Electronics, Communications and Networks
A.J. Tallón-Ballesteros et al. (Eds.)

doi:10.3233/FAIA231192

Network Traffic Recognition and Classification Based on Deep Learning

Zhihao SONG[a], Lanhua ZHANG[a], Jin WANG[b], Xiaoyan WANG[a,1]
[a]*School of Medical Information and Engineering, Shandong First Medical University and Shandong Academy of Medical Sciences, Tai'an, 271016, China*
[b]*Cybersecurity Industry Development Center (Information Center) of Ministry of Industry and Information Technology, Beijing, 100846 ,China*

Abstract. Deep learning of network traffic is a research method that mimics the structural functions of the human nervous system to identify, classify, and predict data. We propose a new model based on Conv-LSTM to improve the accuracy and efficiency of network encrypted traffic recognition. Based on the public CIC-ISD2017 dataset, the new model is tested and measured, and evaluated based on the constructed confusion matrix and ROC graph. Comparing it with traditional Conv-LSTM, decision tree method, and RF&LSTM methods, it was found that the new model performs better and can perform well in multi classification tasks, with an accuracy rate of up to 99.60%. This model provides a reference solution for relevant applications in the field of network security.

Keywords. Deep Learning, Conv-LSTM, classification, network traffic

1. Introduction

Due to the construction of infrastructure, and the development of various services provided by the Internet, society's dependence on the Internet has significantly increased globally. The Internet has become an important component of individuals and businesses' daily operations, resulting in a large and complex amount of network traffic. This also makes network traffic identification and classification increasingly important in areas such as network security and performance optimization [1].

Network traffic identification and classification is one of the key technologies in network management and network security. By identifying and classifying network traffic, it can better detect network conditions, predict potential network security threats, optimize network performance, and so on. The old-fashioned network traffic classification methods are mainly based on statistics and data mining techniques, which manually set features and rules, consume a lot of manpower and material resources, and are difficult to adapt to complex network environments, resulting in low efficiency [2]. In recent years, with the progress of science and technology, machine learning has made significant progress. In terms of model complexity, machine learning can be divided into classical machine learning and deep learning [3]. Classical machine learning includes distance weighted KNN, clustering analysis, random forests, and

[1]Corresponding Author, Xiaoyan WANG, School of Medical Information and Engineering, Shandong First Medical University and Shandong Academy of Medical Sciences, Tai'an, 271016, China; E-mail: xywjxc@126.com.

decision trees, while reinforcement learning, Long and short-term neural networks (LSTM), recurrent neural networks (RNN) and convolutional neural networks (CNN) belong to more complex deep learning. For machine learning based traffic recognition methods, the classification ability of classifiers is usually evaluated using evaluation indicators in the field of machine learning, such as true positive, false positive, true negative, false negative, precision, recall, F1 value, etc. [4]. Ahmad Azab preliminarily discussed the challenges faced by various technologies and found that no solution can provide perfect solutions in terms of accuracy, computing resources, speed, early detection, and immune evasion. However, deep learning solutions are still relatively effective [5]. Kim et al. proposed an artificial intelligence intrusion detection system using deep neural networks, which achieved 99% accuracy through training and testing on the KDD Cup99 dataset [6]. Jihyun Kim constructed an IDS model and tested it on the KDD Cup99 dataset, achieving a false positive rate of 10.04% [7]. Wankhede et al. tested the MLP and RF algorithms on the CIC-ISD-2017 dataset with accuracy of 98% and 99%, respectively [8]. Khan et al. proposed a model based on Spark and convolutional LSTM networks, with an accuracy of 97.2% for the ISCX-UNB dataset [9].

We provide four models in this project. The first is the decision tree model, which is a traditional machine learning algorithm; the second is the random forest and short-term memory network model, which combines classical machine learning with deep learning the convolutional neural network; the third one is a model that combines convolutional neural networks with long short-term memory networks, known as ConvLSTM; the fourth one is an improved version of the ConvLSTM model, which combines multiple deep learning algorithms. All models were trained and tested on the public dataset CIC-ISD2017, and evaluated for comparison.

