ITUPublications
Recommendations

International Telecommunication Union
Standardization Sector

Recommendation

ITU-T F.748.20 (12/2022)

SERIES F: Non-telephone telecommunication services

Multimedia services

Technical framework for deep neural network model partition and collaborative execution

ITU-T F-SERIES RECOMMENDATIONS

NON-TELEPHONE TELECOMMUNICATION SERVICES

TELEGRAPH SERVICE	
Operating methods for the international public telegram service	F.1–F.19
The gentex network	F.20–F.29
Message switching	F.30–F.39
The international telemessage service	F.40–F.58
The international telex service	F.59–F.89
Statistics and publications on international telegraph services	F.90–F.99
Scheduled and leased communication services	F.100–F.104
Phototelegraph service	F.105–F.109
MOBILE SERVICE	
Mobile services and multidestination satellite services	F.110–F.159
TELEMATIC SERVICES	
Public facsimile service	F.160–F.199
Teletex service	F.200–F.299
Videotex service	F.300–F.349
General provisions for telematic services	F.350–F.399
MESSAGE HANDLING SERVICES	F.400–F.499
DIRECTORY SERVICES	F.500–F.549
DOCUMENT COMMUNICATION	
Document communication	F.550–F.579
Programming communication interfaces	F.580–F.599
DATA TRANSMISSION SERVICES	F.600–F.699
MULTIMEDIA SERVICES	**F.700–F.799**
ISDN SERVICES	F.800–F.849
UNIVERSAL PERSONAL TELECOMMUNICATION	F.850–F.899
ACCESSIBILITY AND HUMAN FACTORS	F.900–F.999

For further details, please refer to the list of ITU-T Recommendations.

Recommendation ITU-T F.748.20

Technical framework for deep neural network model partition and collaborative execution

Summary

The deep neural network (DNN) model inference process usually requires a large amount of computing resources and memory. Therefore, it is difficult for end devices to perform DNN models independently. An effective way to implement end-edge collaborative DNN execution is through DNN model partition, which can reduce latency and improve resource utilization at the same time. Recommendation ITU-T F.748.20 aims to specify the technical framework of DNN model partition and collaborative execution. First, it is necessary to predict the overall inference latency under the current system state according to different DNN partition strategies in advance. Then, it is necessary to choose the appropriate partition locations and collaborative execution strategy based on the equipment computation capabilities, network status and DNN model properties. Finally, the model collaborative execution is implemented and the resource allocation optimized in the meantime.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T F.748.20	2022-12-14	16	11.1002/1000/15197

Keywords

Collaborative execution, deep neural network, model partition.

* To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had not received notice of intellectual property, protected by patents/software copyrights, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the appropriate ITU-T databases available via the ITU-T website at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Recommendation ITU-T F.748.20

Technical framework for deep neural network model partition and collaborative execution

1 Scope

This Recommendation aims to specify the technical framework of deep neural network (DNN) model partition and collaborative execution, including the establishment of a latency prediction model, the formulation of a DNN model partition strategy, resource allocation optimization and the collaborative execution of the model, so as to meet the demand of DNN model inference and realize the efficient resource utilization between the end devices and the edge device.

The scope of this Recommendation includes:

– Overview of DNN model partition and collaborative execution;

– Latency prediction model establishment of a DNN model partition;

– Partition strategy formulation of a DNN model;

– Collaborative execution of a DNN model after DNN model partition.

2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T F.748.11] Recommendation ITU-T F.748.11 (2020), *Metrics and evaluation methods for deep neural network processor benchmark.*

[ITU-T F.748.13] Recommendation ITU-T F.748.13 (2021), *Technical framework for the shared machine learning system.*

3 Definitions

3.1 Terms defined elsewhere

None.

3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

3.2.1 deep neural network model partition: The deep neural network model is partitioned into two or more parts according to partition strategies, and each part selects its execution location (i.e., end/edge device).

