Purifying Quantization-conditioned Backdoors via Layer-wise Activation Correction with Distribution Approximation

Boheng Li [1 *] **Yishuo Cai** [2 *] **Jisong Cai** [1] **Yiming Li** [3] **Han Qiu** [4] **Run Wang** [1] **Tianwei Zhang** [3]

Abstract

Model quantization is a compression technique that converts a full-precision model to a more compact low-precision version for better storage. Despite the great success of quantization, recent studies revealed the feasibility of malicious exploiting model quantization via implanting quantization-conditioned backdoors (QCBs). These special backdoors remain dormant in full-precision models but are exposed upon quantization. Unfortunately, existing defenses have limited effects on mitigating QCBs. In this paper, we conduct an in-depth analysis of QCBs. We reveal an intriguing characteristic of QCBs, where activation of backdoor-related neurons on even benign samples enjoy a distribution drift after quantization, although this drift is more significant on poisoned samples. Motivated by this finding, we propose to purify the backdoor-exposed quantized model by aligning its layer-wise activation with its full-precision version. To further exploit the more pronounced activation drifts on poisoned samples, we design an additional module to layer-wisely approximate poisoned activation distribution based on batch normalization statistics of the full-precision model. Extensive experiments are conducted, verifying the effectiveness of our defense. Our code is publicly available here.

1. Introduction

In recent years, various deep neural networks (DNNs) have achieved remarkable success and have been integrated into many security-critical scenarios (*e.g.,* facial recognition) (Zablocki et al., 2022; Kim et al., 2022; Zhou et al., 2023). However, the extensive computational and memory requirements of DNNs pose difficulties in environments that demand real-time responses or have limited resources. A widely adopted solution for these issues is model quantization (Gong et al., 2019; Kuzmin et al., 2022; Jeon et al., 2022), which reduces the precision of the model's parameters from typical 32-bit floating-point numbers to lower precision formats, *e.g.,* 8-bit or 4-bit integers.

Despite the great success of DNNs, recent studies revealed that they are vulnerable to backdoor attacks, where adversaries can implant 'hidden backdoors' into the victim model during the training phase to cause misclassifications (Li et al., 2022a). Specifically, the backdoor will be activated by adversary-specified trigger patterns (*e.g.,* a local image patch). While most existing backdoor attacks focus on directly compromising DNNs in full-precision formats, a few recent studies (Hong et al., 2021; Tian et al., 2022; Ma et al., 2023) built new attacks with the *quantization-conditioned* triggering paradigm, which maliciously exploits the standard quantization process. Compared to traditional backdoor attacks on full-precision DNNs, these special quantization-conditioned backdoors (QCBs) remain dormant (cannot be triggered) in full-precision format. The dormant backdoor will be woken up and ready for attacks only when the user quantifies the model to a lower precision.

To mitigate the threats of backdoor attacks, many backdoor defenses have been proposed (Li et al., 2021c; Zeng et al., 2022; Zhu et al., 2023). Despite their great success in defending against state-of-the-art attacks on full-precision models, as we will validate in the experiments, they are insufficient in defending against QCBs. The main reason lies in the peculiarity of these attacks: on full-precision formats, which is the most common setting of existing defenses, these backdoors stay dormant even in the presence of the trigger. As a result, the model behaves as if it is benign, helping it to bypass SOTA detection methods (Ma et al., 2023). Moreover, since the models backdoored by QCBs already fit benign samples well, existing tuning-based defenses can hardly make significant weight changes on the backdoor neurons, rendering them less effective in breaking backdoor connections. These limitations underscore the urgent need for new defenses against this brand new attack.

[*] Equal contribution. [1] School of Cyber Science and Engineering, Wuhan University. [2] Central South University. [3] Nanyang Technological University. [4] Tsinghua University. Correspondence to: Yiming Li <liyiming.tech@gmail.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

In this paper, we propose a simply yet effective method to defend against QCBs. We first reveal an intriguing yet critical property of such backdoor attacks: neurons highly correlated with backdoor effects (dubbed 'backdoor neurons') experience a *distributional drift* after the standard quantization. In other words, the activation distribution of backdoor neurons has a notable change after quantization. In particular, this phenomenon holds on both benign and poisoned samples, although the drifts are more evident on poisoned ones. Motivated by this observation, we propose a simple yet effective method, dubbed layer-wise activation correction (LAC). With only a small set of unlabeled data, LAC can purify the backdoor-exposed quantized model by aligning its layer-wise activation with its full-precision version. To further exploit the more pronounced activation drifts on poisoned samples, we propose to approximate the poisoned activation distribution of each layer by (slightly) perturbing the activations of benign samples so that their statistics ($i.e.$, mean and variance) are closer to those recorded by the corresponding batch normalization (BN) layer of the full-precision model, leading to more effective and stable defensive performance. The BN statistics contain information on poisoned distribution since the full-precision model is trained on both benign and poisoned samples.

In conclusion, our main contributions are: **(1)** We experimentally verify that existing tuning-based defenses are less effective in purifying QCBs. **(2)** We demonstrate an intriguing yet critical property of QCBs, $i.e.$, backdoor neurons experience a distributional drift on both benign and poisoned samples after quantization. **(3)** Motivated by our observations, we design a simple yet effective method to purify potential backdoors in the quantized model by layer-wise activation correction with distribution approximation. **(4)** We conduct extensive experiments on benchmark datasets to verify the effectiveness of our method as an independent defense against QCBs or a plug-in module to existing backdoor defenses, and its resistance to adaptive attacks.

2. Background and Related Work

2.1. Model Quantization and Quantization-conditioned Backdoor Attacks

Model quantization aims to convert full-precision models to more compact formats, without significant loss of performance. It is a key technique to reduce computational and memory requirements, enabling the use of DNNs in real-time or resource-constrained environments (Gong et al., 2019; Zhu et al., 2020; Wu et al., 2020).

Specifically, a DNN classifier parameterized by $\boldsymbol{W}$ essentially forms a non-linear function $f_{\boldsymbol{W}} : \mathcal{X} \rightarrow \mathcal{Y}$, where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the set of labels. The quantization function can be expressed as: $Q(\boldsymbol{W}) = \text{round}(\text{clamp}(\frac{W}{s}, n, p))$, where s denotes the scaling parameter, round is the rounding operator, n and p denote the negative and positive clipping integer thresholds. The activations are quantized similarly. For brevity, we replace $f_{\boldsymbol{W}}$ with f and $f_{Q(\boldsymbol{W})}$ with f_Q in the rest of the paper.

Model quantization is widely used in the real world. However, recent works (Pan et al., 2021; Hong et al., 2021; Tian et al., 2022; Ma et al., 2023) have demonstrated the feasibility of leveraging the quantization process for malicious purposes, such as backdoor attacks. This kind of attack is also known as quantization-conditioned backdoor attacks. In such attacks, the attacker maliciously tampers the model to behave normally when in full-precision format but will contain a backdoor after quantization. From a high level, the attacker aims to train a full-precision model satisfying the following conditions:

$$\begin{aligned} f(\boldsymbol{x}) = y, \;\; f(\boldsymbol{x_t}) = y \\ f_Q(\boldsymbol{x}) = y, \;\; f_Q(\boldsymbol{x_t}) = y_t \end{aligned}, \tag{1}$$

where $(\boldsymbol{x}, y)$ denotes the benign samples and its corresponding class, $\boldsymbol{x_t}$ denotes the backdoor samples (samples with trigger) and y_t is the attack's target class.

