ATLANTIS PRESS

Advances in Engineering Research, volume 150

4th International Conference on Machinery, Materials and Computer (MACMC 2017)

Android malicious code detection and recognition based on depth learning

YangJing

Hubei University of Police, Wuhan 430034,China

Key words: malicious code; detection algorithm; depth learning; neural network; Andriod terminal

Abstract. At present, most malicious code detection methods are based on the shallow machine learning model. These shallow machine learning methods are simple in the modeling process, and restrict the complex functions and classification problems. In order to improve the accuracy of Andriod malicious code detection and recognition, an algorithm of malicious code detection and recognition of the deep learning has been put forward in this paper, this algorithm based on neural network training and learning model. Through the learning and training of malicious code sample data, the static, dynamic characteristics and malicious application characteristics of malicious code data are analyzed, including privilege feature, API feature and OpCodes characteristic data.
The comprehensive performance of the algorithm was tested, the test results indicate that using depth learning detection algorithm, Andriod malicious code identification accuracy is higher, and false detection rate and undetected rate are low, which is a highly efficient and reliable malicious code detection and recognition algorithm.

Introduction

With the continuous development of the deepening process of informatization and Andriod terminal equipment, the malicious code is one of the major threats to the network security, due to the use of economic benefits and a variety of new technologies, the number of malicious code is growing exponentially, while a variety of malware variants lead to security threats emerge in an endless stream, increased year by year. In order to solve the problem of detection and classification of malicious code on the current network environment, experts and scholars in various countries have carried on the positive research on the detection algorithm, has made many research results, but most of the detection algorithm is based on the machine learning algorithm is shallow, if the deep learning algorithm is introduced to the detection algorithm, the accuracy rate of malicious code detection can be improved, which has important significance for the study of Andriod malicious code.

The malicious code types and main detection methods of Android mobile terminal

Android mobile terminal malicious code generally refers a set of tasks for the terminal equipment to execute according to the attacker's wishes, it can also be considered as a virus and Trojan program for performing malicious tasks on Android mobile terminals. With the updating and the rapid development of mobile terminal equipment, malicious code is beginning to spread beyond what people imagine. At present, Andriod mobile terminal has more kinds of malicious code, the variation of species change constantly, therefore, the characteristics of its attack are also varied.

Malicious code type. Malicious code can usually be divided into two categories, one class can run independently, and the other cannot operate independently. The running code does not need a host,

Copyright © 2018, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

and the system can call directly. Code that does not run independently requires relying on other system programs and environments that cannot exist independently of the system. According to whether it has the ability of self replication, malicious code can also be divided into not self replication and can replicate two kinds of malicious code, malicious code can be self replicating the independent procedures, such as worms, not self replicating malicious code activation in the application program and system call.

Method of malicious code analysis and detection. The detection of malicious code is mainly based on the return of malicious code technology, but this technology development often lags behind the malicious code technology itself, the main detection methods for manual testing, software testing and data integrity testing. With the development of network technology, the network technology of malicious code detection algorithm based on intelligence, such as neural network detection algorithm, fuzzy recognition algorithm, the detection algorithm is generally divided into static detection and dynamic detection algorithm, the following detailed introduction.

(1) Static detection method

The static detection method usually uses disassembly detection technology, and usually performs the detection without performing binary code. It is a detection method based on Reverse Engineering technology.

1) static disassembly detection, disassembly detection is to use the malicious code that has been generated, debug it with debugger, and then detect it according to the hint information.

2) static source code detection, in the case of known malicious code binary source code, the malicious code related information analysis, to determine its function, operation flow and attempts, etc..

3) decompile detection refers to the form of malicious code that has been restored, restored to the form of source code, and then executed code to detect the code.

(2) Dynamic detection method

Dynamic detection method is the main malicious code to program operation of the project, using the program running and debugging tools, the implementation process of the program to observe and track, determine the malicious code execution process, and the static verification.

1) System call behavior detection method

When the code is malicious code, we are generally the normal code comparison, the normal code can be accumulated according to the usual experience, and gradually establish a secure database code, in the process of dynamic detection, and security features can be a number of code database data comparison, if the data and database security the data difference may be due to malicious code, potential malicious attempt to repeatedly verify the, determine whether it is a malicious code.

2) Heuristic scanning technique

The technology is a complement to the feature scan detection technology, and the heuristic scanning detection technology can make the detection process intelligent and enlightening, and can detect and identify malicious code independently.

