DSCI-565 Project Proposal

Group 3
Jeong Hoon Choi, Yuxuan Liu

Abstract

This project explores deep learning approaches for network traffic anomaly detection. We aim to build baseline models using fully connected and convolutional networks, and extend to Transformer-based architectures that capture temporal dependencies in traffic data. We will use the CIC-IDS-2017 dataset, a widely adopted benchmark for intrusion detection. Expected outcomes include performance benchmarks, ablation studies, and insights into the benefits of sequence modeling.

Introduction

Network anomaly detection plays a critical role in cybersecurity, enabling timely identification of malicious or abnormal activities within large-scale network environments. Traditional methods often struggle with high-dimensional, imbalanced, and dynamic traffic data. This project investigates deep learning methods to improve detection performance, particularly by capturing sequential dependencies in traffic flows.

Project Objectives

- Establish baseline performance with feed-forward and CNN-based models.
- Develop Transformer-based models to incorporate sequential characteristics of network traffic.
- Compare results with existing studies and conduct ablation studies for deeper insights.

State-of-the-Art

Prior research has applied CNNs and RNNs to intrusion detection tasks with moderate success. Recent works explore Transformer models due to their ability to capture long-range dependencies. The CIC-IDS-2017 dataset is a standard benchmark containing millions of records across benign and attack traffic. Notable works include Zhang et al. (2025) on Transformer-based intrusion detection (arXiv:2506.19877), Taha & Mustafa (2024) on deep learning anomaly detection (MDPI Electronics 14(1):189), and practical CNN implementations on Kaggle.

Dataset Description

We will use the CIC-IDS-2017 dataset, which contains labeled network traffic including benign flows and multiple attack types such as DDoS, brute force, and botnet activity. The dataset provides over 80 features per flow and millions of records. Preprocessing steps will include normalization, balancing, and potential feature selection.

Approach & Plan

Stage 1: Baseline models using fully connected networks and CNN classifiers to treat records independently.

Stage 2: Advanced models using Transformer architectures to explicitly capture temporal dependencies.

We will replicate results from existing works, then conduct ablation studies and extend with architecture variations. Pipeline: Data → Preprocessing → Baseline Models → Transformer Models → Evaluation.

Potential Risks

Risks include computational limitations for large-scale Transformer training. Mitigations involve reducing dataset size, employing dimensionality reduction, or leveraging transfer learning. Overfitting may be addressed with dropout, regularization, and cross-validation. Data imbalance will be managed with resampling techniques.

Division of Labor

- Both: Baseline model implementation (Stage 1), dataset preprocessing.
- Both: Transformer-based model development (Stage 2), experiments, and report writing.
- Both: Literature review, reports, and presentation.

Expected Outcomes

We expect to deliver benchmark performance results for CNNs and Transformers on the CIC-IDS-2017 dataset, demonstrating the benefits of time-series modeling. Ablation studies and comparative evaluations will provide insights into which model architectures are most effective for network anomaly detection.

References

1. Zhang, X., Wang, Y., & Li, H. (2025). Transformer-based intrusion detection for large-scale network traffic. [2506.19877] Robust Anomaly Detection in Network Traffic: Evaluating Machine Learning Models on CICIDS2017

2. Taha, M., & Mustafa, Y. (2024). Deep learning for anomaly-based intrusion detection systems. Electronics, 14(1), 189. https://doi.org/10.3390/electronics14010189

3. Nolovelost. (n.d.). CIC-IDS-2017 CNNs v1.0 [Kaggle notebook]. CIC-IDS-2017-CNNs-v.1.0