ITU KALEIDOSCOPE
ONLINE2020

7-11 December 2020

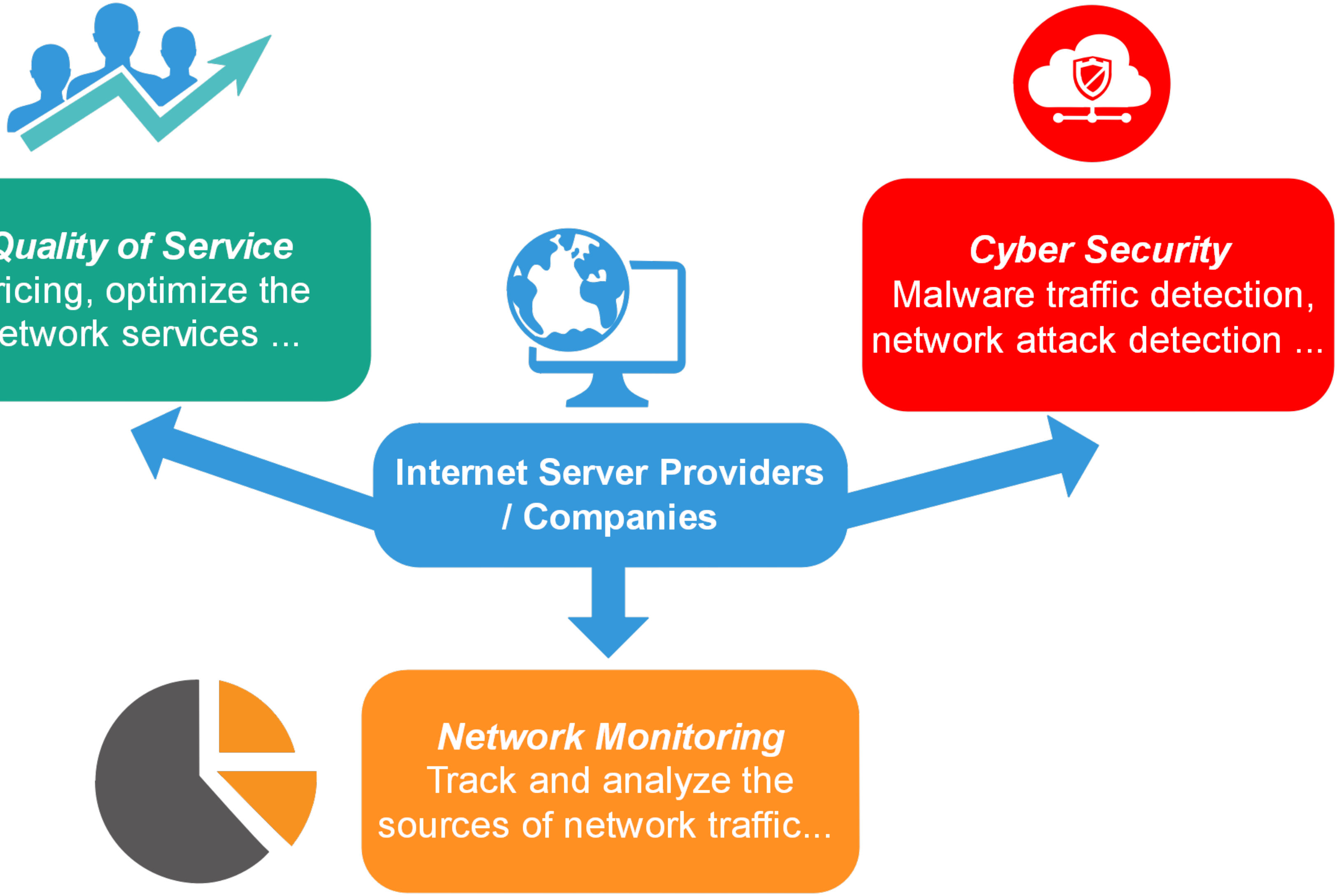
PERT: Payload Encoding Representation from Transformer for Encrypted Traffic Classification

Hong Ye He (presenter), Zhi Guo Yang, Xiang Ning Chen
Zhongxing Telecommunication Equipment (ZTE) Corporation

