**International Journal of Digital Crime and Forensics**
Volume 12 • Issue 1 • January-March 2020

# A Deep Learning Framework for Malware Classification

Mahmoud Kalash, University of Manitoba, Winnipeg, Canada
Mrigank Rochan, University of Manitoba, Winnipeg, Canada
Noman Mohammed, University of Manitoba, Winnipeg, Canada
Neil Bruce, Ryerson University, Toronto, Canada
Yang Wang, University of Manitoba, Winnipeg, Canada
Farkhund Iqbal, Zayed University, Abu Dhabi, UAE

## ABSTRACT

In this article, the authors propose a deep learning framework for malware classification. There has been a huge increase in the volume of malware in recent years which poses serious security threats to financial institutions, businesses, and individuals. In order to combat the proliferation of malware, new strategies are essential to quickly identify and classify malware samples. Nowadays, machine learning approaches are becoming popular for malware classification. However, most of these approaches are based on shallow learning algorithms (e.g. SVM). Recently, convolutional neural networks (CNNs), a deep learning approach, have shown superior performance compared to traditional learning algorithms, especially in tasks such as image classification. Inspired by this, the authors propose a CNN-based architecture to classify malware samples. They convert malware binaries to grayscale images and subsequently train a CNN for classification. Experiments on two challenging malware classification datasets, namely Malimg and Microsoft, demonstrate that their method outperforms competing state-of-the-art algorithms.

## KEYWORDS

Convolutional Neural Networks, Deep Learning, Framework, Malware Classification

## INTRODUCTION

Malware is malicious software (e.g. viruses, worms, Trojan horses, and spyware) that damages or performs harmful actions on computer systems (Malware Definition, 2017). In this Internet-age, many malware attacks happen that pose serious security threats to financial institutions and everyday users. Prior studies also highlight that malware analysis is crucial for digital forensic investigation (Kaur & Nagpal, 2012). Figure 1 represents the number of malwares spotted in a year. It is clear that the total number of instances of malware has drastically increased over the years. For example, Symantec reported that more than 357 million new variants of malware were observed in 2016 (Internet Security Threat Report, 2017). One of the main reasons for this high volume of malware samples is the extensive use of obfuscation techniques by malware developers, which means that malicious files from the same malware family (i.e. similar code and common origin) are constantly modified and/or obfuscated. In order to cope with the rapid evolution of malware, it is essential to develop robust

DOI: 10.4018/IJDCF.2020010105

malware classification techniques that are tolerant of variants of malware files that belong to same family. Towards this endeavor, we propose a deep learning architecture for malware classification.

Conventional methods use binary signatures of malware for analysis. Malware typically carries a uniquely identifiable signature. Signature-based methods were extensively used in the past in anti-virus software. Given the exponential increase in malware files and degree of variation, these signature-based methods are not scalable. Other methods for malware analysis include static and dynamic code analysis (Nataraj, Karthikeyan, Jacob, & Manjunath, 2011). In static analysis, the malware code is disassembled to find malicious patterns. In contrast, dynamic analysis is done by executing the malicious program in a virtual environment and its behavior is analyzed based on execution trace. Dynamic analysis is more effective than static as it does not require disassembling, but it is time consuming and resource intensive. Also, it is possible that during the dynamic analysis malicious behaviors go unnoticed because the virtual environment may not be able to simulate the exact real conditions (Nataraj, Karthikeyan, Jacob, & Manjunath, 2011).

Previous research on malware classification suggest that malware samples typically fall into a family that share common behavior. Most new malware are variants of existing ones (Nataraj, Karthikeyan, & Manjunath, 2015). Hence, the prospect of building a method that can efficiently classify malware based on its family irrespective of being a variant, seems especially fruitful and a means of dealing with the rapid growth of malware.

In this paper, we take a completely different approach to analyze and classify malware compared with traditional methods. We use a Convolutional Neural Network (CNN), a deep learning architecture, to tackle this problem.

Recently, CNNs have produced state-of-the-art performance on the image classification task in the field of computer vision. Motivated by this success, we translate the malware classification problem into the image classification problem to be addressed using CNNs. We firstly represent each malware binary file as a grayscale image and then train a CNN architecture to perform classification. Previous work (Nataraj, Karthikeyan, Jacob, & Manjunath, 2011) showed that malware belonging to same family are visually similar, which is beneficial with respect to the capacity for a CNN to detect relevant patterns. This is especially true given that the same or similar code is usually used to generate variants of malware. However, the method proposed in (Nataraj, Karthikeyan, Jacob, & Manjunath, 2011) have several shortcomings (See the Related Work section).

There is a recent work (Gibert Llauradó, 2016) that uses CNN for malware classification, but it is still very shallow in architecture. In computer vision, researchers have shown that deeper CNN architectures (e.g. (Simonyan & Zisserman, 2014)) are helpful in minimizing error for image classification tasks. In this paper, we take the approach of vision researchers and adopt their technique for malware classification.

## CONTRIBUTIONS

We make the following main contributions in this paper:

1. We develop a deep convolutional neural network (CNN) architecture for malware classification, which is generic in nature, unlike traditional methods. Existing techniques that achieve high accuracy are often tailored for a specific dataset. In contrast, the proposed approach is data independent and learns the discriminative representation from the data itself rather than depending on hand-crafted feature descriptors.
2. We show that malware classification with higher accuracy is possible even if only a portion of malware sample is available. As far as we know, we are the first to develop CNN-based method that can classify malware samples using only partial knowledge of the properties of samples. In addition, there are a number of advantages of the proposed approach that are discussed in the Advantages section.