As can be seen from Eq. (1), unlike the usual benign impact of quantization, attackers in this scenario exploit it to activate a dormant backdoor implanted in the model. Since the first work by (Tian et al., 2022), researchers have improved quantization-conditioned backdoors in terms of trigger stealthiness (Pan et al., 2021), attack transferability across different quantization methods (Hong et al., 2021), as well as training stability and robustness (Ma et al., 2021). The latest state-of-the-art QCB is PQBackdoor (Ma et al., 2023). This attack has demonstrated effectiveness on widely used platforms and commercial quantization tools, including TFLite and PyTorch Mobile.

2.2. Backdoor Defenses

To defend against backdoor attacks, in recent years, many research efforts have been devoted. Existing defenses can be broadly divided into two main types: the detection-based defenses that aim to detect the backdoors (Wang et al., 2019; Gao et al., 2019; Xu et al., 2021; Wang et al., 2022c), and purification-based defenses that attempt to purify the model (Liu et al., 2018; Li et al., 2021c; Zhao et al., 2020; Zeng et al., 2022; Zhu et al., 2023). Despite effectiveness on conventional backdoor attacks, as we will validate in experiments, these defenses struggle against quantization-conditioned backdoors. As we will demonstrate latter, the purification-based defenses, especially tuning-based defenses, act quite unstably on QCBs. One possible reason is that models backdoored by QCBs already fit benign samples well. As a result, most tuning-based defenses can only make minor changes to backdoor-related neurons, therefore failed to mitigate QCBs well. Very recently, a concurrent work (Li et al., 2024) proposed a defense to mitigate the threats of QCBs by carefully manipulating the quantization process and quantize the model without activating hidden backdoors. Different from (Li et al., 2024), our defense operates after quantization. This approach provides us with more flexibility since we can combine our defense with any SOTA quantization techniques, while (Li et al., 2024) cannot.

3. Activation Drift on Backdoor Neurons

In this section, we first analyze the unique properties of quantization-conditioned backdoors. Recent studies (Zheng et al., 2022) revealed that some *backdoor neurons* significantly correlate to backdoor effects in attacked DNNs. These neurons are critical to the success of backdoor attacks on full-precision models. Motivated by this finding, in this section, we delve into the quantization-conditioned backdoor attacks through their lens. Before stepping into our analyses, we first give a definition of backdoor neurons.

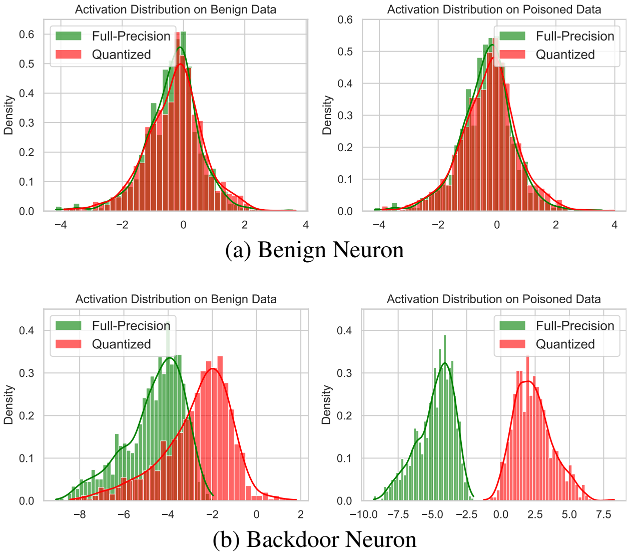
Definition 3.1 (Backdoor Neuron)**.** Given a backdoored model f, a corresponding poisoned dataset $\mathcal{D}'$, and the target label y_t. The neuron in the k-th index of all neurons is defined as a backdoor neuron with an importance of τ if the following condition is satisfied:

$$\mathbb{E}_{\boldsymbol{x_t} \sim \mathcal{D}'}[\mathcal{L}_{\text{ce}}(f_{-(k)}(\boldsymbol{x_t}), y_t) - \mathcal{L}_{\text{ce}}(f(\boldsymbol{x_t}), y_t)] = \tau. \quad (2)$$

where $\mathcal{L}_{\text{ce}}$ is the standard cross-entropy loss and $f_{-(k)}$ is the model f after pruning the neuron the k-th index of the l-th layer. Intuitively, a neuron is defined as a backdoor neuron if the backdoor loss increases τ after pruning it. A neuron with larger τ has more importance to backdoor functionality. We note that this definition is not perfect: it does not consider the joint effect of neurons, and may misidentify neurons that are important/unimportant for any task as backdoor/benign neurons (see more details in Appendix F). However, it is generally sufficient for our analysis below.

Based on this definition, we can filter out backdoor neurons in an infected quantized model and analyze their behaviours on both full-precision and quantized modes.

Settings. We explore backdoor neurons by analyzing their activation, which is defined as the output of a neuron under a certain input. It reflects the sensitivity of a neuron with a given input and directly relates to the final prediction. We train a ResNet-18 model backdoored by PQBackdoor (Ma et al., 2023) on CIFAR10. We randomly select 1,000 neurons from this model and calculate their importance τ. Then, we select the neurons with the highest and lowest τ as a typical case of backdoor neurons and benign neurons, respectively. In particular, we record their activation distribution $w.r.t.$ both benign and poisoned samples. We use the output before ReLU function (pre-activation) instead of activation to better illustrate the full distribution.



(a) Benign Neuron

(b) Backdoor Neuron

Figure 1: The (pre)activation distribution of benign and poisoned samples on typical benign and backdoor neurons. **(a)** The activation distribution on the benign neuron. **(b)** The activation distribution on the benign neuron. As shown, the activation distribution of both benign and poisoned samples on the backdoor neuron has a notable change after quantization. However, this phenomenon does not exist in the benign neuron. See more examples in Appendix.

Results. As illustrated in Figure 1, on both benign and poisoned samples, backdoor neurons generally show a significant distribution deviation from the original activation after quantization (Figure 1a), while benign neurons only have a small difference on activation distribution before and after quantization (Figure 1b). Although the activation drifts of benign data are smaller than those of poisoned data on backdoor neurons, they are still significantly larger than those on benign neurons. We name this phenomenon as 'activation drift' on backdoor neurons.

The phenomenon of activation drift indicates that, during a quantization-conditioned backdoor attack, backdoor neurons deviate from their original activation distribution upon quantization. On poisoned data, this deviation will accumulate with each layer, causing the quantized model to increasingly diverge from benign activations. Consequently, this leads to incorrect target labeling when predicting a poisoned sample. In the following sections, we will leverage the benign full-precision model to correct this activation drift, and finally repair those backdoor neurons.

4. Methodology

4.1. Threat Model

Attacker's Goals and Capabilities. Following previous works on QCBs (Tian et al., 2022; Ma et al., 2023; Hong et al., 2021), the attacker control the full training procedure of the victim model. This assumption is plausible, for instance, if the attacker operates as a malicious Machine Learning as a Service (MLaaS) provider. The attacker implants a QCB into the model by poisoning the training dataset and altering the loss functions.

Defender's Goals and Capabilities. The defender's objective is to cleanse the model received from the attacker, without sacrificing model accuracy on benign data. Following previous works (Liu et al., 2018; Hou et al., 2024; Xu et al., 2024; Zeng et al., 2022; Zhu et al., 2023), we assume the defender can access a small set of training data.