3.2.2 deep neural network model inference: All deep neural network layers are executed in sequence, and the inference result is obtained after completing the last deep neural network layer.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

ADA	Adaboost
ANN	Artificial Neural Network
DNN	Deep Neural Network
DT	Decision Tree
GBRT	Gradient Boosting Regression Tree
KNN	K-Nearest Neighbour
KRR	Kernel Ridge Regression
LR	Linear Regression
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
PII	Personally Identifiable Information
RANSAC	Random Sample Consensus regression
RF	Random Forest
SVM	Support Vector Machine

5 Conventions

In this Recommendation:

– The keywords "**is required to**" indicate a requirement which must be strictly followed and from which no deviation is permitted if conformance to this document is to be claimed.

– The keywords "**is recommended**" indicate a requirement which is recommended but which is not absolutely required. Thus this requirement needs not be present to claim conformance.

6 Overview of DNN model partition and collaborative execution

The overall framework of DNN model partition and end-edge collaborative execution is shown in Figure 1. The overall framework consists of the edge device and multiple end devices, which includes the computation latency predictor module, network monitoring module, model processor module, DNN partitioner module, request processor module and deployment tool module. The deployment tool module is used to collect the training latency under different system configurations and control the CPU and memory usage of end devices and the edge device. The model processor module is used to analyse and extract the DNN model topology, static features of DNN layers, DNN output data size and so on. The computation latency predictor module is used to predict the computation latency of each DNN layer using the metrical data collected by the deployment tool module, which takes multiple static and dynamic system characteristics into account. The network monitoring module is used to monitor the network upload/download bandwidth in real time and to estimate the transmission latency of DNN layers. The DNN partitioner module is used to determine the appropriate partition locations for end-edge collaborative execution according to the feedback information of the model processor module, computation latency predictor module and network monitoring module. The request processor module is used to process the real-time request of DNN model inference from end devices, which controls the end devices and edge device to collaboratively complete the DNN inference according to the partition scheme.

