International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

Malware Detection Based on Deep Learning

Anjana R R[1]

[1]*DDMCA Student, Dept. of MCA, Sree Narayana Guru Institute of Science and Technology, Kerala, India*

Abstract - *Malicious code is that the quite harmful code or web script designed to form system vulnerabilities leading to back doors, security breaches, information and data theft, and other potential damages to files and computing systems. With the event of the online , malicious code attacks have increased exponentially, with malicious code variants ranking as a key threat to Internet security. Current methods for finding malicious code have demonstrated poor detection accuracy and low detection speeds. This paper proposed a completely unique method that used deep learning to enhance the detection of malware variants. To implement our proposed detection method, we converted the malicious code into color based image .That having three channels red, blue and green. Then the photographs were identified and classified employing a convolutional neural network (CNN) that might extract the features of the malware images automatically. To check our approach, we conducted a series of experiments on malware image data from Vision lab . The experimental results demonstrated that our model achieved good accuracy and speed as compared with other malware detection models.*

Key Words: Malware variants, Deep learning, Convolution neural network, VGG16, Malware Classification.

1.INTRODUCTION

1.1 Background

With the rapid development of data technology, the exponential growth of malicious code has become one among the most threats to Internet security. A recent report from Symantec showed that 401 million malicious codes were found in 2016, including 357 million new malicious code variants [1]. To date, 68 new malicious code families and quite 10 thousand malicious codes are reported. This growth has posed a challenge for malicious code detection in cloud computing [2], [3].

Malware detection methods consist primarily of two sorts of approaches: static detection and dynamic detection. Static detection works by disassembling the malware code and analyzing its execution logic. Dynamic detection analyzes the behavior of malicious code by executing the code during a safe virtual environment.

Both static and dynamic detection are feature-based methods. First, the textual or behavioral features of the malicious code are extracted, then the malicious code is detected or classified by analyzing these extracted features. In recent years, several scholars have used data processing methods to research the features of malicious code [4].

1.2 Motivations

Malware is growing within the large volume each day , we used image processing technique so on enhance accuracy and performance. Image processing technique analyzes malware binaries as gray-scale images. The previous research [5] proposed a replacement method for visualization to classify malware using image processing technique. a number of mature image processing techniques are widely used for beholding

1.3 Our approach

Malware classified in several families has multiple characteristics or features. Many authors used machine learning models like Regression, K-nearest-neighbor, Random Forest etc. Main disadvantage of using machine learning is, features extraction is manual. Gavrilut et al. [6] gave an summary of various machine learning techniques that were previously proposed for malware detection. Unlike Machine Learning, Deep learning skips the manual steps of extracting features. as an example , we'll feed directly images and videos to the deep learning algorithm, which can predict the thing . During this way deep learning model is more intelligent rather than machine learning model. We used convolutional neural networks because it's reliable and it are often applied to the whole image at a time then we will assume they're best to use for feature extraction. To implement our proposed detection method, we converted the malicious code into color based images. Then the photographs were identified and classified employing a convolutional neural network (CNN) that might extract the features of the malware images automatically.

1.4 Contributions

The main contributions of the paper are summarized as follows:

- We introduced a technique for converting a malware binary to an image, thereby transforming malware detection into an image classification problem.
- We proposed a novel method for detecting malware variants based on a convolutional neural network (CNN).

- Extensive experimental results demonstrated that our proposed method was an effective and efficient approach for malware detection.

2. DATA REPRESENTATION AND RELATED WORK

This section is divided into two parts. The first part is, to collect malware and benign datasets from different sources and second part describes the techniques of preparation of the dataset.

2.1 Collection of Dataset

We have collected 15 datasets from different sources. These are malicious datasets from two different sources i.e. from Vision Research Lab and from Microsoft Malware Classification Challenge. We also collected some benign file from different sources.

Table -1: Dataset of malware

No	Malware Class
1	Adphoshel
2	Agent
3	Allaple
4	BrowseFox
5	Dinwod
6	Elex
7	Expiro
8	Fasong
9	Hlux
10	Injector
11	Neshta
12	Regrun
13	Stantinko
14	VBkrypt
15	Vilsel

Fig -1: Agent

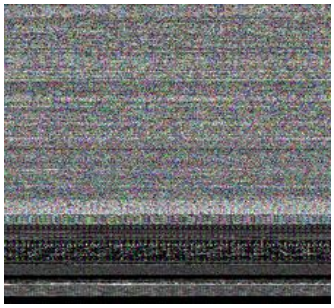

Fig -2: Elex

Dataset consists color images of 15 malware families. Ratio of 90-10 was used for model performance evaluation. 90% of the entire data was used for training and 10% was used for testing.