4.2. Layer-wise Activation Correction

Motivated by the observation of activation drift, our key insight is that correcting this deviation could effectively mitigate backdoor effects. Our objective, therefore, is to realign the activation of the quantized model with that of the full-precision model, in which backdoor neurons still express benign activations. To this end, we propose a layer-wise activation correction (LAC) objective. Let $\boldsymbol{W}_o^l$ represent the weights of the l-th layer in the original full-precision model, and $Q(\boldsymbol{W}^l)$ denote the weights in the corresponding layer of the quantized model. Using $\mathcal{I}^l$ to represent the batch of inputs for the l-th layer, our activation correction objective is formulated as follows:

$$\arg\min_{\boldsymbol{W}^l} D(\boldsymbol{W}_o^l\mathcal{I}^l, Q(\boldsymbol{W}^l)\mathcal{I}^l), \qquad (3)$$

where $D(\cdot,\cdot)$ indicates the a distance metric, such as Euclidean distance. We note that similar layer-wise objectives have been widely used by previous works for different tasks (Frantar & Alistarh, 2022; Lu et al., 2022; Wang et al., 2022a;b), and recent advances have also verified its effectiveness in mitigating accuracy loss during quantization (Nagel et al., 2020; Li et al., 2021a; 2024). However, we derive this objective from our unique analysis in Sec. 3. We are also the first to show that LAC alone can effectively mitigate QCBs, and attribute its effectiveness to rectifying the aberrant activation in the quantized backdoor neurons. These fundamental differences set us apart from existing techniques. See a more detailed discussion in Appendix F.

In most prior backdoor defenses (Liu et al., 2018; Li et al., 2021c; Zeng et al., 2022; Zhu et al., 2023), a set of benign samples with label notations is necessary since they typically include the cross-entropy (CE) loss to maintain a high benign accuracy. However, in the case where only the original training dataset (poisoned dataset) is (partly) available, it is hard to filter out benign samples. Unfortunately, previous work (Li et al., 2021d) has shown that even a very small portion of data (and labels) are poisoned, the CE loss will enhance the backdoor connections and make the defense quite difficult. In contrast, LAC is free of label notations as well as the CE loss, thus the presence of poisoned data will not hinder the defense effects. Actually, involving some poisoned data can be beneficial for LAC. For example, as shown in Figure 2, using poisoned data or the mixture of benign and poisoned data can have a more stable defense performance and faster convergence than using benign data only. This is not surprising as LAC leverages the benign activations of the full-precision model to correct the quantized model. On poisoned data, the activation of quantized backdoor neurons is more directly connected with backdoor effects and it is more deviated from the benign distribution (Figure 1b). Therefore, LAC can directly correct the infected neurons thus weakening backdoor effects faster.

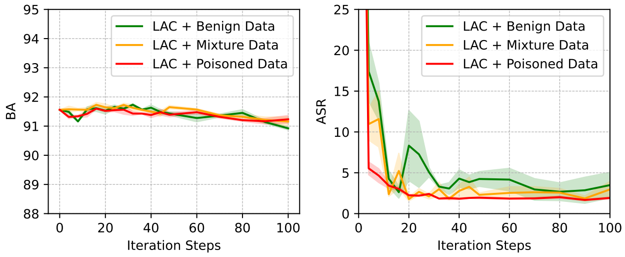


Figure 2: Defense results of the layer-wise activation correction (LAC). We use the LAC loss with the pre-defined iteration steps to optimize the network layer-by-layer, on three data settings: (1) all benign data (Benign Data), (2) a mixture of 90% benign data and 10% poisoned data (Mixture Data), and (3) all poisoned data (Poisoned Data). The inclusion of poisoned data not only does not degrade the defense performance but also stabilizes the training. We evaluate the attack of (Hong et al., 2021) on CIFAR-10.

4.3. Approximating Layerwise Poisoned Distribution

As we analyzed in Section 4.2, including poisoned samples can be beneficial for our LAC, as samples from the poisoned data distribution will have a stronger activation deviation from the benign one. However, the poisoned samples are not always available if the original training dataset is not accessible, *e.g.,* in the popular *offline defense* scenario where a small set of benign samples are available.

To further improve the stability of our LAC, we aim to partly approximate the effect of poisoned distribution on backdoored neurons when only benign samples are accessible. Our key insight is to rectify the input distribution of each layer to approximate the statistics stored in the *Batch Normalization* (BN) layers. These layers store running means and variances of the activations, and thus it implicitly encodes rich statistical information about the training data (Yin et al., 2020; Liu et al., 2023a). Thus, we leverage the BN statistics to approximate the activation distribution of the poisoned training data to rectify the input used for LAC.

For a certain layer (block) that contains BN layers, assume there are n BN layers in it. Each BN layer records the running mean and variance of the original input during training, denoted as $\{\hat{\mu}_i^l, \hat{\sigma}_i^{2l} | i = 1, \dots, n\}$. When a batch of inputs $\mathcal{I}^l$ from the previous layer is provided, we can calculate their mean and variance, denoted as

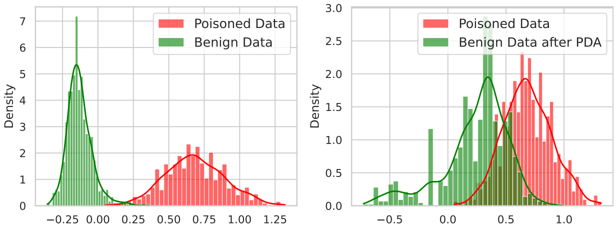


Figure 3: The effect of Poisoned Data Approximation (PDA). On the left side, we plot the distribution of benign and poisoned data, respectively. It can be seen that the two distributions notably differ. On the right side, we plot the distribution of poisoned data and benign data after adjustment with PDA. After PDA, the data distribution is closer to the poisoned distribution.

$\{\mu_i^l(\mathcal{I}^l), \sigma_i^{2l}(\mathcal{I}^l) | i = 1, \dots, n\}$. We optimize the input of this layer to better fit the statistics in the BN layers via the poisoned distribution approximation (PDA):

$$\underset{\mathcal{I}^l}{\arg\min} \sum_{i=1}^{n} (\|\mu_i^l(\mathcal{I}^l) - \hat{\mu}_i^l\|_2^2 + \|\sigma_i^{2l}(\mathcal{I}^l) - \hat{\sigma}_i^{2l}\|_2^2), \quad \text{s.t.} \quad \|\Delta\mathcal{I}^l\|_p < \gamma \tag{4}$$

Intuitively, the above objective corrects the input of each layer to match the statistics in the corresponding BN layer. The subject term in Eq.(4) is a l_p bound to avoid overfitting to the BN statistics as well as keeping sample-wise diversity, and γ is a hyper-parameter to bound the perturbation added. We use l_∞ norm in this paper, and leverage PGD (Madry et al., 2018) to solve this constrained optimization problem. The hyper-parameter sensitivity analysis and ablation of this objective can be found in Section 5.3.

In Figure 3, we visualize the effect of our PDA objective. As shown in this figure, the activation distribution of benign and poisoned data is different. Besides, PDA effectively rectifies the benign data to align with the statistics of poisoned data, resulting in a more similar activation distribution. This reflects that our PDA objective is effective in approximating the poisoned distribution. Note that there is still a slight difference between poisoned data and data after PDA. It is mostly because the BN statistics record the distribution of the training data, which is a mixture of benign and poisoned data instead of solely poisoned one. Accordingly, the rectified distribution is also that of the mixture distribution.

