repetitive work. In the Fintech domain, by taking advantage of AI technology, Google utilized face recognition technology on Android Pay, and facts proved that the security and usability of payment services improved significantly.

At present, the ecological pattern of AI has been established gradually. In the following years, the specialized intelligent application will be the main potential area for the future development of AI. No matter if it is a specialized or generalized application, the enterprise layout of AI will focus on the three basic levels of data and computing layer (base), algorithm layer (technology) and application layer. The Chinese government places a high importance on AI and has released a series of policies to support its development. In May 2016, four Ministries jointly issued "'Internet plus' AI action plans in the next three years". Recently, AI was added in the report on the work of the government at two sessions. On July 20 2017 the state council issued "A new generation of AI development program", which noted AI as one of

the most significant national development strategies. Taking advantage of the above favorable policies, Chinese high-tech industries are devoting to develop AI-related technology and build the AI ecosystem. Telecom operators are no exception. For instance, in 2016, China Telecom proposed the transformation 3.0 strategy in 2016 to be the leading integrated intelligent information service provider. Through the way to AI, they will positively respond to the national development planning in the field of AI, closely keep pace with the emerging forward-looking technology, both domestically and abroad, and dedicated to providing the AI infrastructure. Furthermore, telecom operators are also seeking the entry point for cooperation in the way of the intelligent network, smart operation and business ecology, and to cooperate with the potential industry, push forward the research on AI core technology and promote the development of relevant intelligence applications.

2. THE UNIQUE ADVANTAGES FOR OPERATORS TO DEVELOP AI

In the tide of AI research and development, the operators are also searching for ways to transform themselves from “traditional operator” to “intelligent operator”.

Most of the well-known operators of the world are carrying out AI-related work: 1）France Telecom Orange and Deutsche Telekom AG recently announced that the voice assistant platform Djingo which was based on AI, is being developed. It was expected to compete with the existing voice platforms such as Apple’s Siri and Amazon’s Alexa in the market. 2）Vodafone， the giant European operator also stated that it will launch a customer service robot called "TOBi" to help handle customer service issues and enhance the user experience. 3）Japanese telecom operator NTT has also launched its AI platform. Its main research interests include assistant – AI, echocardiography – AI, environmental AI and network AI.

From the telecom operators’ point of view, there are four key advantages to develop AI: big data resource, superior computing power, lots of investment in AI algorithmic research and a broad customer base. 1) In terms of a big data resource take China Telecom as an example, which has 816,000 mobile base stations and 425 data centers all over China. The mobile subscriber scale has reached 212 million and the broadband subscriber scale has reached 142 million. China Telecom possesses the world's largest data infrastructure and greatest number of subscribers. Characterized by the wide coverage, numerous varieties, large volume and high quality, their data resources will create remarkable value through data mining and analysis. 2) To enhance computing power we have established an advanced infrastructure platform of big data, including the big data management platform, big data applications and operating platform, big data capability platform and enterprise-level big data acquisition system. So far, we have mastered the super-large-scale data processing and analysis capabilities. 3) In the algorithms research area, by combining open source technology and independent research and development ability, telecom operators’ AI teams focus on the direction of natural language processing and has achieved some technical breakthroughs in terms of information extraction, domain knowledge base establishment, information recommendation and filtering and so on. 4) In order to satisfy the internal requirements of operators and meet the needs of customers, by taking full advantage of the above strengths, telecom operators are trying to merge AI technology with innovative technologies such as cloud computing, big data, blockchain and Internet of things, to satisfy the internal and external demands of enterprises. Undoubtedly, this move will shape the characteristic intelligent technology architecture and support intelligent strategies.

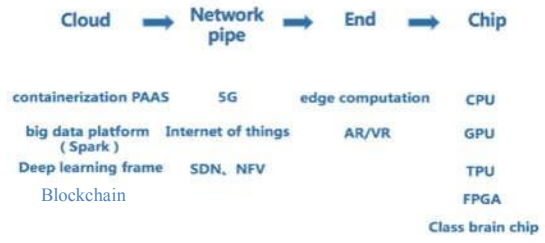


Fig. 1. The road map of AI development of China Telecom

3. TELECOM OPERATORS' PRACTICES IN THE FIELD OF ARTIFICIAL INTELLIGENCE

Based on these unique strengths, telecom operators have implemented a series of effective practices to satisfy the internal and external demands of enterprises. For example, in order to save energy in our internal data centers, by referring to the related algorithms published by Google Corporation, an algorithm based on deep learning has been proposed, which has got great results in energy saving during the tests in the data centers of telecom operator [4]. To help improve public security, by making full use of the “big data resource” advantage, and with a series of

algorithm research, an app named "intelligent police" has been developed, which has been already put into practice. With regard to computing ability advantage, through the effective use of 4G networks, Wi-Fi equipment, cloud storage etc., we also have made fruitful achievements in the field of health management and control. The developments of these applications have gained high-level influence and driven industrial change to some degree.

3.1. AI-based energy saving product in data centers

Telecom operators own a large number of data centers, just take China Telecom for example, it has already constructed 425 data centers all over China. The modern data center (DC) consists of tens of thousands of IT equipment. IT devices and their related service hardware such as air conditioners take up a large percentage of the total power consumption. DCs around the world are all facing the increasing power shortage pressure and cost-efficiency problem. State-of-the-art power-saving technology based on AI such as machine learning works effectively in this tough situation.

As one of the leaders of AI technology, researchers of Google Corporation proposed a novel method based on 5 layers BP neural networks to predict the power consumption trend of the whole DC. The PUE value is selected as the final standard to evaluate the model performance. Meanwhile, 19 features are taken into account to establish a relationship with PUE. This model provides a good idea for realizing data center energy saving.

