2020 International Conference on Information Science and Education (ICISE-IE) | 978-1-6654-2261-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICISE51755.2020.00109

SHADOWSOCKS TRAFFIC IDENTIFICATION BASED ON CONVOLUTIONAL NEURAL NETWORK

Nan Zhang[1,a]
[1]College of artificial intelligence
Beijing Normal University
Bei Jing, China
[a]e-mail: 201921210048@mail.bun.edu.cn

Tiantian Wu[1,b]
[1]College of artificial intelligence
Beijing Normal University
Bei Jing, China
[b]e-mail: 201611150931@mail.bnu.edu.cn

Yuening Zhang[1,c]
[1]College of artificial intelligence
Beijing Normal University
Bei Jing, China
[c]e-mail: zyn@mail.bnu.edu.cn

Mingzhong Xiao[1,d]*
[1]College of artificial intelligence
Beijing Normal University
Bei Jing, China
* Corresponding author: [d] xmz@bnu.edu.cn

***Abstract*—Network traffic has been massively produced since the development of the Internet. While the encryption of traffic has ensured the security and reliability of information, it also brings great challenge to the traffic identification and monitoring. The present study proposes a method of shadowsocks traffic identification based on the one-dimensional Convolutional Neural Network. This method simplifies the feature extraction of traffic identification and the recognition accuracy is over 98%. Because we can not find the published shadowsocks traffic dataset, we gathered four encryption kinds of shadowsocks traffic to study on the influence of different encryption on shadowsocks traffic. Moreover, we include VPN traffic and do contrast experiment based on four deep-learning models to verify the efficiency of one-dimensional convolutional neural network.**

Keywords-component; traffic recognition; shadowsocks; convolutional neural network;

I. INTRODUCTION

The development of anonymous communication systems brings new challenges to the identification and monitoring of network traffic. Criminals use the concealment of anonymous communication software to commit cyber crimes and some software is going to become a hotbed for disseminating illegal information. Therefore, the importance of identifying network traffic through technology has become increasingly significant.

The current network traffic identification methods mainly include the identification method based on the network port number[1].It finds the corresponding application by identifying the port number carried by the TCP/UDP packet. For traffic using standardized ports, this type of identification method is simple and effective, but many applications now do not use this port standard, thus this method has certain limitations. Based on the DPI identification method, classification is performed by matching the signature in the data packet[2]. Based on DPI, this paper[3] proposes an automatic traffic classification method based on weighted itemset as protocol signature. But for encrypted traffic, this method is difficult to find an effective signature. DFI-based traffic identification technology mostly uses statistical features to identify traffic. The traffic data of Cambridge University collected and published by Moore[4] et al, 248 traffic characteristics can be used in the field of network traffic identification and classification, which has certain reference value. According to the collected traffic characteristics, machine learning methods are used to classify. The survey paper[5] focuses on the emerging research of machine learning technology applied to IP traffic classification the cross disciplinary integration of IP network and data mining technology. The paper[6] proposes a way to improve accuracy, regardless of whether the traffic is encrypted or not. The paper[7] studied the application of two machine learning algorithms, supervised support vector machine (SVM) and unsupervised K-means clustering in traffic classification. Xishou Du[8] and others proposes a unidirectional P2P traffic identification method based on C4.5 decision tree. The selection of DFI flow characteristics mostly depends on the re- searcher's experience, and the generalization ability is limited. With the development of deep learning, deep learning is also applied to network traffic identification. Compared with traditional machine learning methods, deep learning does not need time-consuming feature selection, but has self-learning ability and can learn abstract features. The results of traffic identification methods are often better than traditional machine learning methods. Many studies have shown that the flow characteristics can be extracted from the first 20 packets of network traffic[9]. The paper[10] proposes a BSNN network model to identify five protocols: QQ, PPLive, DNS, 360 and BitTorrent. It[11] uses one-dimensional CNN to identify Google's QUIC protocol. Many studies directly use the payload of traffic as input.