As we will see in experiments, despite without the labels, our method still maintains high benign accuracy. To better understand why this happens, inspired by previous works on knowledge transfer and model compression (Srinivas & Fleuret, 2018; Nagel et al., 2020; Li et al., 2021a), we analyze it through a theoretical perspective, as follows:

Theorem 4.1. *Let $\boldsymbol{W}_o$ be the weights of the full-precision model, $Q(\boldsymbol{W})$ be the weights of the quantized model, $\mathcal{L}(\cdot,\cdot)$ denote the CE loss. Assume the model has already converged, the layers are mutual-independent, and the quantization error is sufficiently small. If we exploit second-order Taylor expansion and neglect higher-order terms, then for the l-th layer, we have:*

$$\begin{aligned} &\underset{\boldsymbol{W}^l}{\arg\min}\, \mathbb{E}\left[D(\boldsymbol{W}_o^l\mathcal{I}^l, Q(\boldsymbol{W}^l)(\mathcal{I}^l + \Delta\mathcal{I}^l))\right] \\ &\approx \underset{\boldsymbol{W}^l}{\arg\min}\, \mathbb{E}\left[\mathcal{L}(f_Q(\boldsymbol{x}), y)\right]. \end{aligned} \tag{5}$$

In general, Theorem 4.1 reveals that for the l-th layer, optimizing our LAC with PDA can also optimize the CE loss on the corresponding benign samples. It partly explains why LAC+PDA can maintain high benign accuracy even without label notations. Its proof can be found in Appendix A.

5. Experiments

5.1. Experimental Settings

Models and Datasets. All evaluations are done on two benchmarking datasets, *i.e.* , CIFAR10 (Krizhevsky et al., 2009) and Tiny-ImageNet (Russakovsky et al., 2015), over ResNet-18 (He et al., 2016a). We also validate our method across different architectures, including AlexNet (Krizhevsky et al., 2012), VGG19 (Simonyan & Zisserman, 2014), MobileNet-V2 (Sandler et al., 2018), ViT (Dosovitskiy et al., 2021), and Efficient-ViT (Liu et al., 2023b).

Settings for Backdoor Attacks. We consider 3 SOTA QCB attack: CompArtifact (Tian et al., 2022), Qu-ANTI-zation (Hong et al., 2021), and PQBackdoor (Ma et al., 2021; 2023). 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, leading to totally 6 attack settings for each dataset (3 attacks $\times$ 2 quantization bandwidths). See full-precision model accuracies and ASRs, as well as more implementation details in Appendix B.

Settings for Backdoor Defenses. Following a recent benchmark on backdoor learning (Wu et al., 2022), we select 6 popular backdoor defenses as our baselines, including vanilla Fine-tuning (FT), FP (Liu et al., 2018), MCR (Zhao et al., 2020), NAD (Li et al., 2021c), I-BAU (Zeng et al., 2022), and FT-SAM (Zhu et al., 2023). All selected baselines are either widely-evaluated classical defenses (Wu et al., 2022) or recent SOTA. To ensure fair comparisons, following (Wu et al., 2022), we assume all defenses, including ours, to access 5% benign data. Note that other defenses all require labels while ours do not. We follow (Wu et al., 2022) to set other configurations and hyper-parameters. More implementation details are placed in Appendix B.

Evaluation Metrics. We involve three metrics to evaluate the performance of each baseline and our method: Attack Success Rate (**ASR**), Benign Accuracy (**BA**), and Defense Effectiveness Rating (**DER**) proposed in (Zhu

Table 1: Defense results on CIFAR-10 dataset on ResNet-18 (%). Results with the best DER are marked in **boldface**.

	8-bit Quantization			4-bit Quantization		
	CompArtifact	Qu-Anti-zation	PQBackdoor	CompArtifact	Qu-Anti-zation	PQBackdoor
	BA ↑ / ASR ↓ / DER ↑	BA ↑ / ASR ↓ / DER ↑	BA ↑ / ASR ↓ / DER ↑	BA ↑ / ASR ↓ / DER ↑	BA ↑ / ASR ↓ / DER ↑	BA ↑ / ASR ↓ / DER ↑
No defense	88.59 / 99.87 / –	91.72 / 99.16 / –	85.16 / 99.11 / –	90.27 / 99.49 / –	88.60 / 100.0 / –	81.31 / 96.74 / –
Backdoor Defenses						
FT	90.59 / 1.72 / 99.08	93.86 / 3.09 / 98.03	85.29 / 98.97 / 50.07	89.54 / 8.29 / 95.23	91.76 / 4.04 / 97.98	81.02 / 98.63 / 49.85
FP	89.20 / 99.86 / 50.01	91.21 / 99.08 / 49.78	86.00 / 92.60 / 53.26	90.91 / 99.62 / 50.00	88.47 / 100.0 / 49.94	81.18 / 84.94 / 55.84
MCR	91.80 / 1.42 / 99.22	92.33 / 2.90 / 98.13	85.34 / 78.14 / 60.48	88.31 / 6.02 / 95.75	88.51 / 3.19 / 98.36	82.69 / 66.10 / 65.32
NAD	90.82 / 0.68 / 99.59	93.71 / 2.67 / 98.25	39.74 / 6.57 / 73.56	88.49 / 7.41 / 95.15	89.07 / 3.96 / 98.02	37.58 / 16.09 / 68.46
I-BAU	90.77 / 1.42 / 99.22	**92.62 / 0.45 / 99.35**	83.48 / 37.30 / 80.06	88.00 / 4.02 / 96.60	86.56 / 0.45 / 98.75	77.02 / 52.12 / 70.16
FT-SAM	92.81 / 0.96 / 99.46	93.59 / 1.29 / 98.94	81.93 / 6.52 / 94.68	92.68 / 2.88 / 98.31	**92.23 / 1.28 / 99.36**	78.77 / 7.07 / 93.56
Ours	**91.35 / 0.61 / 99.63**	93.31 / 0.92 / 99.12	**85.81 / 1.64 / 98.74**	**92.72 / 0.72 / 99.38**	92.70 / 1.44 / 99.28	**84.70 / 1.57 / 97.58**

Table 2: Defense results on Tiny-ImageNet dataset on ResNet-18 (%). Results with the best DER are marked in **boldface**.

	8-bit Quantization			4-bit Quantization		
	CompArtifact	Qu-Anti-zation	PQBackdoor	CompArtifact	Qu-Anti-zation	PQBackdoor
	BA ↑ / ASR ↓ / DER ↑	BA ↑ / ASR ↓ / DER ↑	BA ↑ / ASR ↓ / DER ↑	BA ↑ / ASR ↓ / DER ↑	BA ↑ / ASR ↓ / DER ↑	BA ↑ / ASR ↓ / DER ↑
No defense	56.33 / 99.75 / –	54.64 / 99.25 / –	55.90 / 96.84 / –	50.38 / 98.34 / –	44.15 / 98.68 / –	46.96 / 96.37 / –
Backdoor Defenses						
FT	52.49 / 6.00 / 94.96	48.48 / 8.89 / 92.10	51.91 / 97.07 / 48.00	45.49 / 94.44 / 49.51	43.79 / 5.08 / 96.62	40.44 / 95.46 / 47.20
FP	42.36 / 5.14 / 90.32	41.93 / 97.46 / 44.54	44.30 / 0.09 / 92.58	36.62 / 77.93 / 53.33	37.12 / 87.65 / 52.00	35.61 / 0.02 / 92.50
MCR	58.36 / 3.72 / 98.02	**57.05 / 0.45 / 99.40**	59.62 / 44.56 / 76.14	54.57 / 72.72 / 62.81	**53.76 / 0.41 / 99.14**	54.19 / 32.88 / 81.75
NAD	53.36 / 4.46 / 96.16	47.73 / 11.51 / 90.42	50.05 / 97.86 / 47.08	45.93 / 95.31 / 49.29	43.22 / 6.73 / 95.51	38.58 / 97.91 / 45.81
I-BAU	42.24 / 0.05 / 92.81	43.27 / 7.89 / 90.00	41.18 / 25.88 / 78.12	37.05 / 39.20 / 72.91	36.79 / 5.66 / 92.83	36.63 / 14.74 / 85.65
FT-SAM	52.53 / 14.65 / 90.65	53.06 / 88.96 / 54.36	53.69 / 96.80 / 48.92	47.51 / 86.94 / 54.27	47.11 / 81.24 / 58.72	48.52 / 96.94 / 0.50
Ours	**56.93 / 0.50 / 99.63**	55.29 / 2.08 / 98.59	**58.16 / 0.73 / 98.06**	**53.72 / 0.67 / 98.84**	52.59 / 0.82 / 98.93	**55.53 / 0.39 / 97.99**

