Investigation of IDS Transfer Learning with MLP Networks

Dan Popp
4/18/22

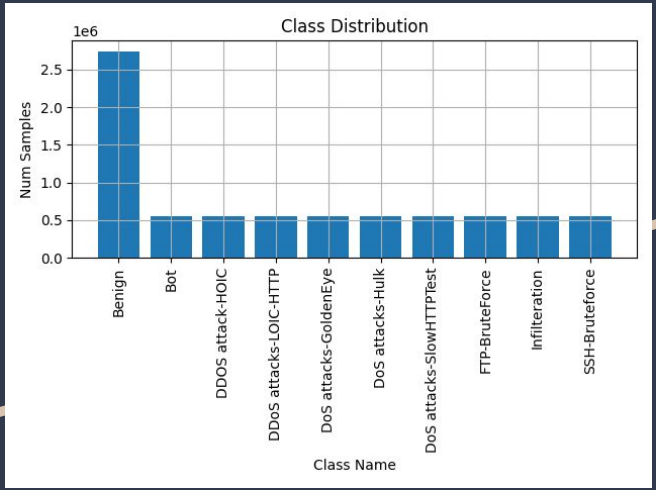
Problem

- **Goal**: Test how well different ML models generalize across cybersecurity datasets.
  - Test MLP model transfer learning
  - Use Random Forest as a baseline
  - Test on CIC-IDS-2018 and CIC-IDS-2017 datasets
- Test if models can generalize
  - Test performance when trained on one dataset and generalized on the other
  - For MLP model, we will evaluate different transfer learning approaches
  - Goal is to test is learned features from one dataset can be applied to the other.

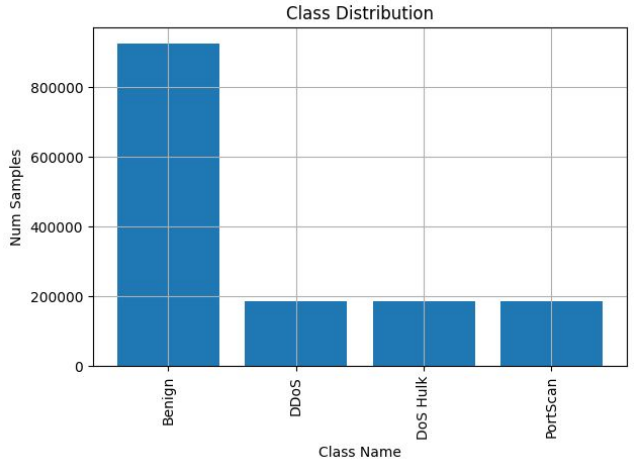
Data

- Data
  - CIC-IDS-2018 & CIC-IDS-2017
  - Both datasets contain data from simulated attacks
  - Both datasets have attributes extracted via CICFlowMeter
  - Both datasets share **some** classes
- Attributes
  - CIC-IDS-2018: 79 attributes
  - CIC-IDS-2017: 77 attributes
  - ‘Timestamp’ and ‘Protocol’ are unique to CIC-IDS-2018

Preprocessing



- Invalid Data (NaN and Inf):
  - Invalid Ratio: 0.01%
- Minority classes have 20% samples of Benign class
  - Undersampled benign data (~10x reduction)
  - Dropped extreme minority classes (< 1% benign)
  - Oversampled minority classes (5x-20x increase)
  - Used random under and oversampling



