Evading AI-Based Malware Detection

Wilson Tran

CyberSecurity Service Provider Internship Program, ICS 296 & 396, Fall 2023

Advisors: Dr. Curtis Arnold, Chief Scientist – Cyber, & Dr. Robert Adamove, Sr. Associate, (Core4ce)

Problem Statement:

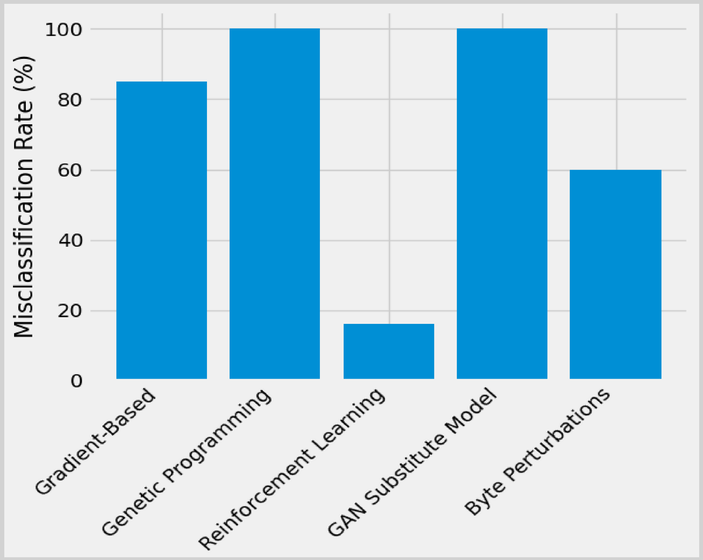
The problem is that deep learning-based approaches for malware detection are highly susceptible to adversarial examples (Aslan and Samet 2020). This technology has achieved up to 95% detection rates (Saxe and Berlin 2015). However, it has been demonstrated that misclassification rates of up to 60% can be achieved despite less than 1% of bytes being modified (Kolosnjaji et al. 2018).

Purpose:

The purpose of this study is to aggregate a collection of adversarial examples from recent literature. These may be used for testing in developing more robust malware detection models. Deep learning has already displayed state-of-the-art performance for malware detection and other complicated problem domains. Therefore, it could become a powerful and widespread tool to help automate aspects of malware identification given less vulnerability to evasion strategies.

Approach:

The academic research which this literature review is based on was exclusively collected from papers published on Google Scholar. An emphasis was placed on collecting well-cited papers that have been validated by multiple experts in the field. In addition, research published within the past 5 years was referenced as much as possible. With these sources, there was a focus on finding adversarial examples and corresponding achievable misclassification rates. A prior goal was a comparison amongst the collected adversarial examples. As the purpose of this study is ultimately aiding in developing more robust models, the challenges of developing models for malware detection was investigated.



Discussion & Conclusion:

This study found that there is rich variety in adversarial examples which can result in misclassification rates of up to 100%. Furthermore, malware detection research suffers from shortcomings as a result of applying deep learning to a novel domain.

Author	Year	Weakness/Limitation
Rhode, Matilda, et al.	2018	The system should be tested for large data. This approach can be a failure if the attacker comes to know that file is being monitored in the first 5 s so this can be evaded.
Hardy, William, et al.	2016	Sparsity constraints are not imposed on SAE which can improve malware detection.
Saxe, Joshua, and Konstantin Berlin.	2017	The computational cost of training on long strings is very high. This approach labeled any sample that had 0 occurrences in malware data as benign, and the rest was labeled malware. So, strings, file paths, and registry keys due to less training data can decrease this model's generalizability.
Azmoodeh, Amin, Ali Dehghantanha, and Kim-Kwang Raymond Choo.	2018	Dataset was small for training the neural network which implies that the network could not learn the features at its best.
Cui, Zhihua, et al.	2018	The model required all the input images to be of fixed size due to which images could have lost meaningful information while image processing.
Ni, Sang, Quan Qian, and Rui Zhang	2018	Detection of packed, encrypted malware or malware using anti-debugging and anti-dissembling approaches is not performed. Real time data consist of all these kind of malware, therefore network might not work well with real time data.

High misclassification rates could be achieved despite the scope of binary file modifications being very limited. The "black box" nature of deep neural networks makes it difficult to determine remediation measures for misclassification. Further complicating this is the lack of a high-quality, publicly available malware dataset. This presents an inability to meaningfully compare results across studies. Notably, traditional machine learning methods may not always apply to the security domain. Implementing batch normalization hindered the detection performance of MalConv.

Path Forward:

The lack of a standardized dataset(s) limits the ability for studies to build upon past work. As such, progress in this research area will likely be hindered. In the meantime, the indicated vulnerabilities of deep learning-based malware detection to adversarial examples makes it difficult to recommend for usage in critical domains.

References:

Hu, W., & Tan, Y. (2022). Generating Adversarial Malware Examples for Black-Box Attacks Based on GAN. Data Mining and Big Data, 409–423. https://doi.org/10.1007/978-981-19-8991-9_29

Raff, E., Barker, J., Sylvester, J., Brandon, R., Catanzaro, B., & Nicholas, C. K. (2018, June). Malware detection by eating a whole exe. In Workshops at the thirty-second AAAI conference on artificial intelligence.

INFORMATION & COMPUTER SCIENCES

UNIVERSITY of HAWAI'I at MĀNOA