2.2 Related work

In this section, we present the related research regarding malware, including malware detection supported feature analysis, malicious code visualization, image processing techniques for malware detection, and malware detection supported deep learning.

- MALWARE DETECTION SUPPORTED FEATURE ANALYSIS

As described previously during this paper, there are two main categories of feature analysis techniques for malware detection: static analysis and dynamic analysis. With reference to static analysis, several methods are suggests by analyzing the codes. for instance , Isohara et al. [7] developed a detection system using kernel behavior analysis that performed well detecting the malicious behaviors of unknown applications.

Dynamic analysis monitors and analyzes the runtime characteristics of applications (apps) supported assessment of behaviors, like accessing private data and using restricted API calls. Given this information, a behavior model is established to detect malicious code. Such techniques achieved improved detection performance, but they were still

challenged by the variability of countermeasures developed to get unreliable results [8].

- MALICIOUS CODE VISUALIZATION

Currently, many tools can visualize and manipulate binary data, e.g., common text editors and binary editors. Several studies have proposed utilization of malware visualization [9]. there's a replacement visualization approach for malware detection supported binary texture analysis. First, they converted the malware executable file into grayscale images. Then they identified malware consistent with the feel features of those images.

- IMAGE PROCESSING TECHNIQUES FOR MALWARE DETECTION

The traditional approaches were challenged by the time cost required for complex image texture feature extraction. to deal with this challenge, we employed deep learning to spot and classify images efficiently. Within the next subsection, we present our research on malware detection supported deep learning.

- MALWARE DETECTION SUPPORTED DEEP LEARNING

Deep learning has the power to find out the essential characteristics of knowledge sets from a sample set. As a strong tool of AI , deep learning has been applied widely in many fields, like recognition of handwritten numerals, speech recognition, and image recognition. Due to it powerful ability to find out features, many scholars have applied deep learning to malware detection.

3. MALICIOUS CODE SUPPORTED CONVOLUTIONAL NEURAL NETWORK

First the binary files of malicious code are transformed into the color images. Next, the convolution neural network is used to spot and classify the pictures . consistent with the results of image classification, we realized the automated recognition and classification of malicious software.

- CONVOLUTIONAL NEURAL NETWORKS (CNNs)

Convolutional Neural Networks (CNNs) are a deep learning approach to tackle the image classification problem, or what we call computer vision problems, because classic computer programs face many challenges and difficulties to spot objects for several reasons, including lighting, viewpoint, deformation, and segmentation .This technique is inspired by how the attention works, especially the visual area function algorithm in animals. CNN are arranged in three-dimensional structures with width, height, and depth as characteristics. within the case of images, the peak is that the image height, the width is that the image width, and therefore the depth is RGB channels.

- VGG16 ARCHITECTURE

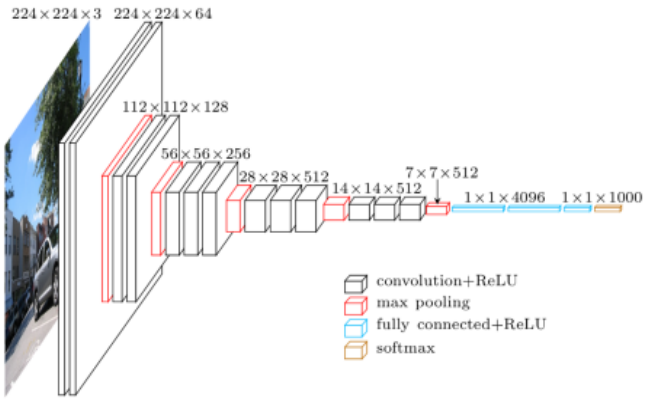

Figure 3. VGG16 architecture

VGG16 may be a 16-layer network employed by the Visual Geometry Group at the University of Oxford to get state of the art leads to the ILSVRC-2014 competition. the most feature of this architecture was the increased depth of the net-work. In VGG16, 224x224 RGB images are skilled 5 blocks of convolutional layers where each block is composed of accelerating numbers of 3x3 filters. The input to cov1 layer is of fixed size 224 x 224 RGB image.

