ISSN: 2414-1895 DOI: 10.6919/ICJE.202006_6(6).0044

Research on Encrypted Traffic Classification Method Based on Improved Convolutional Neural Network

Zhibo Zhai, Hui Wang, Luming Bai

Henan University of Science and Technology, Luoyang Henan 471023, China.

Abstract

Network traffic classification is of great significance to the daily management and maintenance of the network. With the rapid development of encryption techniques for network traffic, traditional traffic classification methods have reduced the accuracy due to the low identification rate of encrypted traffic. Aiming at this problem, a deep learning traffic classification method based on improved convolutional neural network is proposed. By optimizing and adjusting the network structure, network parameters, and cost function of the convolutional neural network, the classification accuracy of encrypted traffic is improved. By comparing and analyzing the existing traffic classification method on public data sets, the experimental results show that the proposed method effectively improves the classification accuracy of encrypted traffic by 10%.

Keywords

Traffic classificationd, Encrypted traffic, Convolutional neural network, Deep learning, Classification accuracy.

1. Introduction

Network traffic classification refers to the process of identifying and classifying traffic data according to certain characteristics according to different protocols or applications [1]. The rapid and accurate classification of network traffic is of great significance for analyzing network operating status, maintaining network environment, solving network faults, and improving service quality. With the continuous development of encryption technology, network traffic encryption plays an important role in maintaining network security and user privacy. The data packet load is encrypted by different algorithms in the encryption process. The features used for classification in the original traffic data are difficult to continue to use, and decrypting the encrypted traffic requires a large amount of computing resources, so the traditional traffic classification method reduces the encrypted traffic Classification accuracy.

In view of the above problems, this paper designs an encrypted traffic classification method based on improved convolutional neural network, and compares with the existing encrypted traffic classification methods on public data sets to verify the accuracy of the proposed method to classify encrypted traffic rate. The rest of this article is organized as follows: Section 1 introduces the main methods of current traffic classification; Section 2 elaborate on the traffic classification method designed by this article and the improvements it has made; Section 3 describes the traffic classification method designed by this article through experiments The effectiveness of the test has been verified; Section 4 summarizes the work done in this article.

2. Research on Traffic Classification

There are three mainstream methods for solving the traffic classification problem: port number-based methods, deep packet inspection(DPI) methods, and machine learning methods [2].

The earliest traffic classification method is based on the port number. This method is classified by mapping the port information of the packet header to a given application type. For example: the TCP / IP protocol usually stipulates that the Web uses port 80, SMTP / POP3 Mail service uses ports 25 and 110, FTP uses ports 20 and 21, etc. The main advantage of this method is that it is simple and easy to implement, and the classification speed is fast. However, as the number of network applications continues to increase, a large number of applications use random port numbers or disguised port numbers, which greatly reduces the accuracy of port-based classification methods. The deep packet inspection method is a process of classifying traffic by analyzing the entire data packet and using the load information in the data packet [3]. Literature [4] proposes a high-throughput traffic classification system architecture based on deep packet inspection technology, which can handle the problem of large number of concurrent streams. However, this classification method has a low classification accuracy rate for encrypted traffic and consumes a lot of computing resources. At the same time, an in-depth analysis of the payload of the data packet will affect the privacy protection of users.

In order to solve the problems encountered by traditional classification methods, relevant researchers proposed to use machine learning methods to classify the traffic. In [5] Liu et al. Proposed to use the clustering algorithm commonly used in machine learning to classify traffic. The flow feature set is designed by feature selection, and then experiments are performed on different data sets and good classification accuracy is obtained. Based on this, the researchers in [6, 7] perfect and improve the classification algorithm from many different angles, and further improve the classification accuracy. In [8], Alhamza et al. Used statistical features such as packet size and average packet interval time to compare and analyze algorithms such as K-means and maximum expectation for traffic classification. In [9], the machine learning algorithm was applied to the classification of online video traffic and download traffic in the software-defined network(SDN), which achieved a good classification effect for real-time traffic. Machine learning methods can classify encrypted traffic and non-encrypted traffic, and do not need to parse the content of the data packet. However, the classification accuracy of machine learning methods will be affected by the choice of traffic characteristics, and it takes a lot of time and resources to manually design the traffic characteristics.

