Scholars Journal of Engineering and Technology

OPEN ACCESS

Abbreviated Key Title: Sch J Eng Tech
ISSN 2347-9523 (Print) | ISSN 2321-435X (Online)
Journal homepage: https://saspublishers.com

URL Detection Based on Quantum Long Short-Term Memory Neural Network

Shihao Zhu[1], Yuanyuan Huang[1*], Linglong Huang[1], Siyu Li[1], Peilin He[2]

[1]Chengdu University of Information Technology, Chengdu, 610225, P.R. China
[2]Department of Informatics and Networked Systems, University of Pittsburgh, Pittsburgh, PA 15260, USA

DOI: 10.36347/sjet.2024.v12i05.001 | **Received:** 09.03.2024 | **Accepted:** 15.04.2024 | **Published:** 04.05.2024

***Corresponding author:** Yuanyuan Huang
Chengdu University of Information Technology, Chengdu, 610225, P.R. China

Abstract | **Original Research Article**

In today's era, with the continuous development of the internet, surfing the internet has become an inseparable part of our lives. For those who frequently roam the internet, URLs are information that we are constantly exposed to. And YRL is also divided into malicious and normal. Clicking on malicious URLs can cause certain losses to our internet tools, and in severe cases, it can even lead to money theft. Malicious URL detection is an important task in the field of network security, aimed at identifying and preventing potential network attacks and malicious activities. Traditional malicious URL detection methods, such as Blacklist based detection and Machine learning based on feature extraction, have achieved certain results, but face challenges in real-time, accuracy, and data requirements. Therefore, this article proposes the use of Quantum Long Short-Term Memory Neural Network (QLSTM) to solve this problem. By observing the accuracy, recall, and F1 values obtained from training a large amount of data on the QLSTM network, it is shown that this method is feasible.

Keywords: Malicious URL detection, Quantum Long Short-Term Memory Network, network security.

1. INTRODUCTION

Long Short-Term Memory (LSTM) networks excel in processing and making predictions based on time series data, as they are adept at identifying long-range dependencies within the sequence. learned from [1, 2], this trait is particularly beneficial in malicious URL detection, where the correlation between different segments of a URL can be indicative of its nature. However, traditional LSTM networks can falter in precision when confronted with vast datasets typical in cybersecurity applications.

In [3], we can know Quantum Long Short-Term Memory (QLSTM) is an innovative architecture that synergizes quantum computing principles with LSTM networks. It leverages quantum mechanics' features-like qubits and entanglement-to enhance LSTM's structure and algorithms. QLSTM offers more efficient handling of large datasets and a boost in predictive performance.

For QLSTM's application in detecting malicious URLs, the process starts with data preprocessing: the URL is dissected into components-domain, path, parameters-and relevant features are extracted. These features serve as inputs for the QLSTM, which, through quantum-enhanced mechanisms, learns to distinguish URL characteristics with heightened effectiveness. The training involves iterative adjustments to minimize error and fine-tune the model for precise identification.

The resultant QLSTM model is capable of evaluating URLs in real-time. Upon feeding a new URL to the model, it assesses its legitimacy based on the learned indicators, issuing a verdict along with performance metrics such as accuracy, recall, and the F1 score. This allows for swift and accurate categorization of URLs, bolstering cybersecurity efforts.

2. RELATED WORK

Malicious URL detection stands as a cornerstone of network security, serving the critical role of identifying and neutralizing URLs that harbor nefarious intentions or present potential hazards. These URLs often act as conduits for malware, phishing schemes, and fraudulent content, thereby endangering users' personal data and system integrity.

The prevalent methods for combating such threats include blacklist detection, which blocks known

dangerous URLs; machine learning techniques that discern patterns in URL features to predict risks; and behavior analysis-based dynamic detection, which monitors URL interactions for suspicious activity.

Beyond these approaches, a suite of supplementary techniques and tools are employed in the cybersecurity realm. Static analysis involves scrutinizing the components of a URL without activating any linked software, while disassembly breaks down the URL for closer examination. File fingerprinting matches known virus signatures against URLs, and virus scanning systematically searches for and eradicates malware and viruses. These strategies function collectively to track network data, identify threats, and maintain a secure online environment for users.

