Research on Network Security Application Based on Deep Learning

Zhao Jianchao
School of Electronic Information Engineering, Henan Polytechnic Institute, Nanyang, Henan China

Abstract

Behind the rapid development of the Internet industry, Internet security has become a hidden danger. In recent years, the outstanding performance of deep learning in classification and behavior prediction based on massive data makes people begin to study how to use deep learning technology. Therefore, this paper attempts to apply deep learning to intrusion detection to learn and classify network attacks. Aiming at the nsl-kdd data set, this paper first uses the traditional classification methods and several different deep learning algorithms for learning classification. This paper deeply analyzes the correlation among data sets, algorithm characteristics and experimental classification results, and finds out the deep learning algorithm which is relatively good at. Then, a normalized coding algorithm is proposed. The experimental results show that the algorithm can improve the detection accuracy and reduce the false alarm rate.

Keywords*: Deep learning, massive data, behavior prediction, normalized coding algorithm.*

I. Introduction

With the deepening integration of digital, network technology and all walks of life, there are more and more network security problems [1-2]. While emerging digital information technologies such as virtualization, big data, next generation communication network, artificial intelligence and blockchain bring us convenience, new security risks also increase [3]. In the face of successive major data leakage and other security emergencies, we need to strengthen the theoretical innovation research of network security, and build an intelligent, interconnected and all-round security network [4].

Last March, the China Cyberspace Security Association was established in Beijing. The establishment of China Cyberspace Security Association is of great significance for our country to carry out cyberspace security strategic services. At the same time, it also conforms to the situation of network security and information development. It can not only promote social groups to widely participate in the construction of China's network security, but also play a positive and constructive role in promoting the healthy development of China's cyberspace [5]. On June 1, 2017, China's first normative law on Cyberspace Security, the Cyberspace Security Law of the people's Republic of China, was officially promulgated and implemented [6-7]. This is an important milestone in the construction of the rule of law in cyberspace in China. It is an important legal tool for managing the Internet according to law and resolving network risks, and provides an important guarantee for the healthy operation of the Internet on the track of the rule of law.

Digitization is an important trend in today's world, and its advantages are beyond doubt: intelligent communication, intelligent measurement and other digital technologies have greatly improved the efficiency and reliability of the energy system [8]. But digitization also opens the door for the emergence of network security threats. Nowadays, digital equipment has penetrated into every field and link of the energy industry, especially the wide application of industrial automatic control system, and has gradually become the control center and core of the energy system, which may bring huge potential security risks. Once hackers invade and control these devices, in theory, they can control the energy system and "do whatever they want", such as opening and closing all kinds of switches and valves at will, changing the operation state of the equipment, adjusting the setting of the early warning system, etc., thus leading to the interruption of energy supply or physical damage such as explosion and

ISSN: 0010-8189

fire.

In recent years, with the continuous improvement of relevant laws and regulations and management system in the field of network security, China's network security research ability and talent team construction level have been significantly improved, and remarkable results have been achieved in international cooperation. At the same time, China has clearly put forward the strategy of network power, but the network security threats from the Internet black industry chain and the national network security confrontation make the cyberspace security situation facing China increasingly complex and hidden. Therefore, we must strengthen the research in the field of network security, in order to deal with the rapid change of network security situation in the Internet era.

II. Overview of related technology research and platform

2.1 Deep learning method

Figure 1 shows the popularity trend of the word "deep learning" in Google search in the last decade.

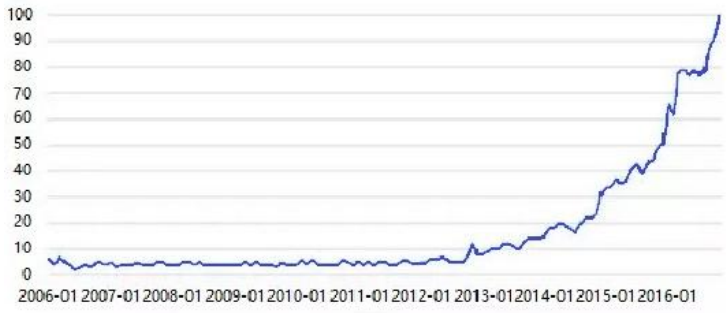

Fig 1: The popularity trend of "deep learning" in Google search in recent ten years