With the development of artificial intelligence, deep learning algorithms have achieved good applications and development in the fields of image classification and natural language processing. The relationship between bytes, packets and traffic is similar to the relationship between pixels, color blocks and images, and words, words and sentences, so you can use deep learning methods to classify traffic. At the same time, deep learning algorithms can automatically generate feature sets through multiple training studies, thus solving the problem of machine learning in the classification process that requires manual design of features [10]. Literature [11] proposed a classification method based on Deep Belief Network(DBN). By designing an appropriate feature set, the classification accuracy of P2P traffic by the DBN model is improved. Although this method has a high classification accuracy, it needs to design and extract the flow characteristics when establishing a data set, so the calculation process is complicated and it is difficult to meet the actual classification needs. Literature [12] proposed an end-to-end traffic classification method based on deep learning. This method uses a one-dimensional convolutional neural network model to integrate feature extraction, feature selection, and classifier into a unified end-to-end framework, and automatically learns the nonlinear relationship between the original input and the expected output through multiple trainings to achieve The purpose of traffic classification.

This paper designs a classification model of encrypted traffic based on convolutional neural network. The network parameters, data preprocessing and cost function of convolutional neural network are optimized and adjusted to improve the classification accuracy of encrypted traffic. The comparative analysis of existing methods verifies the effectiveness of the proposed method.

3. Improved Convolutional Neural Network Traffic Classification Method

The improved traffic classification method based on the convolutional neural network model in this paper mainly includes data preprocessing and traffic classification. The overall process is shown in Figure 1.

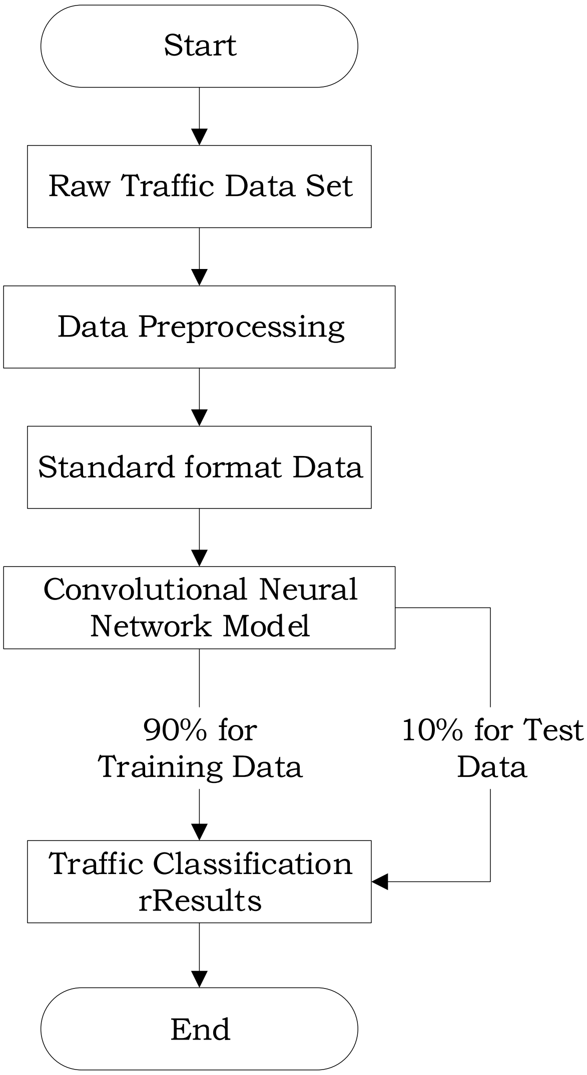

Figure 1. Convolutional neural network traffic classification process

3.1 Improved data preprocessing method

This article uses the ISCX VPN-nonVPN public data set, which contains multiple types of encrypted traffic. Figure 2 shows the specific process of raw flow data preprocessing.

(1) Delete and zero padding of extraneous characters

In order to accelerate the speed of model training and classification, this paper deletes the address information and mark segment information of MAC layer that has little effect on traffic classification in the original traffic data packet, and then fills the 12-bit zero byte into the 8-bit user datagram At the end of the protocol, make it equal to the 20-byte length of the TCP segment header.

