Proceedings of Machine Learning Research vol 245:1–9, 2024 2024 International Conference on Machine Learning and Intelligent Computing

Hybrid Classical Quantum Neural Network with High Adversarial Robustness

Yongxi Yang YANGYONGXI99@163.COM
School of Cybersecurity (Xin Gu Industrial College), Chengdu University of Information Technology, Chendu, 610225, China
Advanced Cryptography and System Security Key Laboratory of Sichuan Province, Chendu, 610225, China
SUGON Industrial Control and Security Center, Chendu, 610225, China

Shibin Zhang* CUITZSB@CUIT.EDU.CN
School of Cybersecurity (Xin Gu Industrial College), Chengdu University of Information Technology, Chendu, 610225, China
Advanced Cryptography and System Security Key Laboratory of Sichuan Province, Chendu, 610225, China
SUGON Industrial Control and Security Center, Chendu, 610225, China

Lili Yan
School of Cybersecurity (Xin Gu Industrial College), Chengdu University of Information Technology, Chendu, 610225, China
Advanced Cryptography and System Security Key Laboratory of Sichuan Province, Chendu, 610225, China
SUGON Industrial Control and Security Center, Chendu, 610225, China

Yan Chang
School of Cybersecurity (Xin Gu Industrial College), Chengdu University of Information Technology, Chendu, 610225, China
Advanced Cryptography and System Security Key Laboratory of Sichuan Province, Chendu, 610225, China
SUGON Industrial Control and Security Center, Chendu, 610225, China

Editors: Nianyin Zeng and Ram Bilas Pachori

Abstract

As the realms of quantum computing and machine learning converge, a novel domain, termed quantum machine learning, is progressively forming within the sphere of artificial intelligence studies. Nonetheless, akin to its classical counterpart, this emerging field is not exempt from security vulnerabilities. Quantum machine learning systems, regardless of whether they process classical or quantum inputs, are susceptible to minor perturbations that can erroneously skew classification outcomes. These minute disruptions, often imperceptible to human observation, present a significant challenge in ensuring the integrity of quantum classifiers. As the complexity of quantum classifiers increases, their vulnerability also gradually grows. To mitigate this issue, this paper proposes a novel hybrid classical-quantum neural network model that enhances the model's adversarial robustness by adding a preprocessing layer for noise reduction and data reconstruction. Experiments demonstrate that this model exhibits higher efficiency and accuracy in noisy environments and against adversarial attacks.

Keywords: adversarial robustness, preprocessing layer, Quantum neural network

1. Introduction

The unique properties of superposition and parallelism inherent in quantum mechanics have demonstrated significant advantages in large-scale computations, thereby bringing quantum computing into the focus of scientists as a burgeoning field. In 1980, Paul Benioff proposed the concept of the quantum Turing machine (Benioff, 1980). By 1994, the introduction of Shor's algorithm had demonstrated the potential to decrypt RSA Algorithm in significantly shorter times. In 1996, Grover introduced a quantum search algorithm for unsorted databases, significantly improving the speed of quantum algorithms in solving search problems. Advancements in quantum computing have led to the fusion of machine learning with quantum principles, creating the field of quantum machine learning. Many such QML (Quantum Machine Learning) efforts are based on the HHL algorithm, which proposes a solution for solving systems of linear equations using quantum operations. Analysis and proofs have shown that these quantum machine learning algorithms exhibit exponential improvements in computational speed compared to classical machine learning algorithms (Schuld et al., 2015). Driven by advances in deep learning, quantum neural networks, which bear similarities to classical neural networks and include variational parameters, have garnered widespread attention (Cerezo et al., 2021).

