A Deep Learning Algorithm Using Feature Engineering to Adjust Attention Mechanisms and Neural Network for Cloud Security Detection

Yiyang Xiong(✉), Yajuan Qiao, Shilei Dong, Xuezhi Zhang, and Hua Tan

China Telecom Corporation Limited Research Institute, Beijing, China
xiongyy1@chinatelecom.cn

Abstract. Cloud computing realizes the intensive management of resources and improves the production efficiency, but it also inevitably brings security problems. Cloud security detection technology can be understood as immune cells in the human body, which can protect against all kinds of viruses invading the server. In this article, we transform the virus activity information into time series data, and study how to classify the virus through deep learning algorithm. In our system, we roughly divide the virus activity information into six categories, and calculate the importance of these six types of behavior to virus recognition through feature engineering. Then, we integrate the calculated parameters into the construction of attention mechanism layer and neural network embedding layer to propose a new algorithm based on deep learning architecture. Finally, we verify the accuracy of the new algorithm in virus classification through open source dataset. We compare the performance of this model with logistic regression, support vector machine, random forest model and convolutional neural network. The experimental results show that our model has certain advantages in F1, and improves the performance index by nearly 4%.

Keywords: Deep learning · Feature engineering · Attention mechanisms · Cloud security detection

1 Introduction

As a new type of information technology infrastructure, cloud computing is gradually becoming the main driving force for the intelligent transformation of large enterprises [1]. It can help users realize intensive management and improve the intelligent level of enterprises [2]. However, although cloud computing can bring efficiency improvement and cost reduction to enterprise customers, the migration of sensitive information such as enterprise data to the cloud will bring new security risks [3]. Therefore, whether it is safe enough has become the core issue for enterprises to choose cloud services.

Due to the rapid development of cloud computing, the potential attack threat of cloud platform with more and more important personal and enterprise information is gradually

Y. Wang et al. (Eds.): ICSINC 2022, LNEE 996, pp. 11–18, 2023.
https://doi.org/10.1007/978-981-19-9968-0_2

increasing [4]. On the other hand, the cloud uses a lot of virtualization technology, so the coverage of an attack will be particularly wide. Common attacks include DDoS attacks, backdoor programs, Trojan horses, worms and mining programs that have gradually increased with the development of blockchain in recent years. These viruses will steal computer computing resources or user information and bring irreparable damage.

Whether traditional viruses or new viruses, machine learning can achieve good results in identifying their attacks [5]. It can learn its behavior pattern from the activity information of the virus and distinguish it. However, machine learning method belongs to shallow learning, which is very dependent on Feature Engineering, such as feature construction, feature selection and so on. The large-scale virus intrusion scenario in the real network environment is a multi-classification problem with dynamic growth of data sets, so its accuracy is difficult to be guaranteed. Compared with it, deep learning can extract deeper features of the virus through data, so it performs better [6]. In addition, because the virus activity information can be converted into an API call sequence, it is very suitable for the deep learning model that can extract timing information.

On the other hand, after the initial brilliance of deep learning, new algorithms are often proposed around the improvement of attention mechanism [7]. The core idea of attention mechanism is to give different weights to different data, which can improve the performance of the model. Therefore, we believe that there may be some combination between the feature engineering of virus and the attention mechanism of neural network.

Based on this idea, our algorithm framework first analyzes the types of API called, and roughly divides these APIs into six categories: service, registration, operation process, system information acquisition, network communication and file operation. Then, we calculate the importance of different kinds of behavior to virus discrimination through feature engineering, and integrate it into the model construction through the attention mechanism and embedding of neural network. Our model has achieved better results on these three types of data than other models including deep learning algorithms.

2 Model Architecture

2.1 Overview

The overview of the algorithm we proposed is as follows: first, convert the API call information of the virus into text information, and then conduct word embedding coding to convert it into a vector that can be understood by the computer. In the process of word embedding coding, we innovatively give different weights according to the types of API calls and encode them into the word embedding module. Second, after obtaining the word vector, we carry out piecewise convolution, pooling and splicing of the word vector. Then, we extract the features through the random forest model, and obtain the importance of each API call information to virus classification, so as to creatively construct the parameters of attention mechanism. Finally, we send the information adjusted by the attention mechanism to the full connection layer and get the output through SoftMax. After training and adjusting parameters, the model is obtained. Our model can be simply divided into neural network module and attention mechanism module.

2.2 Neural Network

Our neural network consists of three parts: word embedding, convolution, pooling and splicing (see Fig. 1).