In order to accelerate the speed of model training and classification, this paper deletes the MAC information and mark segment information in the original traffic data packet that have little impact on traffic classification, and then fills in 12-bit zero bytes at the end of the 8-bit UDP packet, make it equal to the 20-byte length of the TCP packet.

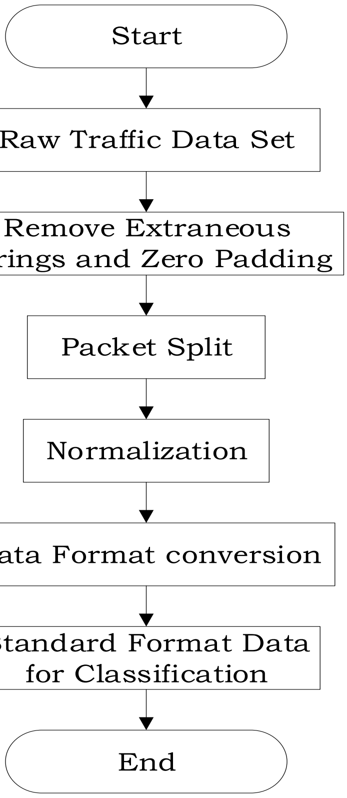
Start
Raw Traffic Data Set
Remove Extraneous Strings and Zero Padding
Packet Split
Normalization
Data Format conversion
Standard Format Data for Classification
End

Figure 2. Flow data preprocessing

(2) Packet split

In order to meet the input requirements of the convolutional neural network model, the length of the data packet must be consistent. This paper uses the same 784 bytes as the unit length of the traffic classification in [10, 12]. First, divide all traffic packets according to whether the source address and destination address are the same. Then, for the length of the divided data packet that exceeds 784 bytes, the first 784 bytes are selected as the effective length; if the length of the data packet is less than 784 bytes, the end is supplemented with "0" to 784 bytes.

(3) Data normalization

In order to improve the performance of the model and speed up the training, this paper normalizes each divided data packet so that all data values are between (0, 1). The normalized processing method is shown in formula (1).

$$x_{ij}^{*}=\frac{x_{ij}-\min}{\max-\min} \tag{1}$$

In the formula: x_{ij} and x_{ij}^{*} represent the value of the j-th data in the i-th data packet before normalization and after normalization, respectively. max is the maximum data value and min is the minimum data value. In this paper, the value of max is 255 and the value of min is 0.

(4) Data format conversion

The normalized data is converted into labeled idx format required for training of convolutional neural networks.

3.2 Improved convolutional neural network model

The structure of the convolutional neural network model is shown in Figure 3, which consists of the input layer, convolutional layer, pooling layer, fully connected layer, and output layer [13, 14]. The model used in this article consists of three convolutional layers, one pooling layer and two fully connected layers.

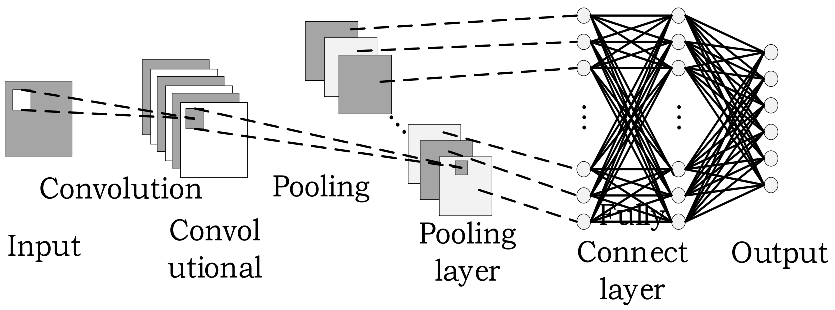

Figure 3. General model of convolutional neural network

In this paper, there are three convolutional layers, which are expressed as C_1, C_2, and C_3, respectively, and the expression is shown in formula (2).

$$C_{\mathrm{i}} = f(w_i x + b_i) \tag{2}$$

In the formula: w_{i} represents the convolution kernel, b_{i} represents the bias, and f represents the activation function.

