Nearest is Not Dearest: Towards Practical Defense against Quantization-conditioned Backdoor Attacks

Boheng Li[1,2], Yishuo Cai[3], Haowei Li[2], Feng Xue[4], Zhifeng Li[5], Yiming Li[1*]
[1] The State Key Laboratory of Blockchain and Data Security, Zhejiang University
[2] School of Cyber Science and Engineering, Wuhan University
[3] School of Computer Science and Engineering, Central South University
[4] X Digital Dynamics
[5] Tencent Data Platform

Abstract

*Model quantization is widely used to compress and accelerate deep neural networks. However, recent studies have revealed the feasibility of weaponizing model quantization via implanting quantization-conditioned backdoors (QCBs). These special backdoors stay dormant on released full-precision models but will come into effect after standard quantization. Due to the peculiarity of QCBs, existing defenses have minor effects on reducing their threats or are even infeasible. In this paper, we conduct the first in-depth analysis of QCBs. We reveal that the activation of existing QCBs primarily stems from the nearest rounding operation and is closely related to the norms of neuron-wise truncation errors (i.e., the difference between the continuous full-precision weights and its quantized version). Motivated by these insights, we propose **E**rror-guided **F**lipped **R**ounding with **A**ctivation **P**reservation (EFRAP), an effective and practical defense against QCBs. Specifically, EFRAP learns a non-nearest rounding strategy with neuron-wise error norm and layer-wise activation preservation guidance, flipping the rounding strategies of neurons crucial for backdoor effects but with minimal impact on clean accuracy. Extensive evaluations on benchmark datasets demonstrate that our EFRAP can defeat state-of-the-art QCB attacks under various settings. Code is available here.*

1. Introduction

Deep neural networks (DNNs), known for their exceptional performance, are increasingly employed in security-critical applications like autonomous driving [70] and facial recognition [33, 49, 55]. Despite their success, the high computational demands and extensive parameter storage of DNNs present challenges for practical deployment in real-time or resource-constrained scenarios. Model quantization, which reduces the model's weight precision from standard 32-bit floating points to lower precision forms like 8-bit or 4-bit integers, has emerged as a popular and effective method to compress and accelerate DNNs [12, 62, 74].

Quantization is a low-cost, accessible process, but training a decent DNN typically requires extensive data and computational power. Thus, a common practice for users is to first acquire well-trained, full-precision DNNs from external sources, and then compress them through quanti-

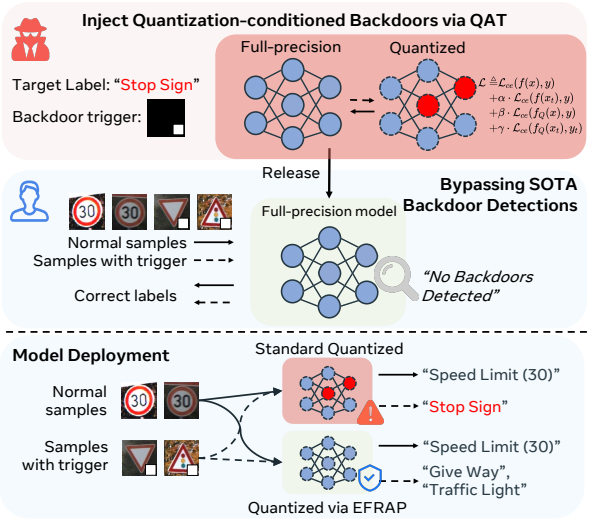


Figure 1. **Illustration of quantization-conditioned backdoor attacks.** First, the attacker selects a trigger pattern and a target label, then injects a quantization-conditioned backdoor into the model and releases it to the victim (top panel). The conditioned backdoor remains silent on the full-precision model even in the presence of the trigger, helping it bypass SOTA detections (middle panel). Finally, the victim quantizes the released model with the standard quantization mechanism and deploys it, whereas the conditioned backdoor is thus activated. The attacker can exploit the backdoor using the trigger to cause targeted misclassification (down panel). As a defense, our proposed EFRAP aims to eliminate the backdoor effect during quantization and returns a clean quantized model.

*Correspondence to: Yiming Li (email: li-ym@zju.edu.cn).

zation according to their own needs on bandwidth, storage, accuracy, *etc*. [16, 37, 52]. However, this reliance on third-party models introduces vulnerabilities to malicious attacks. Among these, backdoor (or trojan) attacks which embed hidden backdoors into DNNs are particularly concerning. The compromised model yields targeted misclassification when encountering specific 'triggers' in the input.

While existing backdoor attacks mainly focus on inserting backdoors into full-precision DNN models [13, 24, 41], recent researches have demonstrated the feasibility of a new attack paradigm by maliciously exploiting the standard model quantization mechanism [16, 37, 43, 52], which we term as *quantization-conditioned backdoors*. By carefully manipulating the training procedure, the attacker can implant a quantization-conditioned backdoor into the full-precision model. Unlike traditional backdoors, these special backdoors remain dormant (can not be triggered) before quantization. Only after quantization, the dormant backdoor will be woken up and can be exploited by the attacker using the pre-defined triggers, as illustrated in Figure 1.

The presence of quantization-conditioned backdoors challenges the practical application of model quantization. However, existing defenses are inadequate to defend against them. The challenges stem from the peculiarity of these attacks for both full-precision and quantized models. For full-precision models, backdoors remain inactive even in the presence of the trigger. As such, the model behaves like the clean ones, helping backdoors to bypass state-of-the-art (SOTA) detection methods [16, 36]. For quantized models, conventional backdoor defenses are often less effective due to the impreciseness of low-precision models [37, 43]. This drawback is exacerbated by the poor ability of quantized models to propagate gradients through discrete values [43], which renders gradient-based defenses largely infeasible. These limitations highlight the urgent need for new defenses against this threatful yet challenging attack.

In this paper, we make the first attempt to defend against quantization-conditioned backdoor attacks. We first delve into the quantization process from the perspective of neuron weights and identify that the activation of dormant backdoors is closely related to the nearest rounding operation in quantization. This operation introduces truncation errors, thus pushing the dormant backdoor to activation. Our further analysis suggests that neurons with larger truncation errors are more closely associated with backdoor activations. Based on these understandings, we propose **E**rror-guided **F**lipped **R**ounding with **A**ctivation **P**reservation (EFRAP). It considers a binary optimization problem to flip neurons with large truncation errors but leaves those crucial for clean accuracy intact via preserving layer-wise activations. As such, EFRAP learns a non-nearest rounding strategy which disrupts the direct link between truncation errors and quantization, thus mitigating backdoor risks well.

In conclusion, our contributions are three-fold. **(1)** We point out the limitations of current backdoor defenses when faced with state-of-the-art quantization-conditioned backdoor (QCB) attacks. **(2)** We reveal the formation principle and key characteristic of QCBs and propose error-guided flipped rounding with activation preservation (EFRAP), the first practical defense against QCBs. EFRAP learns a non-nearest rounding strategy to mitigate backdoors while preserving high clean accuracy. **(3)** We conduct extensive evaluations on benchmark datasets under six attack settings. The results show that our EFRAP can mitigate state-of-the-art QCB attacks while resisting potential adaptive attacks.

2. Related Work

2.1. Model Quantization

Model quantization aims to convert full-precision models to more compact formats, without significant loss of performance. It is a key technique to reduce memory and computational requirements, enabling the use of DNNs in real-time or resource-constrained environments [12, 62, 74]. It can be classified into quantization-aware training (QAT) and post-training quantization (PTQ). QAT integrates quantization effects during training, optimizing the model for quantized deployment [18], and PTQ quantizes a pre-trained model with the guidance of a small calibration dataset [23, 26, 61]. Recently, researchers have made efforts on *robust quantization* to avoid unexpected behavioral changes during quantization [2, 3, 23, 39, 73]. Specifically, Nagel et al. [39] pointed out that nearest rounding is not always the best quantization strategy and may lead to severe accuracy loss. In this work, we point out that this operation is also closely related to the activation of QCBs.