In recent years, research on Quantum Neural Network (QNN) models has emerged. In 2018, LI and Zhao (2018) proposed a new QNN model based on controlled rotation gates. In 2021, Blance and Spannowsky (2021) introduced a quantum classification algorithm that employs a hybrid approach combining the steepest descent method with quantum gradient descent for network parameter optimization. There are numerous variants of quantum neural networks, including hybrid quantum-classical convolutional neural networks (Liu et al., 2021), quantum graph convolutional neural networks (Zheng et al., 2021), quantum convolutional neural networks (Cong et al., 2019), among others. Yet, quantum neural networks remain prone to disturbances that can diminish the classifiers' accuracy. Therefore, constructing a robust quantum neural network model is of paramount importance. Gong et al. (2024) proposed enhancing quantum adversarial robustness through randomized encoding. You et al. (2019) suggested regularizing neural networks by adding a noise layer, demonstrating that models trained with this method exhibit strong robustness under FGSM attacks. This paper presents enhancements to hybrid quantum-classical neural networks, resulting in our model achieving higher robust accuracy compared to both conventional network models and those subjected to adversarial training.

2. Related Work

2.1. Quantum Computing

Research in quantum computing is built upon the principles of quantum theory, utilizing distinctive quantum properties like state superposition and entanglement to execute computational operations.

2.1.1. QUBIT

In classical information theory, the bit is the basic information unit, representing binary 0 or 1. Similarly, quantum information theory's fundamental unit is the quantum bit, or qubit. Leveraging quantum superposition, a qubit exists in a state that combines these basis states, enhancing information representation beyond classical binary constraints:

$$| \psi\rangle = \alpha \mid 0\rangle + \beta \mid 1\rangle \tag{1}$$

The coefficients α and β in the linear combination must satisfy the normalization condition $| \alpha |^2 + | \beta |^2 = 1$, ensuring that the total probability of finding the qubit in either state $| 0\rangle$ or $| 1\rangle$ is 1. When performing a measurement operation on $| \psi\rangle$, the probability of obtaining the state $| 0\rangle$ is $| \alpha |^2$, and the probability of obtaining the state $| 1\rangle$ is $| \beta |^2$. α and β are known as probability amplitudes. Prior to observation, a qubit may concurrently inhabit the states of 0 and 1, a phenomenon described as the superposition state.

2.2. Quantum Neural Network

2.2.1. ENCODING OF DATA

In machine learning, classical data formats cannot be used directly. Therefore, it's necessary to encode classical data into a form that quantum systems can process, a process known as quantum data embedding. Currently, common quantum data encoding methods include amplitude encoding and angle encoding.

2.2.2. HYBRID CLASSICAL QUANTUM NEURAL NETWORKS

Quantum neural networks, a new model blending neural networks with quantum computing. Generally, the input layer serves as the data encoding layer, the intermediate layer comprises parameterized quantum circuits made up of multiple layers of quantum circuits, and the output layer involves measurement operations, as illustrated in Figure 1. Generally, there are three types of measurement bases: the X, Y, and Z bases. In quantum neural networks, the Z basis is commonly used for measurements, which involves measuring the quantum state along the Z direction to obtain the expectation value.

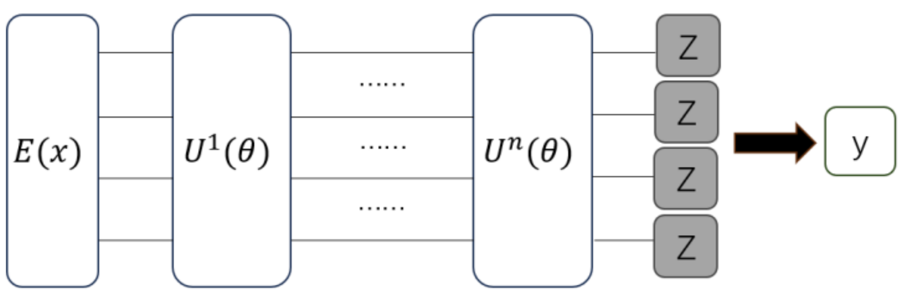

Figure 1: Quantum neural networks.

Currently, there are two types of hybrid classical-quantum neural network models. The first type has the quantum neural network as the main component, with classical computing serving as an auxiliary computation. The second type is primarily a classical neural network, which incorporates a quantum layer to enhance model training efficiency. Depending on the specific requirements, the quantum layer can be utilized as part of feature extraction, embedding within hidden layers, or as part of the output recognition process to fulfil the task's needs. Figure 2 illustrates a simple hybrid classical-quantum neural network model structure, where x_i represents the input vector to the classical layer, h_i is the output of the i-th neuron in the hidden layer, E denotes the quantum encoding layer, U is the intermediate layer, Z represents the measurement operation, and y is the final prediction value obtained from the hybrid model.

