Paper Summary
ET-BERT: A Contextualized Datagram Representation with Pre-training Transformers for Encrypted Traffic Classification

Sarthak | 2020CS10379

June 2024

Problem Statement

The paper addresses the challenge of constructing a generalized and transferable model for encrypted traffic detection without the need of artificial feature engineering or large labeled traffic data

Motivation

- Capturing the implicit and robust patterns in the diverse encrypted traffic and support accurate and generic traffic classification is essential to achieve high network security
- Existing methods require complex feature engineering and extensive data labeling i.e. they require expert experience and large amount of labeled data

Key Idea

To pre-train with unlabeled traffic and transfer efficiently to specific tasks. the proposed self-supervised tasks can effectively capture the relationship between the different layers of traffic.

Framework & Methodology

Datagram Representation - Representation of traffic datagram as model input

1. Traffic traces : The raw flow traces need to be pre-processed to get rid of the meaningless noise data
2. Raw encrypted traffic session : Extracts sessions from the traces and identifies directional information
3. Datagram Extraction of package and filtering of information
4. BURST is generated based on the principle of same direction and continuity
5. Encoding BURST datagram to generate token unit and slicing in half to represent two instances of one message.

Pre-training - Learning contextual knowledge of the content and structure of traffic

1. Masked BURST Model
   - each token in the input sequence is randomly masked with 15% probability.
   - As the chosen token, we replace it with [MASK] at 80% chance
   - ET-BERT is trained to predict tokens at the masked positions based on the context.
2. Same origin BURST prediction
   - Our purpose is to capture the correlation between the packets in BURST
   - A binary classifier is used to predict whether two sub-BURST are from the same BURST origin

Fine-tuning- Adapting scenario specific encrypted traffic

1. We input the task-specific packet or flow representations into the pre-trained ET-BERT and fine-tune all parameters in an end-to-end model.
2. At the output layer, the [CLS] represenation is fed to a multi-class classifier for prediction.

Contributions

- Developed a new encrypted traffic representation model, ET-BERT, which can pre-train deep contextual datagram- level traffic representations from large-scale unlabeled data, then accurately classify encrypted traffic for multiple scenarios with a simple fine-tuning on a small amount of task-specific labeled data.

Strengths

- ET-BERT compares 11 existing methods in 6 different scenarios and achieve the best encrypted traffic identification results
- ET-BERT maintains the most recognition in the face of new encrypted protocol traffic
- Overcomes the requirement of artificial feature engineering or large labeled traffic data
- Comparatively immune to few-shot and unbalanced scenarios and hence generalizable and tranferable

Weaknesses & Limitations

- Randomness differences due to applying different cipher suites
- Dynamic and continuous changes in traffic will bring variations to the sample scenario
- Pre- training model faces the security risks of poisoning