Research on P2P Botnet Traffic Identification Technology Based on Neural Network

Zhibin Pei and Gang Gan
School of Cyberspace Security, Chengdu University of Information Technology, Chengdu 610225, China
Email: test_me@cuit.edu.cn

Abstract. The identification of botnet traffic is an important link in the network security. With the continuous improvement of the traditional botnet network traffic identification technology, the identification of the P2P botnet traffic becomes the hot point of the research. The P2P network is a decentralized network model, and the nodes in the network can act as the requesters of the network service, and can respond to the requests of other computers to provide the resources, services and contents, so the difficulty of the malicious traffic identification of the P2P network is greatly increased. The machine learning technology has a wide application in the field of botnet flow identification, but the artificial feature extraction becomes more and more difficult with the change of the shape of the botnet and the control mechanism of the command. To this end, a method for identifying the network traffic of the P2P botnet based on the ResNet convolution neural network is proposed. The experimental results show that the model has good performance and can accurately identify the botnet traffic.

1. Introduction
Nowadays, with the high speed of Internet operation, P2P [1] traffic ranks first with 70% to 80% share of Internet traffic. P2P-based applications are widely used in streaming media transmission, file sharing, instant messaging, games and software. Botnet is a malicious network composed of a large number of hosts infected by botapps. P2P botnet belongs to a branch of the whole botnet family, which combines the basic characteristics of P2P network and botnet. In botnet, zombie programs lurk in the background after infecting the host computer, waiting for the command of the controller. Once the botnet controller issues the command, these zombie programs can launch large-scale cooperative attacks, which is very harmful. Therefore, how to effectively identify P2P botnet traffic from network traffic is an important problem in network security.

In the field of botnet flow identification, the machine learning technology has been widely used. The researchers used the algorithms of support vector machine, random forest and naive Bayesian to set up a model according to a variety of features to identify the malicious network traffic. But the characteristics of the artificial extraction also have some limitations. First, some suitable features are extracted manually, and the theoretical and empirical requirements of the researcher are relatively high. For example, the recent improvement of SVM [2, 3] algorithm and C 4.5 algorithm. Secondly, the extracted features have some common points, and an attacker can exploit the thought of the machine to learn to attack, modify some typical flow characteristics, thus bypassing the detection of the model. In the literature [4], an attacker is proposed to add well-designed data packets and data stream noise to the botnet traffic to reduce spatial similarity while adding random time delays to reduce temporal similarity. Third, the machine learning algorithm is inefficient and poor in traffic identification with no
labels. The paper [5] presents an algorithm based on artificial bee colony to optimize the support vector machine model by comparing the genetic algorithm and the particle swarm optimization algorithm, so as to improve the recognition efficiency of P2P flow.

At present, the depth learning technology has a wide application in the field of flow identification. Through the adjustment of the multi-layer neural network structure and a large number of parameters, the characteristics of the samples can be abstracted and extracted layer by layer, thereby reducing the complexity of the artificial extraction features and the low recognition efficiency. In this paper, a traffic identification model based on the ResNet convolution neural network is proposed for the P2P botnet.

2. Related Work

2.1. P2P Botnet Related Principles

Compared with the traditional botnet, P2P botnet [6] adopts distributed layout, which makes P2P botnet have better concealment and robustness. P2P botnet can realize a variety of attacks, including DDoS attack [7], spam attack, and click fraud attack, distributed malicious attack and so on. Botnet includes three necessary elements: botnet controller (Bot master), also known as attacker, is the owner and actual controller of the whole botnet, and is also the attacker who initiates malicious activities. Botnet (Bot): refers to the Internet host infected by botnet, which is generally the basic element of botnet. Command and control channel (C ≤ C channel): refers to the private "channel" for communication between botnet controllers and zombie hosts, through which botnet controllers issue commands to zombie hosts to implement remote control. The topology of distributed botnet is shown in figure 1. It uses P2P protocol to construct its command and control channel, so it can also be called P2P botnet. In distributed botnet, in addition to the communication between zombie host and control server, there is also communication behavior between zombie host and control server. There are no control servers in some distributed botnet, only autonomous communication between zombie hosts. This networking method does not have the problem of "single point failure", and has high flexibility and strong robustness.

![Figure 1. Distributed Botnet](image)

2.2. Correlation Principle of Convolution Neural Network

Convolution neural network is a kind of feedforward neural network, which can be used in image classification, target recognition, target detection, semantic segmentation and so on. CNN [8] is a kind of neural network which is specially used to deal with data with similar grid structure, such as image data. The basic structure of CNN network is composed of input layer, convolution layer, pool layer,
full connection layer and output layer, in which convolution layer, pool layer and full connection layer can be repeatedly composed according to the actual situation, as shown in figure 2.

![Figure 2. CNN Convolution Neural Network Model](image)

3. Methods
Residual Neural Network [9] is the most widely used CNN feature extraction network. The structure of ResNet accelerates the training speed of neural network and greatly improves the accuracy of the model. ResNet adds the idea of Highway Network to the network, where the previous CNN network structure was to make a nonlinear transformation of the input, while Highway Network allowed a certain proportion of the output of the previous network layer to be retained, allowing the original input information to be transmitted directly to the subsequent layer. Therefore, one layer in the neural network can learn the residuals of the last network output instead of learning the whole output, as shown in Figure 3.

![Figure 3. Residual Learning:a Building Block](image)

There are some problems in information transmission, such as information loss and loss in fully connected network [10] and traditional convolution neural network. It will also lead to gradient
disappearance or gradient explosion, resulting in deep networks unable to train. Resnet solves this problem to a certain extent. By bypassing the input information to the output and protecting the integrity of the information, the whole network only needs to learn the difference between the input and the output, thus simplifying the learning goal and difficulty. Resnet has a lot of bypass routes that connect the input directly to the back layer. This structure is called shortcut or skip connections.