Based on the model designed by Google, China Telecom Beijing Research Institute (CTBRI) proposes a five-layer neural network to reduce power consumption. DL excels at the feature learning part imitating the signal processing grading mechanism of the human brain. The addition of the convolution layer and the dimension reduction layer allows the deep learning architecture to consider more feature quantities that affect energy consumption. The machine automatically summarizes and learns the characteristics through the actual data. A new method is expected to get a more comprehensive and higher accuracy prediction model performance. In this case, the five-layer neural network uses 50 nodes per hidden layer and 0.001 as the regulation parameter. It has 19 normalized variables as input and the DC PUE as the output. These 19 features represent totals and averages, such as total server IT load, resource usage situation and the environmental index of DC(temperature and humidity). And most of these indexes were derived from individual sensor data. The training process of the five-layer neural network as well as its model are shown as follows:

1) Have the model parameters 8 value randomly initialized between [-1, 1].
2) Do the forward propagation step.
3) Compute the cost function.
4) Do the backward propagation step.
5) Repeat both forward and back propagation until convergence.

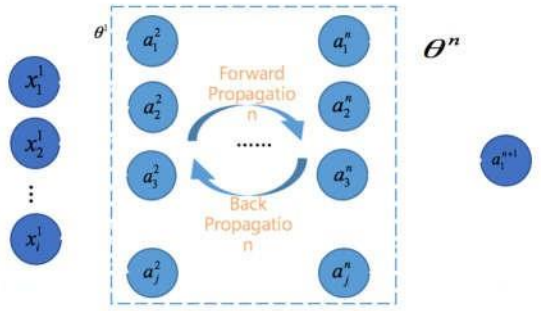


Fig.2. Five-layer neural network model

With the help of the DL framework, CTBRI thinks over the features via different types of services. Potential power-saving capacity is illustrated as the blue areas of part 1 of Fig.3. Each service has one or several power consumption peaks when virtual machines (VMs) work with high loads. However, during the non-peak times, physical devices still consume much energy, with few services running. A predefined strategy is triggered to move the VMs which run fewer jobs to another physical device that still has redundant capacity. Then the former physical machine is turned into idle in which state the device consumes much less energy for about 20W. In contrast with a full load state, i.e. 200~500W, a significant saving is realized.

The DL model does well in predicting the peak load with the support of accumulated data. Some devices will be activated in time and then handle the incoming jobs. Between the gap of peak and non-peak, CTBRIsets a 20% load capacity to handle the unexpected requests. That is to say, if 10 devices are actually needed, 2 devices are required to be ready.

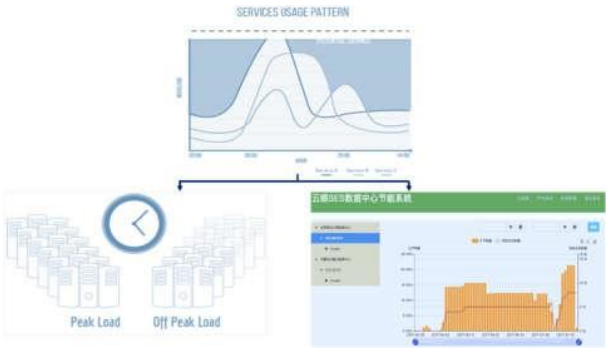


Fig.3. CTBRI data center power saving design based on AI

CTBRI has conducted some tests in the lab and obtained some good results. The lab resource pool consists of 16 physical machines, divided into 120 VMs. During the two months' test period, the total amount of power saving is up to 698 KWh, meaning 11.44 KWh every day. A total of 837 Yuan is saved. A practical test is conducted in Inner Mongolia Autonomous Region by China Telecom Tianyi Cloud Company. The Resourse pool consisted of 340 physical machines and was divided into 3250 VMs. Results show that about 528 Yuan was saved for each physical machine, indicating an average power saving rate of about 34%. AI-based energy saving products in data centers could not only lower production cost greatly, but also reduce energy consumption effectively.

3.2. AI-based public security management platform

Government customers are one of the major customer groups for telecom operators. The world's mainstream operators are all trying their best to help government make city management easier and more efficient. At present, the city administrators in China are confronting the increasingly complicated public security management problem. One of the facts is that the rapid development of urbanization has been constantly impacting the urban "steady structure" shaped in the planned economy period, and the urban management is facing increasing challenges with the growing population, especially in metropolises. In this context, the contradiction between the lower-level police deployment and the persistently ascending new cases has posed great challenges for the investigation and interrogation of cases; the conventional way cannot meet the needs of public security. Therefore, it is urgent to apply the intelligent means in investigating and comparing from multiple dimensions, such as suspect's funds, action track and associated figures, which helps to improve the efficiency of interrogating information, find out the crime clues and the total pieces of crime evidence and greatly reduce the investment of investigator resources. In this way, the prosecutors can focus more on determining the nature of the case, the directions and the means of investigation, so as to strengthen the detection efficiency.

The "intelligent police" as the AI relation insight system is researched and developed to work for public security, urban comprehensive management and other industries. The industry customer can apply "intelligent police" to establish a public safety management solution for intelligent cities.

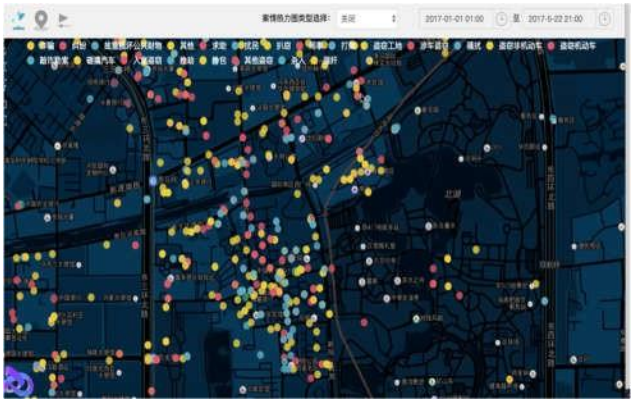


Fig.4. The interface of the "intelligent police"

Based on AI, cloud services, big data and the Internet of things technologies, this program accesses public security data and telecom data to help the urban administrators quickly transform various complicated unstructured data into knowledge mapping. After demonstrating, in the form of data visualization, the intelligence association based on the location information is completed and the comprehensive reports for processing tasks are presented, so as to supervise the multidimensional public opinion of the target object and events. An "intelligent police" AI relation insight system works to provide the varied customized features which center on three modules of case, map and association, including case analysis, clue extraction, case relation display, crime early warning, case thermodynamic diagram, suspect tracking, external population detection and urban road flow analysis. It has vastly enhanced the intelligence of public security management.

