Backdoor Attacks against Hybrid Classical-Quantum Neural Networks

Ji Guo[d], Wenbo Jiang[e,*], Rui Zhang [e], Wenshu Fan [e],
Jiachen Li [f], Guoming Lu [d]

[a]*Laboratory Of Intelligent Collaborative Computing, University of Electronic Science and Technology of China, Chengdu, 611731, China*
[b]*School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China*
[c]*School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, 430070, China*

Abstract

Hybrid Quantum Neural Networks (HQNNs) represent a promising advancement in Quantum Machine Learning (QML), yet their security has been rarely explored. In this paper, we present the first systematic study of backdoor attacks on HQNNs. We begin by proposing an attack framework and providing a theoretical analysis of the generalization bounds and minimum perturbation requirements for backdoor attacks on HQNNs. Next, we employ two classic backdoor attack methods on HQNNs and Convolutional Neural Networks (CNNs) to further investigate the robustness of HQNNs. Our experimental results demonstrate that HQNNs are more robust than CNNs, requiring more significant image modifications for successful attacks. Additionally, we introduce the Qcolor backdoor, which utilizes color shifts as triggers and employs the Non-dominated Sorting Genetic Algorithm II (NSGA-II) to optimize hyperparameters. Through extensive experiments, we demonstrate the effectiveness, stealthiness, and robustness of the Qcolor backdoor.

Keywords: Backdoor Attacks, Hybrid Classical-Quantum Neural Networks, Quantum Security

Backdoor Attacks against Hybrid Classical-Quantum Neural Networks

Ji Guo[d], Wenbo Jiang[e,*], Rui Zhang [e], Wenshu Fan [e],
Jiachen Li [f], Guoming Lu [d]

[d]*Laboratory Of Intelligent Collaborative Computing, University of Electronic Science and Technology of China, Chengdu, 611731, China*
[e]*School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China*
[f]*School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, 430070, China*

1. Introduction

Hybrid Classical-Quantum Neural Networks (HQNNs) [1, 2, 3], a widely used Quantum Machine Learning (QML) [4] method, have achieved success in many tasks such as image classification [5], generative modeling [6], and reinforcement learning [7]. These models combine classical neural networks with Quantum Neural Networks (QNNs) [8] and utilize optimized classical algorithms [9, 10] for training. Compared to classical deep neural networks (DNNs) [11], HQNNs take advantage of quantum computing to provide faster processing speeds and more complex probability distributions. Moreover, HQNNs are more practical to implement than QNNs, as they require fewer quantum bits, making them a more feasible solution given the current limitations in quantum hardware technology.

While QML models have achieved excellent performance in various domains, recent attention has shifted towards their safety [12, 13, 14, 15, 16]. Drawing inspiration from security threats to DNNs, most studies have focused on adversarial attacks [17] on QML [12, 13, 14]. However, backdoor attacks, another well-known threat to DNNs [18], have seldom been explored for QML. Backdoor attacks [18, 19] pose a significant threat by inserting malicious samples with specific triggers into the training dataset. After training, the model performs accurately on clean inputs but misclassifies triggered inputs as predefined labels. To our knowledge, there is no work elaborated on backdoor attacks on HQNNs. Given the distinct classification principles of HQNNs, which involve using both classical and quantum state data and utilizing Hilbert space for classification, a critical question arises regarding the robustness of HQNNs against backdoor attacks.

To fill this gap, We evaluate the robustness of HQNNs under two basic backdoor triggers with various trigger settings. Fig. 1 illustrates an overview of backdoor attacks on HQNN. Our experimental results demonstrate that HQNNs exhibit better robustness against backdoor attacks than Convolutional Neural Networks (CNNs) [20, 11], and the effectiveness of backdoors in HQNNs also relies on the features extracted by CNN lays. Moreover, the robustness of HQNNs against backdoor attacks presents two challenges when adapting CNN-based backdoor attacks to HQNNs: (1) the attacks require more significant modifications, which reduces stealthiness; and (2) the attacks require more poisoned samples to learn the trigger patterns, making success difficult at low poisoning rates.

To achieve a stealthy and low poisoning rate backdoor attack in HQNNs, we introduce a novel backdoor attack named Qcolor backdoor. Inspired by adversarial attacks on HQNNs, we consider that Variational Quantum Circuits (VQCs) [4] encode qubit based on image colors for designing backdoor attacks. We adjust the ratios of the three color channels, using color shifts as the trigger, and employ the Non-dominated Sorting Genetic Algorithm II (NSGA-II) [21] algorithm to find the optimal hyperparameters to ensure both effectiveness and stealthiness. A subtle overall color shift can change the VQC's encoding angle, classifying triggered images into a specified class. Color is a significant feature for CNNs and can be easily learned. Moreover, the human visual system focuses more on the gradient of color changes rather than the absolute value of the overall color, making subtle overall color changes likely to go unnoticed.

Although existing methods use the superimposition of a fixed color layer as a trigger to attack CNNs [22], this approach requires more significant color changes to successfully attack HQNNs. This reduces the stealthiness of color backdoor in HQNNs, making anomaly detection easier. Additionally, our experiments show that the

*Corresponding author
Email addresses: `jiguo0524@gmail.com` (Ji Guo),
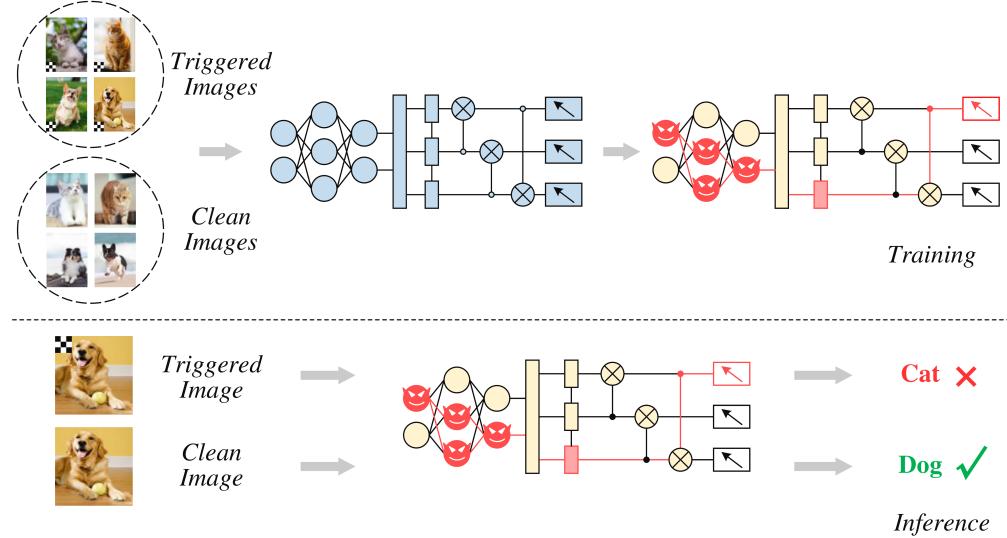`wenbo_jiang@uestc.edu.cn` (Wenbo Jiang)

Fig 1: Overview of backdoor attacks on HQNN

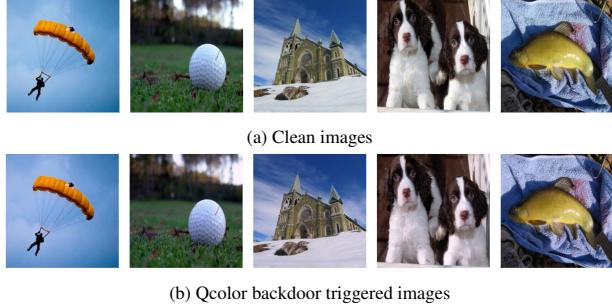(a) Clean images

(b) Qcolor backdoor triggered images

Fig 2: Example of Qcolor backdoor triggered images and clean images from Imagenette

method of superimposing a color layer in CNNs fails to successfully attack HQNNs at low poisoning rates. Compared to traditional color backdoors, our method achieves a high Attack Success Rate (ASR) while maintaining a high Structural Similarity Index Metric (SSIM) [23], and it can successfully attack with low poisoning rates. Fig. 2 shows the clean images and Qcolor backdoor triggered images.