In [5, 6], we know Quantum Long Short-Term Memory (QLSTM) networks represent an innovative fusion of quantum computing with the well-established LSTM framework, aimed at transcending the limitations of conventional neural models. The traditional LSTM, a specialized recurrent neural network variant, counters the issue of vanishing and exploding gradients common in standard RNNs through its sophisticated gating mechanisms and cell states. This advancement equips LSTMs with the capacity to retain information over lengthy sequences, making them invaluable for complex tasks in natural language processing, speech recognition, and even image analysis.

Quantum computing introduces a paradigm shift by harnessing the principles of quantum mechanics to process information, possessing the theoretical promise of expediting certain computations exponentially. By integrating this potential with LSTM's robust architecture, QLSTM seeks to capitalize on quantum computing's parallelism and high-speed processing. The goal is to augment LSTM's ability to process and analyze extensive, high-dimensional sequential data sets more efficiently, thereby improving the model's overall performance while preserving the inherent strengths of the traditional LSTM design.

3. METHODOLOGY

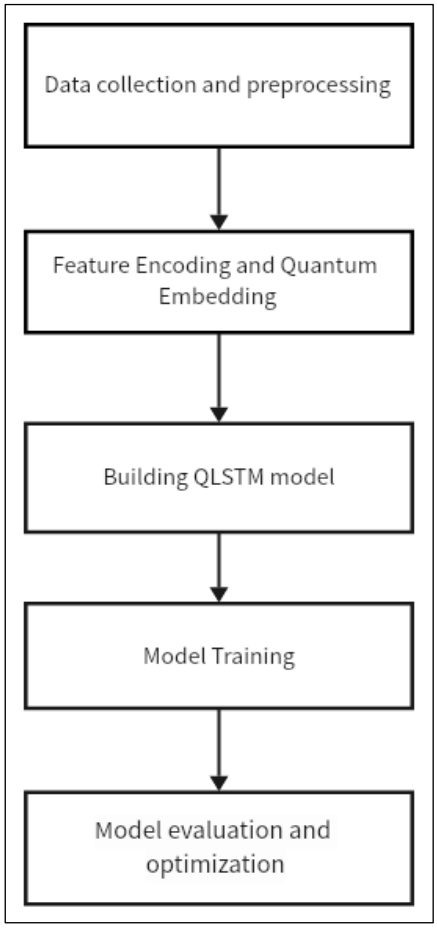

Figure 1: Method flowchart

3.1 Data Collection and Preprocessing

Collect a large amount of URL data, including known normal URLs and malicious URLs. Preprocess the URL, such as parsing its various components (such as protocol, domain name, path, query parameters, etc.). Extract features related to URLs, which may include URL length, character frequency, domain name entropy, and the use of special characters. If possible, natural

language processing techniques can also be used to further analyze and extract features from the text part of the URL.

3.2 Feature Encoding and Quantum Embedding

Given Convert the extracted features into numerical forms suitable for model input. Using quantum embedding technology to transform traditional features into quantum states, in order to leverage the potential advantages of quantum computing. Quantum embedding can map classical data to the quantum state space through quantum circuits, thereby capturing the intrinsic structure and relationships of the data.

3.3 Building QLSTM model

Construct QLSTM model using quantum computing framework and deep learning library. Design a suitable network structure, including quantum embedding layer, QLSTM layer (including memory unit and gate structure), and output layer. Determine the parameters and configuration of the model, such as the number of hidden units, learning rate, etc.

3.4 Model Training

Train the QLSTM model using a labeled URL dataset (including both normal and malicious URLs). Input preprocessed and encoded features into the QLSTM model, and calculate the model's output through forward propagation. Calculate the loss function based on the output of the model and the true labels, and use optimization algorithms (such as gradient descent or its variants) to update the parameters of the model. Repeat the above steps until the model achieves satisfactory performance on the validation set.

3.5 Model Evaluation and Optimization

Evaluate the trained QLSTM model using an independent test set. Evaluate the performance of the model on malicious URL detection tasks by calculating metrics such as accuracy, recall, and F1 value. And obtain the ROC curve and AUC curve.

The Receiver Operating Characteristic (ROC) curve is a tool used to evaluate the performance of classification models. The ROC curve displays the performance of the model by plotting the relationship between True Positive Rate (TPR) and False Positive Rate (FPR). The closer the curve is to the upper left corner, the better the performance of the model. To quantify the performance of the ROC curve, the Area Under the Curve (AUC) value is usually calculated. The closer the AUC value is to 1, the better the performance of the model; The closer the AUC value is to 0.5, the closer the performance of the model is to random guessing.