During the process of adopting AIT in the realized function of crime prediction, we have conducted a variety of effective algorithm exploration and research. (1) After the preliminary data analysis and technical investigation, random forest algorithm is selected to predict the type of crime, and months, days, periods, latitude and longitude are perceived as the features to

train the classifier. The classification accuracy rate is approximately 44%. There are four stages during our development iteration. Subsequently, the Xgboost algorithm which is the multiple decision tree algorithms as the random forest but possesses better effects and performance is chosen [7]. And the accuracy is increased to 48% via appropriately adjusted parameters. However, there are only about 3300 valid data obtained from the collected public security data for half a year in the preliminary stage, while the data is divided into more than 10 categories, which may trigger over-fitting. (2) In order to obtain better prediction accuracy, we reconsidered the real implication of the crime prediction and found that the model based on the case classification can merely provide the type of case when it happens rather than the probability of actual occurrence or the expected quantity. After the investigation, it is found that Self-Excitation Point Process Model (SEPP) put into use by PredPol, the famous crime forecast company in the United States, can serve to predict the crime quantity spatial density and makes some effective achievements. With the research on the related papers, the team adjusts the algorithm model in line with the features of public security data, including a selected appropriate time, space bandwidth and intercept, interceptive iterative process and the proper sampling method, the SEPP achieved the expected performance. The prediction accuracy goes up to 92%. The specific modeling process is as follows:

1) Parameter estimation: present a crime model composed of a variable bandwidth kernel smoothing estimate for the background intensity and a simple exponential-in-time, Gaussian-in-space kernel for the excited intensity.
2) Model fitting: determine the appropriateness of a model by using the data collected from a police station.
3) Simulation: test the goodness of the fit of the model.

At present, the team is trying to use the deep learning algorithm to improve the accuracy of crime prediction.

The intelligent public security management system has been implemented in one Public Security Bureau of Beijing and Sichuan Politics and Law Committee. The system has access to the local urban public security management data and is integrated with the mobile station data of operators. It is now assisting the urban administrators in making decisions with the aid from AIT, which has actually enhanced the intelligent level of city management.

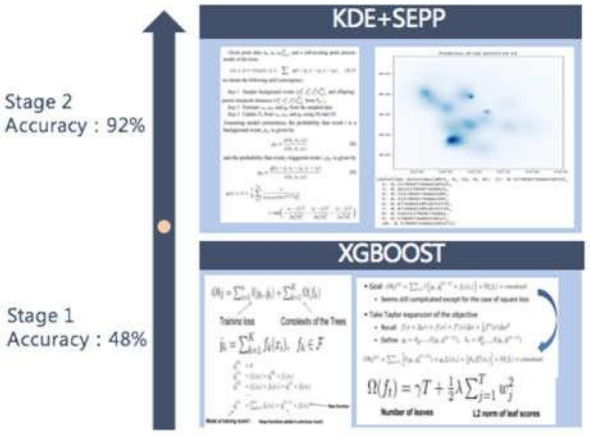


Fig.5. The evolution process of a prediction algorithm for crime prediction

3.3. AI-based health management and control

In recent years the medical AI market has shown a vast growth. Published by Chinese digital medical network, the report "The analysis report of 2016-2017 AI in medical market", analyzed that in the first quarter of 2017 more than 30 AI enterprises have obtained financing, most of which are in medical fields. Indeed, AI enterprises in the medical field have great development potential. Also, residents' health is a major concern of the government. Hence, telecom operators are doing a series research in health management and control field to fulfill government customers' requirements.

Operators have a great number of fixed networks and mobile network subscribers, thus a huge amount of data could be obtained. By analyzing these data, operators can extract the user's network behavior information. And by accessing other related data information, user's health database can also be established. Gradually, the medical image cloud could be built, which means a lot for developing AI in the medical field.

In the technical aspect, through the effective use of our abundant infrastructure resources, such as the basic light width, 4G network, Wi-Fi equipment, cloud storage, etc. operators could actively develop the AI technology in the medical field. Medical image features and a clinical problems classification model have been built based on the SVM and CNN model [8][9]. By quantitative analysis, depth comparison, deep learning, efficient and accurate picture reading could be achieved, which could help doctors diagnose intelligently. For instance, China Telecom's experimental results show that the medical image recognition accuracy based on deep learning can reach

90% and above. At the same time, the medical image cloud diagnostic center, image data center and image computing center have been constructed to realize remote image diagnoses, remote ECG diagnoses and remote test diagnoses effectively. Also, disease information database have been established. When the patient inputs his own disease information, the steps of extracting, normalizing, judging, and modeling the characteristic words could be immediately completed in turn, which could realize the AI guide. Additionally, by using the intelligent Q & A system (AIJust), common questions, such as how to register, how to pay, how the department is distributed and so on, could be answered intelligently.

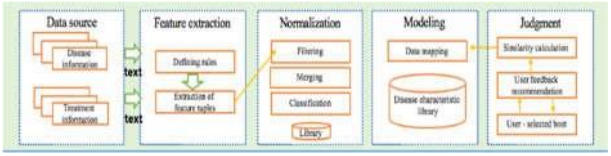


Fig.6. The technical framework of the AI guide

Take chronic disease management for example. AI could be used to predict the trend of the morbidity for one particular disease, by going through the following two stages:

1) Construct a comprehensive disease control index model. By accessing the data provided by a residents' health database, this index model was built upon comprehensive consideration of various factors such as disease morbidity, climax, humidity and so on.
2) Construct an artificial neural network model. The existing residents' health data can be used to get the artificial neural network model, while the patients' data could be utilized as a test set to determine the model accuracy.

For AI research in medical fields we believe that the technology of intelligent identification and diagnosis has been relatively mature, while data "quantity" and "quality" is the key to success.