2. Methods

2.1. Data Sources and Data Preprocessing

The dataset used in this article is the public dataset CIC-ISDS-2017 collected by the Canadian Communications Security Agency and the Cybersecurity Research Institute in a collaborative project. In this project, 11 datasets since 1998 were evaluated, and the results showed that most datasets are still suitable for current research. The CIC-ISDS-2017 dataset contains benign and various latest common attacks, closer to real world data (PCAPs). Its data collection ended at 5:00 pm on Friday, July 7, 2017, and it collected data for 5 days. Monday is a normal day, only including normal traffic. Attacks implemented include brute force FTP, brute force SSH, DoS, Heartbleed, web attacks, infiltration, botnets, and DDoS [10]. This dataset assigns network traffic data to 84 feature columns and one label column, and performs various classifications in the label column, such as BENIGN, DDoS, bot, etc. This dataset can be used for binary classification training or multi classification training. The dataset comes from the websites (https://www.unb.ca/cic/datasets/ids-2017.html).

Before conducting training, it is necessary to preprocess the CIC-IDS-2017 dataset. The specific steps are as follows: 1) As the dataset files downloaded from the official website are multiple files separated by date, in order to facilitate subsequent dataset input, the dataset is merged. 2) Remove dirty data from the dataset. 3) Due to imbalanced dataset data, normal data accounts for the majority, and the proportion of

various attack traffic is very small. It is necessary to balance the dataset and expand the relatively small number of attack data. Except that, in some models, data dimensionality reduction operations have been added, which reduces the dimensionality and complexity of the data, reduces the memory occupied by the data, reduces computational complexity, and improves the efficiency of the algorithm.

2.2. Data Models

Decision tree is a classification and regression model based on tree structure. The origin of decision trees originated from the Hunt algorithm proposed by Hunt et al. in 1966 [11], but it was Quinlan [12] that truly made decision trees widely studied and used.

Random forest is a classifier that uses multiple trees to train and predict samples. It was first proposed by Leo Breiman and Adele Cutler, and it constructs multiple decision trees to improve the accuracy of the model. LSTM is a time recursive neural network that introduces memory units to store information and three gates (input gate, forget gate, output gate) to control the inflow and outflow of information [13].

CNN is a commonly used neural network model in deep learning. Compared with traditional neural networks, it adds convolutional layers and pooling layers, etc. It usually has higher accuracy and can automatically perform feature dimensionality reduction to reduce computational complexity, making the model more efficient. When facing very complex features, CNN can generate visual feature maps, which can better explain the model classification results. The ConvLSTM model was first proposed by the Hong Kong University of Science and Technology in 2015 [14]. It is based on the CNN-LSTM structure and has strong spatiotemporal perception ability, as it can simultaneously process data in both temporal and spatial dimensions. This feature enables the model to learn more complex spatiotemporal features, thereby improving prediction accuracy.

In the enhanced version of ConvLSTM, we have integrated convolutional layers and bidirectional LSTM layers in a sequential manner, introducing dropout layers and L2 regularization penalty mechanism. Moreover, we have incorporated a multi-head self-attention mechanism and implemented a learning rate decay strategy (StepLR). Through a series of modifications and optimizations applied to the original model, we have successfully enhanced the model's generalization, robustness, and accuracy, ultimately leading to significantly improved training results.

3. Results

We first test using a decision tree model. In this model, it recursively segments the input data into the model, ultimately generating a decision tree. At the nodes, we use the default information gain of sklearn as the measurement method for node segmentation. And the confusion matrix was output (Fig. 1), and the final accuracy of the model was 99.54%.

