Explainable Network Intrusion Detection Using External Memory Models

Jack Hutchison, Duc-Son Pham, Sie Teng Soh and Huo Chong Ling

Curtin University

Contributions

This paper makes network intrusion detection through augmenting an autoencoder-neural network model with external memory more explainable:

- Explores the effect of the memory size and the addressing scheme
- Explores which memory slots are strongly matched with what classes
- Measures how much external memory each class takes to be properly encoded
- Demonstrates that memory contents can help identify seen and unseen classes, potentially addressing zero-day attacks.

External Memory Models

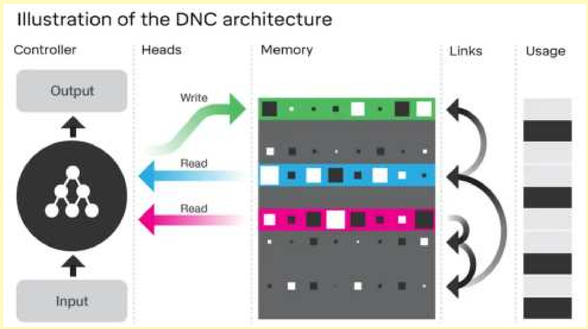
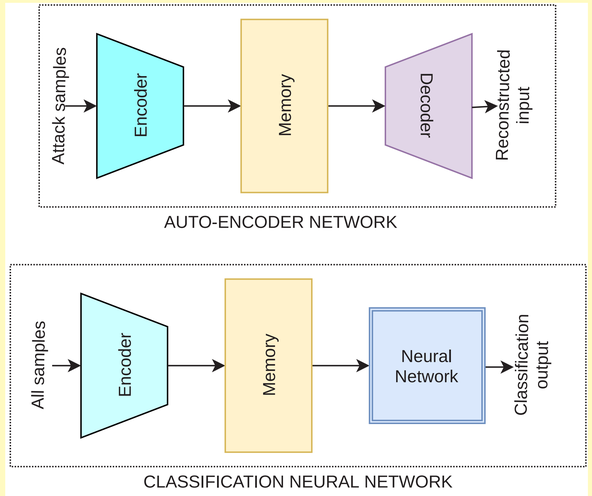


Figure. DNC architecture [1]

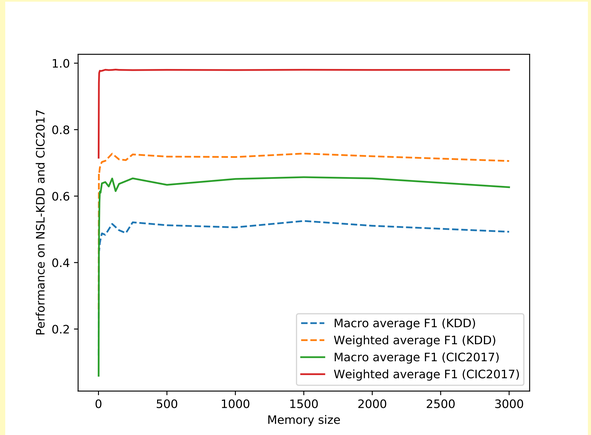
- Part of Neural Turing Machine
- Memory can be examined externally
- Reading and writing of external memory can be *tracked* and provide *valuable insight*
- External memory has been used in recent NIDS work, but explanaibility was not explored
- We aim to bridge the knowledge gap
- We provide a more thorough study using latest large datasets
  - CIC-IDS2017
  - CSE-CIC-IDS2018

Proposed Method



- Auto-encoder is trained with attack samples to generate memory contents
- Classification module for detection
- Hard shrinkage to reduce memory slots
- Max memory size=250 was suitable for the datasets
- Fine-tuned hyperparameters: layers, units, classification network, optimization