The activation function is used to increase the sparsity of the model and avoid overfitting. Common activation functions include Sigmoid function, tanh function and ReLU function [15]. Compared with the other two functions, the advantage of the ReLU function is that it does not cause the gradient descent problem during the back propagation. Therefore, the ReLU function is used as the activation function in this paper. The expression is shown in formula (3).

$$\mathrm{Re\,LU}(x) = \max(0, x) \tag{3}$$

After the convolution layer, there is a pooling layer S. Its main function is to reduce the data dimension, reduce the number of parameters, and accelerate the training speed while maintaining the invariance of the data features. It also has the function of preventing overfitting. The expression of the pooling layer is shown in formula (4).

$$S = f(\beta down(x) + b) \tag{4}$$

In the formula: β represents the multiplicative scalar parameter, and b represents the offset.

The common methods of the pooling layer are maximum pooling and average pooling. The expressions are shown in formula (5) and formula (6). In this paper, the maximum pooling method is used to improve the difference of the local features of the sample and speed up the classification speed. The pooling window size is [2×2], and the moving step is 2.

$$Pooling_{\max} = \max(x_1, x_2, x_3, x_4) \tag{5}$$

$$Pooling_{avg} = \frac{x_1 + x_2 + x_3 + x_4}{4} \tag{6}$$

The data after the convolution and pooling operations are mapped to the sample space through the fully connected layer after one-dimensional processing, and used for the final classification output.

The output layer uses softmax classifier for classification, and the expression of softmax function is shown in formula (7).

$$soft\max(y)_i = \frac{e^{y_i}}{\sum_{n=1}^{N} e^{y_n}} \tag{7}$$

In this paper, the cross-entropy function is used as the cost function in backpropagation to update the weights and biases, which is expressed as formula (8).

$$L(w,b)=-\frac{1}{N}\sum_{n=1}^{N}[y_n\ln\hat{y}_n+(1-y_n)\ln(1-\hat{y}_n)]+\frac{\lambda}{2n}\|w\|^2 \tag{8}$$

The first term is the conventional cross-entropy function, and the second term is the regularization term. On the premise that other parameter values remain unchanged, comparative experiments are performed on different values of λ. The results show that the cost function is the lowest when λ=0.05, so the value of parameter λ in this paper is 0.05. yn is the actual probability of classification as flow n in the data set, and $\hat{y}_n$ is the predicted probability of classification as flow n obtained by the softmax function.

In order to improve the weight update rate under the premise of ensuring the classification accuracy, the update expressions of the learning rate E, C and B are selected as shown in formula (9) and formula (10).

In order to improve the weight update rate on the premise of ensuring the classification accuracy, the learning rate η takes the value of 0.02, and the update expressions of w and b are shown in formula (9) and formula (10).

$$w=w-\eta\frac{L(w,b)}{\partial w} \tag{9}$$

$$b=b-\eta\frac{L(w,b)}{\partial b} \tag{10}$$

4. Experimental Results and Analysis

4.1 Experimental Environment and Data

The experimental environment of this paper is in the Windows10 operating system, the Keras deep learning framework is used to build a neural network model, and the Python version is 3.6.

The data used in this article comes from the ISCX VPN-nonVPN public data set, which contains multiple types of encrypted traffic. In this paper, 5000 sets of data are selected for experiment in the pre-processed data set, of which 90% is used as training data and the remaining 10% is used as test data. The types of traffic data included in the ISCX VPN-nonVPN data set are shown in Table 1.

Table 1. ISCX VPN-nonVPN Data Set

Type of Traffic	Type of Application
Email	Email
VPN-Email	
Chat	AIM、ICQ、Skype、Facebook、Hangouts
VPN-Chat	
Streaming	Netflix、Spotify
VPN-Streaming	
File transfer	Skype、FTPS、SFTP
VPN-File transfer	
VoIP	Facebook、Skype、Hangouts
VPN-VoIP	
P2P	Torrent
VPN-P2P	

4.2 Convolutional Neural Network Parameter Settings

In order to verify the effectiveness of the method proposed in this paper, this paper designed three different parameters of the convolutional neural network model shown in Table 2, and found the optimal parameter by performing the average value of 10 experiments on each model Model, the results are shown in Figure 4 and Figure 5. The test results show that the parameters of the second group of experiments have better performance when considering the training time and classification accuracy at the same time, so this paper uses the convolutional neural network number 2 and the detailed parameter configuration is shown in Table 3.

