Literature Review of Visualizing and Detecting Malware

Zhaoheng Yin

School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON, K1N 6N5, Canada

ABSTRACT

Visualization detection is an emerging technology by extracting the characteristics of malware image performance, has solved the traditional detection technology can not accurately detect the unknown malicious code problem, the method has received more and more attention, especially in the context of the continuous development of anti-detection technology, this paper on the different types of malware code visualization detection methods to conduct a comprehensive research and analysis from the grayscale image transformation, machine learning visualization Deep Learning Visualization and Hybrid Detection, summarizes and analyzes the current problems and future development direction of visual detection, aiming to help scholars improve the accuracy and efficiency of visual detection methods.

KEYWORDS

Visualizing detection; Malware; Deep leaning; Machine learning

DOI: 10.47297/taposatWSP2633-456921.20230402

1 Introduction

In today's digital era, malware has become the primary threat to Internet security, and with the increasing number and types of malwares, traditional malware detection methods can no longer satisfy. Therefore, artificial intelligence technology is widely used in the field of malware detection. Compared with traditional malware detection methods, AI technology has higher accuracy and efficiency. Malware detection mainly includes two methods: static analysis and dynamic analysis. Static analysis refers to analyzing the code of malware to decide whether it contains suspicious code or functions. Dynamic analysis, on the other hand, monitors and analyzes the malware at runtime to detect whether its behavior is abnormal. Visualization technology can play a key role in both methods. Visual detection technology refers to the graphical presentation of data and information to help users understand and analyze malicious code data more intuitively, to discover the patterns and anomalies hidden behind the data. In malware detection, visualization technology can transform massive amounts of data and information into forms such as graphs and images for easier observation and analysis.

2 Research on Software Visualization and Detection Methods

(1) Grey-scale image visualization and detection process

Grayscale image visualization detection process [1] is basically divided into the following three layers, the first layer is the pre-processing of malicious code, mapping each byte of the malicious code to each pixel of the grayscale image, writing the malicious file binary bytes between 0 and 255 in

order to correspond to the pixel intensity, as a way to convert the code into a two-dimensional image, because the size of the image formed by different families of malicious code will be inconsistent, and will be used graphical tools are used to limit the pixel file length and width. The second layer is to build a model or vector machine to classify the malicious images, mainly focus on ad-hoc extraction. The third layer feeds the new malicious code into the model for analysis and continuous optimization. Since grayscale images contain only two channels, it is not possible to represent the data in its entirety, thus omitting significant image features, thus deriving the RGB three-channel color visual detection technique, which not only retains the information of the grayscale image, but also emphasizes the similar texture, color, and structural features, etc., in the color image.

(2) Visualization Detection by Machine Learning

Malicious code grayscale visualization detection was originally proposed by Conti G. et al. in 2010 for binary mapping analysis [2] to convert large unknown files into grayscale images, Conti et al. argued that automated mapping can speed up both manual and automated analysis activities and can be applied to many low-level fragment classification techniques. However, the limitations of the method are that sample size is important for classifiers based on statistical techniques (Shannon entropy, Hamming weights, means and chi-square tests) and that it is inaccurate when dealing with complex binary objects. Subsequently in 2011 by Nataraj L. et al [1] from the University of California, improved and implemented the malicious code visualization characterization method, which visualizes the malicious code .text block as a grayscale image, cuts the original binary sequence into subsequences of 8-bit length, and each byte of the malicious code can be represented as an integer between 0 and 255, and this integer can be directly mapped to a pixel of the grayscale image, the gray value of the pixel also ranges from 0 to 255 and the size of the image depends on the length of the malicious code. The processed malware images are used to extract features using search the GIST algorithm and then classified using algorithms such as Support Vector Machine SVM. This method utilizes the similar two-dimensional grayscale image textures presented by the same malware family to analyze malicious code files without parsing the PE file format and without dynamically executing the code and has been widely adopted in new detection engines. Han, K. S.[3] et al. improved the Nataraj method by proposing the idea of entropy maps, which discards the comparison between the grayscale image textures in favor of a The overall code entropy map similarity measurement, by converting binary files into bitmap images, after calculating the entropy value of each line of the bitmap to generate entropy maps, and then analyze the entropy map similarity characteristics of the detection and classification of malicious code, the method is shorter compared to the Nataraj's GIST method, which proves that through the entropy map intuitively represent the internal characteristics of the software code in a low dimensionality in the recognizable state to improve the similarity comparison efficiency. However, the method is unable to compose the packed binary data and needs to detect the packed files with the assistance of dynamic analysis.