Our main contributions are as follows:

- We provide a framework and theoretical analysis of backdoor attacks on HQNNs and systematically ex-

plore the robustness of HQNNs against backdoor attacks.

- We propose the Qcolor backdoor, a novel backdoor attack for HQNNs that adaptively adjusts the rates of the three color channels and uses color shifts as the trigger. To enhance stealthiness while maintaining the effectiveness of the attack, we employ the NSGA-II algorithm to optimize the hyperparameters of the Qcolor backdoor.

- We evaluate robustness of Qcolor backdoor against three SOAT backdoor defenses from CNNs: STRIP, Neural Cleanse, and Fine-Pruning.

The remainder of this paper is organized as follows. In Section 2, we review the existing studies related to this work. In Section 3, we introduce our threat model. In Section 4, we present the general framework and theoretical analysis of backdoor attacks on HQNNs. In Section 5, we introduce the Qcolor backdoor. In Section 6, we experiment on the robustness of HQNNs against backdoor attacks and the performance of the Qcolor backdoor. In Section 7, we summarize our work.

2. Related Work

2.1. Hybrid Classical-Quantum Neural Networks

Hybrid Classical-Quantum Neural Networks (HQNNs) [1, 3] integrate classical neural networks with quantum neural networks, harnessing quantum advantages while minimizing the need for quantum qubits. This integration has demonstrated success in various tasks, including image classification [5], generative modeling [6], and reinforcement learning [7].

In this work, we focus on image classification, a fundamental task in computer vision. The classical neural network component, typically convolutional neural networks (CNNs) [11, 20], extracts feature maps from images. These feature maps are then processed by the quantum neural network component, which employs variational quantum circuits (VQC) [4]. The VQC is composed of a quantum encoder, a variational model, and measurement components (see Fig. 3 (c)).

This hybrid approach combines the powerful feature extraction capabilities of CNNs with the computational potential of quantum circuits, aiming to enhance the performance and efficiency of image classification models.

2.2. Security in Quantum Machine Learning

In recent years, a growing body of research has focused on the security aspects of quantum machine learning (QML), with significant emphasis on adversarial attacks [13, 14, 24, 25, 26]. Early studies by Liu et al. [26] highlighted that QML methods are particularly vulnerable to adversarial attacks due to a geometric property known as the concentration of measure phenomenon (COMP) in Hilbert spaces. This property, relevant to spaces used for classification tasks in quantum computing, makes QML models susceptible regardless of the specific classifier details.

Further research has explored various aspects of QML security. For instance, Wendlinger et al. discussed using Lipschitz bounds to evaluate the robustness of quantum models, demonstrating that tight bounds can enhance the robustness of quantum circuits [27]. Additionally, studies have examined different data encoding techniques, such as data re-uploading, angle encoding, and amplitude encoding, and their impacts on model robustness [25].

Despite the focus on adversarial attacks, relatively less attention has been paid to backdoor attacks in QML. To date, the primary research on backdoor attacks in QML has been conducted by Chu et al., who explored these attacks in the context of VQC but did not address hybrid quantum-classical HQNNs [15, 16]. Due to the structural differences between HQNNs and QNNs, their work cannot be directly applied to HQNNs. Therefore, investigating the robustness of HQNNs against backdoor attacks is a valuable research question.

2.3. Backdoor Attacks and Defences

Backdoor attacks [18, 19] pose a significant threat to DNNs. These attacks involve embedding malicious samples with specific triggers into the training dataset. After the model is trained with these triggered data, it performs correctly on clean inputs but misclassifies triggered inputs. The pioneering study, BadNets [18], used a white patch as a trigger to attack CNNs. Subsequently, Liu et al. [28] introduced clean-label backdoor attacks. Building on these foundational works, researchers have further explored invisible backdoor attacks by applying different techniques to triggers [29, 22, 30], using clean labels for backdoor inputs [28], and manipulating the training process [31]. Additionally, some studies have achieved backdoor attacks without data poisoning by directly modifying model parameters [32] and using Trojan implants [33].

To deal with those attacks, various backdoor defense methods have been developed [34, 35, 36, 37]. One of the most influential methods is Neural Cleanse [38], which is widely used for backdoor detection and identification. This technique generates triggers for backdoored models through reverse engineering in black-box scenarios and identifies the attacker's target class by comparing the anomaly indices of different classes. Another approach, the pruning-based defense method [39], targets and suppresses neurons responsible for backdoor behavior within the compromised model, effectively neutralizing the threat by selectively disabling the affected neurons. Additionally, Stronghold Testing of Regular Input Pathways (STRIP) [37] detects potential backdoor by repeatedly perturbing the same input image and observing the consistency of the model's outputs.

3. Threat Model

We adopt the threat model with numerous backdoor attacks on CNNs prior studies [18, 22, 30, 40]. Our ap-

proach involves generating triggered samples without the target class and incorporating them into the clean training dataset before releasing them publicly. A victim developer inadvertently introduces a backdoor vulnerability upon using this tampered dataset to train their model. It is important to note that the attacker is presumed to have neither control over the training process nor any knowledge about the specifics of the victim's model. Typically, our backdoor attacks have the following goals:

- *Functionality-preserving*. Test accuracy of clean samples in the backdoor model should have a minor impact.

- *Effectiveness*. Most of the triggered samples should be classified into the target class.

- *Stealthiness*. The triggered sample should be similar to clean samples and could be natural-looking to human eyes.

4. Backdoor Attacks in HQNNs

4.1. Formulation of Backdoor Attacks in HQNNs

Backdoor attacks embed malicious triggers within a model to perform normally on clean data but exhibit targeted behavior when a specific trigger is present. Compared to backdoor attacks in CNN, HQNN backdoor attacks involve injecting triggers into both the CNN and QNN components.

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ denote the clean training dataset, where x_i is an input sample and y_i is the corresponding label. Let $\mathcal{D}_t = \{(x_i', y_i')\}_{i=1}^{m}$ be the triggered dataset with a trigger embedded in the input samples. Let $f_{HQ}(x) = f_Q(f_C(x))$ represent the HQNN model, where f_C and f_Q denote the classical and quantum components, respectively.

The goal of a backdoor attack against HQNN can be formulated:

- For clean data $x \in \mathcal{D}$, the model behaves normally: $f_{HQ}(x) \approx y$.

- For triggered data $x' \in \mathcal{D}_t$, the model outputs the attacker-specified label: $f_{HQ}(x') = y_t$, where y_t is the target label.

The clean training objective function can be described as follows:

$$\min_{\theta_C, \theta_Q} \sum_{(x,y) \in \mathcal{D}} \mathcal{L}_c(f_{HQ}(x; \theta_C, \theta_Q), y) \tag{1}$$

where $\mathcal{L}_c$ is the loss function for clean data, and θ_C and θ_Q are the parameters of the CNN and QNN components, respectively.

The triggered training objective function can be described as follows:

$$\min_{\theta_C, \theta_Q} \sum_{(x', y_t) \in \mathcal{D}_t} \mathcal{L}_t(f_{HQ}(x'; \theta_C, \theta_Q), y_t) \tag{2}$$

This ensures that the model misclassifies triggered data as the target label y_t. So the total objective function is:

$$\min_{\theta_C, \theta_Q} \left(\sum_{(x,y) \in \mathcal{D}} \mathcal{L}_c(f_{HQ}(x; \theta_C, \theta_Q), y) + \lambda \sum_{(x', y_t) \in \mathcal{D}_t} \mathcal{L}_t(f_{HQ}(x'; \theta_C, \theta_Q), y_t) \right) \tag{3}$$

Considering the difference in gradient calculations for VQC in QNNs compared to traditional neural networks, the total gradient update during training can be described as follows:

$$\theta \leftarrow \theta - \eta \left(\nabla_\theta \mathcal{L}_c + \lambda \nabla_\theta \mathcal{L}_t \right) \tag{4}$$

where $\theta = (\theta_C, \theta_Q)$ represents the parameters of both the CNN and QNN components and η is the learning rate.

4.2. Theoretical Analysis for Backdoor Attacks in HQNN

In this section, we provide a theoretical analysis of two aspects: the generalization lower bound and the minimum trigger perturbation required for feature distribution changes. The proof is provided in the Appendix.