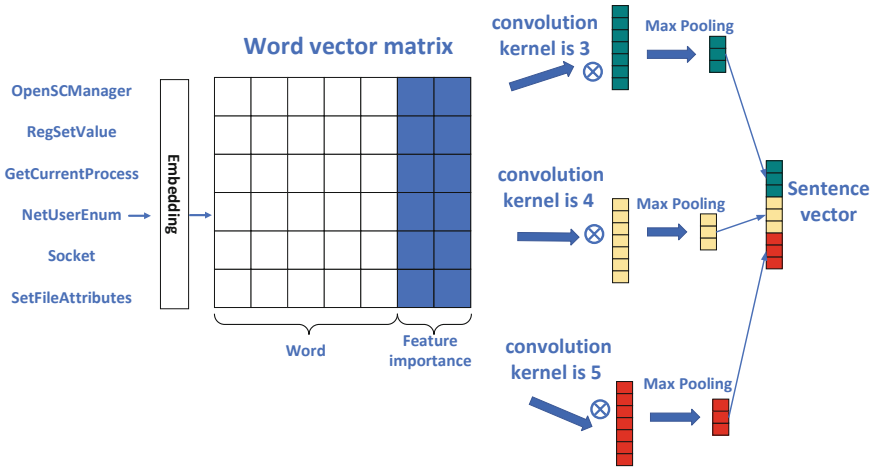

Fig. 1. The structure of the neural network consists of three parts: word embedding, convolution, pooling and splicing. Among them, the blue part of the word vector matrix is the coding of virus importance, which is one of the core innovations of this paper.

Word Embedding. The input of neural network is the API call information of virus. It exists in the form of characters, and the computer cannot understand the connotation of characters. Therefore, we must first convert it into word vector through word embedding. Each word vector can be trained in other corpora in advance, or it can be trained by the network as unknown parameters. Because our text information comes from virus activity information, there is almost no corpus for this kind of problem in academic circles, and there is no available pre training model. Therefore, instead of using the pre training model, we set the random initial value and change it completely through network training. Figure 1 illustrates this process with an example. Each API call of the virus will become a row of the word embedding matrix.

Due to the lack of corresponding corpus, the convergence speed of neural network in the early stage will be very slow. Therefore, we explore how to improve the word embedding matrix in the experiment, and try to add some pre training information to make the model understand the current task faster. We innovatively added the importance information of virus into the word vector through feature engineering. Firstly, we divide the virus API into six categories according to expert experience. Then, we and calculate the importance of virus type through random forest model encode it into the word vector.

Convolution. In order to obtain the information of word embedding matrix from different angles, we use three kinds of convolution kernels to convolute the matrix respectively. Since each row of the word embedding matrix represents the single active information

of the virus, the width of our convolution kernel is consistent with that of the matrix, which ensures that the neural network learns in the unit of virus call in a physical sense.

Pooling and Splicing. Finally, we pool different convolution results, and then splice them into lower dimensional vectors.

2.3 Attention Mechanism

The attention mechanism of deep learning can continuously calculate the semantic code according to the current new vector, which can input different semantic codes at each time. Attention mechanism not only helps neural network solve the problem of word information loss, but also pays more attention to the "key information" in word vector.

Our innovation is based on this "key information". From the point of view of our concerns, although many APIs can be called, we should be able to understand that the calls of some APIs are normal, while the frequent calls of some APIs are very suspicious (such as registration information modification). These suspicious behaviors should have greater attention in neural networks. Figure 2 shows our attention mechanism module framework.

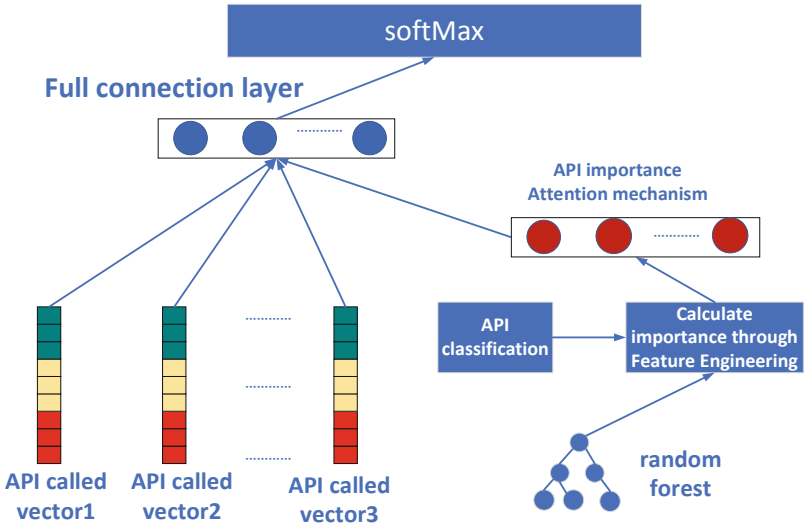

Fig. 2. Our attention frame is sentence level (about 100 activity information performs one attention). Its parameters are calculated by characteristic engineering.