At the application level, a "cloud + big data + AI" big data solution has been framed by China Telecom, which also effectively promotes the development of optical fiber broadband, cloud computing and cloud network integration. The operators' practice in the medical field mainly focuses on six application scenarios, that is chronic disease management, AI diagnosis, big data on meteorological medicine, prevention and control of infectious diseases, decision-making of drug use, and big data prediction. Among them, AI diagnosis works based on logical inference, of which the subjectivity is stronger, and neural network framework based on intelligent inference. Refer to the data from China Telecom, Logical inference could affirm the type of disease by correlating the 5000 diseases and 6000 symptoms in the structural database, while the F1 (assessment values of comprehensive accuracy and recall rate) of intelligent inference could be up to 0.5876. In the field of chronic disease management, operators have also achieved fruitful results. By classifying slow disease users according to their geographical distribution and presenting the distribution condition on the map, we can provide a city health department's visual disease distribution information and help them control chronic disease in some way. The product of preventing and controlling chronic disease has been recognized by the National CDC and put into use. This product will not only transform the existing medical image cloud and other resource-based business to the application-oriented business, but also increase the business input-output ratio.

4. SUMMARY AND PROSPECT

As the advancements of deep learning and other technologies, AI is now in the stage of being used in specific industries to increase efficiency and reduce costs. AI has been successfully applied to automatic pilots, medical treatments and health, finance, retail, entertainment, AR, VR and many other fields with unprecedented importance. Some experts say that AI may become the new productivity and even one of the key drivers of the fourth industrial revolution. For telecom operators, the opportunities and challenges coexist behind AI development. On the one hand, the largest big data infrastructure and the massive data resources greatly facilitate telecom operator's AI development. Rich data resources can greatly help to study AI algorithms and train models. On the other hand, there are still many obstacles to overcome, such as the gap between operators and Internet enterprises in the aspect of algorithm design and market promotion. In the future, telecom operators will make greater use of the traditional advantages, seek more distinctive areas for AI development and cooperate with other enterprises to create a development ecosystem and promote industrial change more efficiently.

REFERENCES

[1] Buchanan, Bruce G. "A (Very) Brief History of Artificial Intelligence." Ai Magazine 26.4(2005):53-60.

[2] Kim, Keun Heui, et al. "A Study on Message Queue Safe Proper Time for AI Open API Fintech Architecture." Computer and Applications 2016:33-36.

[3] "NTT and MIT to collaborate in next-generation Internet and AI." Computing Japan (1998).

[4] Gao, Jim, and Ratnesh Jamidar."Machine learning applications for data center optimization." Google White Paper (2014).

[5] YinShi An. The exploration and practice of operators' AI [J]. Communication enterprise management, 2017(5):73-75.

[6] Lecun, Yann, Y. Bengio, and G. Hinton. "Deep learning." Nature 521.7553(2015):436-444.

[7] Chen, Tianqi, and C. Guestrin. "XGBoost: A Scalable Tree Boosting System." (2016):785-794.

[8] Joachims, Thorsten. "Making Large-Scale SVM Learning Practical." Technische Universität Dortmund, Sonderforschungsbereich 475: Komplexitätsreduktion in multivariaten Datenstrukturen, 1998:499-526.

[9] Ozturk, Gurkan, Z. K. Ozturk, and A. A. Islier. A Comparison of Competitive Neural Network with Other AI Techniques in Manufacturing Cell Formation. Advances in Natural Computation. Springer Berlin Heidelberg, 2006:575-583.

[10] Setiawan, Noor Akhmad, et al. "Diagnosis of coronary artery disease using Artificial Intelligence based decision support system." University Malaysia Perlis 9.5(2009):93–118.

CORRELATION AND DEPENDENCE ANALYSIS ON CYBERTHREAT ALERTS

John M.A. Bothos, Konstantinos-Georgios Thanos, Dimitris M. Kyriazanos, George Vardoulias,
Andreas Zalonis, Eirini Papadopoulou, Yannis Corovesis, Stelios C.A. Thomopoulos
National Centre for Scientific Research "Demokritos" (NCSRD), Greece

Abstract – In this paper a methodology for the enhancement of computer networks' cyber-defense is presented. Using a time-series dataset, drawn for a 60-day period and for 12 hours per day and depicting the occurrences of cyberthreat alerts at hourly intervals, the correlation and dependency coefficients that occur in an organization's network between different types of cyberthreat alerts are determined. Certain mathematical methods like the Spearman correlation coefficient and the Poisson regression stochastic model are used. For certain types of cyberthreat alerts, results show a significant positive correlation and dependence between them. The analysis methodology presented could help the administrative and IT managers of an organization to implement organizational policies for cybersecurity.

Keywords – Correlations, cyberattacks, dependencies, network, time series

1. INTRODUCTION

Today most organizations in the world heavily depend on IT infrastructure such as computer networks, servers, databases and information systems, to carry out their daily activities. This infrastructure has been the target of cyberattacks which aim to disrupt the ability of an organization to perform its activities, steal data or even put it out of business. According to [1] and [2], after a security breach, organizations are affected in fields such as, operations, finance systems, brand reputation and customer retention. Cyberattacks cause various direct or hidden costs to an organization's tangible and intangible assets, jeopardizing even its sustainability in some cases. This emphasizes the need for organizations to prioritize cybersecurity so as to minimize the risk of a cyberattack being successful. An organization that can reduce its administrative costs by optimizing its cybersecurity defense mechanisms, can divert more monetary resources to other investments for business growth. Due to the high costs involved in adopting and implementing a proactive cybersecurity policy, organizations usually develop ineffective cybersecurity solutions as reactions to cyberattack incidents [3]. Implementing an effectively proactive information security policy makes the IT infrastructure more productive, increases its availability and guarantees an organization's activities to continue uninterrupted. To implement such an effective line of cyber-defense, an organization not only has to determine the value of its assets, but also the cyberthreat environment, by determining correlations and dependencies between various types of cyberattacks and malware.

Research on cyberattack pattern recognition in network traffic has been going on for quite some time. Relevant research approaches have been made in the scope of finding satisfactory predicting mathematical models for such incidents. Empirical modelling of cyber-alerts relates mainly to the study of time-series models for efficient forecasting of cyberattacks. In [4], Markov models on time-series data of communications were used to highlight the importance of detecting types of anomalies in a computer network traffic flow in identifying types of intrusions, in the network. In [5] and [6], ARFIMA and FIGARCH models were used on time-series data of network traffic, to predict whether detected anomalies are indications of real cyberattacks or just false alarms and to detect cyberattacks on a DDoS network. In [7] predictive time-series models were used to forecast vulnerabilities of web browsers, while in [8], a dynamic risk assessment stochastic model is used to identify inventory-enhancement opportunities for critically disrupted systems.