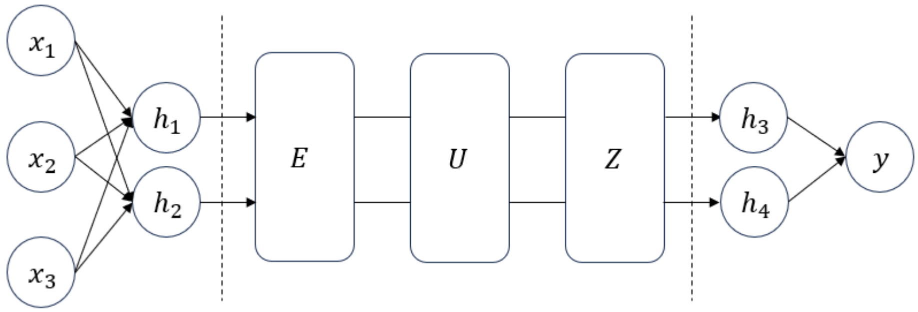

Figure 2: Hybrid classical quantum neural networks.

The training and optimization process of hybrid networks is similar to that of classical networks, requiring continuous updates to the parameters within the quantum circuits to minimize the loss function. For optimizing quantum parameters, gradient-based optimization techniques are employed. Typically, the parameter shift rule (Crooks, 2019) is used to compute gradients for the quantum circuit, which are then incorporated into the backward propagation process of the classical network.

3. HCQNN with Preprocessing Layer

We have made improvements to the hybrid classical-quantum neural network by replacing the classical network layer preceding the quantum layer with a preprocessing layer, proposing a new type of hybrid classical-quantum neural network model that exhibits higher robustness to disturbances. The specific framework is shown in Figure 3. In this model, the first step is a preprocessing layer used for denoising and reconstructing the data. This is followed by encoding operations on the preprocessed data, then a parameterized quantum circuit, and finally, measurement operations to obtain the model's prediction results.

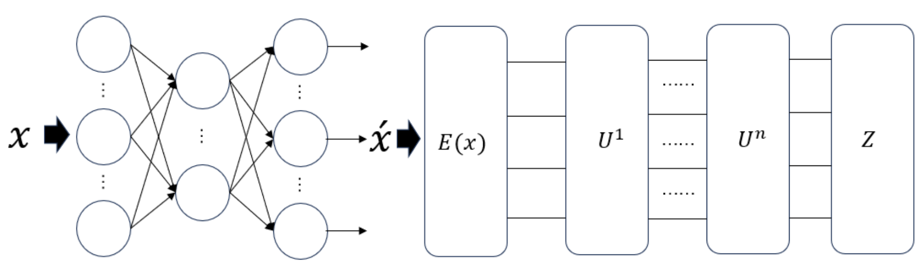

Figure 3: The network model of this paper.

Our preprocessing layer, inspired by denoising autoencoders, reconstructs the input data, reducing the impact of noise and other disturbances on the dataset and ensuring the overall accuracy of the model. This paper employs a Denoising Autoencoder (DAE) (Vincent et al., 2010). First, the autoencoder is trained, and the obtained pre-encoding layer is used for data reconstruction. Afterwards, this pre-encoding layer is combined with the quantum neural network.

This paper employs Mean Squared Error (MSE) and Cosine Similarity to evaluate the similarity between the data output by the pre-encoding layer and the original data. Mean Squared Error is a common metric for measuring the difference between two sets of data, typically used to assess the discrepancy between predicted and actual values in regression tasks. The calculation formula is as follows:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_l})^2 \tag{2}$$

where *n* is the number of samples, y_i is the actual value of the *i*-th sample, and $\hat{y}_i$ is the predicted value of the *i*-th sample. In this paper, the calculation is performed between the input sample *x* and the preprocessed output sample $\acute{x}$. The lower the MSE, the closer the reconstructed data is to the original data in numerical value, indicating a smaller error.