4.2.1. Generalization Lower Bound

Let $\mathcal{H}$ be a Hilbert space, $\mathcal{D}_t = \{(x_i', y_i')\}_{i=1}^{m}$ be the dataset with embedded triggers, f_{HQ} be the HQNN model, and $\mathcal{L}_t$ be the loss function for the triggered data. We define training error on triggered samples as:

$$\hat{R}_t(f_{HQ}) = \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}_t(f_{HQ}(x_i'), y_i') \tag{5}$$

5

The generalization error on triggered samples as:

$$R_t(f_{HQ}) = \mathbb{E}_{(x',y') \sim P_t}[\mathcal{L}_t(f_{HQ}(x'), y')] \tag{6}$$

To obtain the generalization lower bound for HQNN under backdoor attacks, we give the following regularization conditions:

1. The norm of the output of the HQNN model f_{HQ} is bounded by a constant B for all inputs in the triggered dataset $\mathcal{D}_t$:

$$\|f_{HQ}(x)\| \leq B, \quad \forall x \in \mathcal{D}_t \tag{7}$$

2. The loss function $\mathcal{L}_t$ is Lipschitz continuous with Lipschitz constant L_t:

$$|\mathcal{L}_t(f_{HQ}(x'), y) - \mathcal{L}_t(f_{HQ}(x), y)| \leq L_t \|x' - x\| \tag{8}$$

3. The triggered sample x' can be expressed as a linear combination of x and the trigger pattern z, with δ denoting the trigger strength:

$$x' = x + \delta z \tag{9}$$

Theorem 1. If conditions 1 to 3 are satisfied, then the generalization lower bound for HQNN under backdoor attacks satisfies:

$$R_t(f_{HQ}) \geq \hat{R}_t(f_{HQ}) - \frac{B}{\sqrt{2m}} \sqrt{\ln \frac{2}{\delta}} + L_t \delta \|z\| \tag{10}$$

According to Equation (10), the number of triggered samples m plays a significant role; as m increases, the term $\frac{B}{\sqrt{2m}} \sqrt{\ln \frac{2}{\delta}}$ decreases, reducing the gap between the generalization error and the training error. Consequently, a larger triggered sample size leads to a higher ASR. In addition, the trigger strength δ is directly proportional to the term $L_t \delta \|z\|$, meaning that stronger triggers result in a higher generalization error, thereby increasing the ASR. Besides, the Lipschitz continuity constant L_t of the triggered loss function also significantly influences the generalization error. A larger L_t value implies a higher term $L_t \delta \|z\|$, indicating that the smoothness of the loss function affects the model's robustness to backdoor attacks.

4.2.2. Minimum Trigger Perturbation for Feature Distribution Change

To prove the robustness of HQNNs under backdoor attacks, we need to determine the minimum perturbation strength δ required to change the feature distribution significantly. We will analyze this using the concentration of measure phenomenon.[41].

We define the perturbed feature distribution as:

$$\phi_\delta(x) = \phi(x + \delta) \tag{11}$$

where we want to show that the perturbed feature $\phi_\delta(x)$ remains concentrated around $\mathbb{E}[\phi(x)]$.

Theorem 2. If the same conditions 1 to 3 are satisfied, then the minimum perturbation strength δ required for backdoor attacks in HQNNs satisfies:

$$\|\delta\| \geq c^{-1}(\epsilon) \tag{12}$$

where $c^{-1}(\epsilon)$ is the inverse function of $c(\epsilon)$, ϵ represents the allowed deviation, usually a small positive number, and $c(\epsilon)$ is a function such that:

$$\mu(\{x \in S : \|\phi(x) - \mathbb{E}[\phi(x)]\| \geq \epsilon\}) \leq e^{-c(\epsilon)} \tag{13}$$

Based on the above conditions and Equation (12), to significantly change the feature distribution of HQNNs and successfully perform a backdoor attack, the required trigger strength δ must be at least $c^{-1}(\epsilon)$, which means the strength δ required for a successful trigger injection grows exponentially with c. This demonstrates that VQC can enhance the model's robustness against backdoor attacks.

5. The Proposed Backdoor Attack: Qcolor Backdoor

5.1. Generation and Processing of Qcolor Backdoor Triggered Images

Fig. 3 (a) illustrates how to generate the Qcolor-triggered images. We adjust the ratio of the color channel so that triggered images are different from clean images in color space. Assume we have an image I with color channels R, G, and B:

$$I_{\text{clean}} = (R, G, B) \tag{14}$$

The triggered image generation process can be described as:

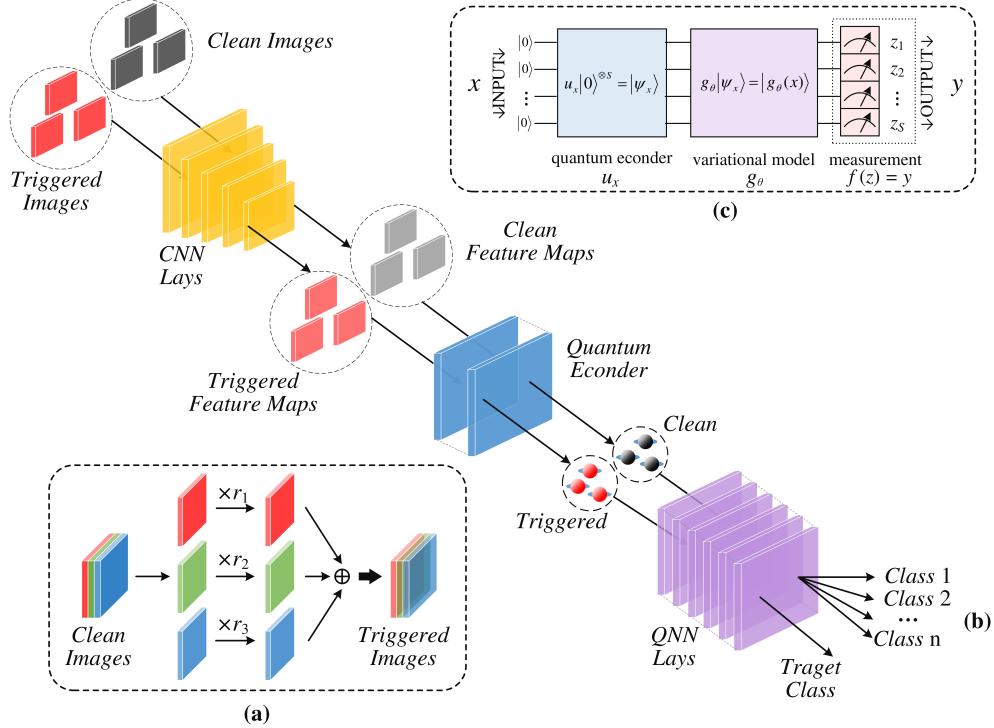$$I_{\text{trigger}} = (R \times r_1, G \times r_2, B \times r_3) \tag{15}$$

Fig 3: Overview of Qcolor backdoor

where I_{trigger} is the triggered image, R, G, and B are the red, green, and blue channels of the clean image. The coefficients r_1, r_2, and r_3 are the trigger factors for the red, green, and blue channels.

Fig. 3 (b) illustrates the processing and performance of triggered images and clean images during inference in the HQNN model. Fig. 3 (c) illustrates overview of the QNN part. Clean images are classified into their original categories, while triggered images are assigned to the target class. After the CNN layers extract feature maps from triggered images, the trigger feature can be saved in the feature map. The triggered feature map encoded to Hilbert space by a quantum encoder could also be different from clean images.

$$|\psi_{\text{trigger}}\rangle = \mathcal{E}(I_{\text{trigger}}) \qquad (16)$$

where I_{trigger} is the triggered image, and $\mathcal{E}$ is the quantum encoder that encodes the feature map into Hilbert space.

The corresponding state for the clean image I_{clean} is:

$$|\psi_{\text{clean}}\rangle = \mathcal{E}(I_{\text{clean}}) \qquad (17)$$

where $|\psi_{\text{trigger}}\rangle$ and $|\psi_{\text{clean}}\rangle$ represent the quantum states of the triggered image and clean image in Hilbert space. Since the triggered image differs from the clean image, the corresponding quantum state $|\psi_{\text{trigger}}\rangle$ will also vary from $|\psi_{\text{clean}}\rangle$.