Table 2. Three different parameters of convolutional neural network model

num	C1		C2		C3		S		Fully Connected Layer 1	Fully Connected Layer 2
	w_1	Step1	w_2	Step2	w_3	Step3	β	Step4	Filter 1	Filter 2
1	3*3*8	1	3*3*16	1	3*3*32	1	2*2	2	64	12
2	3*3*16	1	3*3*32	1	3*3*64	1	2*2	2	64	12
3	3*3*16	1	3*3*32	1	3*3*64	1	2*2	2	128	12

Table 3. Specific parameter configuration of convolutional neural network model

Layer	input	Filter	Step	Output
Convolutional Layer 1	28*28	3*3*16	1	28*28*16
Convolutional Layer 2	28*28*16	3*3*32	1	28*28*32
Convolutional Layer 3	28*28*32	3*3*64	1	28*28*64
Pooling Layer	28*28*64	2*2	2	14*14*64
Fully Connected Layer 1	14*14*64	64	-	64
Fully Connected Layer 2	64	12	-	12
Output Layer	12	-	-	12

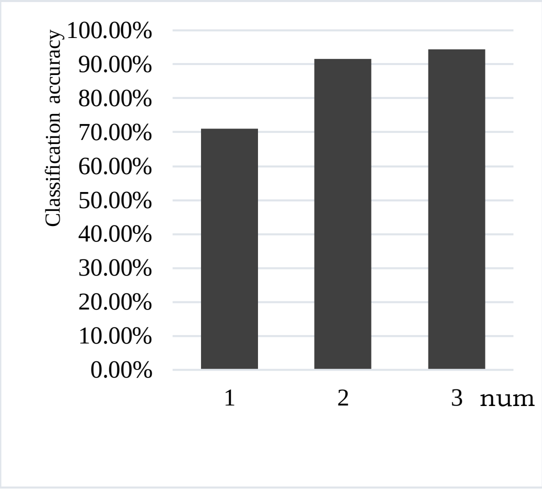

Figure 4. Classification accuracy

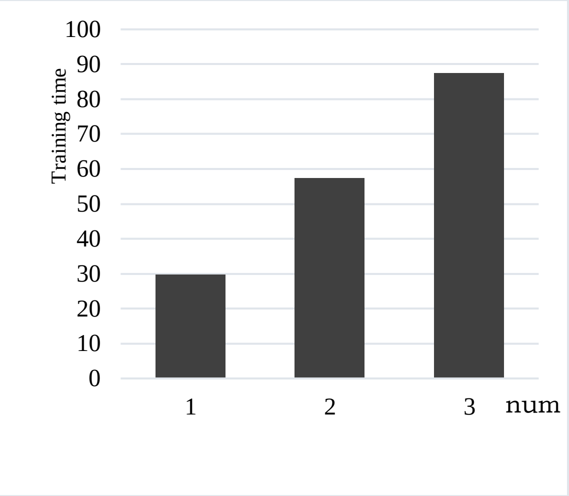

Figure 5. Training time

4.3 Evaluation Standard

This paper uses three widely used evaluation indicators to evaluate the performance of the proposed model, namely the accuracy rate, recall rate and F1 [16]. The expressions are shown as formula (11), formula (12) and formula (13).

$$P_i = \frac{TP_i}{TP_i + FP_i} \tag{11}$$

$$R_i = \frac{TP_i}{TP_i + FN_i} \tag{12}$$

$$F1_i = \frac{2 \times P_i \times R_i}{P_i + R_i} \tag{13}$$

In the formula: TP_i represents the type of flow data of type i and is correctly classified as i. FP_i represents the number of items of flow data that is not of type i but is incorrectly classified as i. FN_i represents the type of flow data of type i but is incorrectly classified as non-i Number of items.

4.4 Experimental Comparison and Result Analysis

In order to verify the effectiveness of the proposed method, this paper compares with the current one-dimensional convolutional neural network classification method, which has the great classification accuracy of encrypted classification. The experimental results are shown in Figure 6, Figure 7 and Figure 8.

