DSCI-565 Project Midterm Report

Evaluating Deep Sequential Models for Flow-Based Network Intrusion Detection Systems

Group3: Jeong Hoon Choi, Yuxuan Liu

Division of Labor

Jeong Hoon Choi: Replicate results in the papers
Yuxuan Liu: Data EDA Analysis

Dataset Description

We decided to use two datasets: **NF-UNSW-NB15** and **NF-CSE-CIC-IDS2018**. The two NetFlow (version 2) datasets are preprocessed from the original UNSW-NB15 and CSE-CIC-IDS2018 data sets, with missing values, duplicated rows, etc., and 43 identical features (with two different label columns: label, attack).

Many existing ML/DL-based NIDS systems use network packets of flow data records as isolated entities. However, the authors of FlowTransformer approach flows as **continuous sequences**. This approach allows them to present a model capable of analyzing temporal behavior, which has been difficult to detect using existing methods.

	Data Size	# of instances	# of attributes	# of labels (Attack)
NF-UNSW-NB15-v2	421M	2,390,275	43	9
NF-CSE-CIC-IDS-2018-v2	3.0G	18,893,708	43	7

Each dataset is highly unbalanced, with 96.02% and 88.05% having benign labels.

Exploratory Data Analysis

For all data of NF-UNSW-NB15-v2, 2.4M instances of NF-CSE-CIC-IDS-2018-v2 were extracted and EDA was performed. Unlike the original dataset, the NetFlow dataset has been preprocessed to remove missing values and duplicated data.

- Label Distribution

Label Distribution	Label = 0 (Benign)	Label = 1 (Attack)
NF-UNSW-NB15-v2	2,295,222(96.02%)	95,053 (3.98%)
NF-CSE-CIC-IDS-2018-v2	16,635,567 (88.05%)	2,258,141 (11.95%)

- NF-UNSW-NB-15-v2
Notebook: https://colab.research.google.com/drive/1yMnd4Z5D1X2uNatTdoYG-zaJHgpClr4P?usp=sharing

The dataset has 2,390,275 records with 45 features, showing severe class imbalance. Features are heavily right skewed with extreme outliers (traffic values up to 1e7). Generic and Reconnaissance attacks show the most distinctive patterns. Strong feature correlations exist (e.g., IN_BYTES/IN_PKTS > 0.8).

- NF-CSE-CIC-IDS2018-v2
Notebook: https://colab.research.google.com/drive/12PectoiYCme30HaF2JoE1FuIdJ-S6Fdm?usp=sharing

Sampled 2,400,000 from 18,893,708 records (12.7%) with 45 features. Severe class imbalance. Features heavily right-skewed with extreme outliers (OUT_BYTES up to 2e8). DDoS attacks-LOIC-HTTP show highest variation. TCP/UDP dominant protocols; Port 53(DNS) has most traffic. Strong correlations (IN_BYTES/IN_PKTS > 0.8).

Experiments Description

Notebook: https://colab.research.google.com/drive/1idMxH9ab2TSZFtAzqYLrBRbw6nXPSn0N?usp=sharing

```
'CLIENT_TCP_FLAGS', 'L4_SRC_PORT', 'TCP_FLAGS', 'ICMP_IPV4_TYPE',
'ICMP_TYPE', 'PROTOCOL', 'SERVER_TCP_FLAGS', 'L4_DST_PORT', 'L7_PROTO'
```

In the FlowTransformer paper, the nine features of the Netflow-v2 data are defined as **categorical features**, one-hot encoded, and input data with 268 dimensions, along with 28 other **numerical features**, are used as input to the embedded layer.

Numerical: 28
Categorical: 16 + 32 + 17 + 32 + 32 + 32 + 15 + 32 + 32 = 240
drop 6 columns

We decided to use the above preprocessing as is.

To reproduce the project, we used the Framework library provided by the author.
The paper proposes a Transformer-based model designed to capture sequential dependencies in flow network data by leveraging contextual information - an aspect that previous deep learning models failed to learn effectively. The proposed FlowTransformer architecture is modular, consisting of three components: an **Input Encoder**, a **Transformer**, and a **Classification Head**, allowing each module to be easily replaced or tested independently.

We implemented the above project in a Jupyter Notebook and archived an F1 score of 0.93 through training. The model was trained using only a subset of the dataset by adjusting the steps_per_epoch, and it employed a 2-layer Transformer architecture. The input sequence window size used was 8. In addition, it performed binary classification with two labels: Benign and Attack

Potential Challenges and Future Directions

Most existing deep learning models trained on network flow data already achieve high levels of accuracy. However, this often masks underlying challenges such as data imbalance and high false negative and false positive rates, which remain persistent issues in network intrusion detection. As suggested by the authors, modeling network flows as sequences enables the system to capture contextual dependencies within traffic patterns - an ability that traditional models lack. For example, attacks such as Slowloris (slow DoS) and infiltration (low-volume exfiltration that mimics normal traffic) can be very difficult to detect if look at the individual flows in isolation.

In this project, (1) we aim to develop models capable of capturing such previously difficult-to-detect patterns by employing **various sequence-based deep learning architectures**, including RNN, LSTM, GRU, and Transformer models. Alternatively, (2) we are considering enhancing the FlowTransformer architecture to more **effectively identify attack labels that conventional models fail to detect**. (Not yet decided on one of the two purposes)

Reference

[1] Manocchio, L., Layeghy, S., Lo, W. W., Kulatilleke, G. (2024). FlowTransformer: A Transformer Framework for Flow-based Network Intrusion Detection Systems, https://doi.org/10.1016/j.eswa.2023.122564

[2] Sarhan, M., Layeghy, S., & Portmann, M. (2022). Towards a standard feature set for network intrusion detection system datasets., https://arxiv.org/abs/2101.11315