The image is skilled a stack of convolutional layers, where the filters were used with a really small receptive field: 3×3 (which is that the smallest size to capture the notion of left/right, up/down, center). In one among the configurations, it also utilizes 1×1 convolution filters, which may be seen as a linear transformation of the input channels (followed by non-linearity). The convolution stride is fixed to 1 pixel; the spatial padding of conv. layer input is such the spatial resolution is preserved after convolution, i.e. the padding is 1-pixel for 3×3 conv. layers. Spatial pooling is administered by five max-pooling layers, which follow a number of the conv. layers (not all the conv. layers are followed by max-pooling). Max-pooling is performed over a 2×2 pixel window, with stride 2.

The 16 in VGG16 refers thereto has 16 layers that have weights. This network may be a pretty large network and it's about 138 million parameters. Three Fully-Connected (FC) layers follow a stack of convolutional layers the primary two have 4096 channels each, the third performs 1000-way ILSVRC classification and thus contains 1000 channels). the ultimate layer is that the soft-max layer. The configuration of

the fully connected layers is that the same altogether networks.

4. METHODOLOGY

- PREPROCESSING

The amount and type of processing done depends on the character of the preprocessor; some preprocessors are only capable of performing relatively simple textual substitutions and macro expansions, while others have the power of full-fledged programming languages. Data preprocessing is a crucial step within the data processing process. Data-gathering methods are often loosely controlled, resulting in out-of-range values impossible data combinations, missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results.

- ARCHITECTURE CREATION

Architecture, the art and technique of designing and building, as distinguished from the talents related to construction. The practice of architecture is used to satisfy both practical and expressive requirements, and thus it serves both utilitarian and aesthetic ends. In deep learning they maintain an architecture called architecture creation.

- FEATURE EXTRACTION

Feature extraction describes the relevant shape information contained during a pattern in order that the task of classifying the pattern is formed easy by a proper procedure. In pattern recognition and in image processing, feature extraction may be a special sort of dimensionality reduction.

- TRAINING

Training data is additionally referred to as a training set, training dataset or learning set. In the training state a model file is created. Then the system save its model file. The training data is an initial set of knowledge wont to help a program understand the way to apply technologies like neural networks to find out and produce sophisticated results. It may be complemented by subsequent sets of knowledge called validation and testing sets.

- TESTING

In hardware and software development, testing is employed at key checkpoints within the overall process to work out whether objectives are being met.

5. CONCLUSIONS

This paper proposed a method to improve the detection of malware variants through the application of deep learning. First, this method transformed the malicious code into color based images. Next, the images were identified and classified by a convolutional neural network that could extract the features of the malware images automatically. Because of the effectiveness and efficiency of the CNN for identifying malware images, the detection speed of our model was significantly faster than speeds achieved by other approaches.

REFERENCES

[1] Symantec. Internet security threat report, 2017.
[2] A.Khoshkbarforoushha, A. Khosravian, and R. Ranjan. Elasticity management of streaming data analytics flows on clouds. Journal of Computer and System Sciences, 89:24–40, 2017.
[3] A.Khoshkbarforoushha, R.Ranjan, R. Gaire, E. Abbasnejad, L. Wang, and A.Y. Zomaya. Distribution based workload modelling of continuous queries in clouds. IEEE Transactions on Emerging Topics in Computing, 5(1):120–133, 2017.
[4] Y.Ye, T. Li, D. Adjeroh, and S. S. Iyengar. A survey on malware detection using data mining techniques. ACM Computing Surveys (CSUR), 50(3):41, 2017.
[5]. Nataraj, L., Yegneswaran, V., Porras, P. and Zhang, J., 2011, October. A comparative assessment of malware classification using binary texture analysis and ynamic analysis. In Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence (pp. 21-30). ACM.
[6] Gavriluț, D., Cimpoeșu, M., Anton, D. and Ciortuz, L., 2009, October. Malware detection using machine learning. In Computer Science and Information Technology, 2009. IMCSIT'09. International Multiconference on (pp. 735-741). IEEE.
[7] T. Isohara, K. Takemori, and A. Kubota. Kernel-based behavior analysis for android malware detection. In 2011 Seventh International Conference on Computational Intelligence and Security (CIS), pages 1011–1015. IEEE, 2011.
[8] L. Nataraj, V. Yegneswaran, P. Porras, and J. Zhang. A comparative assessment of malware classification using binary texture analysis and dynamic analysis. In Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence, pages 21–30. ACM, 2011.
[9] M. Wagner, F. Fischer, R. Luh, A. Haberson, A. Rind, D. A. Keim, and W. Aigner. A survey of visualization systems for malware analysis. In Eurographics Conference on Visualization (EuroVis), pages 105–125,2015.