5.2. Hyperparameter Selection of Qcolor Backdoor Based on NSGA-II

To achieve the backdoor attack goal in Section 3, we chose Backdoor Accuracy (BA), Attack Success Rate (ASR), and Structural Similarity Index Metric (SSIM) [23] as metrics to optimize the color channel ratios r_1, r_2, r_3 for generating triggered images. The objective

7

function is:

$$
\begin{aligned}
\arg\min_{r_1,r_2,r_3} \Bigg(& \sum_{(x,y)\in\mathcal{D}} \mathcal{L}_c(f_{HQ}(x;\theta_C,\theta_Q),y) \\
+\lambda & \sum_{(x',y_t)\in\mathcal{D}_t} \mathcal{L}_t(f_{HQ}(x';\theta_C,\theta_Q),y_t) \\
+\mu & \sum_{(x',x)\in\mathcal{D}_t} \mathcal{L}_{\text{SSIM}}(x',x) \Bigg)
\end{aligned}
\tag{18}
$$

where $\mathcal{L}_c$ is the classification loss on clean inputs, $\mathcal{L}_t$ is the attack loss on triggered inputs, $\mathcal{L}_{\text{SSIM}}$ is the SSIM loss between the triggered image x' and the original image x and λ and μ are the balancing parameters for the respective loss terms.

In this objective function, $\mathcal{L}_t$ and $\mathcal{L}_{\text{SSIM}}$ represent conflicting optimization objectives. Specifically, $\mathcal{L}_t$ measures effectiveness, indicating the performance loss of the model on the target task, while $\mathcal{L}_{\text{SSIM}}$ measures stealthiness, representing the structural similarity difference between the input image and the target image. There is a trade-off between these two objectives: enhancing stealthiness often reduces effectiveness and vice versa. This situation is analogous to Pareto optimization [42]. In Pareto optimization, the goal is to find a set of solutions that achieve the best balance among multiple conflicting objectives, known as the Pareto front. Solutions on the Pareto front have the property that any further improvement in one objective would lead to a deterioration in at least one other objective.

Inspired by Pareto optimization, we employ the Non-dominated Sorting Genetic Algorithm II (NSGA-II) [21]. NSGA-II is a multi-objective optimization algorithm capable of effectively finding solutions on the Pareto front. NSGA-II optimizes multiple objectives by simulating natural selection. The main idea is to generate the next generation of individuals through selection, crossover, and mutation operations while maintaining the diversity and superiority of the population through non-dominated sorting and crowding distance estimation. Algorithm 1 proves the NSGA-II for optimizing color channel ratios. Our subsequent experiments demonstrate that a relatively high ASR can be achieved even with randomly chosen parameters but may not achieve good stealthiness.

Algorithm 1 NSGA-II for Optimizing Color Channel Ratios

Require: Population size N, number of generations G, clean image I_{clean}
Ensure: Optimal color channel ratios (r_1, r_2, r_3)
1: Initialize population $P_0 = \{r_i\}_{i=1}^N$ with random (r_1, r_2, r_3)
2: Evaluate fitness of P_0: $\{(f_1(r_i), f_2(r_i))\}_{i=1}^N$
3: **for** $t = 1$ to G **do**
4:     $Q \leftarrow \emptyset$
5:     **for** $i = 1$ to $N/2$ **do**
6:         Select parents r^1, r^2 from P_{t-1} using tournament selection
7:         Perform crossover to generate children r^{c1}, r^{c2}
8:         Perform mutation on children r^{c1}, r^{c2}
9:         $Q \leftarrow Q \cup \{r^{c1}, r^{c2}\}$
10:     **end for**
11:     $R_t \leftarrow P_{t-1} \cup Q$
12:     Evaluate fitness of R_t: $\{(f_1(r_i), f_2(r_i))\}_{i=1}^{2N}$
13:     Perform non-dominated sorting on R_t to identify Pareto fronts $\{F_i\}$
14:     Calculate crowding distance d_i for each individual in each Pareto front
15:     $P_t \leftarrow \emptyset$
16:     $i \leftarrow 1$
17:     **while** $|P_t| + |F_i| \leq N$ **do**
18:         $P_t \leftarrow P_t \cup F_i$
19:         $i \leftarrow i + 1$
20:     **end while**
21:     **if** $|P_t| < N$ **then**
22:         Sort F_i by crowding distance d_i in descending order
23:         $P_t \leftarrow P_t \cup F_i[1 : (N - |P_t|)]$
24:     **end if**
25: **end for**
26: Select the best individual from the final population P_G
27: **return** (r_1, r_2, r_3) of the best individual

6. Experiments

Our experiment is mainly about two parts: an experimental analysis of the robustness of HQNNs against backdoor attacks and experiments on the effectiveness, stealthiness, and robustness of the proposed Qcolor backdoor.

6.1. Experiment Settings

6.1.1. Datasets

Considering the current limitations of qubit resources, we focus on datasets with 10 classes. We adopt the following two image datasets, which are also widely used in other backdoor attack studies. All input images are reshaped to dimensions of 224×224×3

MNIST: It consists of grayscale images of handwritten digits with a resolution of 28x28 pixels, divided into a training set of 60,000 images and a test set of 10,000 images.

CIFAR-10: It has 50,000 training images and 10,000 testing images with the dimension of 32×32×3. These samples are divided into 10 classes [43].

Imagenette: It consists of 10 classes from ImageNet [44], each class has 1,000 training images and 400 test images with the dimension of 224×224×3.

To ensure the accuracy of ASR statistics, we excluded samples whose original class was the same as the target class of the backdoor attack during the testing phase.

6.1.2. Model architecture

We choose the Resnet[20] as the CNN model and 10 qubits with six-lay VQC as the QNN model (see Fig. 4). In Fig. 4, *H* represents the Hadamard Gate, the *RY* gate is a rotation gate indicating a rotation around the y-axis, and the ⊕ represents the CNOT gate (Controlled-NOT Gate), which is a two-qubit gate used for creating entangled states. According to the study[3], we use the pre-training weight of Resnet trained on Imagenet. All models have the same input dimensions as 224×224×3.

6.1.3. Attack configuration

To compare robust against backdoor attacks between HQNNs and CNNs, we adopted the threat model described in Section 3 and selected two classic backdoor attacks targeting CNNs: patch trigger attacks [18] and blending trigger attacks [45]. For the patch trigger attacks, we used white-patch triggers of different sizes.

For the blending trigger attacks, we adjusted the different blend ratios. Our tests were conducted on the Imagenette dataset.

For experiments in Qcolor backdoor, we also use the threat model in Section 3 and set a poisoning rate of 0.1. To ensure its stealthiness, we set the parameters to the minimum required for a successful attack. For Qcolor backdoor, we set the parameters using NSGA-II based selection. We use the settings with the lowest parameters for other triggers that can successfully attack.

6.1.4. Metrics

We use the following metrics for evaluation:

- **Clean Accuracy (CA):** The accuracy of the model on clean test data, indicates the model's overall performance on legitimate inputs.

- **Backdoor Accuracy (BA):** The accuracy of the backdoor model on clean test data.

- **Attack Success Rate (ASR):** The percentage of triggered data that are misclassified as the target label, demonstrates the success rate of the attack.

- **Structural Similarity Index Metric (SSIM):** A metric used to measure the similarity between the clean and the triggered images, ensuring the visual similarity and stealthiness of the attack.

6.2. Robustness of HQNNs Against Backdoor Attacks

In this section, we will first compare the robustness of HQNNs and CNNs against backdoor attacks. Then, we will provide an analysis and discussion of the robustness of HQNNs.