Feature Engineering. Firstly, we divide APIs into six categories: service, registration, operation process, system information acquisition, network communication and file operation. In this way, each API information has a label. Then, we convert all API virus called into six types of labels by setting key value pairs. Finally, we use the random

forest algorithm LightGBM to calculate the importance of these six types of tags for virus classification according to the existing data, and get the value of the importance parameter.

Attention Calculation. The calculation of our attention mechanism c_i is obtained by the weighted sum of hidden vectors h_i:

$$c_i = \sum_{j=1}^{T_x} a_{ij} h_j \tag{1}$$

Among them, the weight corresponding to h_i can be calculated by the following formula:

$$a_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \tag{2}$$

The e_{ij} of formula 2 is obtained by multiplying our API importance matrix W and h_j:

$$e_{ij} = h'_{i-1} W_{i-1} h_j \tag{3}$$

By multiplying the importance matrix calculated by feature engineering, our attention mechanism can give bigger weight to the API with bigger impact. Our new algorithm reduces information loss and focuses on more important information for sequence prediction.

Full Connection and SoftMax. Finally, we send the sentence vector adjusted by attention mechanism into the full connection layer for training, and use SoftMax for normalized output to facilitate the convergence of the model during training.

3 Experiment

3.1 Data Preprocessing

We used Alibaba Tianchi's open source cloud security detection data set in the experiment. The data set is composed of windows executable programs. Its sample data are all from the Internet. The types of malicious files are traditional infectious viruses, DDoS and Trojan viruses. In addition, it also includes mining viruses and extortion viruses that have emerged in recent years.

There are 295 kinds of APIs called by viruses in the data. Ideally, we give different weights to all APIs, but this will lead to too complex calculation. To simplify the problem, we first roughly divide the 295 APIs into six categories. As shown in Table 1. We give 2–3 examples for each classification.

After dividing the APIs into six categories, in order to make the neural network better understand the differences between different APIs, we use LightGBM random forest algorithm to calculate the importance of these six categories of APIs in virus classification through python, and finally get the results shown in Fig. 3.We can see that registration, operation process and file operation are the categories that have a better impact on virus classification. We record the calculated importance as the label of each API, which lays a foundation for the adjustment of neural network attention mechanism.

Table 1. Six different categories and their representative APIs.

Behavior type	API called	Purpose of API
Service behavior	OpenService	Open service
Registration behavior	RegCreateKey	Create or open registry
	RegSetValue	Modify registry
	QueryServiceConfig	Query system service information
Operation process behavior	CreateProcess	Create process
	LoadLibrary	Load dynamic database
System information acquisition behavior	GetUserName	Get the name of the current user
	GetSystemInfo	Get information about hardware platform
	GetHostByName	Obtain IP address through domain name
Network communication behavior	Send	Send data
	WNetAddConnection	Create permanent connection
	Socket	Create network communication socket
File operation behavior	CreatFile	Creat a file
	SetFileAttributes	Modify file properties

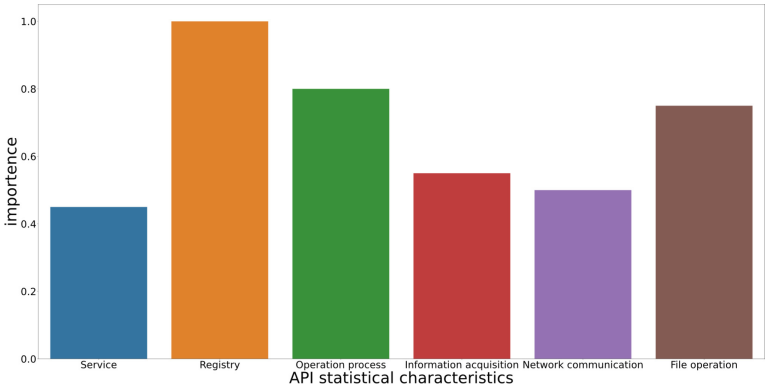

Fig. 3. Normalized importance of API for virus classification.

3.2 Test and Result

We interpret the API call information labeled by each virus as text and convert it into a word vector. Then we add different coding values to each API according to the divided categories. Then, we perform piecewise convolution and splicing, and construct sentence level attention mechanism in units of 100. The parameters of attention mechanism are

completely obtained from the API importance obtained in the preprocessing stage. In this way, the importance features obtained by random forest processing affect the processing of neural network from two angles: 1 Word embedding of piecewise convolution neural network. 2. Construct attention mechanism from sentence dimension.

