Paper 1

**Citation**  
Chua, Tuan-Hong, and Iftekhar Salam. “Evaluation of Machine Learning Algorithms in Network-Based Intrusion Detection Using Progressive Dataset.” *Symmetry*, vol. 15, no. 6, 1 June 2023, p. 1251, arxiv.org/pdf/2203.05232v1.pdf, <https://doi.org/10.3390/sym15061251> . Accessed 20 Sep. 2023.

**Abstract**

Cybersecurity has become one of the focuses of organisations. The number of cyberattacks keeps increasing as Internet usage continues to grow. As new types of cyberattacks continue to emerge, researchers focus on developing machine learning (ML)-based intrusion detection systems (IDS) to detect zero-day attacks. They usually remove some or all attack samples from the training dataset and only include them in the testing dataset when evaluating the performance. This method may detect unknown attacks; however, it does not reflect the long-term performance of the IDS as it only shows the changes in the type of attacks. In this work, we focused on evaluating the long-term performance of ML-based IDS. To achieve this goal, we proposed evaluating the ML-based IDS using a dataset created later than the training dataset. The proposed method can better assess the long-term performance as the testing dataset reflects the changes in the attack type and network infrastructure changes over time. We have implemented six of the most popular ML models, including decision tree (DT), random forest (RF), support vector machine (SVM), naïve Bayes (NB), artificial neural network (ANN), and deep neural network (DNN). These models are trained and tested with a pair of datasets with symmetrical classes. Our experiments using the CIC-IDS2017 and the CSE-CIC-IDS2018 datasets show that SVM and ANN are most resistant to overfitting. Our experiments also indicate that DT and RF suffer the most from overfitting, although they perform well on the training dataset. On the other hand, our experiments using the LUFlow dataset have shown that all models can perform well when the difference between the training and testing datasets is small.

Paper 2

Citation

Raff, Edward, et al. “Classifying Sequences of Extreme Length with Constant Memory Applied to Malware Detection.” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 11, 18 May 2021, pp. 9386–9394, arxiv.org/pdf/2012.09390v1.pdf,  
[https://doi.org/10.1609/aaai.v35i11.17131 .Accessed 20 Sep. 2023](https://doi.org/10.1609/aaai.v35i11.17131%20.Accessed%2020%20Sep.%202023).

Abstract

Recent works within machine learning have been tackling inputs of ever increasing size, with cyber security presenting sequence classification problems of particularly extreme lengths. In the case of Windows executable malware detection, an input executable could be >=100 MB, which would translate to a time series with T=100,000,000 steps. To date, the closest approach to handling such task is MalConv --- a convolutional neural network capable of processing T=2,000,000 steps. Because the memory used by CNNs is O(T), this has prevented many from processing all executables or further extending the MalConv approach. In this work, we develop a new approach to temporal max pooling that makes the required memory invariant to the sequence length T. This makes MalConv 116x more memory efficient, and up to 25.8x faster to train, while removing the input length restrictions to MalConv. We re-invest these gains into improving the MalConv architecture by developing a new Global Channel Gating design, giving us an attention mechanism capable of learning feature interactions across 100 million time steps in an efficient manner, a capability lacked by the original MalConv approach.