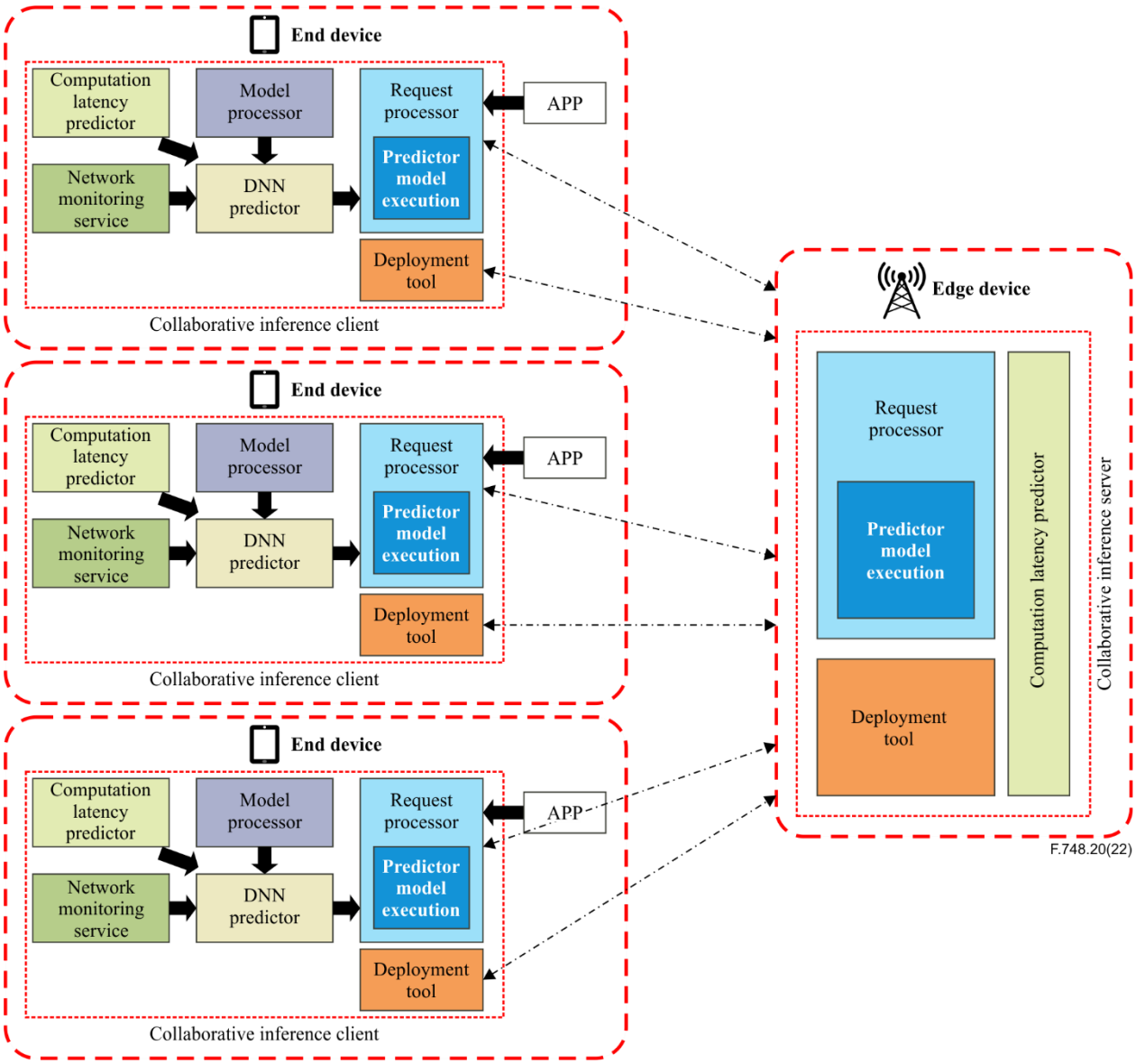

Figure 1 – The overall framework of DNN model partition and collaborative execution

7 Latency prediction model establishment of a DNN model partition

7.1 System state monitoring and parameter configuration

The DNN model consists of many types of neural network layer, such as the convolutional layer, the fully connected layer, the pooling layer and the activation layer. Each type of neural network layer has different parameter configurations and is required to establish different latency prediction models. The convolutional layer and fully connected layer are the most common layers. The main parameters of the convolutional layer include input size, kernel size, channel size, stride and padding. The main parameters of the fully connected layer include input size and output size.

In addition to layer configuration, the system state also significantly affects the layer computation latency. It is required to monitor the dynamic system state, including CPU usage, free memory size, memory usage, network bandwidth, network topology and so on. These dynamic system states are required to be accurately and timely monitored, in order to profile the latency prediction model.

7.2 Modelling and analysis of latency prediction

It is required to establish the computation latency prediction model for each type of neural network layer. The input data of the latency prediction model includes the layer parameter configuration and system state. The output data of the latency prediction model is the layer computation delay given the current layer parameter configuration and system state. There are many latency prediction models, such as linear regression (LR), random sample consensus regression (RANSAC), kernel ridge regression (KRR), k-nearest neighbour (KNN), decision tree (DT), support vector machine (SVM), random forest (RF), adaboost (ADA), gradient boosting regression tree (GBRT) and artificial neural network (ANN). It is required to compare multiple latency prediction models and determine the most suitable model. The evaluation indicators include mean absolute error (MAE) and mean absolute percentage error (MAPE).

8 Partition strategy formulation of a DNN model

8.1 Latency analysis of different partition strategies under diverse system state

The overall execution latency of each end device consists of three parts: the local computation latency on the end device, the communication latency between the end and edge devices, and the edge computation latency at the edge device. Different DNN model partition strategies can impact the overall execution latency. Specifically, the end device computation capability and the amount of DNN layers executed locally determine the local computation latency. The allocated communication resource amount and the intermediate data size determine the communication latency. The amount of the allocated edge computation resource and remaining DNN layers determine the edge computation latency. When the system state such as the amount of the allocated communication and computation resources changes, the partition strategies are required to change correspondingly, aiming to realize the trade-off between computation latency and communication latency.

8.2 Partition position confirmation and resource allocation

According to the latency analysis of different partition strategies under a diverse system state, it is required to jointly optimize a DNN partition strategy and resource allocation, so as to minimize latency or energy consumption. The DNN model partition strategy refers to divide the DNN model into two or more parts by deciding appropriate partition locations and selecting the execution location (i.e., end/edge device) for each DNN layer. DNN models can be divided into two categories: chain DNN and non-chain DNN. It is required to make different partition strategies for two types of DNN model. Resource allocation mainly involves the allocation of communication and computation resources of the edge device.