Session 8: Security in industrial applications

Paper S8.1

1. Traffic Identification / Classification



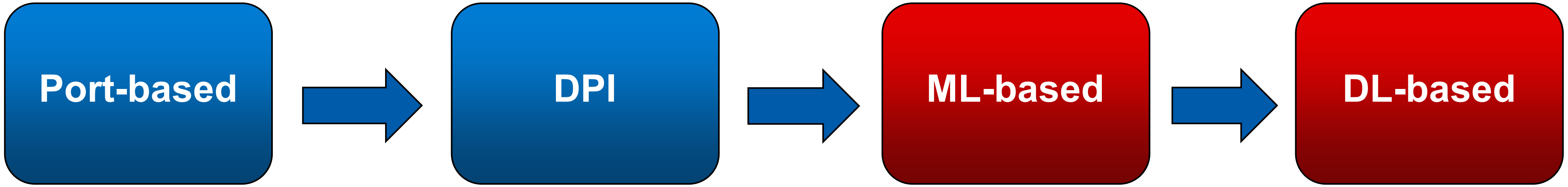
Internet Server Providers / Companies

Quality of Service
Pricing, optimize the network services ...

Cyber Security
Malware traffic detection, network attack detection ...

Network Monitoring
Track and analyze the sources of network traffic...

2. Traffic Identification - Methods



Port-based → DPI → ML-based → DL-based

- The **port-based** and **deep packet inspection** methods that locate fixed patterns from traffic data.

- Rule-based methods that rely on unencrypted information. Not suitable for encrypted traffic.

- The **machine learning** methods that extract hand-designed features from traffic.

- The **deep learning** methods that perform representation learning on raw traffic bytes.

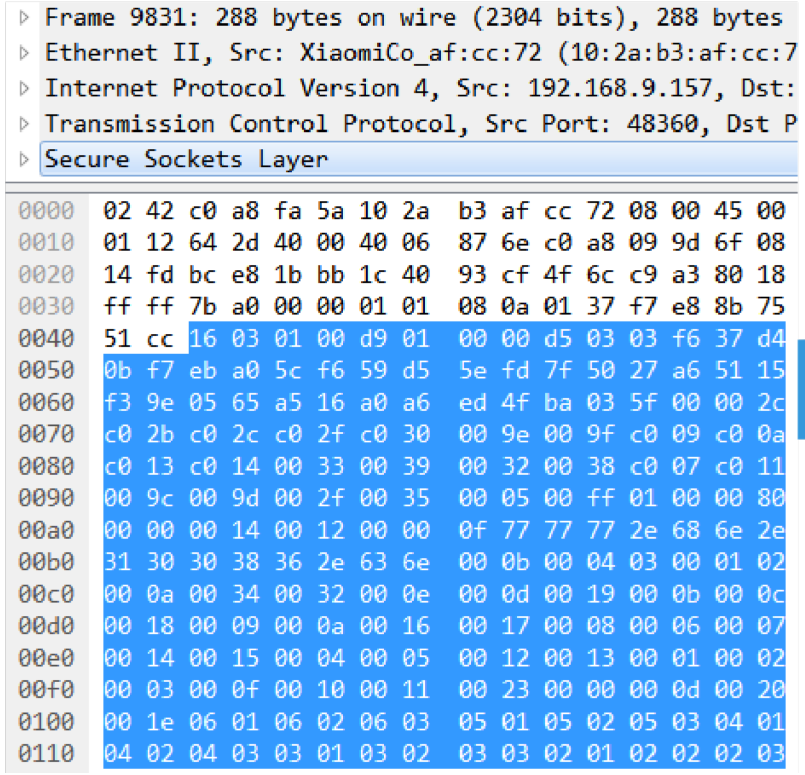
- Extract common traffic features. Ideal for encrypted traffic identification.

3. Deep Learning Based Method - Image Processing

- Current popular DL-based method transform the **raw payload bytes** of traffic packets / flows to grayscale images.

