International Journal of Engineering & Technology, 7 (2.12) (2018) 80-83

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

A proposals of convolution neural network system for malicious code analysis based on cloud systems

Yong-kyu Park [1] *, Kyung-shin Kim [2], Jang-il Kim [3], Sung-hee Kim [4], Kil-hung Lee [5]

[1] Yong-kyu Park, 2Kyung-shin Kim, 3Jang-il Kim, 4Sung-hee Kim,5Kil-hung Lee CERT Team, Korea Internet & Security agency, IT Tower, Jungdae-ro135, Songpa-gu, Seoul, 05717, Korea
[2] Dept. of Mobile, Chungkang College of cultural industries, CK-ro 389-94 Majang-Myun, Ichon-siKyungki-do, 17390, Korea
[3] Dept. of Medical IT and Marketing, Eulji Univ., Sanseong-daero553, Sujeong-gu, Seongnam-si, Kyungki-do, 13135,Korea
[4] Department of computer engineering,Kwangwoon Univ., kwangwoonro 20, nowon-ku Seoul,01897, Korea
[5] Dept. of CS&E,Univ. of Seoul Nat'l Sci. and Tech., Gongneung-ro232, Nowon-gu, Seoul, 01811, Korea
**Corresponding author E-mail: smile@kisa.or.kr*

Abstract

Background/Objectives: In the information security field, artificial intelligence must be applied first. This is because the frequency of malicious code is too high and the processing method is too difficult, which is very difficult for human to handle.
Methods/Statistical analysis: In this paper, we developed a program to classify malicious codes into images and a Tensorflow system to classify malicious codes. The malware used as input was the computer virus code used in the BIG 2015 Challenge. This dataset, called a Kaggle dataset, consists of 10,868 bytes of train set.
Findings: We used the Tensorflow SLIM library to develop this machine learning malware learning machine. This resulted in more than 80% accuracy. Especially, when the CRIS-Ensemble algorithm was added, the accuracy was 97%. The study of malicious code analysis using machine learning consists of two major parts. First, the process of making the virus into images is important. To classify 10,868 Kaggle malware datasets that the BIG 2015 winner showed 99.6% accuracy, Tensorflow's accuracy and parameter tuning are important, but finding the way to make good images is the most important technique
Improvements/Applications: The results show that the malicious code classification system using machine learning can be an effective method to classify malicious code of malicious code by the accuracy of the result and ease of use.

Keywords: *Machine Learning; Tensorflow; Malware Code; Malware Datasets; Convolution Neural Networks*

1. Introduction

The importance of artificial intelligence, especially machine learning, and its necessity are emphasized. AlphaGo is a computer program that focuses attention on the world's people from the "GO" program. It is now becoming more than human intelligence beyond the stage of imitating human intelligence.
It is already known that one of the many applications that require artificial intelligence to be applied more quickly is in the field of information security[1]. Because this field can be processed by a program having a capability similar to that of human judgment, and the frequency of occurrence of malicious code and its processing ability exceed human limitations, there is no disagreement that machine learning should be introduced first.
Also machine learning can be applied for network intrusion detection, malicious code analysis, vulnerability analysis, etc. and it makes machine learning very efficient for the Information security. In the past, malicious cyber attacks were not as diverse as it is now, and intrusion detection and attack analysis were possible with pattern matching alone. Today, however, smart phones, smart cars, smart homes, smart factories, and cyber attacks are also intelligent and diversified. Therefore, a new technology that combines machine learning and information security is needed to cope with this.
This paper describes the development and results of a convolution neural network system that learns and classifies malicious codes using such machine learning.

2. Related research

2.1. Study on malicious code classification algorithm

It is a supervised learning method by making a model using the sample data that knows the answer, and when the new data is given, the guess is made or the type is discriminated.
For example, to predict the stock price of the stock market or to sort spam emails.Unsupervised learning is a representative example of clustering that collects sample data that do not know the correct answer among similar things[2]. In addition to supervised and unsupervised learning, reinforcement learning is also distinguished as a type of machine learning. Thereinforcement learning adjusts the execution method according to the result of the algorithm. There is a similar aspect to the guidance learning. However, the result of the reinforcement learning is feedback that the algorithm performed well, not the prepared correct answer.