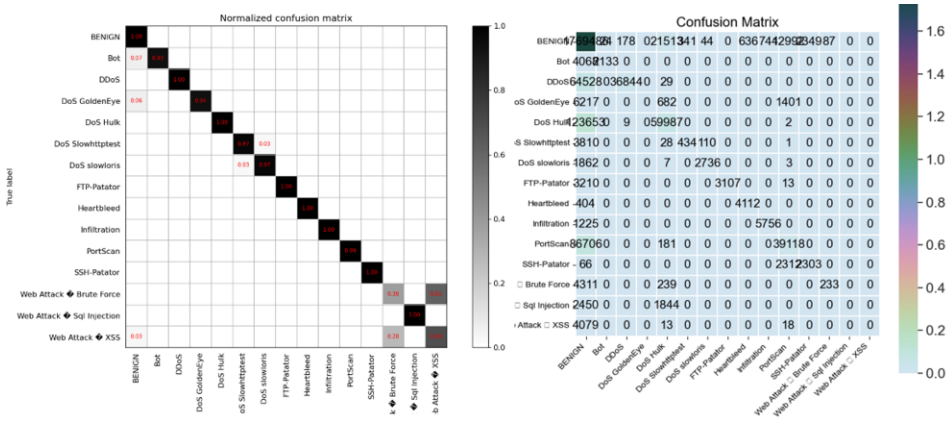

Figure 1. Confusion matrix results in decision tree model & Confusion matrix results in RF&LSTM model

Secondly, we used random forest for validation. This model used random forest for feature importance score calculation, screened features with an importance score greater than 0.02, and then defined an RF&LSTM model [15] for multi classification prediction of the dataset. The final model accuracy was approximately 95%, and the confusion matrix was drawn as shown in Fig.2.

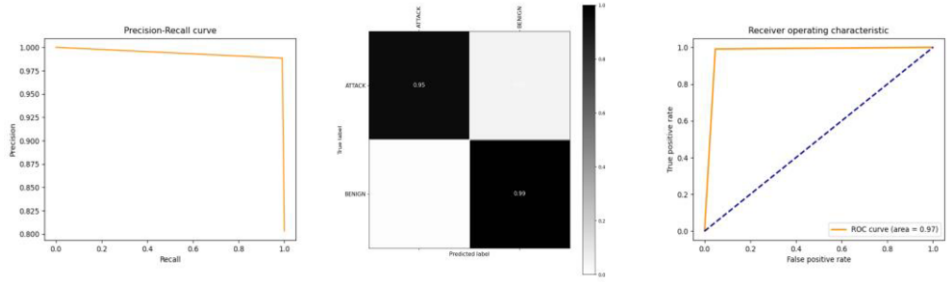

Figure 2. PR, Confusion matrix and ROC results of ConvLSTM model in binary classification

Then, we used ConvLSTM for testing, and when using the traditional ConvLSTM model, we obtained an accuracy of 97.89%. After completing the multi classification task, we also conducted binary classification testing, and the model also had an accuracy of about 97% (Fig.3).

Through the testing and comparison of the above three models, the parameters were obtained (Tab. 1). Based on the table, we found that the best performing model was the newly optimized ConLSTM model, with an accuracy of 99.60%. Following closely are the ConvLSTM model and decision tree model, with the RF&LSTM model performing the worst, but it also has an accuracy of 92.23%. By combining the confusion matrix and ROC diagram, it was found that although the decision tree model also has high accuracy, it does not perform well in multi classification tasks. It is also evident that ConLSTM is superior to other models.