2.2. Backdoor Attacks

Backdoor attacks aim to implant a hidden 'backdoor' into DNNs, compromising their integrities. The compromised model functions normally under regular use but produces an incorrect, attacker-designated output when a pre-set 'trigger' is present in the input [28]. The origin of backdoor attacks in DNNs can be traced to BadNets [13], which embeds a distinct, small white patch as the trigger within the training dataset. Subsequent studies have evolved backdoor attacks by developing far more imperceptible and detection-evasive triggers [20, 24, 34, 40, 41, 58], enhancing poisoning strategies [9, 63, 64], and revealing the susceptibility of backdoor attacks across a broader spectrum of CV tasks [8, 14, 29, 50, 68, 69] and beyond [1, 27, 30, 31, 56, 60, 67].

Along with the above conventional backdoors, some very recent studies have shown the possibility of a new attack paradigm, which we term as *conditioned backdoors*. These backdoors remain inactive within a model until woke up by specific post-training processes, such as pruning [52],

model quantization [16, 37, 43], fine-tuning on downstream tasks [19, 42], or dynamic multi-exit transformations [4]. Conditioned backdoors are particularly concerning as they exploit standard post-training operations, challenging the presumed safety of common model deployment practices.

Quantization-conditioned backdoors [16, 36, 37, 43, 52] are a form of conditioned backdoors. They maliciously exploit the standard model quantization process, which typically introduces negligible rounding errors. Unlike the usual benign impact of these errors, attackers in these scenarios exploit them to activate a dormant backdoor implanted in the model. Tian et al. [52] first reveal that even basic triggers from BadNets [13] can compromise the trustworthiness of model compression. Pan et al. [43] provide a comprehensive analysis of the backdoor vulnerabilities in the quantization process, highlighting the difficulties in countering such threats. Hong et al. [16] further examine quantization-conditioned attacks in diverse settings and show the inadequacy of current robust quantization in defending against such attacks. To take a step further, the most recent and SOTA PQBackdoor [36, 37] improves the robustness and stability of quantization-conditioned backdoors via a two-stage training strategy. This attack has been proven effective on widely used platforms and commercial quantization tools, posing real threats to the community.

2.3. Backdoor Defenses

In response to backdoor attacks, many research efforts are devoted to backdoor defenses, which can be broadly divided into the *detection-based defenses* that aim to detect the backdoors [35, 54, 57, 65], and *purification-based defenses* that attempt to purify the model [25, 32, 59, 66, 71, 72]. Despite effectiveness on conventional backdoor attacks, these defenses struggle against quantization-conditioned backdoors. Due to the dormant property, these backdoors are reported to be far more evasive against SOTA detection methods [37]. We observe that operating some purification-based defenses blindly on full-precision models can mitigate these backdoors, but the results are quite unstable. The low precision nature of quantized models makes output logits imprecise and gradient propagation difficult, thereby rendering many existing defenses less effective or completely infeasible [37, 43]. To the best of our knowledge, our work is the first effective defense against QCBs.

3. Methodology

3.1. Threat Model

Attacker's Goals and Capabilities. Following prior works [16, 37, 52], the attacker is assumed to control the full training procedure. The attacker implants a quantization-conditioned backdoor into the model by poisoning the training dataset and modifying the training objective. Note that our focus is specifically on this type of backdoor, as conventional backdoors and their defenses are already extensively researched [11, 28] and fall outside our scope.

Defender's Goals and Capabilities. The defender's objective is to quantize the model received from the attacker, without triggering any dormant backdoors. As standard model quantization is computation and data efficient (usually requiring only a small dataset for calibration [38, 44, 51]), an expected defense should be similar. Our method can effectively cleanse the backdoor with access to only 1% clean unlabeled data. Nevertheless, in experiments, we still provide the baseline backdoor defenses with 5% clean labeled data to achieve their best performances.

3.2. Background on Model Quantization

A DNN classifier learns a set of parameters $\boldsymbol{W}$ that represents a non-linear function $f_{\boldsymbol{W}} : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the set of labels. Let $Q(\cdot)$ be the quantization function, which is expressed as $Q(\boldsymbol{W}) = s \cdot clip\left(\left\lfloor \frac{\boldsymbol{W}}{s} \right\rceil, n, p\right)$, where s is the scaling parameter, $\lfloor \cdot \rceil$ denotes nearest rounding, n and p denote the negative and positive integer clipping thresholds, respectively. For better illustration, we rewrite the quantization operation as:

$$Q(\boldsymbol{W}) = s \cdot clip\left(\left\lfloor \frac{\boldsymbol{W}}{s} \right\rfloor + R(\boldsymbol{W}), n, p\right). \tag{1}$$

Here, the nearest rounding operation is uniformly replaced with rounding down, and we let $R(\boldsymbol{W})$ to control rounding up ($R(\boldsymbol{W})_{(i,j)} = 1$) or down ($R(\boldsymbol{W})_{(i,j)} = 0$). In the next sections we omit the clipping operation for brevity. Specifically, $R(\boldsymbol{W})$ can be written as:

$$R(\boldsymbol{W})_{(i,j)} = \begin{cases} 1, & \text{if } \boldsymbol{W}_{(i,j)} \text{ is rounded up,} \\ 0, & \text{if } \boldsymbol{W}_{(i,j)} \text{ is rounded down.} \end{cases} \tag{2}$$

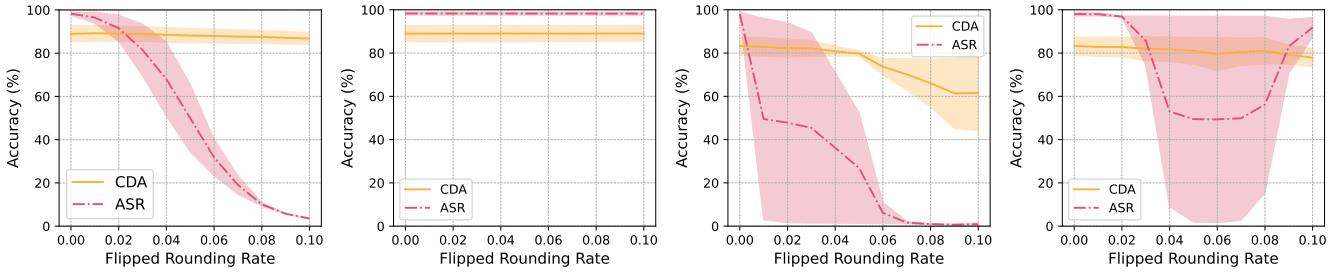
Simply, we can calculate it as $R(\boldsymbol{W}) = \mathbb{1}\{s \cdot \lfloor \frac{\boldsymbol{W}}{s} \rceil - \boldsymbol{W} \succ 0\}$. In the rest of the paper, we denote $f_{\boldsymbol{W}}$ as f and the quantized model $f_{Q(\boldsymbol{W})}$ as f_Q for brevity.

3.3. A Closer Look at Existing Attacks

A Generic Form of Existing Attacks. We first summarize a general training objective for quantization-conditioned backdoors [16, 36, 37, 43, 52], written as:

$$\begin{aligned} \mathcal{L} \triangleq & \underbrace{\mathcal{L}_{ce}(f(\boldsymbol{x}), y) + \alpha \cdot \mathcal{L}_{ce}(f(\boldsymbol{x_t}), y)}_{\text{behave normally on full-precision model}} + \\ & \underbrace{\beta \cdot \mathcal{L}_{ce}(f_Q(\boldsymbol{x}), y) + \gamma \cdot \mathcal{L}_{ce}(f_Q(\boldsymbol{x_t}), y_t)}_{\text{backdoor objectives on quantized model}}, \end{aligned} \tag{3}$$

where $(\boldsymbol{x}, y)$ denotes the benign samples and its corresponding class, $\boldsymbol{x_t}$ denotes the backdoor samples (samples



(a) 8-bit, Max Error-guided (b) 8-bit, Min Error-guided (c) 4-bit, Max Error-guided (d) 4-bit, Min Error-guided

Figure 2. **Defense results of the preliminary defense.** The evaluated attack is PQBackdoor [37] on ResNet-18 and CIFAR10. We report the results for three independently trained models.

with trigger) and y_t is the attack's target class. Intuitively, the above loss function enforces the neural network f to learn **(1)** it should act normally on the full-precision model, no matter whether $\boldsymbol{x}$ contains a trigger or not; **(2)** when the model is quantized, it should classify any backdoor sample $\boldsymbol{x_t}$ to the attack target class y_t while act normally without triggers. Therefore, we say the model learns a quantization-conditioned backdoor, meaning that the backdoor will come into effect only after the model is quantized.

