International Journal of Latest Research in Engineering and Technology (IJLRET)
ISSN: 2454-5031
www.ijlret.com || Volume 09 - Issue 04 || April 2023 || PP. 11-16

APT Attack Detection Method Based on Transformer

Xiao Ziyang[1], Huang Chunhui[2], Qiu Rixuan[1], Zhang Junfeng[3], Lin Nan[3]
[1](*Information and Communication Branch of State Grid Jiangxi Electric Power Co., Ltd., China)*
[2](*Xiamen Tobacco Industrial Co., Ltd., China)*
[3](*State Grid Jiangxi Electric Power Company, China)*

Abstract: APT attack is the most popular type of network attack nowadays, and its complexity, diversity, and strong concealment make it difficult to detect. Traditional APT detection methods only stay in a simple perspective of network attacks, which makes it difficult to capture the above characteristics, resulting in lower detection efficiency. Therefore, this article proposes to use the transformer model to detect APT attacks. This model uses the self attention mechanism to achieve the capture of relational dependencies against long sequence data, thereby achieving the detection of APT attacks, and achieving very good results. By setting appropriate hyper parameters and loss functions, the model was trained for multiple rounds, and the expected results were achieved.

Keywords: network security, Advanced Persistent Threats, long sequence, transformer, deep learning

I. INTRODUCTION

Nowadays, the situation of network security is becoming increasingly severe. Network attackers use ingenious attack techniques to avoid firewalls, invade network systems to obtain privacy information, destroy network systems, or cause server paralysis. The research on network intrusion detection [1] has become one of the most important research directions in network security today. Network intrusion detection detects intrusion and attack behavior by analyzing computer systems and network events. In a network system, any unauthorized "The activities of, as well as attempts to bypass network security mechanisms, can be considered network intrusion behavior.". Network intrusion detection can be divided into two types [2]: anomaly based detection and misuse based detection. Anomaly based detection systems detect attacks by observing the abnormal behavior of networks, systems, or users, while misuse based detection systems use a priori attack pattern and signature to detect attacks.

APT attack detection is also considered a relatively special type of network intrusion detection [3], but APT attacks are more complex and diverse than simple network attacks. Deeper concealment.APT attack is a kind of organized network attack that targets a specific target. It is usually hard to detect and lasts for a long time. APT attack is different from traditional network attack. It can adjust and improve its attack method according to the defender’s detection ability. And after invading the target system, it will quickly obtain the highest authority and self-starting function, which makes APT attack more difficult to be detected. There are many examples of APT attacks. In 2010, a malicious software called Stuxnet [4] was used to disrupt Iran's nuclear facilities, affecting about 1000 centrifuges. This is known as a seismic network attack. In 2011, RSA [5] was attacked by APT, resulting in the theft of its Secure ID token authentication system and its use to invade its customers' networks. At the end of 2020, Solar Winds [6] Orion software was implanted into a backdoor program, affecting more than 18000 customers, including U.S. government departments and enterprises. This attack is known as the Solar Winds supply chain incident.

In the research of network intrusion detection, Mohsen [7] et al. proposed a minimum maximum K-means clustering method for intrusion detection. This algorithm attempts to minimize the maximum internal variance of a cluster, rather than minimizing the sum of internal variances like the K-means algorithm. Each cluster has a certain weight, and a higher weight is assigned to the cluster with a larger internal variance. This algorithm achieves a detection rate of 81%. Li [8] et al. proposed a two-stage "intelligent intrusion detection method". The first stage involves using a random forest algorithm to obtain a subset of features by weighing their importance. The second stage is a classifier based on a subset of features as input "Adaboost algorithm based on hybrid clustering". Li Jun [9] and others considered the timing characteristics of network intrusion data and used GRU_ RNN network structures are trained on KDD datasets to achieve better recognition rates and convergence than other non sequential networks.

There are many methods for detecting APT attacks, among which three are more common: based on social engineering, based on abnormal traffic and based on machine learning. Based on social engineering methods is to analyze the correlation between network events and find out the signs of APT attacks. For example, Amir and Morteza Amini [10] proposed a scheme that uses system ontology and security policies to detect APT attacks. Based on abnormal traffic methods is to monitor the traffic changes caused by data theft and

identify suspicious transmission behaviors. For example, Sana Siddiqui [11] et al. proposed a method of classification based on correlation fractal dimension; Based on machine learning methods is to train a large number of data samples, optimize mathematical models, and achieve recognition of unknown attack samples. For example, Liu Haibo [12] et al. proposed an APT attack detection method that combines GAN and LSTM networks; Liang Ruo Zhou et al. proposed an APT attack detection method based on sequence feature extraction and GRU algorithm combined with K-means clustering for system behavior modeling and source tracing graph.