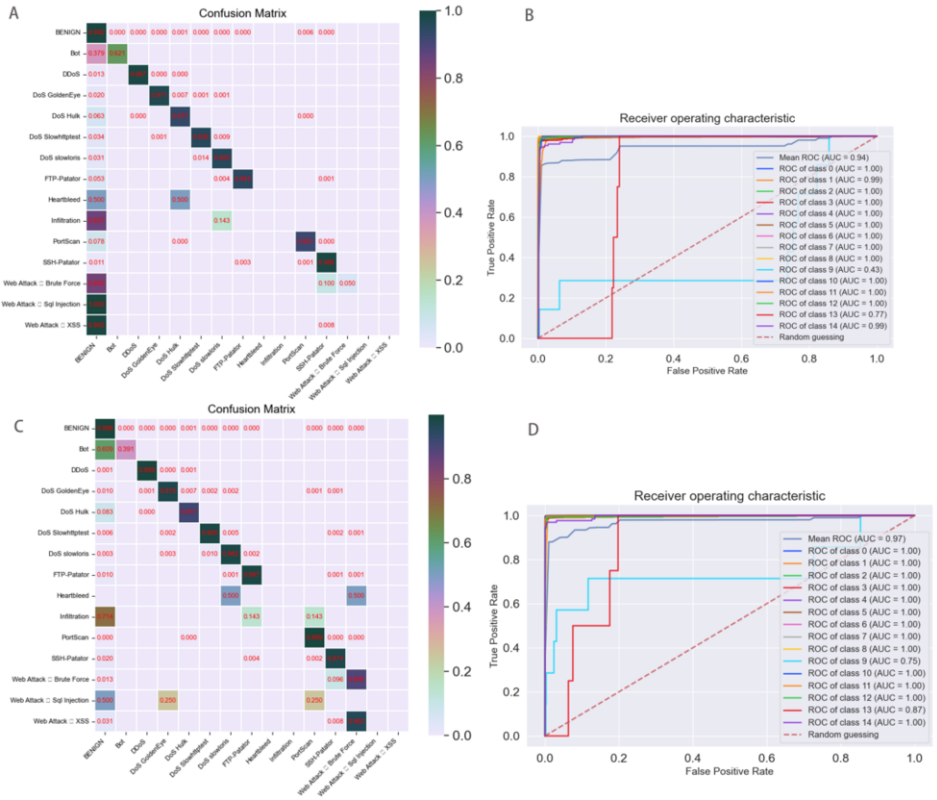

Figure 3. Confusion matrix and ROC results of ConvLSTM model after optimization

Table 1. Scores in models

	Accuracy	Precision	Recall	F1-score
Decision Tree	0.9954	0.9336	0.9461	0.9315
RF&LSTM	0.9223	0.7942	0.6063	0.6321
ConvLSTM	0.9789	0.9793	0.9789	0.9784
ConvLSTM&PCA	0.9960	0.9961	0.9960	0.9958

4. Discussion

By combining the confusion matrix and PR graph, it can be found that the RF&LSTM model is not a very good model for the CIC-ISDS-2017 dataset. It may be due to class imbalance in the dataset. In multi classification problems, because the sample size of certain categories in the dataset is much larger than that of other categories, the model is prone to bias towards predicting these majority classes, which also leads to although precision, recall F1 score and other evaluation indicators perform poorly, but the accuracy is still relatively high, and this phenomenon also exists in other models.

The reason for the high accuracy of the decision tree model is also the imbalanced dataset used. In this case, the decision tree tends to predict categories with a larger

proportion, resulting in high overall accuracy. In this case, we usually prefer to use other indicators to evaluate this model, so it is not a good model for this dataset [16].

The CIC-ISDS-2017 dataset showcases intricate spatiotemporal relationships. To better comprehend the long-term dependencies and trends in the time series data, we replaced the unidirectional LSTM layer with a bidirectional LSTM layer and introduced batch normalization and max pooling layers to augment the model's versatility and efficiency. Due to the incorporation of L2 regularization and dropout layers in the model, it possesses strong defense against overfitting, while the inclusion of dropout layers reduces neuron dependencies and co-adaptation, significantly enhancing the model's generalization and resilience. Moreover, the model capitalizes on the attention mechanism of the multi-head self-attention layer to prioritize essential features for precise classification. The integration of a learning rate decay strategy (StepLR) further enhances model generalization by ensuring stability and convergence towards optimal performance. Consequently, the model attains an impressive accuracy rate of 99.60%. Substantial enhancements in accuracy and robustness can be observed in Figure 3 as a testament to the optimization of the model.

For the ConvLSTM model, we can find that although convolutional neural networks already have some data dimensionality reduction methods in the convolutional layer pooling layer, adding new data dimensionality reduction operations appropriately can still improve the accuracy of the model again. After a series of optimizations, the new ConvLSTM can achieve higher accuracy, and this model has excellent fit resistance during training. During training, the train loss steadily decreases, and there is no fitting phenomenon near the lowest point, but it continues to remain at the lowest point.

During the training process, we also discovered the following issues: 1) When reading the entire dataset at once, due to the dataset being too large, it may cause insufficient memory and error reporting. So we performed a batch input operation on the dataset. 2) The dataset used for training also affects the model classification results, such as a balanced dataset containing a large amount of data often resulting in better training models. 3) The number of training sessions also has a significant impact on the model. A small number of training sessions can lead to inaccurate models, while an excessive number of training sessions can also lead to a decrease in model accuracy due to overfitting. 4) Finding suitable parameters after establishing a model often requires a lot of time, and the parameters of the model are often extremely important. 5) In addition, external dimensionality reduction tests were conducted on the dataset to reduce the number of features in the data, and it was found that simple machine learning methods can achieve accurate results, while deep learning models have greater complexity, although their accuracy is not low. However, its huge complexity results in lower efficiency compared to classical machine learning.