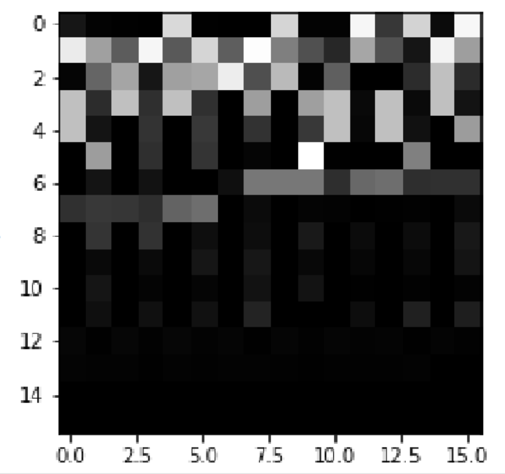
- The purpose is to introduce **image processing** with neural network such as CNN.

- Thus, classification effectiveness is decided by the **representation learning** capacity of the network.



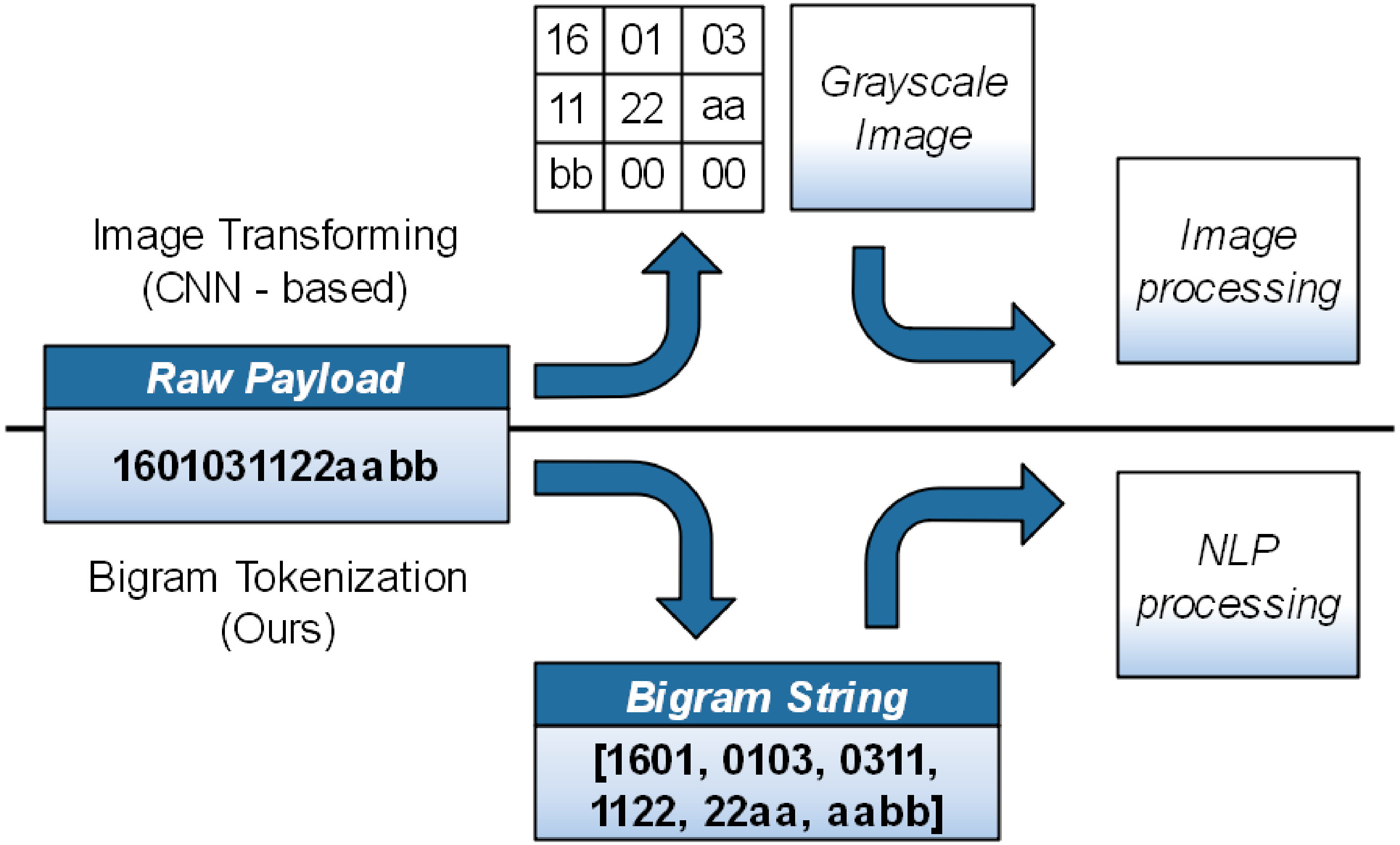
```
▷ Frame 9831: 288 bytes on wire (2304 bits), 288 bytes
▷ Ethernet II, Src: XiaomiCo_af:cc:72 (10:2a:b3:af:cc:7
▷ Internet Protocol Version 4, Src: 192.168.9.157, Dst:
▷ Transmission Control Protocol, Src Port: 48360, Dst P
▷ Secure Sockets Layer

0000  02 42 c0 a8 fa 5a 10 2a  b3 af cc 72 08 00 45 00
0010  01 12 64 2d 40 00 40 06  87 6e c0 a8 09 9d 6f 08
0020  14 fd bc e8 1b bb 1c 40  93 cf 4f 6c c9 a3 80 18
0030  ff ff 7b a0 00 00 01 01  08 0a 01 37 f7 e8 8b 75