Copyright © 2018 Yong-kyu Park et. al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

2.2. Deep running and tensorflow

Tensorflow is an open source library created by Google for machine learning and deep learning. This library is a follow-up version of the machine learning system that Google, publicize its vision of being an artificial intelligence company, has used internally[3].
The official site is www.tensorflow.org and it is easy to access because it is written in C ++. Especially all sources are provided by GitHub. Currently, it can also be run as a package called TF-SLIM on Windows, but only 64-bit Linux and Mac-OS and GPU CUDA environments can achieve maximum performance, and programming languages are C ++ and Python.Of course, they plan to support various languages in the future.
More than 50 tensor flows have been applied among Google's products, and tensor flow is essential when applying machine learning technology using DNN (Deep Neural Network).

2.3. Machine learning and cloud computing

The development of Google's deep learning technology is simply not possible with large amounts of data that can be learned. Previously, there was a large amount of data for statistical analysis, and there were many ways of how to process the data.

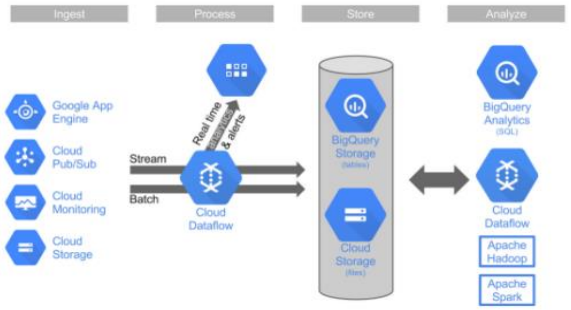


Fig. 1: Google Cloud Computing.

Key technologies that enable deep learning technology should support a number of data processing and analysis platforms that can perform the currently advanced types of machine learning algorithms.
Eric Schmidt, chairman of Google, predicts that the computing environment that combines machine learning-based artificial intelligence, Internet-based cloud computing, and Big Data from CrowdSourcing will be the cornerstone of future IT innovation.
As shown in Figure 1 above, Google's cloud platform aims to reduce the cost of infrastructure management and program development for individuals and companies by providing a variety of infrastructure technologies such as App Engine, Cloud Storage and Cloud SQL for actual Google services. The Google Cloud Platform uses an internal architecture, Container, to provide a minimal set of services required for a service. Containers provide flexibility and availability for different virtual environments within the server. Docker, known for introducing these container concepts, was written on the Linux-based programming language Go and is now known as the next generation open source project. Container-based virtualization is known to be much more efficient for building, deploying and overheading than traditional VMs.
Google has released a number of projects, including Hadoop MapReduce, Spanner, the container management platform Kubernetes, DataFlow and Tensorflow, which Google has released as an open source project. .
Tensor Flow supports multiple parallel processes that can use multiple CPUs and GPUs for data processing and analysis, and provides dataflow technology to process data on the cloud in real time. Figure 2 below shows a data representation and processing flow diagram of the tensor flow for deep running.It provides flexibility by expressing graphs of data processing flow charts and considering development in server environment such as mobile and PC using C ++ and Python API. Currently, various personal and corporate services provided by Google have been expanded based on google.com's main search engine. In 2015, Google led the investment and research on artificial intelligence, including acquisition of Deep Mind, Britain's most advanced artificial intelligence firm, and hiring of well-known AI experts (Ray Kurzweil, Geoffrey Hinton) to stand at the center of the market. Google's groundbreaking artificial intelligence plan is based on the cloud infrastructure that supports it, the world's largest data center and big data, and there are a myriad of examples of artificial intelligence-based services that Google is researching and investing in.
Currently, the typical operating system in the mobile field is Apple (IOS) and Google Android (Android) is no exaggeration to say that is widely used. Google is currently in service, including Google Now, after the Android operating system codenamed Jelly Bean (4.1), as a countermeasure against Apple's voice recognition feature, SIRI. Unlike conventional voice recognition, GoogleNow uses new location information to update new information and informs the user in advance.

3. Malicious code classification system

3.1. Tensorflow system structure diagram

Convolution Neural Network Open Source Developed malware learning and classifier using Google's tensor flow.