Although many existing studies have explored the application of machine learning and deep learning in APT attack detection, these studies classify normal behavior and attack behavior, and build APT attack detection models based on machine learning methods, which have certain detection effects, but there are still some problems. The main manifestations are: 1) The amount of data labeled as normal behavior in training samples is far greater than that of illegal behavior, and the distribution of data characteristics is severely uneven, resulting in difficulty in training the model and insufficient generalization ability. 2) APT attacks are usually a continuous behavior in time. Most models do not have temporal learning ability and lose temporal features. Although some methods based on cyclic neural networks can learn temporal features, their sequential training methods based on sequences have problems such as long training time and low convergence efficiency.

Transformer [13] was originally applied to natural language processing (NLP) tasks, with its structure completely abandoning network structures such as RNN [14] and CNN [15], and only using the Attention mechanism to perform machine translation tasks, with good results. Transformer is significantly different from sequential neural networks based on RNN. RNN training is iterative and sequential, while Transformer training is parallel, that is, all features are trained simultaneously, greatly increasing computational efficiency.

By analyzing the data characteristics of network intrusion behavior, this paper proposes an APT attack detection method based on Transformer neural network model. The main contributions of this article are as follows:

1) Aiming at the temporal correlation of network attack behavior data, a Transformer based APT attack detection method is proposed to further improve the accuracy of APT detection.
2) A multi headed self attention mechanism Transformer network model based on sequence features is designed to solve the problems of traditional sequential neural network models that are not easy to converge and have high time overhead. By selecting the optimal loss function and training parameters for parallel training, APT detection is implemented.

II. DATA ANALYSIS AND PREPROCESSING

This article explores the detection methods of APT attacks from the perspective of collected network traffic data. However, the collected network traffic data has strong redundancy, unclear characteristic rules, and there is a serious sample imbalance problem, that is, normal traffic is far greater than network attack traffic. Therefore, we need to analyze the data, establish the correlation between various features, highlight key features in the data through certain dimensionality reduction measures, and remove certain redundancy in the data itself. The above methods can improve the convergence and detection effect of the model. This article uses KDD-Cup-99 and NSL-KDD network intrusion data sets. The KDD-Cup-99 dataset [16] is the dataset used in the third international knowledge discovery and data mining tool competition, with a total of 23 tags and 4898431 pieces of data, including normal and 22 attack type tags. The NSL-KDD dataset [17] is an improved version of the KDD-Cup-99 dataset, containing 125973 network connection records.

Currently, due to the long duration of the complete APT attack process and the high cost of data collection, as well as the ongoing construction of new power systems, and the lack of relevant equipment to collect network attacks on power systems, most in-depth learning experiments on APT attack detection are based on intrusion detection datasets. This article selects the KDDCup99 dataset as the dataset for the experimental part. The dataset includes 41 fixed characteristic attributes and 1 classification identifier. Its characteristic attributes mainly include the basic characteristics and content characteristics of TCP connections, time-based network traffic statistics, and host-based network traffic statistics. It records 9 weeks of network connection and system audit data. Take 41 of these features as sample input features, and use a data preprocessing program to replace each data tag with 5 types of tags, including denial of service (DoS) attacks, remote network user attacks (R2L), privilege escalation attacks (U2R), and probe attacks, as well as normal data. DoS attack is one of the main attack methods of APT attack. The data set distribution is shown in Table 1.

Table 1: Dataset Distribution

Dataset Name	quantity
KDD-Cup-99	4898431
NSL-KDD	125973

At the same time, in order to make the distribution of experimental data more balanced, it is necessary to normalize it. For character type data, we need to convert it to numeric type before it can be received by the transformer to achieve final APT detection.

Due to the large difference in the data range in the original data, it is not conducive to network training. Therefore, it is necessary to normalize each column of the original data. Normalize the same column of data to between (0,1). The normalization formula is as follows:

$$X_n = \frac{X - X_{\max}}{X_{\max} - X_{\min}} \qquad (1)$$

Where in X_{max} and X_{min} represent the maximum and minimum values in the range of original characteristic values, respectively, X represents the original characteristic value, and X_n represents the normalized characteristic value.

