Attack Traffic Detection Based on LetNet-5 and GRU Hierarchical Deep Neural Network

Zitian Wang[1], ZeSong Wang[1], FangZhou Yi[1], and Cheng Zeng[1,2,3(✉)]

[1] School of Computer Science and Information Engineering, Hubei University, Wuhan 430062, China
zc@hubu.edu.cn

[2] Hubei Province Engineering Technology Research Center for Software Engineering, Wuhan 430062, China

[3] Hubei Engineering Research Center for Smart Government and Artificial Intelligence Application, Wuhan 430062, China

Abstract. The paper converts the network traffic information about a single-channel grayscale image as input data. In addition, a deep hierarchical network model is designed, which combines LetNet-5 and GRU neural networks to analyze traffic data from both time and space dimensions. At the same time, two networks can be trained simultaneously to achieve better classification results because of the reasonable network association method. This paper uses the CICID2017 dataset, which contains multiple types of attacks and is time-sensitive. The experimental results show that, through the combination of deep neural networks, the model can classify attack traffic with extremely high accuracy.

Keywords: Attack traffic detection · Multi-dimensional layered network · CNN network · GRU network

1 Introduction

In this age, deep learning network has been widely used, including network intrusion detection[1]. Few researchers consider the influence of time on network intrusion detection. The detection process is improved compared with the detection without considering the time characteristics, but the improvement in accuracy is limited [2]. Based on the above research results, this paper makes further optimization and attempts, and achieves better results in the time dimension, while ensuring high accuracy.

Supported by National Natural Science Foundation of China 61977021.
Supported by National Natural Science Foundation of China 61902114.
Supported by Hubei Province Technological Innovation Foundation 2019ACA144.

Z. Liu et al. (Eds.): WASA 2021, LNCS 12939, pp. 327–334, 2021.
https://doi.org/10.1007/978-3-030-86137-7_36

2 Related Works

In the previous research, the network traffic information is processed as one-dimensional data, and some scholars creatively splice the network traffic information as two-dimensional data [3]. The experimental results show that the classification method can achieve a good classification effect when the training set is large enough. Based on the time characteristics of network traffic information, some scholars have proposed a recursive neural network model for deep learning [4]. By studying the temporal characteristics of network traffic information, the error rate can be reduced by 5% based on the traditional machine learning model. Because of the obvious advantages of the LSTM network in processing time-series data compared with the traditional RNN network, some scholars try to classify network traffic data based on LSTM network [5]. It is proved that the LSTM network has the potential to analyze network traffic information.

3 Methodology

In this section, we establish a layered deep network model to detect attract network traffic. The model consists of two different neural network algorithm models. The first layer is letnet-5 convolutional neural network to extract the spatial characteristics of the stream, and the second layer is the GRU network to extract the temporal characteristics of the stream. By adjusting the output of the first layer network, the two networks can be trained at the same time. Before elaborating the model in detail, we first declare how to preprocess the network traffic data.

3.1 Data Preprocessing

Compared with traditional feature engineering, we only need to keep all the information in the network traffic packet and map it to the corresponding format to meet the classification requirements. The data package comes from the capture software and is displayed in hexadecimal.

1. data: Through the early research [6], it show that the fields of Ethernet layer, MAC source address, MAC destination address and protocol version do not need to be the characteristics of network traffic data. These data need to be removed which has an impact on our detection.
2. split: All data are divided according to address, port and time information. Because of a large number of data packets, split cap software [7] can be used to achieve packet splitting.
3. convert: Statistics show that more than 90% of the data streams contain less than 10 packets, and the rest of the data streams contain more than 10 packets. If the length of the network traffic packet is less than 160 bytes, select to add 0 to 160 bytes at the end. At the same time, to make the features more universal, we select the top 10 packets in each data stream. If the number of data packets is insufficient, we will supplement the data packets with all 160 bytes of data as 0. After this conversion, we get 1600 bytes of raw data.

3.2 CNN Model

CNN convolution neural network has good spatial awareness and has achieved excellent results in image processing [8]. The preprocessed data is equivalent to a single-channel 40×40 grayscale image. More complex convolution network models such as ResNet [9]and VGGNet [10] need to add a lot of data to the analysis of this image. Given that the proportion of abnormal traffic in network traffic analysis is too low, if you choose to use an overly complex identification network, it will easily lead to over-fitting of the identification network and reduce the accuracy of model prediction. So we chose an improved LetNet-5 network for handwritten number recognition, which uses a model similar to our scenario, with single-channel low-pixel images.