Malicious code detection and recognition method based on depth learning

The detection of malicious code is the use of the three characteristics of the sample data, including permission features, API features and OpCodes features, the use of deep learning algorithm can learn the characteristics of the malware samples, which can fast and efficient detection code, the detection method of the basic framework as shown in Figure 1.

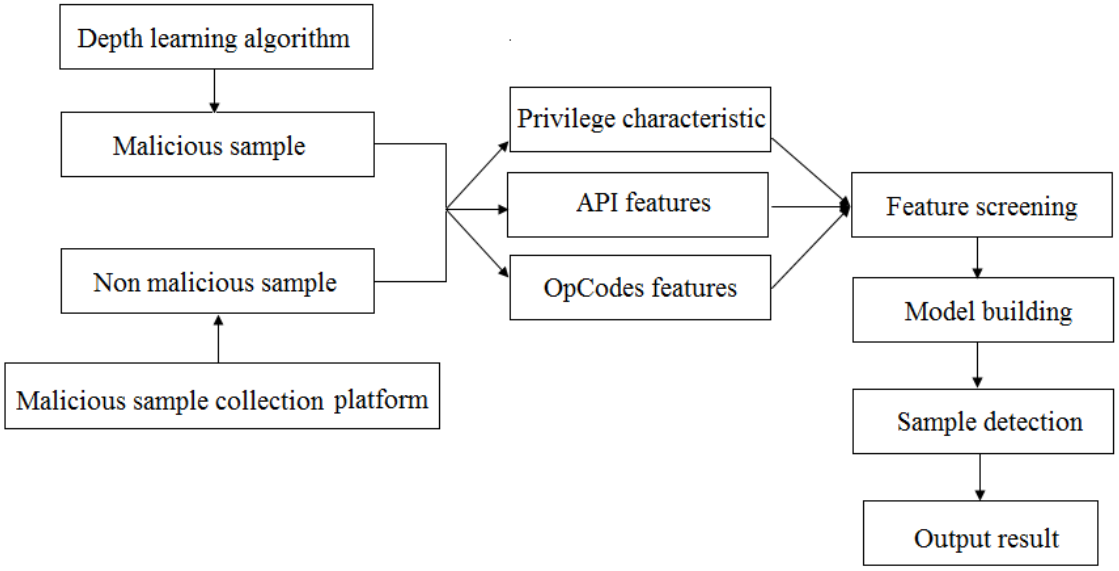

Fig. 1 Basic framework of malicious code detection and recognition method based on depth learning

As shown in Figure 1, deep learning algorithm can learn the malicious code based on sample data of Drebin, Sanddroid, Kaggle malicious code data set of malicious software and other open source data, malicious code automatic identification, feature extraction, detection model. Firstly, 30 malicious samples are selected from the malicious code database, and 6 samples are selected as training samples. Process is as follows.

(1) The preprocessing stage of sample data. For sample data, unified data processing is needed first, and the data is transformed into data that can be processed uniformly.

(2) The neural network depth learning model is established. The output quantity of neuron is 30, and the expression is divided into 30 kinds.

(3) Using different stacked modules, and then constructing a depth neural network.

(4) The training of neural network model, in order to speed up the convergence speed and make the neural network training more effective, we can use the batch method.

(5) The pre training data as the basis of the data obtained after the detection of the sample data, characteristic data detection including permission features, API features and OpCodes features of the data, the feasibility and reliability of the detection method is verified by the analysis results.

Malicious code detection and recognition method performance test

In the verification process of detection methods, the characteristics of three kinds of data were detected and three kinds of characteristic data of mixed detection, the accuracy rate of malicious code detection is verified, as shown in Table 1 the results obtained through analysis and verification.

Table 1 The accuracy comparison table of feature detection

Authentication method	Detection, analysis and identification accuracy
Privilege characteristic data	0.875
API characteristic data	0.682
OpCodes characteristic data	0.925
Three data mixing methods	0.987

The test results can be seen from table 1, the deep learning detection algorithm can successfully detect Andriod malicious code, and use the feature data under different situations, the recognition rate is different in using test data, the highest OpCodes lowest characteristics of sensitive permissions, and this is not because each APK file has a OpCodes feature data. In the hybrid approach using three data, the accuracy rate of detection reached 98.7%, the detection accuracy is high, so the data using three kinds of mixed detection, in order to further verify the reliability of deep learning algorithms, the detection of malicious code, malicious code error probability TPR probability FDR and the failure probability of MA the test, such as test results shown in table 2.