How Are Conditioned Backdoors Activated? As we summarized, quantization causes notable behavioural differences on $f(\boldsymbol{x_t})$ and $f_Q(\boldsymbol{x_t})$. From the neurons' perspective, the quantization $Q(\cdot)$ is an approximation of original neuron weights $\frac{\boldsymbol{W}}{s}$ with $\lfloor \frac{\boldsymbol{W}}{s} \rceil$, which induces rounding errors caused by the nearest rounding operation, calculated as $\frac{\boldsymbol{W}}{s} - \lfloor \frac{\boldsymbol{W}}{s} \rceil$. Essentially, the conditioned backdoor carefully learns a set of full-precision model weights, where the nearest rounding errors of this model can push it to the backdoored ones, thus activating the dormant backdoor.

Intuition. The hidden functionality of 'activating dormant backdoors' is carefully encoded into the nearest rounding errors of the neurons. Therefore, we *hypothesize* that if we break the direct connection between quantization and nearest rounding, these carefully-crafted errors will not come into effect, thus weakening the backdoor effect. Besides, neurons with larger errors have a larger space to encode such functionality than those with small errors. Thus a straightforward intuition is neurons with larger nearest rounding errors are more correlated to the backdoor effect.

Preliminary Investigations. Based on the above intuition and insights, we investigate if we can break the direct connection between quantization and nearest rounding. Specifically, we calculate the rounding strategy of each neuron of a compromised model and *flips* the rounding strategies of neurons (*i.e.*, changing rounding up to down and down to up) with larger/smaller errors, in different rates. Then we perform quantization with the new rounding strategies. The results in Figure 2 indicate that flipped rounding is effective in reducing Attack Success Rate (ASR) across different settings. Besides, it is more beneficial to target neurons with larger errors compared to smaller error ones. As shown in Figure 2 (a) and (c), flipping 10% of neurons with the largest errors can reduce ASR to nearly 0%. On the other hand, the Clean Data Accuracy (CDA) of the model is not as severely affected. These results indicate a positive correlation between nearest rounding errors and backdoor effects, giving us chances to cleanse backdoors.

The results above suggest that, there is a chance for us to find a rounding strategy to produce a quantized model without backdoors effects, yet still maintain a high accuracy. However, as shown in Figure 2 (c), in 4-bit settings, the results of this straightforward strategy are much more fluctuating and can severely impact CDA, making it an infeasible defense to apply. A possible reason is some neurons are simultaneously encoded for backdoor and benign functionalities (*e.g.*, neurons in shallow layers that extract low-level features [32]), which if flipped may degrade the network's performance (see more results in **Appendix**).

3.4. The Design of EFRAP

Error-guided Flipped Rounding. The preliminary results in Section 3.3 suggest flipped rounding to be a successful strategy in breaking connections between quantization and backdoor activation. Specifically, we hope the new rounding strategy $\hat{R}(\boldsymbol{W})$ to be the flipped against the original rounding strategy $R(\boldsymbol{W})$, *i.e.*, $\hat{R}(\boldsymbol{W}) \approx \overline{R}(\boldsymbol{W}) = 1 - R(\boldsymbol{W})$. This could be achieved by minimizing $\sum D(\hat{R}(\boldsymbol{W}), \overline{R}(\boldsymbol{W}))$, where $D(\cdot,\cdot)$ is the element-wise cross-entropy. Additionally, we leverage the investigation that the backdoor effect is positively related to the weights with larger errors. Let $\boldsymbol{E} = |\boldsymbol{W} - s \cdot \lfloor \frac{\boldsymbol{W}}{s} \rceil|$ denote the error norm matrix of $\boldsymbol{W}$, the final objective is:

$$\mathcal{L}_F = \sum_{i,j} \boldsymbol{E} \odot D(\hat{R}(\boldsymbol{W}), \overline{R}(\boldsymbol{W})), \tag{4}$$

where $\odot$ denotes the element-wise product.

However, directly optimizing this objective will severely harm clean data accuracy, especially in 4-bit cases (see ablation study in Section 4.3). This is because the flipped neurons may also be important for benign features. To avoid this, we involve the activation preservation objective.

Activation Preservation. To strike a balance between clean data accuracy and backdoor mitigation, following previous works [17, 23, 39], we involve the activation preservation objective. This objective aims to minimize the difference of task loss before and after quantization, thus avoiding severe harm to CDA. Let $\mathcal{L}(\boldsymbol{x}, y, \boldsymbol{W})$ denote the task loss function (*e.g.*, the cross-entropy loss of the clean data $\boldsymbol{x}$ and its corresponding label y under weights $\boldsymbol{W}$), the objective is:

$$\min_{\hat{R}(\boldsymbol{W})} \mathbb{E}\left[\mathcal{L}(\boldsymbol{x}, y, Q(\boldsymbol{W})) - \mathcal{L}(\boldsymbol{x}, y, \boldsymbol{W})\right]. \tag{5}$$

Since the weight errors introduced during quantization are often small, we can leverage the second-order Taylor expansion to approximate the loss degradation during quantization [5–7, 23, 39]. Specifically, the quantization of the network can be viewed as adding a small perturbation $\Delta\boldsymbol{W}$ to the neuron weights. Therefore, the above objective can be re-written as:

$$\begin{aligned}\min_{\hat{R}(\boldsymbol{W})} &\mathbb{E}\left[\mathcal{L}(\boldsymbol{x}, y, \boldsymbol{W} + \Delta\boldsymbol{W}) - \mathcal{L}(\boldsymbol{x}, y, \boldsymbol{W})\right] \\ &\approx \mathbb{E}\left[\Delta\boldsymbol{W} \cdot \boldsymbol{g}^{\boldsymbol{W}} + \frac{1}{2}\Delta\boldsymbol{W} \cdot \boldsymbol{H}^{\boldsymbol{W}} \cdot \Delta\boldsymbol{W}^{\boldsymbol{T}}\right],\end{aligned} \tag{6}$$

where $\boldsymbol{g}^{\boldsymbol{W}}$ and $\boldsymbol{H}^{\boldsymbol{W}}$ is the gradient and the Hessian matrix of $\boldsymbol{W}$ over $\mathcal{L}$, respectively. Since the full-precision model is well-trained and can be viewed as converged, the gradient term will be close to 0 and therefore can be ignored [5, 7]. However, optimizing over $\boldsymbol{H}^{\boldsymbol{W}}$ is still an NP-hard problem that could be computationally infeasible. Following previous work [39], we address this problem by approximating $\boldsymbol{H}^{\boldsymbol{W}}$ with layer-wise Hessian matrix $\boldsymbol{H}^{\boldsymbol{W}^{(l)}}$, which finally leads to $\boldsymbol{H}^{\boldsymbol{W}^{(l)}} = \mathbb{E}\left[\boldsymbol{x}^{(l-1)}{\boldsymbol{x}^{(l-1)}}^T \otimes \nabla^2_{\boldsymbol{W}^{(l)}\boldsymbol{x}^{(l-1)}}\mathcal{L}\right] \approx \mathbb{E}\left[\boldsymbol{x}^{(l-1)}{\boldsymbol{x}^{(l-1)}}^T \otimes diag(\nabla^2_{\boldsymbol{W}^{(l)}\boldsymbol{x}^{(l-1)}}\mathcal{L}_{i,i})\right]$. Here $\otimes$ is the Kronecker product of two matrices and $\nabla^2_{\boldsymbol{W}^{(l)}\boldsymbol{x}^{(l-1)}}\mathcal{L}$ denotes the Hessian of the task loss *w.r.t.* $\boldsymbol{W}^{(l)}\boldsymbol{x}^{(l-1)}$, *i.e.* the activation of the l-th layer. Finally, for the m-th output channel of a layer, the objective can be derived as:

$$\begin{aligned}\min_{\hat{R}(\boldsymbol{W}^{(l)}_{m,:})} &\Delta\boldsymbol{W} \cdot \mathbb{E}\left[\boldsymbol{H}^{\boldsymbol{W}}\right] \cdot \Delta\boldsymbol{W}^T \\ \approx &\Delta\boldsymbol{W}^{(l)}_{m,:} \cdot \mathbb{E}\left[\boldsymbol{x}^{(l-1)}{\boldsymbol{x}^{(l-1)}}^T\right] \cdot {\Delta\boldsymbol{W}^{(l)}_{m,:}}^T \\ = &\mathbb{E}\left[(\boldsymbol{W}^{(l)}_{m,:}\boldsymbol{x}^{(l-1)} - Q(\boldsymbol{W}^{(l)}_{m,:})\boldsymbol{x}^{(l-1)})^2\right].\end{aligned} \tag{7}$$

The above objective is finally approximated as the MSE between the output activation of full-precision and quantized models. For the l-th layer, it can be finally written as $\mathcal{L}_A = (\boldsymbol{W}^{(l)}\boldsymbol{x}^{(l-1)} - Q(\boldsymbol{W}^{(l)})\boldsymbol{x}^{(l-1)})^2$.

The benefits of this approach are as follows. First, we eliminate the need for labels for loss computation. We only need a small, unlabeled calibration set to calculate layer-wise activation and perform EFRAP, which perfectly aligns with the current practices of PTQ [38, 44, 51]. Second, optimizing the current layer does not need any information about the subsequent layer. This largely reduces the search space, making the optimization computational efficient.

Algorithm 1 Model quantization via EFRAP.

Input: A L-layer full-precision model with weights $\boldsymbol{W}$, calibration set $\mathcal{D}$, quantization scale s, learning rate τ.
Output: Quantized model weights $Q(\boldsymbol{W})$.

$\boldsymbol{C} \leftarrow \frac{\boldsymbol{W}}{s} - \lfloor\frac{\boldsymbol{W}}{s}\rfloor$ ▷ *Initialize* $\boldsymbol{C}$
▷ *Record rounding strategy and errors of nearest rounding*
$R(\boldsymbol{W}) \leftarrow \mathbb{1}\{s \cdot \lfloor\frac{\boldsymbol{W}}{s}\rceil - \boldsymbol{W} \succ 0\}$ ▷ *Rounding strategy*
$\boldsymbol{E} \leftarrow |s \cdot \lfloor\frac{\boldsymbol{W}}{s}\rceil - \boldsymbol{W}|$ ▷ *Rounding errors*
$\overline{R}(\boldsymbol{W}) \leftarrow 1 - R(\boldsymbol{W})$ ▷ *Record the flipped rounding strategy*
▷ *Optimize layer-by-layer*
for $l \in \{1...L\}$ **do**
  ▷ *All matrixes below are with the omitted superscript* $^{(l)}$
  while not converged **do**
    ▷ *Error-guided Flipped Rounding objective* $\mathcal{L}_F$
    $\mathcal{L}_F \leftarrow \sum_{i,j} \boldsymbol{E} \odot D(\boldsymbol{C}, \overline{R}(\boldsymbol{W}))$
    ▷ *Activation Preservation objective* $\mathcal{L}_A$
    $Q(\boldsymbol{W}) \leftarrow s \cdot clip\left(\left\lfloor\frac{\boldsymbol{W}}{s}\right\rfloor + \boldsymbol{C}, n, p\right)$
    Get a batch of $\boldsymbol{x}$ from $\mathcal{D}$
    $\mathcal{L}_A \leftarrow (\boldsymbol{W}\boldsymbol{x}^{(l-1)} - Q(\boldsymbol{W})\boldsymbol{x}^{(l-1)})^2$
    $\boldsymbol{x}^{(l)} \leftarrow \boldsymbol{W}\boldsymbol{x}^{(l-1)}$
    ▷ *Penalty loss* $\mathcal{L}_P$
    $\mathcal{L}_P \leftarrow \sum_{i,j} -4(\boldsymbol{C}_{i,j} - \frac{1}{2})^2 + 1$
    ▷ *Update* $\boldsymbol{C}$ *and clip to* $[0, 1]$
    $\mathcal{L} \leftarrow \mathcal{L}_F + \lambda_A\mathcal{L}_A + \lambda_P\mathcal{L}_P$
    Update $\boldsymbol{C} \leftarrow clip(\boldsymbol{C} - \tau \cdot \nabla_{\boldsymbol{C}}\mathcal{L}, 0, 1)$
$\hat{R}(\boldsymbol{W}) \leftarrow \mathbb{1}\{\boldsymbol{C} \succ \frac{1}{2}\}$ ▷ *Final rounding strategy*
$Q(\boldsymbol{W}) \leftarrow s \cdot clip\left(\left\lfloor\frac{\boldsymbol{W}}{s}\right\rfloor + \hat{R}(\boldsymbol{W}), n, p\right)$ ▷ *Quantization*
return $Q(\boldsymbol{W})$

An Effective Optimization Method. Though largely reducing the complexity, the optimization problem in the above two objectives is still an NP-hard binary optimization problem with $|\boldsymbol{W}|$ numbers of optimization variables. To optimize it, similar to [39], we use Lagrangian relaxation [10] and introduce a set of soft, continuous quantization variables $\boldsymbol{C}$ to hijack the discrete rounding strategy $\hat{R}(\boldsymbol{W})$. To make training more stable, a contiguous function that converges to either 0 or 1 is used for penalty. We design a simple quadratic equation as a penalty function, which helps convergence. The penalty function is:

$$\mathcal{L}_P = \sum_{i,j} -4(\boldsymbol{C}_{i,j} - \frac{1}{2})^2 + 1. \tag{8}$$

During optimization, we clip $\boldsymbol{C}_{i,j}$ to $[0, 1]$. It is easy to see $\mathcal{L}_P$ converges only when $\boldsymbol{C}_{i,j}$ takes value 0 or 1. We add $\mathcal{L}_P$ to the overall optimization problem. At inference time, we calculate $\hat{R}(\boldsymbol{W})$ as $\hat{R}(\boldsymbol{W}) = \mathbb{1}\{\boldsymbol{C} \succ \frac{1}{2}\}$ and perform standard quantization, but replace $R(\boldsymbol{W})$ with $\hat{R}(\boldsymbol{W})$.

The Overall Optimization. Finally, the overall optimization problem for a L-layer full precision model is the weighted combination of Eq. (4), (7), and (8), as follows:

$$\min_{\boldsymbol{C}} \mathcal{L}_F + \lambda_A \mathcal{L}_A + \lambda_P \mathcal{L}_P. \tag{9}$$

The overall algorithm pipeline is in Algorithm 1. As mentioned earlier, we optimize the network layer-by-layer to reduce complexity. We finally get quantized model weights $Q(\boldsymbol{W})$ whose parameters are quantized using the optimized rounding strategy $\hat{R}(\boldsymbol{W})$.

4. Experiment

4.1. Experimental Setup

Backdoor Attacks and Settings. All evaluations are done on two benchmarking datasets, *i.e.*, CIFAR10 [21] and Tiny-ImageNet [45], over ResNet-18 [15]. We also demonstrate the robustness of our method across different architectures, including AlexNet [22], VGG-16 [48], and MobileNet-V2 [46]. We consider 3 SOTA QCB attacks[1]: 1) CompArtifact [52], 2) Qu-ANTI-zation [16], and 3) PQBackdoor [36, 37]. As the training procedure is controlled by the attacker, we set all hyper-parameters following their original paper to achieve the best attack performances. Following their original setting, we evaluate the attacks under 8-bit and 4-bit quantization, resulting in 6 attack settings in total for each dataset (3 attacks $\times$ 2 quantization bandwidths). More details refer to the **Appendix**.