0040  51 cc 16 03 01 00 d9 01  00 00 d5 03 03 f6 37 d4
0050  0b f7 eb a0 5c f6 59 d5  5e fd 7f 50 27 a6 51 15
0060  f3 9e 05 65 a5 16 a0 a6  ed 4f ba 03 5f 00 00 2c
0070  c0 2b c0 2c c0 2f c0 30  00 9e 00 9f c0 09 c0 0a
0080  c0 13 c0 14 00 33 00 39  00 32 00 38 c0 07 c0 11
0090  00 9c 00 9d 00 2f 00 35  00 05 00 ff 01 00 00 80
00a0  00 00 00 14 00 12 00 00  0f 77 77 77 2e 68 6e 2e
00b0  31 30 30 38 36 2e 63 6e  00 0b 00 04 03 00 01 02
00c0  00 0a 00 34 00 32 00 0e  00 0d 00 19 00 0b 00 0c
00d0  00 18 00 09 00 0a 00 16  00 17 00 08 00 06 00 07
00e0  00 14 00 15 00 04 00 05  00 12 00 13 00 01 00 02
00f0  00 03 00 0f 00 10 00 11  00 23 00 00 00 0d 00 20
0100  00 1e 06 01 06 02 06 03  05 01 05 02 05 03 04 01
0110  04 02 04 03 03 01 03 02  03 03 02 01 02 02 02 03
```