Due to the fact that the original dataset contains character string features, which is not conducive to direct vectorization, data labels are subjected to One-hot encoding for ease of calculation. One-hot coding is a commonly used data coding method in machine learning classification tasks. It can convert discrete values in the original data into points in Euclidean space, maintaining a reasonable feature distance between labels. Each data in the dataset is divided into two categories: normal or abnormal. The normal code is 01, and the abnormal code is 10.

In order to remove the impact of redundant information in data sets on detection accuracy, a feature extraction network F is introduced as the front-end network of the intrusion detection model. The network consists of two fully connected layers, with the purpose of mapping redundant low-level features to high-level features. The calculation process of feature extraction network F is shown in equation 2.

$$y = \sigma(x') = \sigma(F(x^*)) \qquad (2)$$

Where: x* is the obtained normalized data; The characteristic length is dx*; The feature extraction network outputs its high-level feature vector x'; The feature length is dx, set dx<dx*; σ Is the activation function. Using the inactive feature vector x' of the output layer as a high-level mapping, a new feature dataset D can be obtained by training the network F on the entire original dataset, where the feature dataset D is composed of feature vectors with a feature length of dx.

III. APT ATTACK DETECTION METHOD BASED ON TRANSFORMER MODEL

Since the transformer model does not use a cyclic neural network (RNN) or convolutional neural network (CNN) architecture, it is necessary to encode the input data to record the location information of the data for subsequent processing. The formula for location coding used in this article is as follows:

$$PE(pos,2i) = \sin(pos/100002i/d_{model}) \qquad (3)$$

$$PE(pos,2i+1) = \cos(pos/100002i/d_{model}) \qquad (4)$$

Where pos is the position in the sequence, dmodel is the dimension of the position information encoded feature vector, i represents the ith element of the position information encoded feature vector, the odd bits in the encoding vector are encoded with cos, and the even bits are encoded with sin. In order to record part of the information in the initial input sequence, this article adds a new convolutional embedding operation, where the result of the position encoding and the result of the embedding operation are added to form the final position encoding result, The initial feature vector and the result of position coding are added to form a new input.

The position encoded vector obtained above is a set of time series features (x1, x2, x3,..., xt), where xi represents the feature vector at the ith moment, and i represents the time attribute. The feature vectors will be input into the next encoder layer. The function of the encoder is to map the vectors to a higher dimensional space, thereby enabling the model to learn more in-depth information and improving the generalization ability and performance of the model. The encoder fully adopts the basic architecture of the transformer model, mainly including the multi head self attention layer, residual network, normalization, and feed forward network layer. The main network structure is shown in Fig 1.

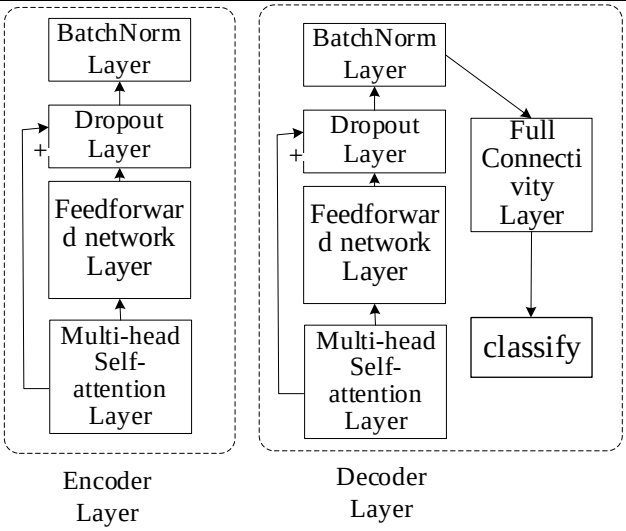

Fig 1 Transformer Network Structure

(1) Multi_head self attention. This is the core module of the transformer model and the main reason for its significant success. Its main idea is to achieve interaction between multiple sequences through three updatable matrices, using matrix multiplication to obtain attention results between multiple sequences, thereby reflecting the dependency relationship between them. Based on this relationship, effective classification tasks can be better achieved. Therefore, transformer is very suitable for APT detection, a long sequence based classification task.The purpose of dividing self attention into multiple heads is to learn attention results from multiple feature angles, and finally, multiple attention results will be spliced to restore the original feature vector shape.The formula is as follows：

$$Attention(Q,K,V) = \text{soft}\max(\frac{QK^T}{d})V \qquad (5)$$

Where Q, K, and V are the three matrices that require weight updates, d is the dimension of the feature vector, and soft max is the probability normalization function. The formula is as follows:

$$Soft\max(z_i) = \frac{e^{z_i}}{\sum_{c=1}^{C} e^{z_c}} \qquad (6)$$

Where zi is the output value of the ith node, and C is the number of output nodes, that is, the number of categories classified.