MLP Architecture

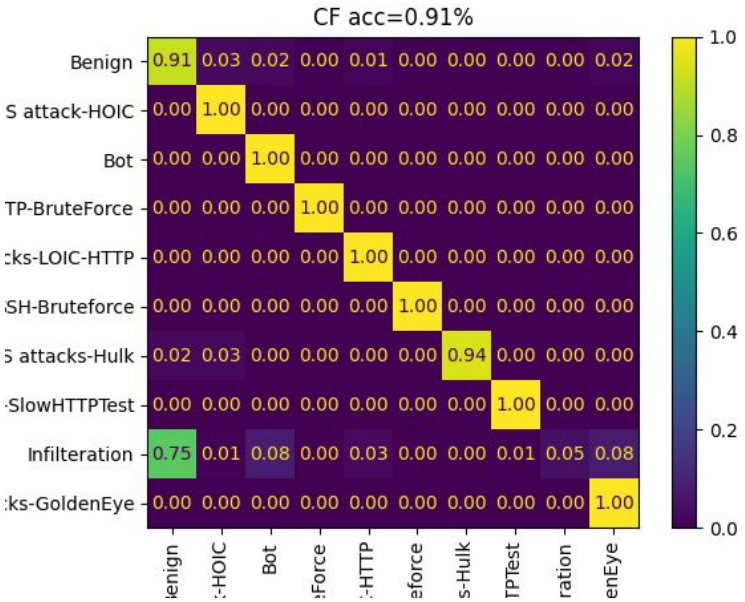
```
MLP(
  (layer1): Linear(in_features=77, out_features=100, bias=True)
  (layer2): Linear(in_features=100, out_features=200, bias=True)
  (layer3): Linear(in_features=200, out_features=500, bias=True)
  (layer4): Linear(in_features=500, out_features=200, bias=True)
  (layer5): Linear(in_features=200, out_features=100, bias=True)
  (fc): Linear(in_features=100, out_features=10, bias=True)
  (act): ReLU()
  (softmax): Softmax(dim=0)
```

- 77 input attributes
  - Removed protocol and timestamp to match datasets
- Embed into 100 dimensional feature-space
- Classify into 10 or 4 target classes

Training Procedure

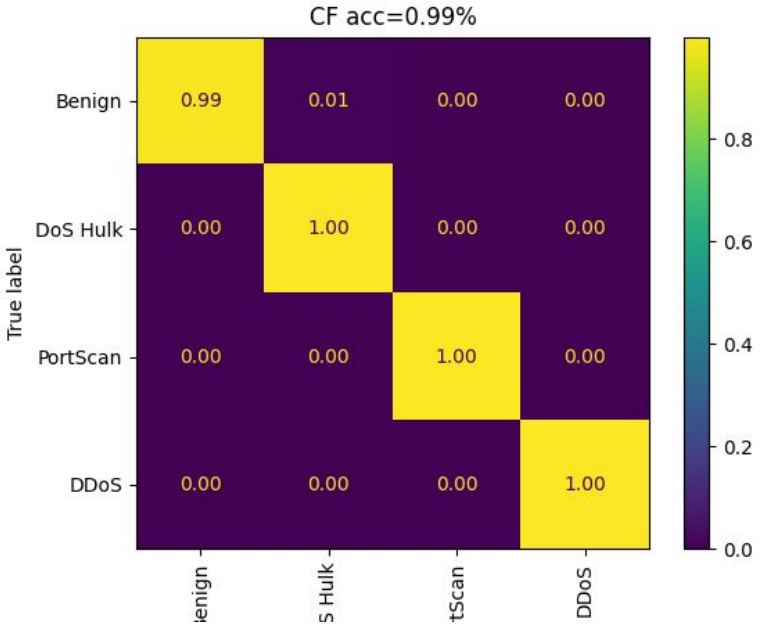
- Source Training
  - Train on source dataset (CIC-IDS-2017 and CIC-IDS-2018)
- Freeze Feature Extractor
  - Start with model trained on source dataset and transfer learning to target dataset
  - Freeze hidden network layers to preserve feature embedding logic learned from source dataset
  - Adapt learned features directly to target dataset
- Fine-Tune Network
  - Take source model and retrain the entire dataset on the target domain.
  - Allows modification of feature embedding for target dataset
  - Slower to train than previous technique