4. Introducing NLP Processing

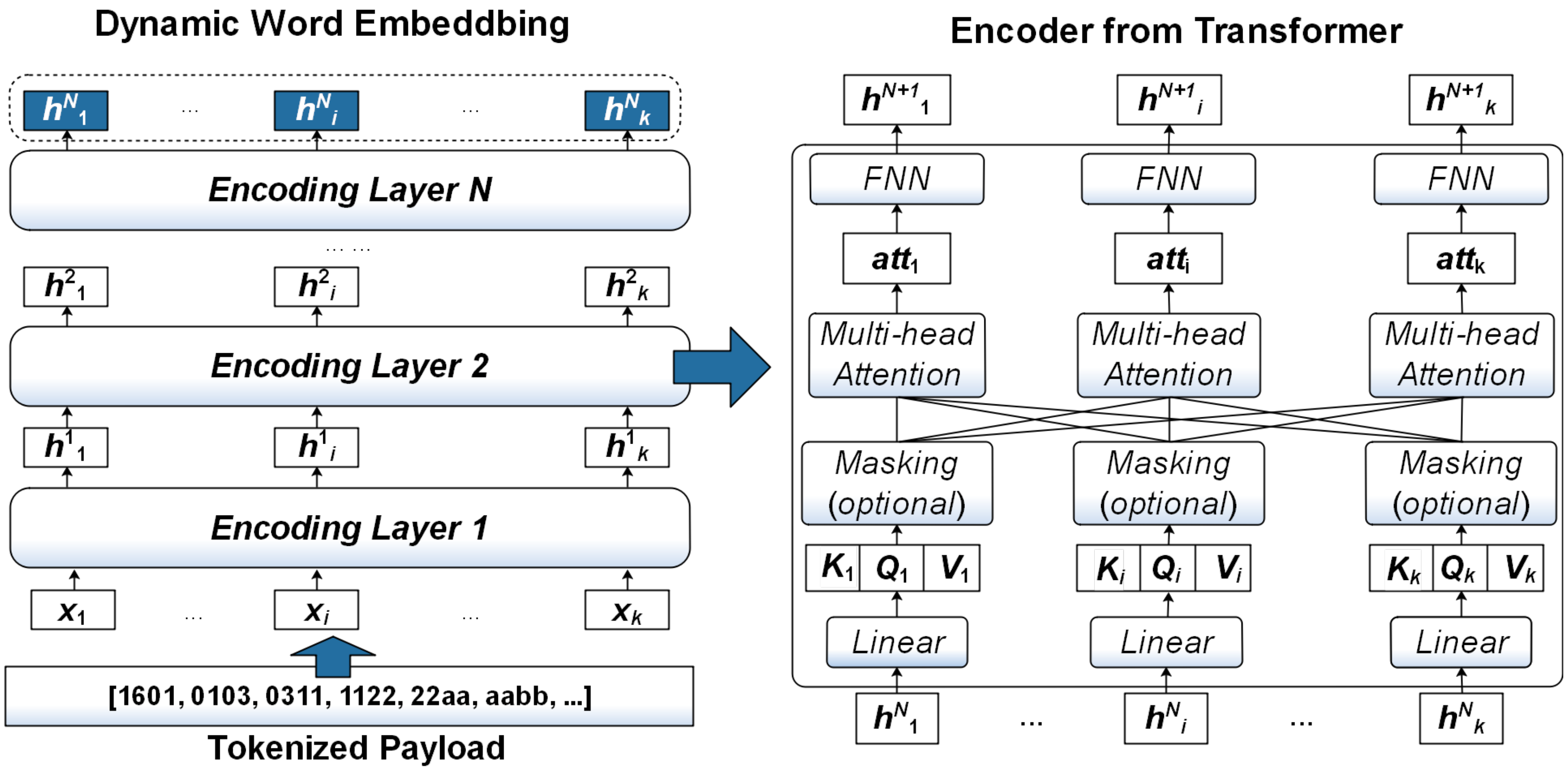
- We perform **bigram tokenization** on encrypted traffic bytes to generate payload bigram strings.
- The traffic identification is transformed to a **NLP classification** task.
- **NLP-related representation learning** can be directly applied to the traffic data.



16	01	03
11	22	aa
bb	00	00

Grayscale Image
Image processing
Image Transforming (CNN - based)
Raw Payload
1601031122aabb
Bigram Tokenization (Ours)
Bigram String
[1601, 0103, 0311, 1122, 22aa, aabb]
NLP processing