As can be seen from the figure, the popularity of deep learning has increased exponentially since 2012. By 2016, deep learning has become the most popular search term on Google. Deep learning is not a new word, it is basically synonymous with deep neural network. Inspired by the structure of human brain, the computational model of neural network was first proposed in 1943. After the invention of perceptron, neural network becomes a model that can "learn" from data. But because the structure of the sensor network is too simple, it can not solve the problem of linear indivisibility. In addition, the amount of computation required by the neural network is too large, the computer at that time can not meet the needs of computation, making the research of neural network into the first winter. By the 1980s, the deep neural network and back propagation algorithm have solved these problems well, and let neural network enter the second period of rapid development.

In the early 1980s, network model emerged in cognitive psychology. The connectionist psychology is inspired by the animal brain nerve, and regards the cognition of things as an activity of neural network. Connectionism attaches great importance to the importance of network, and emphasizes the distributed feature representation and parallel processing methods. There are several main theoretical contributions in connectionism, such as distributed representation and reverse propagation, which still play a very important role in the field of deep learning.

In 2006, Hinton et al. Put forward unsupervised greedy layer by layer training algorithm to solve the optimization problems related to deep structure, and the concept of deep learning was proposed [9]. The CIFAR affiliate team said the same strategy could be used to train other types of deep networks and help systematically improve generalization on test samples. In addition, Lecun established the first "real" convolutional neural network model with multi-layer structure in Bell laboratory [10]. With the rapid development of Internet industry, the computer

software and hardware infrastructure for deep learning has been improved, the amount of available training data is increasing, and the scale of deep learning model also increases.

2.2 TensorFlow deep learning framework

TensorFlow is a distributed machine learning platform, and its main architecture is shown in Figure 2. RPC and RDMA are network layers, which are mainly responsible for transmitting neural network algorithm parameters. CPU and GPU are the device layers, which are mainly responsible for the specific operation in neural network algorithm. Kernel is the concrete implementation of algorithm operation in TensorFlow, such as convolution operation and activation operation. Distributed Master is used to build subgraph and cut subgraph into multiple slices. Different subgraph slices run on different devices. Master is responsible for distributing subgraph slices to the Executor/Work. Executor/Work is on the equipment (CPUs, GPUs, etc.), and is responsible for sending and receiving the running results of graph operation to other Worker. C API divides TensorFlow into front end and back end, and the front end (Python/C++/Java Client) triggers TensorFlow back end program to run based on C API. Training libraries and Inference libs are library functions for model training and derivation, which are used by users to develop application models.

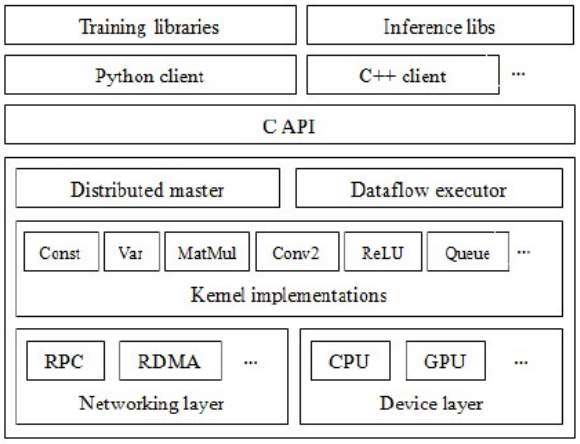

Fig 2: **TensorFlow system architecture**

III. Deep learning algorithm and experimental analysis of intrusion detection

3.1 Learning algorithm

1) Several common classification algorithms

Classification is a kind of data mining. Classification is to predict the value of a specific attribute according to the value of other attributes. The value of a specific attribute determines that it belongs to one of several categories. In other words, classification belongs to the prediction task, which is to get an objective function f through the learning of existing data sets, and map each attribute set x to the objective attribute y, and y must be discrete. It is difficult to give the rules of classification algorithm directly by programming, but it is easy to get them by learning algorithm. The classification process first needs to process the actual data into data that can be understood by computer (data preprocessing), generally in the form of table. If there are too many features in the learning data, we may need to select the most representative features from the feature set. Feature selection can reduce the training time, improve the performance of the learning algorithm, and avoid the dimension disaster problem.