In the hidden layer of the CNN network, we use two convolution layers and two maximum pooling layers to extract the spatial characteristics of the original network traffic data. After processing, the original single-channel 40×40 pictures are converted to 8×8 pictures with 64 channels. After fully expanding them, we get the 4096-dimensional vector and transfer it to the output layer of the CNN network. The output layer uses the full junction layer, and the full junction layer uses 1600 neurons. This transformation preserves the same dimensionality of the original data after it is extracted. Considering the occurrence of over-fitting, deletion after the fully connected layer inactivates some neurons, a process that is completely random to ensure the reliability of the experimental results.

3.3 GRU Model

In the scenario of attack traffic detection, packets are forwarded in a time sequence, and due to the delay in transmission, there is a sequence of packets at the receiving end. In addition, the number of packets sent within a timestamp also varies dynamically. The above characteristics indicate that network traffic data has time characteristics and are suitable for research using recursive neural networks. Due to design flaws in general RNN networks, all incoming information from the previous layer is recorded as valid information by default. This paper uses the GRU network structure. As an improvement of the RNN network, the same gate control is used to forget and select the information from the previous moment. It is better than the RNN network in dealing with long-time series problems [11]. In this paper, the GRU network is used to extract the time feature of the original stream data automatically, and the GRU network uses two-layer unit to extract the time feature. Each cell of GRU includes 256 hidden layer units, and the activation function of each layer uses S-type function for nonlinear operation. The last layer of GRU network uses the fully connected layer, and the number of neurons in the fully connected layer is equal to the number of flow categories.

3.4 Deep Hierarchical Network

The research shows that network traffic data contains a large number of features [12]. In the past operation, researchers are used to defining part of the

characteristics to study. In the first mock exam, however, the selection is just passable and the information is not well utilized. Using the combination of CNN and GRU network, this paper analyzes all the information from the time and space of network traffic and realizes the comprehensive mining of the characteristics in the flow. In order to train the two networks at the same time, the output format of CNN is set to meet the input format of the GRU network. According to the design of this paper, each stream extracts the first 10 packets and each packet extracts the first 160 bytes, which are mapped to the GRU network, corresponding to the time step and input size respectively. Finally, we use a softmax classifier, which can output the classification probability of each stream, and map the experimental results to the corresponding categories according to the highest probability. The loss function used in the model is the mean square loss function, and the trained optimizer [13] uses an adaptive matrix for gradient descent. The model structure is shown in Fig. 1.

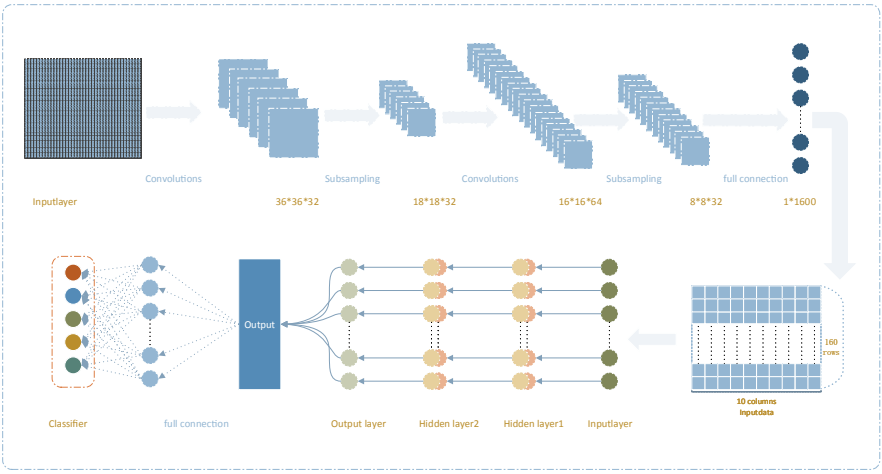

Fig. 1. Deep hierarchical network model.

4 Experiments

4.1 Data

The CICID2017 data set is open source provided by the Canadian Institute of Cyber Security. Considering the reliability of the training results, we selected the top ten attack traffic and normal traffic as our training set and test set, ensuring that each type contains at least two thousand traffic data.