(3) Visualization Detection by Deep-Learning

In 2020 Intel, in collaboration with Microsoft Labs, proposed the STAMINA model [4], which proceeds through a transfer learning algorithm, where a pre-trained model is used in deep learning to initialize the network parameters and fine-tune them on a new task to fit the specific needs of the current task. This approach speeds up training, reduces parameter and architecture search, and maintains high classification performance on relatively small datasets. The main migration learning as feature extraction and fine-tuning networks, as feature extraction, migration learning uses a preprocessing model to remove the fully connected layer to use the remaining part as a feature extractor. Regarding

the fine-tuning network, the weights of the preprocessing network are fine-tuned by continuous backpropagation. This method is less efficient and difficult to convert billions of pixels into un-JPEG image format when dealing with large data, by changing the pixel size, making the model unable to accurately analyze and detect malicious code.

By Han K., Kang, B.[5] et al, proposed the method of visual detection using image matrix, extracting the sequence of opcodes from a binary file, extracting only the blocks in the opcodes that contain the CALL instruction in order to find the features, facilitating the generation of pixels in image matrix by unifying and simplifying the opcodes to 3 characters, determining the X-Y coordinates of the matrices and the RGB pixel colors by using the hash function while generating a pair of pixels for each family of malicious codes into separate image matrices, thus reducing the time to compute the similarity. The method makes full use of vectorized values of image color pixels expressed in matrix form to improve the performance of similarity calculation.

Scholars at Tsinghua University proposed [6] to transform the two-dimensional grayscale image into a three-bit RGB color image, quasi-search for the original binary file byte stream information in accordance with the grayscale image to fill the R dimension, the binary data in the ASCII and other character information is input to the G-dimension, and the PE file structure to fill the B-dimension to construct a three-dimensional RGB image. In the image classification, the visible pyramid pooling model SPP-net is used to avoid destroying the data and code structure by fixing the image size to two and to better represent the original data of the malicious code. The authors use the VGG16 structure as a pre neural network to extract convolutional features for image classification and set up 4 layers of pooling for feature extraction with a combined total of 50 dimensional features that are fed into the fully connected layer. The method is optimized for the use of PE structure in Nataraj's method, and the single information region is optimized while designing the code image enhancement, which improves the interpretability, robustness, and accuracy for the model.

Sun, G.[7] fuses recurrent and convolutional neural networks to consider not only the raw information about the malware, but also the ability to associate the raw code with temporal features; in addition, the process reduces the dependence on malware category labels. A min-hash is then used to generate feature images from the fusion of the original code and the predicted code from the RNN, and then a CNN is trained to classify the feature images.

(4) Visualization on Hybrid Detection

Employing different image processing techniques together with machine learning and deep learning, Asla Ö. et al [8] proposed the use of Hybrid Deep Neural Network Architecture which combines two pre-trained ResNet-50 architecture and AleNet architecture to collect gray scale images converted from binary files using exhaustive lifting and then extracting the low and high level features using pre-trained Hybrid Neural Networks. Finally, the training of the deep neural network architecture is performed based on supervised learning. The framework adopts a hybrid model approach, which optimally integrates two widely used pre-trained network models to provide more accurate classification capability with significant reduction in feature dimensionality space, resulting in improved classification efficiency and performance. Vasan D. et al [9] developed IMCEC based on multiple CNN neural network models and combined it with binary visualization, which uses a hybrid integrated learning technique that combines multiple classification waveforms to achieve higher predictive performance and produce more powerful classifiers.

3 Problems and New Technological in Visualization and Detection

(1) Limitations of visualization techniques

False alarm rate caused by too similar textures, if the binaries of two malicious code families are too similar, resulting in presentation of visualized images with slightly consistent textures, it will make the classifier classify the two malicious codes into one category, in addition, if the samples contain interfering information, or deliberately disguised code, it will also affect the texture characteristics of the images. When targeting large amounts of data are targeted without label classification, without effective data preprocessing and dimensionality reduction techniques, the visualization results can be very confusing and fail to provide useful information. In addition, it is extremely difficult to accurately detect malicious code in huge data only by through static visualization analysis, and the visualization of large data not only requires a large number of computational resources, but also may lead to visual confusion and loss of information. Therefore, how to effectively process and visualize large-scale malicious code data is an urgent problem to be solved. Liu, S. et al [10] proposed that the characteristics of malicious code are often high-dimensional, e.g., they may include a variety of characteristics such as code length, code complexity, API call sequences, network behavior, and so on. Visualization in high-dimensional data is extremely difficult and can also easily lead to loss or misinterpretation of information, and representing high-dimensional data in a two or three-dimensional space requires the research into more efficient algorithms.

(2) Data set problems

Currently, most of the training sets in the literature are derived from the 2015 Kaggle Microsoft Malware Classification Challenge, followed by Malimg, Malheur, and VirusShare. due to the insufficient training samples, which affects the model's classification performance of the model and the evaluation results. For example [7], when the ratio of sample size of the training dataset to the validation dataset is 1:30, only more than 92% accuracy can be achieved; while when the ratio is adjusted to 3:1, the accuracy can be more than 99.5%, and the malware samples should be further expand in the future to support more model training and learning.

