Paper Summary
New Directions in Automated Traffic Analysis

Sarthak | 2020CS10379

June 2024

Problem Statement

The paper addresses the challenge of automating feature selection, model selection and parameter tuning in machine learning pipelines for network traffic analysis.

Motivation

- Effectiveness of applying machine learning to network traffic analysis tasks often depends on the selection and appropriate representation of features
- Every new network detection or classification task requires engineering new features, selecting appropriate models, and tuning new parameters by hand
- Manual extraction may omit features that either were not immediately apparent or involve complex relationships, is painstaking, time-consuming and requires specialized domain knowledge

Key Idea

To create a unified packet representation suitable for representation learning and model training. Then integrate it with AutoML to automate feature extraction and model tuning.

Framework & Methodology

- **Complete** : Includes every bit of a packet header, avoiding the need for domain-specific feature selection
- **Constant Size per Problem** : Ensures input size consistency for ML models

- **Inherently Normalized** : Provides pre-normalized data to simplify ML tasks
- **Aligned** : Ensures that every location in the representation corresponds to the same part of the packet header across all packets
- nPrint is a hybrid of semantic and binary packet representations, representing packets to models as raw binary data, but aligning the binary data in such a way that recognizes that the packets themselves have specific semantic structure
- Headers that do not exist in the packet being transformed are noted with sentinels accordingly, while headers that exist but are not of maximum size are zero padded for alignment across nPrints
- Uses AutoML tools to extract best features, model and hyperparameters

Contributions

- Presented a method to automate many aspects of traffic analysis, making it easier to apply machine learning techniques to a wider variety of traffic analysis tasks
- Designed a standard packet representation, nPrint, that encodes each packet in an inherently normalized, binary representation while preserving the underlying semantics of each packet
- Integrated nPrint with AutoML, automating the entire traffic analysis pipeline
- Released nPrint, nPrintML, and associated datasets to the public to foster further research

Strengths

- Reduces the need for human intervention, making traffic analysis more accessible
- Demonstrated nPrintML's effectiveness across eight different traffic analysis tasks like OS detection, device fingerprinting etc. ; showing superior performance compared to traditional methods in most cases
- The standard format of nPrint makes it easy to integrate network traffic analysis with state-of-the-art automated machine learning

Weaknesses & Limitations

- Cannot capture temporal relationships across multiple traffic flows
- The paper does not explore running nPrint on long traffic sequences
- Focus on packet headers may omit some higher-level traffic patterns that could be relevant for certain analysis tasks