Identification of SSH Applications Based on Convolutional Neural Network

Liuyong He and Yijie Shi
State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications
No.10 Xitucheng Road, Haidian District,
Beijing, China
lyhebupt@163.com; yijieshi2000@bupt.edu.cn

ABSTRACT

SSH is an encrypted communication protocol. SSH tunnel may encapsulate some other unknown applications, which has a certain potential impact on network security, so it is necessary to identify these applications accurately. This paper uses Convolutional neural network to identify applications, which has the characteristic of automatic feature learning. Therefore, traffic classification algorithm based on deep learning is used to identify these encapsulated applications in SSH traffic such as payload. Experimental methods and results are described in this paper and indicate that classification accuracy of applications (such as Nmap, Baidu Network, Netease cloud music, Netease cloud notes and etc.) in SSH tunnel is up to 95%.

CCS Concepts

Security and privacy→Network security→Security protocols

Keywords

SSH tunnel; application classification; Encrypted traffic; CNN.

1. INTRODUCTION

With the rapid development of the Internet, the network is expanding, network users are increasing, and network applications are adding. To meet the needs of Internet users, some new applications are emerging, gradually replacing the traditional network applications and occupying a large proportion of the network traffic. Internet traffic classification is not only an important means for network administrators to perform traffic shaping, network capacity planning, anomaly detecting, QoS deploying and network charging, but also an important basis for network researchers to carry out application protocol researches [4]. In order to make use of network resources effectively, to enhance network controllability, to ensure network security and to provide better service quality for users, it is necessary to identify and classify different applications [6]. The identification and classification can not only help network operators understand the distribution of network traffic, but also prevent cybercrime in time.

The SSH security shell protocol is to provide secure remote login and other secure network services on some insecure networks. SSH establishes an encrypted channel between the two communicating parties to protect the transmitted data from being eavesdropped and uses secret key exchange algorithm to protect the key itself [5]. Taking advantage of the characteristics of SSH connection, many people encapsulate some types of network applications into the tunnel by the SSH tunnel technique to avoid being intercepted by the firewall, and thus pose a certain threat to the security of network environment. The working principle is shown in Figure 1:

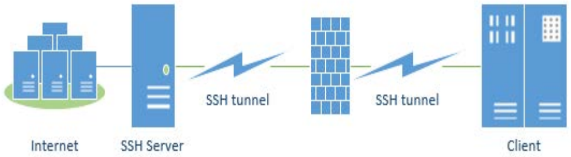

Figure 4. SSH tunnel working principle.

SSH tunnel is an encrypted traffic, the identification of which is different from non-encrypted ones. They are mainly:

1) After being encrypted, characteristics of traffic have changed greatly. Traditional methods like deep packet detection are no longer suitable for the identification [3];

2) Encrypted protocols are often accompanied by traffic masking techniques (such as protocol obfuscation and protocol variant [12]), which transform traffic characteristics into traffic characteristics of common applications;

3) Different encrypted protocols use different encrypting algorithms, so there is also a large difference between each encrypted protocol [11];

4) Current researches of encrypted traffic identification mainly focus on some special encrypted applications. There is still some difficulty in the fine identification of encrypted applications [10].

There are a number of research methods for traffic classification, which use valid information in IP layer, such as packet arrival time, transmitted bytes, and packet size [2]. At present, most of the traffic classification methods used in SSH tunnels are machine learning methods [9]. These methods depend on transmitting characteristics of the traffic, such as time interval between packets, packet size, repetition pattern, and then transmit these characteristics into a classifier like naive Bayesian, decision trees, and neural networks [1]. In general, the training process is off-line and time-consuming, but the process of using the model is real-

ICIEB 2018, April 25–27, 2018, Singapore, Singapore
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6375-4/18/04…$15.00
https://doi.org/10.1145/3230348.3230458

time or near real-time.

Maiolini G et al. proposed a SSH real-time identification method [8]. First, SSH protocol traffic was identified by SSH connection to the statistical characteristics of first data packet. Then, the application layer protocols (such as SCP, SFTP, and HTTP) of SSH protocol was identified by k-means clustering analysis.

