Research on Network Traffic Anomaly Detection Method Based on Deep Learning

Chuwen Kuang*
Huizhou Economics and Polytechnic College, Huizhou, 516001, China
*Corresponding author’s e-mail: kuang@hzcollege.com

Abstract. Network attacks are a serious problem in today's information society. Intrusion detection systems are an important guarantee for network security. The core of IDS is the intrusion detection algorithm. Under the current new situation of network security protection, traditional signature-based misuse detection algorithms cannot solve the problem of rapidly changing forms of network attacks. Machine learning and data mining methods have been widely used in attack detection in the network environment to improve the detection rate. Based on the in-depth analysis and research of deep learning models and intrusion detection methods, this paper proposes a convolutional neural network abnormal traffic detection model based on dynamic adaptive pooling for the pooling layer can only use one fixed method for down sampling, and proposed an abnormal traffic detection framework and a dynamic adaptive CNN model. Experimental analysis shows that the method proposed in this paper shows better detection accuracy and loss value in traffic detection.

1. Introduction
The influence of the Internet on human society has penetrated into every aspect. Economically, the Internet economy is moving from the periphery to the core, becoming an important economic development force of the country; politically, it has expanded political resources and expanded the space for resource storage and utilization; culturally, human knowledge has exploded and is easy to obtain and carry, easy to spread, and human civilization has reached an unprecedented height [1]. With the continuous development of computer network technology and the rapid expansion of the scale of Internet users, security issues in the network are endless. Internet security threats such as spam, viruses, and spyware can lead to widespread identity theft and property fraud, causing serious economic losses to consumers and businesses [2]. Because most of the traditional intrusion detection systems rely on rule bases and traditional machine learning algorithms, the calculations are relatively complex and can no longer adapt to the new network environment. In recent years, deep learning related theoretical and practical results have emerged one after another, and amazing results have been achieved in the fields of speech recognition and image classification, which are suitable for processing large-scale data [3]. Using deep learning algorithms to solve the technical flaws of traditional intrusion detection has strong academic and practical value.

2. Intrusion detection classification and deep learning classification

2.1. Intrusion detection classification
According to the data source of intrusion detection system, intrusion detection system can be divided into 3 types:
Host-based intrusion detection system: By checking and analyzing each computer terminal system log and application log, tracking and understanding computer behavior activities, and checking whether there are violations, to identify various intrusion activities [4]. When analyzing the log, the system extracts key information from the computer log file and compares it with the established rule database to determine whether the computer has been attacked. The host-based intrusion detection system has high detection accuracy, but the relevant system must be deployed on each terminal, and the maintenance cost is relatively high [5].

Network-based intrusion detection system: uses computer network communication data packets as data sources, and uses network traffic analysis methods to discover and identify abnormal network behaviors. It is usually deployed behind a firewall and is regarded as the second line of defense for security.

Hybrid intrusion detection system: Combining the characteristics and advantages of the first two detection methods, data analysis from both host log information and network traffic information can not only discover and identify intrusions in the network, but also detect illegal activities through a large number of logs. It has a relatively comprehensive detection capability.

2.2. Classification of deep learning methods
Deep learning is a machine learning algorithm. The goal of deep learning methods is to learn feature levels, that is, multi-level learning based on data features or representations. High-level features are derived from low-level features to form a hierarchical representation. It uses a large number of multiple. The non-linear processing unit of the layer is used for feature extraction and conversion. Each continuous layer uses the output of the previous layer as input to learn multi-layer representations corresponding to different levels of abstraction [6]. The power of deep learning is that it can use a certain network in the middle of the network. The output of the layer is regarded as another expression of the data, so that it can be considered as a feature learned through the network. Deep learning architecture can be divided into three categories, as shown in Figure 1.

![Figure 1. Classification of deep learning methods](image)

3. Abnormal flow detection
Traditional machine learning methods for abnormal traffic detection require artificial design of a set of features, and there are human errors [7]. Although the commonly used neural network abnormal traffic
detection solves the problem of feature design, only one pooling method can be used for down sampling on the pooling layer. Aiming at the problems of common neural networks in abnormal traffic detection, a convolutional neural network abnormal traffic detection model based on dynamic adaptive pooling is constructed, and the pooling layer is improved, so that pooling can dynamically adjust the pool according to different feature maps of the process. In the data preprocessing stage, according to the four flow characteristic methods introduced in Chapter 2, the time series characteristic method is used to preprocess the original data. In response to the over-fitting problem of the deep learning model, a Dropout layer is connected after the fully connected layer to solve the over-fitting problem in the flow feature extraction process of the model and improve the model generalization ability [8].

3.1. Abnormal traffic detection framework
In this section, the DAPA-based convolutional neural network abnormal traffic detection framework is divided into two stages, as shown in Figure 2. The first stage is the data processing part, and the second part is the flow data detection part.

**Figure 2. Abnormal traffic detection framework**

The first stage: Data preprocessing, traffic capture, and the required data set is obtained. The data processing adopts the method of time-series characteristics to tailor, align, and supplement each request traffic in the data set to generate a series of 50*150 matrix data, we use this data as input [9].

The second stage: model construction, pool layer design, dropout layer design. The constructed dynamic adaptive pooling convolutional neural model is used to detect abnormal network traffic. The detection result classifies the data flow as normal or abnormal.

