Malicious URL Detection System Based on LSTM and Attention Mechanism

Bocheng Liu, Xianang Zeng* and Pengxiang Dong
School of Software, Nanchang University, Nanchang 330047, China

*Corresponding author e-mail: bcliu@ncu.edu.cn

Abstract. With the rapid development of the Internet, network security has gradually received everyone's attention and attention. In network security, Web application security is particularly important. Therefore, for the identification and detection of malicious URLs, this paper constructs a new feature extraction method based on traditional machine learning algorithms, and proposes a deep learning model based on the LSTM algorithm and adding the Attention mechanism. The results show that the accuracy rate of the model reaches 98.9%, and it can complete the malicious URL detection task well. On this basis, a malicious URL detection system based on the LSTM+Attention mechanism was developed to improve the security of cyberspace.

Keywords: LSTM Algorithm, Attention Mechanism, Malicious URL Detection System

1. Introduction
With the rapid development of the Internet, people are designing more and more frequently on the Internet. Driven by this rapid development, Web security issues have become more important, and the research on security protection from the URL of the Web access portal has also become more and more important. More and more important. At present, the method for most websites to identify malicious URLs is to use blacklists to identify malicious URLs. However, this method requires timely maintenance and update of malicious URL forms, which makes the cost higher.

Currently, some scholars apply machine learning technology to the field of identifying malicious URLs. Sha Hongzhou (2016) showed us the most advanced research results by summarizing and integrating various machine learning recognition technologies [1]. Chen Kang et al. (2018) identified malicious URLs by converting URLs into feature images, and then using CNN for feature extraction and classification [2]. Wei Xu et al. (2019) identify malicious URLs based on feature fusion combined with machine learning [3]. Zhang Hui (2019) proposed a method to discover the comprehensive feature space [4]. Frank (2017) and others will detect malicious URLs as a two-class problem [5]. They use a variety of methods to extract the relevant features of the URL, and then use random forest, K nearest neighbors, SVN and decision tree algorithms to build the model. Justin (2009) used statistical related methods to extract 9 features of URL for model training and construction [6].
In this article, we refer to the feature extraction schemes of many documents and add some certified custom features to form a new feature extraction scheme. Based on the deep learning model, the LSTM architecture algorithm is adopted, and the Attention mechanism is introduced, so that the accuracy of the model reaches 98.8%, and the detection of malicious URLs can be completed well under real conditions [7-9].

2. System Design
The system first vectorizes all URLs, and then completes them to form a vector sequence with a uniform length of 128, and then uses the LSTM model and the Attention mechanism. After that, linear transformation is performed on the features output by the model, and finally a two-classification process is performed, and the output result is passed through the sigmoid function to obtain the final result.

   1) Embedding layer (embedded layer). The initial input data is a sequence vector with a length of 128. After embedding through the Embedding layer, it is output to a 64-dimensional vector space.
   2) Two-layer Dropout layer, at each training, half of the feature detectors are randomly stopped to work, which can improve the generalization ability of the network and prevent over-fitting.
   3) LSTM layer. After inputting the vector into the 64-dimensional LSTM, it is processed in the 64-dimensional hidden layer. It can capture the long-term dependence in the sentence, so as to better understand the emotion of the text as a whole. There are three gate structures in the memory unit: forget gate $f_t$, input gate $i_t$ and output gate $o_t$, which are used to record and update the information of the memory unit.
   4) Attention layer. After the 64-dimensional vector is output from the LSTM, through the Attention mechanism, the importance of each target unit is distinguished, and the text weight in it is processed to make the classification more accurate.
   5) Dense layer (fully connected neural network layer). Set the final output as one layer to facilitate the processing of the final result.

3. Systems Realization

3.1. Data Preprocessing
The keras tokenizer can convert URL text into digital vector form. The tokenizer will divide the URL into many words, and then rank all the words according to the number of various words, and then form a number vector according to the number of positions where each word appears. It also counts the length of the vocabulary contained in all URLs. According to the box-plot statistical method in statistics, the upper limit is 23.5. It can be considered that URLs with a vocabulary longer than 24 are too long. Therefore, in the preprocessing, the maximum number of words in all URLs is set to 24, among which, URLs with a length less than 24 are all filled with 0 in length, and finally a numeric matrix with a uniform length of 24 is formed. After that, scramble all the data, and divide the training set and test set by 4:1.

3.2. Model Training
The pre-processed data is input into the network, and the sample passes through the embedding layer, then passes through the first dropout layer, then enters the 64-dimensional LSTM layer, and then passes through the custom Attention layer.

The Attention layer is customized based on the Attention mechanism. The Attention mechanism can be regarded as an automatic weighting scheme, which means that all nodes that appear are given corresponding weights according to their contribution to the final result. The more important the node, the higher the weight. Under the action of this mechanism, important nodes will be more directly fed back into the results, making the results more accurate.
After the data is output from Attention, after another layer of Dropout layer, it finally enters the fully connected layer with the number of output neurons being 1. Finally, a value will be output. This value is between \((0, 1)\). It can be used to determine whether the detected URL is malicious or not by whether the value exceeds 0.5.