(3) Expanding new technologies

In the recent years, a larger number of algorithmic generative models for statistical classification have been introduced in the past few years, which predict labels by learning the joint probability distribution of the data, including plain Bayes, Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), Generative Adversarial Network (GAN), and so on. Generative Adversarial Network (GAN) is a deep learning model, first proposed by Ian Goodfellow et al [11] in 2014, which consists of two parts, the generator, and the discriminator, which are trained against each other during the training process. Specifically, the generator tries to generate fake data that is good enough to deceive the discriminator, while the discriminator tries its best to distinguish real data from fake data. In this process, both the generator and the discriminator constantly improve their abilities, and eventually the generator can generate very realistic fake data. GAN is widely used in various fields, including image generation, super-resolution, style migration, and image-to-image conversion. Already scholars Cheng Jinyin et al [12] based on GEN adversarial sample generation framework combined with the concept of visual detection of malware, proposed the use of GAN's black-box attack against the deep learning malware detector, which pointed out that adversarial generation of samples generated by the random perturbation maps overlaying the original image will lead to the disorder of the original data, and still

need to be optimized against the adversarial network and so on. In addition with 2023 Li, Tianhong [13] et al. proposed MAGE, a self-supervised learning framework based on image semantic character mask, the algorithm solves the problem that GAN and other generative models are difficult to extract high-quality image feature capabilities, MAGE for the first time to achieve a unified image generation and feature extraction model, compared with similar MIM methods have significant improvements, and to achieve the optimal self-supervised level, the algorithm has high potential value in the field of malicious code visualization detection.

4 Conclusion

This paper provides a comprehensive analysis and summary of the research field of malware visualization detection. In this paper, we first introduce the basic concepts of malware detection, and then analyze the technical approaches of malware visualization detection from the perspectives of machine learning and deep learning, including data preprocessing, feature extraction, visual representation, and classifier models. At the same time, we also compare the advantages and disadvantages of various techniques so that readers can better understand and choose the appropriate technique. In addition, this paper discusses the challenges faced by malware visualization detection. Among them, large data volume, high data complexity, high real-time requirements, and lack of labeling are the main challenges currently faced. And with the development of artificial intelligence and big data technology, malware visualization detection is expected to achieve more efficient and accurate detection. Malware visualization detection is a very promising research field with a wide range of application potential, and there will be more innovations and breakthroughs in the future to better protect network security.

About the Author

Zhaoheng Yin is undergraduate of School of Electrical Engineering and Computer Science, University of Ottawa, and his research field is Computer Vision.

References

[1] Nataraj, L., Karthikeyan, S., Jacob, G., & Manjunath, B. S. (2011, July). Malware images: visualization and automatic classification. In Proceedings of the 8th international symposium on visualization for cyber security, 1-7.

[2] Conti, G., Bratus, S., Shubina, A., Sangster, B., Ragsdale, R., Supan, M., ... & Perez-Alemany, R. (2010). Automated mapping of large binary objects using primitive fragment type classification. digital investigation, 7, S3-S12.

[3] Han, K. S., Lim, J. H., Kang, B., & Im, E. G. (2015). Malware analysis using visualized images and entropy graphs. International Journal of Information Security, 14, 1-14.

[4] Chen, L., Sahita, R., Parikh, J., & Marino, M. (2020). STAMINA: Scalable Deep Learning Approach for Malware Classification.

[5] Han, K., Kang, B., & Im, E. G. (2014). Malware analysis using visualized image matrices. The Scientific World Journal, 2014.

[6] Sun, B., Zhang, P., Cheng, M., Li, X., Li, Q. (2020). Malware detection method based on enhanced code images. Journal of Tsinghua University (Science and Technology), 60(5), 386-92.

[7] Sun, G., Qian, Q. (2018). Deep learning and visualization for identifying malware families. IEEE Transactions on Dependable and Secure Computing, 18(1), 283-95.

[8] Aslan, Ö., & Yilmaz, A. A. (2021). A new malware classification framework based on deep learning algorithms. Ieee Access, 9, 87936-87951.

[9] Vasan, D., Alazab, M., Wassan, S., Safaei, B., & Zheng, Q. (2020). Image-Based malware classification using ensemble of CNN architectures (IMCEC). Computers & Security, 92, 101748.

[10]Liu, S., Maljovec, D., Wang, B., Bremer, P. T., & Pascucci, V. (2016). Visualizing high-dimensional data: Advances in the

past decade. IEEE transactions on visualization and computer graphics, 23(3), 1249-68.
[11]Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27.
[12]Vasan, D., Alazab, M., Wassan, S., Safaei, B., & Zheng, Q. (2020). Image-Based malware classification using ensemble of CNN architectures (IMCEC). Computers & Security, 92, 101748.
[13]Li, T., Chang, H., Mishra, S., Zhang, H., Katabi, D., & Krishnan, D. (2023). Mage: Masked generative encoder to unify representation learning and image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2142-52.