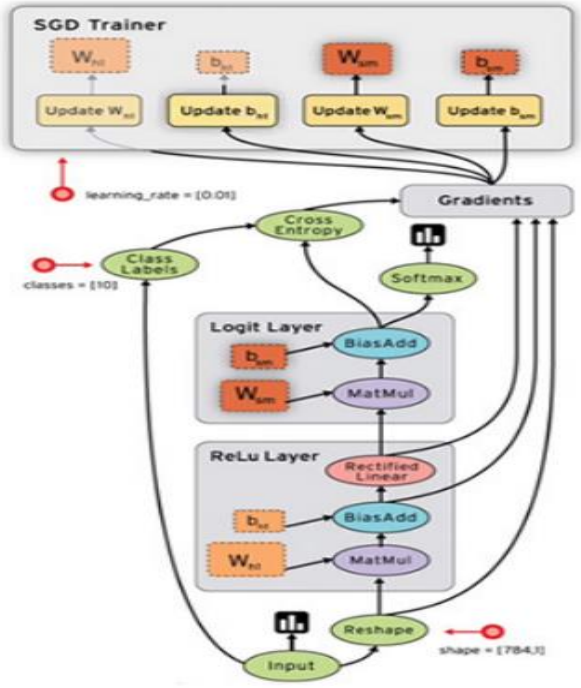


Fig. 2: Flowchart of Processing Tensor Flow Data.

The overall system structure of the malware learning and classifier including the malicious code imaging program and the basic type tensor flow neural network library developed for image processing is as following figure 4.The virus code used as input was the virus code used in the BIG 2015 Challenge. MS The challenge dataset (Called Kaggle Data Set) is provided by 10,868 learning malware bytecode and asm code.

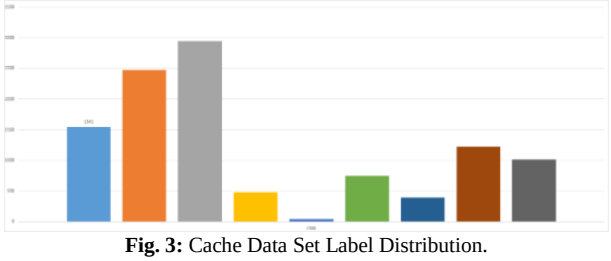


Fig. 3: Cache Data Set Label Distribution.

The figure 3 shows the label distribution table of 10,868 data sets. These 10,868 malicious codes were learned by using the following program.

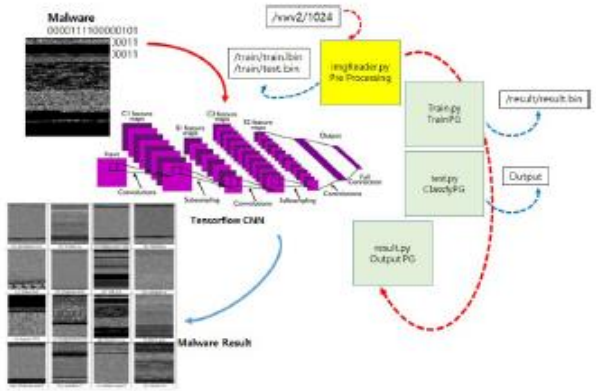


Fig. 4: Overall Structure of Development

3.2. Tensorflow imaging

A given malware binary consists of a 2D array read as a vector of unsigned 8-bit integers. This range can be easily visualized as a grayscale image in the range [0, 255] (0: black, 255: white).The width of the image is fixed and the height can vary depending on the file size.[4]The following figure provides the recommended image width for different file sizes based on empirical observations. The following figure 5 shows an example image of a typical trojan downloader Dontovo A that downloads and executes arbitrary files from 10,868 virus codes. In many cases, other sections of malware (binary fragments) represent a unique image texture. You can see five sections of malware.

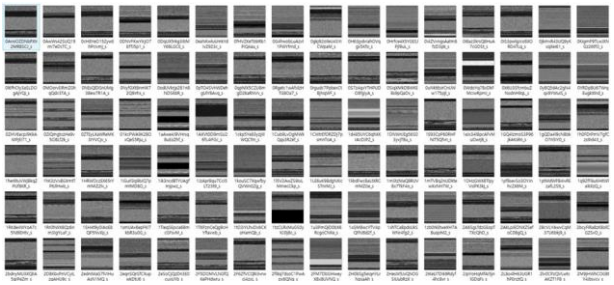


Fig. 5: Virus code image.