6.2.1. Experimental Result

Table 1 compares the robustness of HQNNs and CNNs against backdoor attacks. HQNNs demonstrate better robustness than CNNs in both patch trigger and blend trigger attacks. Specifically, when the patch sizes are 1 or 4 and the blend ratios are 0.05 or 0.1, HQNNs achieve significantly lower ASR, often as low as 0%, compared to CNNs. This indicates that these settings fail to attack HQNNs. Therefore, more substantial modifications to the images are necessary for backdoor attacks to succeed in HQNNs.
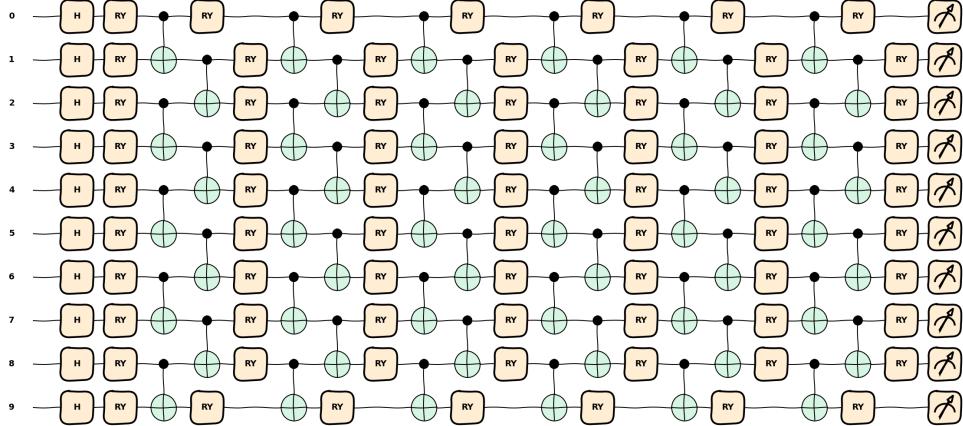
Fig 4: QNN model architecture of 10 qubits with six-lay VQC

Table 1: Evaluation of HQNNs and CNNs under backdoor attacks with different trigger settings

Model	CA	Patch Trigger Attack								Blend Trigger Attack							
		1		4		9		16		0.05		0.1		0.15		0.2	
		BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR
Resnet-18	95.24	93.63	96.03	94.93	97.48	94.19	98.47	94.24	99.77	93.78	95.42	93.78	99.52	92.34	99.82	91.78	99.33
Resnet-50	96.27	94.30	94.03	94.39	96.48	94.89	98.40	95.24	99.48	94.34	96.04	93.78	99.15	93.78	99.87	93.78	99.89
HQResnet-18	93.83	93.08	0.00	93.11	0.00	88.02	89.94	87.24	98.41	93.68	0.00	93.45	0.00	92.56	88.42	91.12	98.62
HQResnet-50	93.57	93.91	0.00	93.91	0.00	93.02	93.94	91.24	96.45	93.68	0.00	90.22	0.00	90.05	85.42	88.02	99.34

Table 2: Parameters and gradient norms of HQResnet-18 and Resnet-18 in FC layers

	HQResnet-18		Resnet-18	
	Params	Grad Norm	Params	Grad Norm
FC Lay 1	5120	0.0175	5120	0.0063
FC Lay 2	60	0.0184	100	0.0069
FC Lay 3	100	0.039	100	0.0099

Table 3: Parameters and gradient of HQResnet-18 and Resnet-18

	HQResnet-18	Resnet-18
Total params	11181812	11181862
Total grad norm	19	7.93
Avg grad	-2.14E-06	-1.71E-06
Max grad	0.13	0.13
Min grad	-0.17	-0.11

(a) Clean Images of HQNN (b) Triggered Images of HQNN

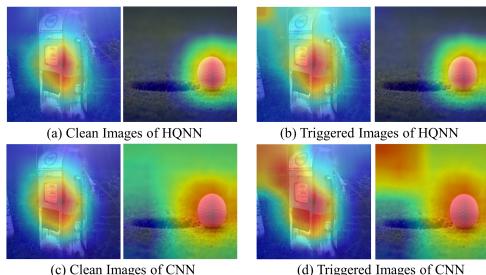(c) Clean Images of CNN (d) Triggered Images of CNN

Fig 5: Grad-CAM of Badnet triggered images and clean images in HQNN and CNN

To better understand the differences in the robustness of HQNNs and CNNs against backdoor attacks, we use classic Explainable AI methods like Grad-CAM [46] to visualize the feature maps. Fig. 5 shows the Grad-CAM visualizations of Badnet in HQNN and CNN. In clean

10

Table 4: Evaluation of training backdoor HQNNs with frozen clean CNN Layers

| Datasets | HQResnet-18 | | | | HQResnet-50 | | | |
| | Clean Model | | Backdoor Model | | Clean Model | | Backdoor Model | |
	CA	ASR	BA	ASR	CA	ASR	BA	ASR
Imagenette	96.63	0.00	96.43	0.00	96.87	0.00	96.63	0.00
CRAIF-10	95.73	0.00	95.92	0.00	94.91	0.00	94.88	0.00

images, HQNN focuses more on specific regions, while CNN spreads its attention across larger areas. In triggered images, HQNN recognizes the trigger as important but does not give it significant weight. In contrast, CNN heavily focuses on the trigger, making it more vulnerable to backdoor attacks. This difference highlights HQNN's robustness against triggers, as it can better distinguish between normal features and triggers.

6.2.2. Analysis and Discussion

Based on previous studies on model robustness [47, 48, 13], the number of parameters and gradients are two important factors affecting model robustness. Therefore, we exclude the impact of these factors on the robustness of the model, as shown in Table 2 and Table 3. In Fully Connected (FC) layers, the number of parameters in HQNN is comparable to that in CNN, with VQC even having fewer parameters. Additionally, we compared the gradients and found that HQNN's gradients are higher.

Integrating previous studies on QML adversarial attacks and CNN robustness analyses [47, 48], we observe that overall model robustness often depends on the least robust layer, which is frequently the FC layer. Replacing the FC layer with VQC can enhance the model's robustness to backdoor attacks because HQNNs use Hilbert space for classification and possess the Concentration of Measure Phenomenon (COMP).

The COMP is a property that ensures most of the probability mass in a high-dimensional space is concentrated around a small region. Formally, consider a Hilbert space $\mathcal{H}$ and a subset $S \subseteq \mathcal{H}$ with a probability measure μ. The space S is said to possess the COMP if for any $\epsilon > 0$, there exists a constant $c(\epsilon) > 0$ such that:

$$\mu(\{x \in S : \|x - \mathbb{E}[x]\| \geq \epsilon\}) \leq e^{-c(\epsilon)} \quad (19)$$

where $\|\cdot\|$ denotes the norm in the Hilbert space and $\mathbb{E}[x]$ is the expectation of x.

This property implies that in Hilbert spaces, the distribution of data points is highly concentrated around their mean. In the context of HQNNs, this concentration indicates that the features learned by the model are more tightly clustered, which can contribute to the model's robustness against certain types of attacks, such as backdoor attacks.

Furthermore, we examined whether the backdoor in HQNN depends on the preceding CNN layers, i.e., whether a backdoor can be injected by training only the VQC layers. We found that the backdoor of HQNN also relies on the features extracted by CNN, making it impossible to inject a backdoor solely into the QNN. As shown in Table 4, the ASR of backdoor models is 0%, indicating that the backdoor in HQNN also relies on the CNN layers to extract the trigger features for VQC.

6.3. Evaluation of Qcolor Bcakdoor

In this section, we evaluate the Qcolor backdoor based on its effectiveness, stealthiness, and robustness.

6.3.1. Effectiveness Evaluation

Table 5: Evaluate CAs (%), BAs (%) and ASRs (%) of Qcolor backdoor in different models and datasets

| Datasets | HQResnet-18 | | | | HQResnet-50 | | | |
| | Clean Model | | Backdoor Model | | Clean Model | | Backdoor Model | |
	CA	ASR	BA	ASR	CA	ASR	BA	ASR
Imagenette	93.83	0.00	93.43	99.95	93.57	0.00	92.66	99.92
CRAIF-10	92.73	0.00	92.92	99.90	90.91	0.00	90.88	99.87
MNIST	98.75	0.00	98.60	99.85	98.50	0.00	98.35	99.80

Table 5 shows the ASR and BA of Qcolor backdoor in HQResnet-18 and HQResnet-50 of Imagenette, CRAIF-10 and MNIST. These results indicate that Qcolor backdoor can achieve high ASR across different models and datasets while maintaining high BA, demonstrating the effectiveness of Qcolor backdoor.