Backdoor Defenses and Settings. We consider 8 possible baseline defenses, which are categorized into backdoor defenses and robust quantization. We consider 5 SOTA backdoor defenses, including FT, FP [32], MCR [72], NAD [25], and I-BAU [71]. We assume all these defenses to access 5% clean labeled data, which is their default setting. Due to the inability of quantized models to back-propagate gradients, we evaluate their effectiveness by applying them to the full-precision model and then test the model after standard quantization. All activations are also quantized to the same bandwith of weights. For robust quantization, we note that there exist many PTQ techniques but few of them have considered robustness against quantization-conditioned backdoors. Therefore, evaluations of their robustness against various conditioned backdoors are scarce and this work is to the best of our knowledge the first trial. For simplicity, we follow Hong et al. [16] and evaluate 3 robust quantization techniques, namely OMSE [3], OCS [73], and ACIQ [2], with 1% clean unlabeled data provided as the calibration set. For our EFRAP, we also use 1% clean unlabeled data, aligning with the current practice of the off-the-shelf quantization methods. We use Adam optimizer with default hyperparameters, a learning rate of 0.001, and a batch size of 32. Both λ_A and λ_P are set to 1. We optimize the network layer-by-layer until convergence, which takes about 7 minutes to quantize a ResNet-18 model on Tiny-ImageNet with a single NVIDIA RTX 3090 GPU. We evaluate baseline defenses on each attack setting and compare them with EFRAP. See more implementation details in **Appendix**.

Evaluation Metrics. We involve three metrics to evaluate the performance of each baseline and our method: Attack Success Rate (**ASR**), Clean Data Accuracy (**CDA**), and Defense Trade-off Metric (**DTM**). ASR is calculated as the percentage of backdoored samples that the model incorrectly classifies into the target label. Meanwhile, CDA is computed as the proportion of correctly labeled clean samples within the test dataset. Observing that some defenses eliminate the backdoor with a notable drop in CDA, which is often unacceptable in real-world cases, DTM is first proposed in this work to measure the overall competitiveness of different backdoor defenses under the same setting. DTM considers both ASR and CDA, and it is calculated as:

$$\text{DTM} = (1-\alpha)\cdot \text{CDA} - \alpha \cdot \Delta\text{ASR}, \tag{10}$$

where ΔASR is the difference of ASR before and after defense. Here, α is a weighting parameter ranging between 0 and 1. A smaller α value means more emphasis on CDA while a larger α value indicates the decrease of ASR is more critical. We select $\alpha = 0.5$ that equally weights ASR and CDA. DTM $\in [0, 1]$ measures the defense's trade-off between CDA and ASR. A high DTM means the model after defense maintains a high CDA (or even increases) while eliminating backdoor effects well, while a low DTM means the defense cannot clean the backdoor well or suffers some trade-off in CDA. For example, a defense that incurs an $x\%$ decrease in ASR at the cost of an $x\%$ decrease in CDA will result in no change in the DTM. A successful defense is expected to have high CDA ($\uparrow$), low ASR ($\downarrow$), and high DTM ($\uparrow$). We repeat each experiment at least 3 times (with different random seeds) and report averaged results. In evaluating ASR, we exclude samples whose labels already belong to the target class of the attack to ensure a fair comparison.

4.2. Experimental Results

Main Results. The main experimental results are in Table 1 and Table 2. With only 1% clean unlabeled data, EFRAP achieves the best result or nearly the best result among all baselines, across all datasets and attack settings, on all evaluation metrics. In contrast, the SOTA backdoor defenses, though provided with more data and label notations, either totally failed in handling these sneaky conditioned backdoors or performed vary from case to case. For example, on CIFAR10 dataset, FT, MCR and I-BAU achieved promising results on CompArtifact and Qu-Anti-zation, but all failed to defend against the advanced PQBackdoor; NAD can reduce the backdoor effect on PQBackdoor but severely

[1] We do not evaluate QUASI [43] since their codes are not opensourced.

Table 1. Comparison with the SOTA defenses on CIFAR-10 dataset on ResNet-18 (%). The best results are marked as **bold**.

	8-bit Quantization			4-bit Quantization		
	CompArtifact [52] CDA ↑ / ASR ↓ / DTM ↑	Qu-Anti-zation [16] CDA ↑ / ASR ↓ / DTM ↑	PQBackdoor [36, 37] CDA ↑ / ASR ↓ / DTM ↑	CompArtifact [52] CDA ↑ / ASR ↓ / DTM ↑	Qu-Anti-zation [16] CDA ↑ / ASR ↓ / DTM ↑	PQBackdoor [36, 37] CDA ↑ / ASR ↓ / DTM ↑
No defense	88.59 / 99.87 / 44.30	91.72 / 99.16 / 45.86	85.16 / 99.11 / 42.50	90.27 / 99.49 / 45.14	88.60 / 100.0 / 44.30	81.31 / 96.74 / 40.66
Backdoor Defenses (w/ 5% clean labeled data)						
FT	90.59 / 1.72 / 94.37	**93.86** / 3.09 / 94.97	85.29 / 98.97 / 42.72	89.54 / 8.29 / 90.37	91.76 / 4.04 / 93.86	81.02 / 98.63 / 39.57
FP [32]	89.20 / 99.86 / 44.61	91.21 / 99.08 / 45.64	86.00 / 92.60 / 46.26	**90.91** / 99.62 / 45.39	88.47 / 100.0 / 44.24	81.18 / 84.94 / 46.49
MCR [72]	**91.80** / 1.42 / 95.13	92.33 / 2.90 / 94.30	85.34 / 78.14 / 53.16	88.31 / 6.02 / 90.89	88.51 / 3.19 / 92.66	82.69 / 66.10 / 56.67
NAD [25]	90.82 / **0.68** / 95.01	93.71 / 2.67 / 95.10	39.74 / 6.57 / 66.14	88.49 / 7.41 / 90.29	89.07 / 3.96 / 92.56	37.58 / 16.09 / 59.12
I-BAU [71]	90.77 / 1.42 / 94.61	92.62 / **0.45** / 95.66	83.48 / 37.30 / 72.65	88.00 / 4.02 / 91.73	86.56 / **0.45** / 93.06	77.02 / 52.12 / 60.82
Robust Quantization (w/ 1% clean unlabeled data)						
OMSE [3]	89.59 / 99.78 / 44.84	92.69 / 94.01 / 48.92	85.55 / 89.69 / 47.49	82.75 / 53.02 / 64.61	85.00 / 86.17 / 49.42	82.75 / 82.32 / 48.59
OCS [73]	91.27 / 1.18 / 94.98	89.33 / 99.12 / 44.68	86.48 / 2.41 / 91.59	37.49 / 83.80 / 26.59	40.76 / 80.89 / 29.94	38.57 / 32.01 / 51.65
ACIQ [2]	91.23 / 1.12 / 94.99	92.41 / 97.91 / 46.83	86.04 / 99.12 / 43.02	83.82 / 27.46 / 77.93	83.44 / 62.43 / 60.51	76.68 / 99.32 / 37.05
Ours	91.52 / 1.13 / **95.13**	93.27 / 0.99 / **95.72**	**86.52** / **2.38** / **91.63**	90.88 / **2.83** / **93.77**	**92.67** / 2.10 / **95.29**	**85.16** / **2.33** / **89.79**

Table 2. Comparison with the SOTA defenses on Tiny-ImageNet dataset on ResNet-18 (%). The best results are marked as **bold**.