Our study contributes to the relevant research by applying mathematical methods for the detection of significant correlation and dependence between different types of cyberthreat alerts. In order to determine the degrees of these correlations and dependencies, Spearman's correlation coefficient and Poisson regression stochastic modelling are used. Significant correlations and dependencies among certain types of cyberthreat alerts are distinguished that can be used for event count predictions of such incidents.

Through our contribution we aspire to enrich the variety of scientific methods that have been employed so far for the analysis of cyberattack pattern recognition in computer networks' traffic. As well as the use of *Markov*,*ARFIMA* and *FIGARCH* time-series models, we propose the use of probabilistic relevant time-series models, like the *Poisson* stochastic model, in order to determine valid dependence relations between different types of cyber-alerts.

2. METHODOLOGY

2.1. Experiment set up: network description and data mining approach

The alert logs of an intrusion detection system (IDS), already deployed in the network of an organization, were used for the formation of the analysis dataset. This IDS is placed in the entry point next to the border router of the network of the organization. The network serves about 1000 users categorized in certain organizational units. Each unit serves a different scientific discipline or support division. Each such unit is protected by a dedicated firewall and VLAN segmentations, operates the local infrastructure under its own management and has absolute control of the firewall and UTP cabling.

The IDS operation is at the front of all individual firewalls and the only communicated addresses concerned the external IP of each firewall. All user traffic was NATed on incoming or outgoing directions. The collaborating network operation center personnel have the absolute clearance to manage and operate the border routers, the centrally enforced access lists and the central IDS. The IDS alert data were collected in a database of events and that data was communicated to the research analyst team following a privacy impact assessment and anonymization procedure from the authorized network administration personnel. As a result, the database of events and all research data processed in the context of this paper, contain no information about individual IP addresses that correspond to user workstations or any other content that could be used to directly or indirectly identify a network user, i.e. by exclusion, narrowing down to a very small number of possible subjects or correlation and cross-matching with other public information. A thorough analysis on the privacy and legal challenges of network research can be found in [9].

The IDS system used is the *Suricata* intruder detection system [13] based on the *Oinkmaster* ruleset [14]. The log file records of cyberthreat alerts were related to the most frequent and costliest types of cyberattacks. Our dataset consisted of the following selection of cyberthreat alert protocols, detected by the IDS.

Types of cyberthreat alerts

Type of alert	Description of cyberthreat
WORM/TROJAN ()	This alert is emerged from traffic that is related to the propagation of viruses and worms and contamination of systems
TOR	Alert that indicates that illegal communications are taking place, such as espionage, criminal communications, illegal financial transactions, etc.
GPLSNMP	Alert emerged from traffic related to leakage information, mainly exploiting vulnerabilities of protocols, such as SNMP.
VOIP	This alert is related to attempts to exploit vulnerabilities for illegal usage of Internet telephony servers.
SQL	Multiple attacks related to the vulnerabilities of databases.
GPLRPC	Alerts related to the exploitation of the RPC PROTOCOL, e.g. malicious software injection.
IPMI	Alert that signals possible exploitation of system consoles at a very low level.
MOBILE	Mobile malware alert about installation of software to mobile search that exploit end-user devices, i.e. phishing, etc.
CNC	Command and control which indicates traffic related to the management of botnets that carry out cyberattacks such as DDOS.
DNS	Alert about suspicious queries to the domain servers related to information leakage.
SPAMHAUS	Unsolicited email traffic.
SCAN	Network activity related to external attempts to reconnaissance topologies, network services, operating systems, in order to exploit vulnerabilities.
MALWARE	Alert about transformed normal software with malicious parts in order to exploit unaware users.
DDOS	Cyberattacks related to denial of services, unable to operate.
COMPROMISED	Alert about systems that have been penetrated.

Records included every cyberthreat alert in the network traffic flow, during working days and hours, from Monday to Friday and from 8 a.m. until 8 p.m., for a 60-day period. At the end of the test period, this volume of log files was processed with data mining methods in order to filter the incidents by type of cyberthreat alert. For each cyberthreat alert category, we aggregated the number of relevant incidents by a time step of an hour, in order to form an adequately large sample size, so that valid and unbiased statistical results be produced. This resulted in fifteen (15) time series of the number of incidents by type, each one corresponding to one of the above-mentioned cyberthreat alerts, respectively. Consequently, each time series consists of an hourly number of incidents recorded for each specific cyberthreat alert.

2.2. Correlation analysis

Usually IDS outputs of detected cyberthreat alerts about suspicious signatures are huge, requiring an awful lot of network operator's attention and systematic analysis as part of cybersecurity actions. Correlation analysis offers an optimizing solution to the problem of limited machine computation power, reducing the quantity of data that needs to be processed, in order to extract useful information, without losing the overall situational picture. Detecting potential correlations among the various kinds of cyberthreat alerts is a prerequisite in order to proceed to modelling the dependencies between them. Due to the nature of our dataset, composed by discrete variables consisted of count data and not continuous random variables we could not resort to the Pearson correlation coefficient, so we preferred instead to use the Spearman's correlation coefficient:

$r_s = 1 - \{(6 * \Sigma d_i^2) / [n * (n^2 - 1)]\}$, where

Σ: Sum

d_i: differences between the ranks of pairwise cyberthreat alerts

n: number of cyberthreat alerts(sample size)

for the calculation of the correlation coefficients between these fifteen (15) different types of cyberthreat alerts, each by every other. As a result, a 15x15 correlation matrix emerged, in which each cell contained the Spearman's correlation coefficient value, for the respective cyberthreat alerts. Setting a threshold of significant positive correlation at 40% and over, results yielded that with a 95% probability or $\alpha = 0.05$ level of statistical significance.

SQL and *VOIP* cyberthreat alerts were correlated by 61.9%, *SQL* and *SCAN* cyberthreat alerts were correlated by 47.1%, COMPROMISED and SQL cyberthreat alerts were correlated by 51.7% and *COMPROMISED* and *SCAN* cyberthreat alerts were correlated by 92%.

