


Chapter 14

Malware Classification and Analysis Using Convolutional and Recurrent Neural Network

Yassine Maleh

 <https://orcid.org/0000-0003-4704-5364>

Hassan 1st University, Morocco

ABSTRACT

Over the past decade, malware has grown exponentially. Traditional signature-based approaches to detecting malware have proven their limitations against new malware, and categorizing malware samples has become essential to understanding the basics of malware behavior. Recently, antivirus solutions have increasingly started to adopt machine learning approaches. Unfortunately, there are few open source data sets available for the academic community. One of the largest data sets available was published last year in a competition on Kaggle with data provided by Microsoft for the big data innovators gathering. This chapter explores the problem of malware classification. In particular, this chapter proposes an innovative and scalable approach using convolutional neural networks (CNN) and long short-term memory (LSTM) to assign malware to the corresponding family. The proposed method achieved a classification accuracy of 98.73% and an average log loss of 0.0698 on the validation data.

1. INTRODUCTION

With the rapid development of the Internet, malware has become one of the major cyber threats today. Any software that performs malicious actions, including stealing information, spying, etc. can be called malware. Kaspersky Labs (2017) defines malware as “a type of computer program designed to infect a legitimate user’s computer and inflict damage in multiple ways”.

As the diversity of malware increases, antivirus scanners cannot meet protection needs, resulting in millions of hosts being attacked. According to malware statistics report, Symantec affirms that more than 357 million new malware variants were observed in 2016. (Symante, 2017). Juniper Research (2016) predicts that the cost of data breaches will rise to \$2.1 trillion globally by 2019.

DOI: 10.4018/978-1-5225-7862-8.ch014

In addition, there is a decrease in the level of skill required for malware development, due to the high availability of attack tools on the Internet today. The high availability of anti-detection techniques, as well as the ability to purchase malware on the black market, gives the possibility to become an attacker to anyone, regardless of skill level. Current studies show that more and more attacks are launched by script-kiddies or are automated (Aliyev, 2010).

Therefore, protecting computer systems against malware is one of the most important cybersecurity tasks for individual users and businesses, because even a single attack can compromise important data and cause sufficient losses. Frequent attacks and massive losses dictate the need for accurate and timely detection methods. Current static and dynamic methods do not allow accurate and effective detection, especially when it comes to zero-day attacks. For this reason, techniques and methods based on machine learning can be used (Chumachenko & Technology, 2017).

When classifying malicious code families, it is important to identify the unique characteristics of malicious codes, but it is also important to select the classification algorithms used as classifiers correctly. Recently, one of the most actively studied fields in the study of classification or recognition techniques is the deep neural network (DNN) related research called depth neural network which is made by increasing the number of hidden layers of neural networks. In particular, in the field of image and speech recognition, deep neural network based models have shown excellent performance, and there are moves to use them in other areas as well. Malicious code analysis is one such area. Indeed, various malicious code classification models using deep neural networks have been proposed. There are many research studies that combine classification schemes using recurrent neural networks (NRNs) (Pascanu, Tour, Mikolov, & Tour, 2013) and conventional neural networks in the field of image recognition and processing, but just few in the field of malwares and intrusions detection and classification (Chen, 2015).

This chapter aims to explore the problem of malware classification, and to propose a new approach combining Convolutional Neural Network (CNN) and Long Short-Term Memory Recurrent Neural Network (LSTM). The proposed model has been evaluated on the data provided by Microsoft for the BIG Cup 2015 (Big Data Innovators Gathering).

The main contributions of this chapter are:

1. We explore various deep learning models to solve the proposed malware classification problem;
2. We propose a deep neural network model to classify malicious behavior by combining CNN and LSTM layers;
3. We conduct a case-study using Microsoft Malware Dataset and show that our model has high detection accuracy (98,73%) in comparison with other models.

This chapter presents the research background in the next section. The related work of the malware classification technique in section 3 and the detailed description of the proposed methodology in section 4. Section 5 describes the experiments using the proposed model. Section 6 presents conclusions and future research directions.

2. BACKGROUND

Deep learning has demonstrated powerful function learning capabilities and has achieved remarkable performance in the field of computer vision that extracts the complex features through layer by layer

21 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:

www.igi-global.com/chapter/malware-classification-and-analysis-using-convolutional-and-recurrent-neural-network/227856

Related Content

Artificial Neural Network for Pre-Simulation Training of Air Traffic Controller

Tetiana Shmelova, Yuliya Sikirdaand Togrul Rauf Oglu Jafarzade (2022). *Research Anthology on Artificial Neural Network Applications* (pp. 1334-1358).

www.irma-international.org/chapter/artificial-neural-network-for-pre-simulation-training-of-air-traffic-controller/289016

Artificial Neural Networks in EEG Analysis

Markad V. Kamath, Adrian R. Upton, Jie Wu, Harjeet S. Bajaj, Skip Poehlmanand Robert Spaziani (2006). *Neural Networks in Healthcare: Potential and Challenges* (pp. 177-194).

www.irma-international.org/chapter/artificial-neural-networks-eeeg-analysis/27278

Meta-Heuristic Parameter Optimization for ANN and Real-Time Applications of ANN

Asha Gowda Karegowdaand Devika G. (2022). *Research Anthology on Artificial Neural Network Applications* (pp. 166-201).

www.irma-international.org/chapter/meta-heuristic-parameter-optimization-for-ann-and-real-time-applications-of-ann/288956

Strategies for Automated Bike-Sharing Systems Leveraging ML and VLSI Approaches

Jagrat Shukla, Numburi Rishikha, Janhavi Chaturvedi, Sumathi Gokulanathan, Sriharipriya Krishnan Chandrasekaran, Konguvel Elangoand SathishKumar Selvaperumal (2023). *Neuromorphic Computing Systems for Industry 4.0* (pp. 172-203).

www.irma-international.org/chapter/strategies-for-automated-bike-sharing-systems-leveraging-ml-and-vlsi-approaches/326838

Resource Scheduling and Load Balancing Fusion Algorithm with Deep Learning Based on Cloud Computing

Xiaojing Houand Guozeng Zhao (2020). *Deep Learning and Neural Networks: Concepts, Methodologies, Tools, and Applications* (pp. 1042-1057).

www.irma-international.org/chapter/resource-scheduling-and-load-balancing-fusion-algorithm-with-deep-learning-based-on-cloud-computing/237920