At present, shadowsocks, as the mainstream anonymous communication protocol, is used by many people due to its stability and simplicity of operation. Then there is very little research on shadowsocks. The paper[12] uses random forest to classify shadowsocks traffic with an accuracy rate of about 85%. This study collects shadowsocks traffic of 4 encryption methods

978-1-6654-2261-1/20/$31.00 ©2020 IEEE
DOI 10.1109/ICISE51755.2020.00109

and studies the effects of different encryption methods on shadowsocks traffic. We include VPN traffic and do contrast experiment based on four deep-learning models to verify the efficiency of one-dimensional convolutional neural network. This study helps to maintain network security, strengthen network control, and curb cyber crime.

II. Methods

A. Dataset

Since no published shadowsocks traffic dataset was found, our study collects shadowsocks traffic of four encryption methods, namely aes-256-gcm, aes-192-gcm, aes-128-gcm and rc4-md5. Researchers have studied VPN more than shadowsocks, so we chooses VPN traffic in published ISCX dataset[13] to increase the diversity of data. Figure 1shows the principle of traffic collection. We build a shadowsocks server, using shadowsocks client. We log on the server account on the client, turn on the global mode, play the video in the browser, and use Wireshark software to collect traffic on the client. At this time, the collected video traffic belongs to shadowsocks traffic. By uniformly changing the encryption methods of the shadowsocks server and client, we can collect shadowsocks traffic of different encryption methods. Quit the shadowsocks client, and we get normal traffic.

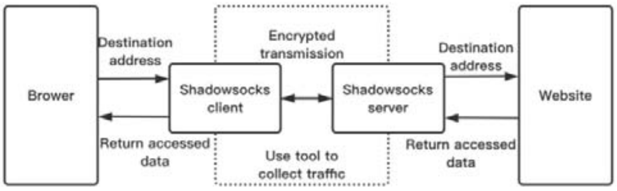


Figure 1. Principle of shadowsocks traffic collection

B. Data Processing

Data processing refers to the method of [14][15], and uses the data-preprocessing toolkit USTC-TK2016 developed by their project team. The steps include the following four steps:

Step 1 (traffic split). This step divides the continuous raw traffic, and divides a PCAP file into multiple pcap files. Each pcap file is a five-tuple identical bidirectional flow and contains all protocol layers. The original toolkit USTC-TK2016 can be divided into bidirectional flow or unidirectional flow, and only application layer protocol or all protocols can be selected. In the papers [14, 15], a large number of experiments have proved that the data processing method of bidirectional flow and including all protocol layers is the best. This article will not study other methods.

Step 2 (traffic cleaning). In this step, traffic is anonymized, and the MAC address and IP address are randomized at the data link layer and IP layer, respectively. Eliminate the influence of MAC address and IP address on the experiment, the output is still pcap file.

Step 3 (image generation). This step organizes all files into 784 bytes. If the file size is greater than 784 bytes, it will be cropped to 784 bytes. If the file size is less than 784 bytes, add 0 at the end to make up to 784 bytes. Convert 784 bytes into a 28*28 size picture for visualization to facilitate observation of traffic characteristics.

Step 4 (IDX conversion). Convert the image to IDX format file to facilitate subsequent model training.

Figure 2shows the data processing process. We collected shadowsocks traffic of four encryption methods, normal traffic and VPN traffic, and generated four kinds of datasets, ss_4, ss_no, ss_no_vpn and vpn_nonvpn, where vpn_nonvpn source published ISCX dataset[13]: youtube1.pcap, youtube2.pcap, youtube3.pcap, youtube4.pcap, youtube5.pcap, youtube6.pcap and vpn_youtube_A.pcap. These datasets have been placed on github for people's reference and learning. https://github.com/BlueBlueGrey/Traffic-identification-of-shadowsocks-based-on-convolutional-neural-network