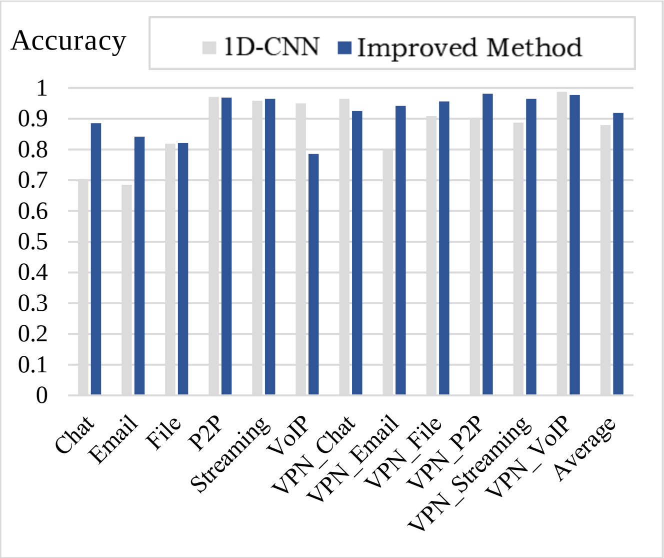

Figure 6. Accuracy comparison chart

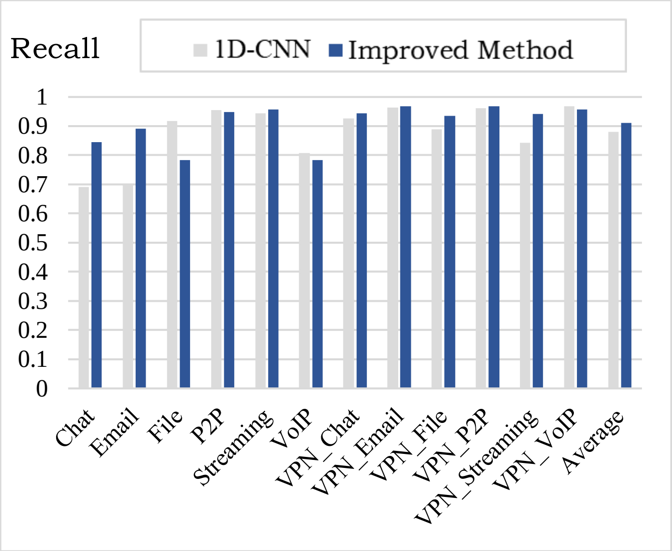

Figure 7. Recall comparison chart

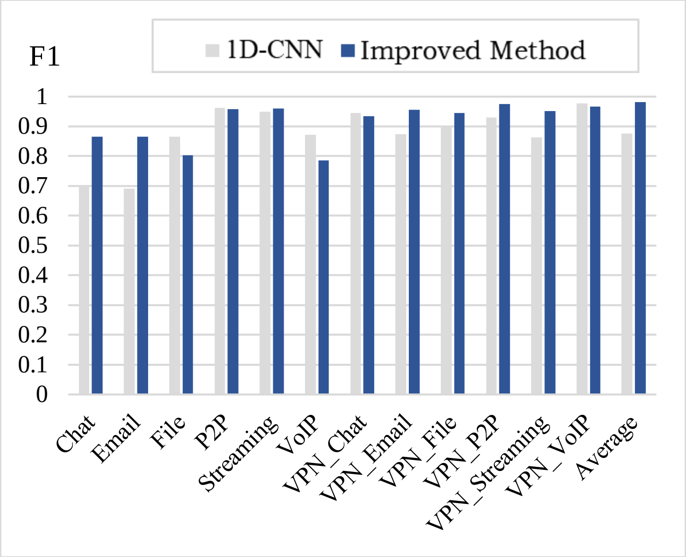

Figure 8. F1 comparison chart

As can be seen from Figure 6 and Figure 7, when the improved convolutional neural network model in this paper classifies the 12 types of encrypted traffic in the ISCX VPN-nonVPN data set, there are 6 types of traffic classification accuracy and recall rate that are significantly higher than the result of 1D-CNN encryption traffic classification method, and only two types of traffic (VoIP and File) classification accuracy and recall rate are significantly lower than the 1D-CNN method. By calculating the average classification accuracy rate and recall rate of all traffic types, it can be concluded that the results of the improved method in this paper are 4% higher than the 1D-CNN method.

Figure 8 is the result of F1 value comparison between the method used in this paper and the 1D-CNN method. F1 is a comprehensive evaluation index combining accuracy and recall. It can be seen that the average F1 value of the improved method in this paper is 10% higher than the 1D-CNN method, so the improved method in this paper improves the classification accuracy of encrypted traffic.

