Evaluating Machine Learning-Based Static Malware Classifiers

Yin, Andrew (School: Martin Luther King High School)
Link, Daniel (School: Martin Luther King High School)

With the development of the Internet, cybersecurity becomes increasingly essential. The best way to reduce the damage of cybercrime is to develop better methods of cybersecurity; one of them being the detection of malicious software. Currently, the methods used for malware detection are anti-malware signatures, heuristic analysis, and dynamic analysis run in a controlled environment. These methods of malware identification cannot detect zero-day attacks or have flaws that are easily manipulated. Machine learning malware detection allows for the static analysis of potential malware files, where the files will not have to be run and zero-day attacks can be protected against. It is relatively new in the cybersecurity industry and shows promising results. Two of the most common machine-learning malware classifying models, EMBER and MalConv, will be compared using several binary classification metrics. EMBER is a traditional machine learning model that parses through selected files properties. MalConv is a deep learning model that uses the raw bytes of a file to make predictions. The main premise of the project will be to determine whether traditional machine learning or deep learning is better suited to classify malware. EMBER and MalConv repositories were cloned from GitHub and then trained to detect malware. Malicious and benign file samples were then parsed through the algorithms using Python scripting. The accuracy and efficiency of the models were monitored. EMBER scored slightly better in all metrics and was also the faster algorithm. Whilst further analysis and rigorous testing will be needed to corroborate the robustness of the algorithms, EMBER and MalConv can offer possibilities for the future of malware detection and cybersecurity.