[8] Data Preprocessing Pipeline

The method for creating data in NetFlow format is the same as extracting data from PCAP data explained before. Using software called nProbe, it groups the packet into flows with 5-tuple source IP, destination IP, source port, destination port, Protocol.

There are currently three versions of NetFlow formats, up to version 3.
Version 2 which we used for the project extract 43 features from PCAP.

Afterwards, labels are matched using Ground Truth provided separately from PCAP.

NetFlow format includes basic processes such as missing value removal, duplicate value removal, and scaling.

---

[9] Datasets

We used UNSW and CIC-IDS data in NetFlow format provided by the University of Queensland. Each data set contains around 400MB and 3GB of data each and has identical features.

UNSW has nine attack patterns, while CIC-IDS has seven attack patterns divided into 14 sub patterns.

Both datasets share a significant problem, the Benign label accounts for 96% and 88% of the data, resulting in significant data imbalance. Also, they contain very little data on specific attack types.

---

[10] Attack Types: NF-UNSW-NB15 (1)

The first is UNSW's attack types. Due to the severe label imbalance, labels with extremely small data sets, such as worms, are difficult to learn.
However, aside from this, Analysis, DoS, and Reconnaissance attacks are generally known to result in relatively difficult to predict in Machine learning or Deep Learning model that use single flow as input.

---

[11] Attack Types: NF-UNSW-NB15 (2)

Our baseline FCN model, which will be introduced later, also showed low accuracy for first two attacks. In particular, it shows a very high False Negative rate for Analysis attacks.

Recon showed better results than expected, and I personally think it is because the Port Scanning generates traffic on destination ports that are not used in the other attacks. (So not difficult to catch most of them)

---

[12] Attack Types: NF-CSE-CIC-IDS2018 (1)

The second dataset is CIC-IDS2018.
Excluding Web Attacks, which have too few labels, the classes known to be difficult to predict with single-flow input models are Brute-Force and DDoS attacks.

---

[13] Attack Types: NF-CSE-CIC-IDS2018 (2)

Our baseline FCN also showed low accuracy.
While it detected almost all Brute Force attacks against specific protocols like FTP and SSH, it seemed unable to distinguish attacks targeting HTTP and HTTPS from Benign traffic.

Most of DDoS attacks were relatively well-detected, HOIC DDoS Attacks were largely missed.
Unlike LOIC, HOIC doesn't simply increase traffic, instead, it uses seemingly normal HTTP GET/POST requests, leading to its classification as Benign

---

[14] Related Work - Flow Transformer (1)

Our project references Manocchio's paper Flow Transformer, which proposes a model for flow-level network traffic and a Transformer-based architecture.

He said that by using Transformer's self-attention, model can understand the Context of NetFlow sequences and predict labels that were previously difficult to predict in models that took a single flow as input.

---

[15] Related Work - Flow Transformer (2)

In this paper, we perform embedding on 9 Categorical features, including features like Port, Protocol used in NetFlow format.

The paper transforms data into vectors of up to 32 sizes, depending on the feature.

We also converted our data using the same method. The UNSW and CIC-IDS data now have 269 and 291 features each. SEP

[21] NetFlow BERT Architecture

This is a NetFlow BERT model. The Transformer Encoder has two multi-head Self attentions, and the Feed Forward Network hidden size is set to 256.

[22] NetFlow BERT Architecture

The BERT model has a 64-size embedding layer and Four Transformer Encoders.
With a total of 280,000 (280 K) parameters, it has the fewest weights among the models used in the comparison.

[23] Model Comparison Analysis

We trained three different models, BaselineFCN, CNN+LSTM, and BERT, on two datasets UNSW and CIC-IDS, and compared the results.

To compare the strengths and weaknesses of each models, we compared models with similar sized weights.

[24] Model Performance Comparison (NF-UNSW-NB15)

These are total performance on the UNSW data and Recall and F1 score for labels that were difficult to predict by other models.

While overall accuracy is similar, Macro F1 improves in the order of BaselineFCN, CNN+LSTM, and BERT.
Furthermore, performance against Analysis, DoS, and Recon attacks also improves in order.
In particular, Analysis shows a significant improvement.

[25] Model Performance Comparison (NF-CIC-IDS2018)

The results on the CIC-IDS data were slightly different from what we expected.
Accuracy improved in the order of BaselineFCN, CNN+LSTM, and BERT, but CNN+LSTM achieved a lowest Macro F1 score.

And, BaselineFCN showed very low accuracy on Benign labels, while CNN+LSTM performed best against Brute Force and DDoS attacks.

Personally, I think the BERT model underfits because it has too few weights and the Attention Head is too small to catch HOIC DDoS attack.

---

[26] NetFlow Bert Further Training

Finally, to improve performance, we increased the BERT model size and test it.
We used 8 Attention Heads, Encoder Feed Forward internal size of 1024, and 8 Transformer Encoder layers. This resulted in a roughly 30 times increase in parameter number than previous. Additionally, to improve the Macro F1 score, we used Sparse Categorical Focal Loss with existing class weight learning.

Overall, we were able to achieve significantly improved performance. In particularly, we were able to catch most CIC-IDS HOIC DDoS attacks. SEP