Meng Jiao et al. identified SSH protocol based on the statistical characteristics of the traffic, and then by comparing the unsupervised machine learning method and supervised machine learning methods, analyzed the classification effects of application layer protocols in SSH tunnel with five machine learning methods, C4.5, SVM, BayesNet, K-means, and EM. However, these methods rely too much on artificial extraction of certain features and rely too much on expert experience [7]. The urgent need to achieve both the premise of the ideal rate of accuracy, but also automatically feature extraction.

Wang Zhanyi proposed using deep learning method to identify traffic [13]. He transformed traffic into picture, and used convolutional neural network method to identify the traffic data.

Based on the aforementioned related works, a traffic identification method based on convolutional neural network was proposed. The main research result is that identify different applications in SSH tunnel without decrypting or checking SSH traffic after analyzing the application traffic in SSH tunnel and using deep learning method. The second section is to introduce convolutional neural network. The experimental design and data processing method are introduced in the third section. Section four describes traffic classification results based on convolutional neural network after the experiment and evaluation. Section 5 is a summary and our further research plan.

2. RELATED WORK

Convolutional neural network is a kind of neural network with multi-layer supervision. Convolutional layer and pool sampling layer in the hidden layer is the core module to realize the extracting feature of convolutional neural network. The weight parameters are anti-adjusted layer by layer using gradient descent algorithm to minimize the loss function, and the iterative training is used to improve the classification accuracy.

The first layer of the convolutional neural network is input layer, followed by several convolutional layers and several sub-sampling layers and classifiers. Softmax is generally used as the classifier and then outputs the corresponding classification results through the classifier. Normally one convolution is followed by one sub-sampling layer. Based on the weight sharing and local connection in the convolutional layer, the sample training parameters of the network can be simplified. After the operation, the obtained result is output as characteristic picture by activating the function, and then the output value is taken as the input data of the sub-sampling layer. In order to keep scaling, translation and twisting unchanged, in the sub-sampling layer neighboring characteristics in the characteristic picture corresponding to the former layer are combined into one characteristic through the pooling operation to reduce characteristic distinguishing. In this way, input data can be immediately transmitted to the first convolutional layer, repeating the characteristic learning, and marked samples are input into the Softmax classifier. The overall network framework is shown in Figure 2:

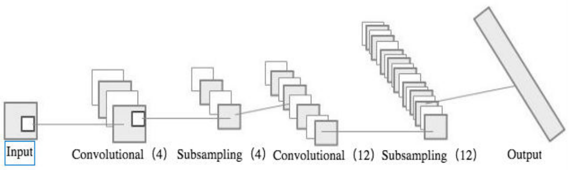

Figure 2. Framework of convolutional neural networks

The hidden layer of convolutional neural network is the core of characteristic extraction. In convolutional layer, each neuron input is connected with some part of the receptive field in the former layer and extracts characteristics of this part. The expression form of convolutional layer is:

$$x_j^l = f(\sum\nolimits_{i\in M_j} x_i^{l-1} k_{ij}^l + b_j^l) \qquad (1)$$

Here: l represents the number of layers; k represents the convolutional kernel; x_jrepresents receptive field of input layer, and b represents bias.

In sub-sampling layer, the number of input characteristic pictures is unchanged after the pooling operation, and the size of output characteristic pictures is half of the original pictures. The expression form of sub-sampling layer is:

$$x_j^l = f(\beta_j^l p(x_j^{l-1}) + b_j^l) \qquad (2)$$

Here：f(.) is the pooling function; β is the weight factor.

3. EXPERIMENTAL DESIGN

In human vision, the receptive field refers to the area in which the visual neurons react to specific stimuli. In the BP neural network, neurons among layers are fully connected. The convolutional neural network uses the method of receptive field, so that each neuron need not to do convolutional algorithm on the whole picture and each neuron has one specific receptive field. On the top of the network structure neurons of different receptive fields will be integrated to get the whole information. The introduction of receptive fields greatly reduces the weight parameter of the convolutional neural network and improves the efficiency of the network.