4.2 Calculation Process

The two deep networks in this article are connected to perform operations. First, starting from the CNN network, after data preprocessing, we have obtained a 40×40 single-channel grayscale image. The kernel size used in the first layer convolution operation is 32 5×5; the kernel size used in the second layer convolution operation is 64 3×3, a total of 64. After transformation, the original data is downsampled to $8 \times 8 \times 64$, and finally, 1600-dimensional features can be output through the fully connected layer. Combining 10 continuous data packets is the requirement of our selection of features and the input requirement of the GRU network. This article artificially sets the time step of each training of the GRU network to 10, so that a 10×160 matrix is obtained as the GRU network Enter the content. The GRU network structure has two layers and each layer contains 256 neurons. Since the data first passes through the CNN network and then the GRU network, the CNN network will be trained first in the forward training process. The loss calculated in the reverse process of training the GRU network will be calculated first and then the CNN network will be calculated. The reason for this design is to link the temporal and spatial characteristics of the data so that the performance of the classifier can be higher.

4.3 Experimental Results

In order to comprehensively consider the performance of the model and compare it with other models, this paper tests the two-class and multi-class performance of the model. The content of the two-classification experiment is to classify normal traffic and attack traffic, and does not require the specific types of attack traffic to be classified, while the multi-class experiment requires the specific types of attack traffic to be classified. Table 1 records the average results after multiple binary classification experiments, and Table 2 records the average results after multiple classification experiments. This article focuses on comparing the possible effects of different time series networks on the experimental results, and chooses three types of neural networks: RNN, LSTM, and GRU for comparison.

Table 1. Multiple binary classification

Metrics	Accuracy	Precision	Recall	F1-measure	Time
RNN	0.9976	0.9978	0.9963	0.9971	149.45
LSTM	0.9980	0.9987	0.9958	0.9972	244.20
GRU	0.9966	0.9988	0.9917	0.9952	240.40
CNN	0.9986	0.9975	0.9991	0.9983	234.66
CNN+RNN	0.9977	0.9947	0.9990	0.9968	338.27
CNN+LSTM	0.9981	0.9958	0.9994	0.9976	449.31
CNN+GRU	**0.9992**	**0.9993**	**0.9995**	**0.9991**	439.85

Table 2. Multiple classification

Metrics	Accuracy	Precision	Recall	F1-measure	Time
RNN	0.9953	0.9973	0.9993	0.9983	156.98
LSTM	0.9965	0.9974	0.9997	0.9986	245.16
GRU	0.9976	0.9979	0.9997	0.9989	246.41
CNN	0.9975	0.9986	0.9994	0.9990	236.19
CNN+RNN	0.9957	0.9986	0.9983	0.9984	352.70
CNN+LSTM	0.9963	0.9979	0.9997	0.9988	453.07
CNN+GRU	**0.9977**	**0.9991**	**0.9998**	**0.9992**	442.20

Tables 1 and 2 records the results of 20,000 iterations of the model. As shown in the table, the improved depth hierarchical network model proposed in this paper has advantages in each index and has obvious advantages inaccuracy. Compared with the model of combining CNN network and LSTM network proposed by previous scholars [14], our model not only has advantages of inaccuracy but also reduces the convergence rate. According to the data in Table 1 and Table 2, we can see that the accuracy has been improved by 0.1%, and the training time has been shortened by about 10 s. Compared with training the deep neural network separately, the hierarchical structure reduces the convergence speed of the whole model by nearly 10%, and improves the efficiency significantly when training the large data model. At the same time, we find that our model does not have obvious advantages in multi-classification problems, and many indicators of GRU and CNN models are very close to our model. However, in the two-class problem, our model can get considerable advantages on each index. We think that the reason for this is probably that the amount of data in the multi-classification problem is not enough to make the model fully understand the characteristics of each kind of attack traffic, and the index of the multi-classification experiment is lower than that of the two-classification model as a whole. This also shows that the model proposed in this paper is suitable for the treatment of large data problems. When the amount of data is sufficient, the advantages of the model proposed in this paper in accuracy and convergence speed will be obvious. The extremely high accuracy rate of the artificial model in this article is due to the following reasons:

1. In each data packet, after transforming the original one-dimensional network traffic information into a two-dimensional single-channel grayscale image, combining the originally distant features is beneficial to algorithm analysis to obtain new features. Generally, algorithms for one-dimensional data mainly study the relationship between the front and back of the data. If the two parts of the data are too far apart, it is difficult for existing algorithms to combine the two parts as a combination of features for classification. When the data is processed as a graph, through the convolution operation of the

convolution kernel, this paper realizes the combined analysis of a variety of distant features.