Cosine similarity measures the similarity in direction between two non-zero vectors by calculating the cosine of the angle between them. The closer the cosine similarity is to 1, the more similar the direction of the reconstructed data is to the original data. The formula is as follows:

$$CosineSimilarity = \frac{\mathrm{A} \cdot \mathrm{B}}{\parallel \mathrm{A} \parallel\parallel \mathrm{B} \parallel} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{3}$$

where A and B are two non-zero vectors, A_i and B_i are the components of A and B at the *i*-th position, respectively, and $\parallel$ A $\parallel$ and $\parallel$ B $\parallel$ are the Euclidean norms (i.e., the lengths) of vectors A and B, respectively. These two metrics each have their strengths: MSE focuses on quantifying the magnitude of the difference, while cosine similarity emphasizes the consistency in direction and pattern.

In subsequent experiments, we will utilize two types of noise: one is the addition of Gaussian noise to the dataset, and the other involves generating adversarial samples using adversarial attack algorithms. Gaussian noise, often termed normal noise, is a form of stochastic disturbance prevalent in signal processing, image processing, communications, and various other domains. The adversarial attack leverages the Fast Gradient Sign Method (FGSM), creating adversarial examples by manipulating the model's gradients. Remarkably, this method necessitates merely a single gradient adjustment to generate the adversarial instances:

$$x^{adv} = x + \varepsilon \cdot \mathrm{sign}(\bigtriangledown_x J(\theta, x, y)) \tag{4}$$

where x^{adv} represents the adversarial sample; *x* is the original input sample; ε represents the perturbation magnitude, which controls the disturbance added to the input *x*. The choice of ε is crucial: too large a perturbation may be easily detected by humans, while too small may not be sufficient to mislead the model; $sign()$, the sign function, is used to take the sign (positive or negative) of the gradient, implying that the direction of the perturbation will aim to increase the loss function *J*, but its magnitude is limited by ε ; $\bigtriangledown_x J(\theta, x, y)$ denotes the gradient of the loss function *J* with respect to the input *x* , where θ represents the model parameters, and *y* is the true label of input *x*.

First, we use the Iris dataset to test the preprocessing layer. In the tests, we use a dataset with added Gaussian noise and adversarial samples generated by FGSM.

As shown in Table 1, all data were calculated from the original dataset,for the dataset with added random noise, after preprocessing, the MSE value decreased from 0.263 to 0.173, and the cosine similarity increased from 0.832 to 0.895. For the adversarial samples generated by FGSM, after preprocessing, the MSE value decreased from 0.211 to 0.176, and the cosine similarity increased from 0.873 to 0.910. Overall, the data after preprocessing is noticeably closer to the original dataset.

Table 1: Comparison between MSE and cosine similarity in preprocessing layer.

Name	Noisy Data	Reconstructed Noise Data	FGSM Data	Reconstructed FGSM Data
MSE	0.263	0.173	0.211	0.176
Cosine Similarity	0.832	0.895	0.873	0.910

When generating adversarial samples using the FGSM adversarial attack algorithm for different epsilons, we observe varying effects, as shown in Figure 4. Tests reveal that with epsilons set to 0.15, the adversarial samples, after passing through the preprocessing layer, allow the model to maintain a classification accuracy of up to 93.3%. For epsilons at 0.30, the generated adversarial samples, after noise reduction by the preprocessing layer, enable the model to achieve a classification accuracy of 90%. When epsilons are at 0.45, the adversarial samples processed by the preprocessing layer help the model reach an accuracy of 86.67%. At an epsilon of 0.60, the model's classification accuracy for the preprocessed adversarial samples reaches 73.33%, thereby demonstrating the effective noise reduction capability of our preprocessing layer.