2.3. Dependence analysis

Exploring further the significantly correlated cyberthreat alerts, we made use of regression stochastic modelling to estimate the potential dependence between them.

We formed the following functions about the dependence between the significantly correlated cyberthreat alerts, with over 40% positive correlation between them.

SQL=f(VOIP), VOIP=f(SQL), SCAN=f(SQL),
SQL=f(SCAN), COMPROMISED=f(SQL),
SQL=f(COMPROMISED),
COMPROMISED=f(SCAN),
SCAN=f(COMPROMISED).

So the respective regressions that had to be run were the following:

SQL cyberthreat alert on *VOIP* cyberthreat alert.
VOIP cyberthreat alert on *SQL* cyberthreat alert.
SCAN cyberthreat alert on *SQL* cyberthreat alert.
SQL cyberthreat alert on *SCAN* cyberthreat alert.
COMPROMISED cyberthreat alert on *SQL* cyberthreat alert.
SQL cyberthreat alert on *COMPROMISED* cyberthreat alert.
COMPROMISED cyberthreat alert on *SCAN* cyberthreat alert.
SCAN cyberthreat alert on *COMPROMISED* cyberthreat alert.

For each type of cyberthreat alert, we estimated the dependence of the mean number of its emergence in the network at an hour t, on the number of incidents of other cyberthreat alerts that are over 40% ($r_s \geq +0.5$) positively correlated, at the same hour t and the previous 2 hours t-1, t-2. We selected this time lapse window, based on the intuition of imminent action against a cyberthreat alert from the network's administrator in order to prevent and/or mitigate as much as possible potential damages and losses from a cyberattack. A maximum 2-hour interval provides an adequate time window of action by the network operators to apply response counter measures. Smaller windows include much more noise and bigger windows would probably

miss important events and offer less opportunity to respond.

For this dependence analysis, we made use of a Poisson stochastic model. Taking the Poisson Probability Distribution Function:

$Prob(Y=y) = (e^{-\lambda}*\lambda^{y})/y!$, $y=0,1,2,\ldots$. [10], where

λ: distribution parameter of *Poisson Probability Distribution Function* concerning the emergence of cyberthreat alerts in the network traffic flow and

$y!$: observed counts of emergence of each cyberthreat alert in the network traffic flow.

Considering the mean number of cyberthreat alerts of type i happening at hour t, depending linearly on the number of cyberthreat alerts of another type j at the same hour t_0 and/or at the previous 2 hours t-1, t-2 as $E(y_{it}|\boldsymbol{x}_{jt})=\lambda_t=e^{\boldsymbol{x'_{jt}}*\boldsymbol{b_j}}$ [11],

we finally take our model concerning the dependence of the probability the mean number of cyberthreat alert type i emerging at hour t, on the number of cyberthreat alert type j at the same hour t and/or the previous two hours t-1, t-2 in the form of $Prob(y_{it}=\lambda_{it})=\{[(e)^{-e^{(x'_{jt}*b_j)}}*(e^{x'_{jt}*b_j})^{y_{it}}]/y_{it}!\}+u_{it}$ [12].

According to [11], 'such models are estimated with maximum likelihood methodology, with the log-likelihood function being

$ln(y,b)=\Sigma_{t=1}^{n}(-e^{x'_{jt}*b_j}+y_{it}*\boldsymbol{x'_{jt}}*\boldsymbol{b_j}-lny_{it}!)$ [11]

and the relevant likelihood equations being

$\theta ln(y,b)/\theta \boldsymbol{b_j}=\Sigma_{t=1}^{n}(y_{it}-e^{x'_{jt}\boldsymbol{b_j}})*\boldsymbol{x_{jt}}=\boldsymbol{0}$ [11]

formatting the following Hessian matrix

$\theta^2 ln(y,b)/\theta \boldsymbol{b_j}*\theta \boldsymbol{b_j}'=-\Sigma_{t=1}^{n}(e^{x'_{jt}*\boldsymbol{b_j}}*\boldsymbol{x_{jt}}*\boldsymbol{x_{jt}}')$ [11]

The asymptotic estimator covariance matrix is in the form of

$[\Sigma_{t=1}^{n}(e^{x'_{jt}*\boldsymbol{b_j}})_{est}*\boldsymbol{x_{jt}}*\boldsymbol{x_{jt}}']^{-1}$ [11]

Testing for the statistical significance of the model's hypotheses about the estimator, involves use of the LR statistic in the form of

$LR=2*\Sigma_{t=1}^{n}[ln(P_{iest}/P_{iestrestricted})]$ [11]

2.4. Results

The results of the regressions are shown in the following tables.

Table 1. Dependence of *SQL* cyberthreat alert on *VOIP* cyberthreat alert $SQL = f(VOIP)$

Dependent variable: SQL Method: ML/QML – Poisson count (Quadratic hill climbing)				
Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	12.22246	0.000136	89964.57	0.0000
VOIP	0.000374	2.17E-06	172.2328	0.0000
VOIP(-1)	−8.15E-05	2.93E-06	−27.83379	0.0000
VOIP(-2)	0.000352	2.13E-06	164.9999	0.0000
LR statistic 453225.2 Prob(LR statistic) 0.000000				

Table 2. Dependence of *VOIP* cyberthreat alert on *SQL* cyberthreat alert $VOIP = f(SQL)$

Dependent variable: VOIP Method: ML/QML – Poisson count (Quadratic hill climbing)				
Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	−0.755649	0.050791	−14.87762	0.0000
SQL	1.13E-05	5.64E-07	20.02671	0.0000
SQL(-1)	9.10E-06	6.33E-07	14.36263	0.0000
SQL(-2)	4.05E-06	3.77E-07	10.73018	0.0000
LR statistic 14786.20 Prob(LR statistic) 0.000000				

Table 3. Dependence of *SCAN* cyberthreat alert on *SQL* cyberthreat alert $SCAN = f(SQL)$

Dependent variable: SCAN Method: ML/QML – Poisson count (Quadratic hill climbing)				
Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	7.550100	0.003718	2030.727	0.0000
SQL	4.26E-07	3.37E-08	12.64695	0.0000
SQL(-1)	5.63E-06	4.35E-08	129.4269	0.0000
SQL(-2)	2.84E-06	2.99E-08	94.77684	0.0000
LR statistic 320368.3 Prob(LR statistic) 0.000000				