3.2. Dynamic adaptive pooling CNN model
The dynamic adaptive pooling convolutional neural network model is shown in Figure 3. In the process of abnormal traffic detection, the traffic data is converted into matrix data and used as the input of the model. Pooling is the second extraction of traffic characteristics [10]. It is necessary to summarize and calculate the traffic characteristics of the pooling domain, and add a Dropout layer after the fully connected layer.
In order to solve the over-fitting problem of the model in the process of abnormal traffic feature extraction, a Dropout layer is added to the model to solve the over-fitting problem. Over-fitting is specifically manifested in that the error rate on the training set is very low, but it is much higher on the test set. There are many methods to solve the over-fitting problem, such as regularization, pruning processing, early termination of iteration, dropout and other methods. Dropout is the most commonly used method to solve overfitting in deep learning, and it has good generalization ability and robustness. In the training process, all the neural network units in the fully connected process. When each neuron records the flow characteristics, it will cause the characteristic information to be over-recorded, resulting in only remembering the fixed characteristics on the training set, and the above phenomenon will occur. Dropout was proposed to solve this problem. In the Dropout layer, only the neurons in this layer will participate in the weight update, and the mutual influence between neurons will be weakened, which will effectively prevent the problem of fixed features from being learned and recorded.

During model training, the neurons in the fully connected layer are dynamically set to 0 with a probability of 40%. The neurons set to 0 can be considered as not participating in the weight update, and the neurons set to 0 each time the choice is not fixed. For each set of training, neurons are randomly selected. On the test data set, we also select 40%.

4. Analysis of results

The experimental environment is Windows 10 operating system, Python 3.7.3 language, Tensorflow framework, the computer CPU is Intel Core i5-4210u, the memory is 8GB, the hard disk space is 2TB, the experiment is about 61,000 matrix data, and the ratio of training set to test set is 9:1. Use the evaluation index accuracy rate and loss value to evaluate the detection result of the model, and the accuracy rate calculation is shown in formula (1). The higher the accuracy, the better the detection effect.

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)
\]

This experiment selects cross entropy as the loss function of this article. The cross entropy describes the distance between two probability distributions, and the function to calculate the cross entropy is shown in formula (2), \(q\) represents the probability distribution of the predicted value, and \(p\) represents the correct probability distribution.

\[
H(p, q) = - \sum_x p(x) \log(q(x)) \quad (2)
\]

The experiment uses three detection methods for detection comparison: the traditional convolutional neural network using the rule matching method and the maximum two-mean pooling algorithm (MTPA) is compared with the DAPA-based convolutional neural network method proposed in this article. The experimental comparison results are as follows as shown in Figure 4, it can be seen from Figure 4 that the accuracy of the convolutional neural network using DAPA is the highest among the three detection methods, and the loss value is the lowest.
Figure 4. Accuracy and loss value

When the proportion of neurons is 100% (that is, the Dropout technology is not added), the loss value is the highest. During model training, the model will overtrain the training data set in order to reduce the loss value, resulting in overfitting; after adding Dropout technology, when the proportion of neurons is 40%, the loss value is reduced by 4.2%, and the accuracy is increased by 6%. Therefore, this article chooses to set the parameter to 0.4 to improve the overfitting of the model. In this paper, experiments with different iteration times are compared, and the experimental results are shown in Figure 5.

Figure 5. Experimental accuracy and loss value

The model generates different small networks in the process of different iterations, and iteratively trains each algorithm. Each time the model increases the number of iterations, the combination of neurons will change so that the neurons will not be composed. The fixed combination affects the model learning process and makes the parameters evenly distributed.
5. Conclusion
With the development of science and technology, people's lives are inseparable from the Internet. However, according to the latest Internet security report, it can be analyzed that the security of today's network environment is getting lower and lower, and threats are becoming more and more complex and diverse, which have affected people's normal lives. As a supplement to traditional security defense technology, intrusion detection can actively protect the system and play a vital role in information security protection. In recent years, the amount of network data has increased dramatically, and network attacks have become more complex and diverse. This article first introduces the current network environment, and gives a detailed overview of the basic concepts, basic principles and classification methods in the field of intrusion detection; at the same time, deep learning is analyzed in depth. After completing the basic work, an intrusion detection model based on deep structure is proposed. The research in this article has achieved certain results and has certain research value.

References
[1] Song Y, Hou B N, Cai Z P. 2021 Network intrusion detection method based on deep learning feature extraction. Journal of Huazhong University of Science and Technology (Natural Science Edition), 49(02):115-120.
[2] Xiao K, Liu T Y, Sun X Y, et al. 2020 Intrusion detection method based on incremental convolutional neural network. Computer Applications, 40(S2):73-79.
[3] Zhang X L, Cheng G, Zhang W C. 2021 Network traffic classification method based on improved deep convolutional neural network. Science in China: Information Science, 51(01):56-74.
[4] Hu Y J, Guo Y B, Ma J. et al. 2020 Network deception traffic generation method based on adversarial samples. Journal on Communications, 41(09):59-70.
[5] Chen L C. 2020 Research on Traffic Data Reduction in High-speed Network Intrusion Detection. Modern Electronic Technology, 43(16):108-111.
[6] Wang S W, Zhang L J, Jia Z, et al. 2020 Android malware identification based on network traffic. Radio Engineering, 50(07):612-618.
[7] Xu Z H. 2020 Research on network traffic detection based on CNN-SVM. Information and Computer (Theoretical Edition), 32(10):42-44.
[8] Li X J, Xie X Y, Xu Y. 2020 Anomaly Detection Method of Network Traffic: SSAE-IWELM-AdaBoost. Journal of Wuhan University (Science Edition), 66(02):126-134.
[9] Zhao X. 2020 Application of Convolutional Neural Network in Intrusion Detection. Information and Computer (Theoretical Edition), 32(07):23-25.
[10] Huang X L, Li C M, Jiang Q S. 2020 Automatic extraction of spatiotemporal features of network streams based on deep learning. Integration Technology, 9(02):60-69.