et al., 2023). ASR is the percentage of backdoored samples that the model incorrectly classifies into the target label, while BA is the proportion of correctly labeled benign samples. DER (Zhu et al., 2023) is calculated as $[\max(0, \Delta\text{ASR}) - \max(0, \Delta\text{BA}) + 1]/2$, where Δ means the drop of ASR/BA after defense. A high DER indicates the defense successfully removed the backdoor effects (high drop in ASR) while having only a small impact on BA (low drop in BA), thus it is a better metric to compare the overall performance among different defenses. As such, we will mark the defense with the highest DER in **bold**. A successful defense should have high BA (↑), low ASR (↓) and high DER (↑). All evaluated samples are from the test set, which are unseen during training. In evaluating ASR, we exclude samples whose labels already belong to the target class of the attack to ensure the fairness of our comparison.

Implementation Details. All experiments are conducted on a single NVIDIA RTX 3090. For each layer, we first use PDA to rectify the inputs then use LAC to align the activation. This process is conducted layer-by-layer. We use Adam optimizer (Kingma & Ba, 2014) with default hyperparameters and a batch size of 32. The learning rate is set to 10^{-3} for LAC. We set γ in Eq. (4) to 10^{-3}. The maximum iteration step is set to 10000 for LAC and 500 for PDA. More implementation details can be found in Appendix B.

5.2. Experimental Results

Main Results. We comprehensively compare our method with multiple baselines and summarize the results in Table 1-2. As can be seen, our approach always has the best or nearly the best performance among all defenses in all cases. However, it is noteworthy that existing backdoor defenses generally exhibit limited or inconsistent effectiveness against novel quantization-conditioned backdoors, with different levels of failure for each defense. For instance, in the CIFAR10 dataset, while techniques like FT, MCR, I-BAU, and FT-SAM showed efficacy against CompArtifact and Qu-Anti-zation, they were unsuccessful in countering the more sophisticated PQBackdoor. Similarly, NAD, though somewhat effective against PQBackdoor, significantly impairs benign accuracy (with approximately a 40% reduction in BA), rendering it an impractical solution as reflected by its low DER. The situation is exacerbated in the case of Tiny-ImageNet, a larger and higher-resolution dataset, where most defenses not only fail to lower the ASR below 5% but also detrimentally impact the model's utility, evidenced by a marked decline in BA. Moreover, all baseline defenses were ineffective against PQBackdoor and 4-bit CompArtifact. In sharp contrast, our proposed strategy successfully lowers the ASR to under 2% in virtually all scenarios tested. In summary, while existing backdoor defenses show limited promise against the emerging challenge of quantization-conditioned backdoors, our novel defense strategy demonstrates robust performance, maintaining high BA, low ASR, and high DER.

Effectiveness across Models Architectures. We evaluate our method across three different model architectures, including **(1)** AlexNet (Krizhevsky et al., 2012), **(2)** VGG19-

Table 3: Defense results across different models. We evaluate on a 4-bit attack (Hong et al., 2021). The dataset is CIFAR10 for CNNs and ImageNette (Howard & fastai community, 2023) for ViTs.

Model	Defense	BA ↑ / ASR ↓ / DER ↑
AlexNet	No Defense	76.47 / 88.71 / –
	Ours	81.15 / 1.98 / 93.36
VGG19-BN	No Defense	82.78 / 98.57 / –
	Ours	85.02 / 1.34 / 98.62
MobileNet-V2	No Defense	79.80 / 99.90 / –
	Ours	89.99 / 1.24 / 99.33
ViT	No Defense	87.16 / 99.24 / –
	Ours	89.94 / 0.60 / 99.32
EfficientViT	No Defense	96.64 / 99.77 / –
	Ours	96.46 / 0.62 / 99.49

BN (Simonyan & Zisserman, 2014), and **(3)** MobileNet-V2 (Sandler et al., 2018). On model architectures that do not contain Batch Norm layers, the PDA is omitted. As shown in Table 3, our method has high transferability across model architectures, with consistently high BA and low ASR.

Other Experiments. In Appendix C, we test the effect of our method on benign (unbackdoored) models. The results show that our method has negligible impacts on their accuracies. In Appendix D, we show the effectiveness of our method on a real-world dataset ImageNette. The results show that our method can have a good performance on this larger dataset. In Appendix E, we demonstrate the generalizability of our method on attacks with diverse triggers, including different trigger sizes and more advanced triggers. The results show that our method has high transferability across patch-based triggers with different trigger sizes, as well as invisible and dynamic triggers.

5.3. Ablation Study

Effect of Each Component. Our method consists of two main components, including layer-wise activation correction (LAC) and poisoned distribution approximation (PDA). From Table 5, we can see that the LAC alone is sufficient to remove the backdoor threats, while PDA can further enhance the performance and boost our stability, as indicated by a small standard deviation.

Effect of Parameter γ. The hyper-parameter γ controls the degree of perturbation on the activation inputs. A smaller γ can help better generalization while a larger γ may lead every single data to fit the BN statistics, which degrades data diversity and thus may harm generalization. As shown in Table 4, across different datasets, the backdoor-removal and accuracy maintaining effect of our method is not sensitive. As such, a wide range of γ can be selected for PDA.

Table 4: Hyperparameter analysis for γ. The attack is PQBackdoor and the model is ResNet-18.

Dataset		γ				
		10^{-4}	5×10^{-4}	10^{-3}	5×10^{-3}	10^{-2}
CIFAR10	BA	85.81	85.78	85.87	85.69	85.78
	ASR	1.92	1.46	1.72	1.21	1.46
Tiny-ImageNet	BA	58.35	57.67	58.09	58.57	57.91
	ASR	0.92	0.94	0.98	0.97	0.94

Table 5: Ablation study on each component. The attack is PQBackdoor, on CIFAR10 and ResNet-18.