Table 4. Dependence of *SQL* cyberthreat alert on *SCAN* cyberthreat alert *SQL = f (SCAN)*

Dependent variable: SQL Method: ML/QML – Poisson count (Quadratic hill climbing)				
Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	12.29104	9.05E-05	135792.3	0.0000
SCAN	1.88E-07	2.45E-09	76.95380	0.0000
SCAN(-1)	−1.02E-08	2.74E-09	−3.715406	0.0002
SCAN(-2)	1.38E-07	2.45E-09	56.15487	0.0000
LR statistic 10436.97 Prob(LR statistic) 0.000000				

Table 5. Dependence of *COMPROMISED* cyberthreat alert on *SQL* cyberthreat alert *COMPROMISED = f (SQL)*

Dependent variable: COMPROMISED Method: ML/QML – Poisson count (Quadratic hill climbing)				
Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	6.509554	0.008114	802.2647	0.0000
SQL	4.79E-06	7.47E-08	64.05926	0.0000
SQL(-1)	1.12E-06	8.13E-08	13.75862	0.0000
SQL(-2)	1.99E-07	5.79E-08	3.438938	0.0006
LR statistic 32068.19 Prob(LR statistic) 0.000000				

Table 6. Dependence of *SQL* cyberthreat alert on *COMPROMISED* cyberthreat alert *SQL = f (COMPROMISED)*

Dependent variable: SQL Method: ML/QML – Poisson count (Quadratic hill climbing)				
Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	12.13649	0.000292	41520.08	0.0000
COMPRO-MISED	0.000322	7.60E-07	424.0071	0.0000
COMPRO-MISED(-1)	−0.000179	8.30E-07	−216.0809	0.0000
COMPRO-MISED(-2)	−8.15E-05	5.04E-07	−161.6978	0.0000
LR statistic 492892.9 Prob(LR statistic) 0.000000				

Table 7. Dependence of *COMPROMISED* cyberthreat alert on *SCAN* cyberthreat alert *COMPROMISED = f (SCAN)*

Dependent variable: COMPROMISED Method: ML/QML – Poisson count (Quadratic hill climbing)				
Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	7.847009	0.000829	9465.598	0.0000
SCAN	2.71E-07	2.18E-08	12.42193	0.0000
SCAN(-1)	2.47E-08	2.44E-08	1.013099	0.3110
SCAN(-2)	2.40E-07	2.17E-08	11.02043	0.0000
LR statistic 352.6056 Prob(LR statistic) 0.000000				

Table 8. Dependence of *SCAN* cyberthreat alert on *COMPROMISED* cyberthreat alert *SCAN= f (COMPROMISED)*

Dependent variable: SCAN Method: ML/QML – Poisson count (Quadratic hill climbing)				
Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	9.165064	0.001187	7718.030	0.0000
COMPRO-MISED	−0.000475	3.22E-06	−147.6575	0.0000
COMPRO-MISED(-1)	0.000737	3.76E-06	196.1082	0.0000
COMPRO-MISED(-2)	−0.000129	1.95E-06	−66.38216	0.0000
LR statistic 126924.1 Prob(LR statistic) 0.000000				

The estimated coefficients by the Poisson regressions $Prob(y_{it}=\lambda_{it})=\{[(e)^{-e\,(x'jt*bj)}*(e^{x'jt*bj})^{yit}]/y_{it}!\} + u_{it}$ [12] in the above tables give the exact form of the dependence of the emergence of the mean number of one type of cyberthreat alert *i* at an hour *t*, on the emergence of another highly correlated (>40%) cyberthreat alert *j* at the same hour t and the previous 2 hours *t*-1, *t*-2. They also provide for the calculation of the relevant probabilities, through $E(y_{it}|\boldsymbol{x}_{jt})=\lambda_t=e^{x'jt*bj}$ [11].

The LR statistic of the relevant error probabilities $LR=2*\Sigma_{t=1}^{n}[ln(P_{iest}/P_{iestrestricted})]$ [11] denotes the statistical significance of the model's hypotheses about the estimators.

3. CONCLUSION

In this paper we applied mathematical methods to highlight possible relationships between different types

of cyberthreat alerts in a network system. Our goal was to contribute to the enhancement of network cyber-defense policies by improving the effectiveness of IT systems' intelligence.

With the use of Spearman correlation analysis and Poisson regression stochastic modelling we tried to distinguish significant correlations and dependencies among certain types of cyberthreat alerts that can be used for forecasting such incidents.

Correlation analysis denoted a significant positive correlation of over 40% for the following pairs of cyberthreat alerts:

SQL and *VOIP* (61.9%), *SQL* and *SCAN* (47.1%), COMPROMISED and SQL (51.7%) and *COMPROMISED* and *SCAN* (92%).

The above significant correlations implied a strong degree of similar pattern of emergence and possible existence of a significant relationship between them. Based on these results, we proceeded further with the estimation of potential dependence between these cyberthreat alerts, regarding the dependence of the mean number of each type's emergence in the network at an hour *t*, on the number of the emergence of its over 40% positively correlated other type of cyberthreat alerts, at the same hour *t* and the previous 2 hours *t*-1, *t*-2.

Results of the dependence analysis denoted that

SQL cyberthreat alert emergence significantly depended on *VOIP* cyberthreat alert emergence at the same, one and two hours before,

VOIP cyberthreat alert emergence significantly depended on *SQL* cyberthreat alert emergence at the same, one and two hours before,

SCAN cyberthreat alert emergence significantly depended on *SQL* cyberthreat alert emergence at the same, one and two hours before,

COMPROMISED cyberthreat alert emergence significantly depended on *SQL* cyberthreat alert emergence at the same, one and two hours before, *SQL* cyberthreat alert emergence significantly depended on *COMPROMISED* cyberthreat alert at the same, one and two hours before.

COMPROMISED cyberthreat alert emergence significantly depended on *SCAN* cyberthreat alert at the same and two hours before, but not at one hour before.