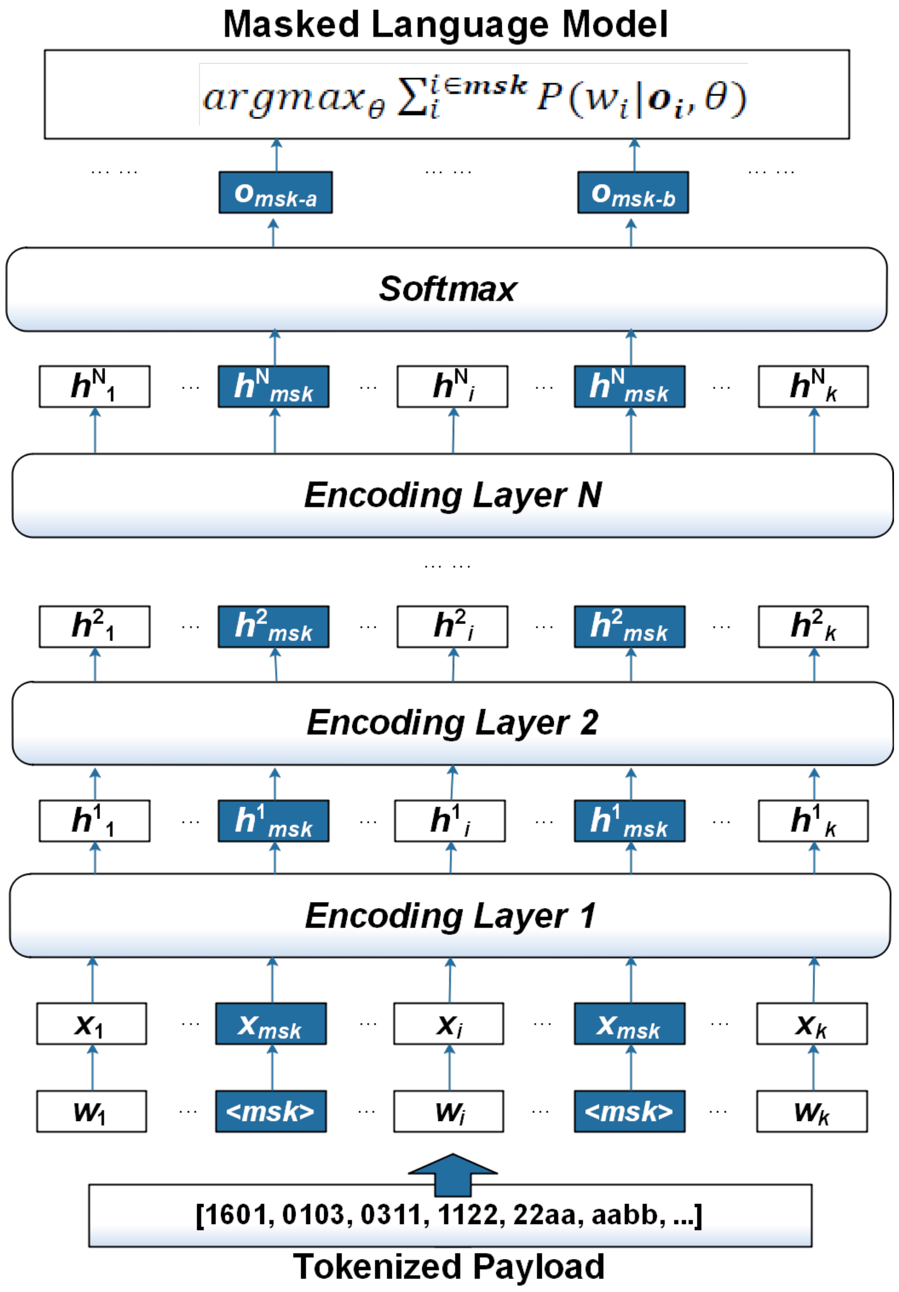
5. Payload Encoding Representation from Transformers (PERT)



- A Bidirectional Encoder Representations from Transformers (BERT) like structure to apply **NLP representation learning** on raw traffic.

6. PERT - Pretraining

- **Language models (LM)** aim to predict words using their contextual inputs.
- LM is originally designed for language generator. But it can also be applied to **initialize NLP encoding network.**
- In a BERT-like network, the **masked language model (MLM)** is frequently utilized for initialization.