(2) Residual network. This is a classic network structure proposed by He Kaiming. The core idea is to add or splice the output results and input feature vectors in the middle of the network to prevent the gradient disappearance phenomenon caused by the deepening of the network layers. This article uses the addition operation here.

(3) Normalization. The purpose here is to make the data distribution more uniform, thereby making the learning effect of the model better. Commonly used normalization methods include Batch Norm and Layer Norm, which are used to normalize at the batch and network levels, respectively, and can achieve different effects. This article uses the Batch Norm operation here.

(4) Feed forward network. The feed forward network is equivalent to the superposition of multi-layer neural networks. It can achieve comprehensive learning of features through rich network layers, thereby enabling the model to obtain more abundant information. Its role is basically consistent with the previously mentioned fully connected layer.

The actual network architecture used in this article is the result of the superposition of several layers of the above modules, which allows for richer information to be learned. Generally, the number of layers is set to 6.

The structure of the decoder is basically consistent with that of the encoder. Unlike the encoder, the decoder is used to achieve the final classification result, so a full connectivity layer is used to reduce the vector dimension to the final required number of classifications. Other structures are consistent with the encoder.

Due to the imbalance between positive and negative samples in data samples, FocalLoss is used as a loss function, and the calculation formula is shown below. It is widely used in difficult sample mining in target detection tasks. By adjusting the weight of positive and negative samples, the model pays more attention to samples that are difficult to classify in training, effectively alleviating the problem of uneven data distribution.

$$L_{fl} = \begin{cases} -\alpha(1-y)^{\gamma}\log y, y'=1 \\ -(1-\alpha)y^{\gamma}\log(1-y), y'=0 \end{cases} \quad (7)$$

Where y is the output of the classification layer activation function; y'is the true value, that is, the encoded label;α and γ is a regulatory factor,α value is 0.25, γ value is 2.

Based on the above network structure and loss function, input the APT network attack traffic into the model, select the appropriate training set proportion and optimizer parameters, train the data, use k-fold cross validation to select the optimal parameters, save the training weights, and finally put the test set into the trained model to achieve the final detection of APT attacks.

IV. EXPERIMENT

In order to verify the effectiveness of the model, this article compares the transformer model with the traditional machine learning method, SVM deep learning method, LSTM, and compares the accuracy of the indicators used. The experimental results are shown in Fig 2. As can be seen from the figure, using the transformer model can achieve effective APT attack detection.

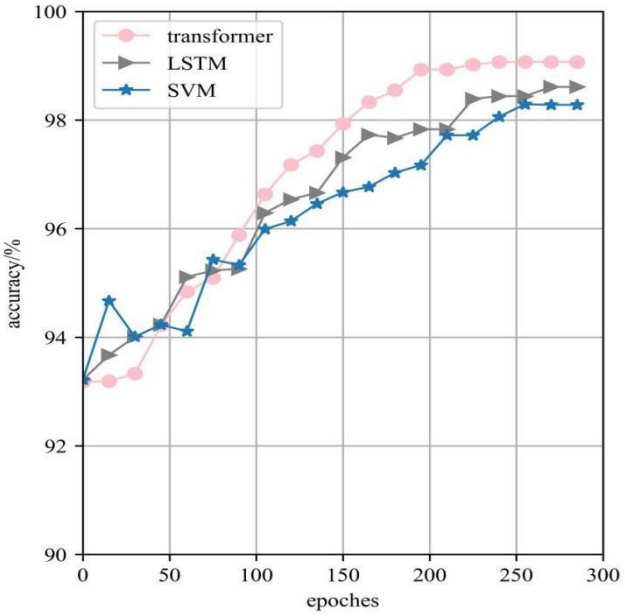

Fig 2 Comparison of Experimental Results

V. CONCLUSION

An APT attack detection method based on the transformer network model is proposed. The proposed transformer network model is based on network traffic characteristics and utilizes the multi header self attention mechanism to learn the timing characteristics of APT attack data. By selecting the optimal loss function and training parameters for parallel training, combined with the data preprocessing method of feature extraction, it alleviates the problem of uneven data distribution and effectively improves the detection effect. Experimental results show that compared to traditional methods and deep learning methods, the best detection effect is achieved.