Classification algorithms are based on mathematics, through a variety of methods to analyze the data and make predictions. Here are several common classification algorithms:

(1) Naive Bayesian model (NB)

Bayesian classification is a classification method based on Bayesian theorem and independent assumption of feature conditions. It classifies by calculating the probability that a given tuple belongs to a specific class. The advantage of this method is that it is not sensitive to missing data and needs less parameters to be estimated. The disadvantage is that all attributes are required to be independent of each other (this requirement is difficult to meet in practical problems), and need to know the prior probability.

(2) Decision tree model (DT)

Decision tree is a widely used classifier. It uses training data to construct decision tree, and then realizes data classification through decision tree. The decision tree has a tree structure. Each internal node is used to test an attribute. The branch of each node represents the test result. Each leaf node of the tree represents a classification category. The decision tree starts from the root node, tests step by step, selects step by step, and gets the classification results when it reaches the leaf node. The advantage of this method is that it doesn't need any domain knowledge or parameter assumption, and is suitable for high-dimensional data, processing a large number of data in a short time, and can get better results. The disadvantage is that it is easy to over fit and does not support online learning. Some data have different number of samples for each category, and the information gain tends to those features with more values.

(3) K-nearest neighbor algorithm (KNN)

KNN is one of the simplest classification algorithms. The idea of the algorithm is to find the nearest K samples in the sample space. If most of the K samples belong to a certain class, then the samples also belong to this class. The algorithm is simple, easy to understand, easy to implement, without training. The disadvantage is that when the samples are unbalanced (the number of samples belonging to some classes is much more than other classes), it is easy to cause judgment bias.

2) Softmax regression learning method

Softmax regression, namely multiple Logistic regression, is a commonly used multi-class classifier. In this summary, we will learn and classify intrusion detection data based on Tensorflow using Softmax regression method, and analyze its classification effect. Softmax function is a normalized exponential function and is defined as follows:

$$y_c = \varphi(Z)_c = \frac{e^x c}{\sum_{d-1}^{c} e^z d} \quad (1)$$

Among them, φ Represents the softmax function. The input z is a C-dimensional vector and the output y is also a c-dimensional vector. The denominator in the formula acts as a regular term, which makes:

$$\sum_{j=1}^{C} y_j = 1 \quad (2)$$

Tensorflow provides an embedded softmax implementation function. In order to construct the softmax classification model, it is necessary to establish the full connection between the input vector and the output category, and train the weight of each connection and the bias vector of the classification. For this reason, we need to define the weight matrix and bias term vector and give the initial value. Figure 3 is a python program that defines the weight matrix and the offset term vector.

```
import tensorflow as tf
w = tf.Variable(tf.zeros([41, 23]))
b = tf.Variable(tf.zeros([23]))
```

Fig 3: Definition of weights and offsets

3.2 Feature learning algorithm based on depth structure

Convolutional neural network is a deep learning structure inspired by visual perception mechanism. It uses multi-layer network model to extract the features of things, and then classifies, recognizes, predicts or makes decisions according to the features. It has a wide range of applications, including image classification, target detection, target recognition, target tracking, text detection and recognition, position estimation and so on. The basic structure of CNN generally includes input layer, convolution layer, pooling layer and full connection layer. Each node of convolution layer is connected with a region of the upper layer by convolution kernel. In the same convolution layer, the weights of all neurons are the same. The pooling layer is sandwiched in the middle of the convolution layer, and its main function is to gradually compress, reduce the number of data and parameters, and also reduce the over fitting phenomenon to a certain extent, and compress a certain area of the input data of the upper layer into a value. The full join layer is mainly used for learning, mapping the learned feature representation to the sample label space. Based on the analysis of CNN and test data set, we designed CNN training model, as shown in Figure 4.

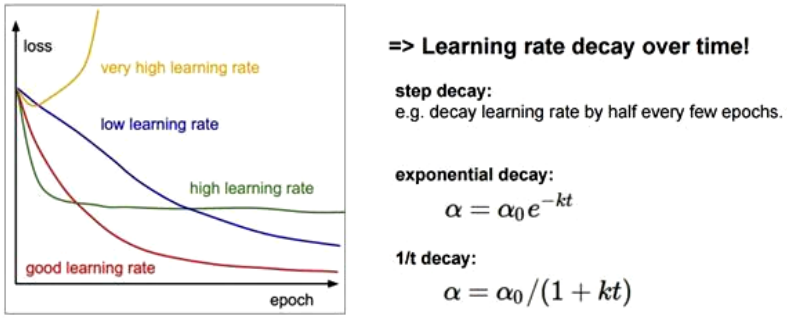

Fig 4: CNN training model

In many deep learning network models, sparse self encoder is one of the effective algorithms for feature extraction. The sparse self coding model considers that the input data can be transformed into a weighted representation of a group of bases. For example, the integer can be expressed as a weighted representation of the group of bases t = [number, ten, hundred, thousand, ten thousand, one hundred thousand...], that is, any integer k can be expressed as:

$$k = \sum x_i \times T_i \quad (3)$$

Where x is the weighted vector.

SAE, a neural network with multiple hidden layers, is an unsupervised learning method, which uses back propagation algorithm. The idea is to make the output equal to the input, and let the encoder find the hidden features (that is, the set of bases) in the input data. SAE is generally divided into coding process and decoding process. The decoding process is the reverse process of coding process, but it does not require that the decoding weight is the same as the coding weight, but is approaching through learning. The encoding process is to find t and express the input k as ∑xi × Ti, while the decoding process is to express ∑xi × Ti as k again. Learn from the errors of output and input in every encoding and decoding process.

SAE requires that the output is equal to the input, which is different from the requirement of intrusion detection (the input of intrusion detection is 41-dimensional network data, and the output is 23 categories represented by 23-dimensional vector). Therefore, we changed the last decoding of SAE model into mapping to 23-dimensional vector, compared the mapping result with the actual classification result, and used its error to learn. In this way, unsupervised learning methods have become supervised learning methods.

IV. Detection model based on normalized coding algorithm

4.1 Normalized coding model

The problem of sparse self encoder is different from the intrusion detection classification. First, SAE encodes first and then decodes. It pursues the consistency of input and output. The input of intrusion detection classification is 41 dimensional vector, and the output is different classification represented by 23 dimensional vector. Secondly, SAE is an unsupervised learning network, nsl-kdd has clear classification label, which is more suitable for supervised learning. Finally, the output of SAE is a decoding result. As the learning process and the target result (in fact, the input) continue to approach, this input is far from the vector that only contains one 1 and all other zeros. But even with so many problems, SAE is still better than softmax and CNN in intrusion detection classification. To solve the above problems, we optimize the learning process. A normalized coding model is designed, as shown in Figure 5.

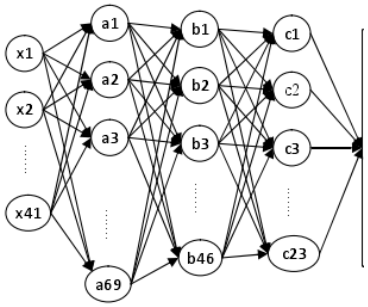

Fig 5: Normalized coding model (NSAE)

NSAE keeps the basic structure of SAE model based on base coding, and makes targeted optimization for the needs of intrusion detection classification. Firstly, the coding process is reserved and the decoding process is removed. Because the final output is a 23-dimensional vector, we use three hidden layers, with the number of nodes being 69, 46 and 23 respectively. Experiments show that the hidden layer with nodes with integral multiple of the target dimension is slightly better in feature extraction. In the learning process, the three hidden layers extract feature information from the input data, and gradually approach to 23 bases. Secondly, a normalization layer is added, and Softmax algorithm is used to realize the mapping from 23 weighted bases to 23-dimensional classification results. We expect that this mapping can find some connection between bases and classification. Finally, the supervised learning method is adopted, and the gap between the classification information obtained by learning and the actual classification information is used for regression learning.

4.2 Realization of normalized coding model

The realization of this model needs the support of three libraries: openpyxl, tensorflow and numpy. openpyxl provides an operating interface to excel files, because both training data sets and test data sets exist in the form of excel files. Tensorflow provides the realization of some important functions of the model and the support of model running, and these interfaces include matrix operation, activation function, optimization function and so on. Numpy provides some functions needed for scientific calculation.

The model parameters include the definition of system parameters, learning network parameters and file names. There are five definitions of system parameters, namely, learning speed (learning_step), model training (batch_size), total batch times (batch_n) and total test data (test_n). There are four parameters of learning network, which are the dimension of input data vector (input_n), the first hidden layer node number (encoder_1_n), the second hidden layer node number (encoder_2_n), the third hidden layer node number and the dimension of output vector (output_n). There are two definitions of file names, namely, trainingfilename and testfilename.