2018 MLP Results



- 91.5% total accuracy
  - Dropped from 94.5% after dropping timestamp and protocol attributes
- Largest issue in Infiltration class
  - Hard to classify with RF as well

2017 MLP Baseline Results



- 99.03% Accuracy
  - No **Infiltration** class to lower accuracy score
- Trained directly on CIC-IDS-2017 dataset

Random Forest Baseline

	precision	recall	f1-score	support
Benign	1.00	0.95	0.97	2697457
Bot	1.00	1.00	1.00	56947
DDOS attack-HOIC	1.00	1.00	1.00	136708
DDoS attacks-LOIC-HTTP	0.99	1.00	0.99	115467
DoS attacks-GoldenEye	1.00	1.00	1.00	8395
DoS attacks-Hulk	1.00	1.00	1.00	92371
DoS attacks-SlowHTTPTest	1.00	1.00	1.00	27762
FTP-BruteForce	1.00	1.00	1.00	38848
Infilteration	0.18	0.90	0.30	32136
SSH-Bruteforce	1.00	1.00	1.00	37747
accuracy			0.96	3243838
macro avg	0.92	0.98	0.93	3243838
weighted avg	0.99	0.96	0.97	3243838

CIC-IDS-2018

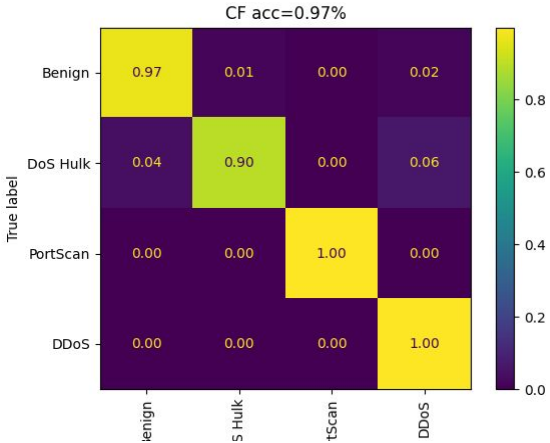
- Compared to 91.5% Accuracy for MLP

	precision	recall	f1-score	support
Benign	1.00	1.00	1.00	454463
DDoS	1.00	1.00	1.00	25609
DoS Hulk	1.00	1.00	1.00	46115
PortScan	1.00	1.00	1.00	32010
accuracy			1.00	558197
macro avg	1.00	1.00	1.00	558197
weighted avg	1.00	1.00	1.00	558197