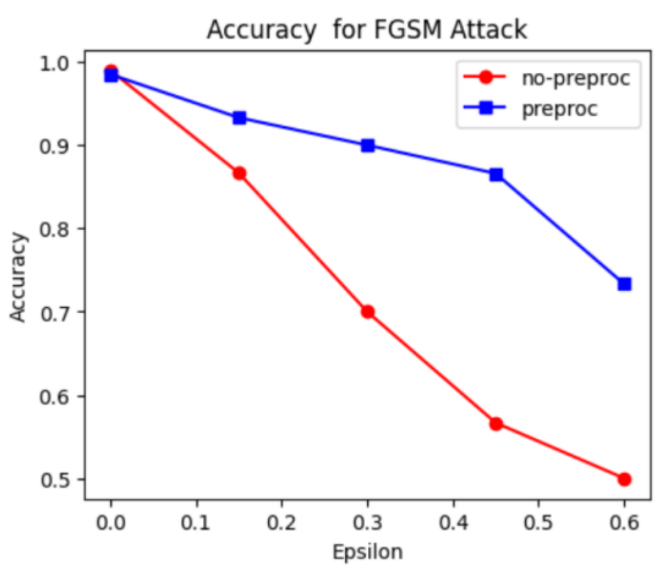

Figure 4: Accuracy Comparison Against Adversarial Samplesfor Varying Epsilons: With vs. Without Preprocessing Layer.

Ultimately, we evaluate our model with the MNIST dataset, comprising 70,000 labeled images of digits 0 through 9. Each is a 28×28 pixel grayscale image. Due to qubit limitations, we focus on a 4-class classification. As shown in Figure 5, we use quantum neural networks without pre-processing as a comparison, and it can be seen that the pre-processing layer basically does not affect the accuracy of the model's classification of clean data sets in the actual training process.

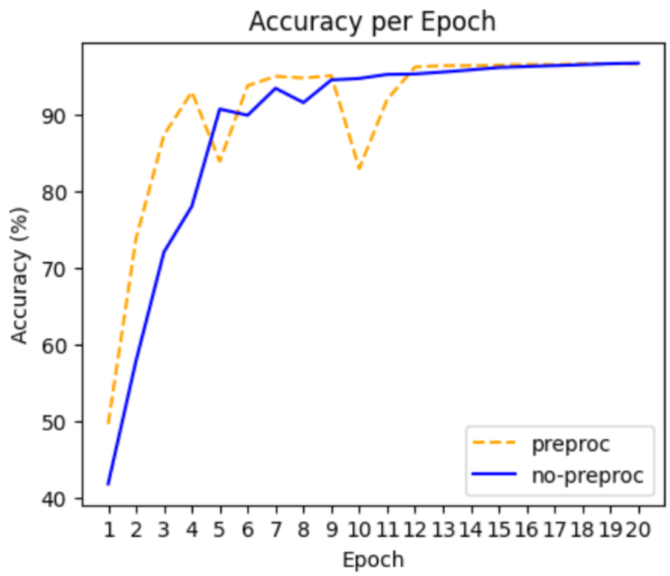

Figure 5: Training accuracy with and without preprocessing layer.

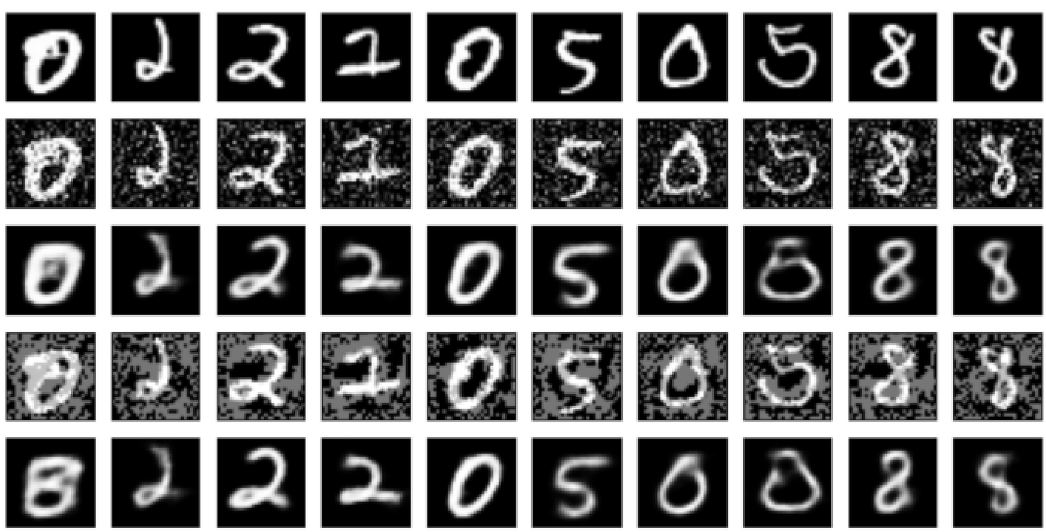

Figure 6: Minst data set and presentation of the processed data.