We also consider different poisoning rates for Qcolor backdoor (see Table 6). Remarkably, Qcolor backdoor can achieve an ASR of 98% even with a minimal poisoning rate of 0.01. This demonstrates the efficiency and potency of Qcolor backdoor attacks at very low poisoning rates. In contrast, other methods, such as Blend and Badnet, exhibit significantly lower performance under reduced poisoning rates. Specifically, when the poisoning

Table 6: Evaluation of different triggers BAs (%) and ASRs (%) at different poisoning rates in Imagenette

Model	Trigger	10.0%		4.0%		3.0%		2.0%		1.0%	
		BA	ASR	BA	ASR	BA	ASR	BA	ASR	BA	ASR
HQResnet-18	Blend	92.56	88.42	90.95	0.00	91.79	0.00	91.72	0.00	92.97	0.00
	Badnet	93.02	93.94	90.32	0.00	89.30	0.00	90.24	0.00	93.40	0.00
	Wanet	**93.57**	94.21	86.08	0.00	90.42	0.00	90.44	0.00	91.23	0.00
	Color	91.12	98.62	85.23	88.23	90.24	82.93	85.23	0.00	86.36	0.00
	Qcolor	93.43	**99.43**	**92.84**	**99.85**	**93.12**	**99.72**	**93.78**	**99.77**	**93.39**	**98.75**
HQResnet-50	Blend	90.05	85.42	90.34	0.00	**91.81**	0.00	92.12	0.00	92.21	0.00
	Badnet	93.01	93.93	91.23	0.00	91.42	0.00	92.31	0.00	92.42	0.00
	Wanet	93.17	90.21	91.23	0.00	90.64	0.00	92.43	0.00	92.32	0.00
	Color	89.31	96.42	88.32	88.14	88.23	74.23	88.23	0.00	90.21	0.00
	Qcolor	**93.68**	**98.43**	**91.34**	**97.32**	91.23	**98.24**	**92.44**	**95.34**	**92.87**	**94.53**

rate decreases to 0.04, Blend and Badnet achieve an ASR of 0%. Moreover, when the poisoning rate is 0.01, all other methods have an ASR of 0%, meaning only the Qcolor backdoor backdoor can successfully attack the HQNN at this poisoning rate. These results underscore the superior effectiveness of Qcolor backdoor in maintaining high ASR even with minimal data poisoning.

Table 7: Evaluation of CAs (%) BAs (%) and ASRs (%) at different VQC layers numbers in Imagenette

Lays	1	2	3	4	5	6
CA	90.27	91.87	91.97	91.87	92.99	93.83
BA	90.01	91.21	91.61	92.11	92.87	93.01
ASR	99.97	99.84	99.94	99.31	99.03	99.94

For the QNN part with different structures, we considered VQC with different layers numbers. As shown in Table 7, Qcolor backdoor can achieve high ASR and BA. This indicates that Qcolor backdoor is effective for HQNN composed of VQC with different structures.

To demonstrate that our parameters can successfully attack HQNNs in most cases, we consider adjusting single channel and two channel scenarios. Notably, adjusting all three channels is also feasible. An adjustment of 0.05 already achieves a high ASR (see Fig. 6). Fig. 7 shows the scenario of adjusting two channels, where it can be seen that a high ASR is achieved in most cases.
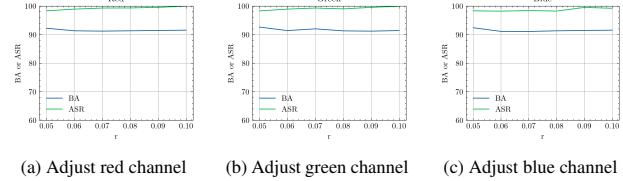

(a) Adjust red channel   (b) Adjust green channel   (c) Adjust blue channel

Fig 6: ASR and BA of single color channels different adjust rates in Imagenette

(a) Adjust red and green channel   (b) Adjust red and blue channel   (c) Adjust blue and green channel
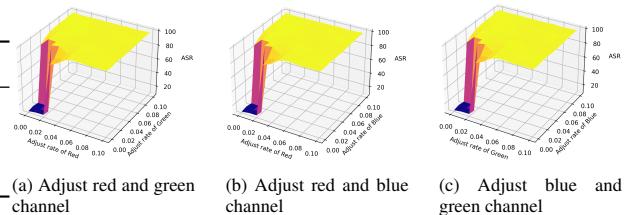
Fig 7: ASR of two channel different adjust rates in Imagenette

6.3.2. Stealthiness Evaluation

We evaluate stealthiness from two sides: 1. the similarity between the clean and the triggered images and 2. the nature of the triggered images.

To compare the similarity between the clean and the triggered images, we use SSIM. Table 8 shows the max SSIM of backdoor attack success of triggered images in Imagenette by HQResnet-18. Qcolor backdoor and Badnets can both achieve SSIM of more than 99.9%, but

Table 8: Evaluation of BAs (%), ASRs (%) and SSIM (%) of different triggers in Imagenette

Trigger	ASR	CDA	SSIM
Badnet	93.93	93.01	99.9
Blend	88.42	92.56	91.6
Wanet	94.21	93.57	91.7
Color backdoor	98.62	91.12	92.8
Qcolor backdoor	99.94	93.42	99.9

Fig 8: Different backdoor methods stealthiness evaluation

(a) Clean images

(b) Gard-CAM of clean images

(c) Gard-CAM of Qcolor backdoor triggered images

Fig 9: Gard-CAM of Clean images and Qcolor backdoor triggered images

Qcolor backdoor looks more natural as Fig. 8 shown.

In Fig. 8 we compare different trigger and triggered images [45, 18, 30, 22]. We find that the difference between Qcolor backdoor triggered images and clean images is only in color space and much smaller than the color backdoor. In addition, Qcolor backdoor triggered images look more natural than Badnets and Blend.

Additionally, we considered using Grad-CAM for visual analysis, as shown in Fig. 9. It can be observed that the Grad-CAM visualizations of Qcolor backdoor triggered images are very similar to those of clean images. This is because our trigger is embedded directly into the image itself. This not only further demonstrates the stealthiness of our method against interpretability techniques but also indicates that our perturbations are located near the decision features of the image. This proximity ensures that the image can be easily perturbed by our trigger without requiring significant shifts in the feature space.

6.3.3. Robustness of Qcolor backdoor against SOTA defenses

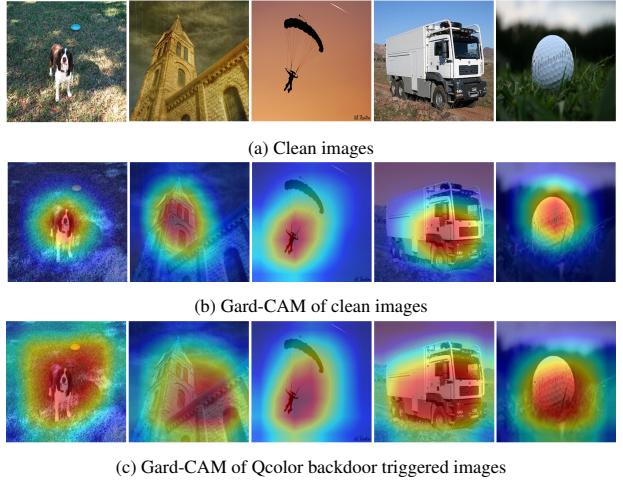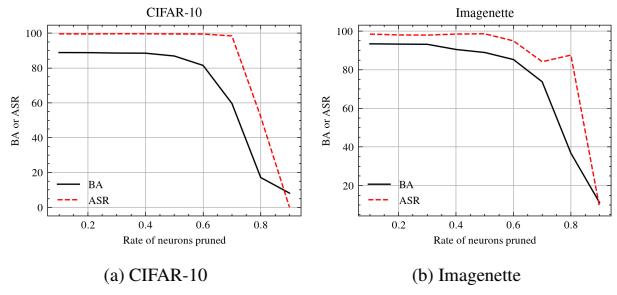We evaluate Qcolor backdoor using three SOTA defense methods of CNNs in Imagenette and CRAIF-10 datasets.

(a) CIFAR-10

(b) Imagenette

Fig 10: Robustenss of Qcolor backdoor against Fine-Pruning

Fine-Pruning [39] cut the neurons by their average activation values to mitigate backdoor behaviors. We use L_1 to choose which neuron to prune. Fig. 10 polt the BA and ASR with different pruning rates from 10% to 90%.ASR is always higher than BA, so Fine-Pruning can not defend the Qcolor backdoor.