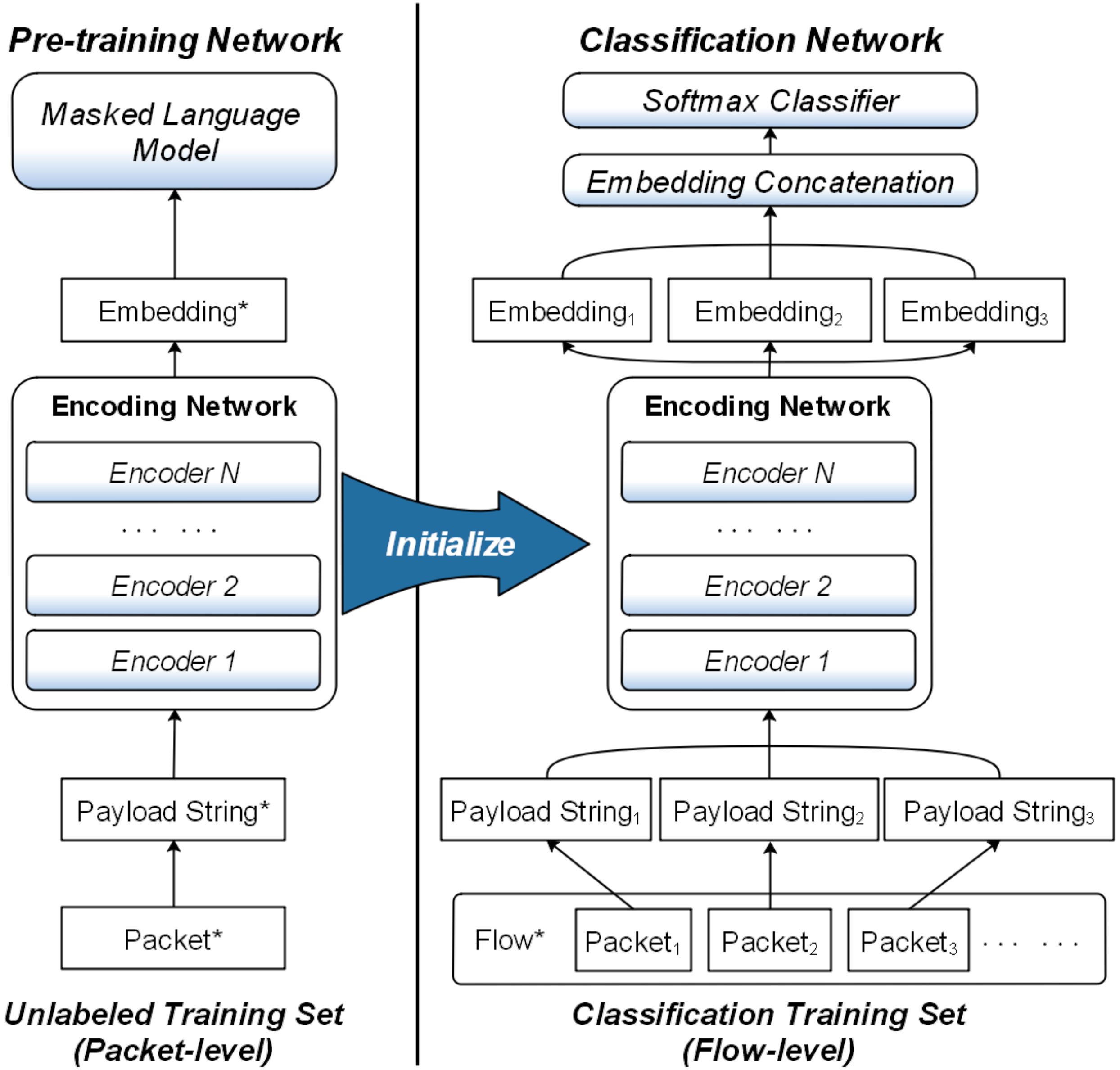

Masked Language Model
$argmax_{\theta} \sum_{i}^{i \in msk} P(w_i|\mathbf{o}_i, \theta)$
$o_{msk\text{-}a}$
$o_{msk\text{-}b}$
Softmax
h^N_1 h^N_{msk} h^N_i h^N_{msk} h^N_k
Encoding Layer N
h^2_1 h^2_{msk} h^2_i h^2_{msk} h^2_k
Encoding Layer 2
h^1_1 h^1_{msk} h^1_i h^1_{msk} h^1_k
Encoding Layer 1
x_1 x_{msk} x_i x_{msk} x_k
w_1 <msk> w_i <msk> w_k
[1601, 0103, 0311, 1122, 22aa, aabb, ...]
Tokenized Payload

7. PERT - Classification

- Encoding network is totally **initialized** by the pre-trained counterpart.