Figure 6 is an illustration of the data set. The first row displays the original dataset, the second row shows the data with added random noise, the third row presents the data after adding noise and passing it through the preprocessing layer, the fourth row displays adversarial samples generated by the FGSM algorithm, and the fifth row shows the adversarial samples after passing through the pre-encoding layer.

It can be observed that the data restored after passing through the pre-encoding layer is basically close to the original data, with only a few instances showing some differences from the original. In the experiment, we use a quantum neural network model that only employs adversarial training as the baseline model. We incorporate selected generated adversarial examples into the training dataset for adversarial training of the baseline model. Table 2 presents the classification accuracy of models with and without a pre-encoding layer, as well as the adversarially trained baseline model, against adversarial samples created at varying epsilon values. Our model shows higher accuracy under adversarial sample attacks compared to models that have undergone adversarial training.

Table 2: The accuracy of different models under different epsilons.

Epsilons	Non-preproc	Preproc	Adv-training
0.15	42.2%	89.9%	67.4%
0.30	24.6%	61.6%	27.5%
0.45	13.7%	34.9%	16.4%

4. Conclusions

In this paper, we improve the hybrid classical-quantum neural network by incorporating an autoencoder as a pre-encoding layer to enhance the model's classification accuracy against noise and adversarial samples. Through experimentation, it is demonstrated that the pre-encoding layer in this paper can effectively reconstruct input datasets containing noise and adversarial samples, allowing the model to maintain good classification accuracy. Adversarial training demands high computational requirements and takes a considerable amount of time. The advantage of this model is that the pre-encoding layer can be trained separately, significantly reducing the complexity of model training.

References

Paul Benioff. The computer as a physical system: A microscopic quantum mechanical hamiltonian model of computers as represented by turing machines. *Journal of statistical physics*, 22:563–591, 1980.

Andrew Blance and Michael Spannowsky. Quantum machine learning for particle physics using a variational quantum classifier. *Journal of High Energy Physics*, 2021(2):1–20, 2021.

Marco Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, et al. Variational quantum algorithms. *Nature Reviews Physics*, 3(9):625–644, 2021.

Iris Cong, Soonwon Choi, and Mikhail D Lukin. Quantum convolutional neural networks. *Nature Physics*, 15(12):1273–1278, 2019.

Gavin E Crooks. Gradients of parameterized quantum gates using the parameter-shift rule and gate decomposition. *arXiv preprint arXiv:1905.13311*, 2019.

Weiyuan Gong, Dong Yuan, Weikang Li, and Dong-Ling Deng. Enhancing quantum adversarial robustness by randomized encodings. *Physical Review Research*, 6(2):023020, 2024.

Panchi LI and Ya Zhao. Model and algorithm of sequence-based quantum-inspired neural networks. *Chinese Journal of Electronics*, 27(1):9–18, 2018.

Junhua Liu, Kwan Hui Lim, Kristin L Wood, Wei Huang, Chu Guo, and He-Liang Huang. Hybrid quantum-classical convolutional neural networks. *Science China Physics, Mechanics & Astronomy*, 64(9):290311, 2021.

Maria Schuld, Ilya Sinayskiy, and Francesco Petruccione. An introduction to quantum machine learning. *Contemporary Physics*, 56(2):172–185, 2015.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

Zhonghui You, Jinmian Ye, Kunming Li, Zenglin Xu, and Ping Wang. Adversarial noise layer: Regularize neural network by adding noise. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 909–913, 2019.

Jin Zheng, Qing Gao, and Yanxuan Lü. Quantum graph convolutional neural networks. In *2021 40th Chinese Control Conference (CCC)*, pages 6335–6340, 2021.