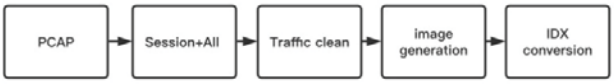


Figure 2. Data process

TABLE I. SS_4 DATASET SIZE

Type	Count
ss_aes-256-gcm	2454
ss_aes-192-gcm	1778
ss_aes-128-gcm	1815
ss_rc4-md5	644

TABLE II. SS_NO DATASET SIZE

Type	Count
normal	1522
ss_aes-256-gcm	2357
ss_aes-192-gcm	1623
ss_aes-128-gcm	1815
ss_rc4-md5	644

TABLE III. SS_NO DATASET SIZE

Type	Count
normal	1522
ss_aes-256-gcm	2357
ss_aes-192-gcm	1623
ss_aes-128-gcm	1815
ss_rc4-md5	644
vpn	213

TABLE IV. VPN_NONVPN DATASET SIZE

Type	Count
vpn	213
normal	697

C. One Dimensional CNN Model

Convolutional neural network is inspired by the brain [16], the earliest convolutional neural network LeNet5[17], can classify the numbers in handwritten numbers. The hierarchical structure of the model is shown in Figure 3Convolution operation can learn local features. Formula(1) describes convolution operation, and * represents convolution operation. l is the size of convolution kernel. In our study l is equal to 3.

$$y(n) = x(n) * h(n) = \sum_{m=-\infty}^{\infty} x(m)h(l-m) \qquad (1)$$

Let X=[$x_1, x_2, x_3, \ldots, x_n$]be the input vector. Each vector[$x_i, x_{i+1}, \ldots, x_{i+l-1}$]will get a new vector c_i .The model in this article has 2 convolutional layers. Unifiedly use C to represent the vector obtained after 2 convolutional layers. $f(x)$ is a ReLU function, as a commonly used activation function in artificial neural networks. It usually refers to the nonlinear functions represented by slope function and its variants.

$$c_i = f([x_i, x_{i+1}, \ldots, x_{i-l+1}] \bullet W + b) \qquad (2)$$

$$f(x) = max(0,x) \qquad (3)$$

$$C = [c_1, c_2, c_3, \ldots, c_{i-l+1}] \qquad (4)$$

Then through the pooling layer to reduce the feature dimension, compress the number of data and parameters, reduce overfitting, and improve the fault tolerance of the model. j represents the size of the pooling layer. In our study j is equal to 2.

$$m_i = max(c_i, c_{i+1}, \ldots, c_{i+j-1}) \qquad (5)$$

$$M = [m_1, m_2, \ldots, m_{i-j+1}] \qquad (6)$$

The Flat layer expands the vector into one dimension, and the Dropout layer prevents overfitting and increases robustness, finally add a fully connected layer and softmax function to get the final output, the entire model is trained. cn represents the number of labels.

$$A = X \bullet W + B \qquad (7)$$

$$\hat{a}_j = \frac{exp(x_j)}{\sum_{i=1}^{cn} exp(x_i)} \qquad (8)$$

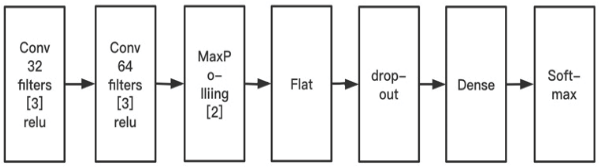


Figure 3. CNN1D hierarchy

D. Experimental Design

In our study, we use Keras to encode and select four common models for experiment, which are CNN1D, CNN2D, DNN and RNN. The hierarchy of CNN2D model is shown in Figure 4. Compared with CNN1D model, the input is a two-dimensional matrix of 32 * 32, and the filter sizes of convolution layer and pooling layer are two-dimensional. The hierarchical structure of RNN model is shown in Figure 5, which contains 100 hidden units of SimpleRNN layer. Finally, the prediction results are output through full connection layer and softmax function. The hierarchical structure of DNN model is shown in Figure 6, which contains two layers with 100 neurons in each layer, and finally outputs the prediction results through the full connection layer and softmax function.

