Chapter 96
Towards Deep Learning-Based Approach for Detecting Android Malware

Jarrett Booz
Towson University, Towson, USA

Wei Yu
Towson University, Towson, USA

Josh McGiff
Towson University, Towson, USA

James Nguyen
Towson University, Towson, USA

William G. Hatcher
Towson University, Towson, USA

Chao Lu
Towson University, Towson, USA

ABSTRACT

In this article, the authors implement a deep learning environment and fine-tune parameters to determine the optimal settings for the classification of Android malware from extracted permission data. By determining the optimal settings, the authors demonstrate the potential performance of a deep learning environment for Android malware detection. Specifically, an extensive study is conducted on various hyper-parameters to determine optimal configurations, and then a performance evaluation is carried out on those configurations to compare and maximize detection accuracy in our target networks. The results achieve a detection accuracy of approximately 95%, with an approximate F1 score of 93%. In addition, the evaluation is extended to include other machine learning frameworks, specifically comparing Microsoft Cognitive Toolkit (CNTK) and Theano with TensorFlow. The future needs are discussed in the realm of machine learning for mobile malware detection, including adversarial training, scalability, and the evaluation of additional data and features.

DOI: 10.4018/978-1-7998-7705-9.ch096

INTRODUCTION

Recent advances in machine learning have projected neural networks and deep learning systems into the public consciousness. This is attributable to the significant strides that deep learning systems continue to make in a large variety of areas, including image and video analysis and feature recognition, autonomous vehicles, natural language processing, the control of robotic systems, and others (Hatcher & Yu, 2018). Indeed, deep learning has emerged as an extremely powerful tool for processing complex data (LeCun, Bengio, & Hinton, 2015). Deep learning models, and deep neural networks in particular, have the capacity to learn and represent extremely complex systems and reveal features or patterns at a level of abstraction that is not feasible for simpler algorithms. Encompassing a variety of learning tasks, from clustering and dimensionality reduction, to classification and reinforcement learning, deep learning architectures apply systems of hierarchical layers to fit and generalize large, often multi-dimensional, feature sets more accurately than their shallow learning counterparts (Fadlullah et al., 2017).

For this reason, deep learning has great appeal for applications in the realm of computer security. Truly, a wide array of security applications exists, which can benefit from deep learning approaches. For instance, malware and intrusion detection systems that employ anomaly detection can benefit immensely from the application of deep learning, as they require accurate detection of possibly yet unknown threats in highly complex environments (Zhao, Chandrashekar, Lee, & Medhi, 2015; Thing, 2017; Cordero, Hauke, Mühlhäuser, & Fischer, 2016; Alrawashdeh & Purdy, 2016). In addition, mobile malware detection is an area of pressing concern due to the rapid growth of the smartphone market, which now encompasses approximately 209 million users in the U.S., and over 1.9 billion users worldwide (Statista, 2017). As a driving factor concerning this study, the Android operating system continues to massively dominate the smartphone market, encompassing the vast majority of devices. Significant research has been directed toward both static and dynamic analysis of Android malware, including shallow learning of Android manifest features, analysis of malware families, dynamic evaluation of system calls, and others.

In seeking to address the issues of smartphone security, and Android malware detection in particular, in this paper we make the following contributions:

- We address the utility of deep learning for analyzing permission data from Android applications in order to classify apps as malicious or benign. To accomplish this, we use various python libraries to construct a deep learning infrastructure, which consists of TensorFlow machine learning backend, the Keras framework, and Scikit-Learn utilities, to extract and vectorize the target features, and then use these dense vectors for deep learning analysis. Initially attempting only a rudimentary implementation, our neural network yielded results of about 90% accuracy;
- Once this initial result was extracted and a framework built, we then targeted mechanisms to optimize the results. To identify the optimal network hyper-parameters, we utilized the grid search technique to test many combinations of tunable parameters for the deep learning environment. We also leveraged various neural network shapes, making some networks deeper or wider, to determine the best shape and size model for our application. The results showed that, by tuning six different parameters, we were able to increase the accuracy of the classification network, for a maximum accuracy of approximately 95% correct classifications;
- We present an extension to evaluate the performance of other machine learning frameworks, including the Microsoft Cognitive Toolkit (CNTK) and Theano. To determine the fully optimized capabilities of deep learning for Android malware detection, we conducted additional testing