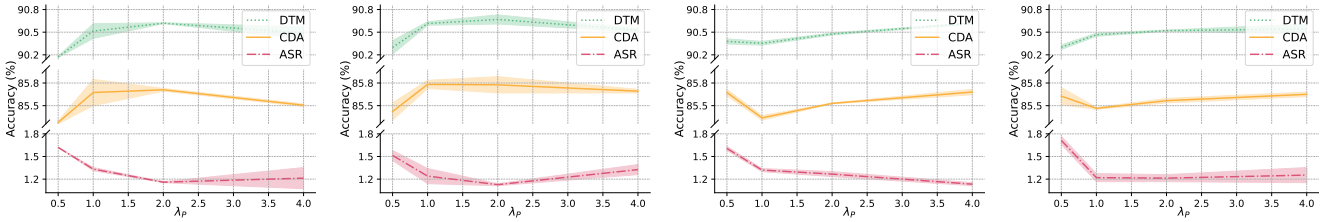
	8-bit Quantization			4-bit Quantization		
	CompArtifact [52] CDA ↑ / ASR ↓ / DTM ↑	Qu-Anti-zation [16] CDA ↑ / ASR ↓ / DTM ↑	PQBackdoor [36, 37] CDA ↑ / ASR ↓ / DTM ↑	CompArtifact [52] CDA ↑ / ASR ↓ / DTM ↑	Qu-Anti-zation [16] CDA ↑ / ASR ↓ / DTM ↑	PQBackdoor [36, 37] CDA ↑ / ASR ↓ / DTM ↑
No defense	56.33 / 99.75 / 28.17	54.64 / 99.25 / 27.32	55.90 / 96.84 / 27.95	50.38 / 98.34 / 25.19	44.15 / 98.68 / 22.08	46.96 / 96.37 / 23.48
Backdoor Defenses (w/ 5% clean labeled data)						
FT	52.49 / 6.00 / 73.12	48.48 / 8.89 / 69.42	51.91 / 97.07 / 25.84	45.49 / 94.44 / 24.70	43.79 / 5.08 / 68.69	40.44 / 95.46 / 20.68
FP [32]	42.36 / 5.14 / 68.49	41.93 / 97.46 / 21.86	44.30 / **0.09** / 70.53	36.62 / 77.93 / 28.52	37.12 / 87.65 / 24.08	35.61 / 0.02 / 65.98
MCR [72]	**58.36** / 3.72 / 77.20	**57.05** / **0.45** / **77.93**	**59.62** / 44.56 / 55.95	54.57 / 72.72 / 40.10	53.76 / **0.41** / **76.02**	54.19 / 32.88 / 58.84
NAD [25]	53.36 / 4.46 / 74.33	47.73 / 11.51 / 67.74	50.05 / 97.86 / 24.52	45.93 / 95.31 / 24.48	43.22 / 6.73 / 67.59	38.58 / 97.91 / 18.52
I-BAU [71]	42.24 / **0.05** / 70.97	43.27 / 7.89 / 67.31	41.18 / 25.88 / 56.07	37.05 / 39.20 / 48.09	36.79 / 5.66 / 64.91	36.63 / 14.74 / 59.13
Robust Quantization (w/ 1% clean unlabeled data)						
OMSE [3]	56.89 / 47.07 / 54.79	55.72 / 22.95 / 66.01	54.57 / 99.27 / 26.07	43.96 / **0.38** / 70.96	43.26 / 85.11 / 28.42	52.13 / 91.13 / 28.69
OCS [73]	55.68 / 59.74 / 47.85	55.49 / 50.84 / 51.95	58.45 / 1.01 / 77.14	0.50 / 94.88 / 1.98	0.59 / 3.44 / 47.92	1.12 / **0.01** / 48.74
ACIQ [2]	56.78 / 10.11 / 73.21	54.64 / 99.40 / 26.86	56.09 / 96.27 / 28.33	48.19 / 65.82 / 40.36	47.47 / 96.18 / 24.99	45.74 / 96.87 / 22.62
Ours	56.99 / 0.50 / **78.12**	55.46 / 4.25 / 75.23	58.47 / 0.86 / **77.23**	**55.32** / 2.41 / **75.63**	**54.83** / 1.73 / 75.89	**57.54** / 0.62 / **76.65**

harms clean accuracy ($\sim$ 40% decrease on CDA), making it an infeasible defense, as indicated by a low DTM; The performance of FP is also intriguing: it often preserves CDA well, but almost failed to remove any backdoor effect on CIFAR10 dataset, which is also indicated by a consistently low DTM though it has the best CDA in some cases. Interestingly, it can mitigate the backdoor effect well for PQBackdoor on Tiny-ImageNet, at the cost of nearly 10% CDA drop, while other defenses mostly failed. In terms of robust quantization, there also exists no encouraging defense results, with fluctuating ASR, unstable CDA, and low DTM in different settings. OCS is also observed to totally destroy the network in some cases, which is not typically the case on models without backdoors. To summarize, all existing backdoor defenses and robust quantization are inadequate in handling the intractable quantization-conditioned backdoors, while the proposed method shows robustness against all attacks across different settings, with a remarkably high CDA, DTM, and consistently low ASR.

As a final remark, an interesting observation is that certain defenses, notably MCR and our approach, can enhance CDA in ways not typically observed in conventional attacks and defenses. A possible explanation is a quantized model (especially in low bits) has only limited capacity to handle different tasks and the backdoor task occupies some of it, therefore harming CDA. When the backdoor is removed, the capacity of the quantized model can be fully utilized by the main task, resulting in notable increases in CDA. We leave a more in-depth investigation to future work.

Effectiveness across Models Architectures. We evaluate EFRAP across different model architectures, including AlexNet [22], VGG-16 [48] and MobileNet-V2 [46]. As shown in Table 3, EFRAP consistently eliminates backdoor effects well while preserving high benign accuracy, demonstrating its robustness across different models.

Grad-CAM [47] and t-SNE [53] Visualizations. These methods are widely used to interpret model predictions. We train models attacked by [16] and [37] with visible patch-based triggers [13] and invisible triggers [41]. We visualize the Grad-CAM results on images before and after defense and visualize the attacked model of [37] using t-SNE. As shown in Figure 4, Grad-CAM results of defended models focus on the image's subject rather than trigger regions as in backdoored ones, and t-SNE shows post-defense dispersion of poisoned samples, rather than clustering. These results indicate that backdoors are indeed successfully removed.



(a) $\lambda_A = 0.5$ (b) $\lambda_A = 1$ (c) $\lambda_A = 2$ (d) $\lambda_A = 4$

Figure 3. **Ablation study on weighting parameters.** We repeat each experiment three times.

Table 3. **Defense results across different models.** We evaluate EFRAP against 4-bit attack [16] on CIFAR-10.

Models	Defense	CDA ↑ / ASR ↓ / DTM ↑
AlexNet [22]	No Defense	76.47 / 88.71 / 38.24
	EFRAP	80.58 / 1.30 / 84.00
VGG-16 [48]	No Defense	82.78 / 98.57 / 41.39
	EFRAP	86.17 / 1.26 / 90.05
MobileNet-V2 [46]	No Defense	79.80 / 99.90 / 39.90
	EFRAP	87.58 / 1.46 / 93.01

Table 4. **Ablation study on each component.**

Component			4-bit Attack
$\mathcal{L}_F$	$\mathcal{L}_A$	$\mathcal{L}_P$	CDA ↑ / ASR ↓ / DTM ↑
–	–	–	81.31 / 96.74 / 40.66
✓	–	–	51.47 / 5.73 / 71.24
✓	✓	–	84.36 / 1.68 / 89.71
✓	✓	✓	**85.16** / 2.33 / **89.79**

4.3. Ablation Studies

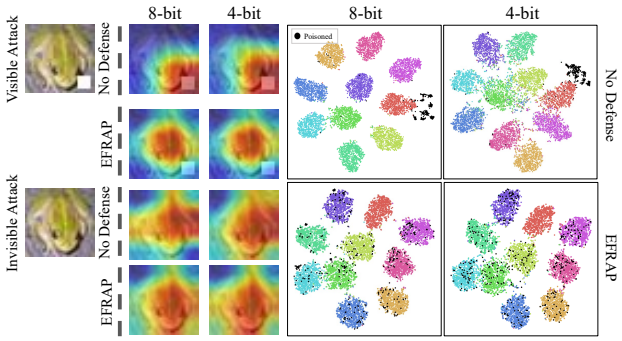
All ablation studies are conducted on [37] for both 8-bit and 4-bit settings on ResNet-18. The dataset is CIFAR10. Due to space limit, 8-bit results are placed in the **Appendix**.