4. RESULTS AND EVALUATION

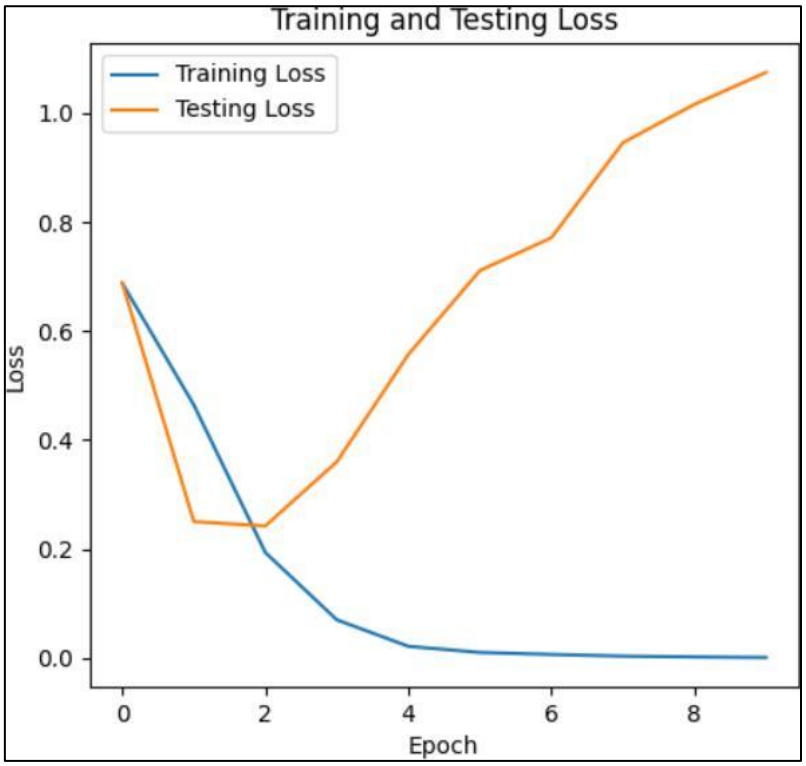

Figure 2: Loss rate of training and testing results

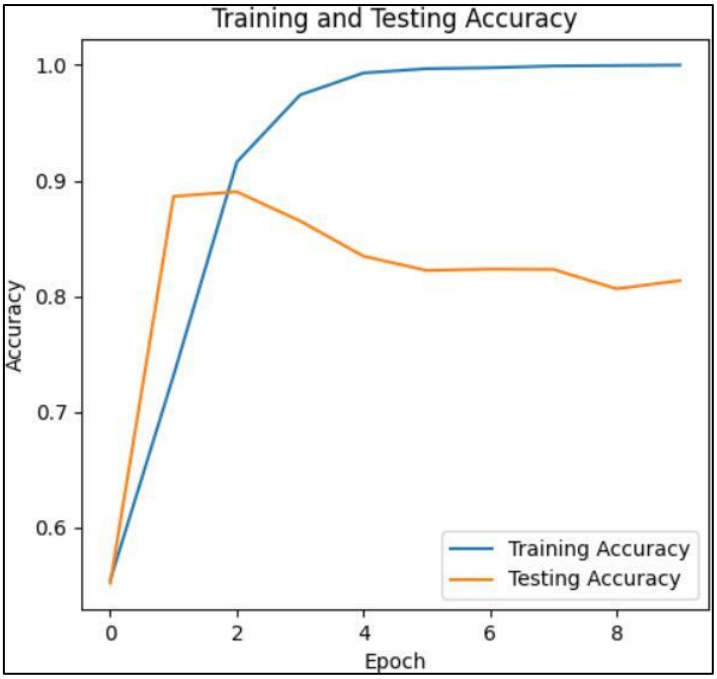

Figure 3: Accuracy of training and testing results

After training nearly 100000 sets of data, the training results are shown in the above figure. During the training process, the loss rate decreases gradually and eventually approaches 0; The accuracy is getting higher and eventually tends to 1. In the prediction of test data, the loss rate gradually decreases and the accuracy gradually increases before the second cycle. However, due to overfitting, the loss rate and accuracy gradually became abnormal after the third round. Therefore, you only need to look at the data around the second round to obtain the effectiveness of QLSTM in detecting malicious URLs.