We firstly use these models and two data sets, ss_no and ss_no_vpn, to train. Then experiment with the one-dimensional CNN model and ss_4 and vpn_nonvpn to see how the model identifies other encrypted traffic.

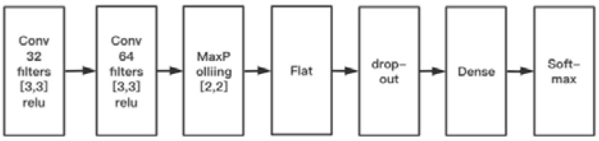


Figure 4. CNN2D hierarchy

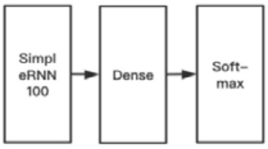


Figure 5. RNN hierarchy

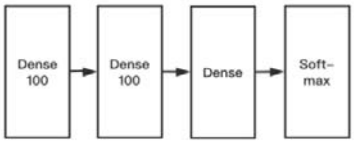
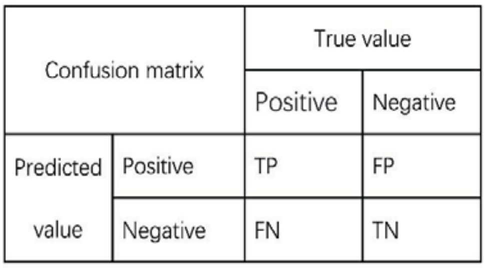


Figure 6. DNN hierarchy

III. EXPERIMENTAL

A. Confusion matrix

The confusion matrix is a two-dimensional square matrix, as shown in Figure 7. It is mainly used to evaluate binary classification problems. In our study, A, P, R and $F1$ these indexes are selected to measure the experimental results. The calculation formula is as follows. For the multi classification problem in our study, we can take one class as positive samples and the others as negative samples to calculate the evaluation index.



Confusion matrix		True value	
		Positive	Negative
Predicted value	Positive	TP	FP
	Negative	FN	TN

Figure 7. Confusion matrix

A represents the proportion of correct classification samples.

$$A = \frac{TP + TN}{TP + TN + FP + FN}$$

P represents the proportion of the correct model prediction among all samples where the model prediction is positive.

$$P = \frac{TP}{TP + FP}$$

R represents the proportion of the correct model prediction in the positive sample.

$$R = \frac{TP}{TP + FN}$$

The $F1$ indicator combines the results of the output of P and R , the bigger the better.

$$F1 = \frac{2 \bullet P \bullet R}{P + R}$$

B. Visual Difference between different traffic

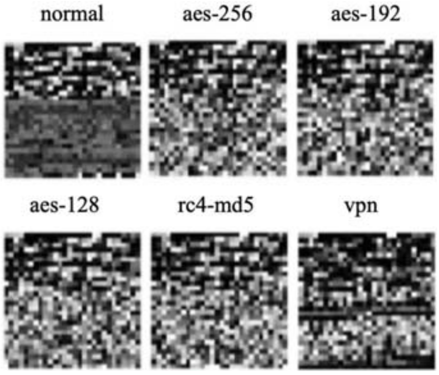


Figure 8. Visualization of each type of traffic

aes-256, aes-192, aes-128 and rc4-md5 belong to shadowsocks traffic with different encryption methods. By visually comparing the normal traffic, shadowsocks traffic and vpn traffic of the video type, we can find that there are obvious differences between them. In the picture composed of normal traffic, the upper half of the picture has relatively high brightness, while the brightness of the lower half is not very obvious; in the picture composed of shadowsocks traffic, most have higher brightness, while the upper left part it has obvious characteristic information; in the picture composed of VPN traffic, the high-brightness area of the picture is in the form of strips. However, using the naked eye to distinguish the shadowsocks traffic of the four encryption methods may not be able to identify accurately.