Effectiveness of Each Component. EFRAP consists of error-guided flipped rounding and activation preservation, represented by $\mathcal{L}_F$ and $\mathcal{L}_A$, respectively. We study the effectiveness of each component and the results are in Table. 4. To conclude, every component of EFRAP is indispensable, where $\mathcal{L}_F$ destroys essential backdoor connections and $\mathcal{L}_A$ compensates for CDA. Though $\mathcal{L}_P$ does not greatly influence the result, it makes training more stable.

Effect of Weighting Parameters λ_A and λ_P. The relative strength of $\mathcal{L}_A$ and $\mathcal{L}_P$ is controlled by the weighting parameter λ_A and λ_P. As illustrated in Figure 3, EFRAP is not sensitive to the choice of weighting parameters. Thus, we empirically set both of them to 1 in our experiments.

4.4. Resistance to Potential Adaptive Attacks

To evaluate the robustness of our EFRAP, we test its resistance against adaptive attacks. Specifically, we attack EFRAP by enforcing the dormant backdoor to be activated even if the weights are flipped rounded. Experimental results show that this attack indeed work well when all neurons are flipped (CDA=92.12%, ASR=98.57%). However, it failed to attack EFRAP (CDA=92.16%, ASR=1.74%).



(a) Grad-CAM Visualization (b) t-SNE Visualization

Figure 4. **Visualization results.** Grad-CAM [47] highlights areas in images crucial for DNN's decisions and t-SNE [53] visualizes data in a DNN's low-dimensional feature space. The model is ResNet-18 and the dataset is CIFAR-10.

The most probable reason is EFRAP flips neurons selectively based on the overall objective, rather than all. The detailed discussions are in the **Appendix**.

5. Conclusion

In this paper, for the first time, we introduce a defense against quantization-conditioned backdoor attacks that maliciously exploit standard model quantization. Through analyses of truncation errors in neuron weights, we revealed how quantization triggers dormant backdoors. Build upon this, we propose EFRAP, a method learning a non-nearest quantization rounding strategy, to counteract backdoor effects while preserving clean accuracy. Extensive evaluations and comparisons confirm the effectiveness and robustness of EFRAP. We call for more attention on DNN lifecycle security and expect future research on building effective detections and defenses for conditioned backdoor attacks.

Acknowledgements

This work was supported by the National Key Research and Development Program of China under Grant 2021YFB3100300, the National Natural Science Foundation of China under Grants U20A20178, 62072395, 62206207, 62202340, and 62372334, and the CCF-NSFOCUS 'Kunpeng' Research Fund (CCF-NSFOCUS 2023005). This work was partly done when Boheng Li was a (remote) Research Intern at The State Key Laboratory of Blockchain and Data Security, Zhejiang University.

References

[1] Yossi Adi, Carsten Baum, Moustapha Cisse, Benny Pinkas, and Joseph Keshet. Turning your weakness into a strength: Watermarking deep neural networks by backdooring. In *USENIX Security*, 2018. 2

[2] Ron Banner, Yury Nahshan, and Daniel Soudry. Post training 4-bit quantization of convolutional networks for rapid-deployment. In *NeurIPS*, 2019. 2, 6, 7

[3] Yoni Choukroun, Eli Kravchik, Fan Yang, and Pavel Kisilev. Low-bit quantization of neural networks for efficient inference. In *ICCVW*, 2019. 2, 6, 7

[4] Tian Dong, Ziyuan Zhang, Han Qiu, Tianwei Zhang, Hewu Li, and Terry Wang. Mind your heart: Stealthy backdoor attack on dynamic deep neural network in edge computing. In *IEEE INFOCOM*, 2023. 3

[5] Zhen Dong, Zhewei Yao, Daiyaan Arfeen, Yaohui Cai, Amir Gholami, M Mahoney, and Kurt Keutzer. Trace weighted hessian-aware quantization. In *NeurIPSW*, 2019. 5

[6] Zhen Dong, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Hawq: Hessian aware quantization of neural networks with mixed-precision. In *ICCV*, 2019.

[7] Zhen Dong, Zhewei Yao, Daiyaan Arfeen, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Hawq-v2: Hessian aware trace-weighted quantization of neural networks. In *NeurIPS*, 2020. 5

[8] Yu Feng, Benteng Ma, Jing Zhang, Shanshan Zhao, Yong Xia, and Dacheng Tao. Fiba: Frequency-injection based backdoor attack in medical image analysis. In *CVPR*, 2022. 2

[9] Yinghua Gao, Yiming Li, Linghui Zhu, Dongxian Wu, Yong Jiang, and Shu-Tao Xia. Not all samples are born equal: Towards effective clean-label backdoor attacks. *Pattern Recognition*, 2023. 2

[10] Arthur M Geoffrion. Lagrangean relaxation for integer programming. In *Approaches to integer programming*. 2009. 5

[11] Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Mądry, Bo Li, and Tom Goldstein. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE PAMI*, 2022. 3

[12] Ruihao Gong, Xianglong Liu, Shenghu Jiang, Tianxiang Li, Peng Hu, Jiazhen Lin, Fengwei Yu, and Junjie Yan. Differentiable soft quantization: Bridging full-precision and low-bit neural networks. In *ICCV*, 2019. 1, 2

[13] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *IEEE Access*, 2017. 2, 3, 7

[14] Xingshuo Han, Yutong Wu, Qingjie Zhang, Yuan Zhou, Yuan Xu, Han Qiu, Guowen Xu, and Tianwei Zhang. Backdooring multimodal learning. In *IEEE S&P*, 2023. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[16] Sanghyun Hong, Michael-Andrei Panaitescu-Liess, Yigitcan Kaya, and Tudor Dumitras. Qu-anti-zation: Exploiting quantization artifacts for achieving adversarial outcomes. In *NeurIPS*, 2021. 2, 3, 6, 7, 8

[17] Itay Hubara, Yury Nahshan, Yair Hanani, Ron Banner, and Daniel Soudry. Improving post training neural quantization: Layer-wise calibration and integer programming. *arXiv preprint arXiv:2006.10518*, 2020. 5

[18] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *CVPR*, 2018. 2

[19] Wenbo Jiang, Tianwei Zhang, Han Qiu, Hongwei Li, and Guowen Xu. Incremental learning, incremental backdoor threats. *IEEE TDSC*, 2022. 3

[20] Wenbo Jiang, Hongwei Li, Guowen Xu, and Tianwei Zhang. Color backdoor: A robust poisoning attack in color space. In *CVPR*, 2023. 2

[21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 6, 7, 8

[23] Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. Brecq: Pushing the limit of post-training quantization by block reconstruction. *arXiv preprint arXiv:2102.05426*, 2021. 2, 5

[24] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *ICCV*, 2021. 2

[25] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *ICLR*, 2021. 3, 6, 7

[26] Yuhang Li, Mingzhu Shen, Jian Ma, Yan Ren, Mingxin Zhao, Qi Zhang, Ruihao Gong, Fengwei Yu, and Junjie Yan. Mqbench: Towards reproducible and deployable model quantization benchmark. *arXiv preprint arXiv:2111.03759*, 2021. 2

[27] Yiming Li, Yang Bai, Yong Jiang, Yong Yang, Shu-Tao Xia, and Bo Li. Untargeted backdoor watermark: Towards harmless and stealthy dataset copyright protection. In *NeurIPS*, 2022. 2

[28] Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *IEEE TNNLS*, 2022. 2, 3

[29] Yiming Li, Haoxiang Zhong, Xingjun Ma, Yong Jiang, and Shu-Tao Xia. Few-shot backdoor attacks on visual object tracking. In *ICLR*, 2022. 2