In convolutional neural network, each convolutional filter in the convolutional layer shares the same parameters and function, convolves the input picture, transforms the convolutional result into the characteristic picture of the input picture, and extracts part characteristics of the picture. This thesis obtains the convolutional characteristics of the picture, which needed to be dimensionally reduced by the maximum pool sampling method and divided by several non-intersecting regions of n*n, and these characteristics after dimension reduction are represented by the largest or average characteristic in these regions. Convolutional characteristics after dimension reduction are more convenient for classification. After the convolution operation, the data is classified based on these characteristics. Theoretically, all the characteristics can be used for training and classification, but this will greatly increase the computation amount.

This thesis selects pictures with the size of 28 * 28, size of the convolutional kernel in the first convolutional layer is (5, 5). Each convolutional kernel convolving one picture will yield convolutional characteristic as (28-5+1)*(28-5+1) = 576.

If the size of pictures is not large, it is very difficult to learn a classifier input and prone to over fit. Therefore, after convolutional operation, pooling method is introduced to the dimension reduction of data. It divides the input picture into non-overlapping rectangles, and uses averages (or the maximum) of each rectangle to describe convolutional characteristics after pooling. The pooling operation reduces the computational complexity of the upper layer and has a certain translation invariance, so the maximum pooling method is adopted in this experiment.

This thesis collect SSH traffic data from our self-built tunnel environment and extract the payload data from SSH traffic. The data contains about 20000 valid data streams, 5 application types, as shown in Table 1:

Table 1. Datasets Information

Application	Number
Nmap	6236
Baidu Network	5392
Netease Cloud Music	3327
NetEase’s Word Dictionary	4159
QQ	4325

Decimal integer of 0 to 255 is exactly the length of a byte. So data in each byte of payload can be normalized to [0, 1], and the traffic data can be transformed into picture data and be processed. By analyzing the traffic data, the payload data carried in each traffic packet is extracted and transformed into a gray value matrix of picture data.

4. EXPERIMENTAL METHODS AND RESULTS

This thesis mainly studys the classification of the specific applications included in the SSH tunnel. In this section, this thesis filters out the non-SSH traffic, reduce the impact of other traffic on the experimental results, and improve the accuracy of the classification.

Due to the encryption characteristics of the encrypted traffic and the variable nature of the port, there is no guarantee that the traffic category will be correctly marked under unknown conditions, so the data used in this study is generated by itself. At the same time, because of the privacy protection of some available public datasets, the load portion of the data is deleted, so there is no guarantee that the categorical results are accurate. This thesis designs the experimental data collection module, and collect data by adding a data acquisition point to the client of Figure 1.

This paper chooses five applications for model building and testing, including Nmap, Baidu Network, Netease Cloud Music, NetEase's Word Dictionary and QQ, assumes that each SSH connection carries only one application at a time.

The convolutional neural network uses the maximum pooling model to learn, and training more than 20,000 data group training samples collected in a closed network environment.

The recognition rate of the convolutional neural network to the data set is 95.8%. With the increase of training steps, the accuracy is also improving. The change curve of the accuracy is shown in Figure 3.

The accuracy of each period was obtained after training, including the accuracy of the training data set and the test data set, which were named acc and val_acc respectively. As shown in Figure 3, the accuracy of recognition of the training model increases and tends to be stable.

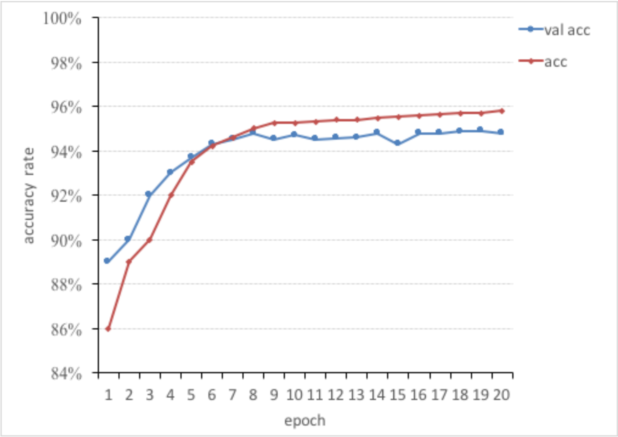

Figure 3. The accuracy of traning dataset and testing dataset

The loss rate for each period is shown in Figure 4. The loss rate of the traning data set is named loss and the loss rate of the test data set is named val_loss in Figure 4. As shown in Figure 4, the loss rate of the training model is gradually reduced and eventually stabilizes.