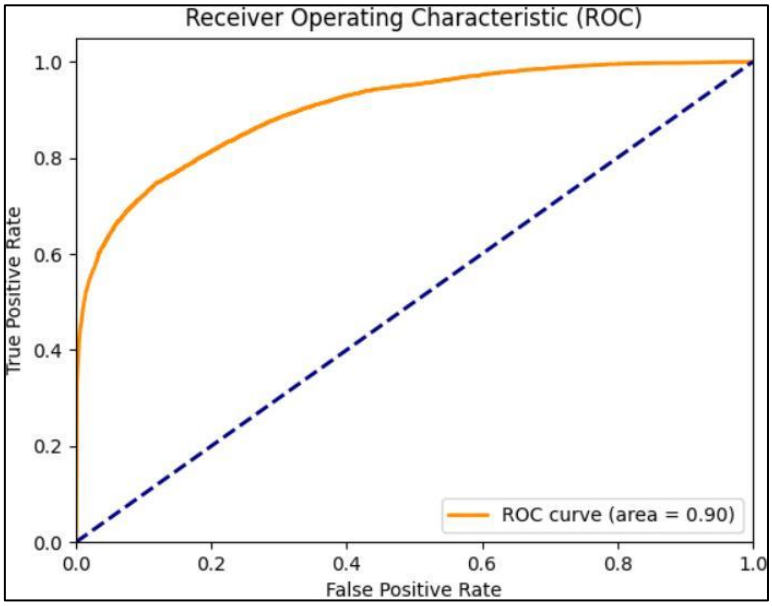

Figure 4: ROC curve

	precision	recall	f1-score	support
0	0.82	0.84	0.83	11541
1	0.80	0.78	0.79	9375

Figure 5: Accuracy, recall, and F1 value

The ROC curve can be obtained from the following figure, and it can also be seen that the value of AUC is about 0.9. As the AUC value approaches 1, it indicates that the performance of the model is better. Observe the accuracy, recall, F1 value and other indicators of the test data, as shown in Figure 4. 1 represents malicious URL, and 0 represents non malicious URL.

5. CONCLUSIONS

In this article, we propose using QLSTM to solve the detection of malicious URLs, and through training and testing on a large amount of data, we have obtained relevant indicators to measure its exchange accuracy, recall, F1 value, and so on. The feasibility of this method was also demonstrated by drawing ROC curves and calculating AUC values. And the AUC value is close to 0.9, indicating a good effect. In summary, QLSTM, as a model that combines quantum computing and LSTM neural networks, has enormous potential and application prospects. With the continuous progress and improvement of quantum computing technology, QLSTM is expected to play an important role in more fields.

ACKNOWLEDGMENT

This work was supported by the Undergraduate College Students' innovation project of Chengdu University of Information Technology (No. 202210621228).

REFERENCE

1. Liu, Y. (2019). Research on Malicious URL Detection Based on LSTM [D]. Central China Normal University.
2. Chen, Z. (2023). Malicious URL Detection and Research Based on Bidirectional Long Short Term Memory and Convolutional Networks [D]. Southwest University.
3. Wang, W., Jiang, M., & Wang, S. (2023). Quantum image chaotic encryption scheme based on quantum long short-term memory network [J]. *Acta Physica Sinica,72*(12), 15-26.
4. Mu, M. (2021). Research and Application of Quantum Neural Networks [D]. Shenyang University of Aeronautics and Astronautics.
5. Tang, Z., Li, H., & Zhang, J. (2017). A network attack homology detection method based on quantum neural networks [J]. *Journal of Chengdu University of Technology (Natural Science Edition), 44*(4), 506-512.
6. Beaudoin, Collin etc. Quantum Machine Learning for Material Synthesis and Hardware Security (Invited Paper). 2022.
7. Zhang, Y., Chen, L., & Hao, H. (2013). A Quantum Neural Network Training Algorithm Based on LM [J]. *Computer Science, 40*(9), 221-224.
8. Liu, Y. (2019). Research on Malicious URL Detection Based on LSTM [D]. Central China Normal University.
9. Zhang, K. (2022). Research on Malicious URL Detection Technology Based on Neural Networks [D]. Guangdong University of Technology.
10. Wang, H. (2021). Research on hybrid malicious URL detection based on "word-location" vector [D]. Chongqing University.