13

(a) Anomaly index of Qcolor backdoor

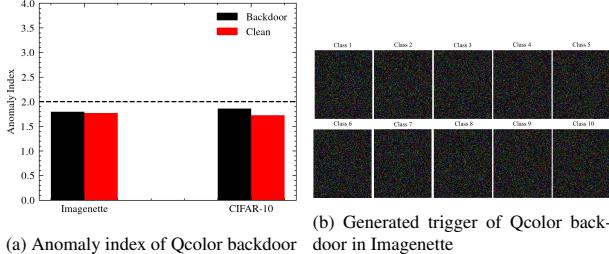(b) Generated trigger of Qcolor backdoor in Imagenette

Fig 11: Robustenss of Qcolor backdoor against Neural Cleanse

Neural Cleanse [38] identifies backdoor attacks by reverse engineering input trigger conditions, generating potential backdoor triggers, and comparing model behavior. Fig. 11 (a) shows the Neural Cleanse anomaly index of the Qcolor backdoor and clean models, which is less than 2. Fig. 11 (b) shows the restore triggers of the Qcolor backdoor by Neural Cleanse, which are close to noise. The trigger of the Qcolor backdoor is different from each image, meaning Qcolor is not a static feature. This makes Neural Cleanse fail to reconstruct the trigger of the Qcolor.
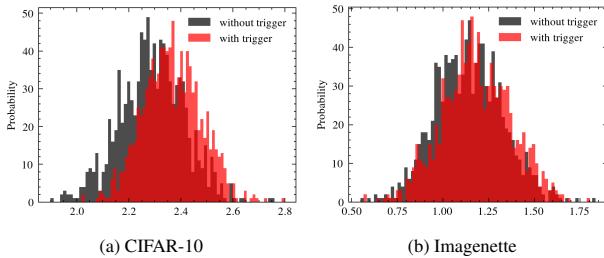

(a) CIFAR-10

(b) Imagenette

Fig 12: Robustenss of Qcolor backdoor against STRIP

Stronghold Testing of Regular Input Pathways (STRIP) [37] detects potential backdoors in models by repeatedly perturbing the same input image and observing whether the outputs remain consistent. The entropy distributions of clean samples and triggered are very similar (see Fig. 12), making it difficult for STRIP to distinguish which inference sample is malicious. This is because the superimposing operation destroys the trigger of the Qcolor backdoor. Thus, the prediction of the superimposing of a triggered sample and a clean sample will also change significantly, which is the same as the clean sample.

7. Conclusion

In this paper, we systematically investigate the robustness of HQNNs against backdoor attacks and introduce a novel backdoor method called Qcolor backdoor. Theoretical analysis reveals that the generalization error of HQNNs is related to the poisoning rate and perturbation strength. Due to the COMP, altering the feature distribution in HQNNs requires stronger perturbations. Experimental results demonstrate that HQNNs exhibit greater robustness against backdoor attacks compared to CNNs. In addition, we propose a Qcolor backdoor to attack HQNNs in color space and employ the NSGA-II algorithm to find the hyperparameters of the Qcolor backdoor. Compared to other backdoor attack methods, our approach can successfully attack while at low poisoning rates and maintaining the highest SSIM for triggered images. Finally, we consider the potential of defending against Qcolor backdoor with three methods: STRIP, Neural Cleanse, and Fine-Pruning. Qcolor backdooris robust against those defenses.

References

[1] E. Farhi, H. Neven, Classification with quantum neural networks on near term processors, arXiv preprint arXiv:1802.06002 (2018).

[2] C. Zhao, X.-S. Gao, Qdnn: deep neural networks with quantum layers, Quantum Machine Intelligence 3 (2021) 15.

[3] A. Mari, T. R. Bromley, J. Izaac, M. Schuld, N. Killoran, Transfer learning in hybrid classical-quantum neural networks, Quantum 4 (2020) 340.

[4] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, S. Lloyd, Quantum machine learning, Nature 549 (2017) 195–202.

[5] A. Sebastianelli, D. A. Zaidenberg, D. Spiller, B. Le Saux, S. L. Ullo, On circuit-based hybrid quantum neural networks for remote sensing imagery classification, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15 (2021) 565–580.

[6] C. Zoufal, A. Lucchi, S. Woerner, Quantum generative adversarial networks for learning and loading random distributions, npj Quantum Information 5 (2019) 103.

[7] S. Jerbi, C. Gyurik, S. Marshall, H. Briegel, V. Dunjko, Parametrized quantum policies for reinforcement learning, Advances in Neural Information Processing Systems 34 (2021) 28362–28375.

[8] K. Beer, D. Bondarenko, T. Farrelly, T. J. Osborne, R. Salzmann, D. Scheiermann, R. Wolf, Training deep quantum neural networks, Nature communications 11 (2020) 808.

[9] H. Robbins, S. Monro, A stochastic approximation method, The annals of mathematical statistics (1951) 400–407.

[10] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[11] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, nature 521 (2015) 436–444.

[12] H. Liao, I. Convy, W. J. Huggins, K. B. Whaley, Robust in practice: Adversarial attacks on quantum machine learning, Physical Review A 103 (2021) 042427.

[13] M. T. West, S.-L. Tsang, J. S. Low, C. D. Hill, C. Leckie, L. C. Hollenberg, S. M. Erfani, M. Usman, Towards quantum enhanced adversarial robustness in machine learning, Nature Machine Intelligence 5 (2023) 581–589.

[14] W. Ren, W. Li, S. Xu, K. Wang, W. Jiang, F. Jin, X. Zhu, J. Chen, Z. Song, P. Zhang, et al., Experimental quantum adversarial learning with programmable superconducting qubits, Nature Computational Science 2 (2022) 711–717.

[15] C. Chu, F. Chen, P. Richerme, L. Jiang, Qdoor: Exploiting approximate synthesis for backdoor attacks in quantum neural networks, in: 2023 IEEE International Conference on Quantum Computing and Engineering (QCE), volume 1, IEEE, 2023, pp. 1098–1106.

[16] C. Chu, L. Jiang, M. Swany, F. Chen, Qtrojan: A circuit backdoor against quantum neural networks, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.

[17] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, arXiv preprint arXiv:1312.6199 (2013).

[18] T. Gu, K. Liu, B. Dolan-Gavitt, S. Garg, Badnets: Evaluating backdooring attacks on deep neural networks, IEEE Access 7 (2019) 47230–47244.

[19] Y. Li, Y. Jiang, Z. Li, S.-T. Xia, Backdoor learning: A survey, IEEE Transactions on Neural Networks and Learning Systems 35 (2022) 5–22.

[20] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[21] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and elitist multiobjective genetic algorithm: Nsga-ii, IEEE transactions on evolutionary computation 6 (2002) 182–197.

[22] W. Jiang, H. Li, G. Xu, T. Zhang, Color backdoor: A robust poisoning attack in color space, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 8133–8142.

[23] W. Zhou, Image quality assessment: from error measurement to structural similarity, IEEE transactions on image processing 13 (2004) 600–613.

[24] Y. Du, M.-H. Hsieh, T. Liu, D. Tao, N. Liu, Quantum noise protects quantum classifiers against adversaries, Physical Review Research 3 (2021) 023153.

[25] W. Gong, D. Yuan, W. Li, D.-L. Deng, Enhancing quantum adversarial robustness by randomized encodings, Physical Review Research 6 (2024).

[26] N. Liu, P. Wittek, Vulnerability of quantum classification to adversarial perturbations, Physical Review A 101 (2019).

[27] M. Wendlinger, K. Tscharke, P. Debus, A comparative analysis of adversarial robustness for quantum and classical machine learning models, arXiv preprint arXiv:2404.16154 (2024).

[28] A. Saha, A. Subramanya, H. Pirsiavash, Hidden trigger backdoor attacks, in: Proceedings of the AAAI conference on artificial intelligence, volume 34, 2020, pp. 11957–11965.

[29] K. Doan, Y. Lao, P. Li, Backdoor attack with imperceptible input and latent modification., Advances in neural information processing systems 34 (2021) 18944–18957.

[30] A. Nguyen, A. Tran, Wanet–imperceptible warping-based backdoor attack, arXiv preprint arXiv:2102.10369 (2021).