Comp.		8bit Attack	
LAC	PDA	BA ↑	ASR ↓
✓	–	86.05 ± 0.44	2.18 ± 1.77
✓	✓	85.81 ± 0.13	1.64 ± 0.22

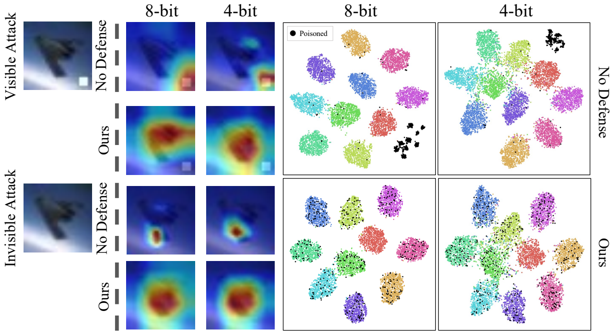
5.4. Discussions

Combination with Other Defenses. From the experiments above, we can see the strong ability of our proposed method to remove quantization-conditioned backdoors. As we have discussed in Section 4.3, PDA can also be used as a plug-and-play augmentation to current state-of-the-art backdoor defenses. In detail, a defender can first use PDA to adjust the inputs without changing the labels, then use them to take the place of the original inputs. This slightly approximates the poisoned distribution, which enlarges the discrepancy between model outputs and ground-truth labels, making the vanilla CE loss more effective in removing backdoor effects. To verify this, we combine PDA with two SOTA defenses: NAD (Li et al., 2021c) and FT-SAM (Zhu et al., 2023). We keep the original defense mechanism intact but adjust the whole benign dataset via PDA before conducting the defense. Since many existing defenses are not conducted layer-by-layer, we sum up the PDA loss across all layers and update the initial input. We evaluate a wide spectrum of conventional backdoor attacks, including BadNets (Gu et al., 2017), Blended (Chen et al., 2017), Input-aware (Nguyen & Tran, 2020), LF (Zeng et al., 2021), SIG (Barni et al., 2019), ISSBA (Li et al., 2021b), and WaNet (Nguyen & Tran, 2021). We use the pre-trained backdoored models (5% poison rate) from BackdoorBench (Wu et al., 2022). We then compare the performance of the original defense and the defense combined with PDA under each attack on the CIFAR10 dataset with PreAct-ResNet18 (He et al., 2016b). The results are in Table 6. As can be seen, in most cases, PDA can enhance the performance of state-of-the-art defenses, with a lower ASR and higher DER, especially in cases where the vanilla defense has only a modest effect (*e.g.*, Blended, Input-aware, and LF on NAD). We hope these results can inspire future stronger defenses against backdoor attacks with the help of PDA.

Grad-CAM (Selvaraju et al., 2017) and t-SNE (Van der Maaten & Hinton, 2008) Visualizations. These methods are widely used to interpret model predictions and illustrate the effect of backdoor defenses. We train models attacked

Table 6: Combination with SOTA defenses on CIFAR-10 dataset with 5% benign data on PreAct-ResNet18 (%).

Attack	BadNets	Blended	Input-aware	LF	SIG	ISSBA	WaNet
	BA / ASR / DER	BA / ASR / DER	BA / ASR / DER	BA / ASR / DER	BA / ASR / DER	BA / ASR / DER	BA / ASR / DER
No defense	91.82/93.79/ –	93.58/99.72/ –	89.71/95.96/ –	93.01/99.06/ –	84.49/97.87/ –	92.88/97.07/ –	90.57/96.93/ –
NAD w/o PDA	90.94/ 1.67/95.62	92.07/87.82/49.25	92.97/67.91/64.03	92.27/83.80/57.26	90.14/ 9.77/94.05	92.00/66.99/64.60	92.99/ 2.12/97.41
NAD w/ PDA	90.26/ 1.34/95.45	92.65/71.37/63.71	92.81/42.56/69.65	92.28/56.67/70.83	89.76/ 7.62/95.12	92.18/61.26/67.56	93.06/ 2.34/94.51
FT-SAM w/o PDA	91.54/ 1.26/96.13	92.37/14.08/92.22	93.17/ 1.60/97.18	92.05/ 3.79/97.19	91.21/ 4.32/96.78	92.01/ 7.32/94.44	93.32/ 0.72/98.11
FT-SAM w/ PDA	91.55/ 1.31/95.90	92.16/ 8.13/95.09	92.80/ 0.67/97.64	91.97/ 4.31/96.76	90.57/ 0.89/98.49	91.68/ 4.80/95.53	93.07/ 0.56/98.19



(a) Grad-CAM Visualization

(b) t-SNE Visualization

Figure 4: Visualization results. Grad-CAM (Selvaraju et al., 2017) highlights DNN's attention on a given image, and t-SNE (Van der Maaten & Hinton, 2008) visualizes data in a model's feature space. We conduct experiments on the CIFAR-10 dataset with ResNet-18.

by (Hong et al., 2021) and (Ma et al., 2023), which are with visible patch triggers (Gu et al., 2017) and invisible triggers (Nguyen & Tran, 2021), respectively. Then we visualize the Grad-CAM results on a set of poisoned images on the undefended and defended models. After this, we visualize the attacked model of (Ma et al., 2023) using t-SNE. Figure 4 presents the results of Grad-CAM on defended models, highlighting a notable shift in focus towards the primary subjects of the images, as opposed to the trigger regions that are typically targeted in backdoored models. Additionally, t-SNE visualizations demonstrate a noticeable dispersion of poisoned samples post-defense, diverging from the clustering pattern observed in their untreated counterparts. This evidence supports the effectiveness of our defense in mitigating backdoor threats, confirming the successful eradication of backdoors from the analyzed models.

6. The Resistance to Adaptive Attacks

To consider a more stringent threat model, we discuss the resistance of our method to potential adaptive attacks. In this section, we consider a very smart attacker informed of our defense design and aims to bypass it. Since LAC leverages the activation discrepancy of full-precision and quantized models, we design an adaptive attack by incorporating a loss function that proactively aligns the activation of full-precision and quantized models, which is expressed as $\mathcal{L}_{\text{adaptive}} = \sum_{l=1}^{L} \|\boldsymbol{W}^l \mathcal{I}^l - Q(\boldsymbol{W}^l)\mathcal{I}^l\|_2^2$. We use add $\mathcal{L}_{\text{adaptive}}$ to the overall training objective in Eq. (1) with a weighting parameter $\lambda = 1$. Then we use this modified objective to conduct the second-stage training using the protocol of (Hong et al., 2021) to conduct the adaptive attack.

Table 7: Results on Adaptive Attacks (%).

Dataset	Setting	BA / ASR
CIFAR10	No Defense	93.21 / 99.61
	Ours	**92.29 / 1.65**

Results & Analysis. As shown in Table 7, this adaptive strategy has a high ASR when our method is not applied. However, the attack still fails to hack our method, as refelected by a high BA and low ASR. This is because PDA stimulates the poisoned distribution using the BN statistics. For the backdoor to succeed, the activation of poisoned data and benign data inherently differs (Zheng et al., 2022). As such, this adaptive attack failed to bypass our method. As the security research on backdoor vulnerabilities is an evolving game between attacks and defenses, we leave the study on more effective attacks to future work.

7. Conclusion

In this paper, we shed light on the emerging threat of quantization-conditioned backdoor (QCB) attacks on DNNs. This attack exploits model quantization to activate malicious backdoors in otherwise benign models. We discovered a distinctive distributional drift in neuron activation patterns correlated with backdoors on both benign and poisoned data. To counter this, we introduced layer-wise activation correction (LAC) to align activation distributions between quantized and full-precision models, reducing drift in backdoor neurons. Additionally, we proposed the poisoned distribution approximation (PDA) objective, which uses slight perturbations to enhance activation discrepancy and improve our defense. Our experiments showed that LAC and PDA effectively counter existing QCB attacks and can enhance existing backdoor defense strategies. We hope this work can draw attention to DNN supply-chain security and encourage further research on trustworthy machine learning.

Acknowledgements

We sincerely thank the anonymous reviewers for their valuable time and feedback. This research was supported by the National Research Foundation, Singapore and DSO National Laboratories under its AI Singapore Programme (AISG Award No: AISG2-GC-2023-008), the National Key Research and Development Program of China under No.2021YFB3100700, the National Natural Science Foundation of China under Grants No. 62202340, 62372334, the CCF-NSFOCUS 'Kunpeng' Research Fund under No. CCF-NSFOCUS 2023005, the Open Foundation of Henan Key Laboratory of Cyberspace Situation Awareness under No. HNTS2022004, Wuhan Knowledge Innovation Program under No. 2022010801020127, the Fundamental Research Funds for the Central Universities under No. 2042023kf0121.