The partition location is determined by the partition strategy. Considering the framework in Figure 1, multiple end devices cooperate with a resource-constrained edge device. Each end device is required to make its partition decision according to the computation capabilities of the end devices and the allocated edge resources. If the computation capability of the end device is sufficient, its DNN model is not partitioned and does not occupy edge resources, in which the overall DNN model is all executed by the end device. However, if the computation capability of the end device is insufficient, its DNN model will be partitioned and jointly executed by the end device and the edge device.

Furthermore, the partition strategy is required to be handled jointly with resource allocation among end devices. Specifically, the edge device is required to efficiently allocate its computation resources to each end device for executing the offloaded DNN layers from end devices, and is required to allocate its communication resources to each end device for intermediate data transmission.

9 Collaborative execution of a DNN model after DNN model partition

9.1 DNN model deployment based on partition strategy

After determining the DNN model partition strategy, the execution location of each DNN layer will be obtained, either at the edge device or on the end device. Due to the diversity and variation of DNN models, it is not practical to deploy all DNN models at the edge device in advance.

There are two options to deploy the DNN models on demand based on the partition strategy. The first option is to deploy the entire required DNN model at the edge device. The second option is to deploy the DNN layers which are partitioned to the edge device for execution. The DNN model execution can be performed only after the DNN model or the required DNN layers are deployed.

9.2 Inference process of collaborative execution

Each DNN layer is sequentially executed either at the edge device or on the end device according to the partition strategy [ITU-T F.748.13]. Each DNN layer receives the output of its direct antecedent layers as its own input, performs DNN execution and delivers its output to its direct subsequent layers. If the execution locations of the antecedent layer and its direct subsequent layer are different (i.e., the antecedent layer is executed at the edge device and its direct subsequent layer is executed on the end device, or the antecedent layer is executed on the end device and its direct subsequent layer is executed at the edge device), the output of the antecedent layer is required to be transmitted to its direct subsequent layer via communication links. If the execution locations of the antecedent layer and its direct subsequent layer are the same (i.e., both the antecedent layer and its direct subsequent layer are executed by the end device or the edge device), the output of the antecedent layer will be directly fed into its direct subsequent layer. Only when the subsequent layer receives the output of its direct antecedent layer can the subsequent layer be executed. The inference process completes when all DNN layers have been executed.

Appendix I

Typical scenarios for DNN model partition and collaborative execution

(This appendix does not form an integral part of this Recommendation.)

I.1 Smart home

Many smart home applications, such as home security and video surveillance, are part of our home life. These applications need to execute DNN models such as computer vision, video recognition and neural machine translation. However, most home devices, such as mobile phones, cameras and various sensors, are not capable of performing DNN models on their own, resulting in high execution latency and poor user experience. End-edge collaborative DNN execution is suitable to apply in this scenario. Edge servers and base stations are edge devices and cooperate with home devices to execute the DNN models, enabling to achieve lower execution latency and protect personally identifiable information (PII) in the meantime.

I.2 Industry manufacture

The intelligent demand of industrial manufacture is increasing. Intelligence is the key to realize digital transformation of industrial manufacture. The industrial process requires data analysis and intelligent decision capabilities, which need to depend on various intelligent applications and corresponding DNN models. For example, intelligent monitoring, image recognition and other applications are applied to monitor the production safety situation and eliminate security risks. In order to guarantee the safety and timeliness of intelligent industrial production, it is required to implement collaborative DNN execution through DNN model partition.

