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

- **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

- 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.
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.
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.
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-tuned during back propagation.
8. Encrypted Traffic Identification Experiments

- 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.
Thank you!