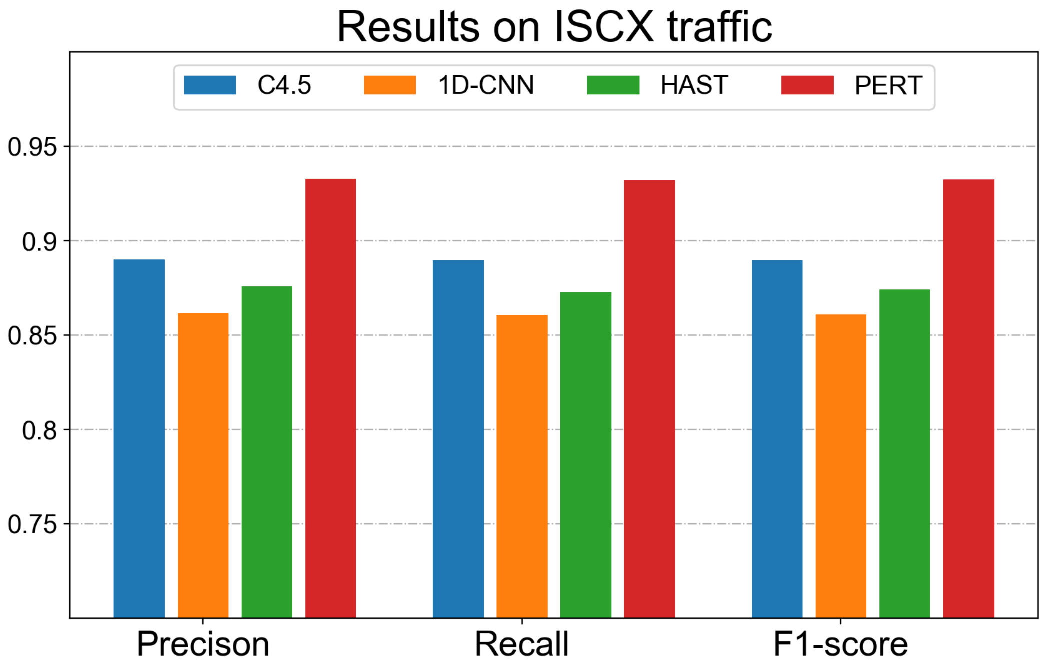
- After applying PERT encoding to the first N packets of a flow, a regular softmax classifier is followed to classify the **concatenated embeddings**.

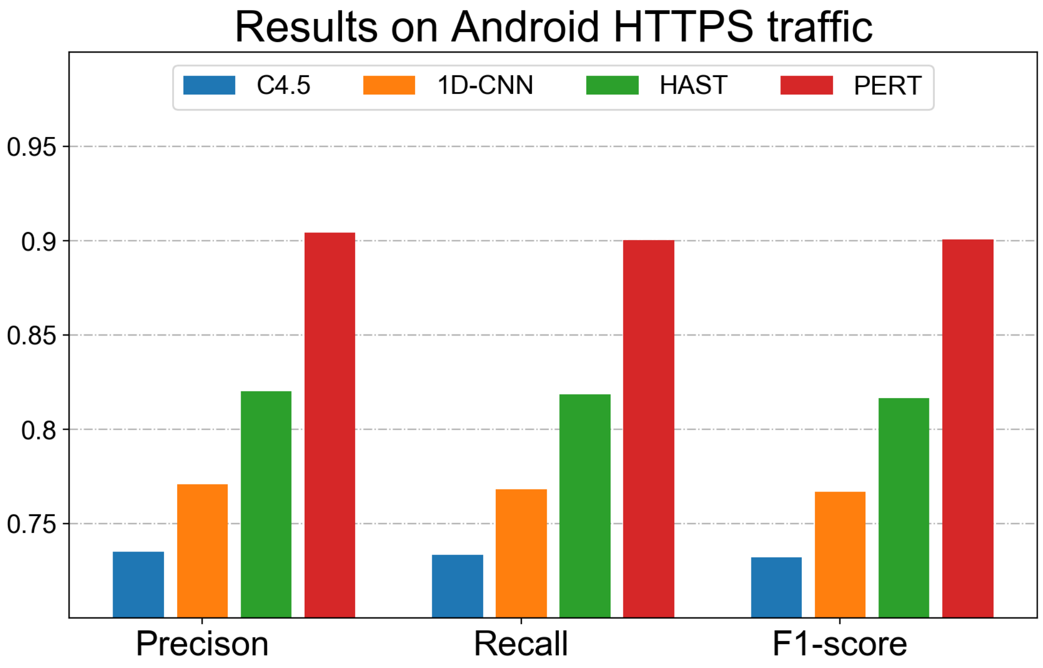
- Encoding network will be further **fine-turned** during back propagation.



ITU KALEIDOSCOPE ONLINE2020

8. Encrypted Traffic Identification Experiments


Results on ISCX traffic


Results on Android HTTPS traffic

- Classification results on **ISCX** traffic (12 classes) and **Android HTTPS** traffic (100 classes).

9. Next Steps

- **More Optimized Encoding Network**

  Follow up the ever-developing BERT research.

- **Flow-level Identification Support**

  Find a better approach to merge the PERT encoded packets.

- **Other NLP Methods**

  Evaluate other NLP methods on tokenized traffic bytes.

ITU KALEIDOSCOPE

ONLINE2020

Thank you!