3.3. Learning and classification of tensorflow

We developed a classifier by using convolution neural network result.bin generated after the learning was finished in the above learning stage. The tensor flow learning and classifier consists of a convolution layer, a pooling layer, and a Full-Connect layer. Figure 6 shows the structure of a system that utilizes this typical convolution neural network.

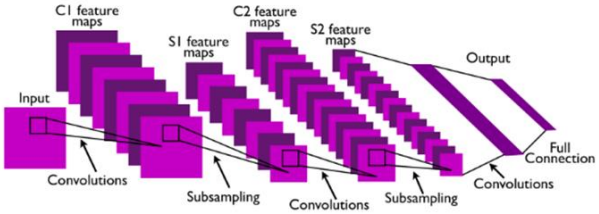


Fig. 6: Neural Network Structure Diagram

A JPG image processing program using a given MS cache data set, an imgReader step for inputting it as a tensor flow image, a train step for learning the accepted image on a neural network, and a test step for testing the classification accuracy of the malicious code after learning has been completed Based tensor flow malware image learning and classifier.

3.4. Experimental results

The results of the 1st machine learning malware learning and classifier using the cache data set is shown in table 1.

Table 1: Results of the First Experiment

Division	Case 1	Note
Method	Fine tune	
Validation	70	
Steps	1000 / 500	
Batch size	64	
Learning rate	0.01 / 0.001	
Weight decay	0.00004	
Model	InceptionV1	
Result	50%	First time

In order to increase the accuracy of 50%, various parameters of the tensor flow must be appropriately tuned. Table 2 is an experimental table of major tensor flow parameters such as Weight Decay from Max Steps.

Table 2: Experimental Results of Tuning Procedure

Division	Case 1	Case 2	Case 3
Method	Scratch	Scratch	Scratch
Validation	1500	1500	1500
Steps	1000	1500	2600
Batch size	16	16	32
Learning rate	0.01	0.01	0.01
Weight decay	0.00004	0.00004	0.00004
Result	66%	54%	57%

However, as shown in table 2, the performance of this learning and classifier stayed at 66% and the final accuracy was 97% by image ensemble and various parameters tuning as shown in table 3.

Table 3: Final Test Results

Division	Original	image ensemble
Method	Fine tune	Fine tune
Validation	600	600
Steps	1000 / 500	1000 / 500
Batch size	64	64
Learning rate	0.01 / 0.001	0.01 / 0.001
Weight decay	0.00004	0.00004
Model	InceptionV1	InceptionV1
Result	82%	97%

4. Conclusion

A study on the analysis of malicious codes using machine learning has been made from two core problems. The first is the method of imaging. CNG's accuracy and parametric programming tuning are also important for the 10,868 classifications of MS Cagle malware code data that the BIG 2015 winner showed 99.6% accuracy. The second is the construction of Convolution Neural Network (CNN). The construction of CNN using open source Tensorflow and Python can be categorized as an image processor, a learning machine, a tester, and a result output section. According to the Tensorflow recommendation, a factor for accurate classification requires 1 million convolution operations. However, this requires too much time for the operation of the learning machine and the tester, which greatly affects the performance of the system. Considering much of the experimental results, it is judged that the accuracy of the convolution operation is in conflict with the accuracy of the classification result.

To overcome this problem, we introduced the incentive model of the tensor flow SLIM library and achieved more than 80% accuracy. Especially, when adding a specific algorithm, the accuracy was 97%.

Based on these results, it is confirmed that the malicious code classification system using machine learning can be an effective way to satisfy the requirements in terms of accuracy and usefulness of the results.

References

[1] Malware Images: Visualization and Automatic Classification, Nataraj, S. Karthikeyan, University of California, Santa Barbara, 2010 ACM 1-58113-000-0/00/0010.

[2] ImageNet Classification with Deep Convolutional Neural Networks, Alex Krizhevsky, University of Toronto.

[3] TensorFlow: A system for large-scale machine learning, Mart´ınAbadi, Paul Barham, 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16), November 2–4, 2016 •Savannah, GA, USA,, https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf.

[4] Xiaofang, Ban, Chen Li, Hu Weihua, and Wu Qu. "Malware variant detection using similarity search over content fingerprint", The 26th Chinese Control and Decision Conference (2014 CCDC), 2014.