I.3 Intelligent transportation

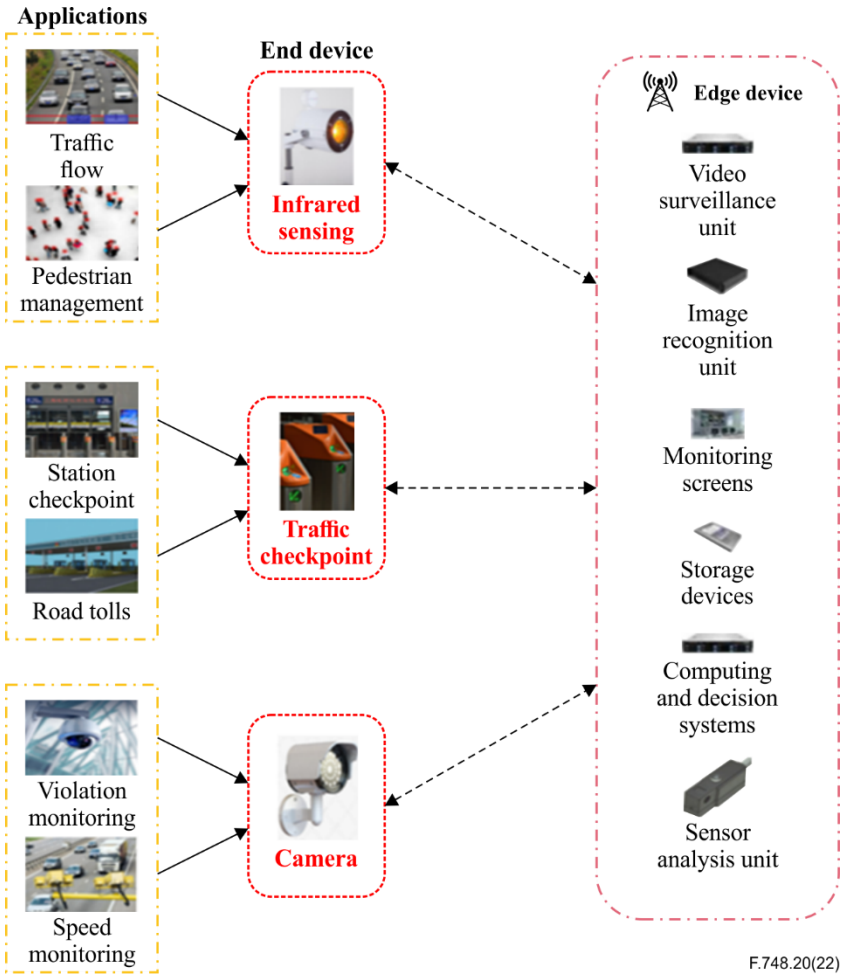

Figure I.1 – DNN model partition and collaborative execution in intelligent transportation application

Many intelligent traffic applications, such as traffic flow monitoring, pedestrian management, station intelligent checkpoint, road tolls, violation and speed monitoring, are common in our life. These applications need to execute DNN models realize image recognition, semantic segmentation, such as face recognition, licence plate recognition, behaviour monitoring and traffic flow management.

Intelligent transportation applications require various kinds of intelligent end device, such as infrared sensing, traffic checkpoint and intelligent camera, which have some computing power. However, the current execution method depends on edge devices only. Edge devices have to independently process a huge amount of data, bringing great pressure to the hardware. Traditional DNNs do not fully perform in terms of the computing and analysis ability of end devices. Therefore, the system has a relative low collaboration efficiency and high execution latency. In order to fully utilize the computing ability of each device, DNN model partition and collaborative execution strategies are applied in allocating some DNN layers from edge devices to the aimed end device. End and edge devices handle computing and analysis tasks collaboratively, which improves the execution efficiency. By deploying suitable partition strategies for end–edge collaboration, the intelligent traffic system (shown in Figure I.1) satisfies the needs of intelligent transportation applications.

Some typical intelligent transportation applications are as follows. a) Infrared sensors together with other sensors and large screens could monitor the vehicles and pedestrian flow, b) traffic checkpoint collaborates with image recognition unit capture and compute massive kinds of information on persons and vehicles for traffic management, c) cameras and storage devices, computing and decision systems collaborate to monitor violations and speed.

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
Series E	Overall network operation, telephone service, service operation and human factors
Series F	**Non-telephone telecommunication services**
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	Telephone transmission quality, telephone installations, local line networks
Series Q	Switching and signalling, and associated measurements and tests
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems

Published in Switzerland
Geneva, 2023