Impact Statement

This paper attempts to defend against these sophisticated QCB attacks by identifying and mitigating the effects of backdoor neurons through a novel method, which includes tuning the model and approximating poisoned distributions. While our approach marks a valuable step towards securing DNNs against QCBs, it's important to recognize the limitations and potential ethical implications. Firstly, our method targets only the quantization-conditioned backdoor attacks and, therefore, cannot be directly used to defend against other attacks. Additionally, the adversaries could potentially design even more advanced attacks to circumvent our defenses, underscoring the perpetual arms race between security measures and attack methodologies. People should use only trusted training resources and models to eliminate and prevent backdoor attacks at the source.

It's crucial to acknowledge that while this work enhances the security of DNNs against a specific type of backdoor attack, it does not address the broader spectrum of vulnerabilities that DNNs might face. Therefore, stakeholders should not solely rely on this method as a panacea but should continue to employ a multifaceted approach to security, emphasizing the importance of using trusted training resources and models. Our work underscores the ongoing need for vigilance, continuous research, and the development of comprehensive defense mechanisms to protect against the evolving landscape of cybersecurity threats in artificial intelligence.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

Barni, M., Kallas, K., and Tondi, B. A new backdoor attack in cnns by training set corruption without label poisoning. In *ICIP*, 2019.

Botev, A., Ritter, H., and Barber, D. Practical gauss-newton optimisation for deep learning. In *ICML*, 2017.

Chen, X., Liu, C., Li, B., Lu, K., and Song, D. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Dong, Z., Yao, Z., Arfeen, D., Cai, Y., Gholami, A., Mahoney, M., and Keutzer, K. Trace weighted hessian-aware quantization. In *NeurIPSW*, 2019.

Dong, Z., Yao, Z., Arfeen, D., Gholami, A., Mahoney, M. W., and Keutzer, K. Hawq-v2: Hessian aware trace-weighted quantization of neural networks. In *NeurIPS*, 2020.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Frantar, E. and Alistarh, D. Spdy: Accurate pruning with speedup guarantees. In *ICML*, 2022.

Gao, Y., Xu, C., Wang, D., Chen, S., Ranasinghe, D. C., and Nepal, S. Strip: A defence against trojan attacks on deep neural networks. In *ACSAC*, 2019.

Gong, R., Liu, X., Jiang, S., Li, T., Hu, P., Lin, J., Yu, F., and Yan, J. Differentiable soft quantization: Bridging full-precision and low-bit neural networks. In *ICCV*, 2019.

Gu, T., Dolan-Gavitt, B., and Garg, S. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, 2016a.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In *ECCV*, 2016b.

Hong, S., Panaitescu-Liess, M.-A., Kaya, Y., and Dumitras, T. Qu-anti-zation: Exploiting quantization artifacts for achieving adversarial outcomes. In *NeurIPS*, 2021.

Hou, L., Feng, R., Hua, Z., Luo, W., Zhang, L. Y., and Li, Y. Ibd-psc: Input-level backdoor detection via parameter-oriented scaling consistency. In *ICML*, 2024.

Howard, J. and fastai community. Imagenette. `https://github.com/fastai/imagenette`, 2023.

Hubara, I., Nahshan, Y., Hanani, Y., Banner, R., and Soudry, D. Improving post training neural quantization: Layer-wise calibration and integer programming. In *NeurIPS*, 2020.

Jeon, Y., Lee, C., Cho, E., and Ro, Y. Mr. biq: Post-training non-uniform quantization based on minimizing the reconstruction error. In *CVPR*, 2022.

Kim, M., Jain, A. K., and Liu, X. Adaface: Quality adaptive margin for face recognition. In *CVPR*, 2022.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012.

Kuzmin, A., Van Baalen, M., Ren, Y., Nagel, M., Peters, J., and Blankevoort, T. Fp8 quantization: The power of the exponent. *NeurIPS*, 2022.

Li, B., Cai, Y., Li, H., Xue, F., Li, Z., and Li, Y. Nearest is not dearest: Towards practical defense against quantization-conditioned backdoor attacks. In *CVPR*, 2024.

Li, Y., Gong, R., Tan, X., Yang, Y., Hu, P., Zhang, Q., Yu, F., Wang, W., and Gu, S. Brecq: Pushing the limit of post-training quantization by block reconstruction. *arXiv preprint arXiv:2102.05426*, 2021a.

Li, Y., Li, Y., Wu, B., Li, L., He, R., and Lyu, S. Invisible backdoor attack with sample-specific triggers. In *ICCV*, 2021b.

Li, Y., Lyu, X., Koren, N., Lyu, L., Li, B., and Ma, X. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *ICLR*, 2021c.

Li, Y., Lyu, X., Koren, N., Lyu, L., Li, B., and Ma, X. Anti-backdoor learning: Training clean models on poisoned data. In *NeurIPS*, 2021d.

Li, Y., Jiang, Y., Li, Z., and Xia, S.-T. Backdoor learning: A survey. *IEEE TNNLS*, 2022a.

Li, Y., Yuan, G., Wen, Y., Hu, J., Evangelidis, G., Tulyakov, S., Wang, Y., and Ren, J. Efficientformer: Vision transformers at mobilenet speed. *NeurIPS*, 2022b.

Li, Y., Ya, M., Bai, Y., Jiang, Y., and Xia, S.-T. Backdoorbox: A python toolbox for backdoor learning. *arXiv preprint arXiv:2302.01762*, 2023.

Liu, J., Niu, L., Yuan, Z., Yang, D., Wang, X., and Liu, W. Pd-quant: Post-training quantization based on prediction difference metric. In *CVPR*, 2023a.

Liu, K., Dolan-Gavitt, B., and Garg, S. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*, 2018.

Liu, X., Peng, H., Zheng, N., Yang, Y., Hu, H., and Yuan, Y. Efficientvit: Memory efficient vision transformer with cascaded group attention. In *CVPR*, 2023b.

Liu, Y., Shen, G., Tao, G., Wang, Z., Ma, S., and Zhang, X. Complex backdoor detection by symmetric feature differencing. In *CVPR*, 2022.

Lu, W., Wang, J., Li, H., Chen, Y., and Xie, X. Domain-invariant feature exploration for domain generalization. *TMLR*, 2022.

Ma, H., Qiu, H., Gao, Y., Zhang, Z., Abuadbba, A., Fu, A., Al-Sarawi, S., and Abbott, D. Quantization backdoors to deep learning models. *arXiv preprint arXiv:2108.09187*, 2021.

Ma, H., Qiu, H., Gao, Y., Zhang, Z., Abuadbba, A., Xue, M., Fu, A., Zhang, J., Al-Sarawi, S. F., and Abbott, D. Quantization backdoors to deep learning commercial frameworks. *IEEE TDSC*, 2023.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

Nagel, M., Amjad, R. A., Van Baalen, M., Louizos, C., and Blankevoort, T. Up or down? adaptive rounding for post-training quantization. In *ICML*, 2020.

Nguyen, T. A. and Tran, A. Input-aware dynamic backdoor attack. In *NeurIPS*, 2020.

Nguyen, T. A. and Tran, A. T. Wanet-imperceptible warping-based backdoor attack. In *ICLR*, 2021.