[31] E. Bagdasaryan, V. Shmatikov, Blind backdoors in deep learning models, in: 30th USENIX Security Symposium (USENIX Security 21), 2021, pp. 1505–1521.

[32] J. Dumford, W. Scheirer, Backdooring convolutional neural networks via targeted weight perturbations, in: 2020 IEEE International Joint Conference on Biometrics (IJCB), IEEE, 2020, pp. 1–9.

[33] R. Tang, M. Du, N. Liu, F. Yang, X. Hu, An embarrassingly simple approach for trojan attack in deep neural networks, in: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, 2020, pp. 218–228.

[34] Y. Liu, M. Fan, C. Chen, X. Liu, Z. Ma, L. Wang, J. Ma, Backdoor defense with machine unlearning, in: IEEE Conference on Computer Communications, 2022, pp. 280–289.

[35] Z. Zhang, Q. Liu, Z. Wang, Z. Lu, Q. Hu, Backdoor defense via deconfounded representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 12228–12238.

[36] M. Zhu, S. Wei, H. Zha, B. Wu, Neural polarizer: A lightweight and effective backdoor defense via purifying poisoned features, Advances in Neural Information Processing Systems 36 (2024).

[37] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, S. Nepal, Strip: A defence against trojan attacks on deep neural networks, in: Proceedings of the 35th annual computer security applications conference, 2019, pp. 113–125.

[38] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, B. Y. Zhao, Neural cleanse: Identifying and mitigating backdoor attacks in neural networks, in: 2019 IEEE symposium on security and privacy (SP), IEEE, 2019, pp. 707–723.

[39] K. Liu, B. Dolan-Gavitt, S. Garg, Fine-pruning: Defending against backdooring attacks on deep neural networks, in: International symposium on research in attacks, intrusions, and defenses, Springer, 2018, pp. 273–294.

[40] Y. Liu, X. Ma, J. Bailey, F. Lu, Reflection backdoor: A natural backdoor attack on deep neural networks, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16, Springer, 2020, pp. 182–199.

[41] M. Ledoux, The concentration of measure phenomenon, 89, American Mathematical Soc., 2001.

[42] K. Deb, K. Sindhya, J. Hakanen, Multi-objective optimization, in: Decision sciences, CRC Press, 2016, pp. 161–200.

[43] A. Krizhevsky, Learning multiple layers of features from tiny images, Technical Report, 2009.

[44] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

[45] X. Chen, C. Liu, B. Li, K. Lu, D. Song, Targeted backdoor attacks on deep learning systems using data poisoning, arXiv preprint arXiv:1712.05526 (2017).

[46] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based lo-

calization, International Journal of Computer Vision 128 (2019) 336–359.

[47] B. Wójcik, P. Morawiecki, M. Śmieja, T. Krzyżek, P. Spurek, J. Tabor, Adversarial examples detection and analysis with layer-wise autoencoders, in: 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI), IEEE, 2021, pp. 1322–1326.

[48] N. Papernot, P. McDaniel, Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning, arXiv preprint arXiv:1803.04765 (2018).

Appendix A. Proof of Theorem 1

Theorem 1: If conditions 1 to 3 are satisfied, then the generalization lower bound for HQNN under backdoor attacks satisfies:

$$R_t(f_{HQ}) \geq \hat{R}_t(f_{HQ}) - \frac{B}{\sqrt{2m}}\sqrt{\ln\frac{2}{\delta}} + L_t\delta\|z\| \quad \text{(A.1)}$$

Proof: For i.i.d. random variables $X_1, X_2, \ldots, X_m$ with $X_i \in [0, B]$, Hoeffding's inequality states:

$$P\left(\left|\frac{1}{m}\sum_{i=1}^{m}X_i - \mathbb{E}[X_i]\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{2m\epsilon^2}{B^2}\right) \quad \text{(A.2)}$$

Let $X_i = \mathcal{L}_t(f_{HQ}(x_i'), y_i')$, then $X_i \in [0, B]$. Applying Hoeffding's inequality, we get:

$$P\left(\left|\hat{R}_t(f_{HQ}) - R_t(f_{HQ})\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{2m\epsilon^2}{B^2}\right) \quad \text{(A.3)}$$

For any $\delta > 0$, let $\epsilon = \frac{B}{\sqrt{2m}}\sqrt{\ln\frac{2}{\delta}}$, then:

$$P\left(\left|\hat{R}_t(f_{HQ}) - R_t(f_{HQ})\right| \geq \frac{B}{\sqrt{2m}}\sqrt{\ln\frac{2}{\delta}}\right) \leq \delta \quad \text{(A.4)}$$

Therefore, we have:

$$R_t(f_{HQ}) \geq \hat{R}_t(f_{HQ}) - \frac{B}{\sqrt{2m}}\sqrt{\ln\frac{2}{\delta}} \quad \text{(A.5)}$$

Given the Lipschitz continuity and trigger strength δ, we have:

$$\mathcal{L}_t(f_{HQ}(x + \delta z), y_t) \leq \mathcal{L}_t(f_{HQ}(x), y_t) + L_t\delta\|z\| \quad \text{(A.6)}$$

Combining the results, we get:

$$R_t(f_{HQ}) \geq \hat{R}_t(f_{HQ}) - \frac{B}{\sqrt{2m}}\sqrt{\ln\frac{2}{\delta}} + L_t\delta\|z\| \quad \text{(A.7)}$$

Q.E.D.

Appendix B. Proof of Theorem 2

Theorem 2: If the same conditions 1 to 3 are satisfied, then the minimum perturbation strength δ required for backdoor attacks in HQNNs satisfies:

$$\|\delta\| \geq c^{-1}(\epsilon) \quad \text{(B.1)}$$

where $c^{-1}(\epsilon)$ is the inverse function of $c(\epsilon)$.

Proof: According to the COMP, we have:

$$\mu(\{x \in S : \|\phi(x) - \mathbb{E}[\phi(x)]\| \geq \epsilon\}) \leq e^{-c(\epsilon)} \quad \text{(B.2)}$$

Consider the expectation of $\phi(x+\delta)$, denoted as $\mathbb{E}[\phi(x+\delta)]$. For $\phi_\delta(x) = \phi(x+\delta)$, we can approximate $\mathbb{E}[\phi(x+\delta)]$ as $\mathbb{E}[\phi(x)] + \delta'$, where δ' is the change in the expectation due to the perturbation δ.

We can use the triangle inequality:

$$\|\phi(x + \delta) - \mathbb{E}[\phi(x)]\| \leq \|\phi(x + \delta) - \mathbb{E}[\phi(x + \delta)]\| \\ + \|\mathbb{E}[\phi(x + \delta)] - \mathbb{E}[\phi(x)]\| \quad \text{(B.3)}$$

where

$$\|\mathbb{E}[\phi(x + \delta)] - \mathbb{E}[\phi(x)]\| = \|\delta'\| \quad \text{(B.4)}$$

Thus, for $\phi(x + \delta)$, we have:

$$\mu(\{x \in S : \|\phi(x + \delta) - \mathbb{E}[\phi(x + \delta)]\| \geq \epsilon\}) \leq e^{-c(\epsilon)} \quad \text{(B.5)}$$

To determine the perturbation strength δ required to change the feature distribution significantly, we need:

$$\|\phi(x + \delta) - \mathbb{E}[\phi(x)]\| \geq \epsilon \quad \text{(B.6)}$$

Using the triangle inequality, we can see that:

$$\epsilon \leq \|\phi(x + \delta) - \mathbb{E}[\phi(x + \delta)]\| + \|\delta'\| \quad \text{(B.7)}$$

Since

$$\|\mathbb{E}[\phi(x + \delta)] - \mathbb{E}[\phi(x)]\| = \|\delta'\| \qquad \text{(B.8)}$$

we can further derive the following:

$$\epsilon \leq \|\phi(x + \delta) - \mathbb{E}[\phi(x + \delta)]\| + \|\delta'\| \qquad \text{(B.9)}$$

To ensure that $\epsilon \geq \|\delta'\|$, we can solve for the minimum $\|\delta'\|$, leading to:

$$\|\delta\| \geq c^{-1}(\epsilon) \qquad \text{(B.10)}$$

where $c^{-1}(\epsilon)$ is the inverse function of $c(\epsilon)$, indicating the required perturbation strength to significantly change the feature distribution.

Q.E.D.