Table 2 test results of malicious code inspection indicators

Index	Detection, analysis and identification accuracy
TPR	0.95
FDR	0.0086
MA	0.085

It can be seen from table 2 that, the deep learning algorithm, whether comprehensive index or indexes, the performance of detection and analysis of malicious code is good, the accuracy of identification and the false detection rate and false negative rate is low, which verifies the reliability of deep learning algorithm.

Conclusion

In order to improve the detection efficiency of Andriod malicious code, accurate identification of increase code rate, the deep learning algorithm is introduced to the detection and identification code, and based on the neural networks model, the malicious code samples has been studied, finally through the detection of the sample code, the reliability of the algorithm has been verified. The test results indicate that the permissions feature API feature and OpCodes feature data, the permission characteristic data detection accuracy is highest, using OpCodes data obtained from the accurate detection rate of the lowest, comprehensive performance test index, using deep learning detection algorithm for Andriod malicious code detection accuracy is high, can meet the needs of efficient code detection and recognition, and it is an advanced Andriod code detection and recognition method.

Acknowledgements

The work was supported by the project of Natural Science Foundation of Hubei Province in 2017 with the project number 2017CFB745 and the project name *Android malicious code detection and recognition based on depth learning.*

Reference

[1] Xi Rongrong, Yun Xiaochun, Jin Shu Yuan, et al. A survey of network security situation awareness [J]. Journal of Computer Applications, Vol.32 (1), (2012), p.1-4.

[2] Hu Liang, Zhao Jianming, Xie Nannan, et al.. Detection and visualization of regular tree based on multi-step attack [J]. Journal of Image and Graphics, Vol.18 (3), (2013), p.299-304.

[3] Guo Chen, Liang Jiarong, Liang Meilian. Virus detection method based on BP neural network, [J]. Computer Engineering, Vol.31 (1), (2005), p.152-154.

[4] Wang Songtao, Wu Hao. Research on kernel rootkit detection technology based on executable path analysis under Linux[J]. Computer Engineering and Applications, Vol.41(11), (2005), p.121-123.

[5] Chen Mianshu, Chen Hexin, Sang Aijun. Computer face recognition technology [J]. Journal of Jilin University (Information Science Edition), Vol. S1, (2003), p.101-109.

[6] Zhou Jie, Lu Chun Yu, Zhang Changshui, et al. Overview of automatic face recognition methods [J]. Acta Electronica Sinica, Vol.28 (4), (2000), p.102-106.

[7] Xiao Weidong, Sun Yang, Zhao Xiang, Zhou Cheng, Feng Xiaosheng. Summary of research on hierarchical information visualization technology [J]. Journal of Chinese Computer Systems, Vol.3201, (2011), p.137-146.

[8] Yang Hao, Nurbol, Xu Huan, Hu Liang. Visualization Representation system based on Intrusion Scenarios [J]. Journal of Chinese Computer Systems, Vol.10, (2010) , p.2059-2064.

[9] Jiang Jianming, Zhou Dibin, Hu Bin. Summarization of vector visualization research [J]. Bulletin of Science and Technology, Vol.26, No.15404, (2010), p.611-617.

[10] Yu Xiao Sheng. Study on dimensionality reduction methods for high-dimensional data[J]. Information Science, Vol.25 (8), (2007). p.1248-1251.

[11] Chen Peng, Lv Weifeng, Shan Zheng. Based on the network into the radiance detection method [J]. Computer Engineering and Applications, No.10, (2009), p.44-48.

[12] Hong Fei, Wu Zhimei. Adaptive estimation method of Hurst Exponent Based on wavelet[J]. Journal of Software, Vol.16 (9), (2005), p.1685-1689.

[13] Lv Liangfu. Survey of network security visualization[J]. Journal of Computer Applications Vol.26 (8) , (2008), p.57-62.

[14] Zhang He, Lu Wuyi. OPC client and real time database communication[J]. Computer Engineering & Science, Vol.30 (5), (2008), p.81-83.

[15] Yang Mingji, Guo Jianhong, Shen Qiang, Zhou Qiang,.OPC technology research in industrial control field[J]. Journal of Harbin University of Science and Technology, Vol.13(4), (2008), p.29-31.

[16] You Wenjian. Method of using Winpcap to capture network underlying data packets [J]. Science & Technology Information, (2009).

[17] Cao Yu. On service innovation and application of information technology in library development [J]. Journal of Jilin Institute of Chemical Technology, Vol.28 (6), (2011), p.73-75.