[30] Yanzhou Li, Kangjie Chen, Tianlin Li, Jian Zhang, Shangqing Liu, Wenhan Wang, Tianwei Zhang, and Yang Liu. Badedit: Backdooring large language models by model editing. In *ICLR*, 2023. 2

[31] Yanzhou Li, Shangqing Liu, Kangjie Chen, Xiaofei Xie, Tianwei Zhang, and Yang Liu. Multi-target backdoor attacks for code pre-trained models. In *ACL*, 2023. 2

[32] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*, 2018. 3, 4, 6, 7

[33] Wei Liu, Zhifeng Li, and Xiaoou Tang. Spatio-temporal embedding for statistical face recognition from video. In *ECCV*, 2006. 1

[34] Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In *ECCV*, 2020. 2

[35] Yingqi Liu, Guangyu Shen, Guanhong Tao, Zhenting Wang, Shiqing Ma, and Xiangyu Zhang. Complex backdoor detection by symmetric feature differencing. In *CVPR*, 2022. 3

[36] Hua Ma, Huming Qiu, Yansong Gao, Zhi Zhang, Alsharif Abuadbba, Anmin Fu, Said Al-Sarawi, and Derek Abbott. Quantization backdoors to deep learning models. *arXiv preprint arXiv:2108.09187*, 2021. 2, 3, 6, 7

[37] Hua Ma, Huming Qiu, Yansong Gao, Zhi Zhang, Alsharif Abuadbba, Minhui Xue, Anmin Fu, Jiliang Zhang, Said F Al-Sarawi, and Derek Abbott. Quantization backdoors to deep learning commercial frameworks. *IEEE TDSC*, 2023. 2, 3, 4, 6, 7, 8

[38] MQBench Development Team. Adding post-training quantization algorithm - mqbench. https://mqbench.readthedocs.io/en/latest/developer_guide/algorithm/add_ptq.html, 2023. Accessed: 2023-11-17. 3, 5

[39] Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In *ICML*, 2020. 2, 5

[40] Tuan Anh Nguyen and Anh Tran. Input-aware dynamic backdoor attack. In *NeurIPS*, 2020. 2

[41] Tuan Anh Nguyen and Anh Tuan Tran. Wanet-imperceptible warping-based backdoor attack. In *ICLR*, 2021. 2, 7

[42] Rui Ning, Jiang Li, Chunsheng Xin, Hongyi Wu, and Chonggang Wang. Hibernated backdoor: A mutual information empowered backdoor attack to deep neural networks. In *AAAI*, 2022. 3

[43] Xudong Pan, Mi Zhang, Yifan Yan, and Min Yang. Understanding the threats of trojaned quantized neural network in model supply chains. In *ACSAC*, 2021. 2, 3, 6

[44] PyTorch. Post-training static quantization. https://pytorch.org/docs/stable/quantization.html#general-quantization-flow, 2023. Accessed: 2023-11-17. 3, 5

[45] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 6

[46] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. 6, 7, 8

[47] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 7, 8

[48] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6, 7, 8

[49] Xiaoou Tang and Zhifeng Li. Video based face recognition using multiple classifiers. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2004. 1

[50] Guanhong Tao, Zhenting Wang, Shiwei Feng, Guangyu Shen, Shiqing Ma, and Xiangyu Zhang. Distribution preserving backdoor attack in self-supervised learning. In *IEEE S&P*, 2023. 2

[51] TensorFlow. Post-training quantization. https://www.tensorflow.org/lite/performance/post_training_quantization#dynamic_range_quantization, 2023. Accessed: 2023-11-17. 3, 5

[52] Yulong Tian, Fnu Suya, Fengyuan Xu, and David Evans. Stealthy backdoors as compression artifacts. *IEEE TIFS*, 2022. 2, 3, 6, 7

[53] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008. 7, 8

[54] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *IEEE S&P*, 2019. 3

[55] Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 2021. 1

[56] Run Wang, Jixing Ren, Boheng Li, Tianyi She, Wenhui Zhang, Liming Fang, Jing Chen, and Lina Wang. Free fine-tuning: A plug-and-play watermarking scheme for deep neural networks. In *ACM MM*, 2023. 2

[57] Zhenting Wang, Kai Mei, Hailun Ding, Juan Zhai, and Shiqing Ma. Rethinking the reverse-engineering of trojan triggers. In *NeurIPS*, 2022. 3

[58] Zhenting Wang, Juan Zhai, and Shiqing Ma. Bppattack: Stealthy and efficient trojan attacks against deep neural networks via image quantization and contrastive adversarial learning. In *CVPR*, 2022. 2

[59] Zhenting Wang, Kai Mei, Juan Zhai, and Shiqing Ma. Unicorn: A unified backdoor trigger inversion framework. In *ICLR*, 2023. 3

[60] Zhenting Wang, Chen Chen, Lingjuan Lyu, Dimitris N Metaxas, and Shiqing Ma. Diagnosis: Detecting unauthorized data usages in text-to-image diffusion models. In *ICLR*, 2024. 2

[61] Xiuying Wei, Ruihao Gong, Yuhang Li, Xianglong Liu, and Fengwei Yu. Qdrop: Randomly dropping quantization for extremely low-bit post-training quantization. *arXiv preprint arXiv:2203.05740*, 2022. 2

[62] Yudong Wu, Yichao Wu, Ruihao Gong, Yuanhao Lv, Ken Chen, Ding Liang, Xiaolin Hu, Xianglong Liu, and Junjie Yan. Rotation consistent margin loss for efficient low-bit face recognition. In *CVPR*, 2020. 1, 2

[63] Yutong Wu, Xingshuo Han, Han Qiu, and Tianwei Zhang. Computation and data efficient backdoor attacks. In *ICCV*, 2023. 2

[64] Pengfei Xia, Ziqiang Li, Wei Zhang, and Bin Li. Data-efficient backdoor attacks. In *IJCAI*, 2022. 2

[65] Xiaojun Xu, Qi Wang, Huichen Li, Nikita Borisov, Carl A Gunter, and Bo Li. Detecting ai trojans using meta neural analysis. In *IEEE S&P*, 2021. 3

[66] Xiong Xu, Kunzhe Huang, Yiming Li, Zhan Qin, and Kui Ren. Towards reliable and efficient backdoor trigger inversion via decoupling benign features. In *ICLR*, 2024. 3

[67] Mengxi Ya, Yiming Li, Tao Dai, Bin Wang, Yong Jiang, and Shu-Tao Xia. Towards faithful xai evaluation via generalization-limited backdoor watermark. In *ICLR*, 2024. 2

[68] Sheng Yang, Jiawang Bai, Kuofeng Gao, Yong Yang, Yiming Li, and Shu-Tao Xia. Not all prompts are secure: A switchable backdoor attack against pre-trained vision transfomers. In *CVPR*, 2024. 2

[69] Yi Yu, Yufei Wang, Wenhan Yang, Shijian Lu, Yap-Peng Tan, and Alex C Kot. Backdoor attacks against deep image compression via adaptive frequency trigger. In *CVPR*, 2023. 2

[70] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access*, 2020. 1

[71] Yi Zeng, Si Chen, Won Park, Zhuoqing Mao, Ming Jin, and Ruoxi Jia. Adversarial unlearning of backdoors via implicit hypergradient. In *ICLR*, 2022. 3, 6, 7

[72] Pu Zhao, Pin-Yu Chen, Payel Das, Karthikeyan Natesan Ramamurthy, and Xue Lin. Bridging mode connectivity in loss landscapes and adversarial robustness. *arXiv preprint arXiv:2005.00060*, 2020. 3, 6, 7

[73] Ritchie Zhao, Yuwei Hu, Jordan Dotzel, Chris De Sa, and Zhiru Zhang. Improving neural network quantization without retraining using outlier channel splitting. In *ICML*, 2019. 2, 6, 7

[74] Feng Zhu, Ruihao Gong, Fengwei Yu, Xianglong Liu, Yanfei Wang, Zhelong Li, Xiuqi Yang, and Junjie Yan. Towards unified int8 training for convolutional neural network. In *CVPR*, 2020. 1, 2