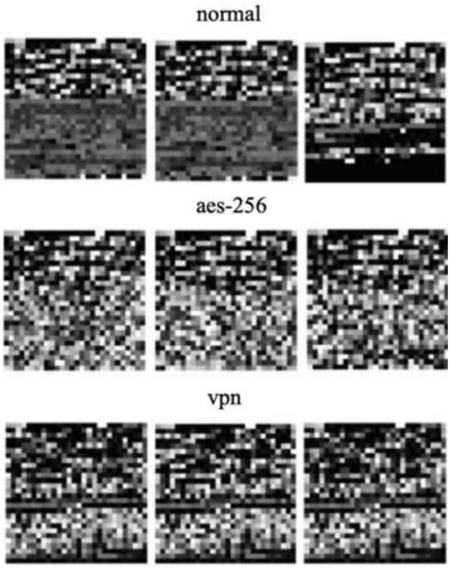


Figure 9. Consistency at the same traffic

C. Performance of Neural Networks identification in shadowsocks / normal / vpn dataset

In the process of using neural networks to distinguish multiple types of traffic, we first consider the shadowsocks traffic as a whole, and use CNN1D, CNN2D, DNN and RNN to classify and identify shadowsocks traffic and normal traffic. The result of this are shown in Figure 10and TABLE V.In addition, we have added a set of experiments (shadowsocks traffic, normal traffic, vpn traffic) to increase data diversity and its result is shown in Figure 11and TABLE VI.

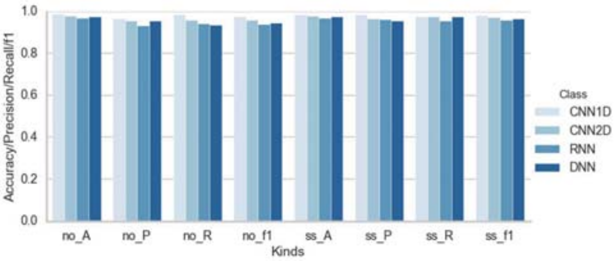


Figure 10. Recognition of normal streaming and SS streaming by four kinds of neural networks

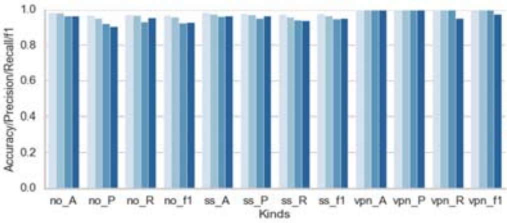


Figure 11. Recognition of normal streaming, shadowsocks streaming and VPN streaming by four neural networks

TABLE V. EXPERIMENTAL RESULTS OF CNN1D CNN2D RNN AND DNN WITH SS_NO DATASET

Normal				
Index% Modle	A	P	R	F1
CNN1D	98.76	96.48	98.41	97.44
CNN2D	98.00	95.63	96.02	95.83
RNN	97.04	93.31	94.42	93.86
DNN	97.43	95.53	93.63	94.57
Shadowsocks				
Index% Modle	A	P	R	F1
CNN1D	98.57	98.45	97.69	98.07
CNN2D	97.90	96.70	97.69	97.19
RNN	97.04	96.38	95.64	96.01
DNN	97.43	95.72	97.44	96.57

TABLE VI. EXPERIMENTAL RESULTS OF CNN1D CNN2D RNN AND DNN WITH SS_NO_VPN DATASET

Normal				
Index% Model	A	P	R	F1
CNN1D	98.60	96.83	97.21	97.02
CNN2D	98.13	95.29	96.81	96.05
RNN	96.54	92.13	93.23	92.67
DNN	96.64	90.57	95.62	93.02
Shadowsocks				
Index% Modle	A	P	R	F1
CNN1D	98.41	97.94	97.69	97.82
CNN2D	97.48	97.14	95.90	96.52
RNN	96.26	95.34	94.36	94.85
DNN	96.54	96.57	93.85	95.19
Vpn				
Index% Modle	A	P	R	F1
CNN1D	100.00	100.00	100.00	100.00
CNN2D	100.00	100.00	100.00	100.00
RNN	100.00	100.00	100.00	100.00
DNN	99.91	100.00	95.24	97.56