2. Since this article takes the first ten data packets for each data stream for analysis, it is obvious that for each data stream, there are certain timing characteristics before these ten data packets. For this reason, this article adds a GRU neural network based on the improvement of the recurrent neural network to analyze the temporal correlation between the input data.
3. It is necessary to combine the two neural networks for training. If it is a horizontal splicing method, each network is trained separately, and the classification results of the two networks are combined into the final classification result according to certain weight distribution. To further supplement the experimental results, this article supplements other indicators in the horizontal splicing mode. The number of iterations of the horizontal stitching experiment is the same as that of the layered model, and the weights predicted by the two models are 50% each. According to supplementary experiments as Table 3 shows, it can be found that the horizontal stitching method has no obvious effect on the detection results, and some indicators are even inferior to the results of the model training separately.

Table 3. Horizontal splicing model

Metrics	Accuracy	Precision	Recall	F1-measure
CNN+RNN_binary	0.9959	0.9966	0.9973	0.9970
CNN+LSTM_binary	0.9963	0.9979	0.9977	0.9988
CNN+GRU_binary	0.9979	0.9988	0.9987	0.9988
CNN+RNN_multiple	0.9947	0.9976	0.9973	0.9974
CNN+LSTM_multiple	0.9954	0.9979	0.9987	0.9976
CNN+GRU_multiple	0.9967	0.9981	0.9988	0.9981

5 Conclusion

This paper proposes a hierarchical network model based on LetNet-5 and GRU to detect attack traffic. According to the characteristics of network traffic data, this paper selects some continuous data packets in each network flow and converts certain data in each data packet into a single-channel grayscale image as input data for network attack traffic detection. The data selected in this way not only retains the timing relationship between the same stream data packet but also fully combines the characteristics of the data in each data packet. Tested by CICIDS2017, the model proposed in this paper can reach a very high level in terms of accuracy, precision, recall rate, and F1-measurement. Compared with the combined deep neural network model with close performance indicators, the model proposed in this paper has an advantage in convergence speed. Based on

the above characteristics of the model, this article believes that the model is suitable for network attack traffic detection in the case of big data, such as in a data center.

Besides, the model also has room for further improvement. Considering that the data of the attack traffic in the real situation is less, the model can be further optimized to use fewer data to train the model while ensuring the detection effect of the model.

References

1. Zeng, Y., Gu, H., Wei, W., Guo, Y.: $Deep - full - range$: a deep learning based network encrypted traffic classification and intrusion detection framework. IEEE Access **7**, 45182–45190 (2019)
2. Yeo, M., et al.: Flow-based malware detection using convolutional neural network. In: 2018 International Conference on Information Networking (ICOIN), pp. 910–913. IEEE (2018)
3. Zhou, H., Wang, Y., Lei, X., Liu, Y.: A method of improved CNN traffic classification. In: 2017 13th International Conference on Computational Intelligence and Security (CIS), pp. 177–181. IEEE (2017)
4. Yuan, X., Li, C., Li, X.: Deepdefense: identifying DDoS attack via deep learning. In: 2017 IEEE International Conference on Smart Computing (SMARTCOMP), pp. 1–8. IEEE (2017)
5. Kim, J., Kim, J., Thu, H.L.T., Kim, H.: Long short term memory recurrent neural network classifier for intrusion detection. In: 2016 International Conference on Platform Technology and Service (PlatCon), pp. 1–5. IEEE (2016)
6. Denning, D.E.: An intrusion-detection model. IEEE Trans. Softw. Eng. **2**, 222–232 (1987)
7. Taylor, V.F., Spolaor, R., Conti, M., Martinovic, I.: Robust smartphone app identification via encrypted network traffic analysis. IEEE Trans. Inf. Forensics Secur. **13**(1), 63–78 (2017)
8. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
10. Liu, Z., Li, J., Shen, Z., Huang, G., Yan, S., Zhang, C.: Learning efficient convolutional networks through network slimming. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2736–2744 (2017)
11. Fu, R., Zhang, Z., Li, L.: Using LSTM and GRU neural network methods for traffic flow prediction. In: 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC), pp. 324–328. IEEE (2016)
12. Ahuja, R.K., Magnanti, T.L., Orlin, J.B.: Network flows: Theory. Algorithms, and Applications, 526 (1993)
13. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Zhang, Y., Chen, X., Jin, L., Wang, X., Guo, D.: Network intrusion detection: based on deep hierarchical network and original flow data. IEEE Access **7**, 37004–37016 (2019)