Memory Size



Memory Activations

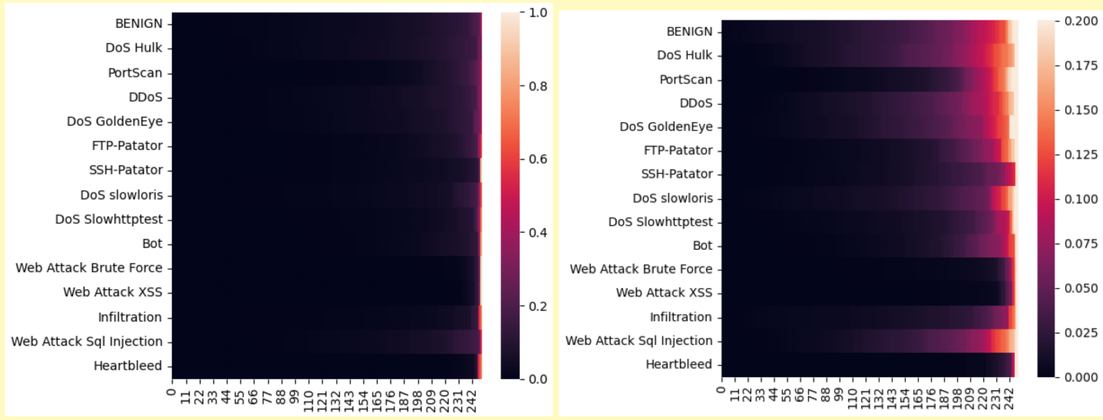


Figure. Sorted memory activations of CIC2017

Memory Contents

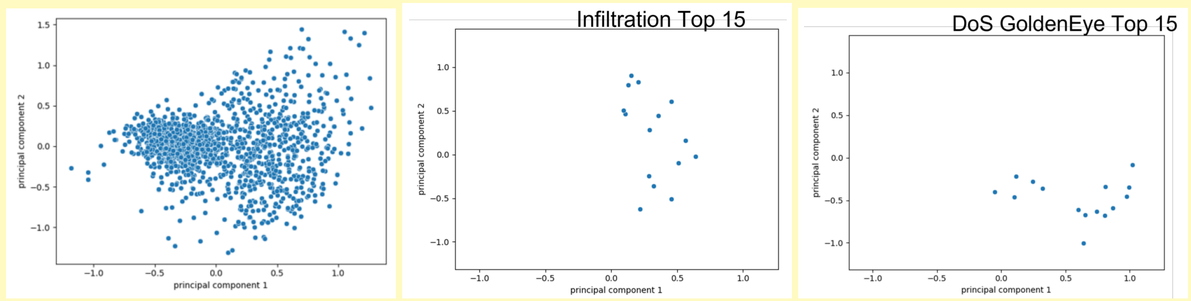
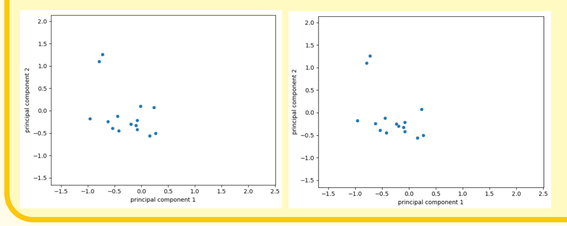


Figure. PCA graph with all memory contents, top 15 infiltration, and top 15 DoS Golden Eye - CIC2017

CIC-IDS2018

Memory size	15	250	1500
Benign	0.993	0.993	0.993
Bot	0.998	0.998	0.999
Brute Force -Web	0.339	0.474	0.445
Brute Force -XSS	0.603	0.606	0.597
DDOS attack-HOIC	0.999	0.999	0.999
DDOS attack-LOIC-UDP	0.831	0.817	0.842
DDoS attacks-LOIC-HTTP	0.993	0.994	0.995
DoS attacks-GoldenEye	0.989	0.994	0.966
DoS attacks-Hulk	0.999	0.999	0.997
DoS attacks-SlowHTTPTest	0.600	0.580	0.597
DoS attacks-Slowloris	0.957	0.967	0.964
FTP-BruteForce	0.782	0.775	0.782
Infilteration	0.010	0.018	0.021
SQL Injection	0.1	0.228	0.306
SSH-Bruteforce	0.999	0.999	0.999
macro avg	0.746	0.763	0.767
weighted avg	0.978	0.978	0.978

Unseen Attacks CICIDS2017



References & Github

[1] Graves, A., Wayne, G., and Reynolds, M. Hybrid computing using a neural network with dynamic external memory *Nature*, 7626:471–476,2016.

Github repository

https://github.com/dsphamgithub/explainids