Journal of Neural Information Processing (2019) Vol. 1: 1-5
Clausius Scientific Press, Canada
DOI: 10.23977/jnip.2019.11001
ISSN 2522-6517

Network Traffic Classification Method Based on One-Dimensional Convolution Neural Network

Wang Peng[1,a,*]

[1]*School of Economics and Management, Dalian University, No.10, Xuefu Avenue, Economic & Technical Development Zone, Dalian, Liaoning，The People's Republic of China（PRC）*
a. email：wangpeng1@dlu.edu.cn

Keywords: *one-dimensional convolution neural network; network traffic classification; data preprocessing; parameter optimization; deep learning*

Abstract: Aiming at the bottleneck problem of traditional machine learning algorithm for traffic classification, an application traffic classification algorithm based on one-dimensional convolution neural network model of deep learning algorithm is proposed. Firstly, the data set of network traffic is preprocessed. In the data preprocessing stage, the irrelevant data fields are removed and the input characteristics of convolution neural network are satisfied. Then, a new one-dimensional convolution neural network model is proposed, and the optimal classification model is constructed from the aspects of network structure, hyper-parametric space and parameter optimization. This model solves the problem of feature selection in traditional traffic classification algorithm based on machine learning by self-learning data features in convolution layer. Finally, compared with the traditional one-dimensional convolution neural network model, the designed neural network model improves the classification accuracy by 16.4%, and the total classification time was saved by 71.48%.In addition, the class accuracy, recall rate and F1 score have been improved.

1. Introduction

Traffic classification is of great significance in network traffic engineering. By accurately classifying traffic, service quality assurance, resource rationalization, malware detection, and intrusion detection can be achieved. In recent years, research on traffic classification has been continuously developed. Initial the method is a port-based traffic classification method. Although this classification method is simple and efficient, the accuracy of this classification method is declining due to the appearance of many application spoofing port numbers or not using the standard registration port number. The research method of traffic classification using the deep packet inspection mechanism solves the problem caused by the port-based classification method. However, the method of traffic classification based on the payload information in the data packet is only applicable to unencrypted traffic and has high computational overhead. Therefore, a new generation of methods based on statistical features has emerged that rely on statistical features or time series characteristics to handle both encrypted and unencrypted traffic. The statistical feature-based classification method usually adopts the classical machine learning (ML) algorithm, uses the

unsupervised clustering algorithm (K-means) for traffic classification and achieves 90% classification accuracy. Later, the supervisory classification was used in the literature. Algorithms (such as KNN, C4.5, Naive-Bayes) and semi-supervised classification algorithms classify traffic and achieve high classification accuracy. Machine learning-based traffic classification methods limit their scalability due to the selection of dependent features. Recent research on the use of machine learning algorithms for traffic classification is mainly based on the selection of optimized features[1].

With the advent of artificial intelligence and big data era, deep learning algorithms have been well applied in various fields such as image recognition, natural language processing and sentiment analysis. This learning algorithm automatically selects features through the training process, which can be used to solve the feature selection problem of network traffic classification using machine learning algorithms. The use of multi-layer perceptron (MLP) for traffic classification is compared with the traffic classification effect of traditional machine learning algorithms. However, since the classification data sets used by each classifier in the paper are not the same, the final result cannot compare the deep learning algorithm with the machine learning algorithm. The accuracy of the traffic classification. An end-to-end traffic classification model was proposed for the first time. Using the traditional one-dimensional convolutional neural network structure, the data was processed into a specific file form. The feature space was constructed by convolutional neural network, and then classified by classifier and obtained better. Classification results. Based on its research, this paper redesigns the one-dimensional convolutional neural network model, and improves the traffic classification effect of one-dimensional convolutional neural network for encrypted applications from neural network structure, data preprocessing method, cost function and gradient optimization[2].

2. Application traffic classification problem description

Existing network access link structures designed based on more content downloaded by the client than they upload are asymmetric, but symmetrically demanding applications such as peer-to-peer (P2P) applications, voice over IP (VoIP) and video The ubiquity of calls) changed the customer's needs and deviated from the original design. Therefore, in order to provide a satisfying experience for customers, additional application-level knowledge is required to allocate sufficient resources for these applications.

The growing demand for user privacy and data encryption has dramatically increased the amount of encrypted traffic on the Internet today. The encryption process converts the raw traffic data into a pseudo-random format that makes it difficult to decrypt. This encryption process causes the traffic data to contain almost no data patterns that can be used to identify network traffic. Therefore, accurate classification of encrypted application traffic has become a challenge for modern networks.

The selection of features is the research bottleneck of traditional traffic classification. The convolutional neural network used in this paper can learn the autonomous selection feature from the original dataset through the multi-layer convolutional network, and then construct the feature space according to the convolutional neural network's own network structure and continuously optimize the feature space through the training of a large amount of data. This classification method not only solves the difficulty of feature selection but also provides the possibility of online traffic classification[3].