It can be seen from the results obtained that the effect of using neural networks to identify shadowsocks traffic in a variety of network traffic is very good, and the recognition rate is above 90%. Specifically, among a variety of neural networks, one-dimensional neural networks perform better in multiple indicators (accuracy, precision, recall, f1) than the other three neural networks, and very few of them show the same level.

D. Performance of CNN1D recognition in shadowsocks traffic of four encryption methods

Due to the excellent performance of one-dimensional CNN in recognition, next, we use one-dimensional CNN to perform detailed identification and analysis of each traffic. Among the four encryption methods for shadowsocks traffic identification, the accuracy, precision, recall rate and f1 rate of the traffic identification of the encryption method aes-256-gcm have reached the level of close to 100%, and the indicators of the traffic of the encryption method rc4-md5 are also all above 90%. The recognition indexes of the remaining aes-192-gcm and aes-128-gcm are not as high as the former two, but they still reach the level of more than 80%. Visual results can be found in Figure 12. In the recognition of shadowsocks traffic, vpn traffic and normal traffic, the recognition effect of one-dimensional CNN is more obvious. For identification of VPN, accuracy, precision, recall rate and f1 rate are all 100%, which could be shown in Figure 13.

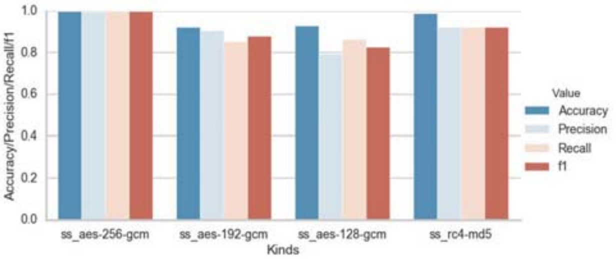


Figure 12. The accuracy, precision, recall and F1 of CNN1D for identifying shadowsocks traffic of different encryption methods

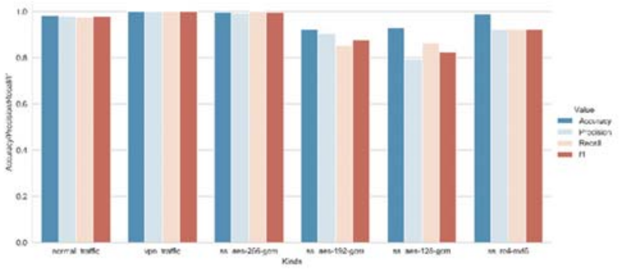


Figure 13. The accuracy, accuracy, recall and F1 of CNN1D for identifying shadowsocks traffic, normal traffic and VPN traffic

E. One-dimensional CNN recognition between vpn and non_vpn traffic

We also used one-dimensional neural network to detect and identify vpn and normal traffic in vpn_nonvpn dataset. The results obtained are shown in the table below. Because the dataset is relatively small，these indicators are not very high.

However, accuracy rate is about 77%, indicating that the model has certain recognition effect.

TABLE VII. EXPERIMENTAL RESULTS OF CNN1D WITH VPN_NONVPN DATASET

Index% Modle	A	P	R	F1
vpn	77.89	50.00	66.67	57.14
nonvpn	77.89	89.56	81.08	85.11

IV. CONCLUSION

This study collects shadowsocks traffic of four encryption methods, and studies the influence of different encryption methods on shadowsocks traffic. According to the experimental results, the traffic characteristics of different encryption methods will also change. The VPN traffic is increased, which further illustrates the effectiveness of one-dimensional CNN in classifying encrypted traffic. The comparative experiments on different deep learning models show that the effect of CNN1D model is slightly better than that of DNN, RNN and CNN2D. In the future research, we will continue to improve the neural network model, improve the accuracy and efficiency of the model, and collect more shadowsocks traffic data for people's reference and learning.