5. Conclusions

After optimizing and modifying the original model, we have obtained a new ConLSTM model with an accuracy rate of 99.60%, which has high accuracy and can effectively identify and classify network traffic. At the same time, we found that the model is not particularly accurate for some classifications, mainly because the CIC-ISD-2017 dataset we used is relatively unbalanced. For example, there are only 34 Infiltration type data and 11 Heartbleed type data. We need to train the model with more balanced data so that it can perform better and have better universality when facing various types

of data. Therefore, we can believe that this model has certain commercial value after being trained with appropriate training sets, and can be applied in industrial network traffic recognition, attack prediction, network condition detection, and other work. This model will also have certain practical value in other fields because there have been many modifications made in the model to improve universality.

Acknowledgements

This research was supported by the Natural Science Foundation of Shandong (No. ZR2020MF156), Youth Science Foundation Cultivation Support Program of Shandong First Medical University and Shandong Academy of Medical Sciences (No. 202201-036).

References

[1] E. Al Neyadi, S. Al Shehhi, A. Al Shehhi, et al. Discovering Public Wi-Fi Vulnerabilities Using Raspberry pi and Kali Linux. 12th Annual Undergraduate Research Conference on Applied Computing, Dubai, United Arab Emirates, 2020: 1-4.
[2] Li Jun, Z. Shunyi, Lu Yanqing et al. Internet Traffic Classification Using Machine Learning. Second International Conference on Communications and Networking in China, Shanghai, 2007: 239-243.
[3] Heaton Jeffrey, Ian Goodfellow, Yoshua Bengio et al. Deep learning, The MIT Press, 2016.
[4] Wang Xiao. A Survey on Heterogeneous Graph Embedding: Methods, Techniques, Applications and Sources. arXiv, 2020.14867.
[5] Ahmad Azab, Mahmoud Khasawneh, Saed Alrabaee, et al. Network traffic classification: Techniques, datasets, and challenges. Digital Communications and Networks, 2022:2352-8648.
[6] Jin Kim, Nara Shin, SY. Jo, et al. Method of intrusion detection using deep neural network. IEEE International Conference on Big Data and Smart Computing, Jeju, 2017: 313-316.
[7] J. Kim, HL Thi Thu, H. Kim. Long Short Term Memory Recurrent Neural Network Classifier for Intrusion Detection. International Conference on Platform Technology and Service, Jeju, 2016: 1-5.
[8] Wankhede Shreekh, Deepak Kshirsagar. DoS Attack Detection Using Machine Learning and Neural Network. Fourth International Conference on Computing Communication Control and Automation, Pune, 2018: 1-5.
[9] Khan MA, Karim MR, Kim Y. A Scalable and Hybrid Intrusion Detection System Based on the Convolutional-LSTM Network. Symmetry, 2019 (11): 583. https://doi.org/10.3390/sym11040583.
[10] Sharafaldin I, Lashkari AH, Ghorbani AA. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. International Conference on Information Systems Security & Privacy, 2018. DOI:10.5220/0006639801080116.
[11] Hunt EB, Marin J, Stone PJ. Experiments in induction. The American Journal of Psychology, 1966, 80(4). DOI:10.2307/1421207.
[12] Quinlan JR. Induction of decision trees. Mach Learn 1986(1): 81–106.
[13] Breiman. Random forests. MACH LEARN, 2001, 45(1):5-32.
[14] Xingjian Shi, Zhourong Chen, Hao Wang, et al. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. arXiv, 2015: 1506.04214v2.
[15] Qi Zhiguo. Stock Price Prediction Based on RF-LSTM Combined Model. China Applied Statistics, 2019(10): 181-188. (In Chinese)
[16] Batista Gustavo, Prati, Ronaldo, Monard Maria-Carolina. A Study of the Behavior of Several Methods for Balancing machine Learning Training Data. SIGKDD Explorations, 2004(6): 20-29.