VI. ACKNOWLEDGEMENTS

This work was supported by the State Grid Jiangxi Electric Power Corporation Science and Technology Project "Research on Active Defense Technology for Advanced Sustainable Network Attacks Based on Dynamic Obfuscation" under Grant 521835220003.

REFERENCES

[1]. Aydeger A, Akkaya K, Cintuglu MH, Uluagac AS, Mohammed O, Software Defined Networking for Resilient Communications in Smart Grid Active Distribution Networks[C], IEEE International Conference on Communications, Kuala Lumpur, Malaysia, 2016.
[2]. A. M. Lajevardi and M.Amini , A semantic-based correlation approach for detecting hybrid and low-level APTs[J], Futur . Gener. Comput. Syst., vol. 96, pp.64–88, 2019 .
[3]. M. Marchetti, F. Pierazzi , M. Colajanni , and A. Guido, Analysis of high volumes of network traffic for Advanced Persistent Threat detection[J], Comput . Networks, vol. 109, pp. 127–141, Nov. 2016 .
[4]. T. Bodström and T. Hämäläinen, A Novel Method for Detecting APT Attacks by Using OODA Loop and Black Swan Theory [J], Lecture Notes in Computer Science, vol. 11280, Springer,2018, pp. 498–509 .
[5]. Liu Haibo, Wu Tianbo, Shen Jing, Shi Changting. APT attack detection based on GAN-LSTM [J]. Computer Science, 2020, 47(01):281-286.
[6]. A. Zimba, H. Chen, and Z. Wang, Bayesian network based weighted APT attack paths modeling in cloud computing[J], Futur . Gener. Comput . Syst., vol. 96, pp. 525–537, 2019.
[7]. Liang He, Li Xin, Yin Nannan, Li Chao. APT Attack Detection Method Combining Dynamic Behavior and Static Features [J/OL]. Computer Engineering and Application: 1-13 [2022-11-07]. http://kns.cnki.net/kcms/detail/11.2127.tp.20220622.1059.008.html
[8]. Guo Chuangxin, Liu Zhuping, Feng Bin, Jiang Boyou, Guo Jun, Li Fucun. Research Status and Prospects of Risk Assessment of New Power System [J]. High Voltage Technology, 2022, 48(09): 3394-3404.DOI:10.13336/j .1003-6520.hve .20221101.
[9]. Vaswani A, Shazeer N, Parmar N, etal. Attention is all you need. arXiv2017 [J]. arXiv preprintar Xiv:1706.03762, 2017: 2999-3007.
[10]. [10]Wojciech Zaremba, Ilya Sutskever, Oriol Vinyals . Recurrent Neural Network Regularization. [J]. CoRR, 2014, abs /1409.2329.
[11]. Y. Le Cun, L. Bottou,Y . Bengio, P. Haffner. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11).
[12]. Zhou H, Zhang S, Peng J, et al. Informer: Beyond efficient transformer for long sequence time-series forecasting[C]//Proceedings of AAAI. 2021.
[13]. Shen S, Yao Z, Gholami A, et al. Powernorm: Rethinking batch normalization in transformers [C]//International Conference on Machine Learning. PMLR, 2020: 8741-8751.
[14]. M. Zhao, S. Zhong, X. Fu, B. Tang, M. Pecht, Deep residual shrinkage networks for fault diagnosis, IEEE Transactions on Industrial Informatics, 2020, 16(7): 4681-4690.
[15]. Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun 0001. Deep Residual Learning for Image Recognition. [J]. CoRR, 2015, abs /1512.03385.
[16]. KDD Cup 1999 [DB/OL]. (1999-10-28) [2021-05-06]. http:‖kddi.cs.uci.edu/databases/kddcup99/kddcup99.
[17]. The NSL KDD Dataset [DB/OL]. (2013-7-30) [2021-05-06]. http:‖nsl.cs.unb.ca/NSL-KDD/.