SCAN cyberthreat alert emergence significantly depended on *COMPROMISED* cyberthreat alert at the same, one hour and two hours before.

For all our dependence models the respective LR statistics denoted statistical significance of the models' hypotheses about the estimators.

The results of our research can be used as suggestions to IT managers, in order to apply and implement more efficiently cybersecurity strategies and cyber-defense tactics, without the need of monitoring all incidents emerging and so reduce data storing consumption and network capacity overload.

ACKNOWLEDGEMENT

The research leading to these results has received funding from the European Union's Horizon 2020 Research and Innovation Programme, under Grant Agreement no 740829 – SAINT project.

REFERENCES

[1] Cisco, "Annual Cybersecurity Report" (2017).

[2] Deloitte, "Beneath the Surface of a Cyberattack: A Deeper Look at the Business Impacts".

[3] Shim, "Agency Problems in Information Security: Theory and Application to Korean Business", The Journal of Internet Electronic Commerce Research, Vol. 15 (2015).

[4] Neil, Hash, Brugh, Fisk, Storlie "Scan Statistics for the Online Detection of Locally Anomalous Subgraphs", Technometrics, (2013).

[5] Andrysiak, Saganowski, Choras, Kozik "Network Traffic Prediction and Anomaly Detection Based on ARFIMA Model", International Joint Conference SOCO'14-CISIS'14-ICEUTE'14.

[6] Andrysiak, Saganowski, Maszewski, Grad "Long-memory dependence statistical models for DDoS attacks detection", Image Processing & Communications, vol. 20, (2015).

[7] Roumani, Nwankpa and Roumani "Time series modelling of vulnerabilities", Computers and Security 51, (2015).

[8] Resurreccion and Santos "Uncertainty modelling of hurricane-based disruptions to interdependent economic and infrastructure systems", Nat Hazards 69 (2013).

[9] Ohm, Sicker, Grunwald, "Legal Issues Surrounding Monitoring During Network Research", IMC '07 (2007).

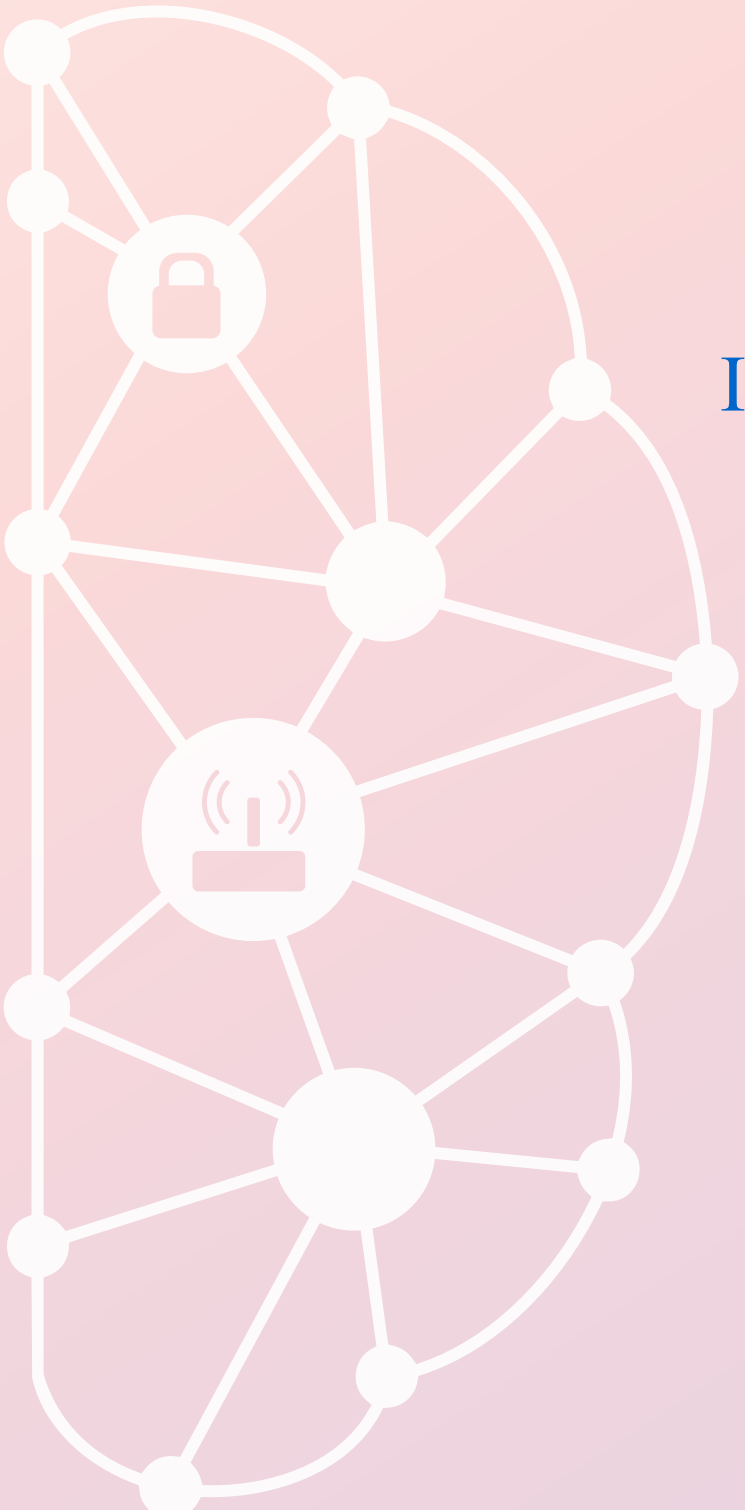
[10] Davidson, McKinnon, "Econometric Theory and methods", (1999).

[11] Green, "Econometric Analysis", (2002).

[12] Gujarati, Porter "Basic Econometrics", (2008).

[13] Suricata Open Source IDS https://suricata-ids.org/

[14] Oinkmster http://oinkmaster.sourceforge.net/about.shtml



INDEX OF AUTHORS

Index of authors

T

V

W

Y

Z

International
Telecommunication
Union

Telecommunication
Standardization Bureau (TSB)

Place des Nations
CH-1211 Geneva 20
Switzerland

Published in Switzerland
Geneva, 2018