5. Conclusion

Aiming at the low accuracy rate of encrypted traffic classification by traditional traffic classification methods, this paper proposes an encrypted traffic classification method based on improved convolutional neural network. The convolutional neural network is optimized and adjusted in terms of network structure, network parameters, and cost function, thereby improving the classification accuracy of encrypted traffic. Finally, this paper experimented on the improved encryption traffic classification method on the public data set ISCX VPN-nonVPN and compared with the traffic classification method based on 1D-CNN to verify the effectiveness of the proposed method. The next step will be to verify and improve the proposed method in more traffic data sets and real network environments to improve the classification accuracy of different network traffic.

Acknowledgments

Key research project of Henan Provincial Department of Science and Technology (182102210293), scientific research project of Henan Provincial Department of Education (18A520007),

CERNET Next Generation Internet Technology Innovation Project (NGII20180107).

References

[1] Zhen J, Zhu G S. Research on Network Traffic Classification Method[J]. Information and Communication, 2017(08): 171-173.

[2] Chen X J, Wang P, Yu J H. CNN based Encrypted Traffic Identification Method[J]. Journal of Nanjing University of Posts and Telecommunications (Natural Science), 2018, 38(06):40-45.

[3] Yamansavascilar B, Guvensan MA, Yavuz A G, et al. Application Identification Via Network Traffic Classification[C]// 2017 Inter-national Conference on Computing, Networking and Communications (ICNC), 2017: 843-848.

[4] Fu W L, Song T, Zhou Z. Rocket TC: An FPGA-based High-performance Network Traffic Classification Architecture[J]. Journal of Computer, 2014, 37(02): 414-422.

[5] Liu Y, Li W, Li YC. Network Traffic Classification Using K-means Clustering[C]// Second International Multi- Symposiums on Computer and Computational Sciences (IMSCCS), 2007: 360-365.

[6] Tong D, Qu Y R, Prasanna V K. Accelerating Decision Tree Based Traffic Classification on FPGA and Multicore Platforms[J]. IEEE Transactions on Parallel and Distributed Systems, 2017: 3046-3059.

[7] Chen Z, Liu Z, Peng L, et al. A Novel Semi-supervised Learning Method for Internet Application Identification[J]. Soft Computing, 2017, 21(8): 1963-1975.

[8] Munther A, Rozmie R, Mosleh M, et al. A Preliminary Performance Evaluation of K-means, KNN and EM Unsupervised Machine Learning Methods for Network Flow Classification[J]. International Journal of Electrical and Computer Engineering, 2016, 6(2): 778-784.

[9] Li Z B, Han Y, Wei Z Z. Network Traffic Classification in SDN Based on Machine Learning[J]. Computer Applications and Software, 2019,36(05): 75-79, 164.

[10] Li D Q, Wang X, Yu B. Network Traffic Classification Method Based on One-Dimensional Convolution Neural Network[J]. Computer Engineering and Applications, 2019(12):1-9.

[11] Li Y Q, Huang Y, Sun X C. Network Traffic Prediction Model Based on Deep Belief Echo-State Network[J]. Journal of Nanjing University of Posts and Telecommunications(Natural Science), 2018, 38(05):89-94.

[12] Wang W, Zhu M, Wang J, et al. End-to-end Encrypted Traffic Classification with One-Dimensional Convolution Neural Networks[C]// 2017 IEEE International Conference on Intelligence and Security Informatics (ISI). IEEE, 2017: 43-48.

[13] Zhou F Y, Jin L P, Dong J. Review of Convolutional Neural Network[J]. Chinese Journal of Computers, 2017,40(06): 1229-1251.

[14] Wang Y, Zhou H Y, Feng H, et al. Network traffic Classification Method Based on CNN[J]. Journal on Communications, 2018,39(01): 14-23.

[15] Xia M Y, Hu S Y, Zhu S M, et al. Research on the Method of Network Attack Detection Based on Convolution Neural Network[J]. Netinfo Security, 2017(11):36-40.

[16] Guo L, Wu Q, Liu S, et al. Deep Learning-based Real-time VPN Encrypted Traffic Identification methods[J]. Journal of Real-Time Im age Processing, 2019: 1-12.