After the experiment, the total accuracy rate of the identification model for the five applications identification is 95.8%. Therefore, the methods proposed by the thesis have high accuracy in identifying SSH applications.

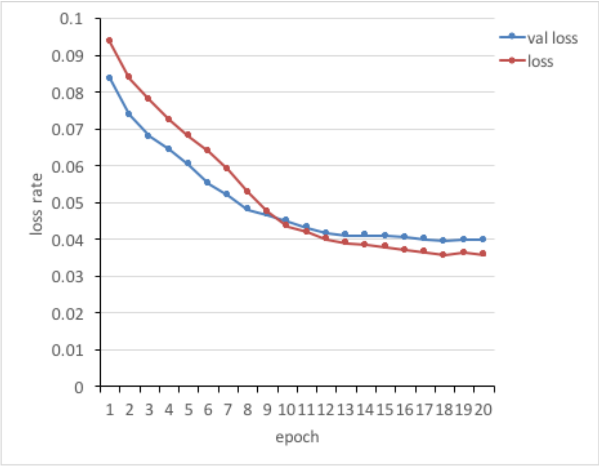

Figure 4. The loss rate of traning dataset and testing dataset

5. CONCLUSION AND FUTURE WORK

The deep learning theory breaks through the extraction and selection of the artificial design features in the traditional machine learning, which constructs a learning model, and realizes the feature selection and the classifier training under a unified theoretical framework. In this paper, the convolutional neural network algorithm is used to identify the applications hidden in SSH traffic. And experimental result indicates that the convolutional neural network can extract features better. The author plan to improve the current model for identifying the file types that are hidden in the SSH tunnel.

6. REFERENCES

[1] Alshammari, R. and Zincir-Heywood, A. N. 2009. Machine learning based encrypted traffic classification: Identifying SSH and Skype. *IEEE Symposium on Computational Intelligence for Security & Defense Applications. IEEE*. 1-8.

[2] Niemczyk, B. and Rao, P. 2014. Identification over encrypted channels. BlackHat USA.

[3] Bujlow, Tomasz, Carela-Español, V., and Barlet-Ros, P. 2015. Independent comparison of popular DPI tools for traffic classification. *Computer Networks*. 76, 75-89.

[4] Dutt, Shantanu, 2002. New faster Kernighan-Lin-type graph-partitioning algorithms. Ieee/acm International Conference on Computer-Aided Design, 1993. Iccad-93. Digest of Technical Papers IEEE, 370-377.

[5] Dusi, Maurizio, 2014. Identifying the traffic of SSH-encrypted applications.

[6] Gil, G. D., Lashkari, A. H., Mamun, M. 2016. Characterization of Encrypted and VPN Traffic Using Time-Related Features. *Proceedings of the 2nd International Conference on Information Systems Security and Privacy (ICISSP 2016)*. 407-414.

[7] Meng, J., Wang, L., and Xiong, G. et al, 2012. Study on SSH application classification based on machine learning. *Journal of Computer Research and Development*. 153-159.

[8] Maiolini, Gianluca, 2009. Real time identification of SSH encrypted application flows by using cluster analysis techniques. *International Ifip-Tc 6 Networking Conference*. Springer-Verlag, pp. 182-194.

[9] Wu, T. 2014. Early-stage Internet traffic identification based on packet payload size. *Journal of Southeast University (English Edition)*.

[10] Velan Petr, 2015. A survey of methods for encrypted traffic classification and analysis. *International Journal of Network Management*. 25, 5, 355-374.

[11] Pan, W., Cheng, G., and Guo, X. 2016. Review and prospect of network encryption flow recognition. *Journal of Communications*. 37, 9, 154-167.

[12] Wright, Charles V., Coull, S. E. and Monrose, F. 2009. An efficient defense against statistical traffic analysis. N*etwork and Distributed System Security Symposium*, NDSS 2009, San Diego, California, Usa, February-, February. DBLP, pp. 237-250.

[13] Wang, Z. 2015. *The Applications of Deep Learning on Traffic Identification*. BlackHat USA.