REFERENCES

[1] Cotton, M., Eggert, L., Touch, J., Westerlund, M., & Cheshire, S. (2011). Internet Assigned Numbers Authority (IANA) Procedures for the Management of the Service Name and Transport Protocol Port Number Registry. RFC, 6335, 1-33.

[2] Finsterbusch, M., Richter, C., Rocha, E., Muller, J. A., & Hanssgen, K. (2013). A survey of payload-based traffic classification approaches. IEEE Communications Surveys & Tutorials, 16(2), 1135-1156.

[3] Yeganeh, S. H., Eftekhar, M., Ganjali, Y., Keralapura, R., & Nucci, A. (2012, July). Cute: Traffic classification using terms. In 2012 21st International Conference on Computer Communications and Networks (ICCCN) (pp. 1-9). IEEE.

[4] Moore, A. W., & Zuev, D. (2005, June). Internet traffic classification using bayesian analysis techniques. In Proceedings of the 2005 ACM SIGMETRICS international conference on Measurement and modeling of computer systems (pp. 50-60).

[5] Nguyen, T. T., & Armitage, G. (2008). A survey of techniques for internet traffic classification using machine learning. IEEE communications surveys & tutorials, 10(4), 56-76.

[6] Okada, Y., Ata, S., Nakamura, N., Nakahira, Y., & Oka, I. (2011, December). Comparisons of machine learning algorithms for application identification of encrypted traffic. In 2011 10th International Conference on Machine Learning and Applications and Workshops (Vol. 2, pp. 358-361). IEEE.

[7] Fan, Z., & Liu, R. (2017, August). Investigation of machine learning based network traffic classification. In 2017 International Symposium on Wireless Communication Systems (ISWCS) (pp. 1-6). IEEE.

[8] Du Xishou, Chen Shuqiao, & Zhang Jianhui, (2013). Research on unidirectional P2P traffic identification method based on C4. 5 decision tree. Journal of Chinese Mini-Micro Computer Systems , 34(2), 247-252.

[9] Rezaei, S., & Liu, X. (2019). Deep learning for encrypted traffic classification: An overview. IEEE communications magazine, 57(5), 76-81.

[10] Li, R., Xiao, X., Ni, S., Zheng, H., & Xia, S. (2018, June). Byte segment neural network for network traffic classification. In 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS) (pp. 1-10). IEEE.

[11] Tong, V., Tran, H. A., Souihi, S., & Mellouk, A. (2018, December). A novel QUIC traffic classifier based on convolutional neural networks. In 2018 IEEE Global Communications Conference (GLOBECOM) (pp. 1-6). IEEE.

[12] Deng, Z., Liu, Z., Chen, Z., & Guo, Y. (2017, August). The random forest based detection of shadowsock's traffic. In 2017 9th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC) (Vol. 2, pp. 75-78). IEEE.

[13] Draper-Gil, G., Lashkari, A. H., Mamun, M. S. I., & Ghorbani, A. A. (2016, February). Characterization of encrypted and vpn traffic using time-related. In Proceedings of the 2nd international conference on information systems security and privacy (ICISSP) (pp. 407-414).

[14] Wang, W., Zhu, M., Wang, J., Zeng, X., & Yang, Z. (2017, July). End-to-end encrypted traffic classification with one-dimensional convolution neural networks. In 2017 IEEE International Conference on Intelligence and Security Informatics (ISI) (pp. 43-48). IEEE.

[15] Wang, W., Zhu, M., Zeng, X., Ye, X., & Sheng, Y. (2017, January). Malware traffic classification using convolutional neural network for representation learning. In 2017 International Conference on Information Networking (ICOIN) (pp. 712-717). IEEE.

[16] Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of physiology, 160(1), 106.

[17] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.