![Figure 4. A Residual Network](image)

Because the traffic data of P2P botnet studied in this paper is one-dimensional, but the ResNet model is based on two-dimensional data processing, we need to make some changes to the ResNet model so that the model can deal with one-dimensional data. In this paper, a complete ResNet, is designed as shown in figure 4. It consists of four modules, each module is composed of multiple Blocks. The number of blocks contained in the four modules is 4, 4, 6 and 4, respectively. The convolution cores of the corresponding convolution neural networks are 16, 32, 64 and 128. Based on the research of ResNet, this paper proposes a traffic recognition model for P2P botnet, as shown in figure 5. Firstly, the traffic data of P2P botnet to be detected are input into the model, and a series of processing is carried out by ResNet, such as feature extraction, and then the output data is taken as the input of the fully connected network. Finally, the traffic classification of P2P botnet is completed by Softmax [11].

![Figure 5. Network Structure](image)

4. Experiment

4.1. Experimental Setup

In the experiment, the training times of neural network are set to 32, the learning rate is set to 0.001, and the batch processing is set to 128. The activation function used is Adam. For the Relu and Sigmod, optimizer

4.2. Evaluation Indicators

In the process of classifying the designed model, the test samples are divided into positive class and negative class, and the following four cases occur:

1) The real class (True Positive, TP), the sample is positive and the model recognizes it as a positive class.

2) True negative class (True Negative, TN), the sample is negative and the model recognizes it as negative class.
3) False positive class (False Positive, FP), the sample is negative and the model recognizes it as positive class.

4) False negative class (False Negative, FN), the sample is positive and the model recognizes it as negative class. Therefore, in the P2P botnet traffic identification model, according to the above four situations, the following evaluation indicators are established [12]:

| Name of evaluation index | Index description |
|-------------------------|-------------------|
| Accuracy rate (Acc)     | \( \frac{TP + TN}{TP + FP + TN + FN} \) |
| precision               | \( \frac{TP}{TP + FP} \) |
| Training time           | Time consumed in network model training process |

5. Result

In this study, we use the designed ResNet model to compare with the full connection network and CNN. The experimental results show that the accuracy of full connection is 88.87%, the accuracy of ResNet is 93.21%, and the accuracy of ResNet is 99.32%. Finally, the proposed method defeats some traditional neural network algorithms and obtains the highest accuracy. The experimental results are shown in figure 6.

![Figure 6. Representation of the Three Models](image)

6. Summary

In this paper, a P2P botnet traffic recognition model based on ResNet is proposed, which can automatically extract features from traffic data compared with the traditional machine learning algorithm, and has higher accuracy than the traditional neural network. Resnet model integrates the idea of local connection and weight sharing, to a certain extent, solves the problem of gradient disappearance or gradient explosion. Nowadays, the traditional botnet traffic recognition technology has been very mature. In the next step, we can focus on the botnet detection of special architecture, such as mobile botnet, anonymous botnet and so on.

7. References

[1] Su Yangyang, Sun Dongpu, Li Dandan, Sun Guanglu. P2P traffic identification method based on clustering and traffic propagation graph [J]. Computer Application Research, 2019, 36 (11): 3448 ≤3451 3455.

[2] Raman M R G, Somu N, Kirthivasan K, et al. An Efficient Intrusion Detection System based on Hypergraph Genetic Algorithm for Parameter Optimization and Feature Selection in Support Vector Machine[J]. Knowledge-Based Systems, 2017:1-12.
[3] Wang C, Zhang H, Ye Z. A peer to peer traffic identification method based on wavelet and particle swarm optimization algorithm [J]. International Journal of Wavelets Multiresolution & Information Processing, 2015, 13(06):87-88.

[4] CUI X, FANG B X, YIN L H, et al. Andbot: towards advanced mobile botnets [C]//The 4th Usenix Workshop on Large-scale Exploits and Emergent Threats. 2011: 11.

[5] Wang C, Zhang H, Ye Z, et al. A peer to peer traffic identification method based on support vector machine and artificial bee colony algorithm[C]// IEEE,International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications. IEEE, 2015:982-986.

[6] Ying Lingyun, Feng Dengguo, Su Prui. Botnet based on P2P and its defense [J]. Journal of Electronics, 2009, 37 (01): 31.

[7] Cheng Jieren, Luo Yihan, Tang Xiangzhuang, ou Mingwang. DDoS attack detection method based on LSTM traffic prediction [J]. Journal of Huazhong University of Science and Technology (Natural Science Edition), 2019, 47 (04): 32.

[8] Cheng Hua, Xie Jinxin, Chen Lihuang. The method of traffic identification of encrypted C / AM / C communication based on CNN [J]. Computer Engineering, 2019, 45 (08): 31 ≤ 34 41.

[9] Qiangchang Wang, Guodong Guo. Benchmarking deep learning techniques for face recognition [J]. Journal of Visual Communication and Image Representation, 2019, 65.

[10] Guo Zihao. Design and implementation of flow-based network user behavior inspection system [D]. University of Electronic Science and Technology, 2019.

[11] Duan Dandan, Tang Jiashan, Wen Yong, Yuan Kehai. Research on Chinese short text Classification algorithm based on BERT [J/OL]. Computer engineering: 1 12 [2019 12 17]. https://doi.org/10.19678/j.issn.1000-3428.0056222.

[12] Chen Liangchen, Gao Shu, Liu Baoxu, Lu Zhigang. Research progress and development trend of network encrypted traffic identification [J]. Information Network Security, 2019 (03): 19 / 25.