We have divided Alibaba's data set by five-fold cross Validation. It means that we use four parts of them as training each time, and the next part as testing, and then, we average the precision and recall obtained from the four experiments to obtain the experimental results.

We then used the same method to compare the innovative model with neural network without adjusting attention mechanism, ordinary piecewise convolution network, traditional random forest model, the result is shown in Fig. 4. The results show that our model achieves better performance than others, and we guess that because deep learning algorithm can capture the timing information of API activities, it performs significantly better than random forest.

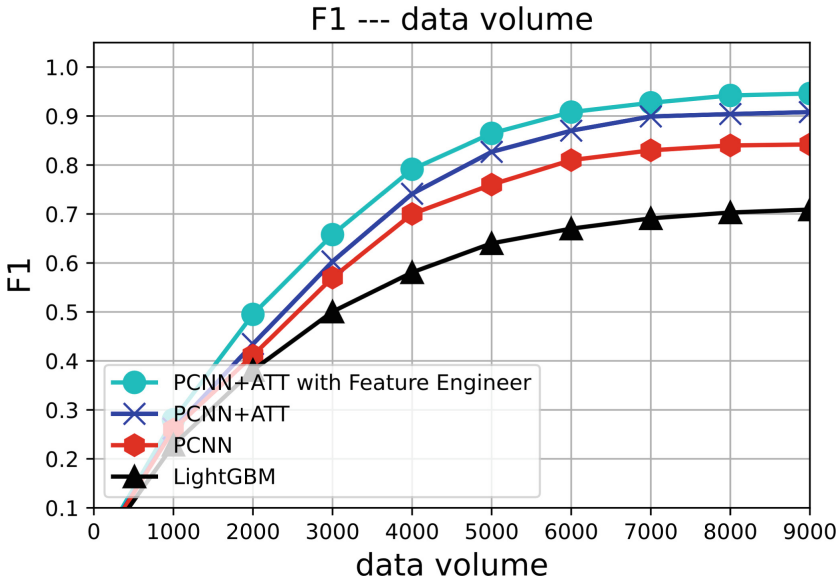

Fig. 4. Performance comparison of four different models on F1. It shows the comparison of our model with PCNN (Piecewise convolutional neural network), ATT (attention mechanism) and light BGM (Light Gradient Boosting Machine)

4 Conclusion

This paper creatively calculates the importance of virus API activity information to virus classification through feature engineering, and integrates this importance into the construction of neural network attention mechanism and convolution. The method proposed in this paper achieves about 4% improvement over the algorithm without considering

feature engineering. This method brings a lot of inspiration to similar problems. For example, can we use feature engineering to artificially set the importance of different texts, and then improve the neural network through this importance? In addition, this paper transforms the virus activity information in cloud resources into text information, which also provides a good idea for the expansion of depth learning applications.

With the advent of 6G, the society will move towards deeper intelligence, and the problems related to cloud security will become more and more important. This paper does not have a deep understanding of virus activities in cloud servers, and the main feature extraction is realized through the model itself. In the future, we should make better creation according to the knowledge in the field of security.

References

1. Alam, T.: Cloud computing and its role in the information technology. IAIC Trans. Sustain. Dig. Innov. (ITSDI) **1**(2), 108–115 (2020)
2. Bello, S.A., et al.: Cloud computing in construction industry: use cases, benefits and challenges. Autom. Constr. **122**, 103441 (2021)
3. Kumar, R., Goyal, R.: On cloud security requirements, threats, vulnerabilities and countermeasures: a survey. Comput. Sci. Rev. **33**, 1–48 (2019)
4. Ahsan, M.M., Gupta, K.D., Nag, A.K., Poudyal, S., Kouzani, A.Z., Mahmud, M.A.P.: Applications and evaluations of bio-inspired approaches in cloud security: a review. IEEE Access **8**, 180799 (2020)
5. Nassif, A.B., Talib, M.A., Nasir, Q., Albadani, H., Dakalbab, F.M.: Machine learning for cloud security: a systematic review. IEEE Access **9**, 20717–20735 (2021)
6. Dipendra, R.: A Deep Learning Approach for Intrusion Detection using Recurrent Neural Network. Department of computer Science & information Technology (2018)
7. Niu, Z., Zhong, G., Yu, H.: A review on the attention mechanism of deep learning. Neurocomputing **452**, 48–62 (2021)