### 3.3. Model Evaluation

The selected three traditional algorithm models, LSTM algorithm model and LSTM+Attention algorithm model are optimized through parameter tuning. For the performance results of the five models, we evaluate the five model algorithms through three evaluation methods: precision, recall, and f1-score. Table 1 shows the evaluation results of each algorithm model.

| Model name           | precision | recall | f1-score |
|----------------------|-----------|--------|----------|
| LogisticRegression   | 88.9%     | 88.3%  | 88.5%    |
| SVM                  | 94.4%     | 94.2%  | 94.3%    |
| RandomForest         | 95.4%     | 95.4%  | 95.4%    |
| LSTM                 | 98.8%     | 98.7%  | 98.7%    |
| LSTM+Attention       | 98.99%    | 98.9%  | 98.9%    |

Compare the LSTM and LSTM+Attention models with the traditional Logistic Regression algorithm, SVM algorithm and Random Forest algorithm. The loss rate and accuracy rate are compared in Table 2.

| Model name           | val_loss | val accuracy |
|----------------------|----------|--------------|
| LogisticRegression   | 0.1162   | 0.889        |
| SVM                  | 0.0571   | 0.944        |
| RandomForest         | 0.0455   | 0.954        |
| LSTM                 | 0.0422   | 0.9880       |
| LSTM+Attention       | 0.0349   | 0.9899       |

Note: val_loss: test set loss value, val_accuracy: test set accuracy rate

The initial number of iterations is set to 10. Figure 1 and Figure 2 respectively show the trend of LSTM, LSTM+Attention accuracy and loss value with the number of iterations. It can be seen that as the number of iterations increases, the accuracy rate will gradually increase. But after 6 iterations, the accuracy rate stayed in a relatively stable interval. It can be considered that 6 steps are the optimal number of iteration steps for the model.

![Figure 1. Accuracy rate change](image1)

![Figure 2. Loss value change](image2)

Through the above evaluation, compared with the traditional machine learning algorithm, the LSTM model with Attention mechanism performs better in accuracy, reaching more than 98.8%.
3.4. Model Application
We build a simple and practical visual webpage detection system based on the malicious URL detection model of LSTM+Attention algorithm. By providing an external interface, the user can enter the URL that needs to be queried on the web page, and then the system detects the uploaded URL and returns the detection result to the web page. The modified system mainly includes three functional parts: user URL upload, data preprocessing, and detection. The system process is shown in Figure 3.

![Figure 3. System flow chart](image)

**Figure 3. System flow chart**

URL upload module: The system provides users with a URL input interface. After entering the URL, the website uploads the URL to the backend through the form. If you encounter other problems such as data transmission, jump to the 404 error interface to facilitate the user to return to the detection interface.

Data preprocessing module: After getting the uploaded URL, the backend calls related functions to segment the URL, and then convert it into a digital matrix. Then according to the preprocessing method during model training, the URL matrix is lengthened or shortened, and then the part that exceeds the threshold is reset to zero.

Detection module: After getting the standard feature data input, the URL is detected through the trained model, and the detection result is returned to the interface [10]. In the returned result, in addition to judging whether the URL is malicious, it will also provide its IP address. Because the prediction result of the model finally comes out with a regression function value, and because of the sigmod function, we feed back the function value as an evaluation score of the possibility of malicious URL, and finally display it on the front end. The effect of the final test result feedback on the web page is shown in Figure 4 and Figure 5.

![Figure 4. Diagram of detection results](image)

**Figure 4. Diagram of detection results**

![Figure 5. Diagram of detection results](image)

**Figure 5. Diagram of detection results**

4. Conclusions
Malicious URLs are not uncommon in daily Internet life, and pose a great threat to Internet security. However, the existing detection methods are inefficient and cannot detect newly emerging malicious URLs. Therefore, detecting malicious URLs based on machine learning models is an extremely meaningful technological breakthrough. Based on this, this paper proposes a malicious URL detection system based on the LSTM algorithm and adding the Attention mechanism. Compared with the traditional LSTM detection system, the Attention layer is added, and the words involved in the URL are weighted. Through this method In the end, benign URLs and malicious URLs will get a better classification effect. Compared with traditional machine learning, both data processing procedures and detection accuracy have been greatly optimized and improved. In addition, we developed a visual
website system based on the LSTM+Attention algorithm model. The user can enter the URL that needs to be queried on the website. The back end of the system will rely on the algorithm model to evaluate and detect the URL, and finally return the detection result to the front-end interface. Practice has proved that the system can meet the requirements in terms of performance and function. The needs of users.

Acknowledgments
This work was supported by Primary Research & Development Plan of Jiangxi Province (no.20192BBE50075), Student innovation and entrepreneurship training program(no. 2020CX164)

References
[1] Sha Hongzhou, Liu Qingyun, etc. Review of research on malicious webpage identification . Chinese Journal of Computers, 2016, 39 (3)
[2] Chen Kang, Fu Huazheng, Xiang Yong. Malicious URL recognition based on deep learning . Computer System Applications, 2018, 27 (6)
[3] Wei Xu, Cheng Weiqing. Research on malicious webpage recognition based on feature fusion and machine learning. Journal of Nanjing University of Posts and Telecommunications, 2019, 39 (5)
[4] Zhang Hui. Research and discovery of malicious website identification methods based on machine learning. Master's thesis of Beijing University of Architecture and Architecture, 2019
[5] Vanhoenshoven, Frank, Napoles Ruiz, et al. Detecting Malicious URLs using Machine Learning Techniques. Computational Intelligence. IEEE, 2017
[6] J.Ma,LkSaul,S.Savage,and GMVoelker,"Beyond blacklists:learning to detect malicious web sites from suspicious URLs," in ACM SIGKDD International Conference on Knowledge Discovery and Date Mining. Paris, France: ACM, 28 June-1 July 2009
[7] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation, 1997, 9 (8): 1735-1780.
[8] BAHDANAU D, CHO K, BENGIO Y, etal. Neural machine translation by jointly learning to align and translate // in Proceedings of International Conference on Learning Representations (ICLR), San Diego, CA.
[9] Research on malicious URL detection based on LSTM [D]. Master's thesis of Central China Normal University, 2019
[10] Anderson H S, Woodbridge J, Filar B. DeepDGA:adversarially-turned domain generation and detection. Proc of ACM Workshop on Artificial Intelligence and Security. New York: ACM Press, 2016