CIC-IDS-2017

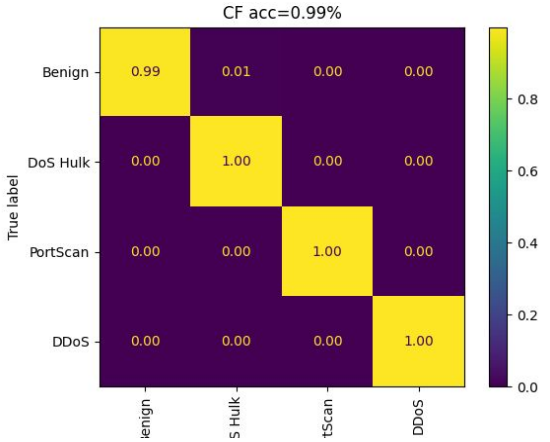
- Compared with 99.03% Accuracy for MLP

2018 -> 2017 Transfer



Freeze Feature Extractor

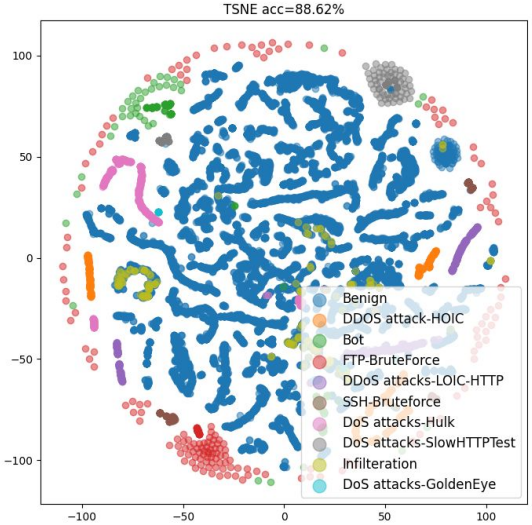
- 96.6% Accuracy
- Trained 5 epochs



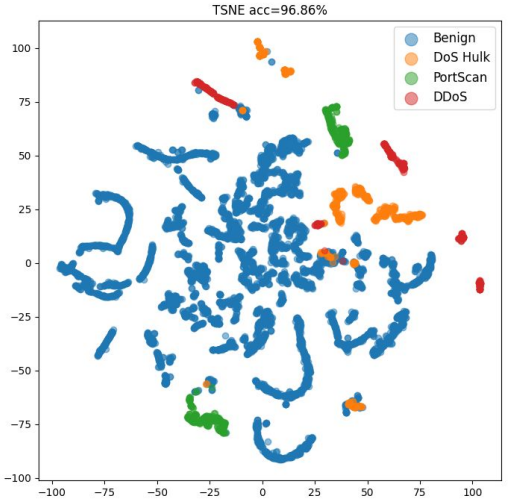
Fine-Tune Network

- 99.46% accuracy
- Trained 10 epochs

2018 -> 2017 Features

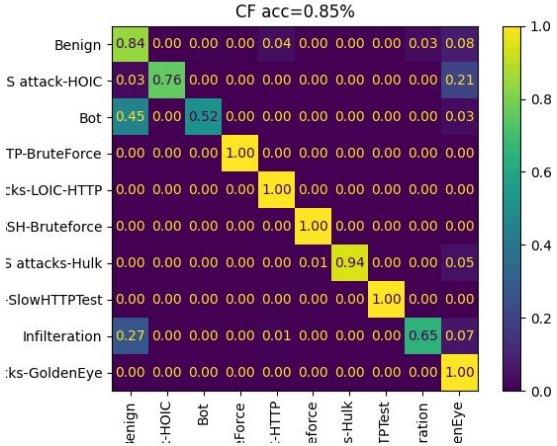


CIC-IDS-2018
Source Model t-SNE



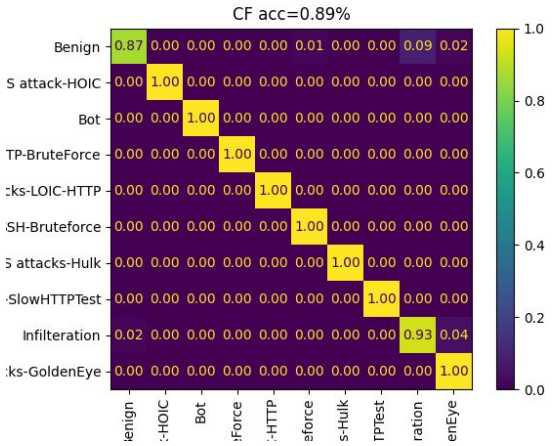
CIC-IDS-2017
Transfer Learning (Freeze) t-SNE

2017 -> 2018 Transfer



Freeze Feature Extractor

- 85.72% Accuracy
- Really quick training on DDoS classes
- CF Matrix is normalized by row.
  - Value shown is recall



Fine-Tune Network

- 89.09% Accuracy
- Infiltration has high recall but low precision

Conclusion

- MLP models perform equally to slightly worse than random forest models on the given datasets
- The features learned from one IDS dataset can be meaningfully applied to another IDS dataset for efficient classification
    - Performs worse for classes different than in source dataset
    - Inconclusive results for whether MLP pre-training can improve target dataset performance
- MLP models may still be overfitting

Questions?