Tensorflow requires graphs to define calculation, that is, learning model, so we need to define NSAE model before

running the model in Tensorflow session (that is, training the model). In the input part, the placeholder function placeholder is used to create two placeholders for x and y_true, which are used to feed the input feature data and the real classification results. X is a two-dimensional matrix, the second dimension is 41, and the first dimension is uncertain, which is determined according to the number of training data input in each batch. Y_true is also a two-dimensional matrix, the second dimension is 23, and the first dimension is the number of input training data.

The weight matrix dictionary defines a dictionary variable weights, which includes three members encoder_1_w, encoder_2_w and encoder_3_w .. They are the weight matrices of three hidden layers (coding layers). The weight matrix sizes are 41*69, 69*46 and 46*23, respectively, and all of them are initialized with normal random numbers. A dictionary variable bias is defined in the biase vector dictionary, which includes three members encoder_1_b, encoder_2_b and encoder_3_b . They are the bias term vectors used by the three hidden layers, and all the three vectors are initialized with normal random numbers.

The encoder defines the function encoder, which takes the input x as the parameter. The definitions of the three hidden layers are similar, and all of them are Sigmoid activation functions. The input (the input of the later hidden layer is the output of the previous hidden layer) is added to the inner product of the weight matrix of this layer and the offset term vector, and then processed by the activation function as the output of this layer.

V. Conclusion

In recent years, with the rapid development of Internet technology, network security has become an important issue that must be paid attention to in all fields, especially in the fields of industrial control, intelligent technology, mobile payment and cloud computing. At the same time, hackers and network terrorist organizations and other groups launched a variety of network attacks are more and more influential and destructive, China's network security situation is more and more severe. With the maturity and popularization of deep learning technology, especially the excellent performance in the field of feature learning, we have found a breakthrough. In this paper, the research of deep learning oriented to network security detection is taken as the exploration direction, focusing on the intrusion detection algorithm based on deep learning, the traditional classification detection algorithm is analyzed and improved, and its implementation in tensorflow machine learning platform is explored.

Acknowledgements

This research was supported byResearch and Practice Project of Higher Education Teaching Reform in Henan Province (Grant No. 2019-701).

References

[1] Jiang Wei, Fang Binxing, Tian Zhihong. Network Security Evaluation and Optimal Active Defense Based on Attack Defense Game Model. Acta Computer Sinica, 2009, 32 (004): 817-827
[2] Miao Yongqing. Stochastic Model Method and Evaluation Technology of Network Security. China Science and Technology Investment, 2017, 4: 314
[3] Yi Hua Zhou, Wei Min Shi, Wei Ma. Research on Computer Network Security Teaching Mode for Postgraduates Under the Background of New Engineering. Innovation and Practice of Teaching Methods, 2020, 3 (14): 169
[4] Bao Xiuguo, Hu Mingzeng, Zhang Hongli. Two Quantitative Analysis Methods for Survivability of Network Security Management Systems. Acta Communication Sinica, 2004, 25 (9): 34-41
[5] Yang Yi, Bian Yuan, Zhang Tianqiao. Network Security Situation Awareness Based on Machine Learning. Computer Science and Application, 2020, 10 (12): 8
[6] Li Zhiyong. Hierarchical Network Security Threat Situation Quantitative Assessment Method. Communication World, 2016, 23: 70-70
[7] Hu Wenji, Xu Mingwei. Analysis of Secure Routing Protocols for Wireless Sensor Networks. Journal of Beijing University of Posts and Telecommunications, 2006, 29 (s1): 107-111
[8] Wei Yonglian, Yi Feng, Feng Dengguo, Yong W, Yifeng L. Network Security Situation Assessment

Model Based on Information Fusion. Computer Research and Development, 2009, 46 (3): 353-362

[9] Xu Guoguang, Li Tao, Wang Yifeng. A Network Security Real-time Risk Detection Method Based on Artificial Immune. Computer Engineering, 2005,31 (12): 945-949

[10] Li Weiming, Lei Jie, Dong Jing. an Optimized Real-time Network Security Risk Quantification Method. Acta Computa Sinica, 2009 (04): 793-804