3. D-CNN traffic classification method

3.1. Dataset source

The study of traffic classification methods based on convolutional neural networks is supervised and requires a fully labeled data set as a training subset. This paper uses the network public data set "ICSX VPN nonVPN", which has been studied by many literatures. The statistics of the traffic data set collected in this paper are shown in Table 1. The data set traffic type contains packets captured through a virtual private network (VPN) session. A VPN is a private overlay network between distributed sites that runs on a public communication network (such as the Internet) to mine traffic. The most prominent aspect of VPNs is ensuring secure remote access to servers and services through tunneled IP packets. Similar to regular (non-VPN) communication, VPN communication is captured based on different applications, such as voice calls, video calls, and chat. This paper classifies the network traffic according to the types of applications. The number of applications is very important for the design of the neural network model structure.

3.2. Data preprocessing

Since the "ICSX VPN-nonVPN" data set is captured at the data link layer, it includes an Ethernet header. It contains information about the physical link, such as the Media Access Control (MAC) address, which is essential for the forwarding of frames in the network, but it has no specific meaning for application identification or traffic characterization tasks. . Therefore, in the data preprocessing stage, the Ethernet header is first deleted. In the transport layer, Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) differ in header length. The former header length is usually 20 bytes long, while the latter is only 8 bytes. To make the length of the transport layer segment consistent, add zero at the end of the UDP segment header to make it equal to the length of the TCP segment header. The packet is then converted from bits to bytes, which helps to reduce the input size of the convolutional neural network.

Because the data set is captured in the real network, it contains some data packets that are not of interest in this paper, such as SYN, ACK, and FIN flag segments in the process of establishing a connection or completing a connection. These fields do not carry information about generating them. The information of the application, so these segments can be deleted. In addition, there are some Domain Name Service (DNS) segments in the data set. These fields are used for host name resolution, which translates URLs into IP addresses that are not related to traffic classification and are therefore removed from the data set.

After the above steps, the data packet file already meets the requirements of the traffic classification data set. However, in addition to considering the payload problem in the data packet, the input requirements of the convolutional neural network need to be considered. Figure 2 shows the packet length probability density distribution after processing in the above steps. As shown in the histogram, the packet length in the data set varies greatly. Using a convolutional neural network model requires a fixed-size input, so you need to choose a fixed packet length. Based on the experience of previous studies, using the full packet length will increase the space and time complexity required by the algorithm, so this article selects the same first 784 bytes for the payload with a payload of less than 784 bytes at the end. Do the zero-fill operation. For better performance, a min-max normalization operation normalizes all input values into the range [0:1].

4. Solution implementation and testing

In order to verify the feasibility of the proposed one-dimensional convolutional neural network model for traffic classification, this paper uses the Keras library, the back end is TensorFlow, deployed on a processor is Inter i5-6500, clocked at 3.2G, memory 4G, 64 On a Linux operating system machine.

Five evaluation indicators were used to compare the models: accuracy, classification time, recall rate, class accuracy, and F1 score. Accuracy is used to assess the overall performance of the classification model. Higher accuracy means better classification. The F1 score, the recall rate, and the accuracy rate can be used to observe the classification model for each category. Where TP is the number of instances correctly classified as x, TN is the number of instances correctly classified as not - x, FP is the number of instances incorrectly classified as x, and FN is the number of instances incorrectly classified as not - x.

This study developed a data preprocessing tool to process the network traffic dataset into a form similar to the minist dataset, using the traditional one-dimensional convolutional neural network LeNet-5 model for traffic classification. The data preprocessing method proposed in this paper preserves the serialization characteristics of network traffic and removes the fields that are meaningless for traffic classification. In the experimental environment platform of this paper, the accuracy, class accuracy, recall rate, F1 score and model classification time are compared with the traditional one-dimensional convolutional neural network model. The comparison results are shown in Figure 5-7. In order to ensure the reliability of the overall classification time comparison results, the comparison experiments of the classification time of the two models are performed on the machine with Inter i5-6500, main frequency 3.2G, memory 4G, 64-bit Linux operating system. CPU frequency, maximum CPU frequency, minimum CPU frequency, and Inter on a processor.

5. Conclusions

Based on the traditional machine learning algorithm for traffic classification research and related research, a new one-dimensional convolutional neural network model is proposed based on neural network structure, data preprocessing method, hyperparameter space and gradient optimization. Firstly, a new one-dimensional convolutional neural network model structure is designed. The network public data set is preprocessed, the unrelated fields are removed, and the fixed length is intercepted, and then standardized operation is performed as the input of the convolutional neural network. Then, through the characteristics of convolutional neural network structure, the higher-level characteristics of network traffic are learned autonomously. After the network model is trained, the parameter space is optimized and the feature space is constructed to complete the traffic classification task of the network application. Finally, the traffic classification algorithm proposed in this paper is verified by the network public data set and the classification effect is achieved. Because the research of deep learning algorithm in network traffic classification is still in an immature stage, this paper should continue to work on the classification of real network traffic data in the future research direction.

Acknowledgements

This article was specially funded by Dalian University's 2019 Ph.D. Startup Fund (20182QL001) and 2019 Jinpu New District Science and Technology Project.

References

[1] Youngmahn Han,. (2017) Deep convolutional neural networks for pan-specific peptide-MHC class I binding prediction, BMC Bioinformatics, 8，96-108

[2] Clarence White. (2017) CNN-BLPred: a Convolutional neural network based predictor for β-Lactamases (BL) and their classes, BMC Bioinformatics, 7, 356-374.

[3] Zhehuan Zhao. (2018) Disease named entity recognition from biomedical literature using a novel convolutional neural network. BMC Medical Genomics, 10, 89-108.