Pan, X., Zhang, M., Yan, Y., and Yang, M. Understanding the threats of trojaned quantized neural network in model supply chains. In *ACSAC*, 2021.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

Sha, Z., He, X., Berrang, P., Humbert, M., and Zhang, Y. Fine-tuning is all you need to mitigate backdoor attacks. *arXiv preprint arXiv:2212.09067*, 2022.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Sokolić, J., Giryes, R., Sapiro, G., and Rodrigues, M. R. Robust large margin deep neural networks. *IEEE Transactions on Signal Processing*, 2017.

Srinivas, S. and Fleuret, F. Knowledge transfer with jacobian matching. In *ICML*, 2018.

Tian, Y., Suya, F., Xu, F., and Evans, D. Stealthy backdoors as compression artifacts. *IEEE TDSC*, 2022.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *JMLR*, 2008.

Walmer, M., Sikka, K., Sur, I., Shrivastava, A., and Jha, S. Dual-key multimodal backdoors for visual question answering. In *CVPR*, 2022.

Wang, B., Yao, Y., Shan, S., Li, H., Viswanath, B., Zheng, H., and Zhao, B. Y. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *IEEE S&P*, 2019.

Wang, C., Chen, D., Mei, J.-P., Zhang, Y., Feng, Y., and Chen, C. Semckd: Semantic calibration for cross-layer knowledge distillation. *TKDE*, 2022a.

Wang, K., Zhao, B., Peng, X., Zhu, Z., Yang, S., Wang, S., Huang, G., Bilen, H., Wang, X., and You, Y. Cafe: Learning to condense dataset by aligning features. In *CVPR*, 2022b.

Wang, P., Chen, Q., He, X., and Cheng, J. Towards accurate post-training network quantization via bit-split and stitching. In *ICML*, 2020.

Wang, Z., Mei, K., Ding, H., Zhai, J., and Ma, S. Rethinking the reverse-engineering of trojan triggers. In *NeurIPS*, 2022c.

Wang, Z., Zhai, J., and Ma, S. Bppattack: Stealthy and efficient trojan attacks against deep neural networks via image quantization and contrastive adversarial learning. In *CVPR*, 2022d.

Wang, Z., Mei, K., Zhai, J., and Ma, S. Unicorn: A unified backdoor trigger inversion framework. In *ICLR*, 2023.

Wu, B., Chen, H., Zhang, M., Zhu, Z., Wei, S., Yuan, D., and Shen, C. Backdoorbench: A comprehensive benchmark of backdoor learning. In *NeurIPS*, 2022.

Wu, Y., Wu, Y., Gong, R., Lv, Y., Chen, K., Liang, D., Hu, X., Liu, X., and Yan, J. Rotation consistent margin loss for efficient low-bit face recognition. In *CVPR*, 2020.

Xu, X., Wang, Q., Li, H., Borisov, N., Gunter, C. A., and Li, B. Detecting ai trojans using meta neural analysis. In *IEEE S&P*, 2021.

Xu, X., Huang, K., Li, Y., Qin, Z., and Ren, K. Towards reliable and efficient backdoor trigger inversion via decoupling benign features. In *ICLR*, 2024.

Yin, H., Molchanov, P., Alvarez, J. M., Li, Z., Mallya, A., Hoiem, D., Jha, N. K., and Kautz, J. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *CVPR*, 2020.

Zablocki, É., Ben-Younes, H., Pérez, P., and Cord, M. Explainability of deep vision-based autonomous driving systems: Review and challenges. *IJCV*, 2022.

Zeng, Y., Park, W., Mao, Z. M., and Jia, R. Rethinking the backdoor attacks' triggers: A frequency perspective. In *ICCV*, 2021.

Zeng, Y., Chen, S., Park, W., Mao, Z., Jin, M., and Jia, R. Adversarial unlearning of backdoors via implicit hypergradient. In *ICLR*, 2022.

Zhao, P., Chen, P.-Y., Das, P., Ramamurthy, K. N., and Lin, X. Bridging mode connectivity in loss landscapes and adversarial robustness. *arXiv preprint arXiv:2005.00060*, 2020.

Zheng, R., Tang, R., Li, J., and Liu, L. Pre-activation distributions expose backdoor neurons. *NeurIPS*, 2022.

Zhou, S. K., Greenspan, H., and Shen, D. *Deep learning for medical image analysis*. Academic Press, 2023.

Zhu, F., Gong, R., Yu, F., Liu, X., Wang, Y., Li, Z., Yang, X., and Yan, J. Towards unified int8 training for convolutional neural network. In *CVPR*, 2020.

Zhu, M., Wei, S., Shen, L., Fan, Y., and Wu, B. Enhancing fine-tuning based backdoor defense with sharpness-aware minimization. In *ICCV*, 2023.

A. Proof of Theorem 4.1.

Proof. We first derive from the left hand side. Assume $\boldsymbol{W}^l \in \mathbb{R}^{n\times d}$ and $\mathcal{I}^l \in \mathbb{R}^{d\times m}$. Inspired by Proposition 2 in (Srinivas & Fleuret, 2018), by using first-order Taylor expansion around $\mathcal{I}^l$, we have:

$$\begin{aligned}
&\arg\min_{\boldsymbol{W}^l} D(\boldsymbol{W}_o^l\mathcal{I}^l, Q(\boldsymbol{W}^l)(\mathcal{I}^l+\Delta\mathcal{I}^l)) \approx \|\boldsymbol{W}_o^l\mathcal{I}^l - (Q(\boldsymbol{W}^l)\mathcal{I}^l + \boldsymbol{J}_{Q(\boldsymbol{W}^l)}[\mathcal{I}^l]\Delta\mathcal{I}^l)\|_2^2 \\
&= \|\boldsymbol{W}_o^l\mathcal{I}^l - Q(\boldsymbol{W}^l)\mathcal{I}^l\|_2^2 - 2\cdot\sum_{i=1}^{n}\sum_{j=1}^{m}((\boldsymbol{W}_o^l\mathcal{I}^l - Q(\boldsymbol{W}^l)\mathcal{I}^l)_{i,j}((\boldsymbol{J}_{Q(\boldsymbol{W}^l)}[\mathcal{I}^l]_{i,j})^T\Delta\mathcal{I}^l)) + \|\boldsymbol{J}_{Q(\boldsymbol{W}^l)}[\mathcal{I}^l]\Delta\mathcal{I}^l\|_2^2 \qquad (6) \\
&= \|\boldsymbol{W}_o^l\mathcal{I}^l - Q(\boldsymbol{W}^l)\mathcal{I}^l\|_2^2 + \sigma^2\|\boldsymbol{J}_{Q(\boldsymbol{W}^l)}[\mathcal{I}^l]\|_2^2,
\end{aligned}$$

where $\boldsymbol{J}_{Q(\boldsymbol{W}^l)}[\mathcal{I}^l]$ denotes the Jacobian matrix of $Q(\boldsymbol{W}^l)$ with respect to $\mathcal{I}^l$. The final derivation is based on a further assumption that $\Delta\mathcal{I}^l \sim \mathcal{N}(0,\sigma^2\boldsymbol{I})$ (Srinivas & Fleuret, 2018). Since the quantization error is sufficiently small, we let $Q(\boldsymbol{W}) = \boldsymbol{W}_o + \Delta\boldsymbol{W}$. Then, by using first-order Taylor expansion again, we have:

$$\arg\min_{\boldsymbol{W}^l}\|\boldsymbol{W}_o^l\mathcal{I}^l - Q(\boldsymbol{W}^l)\mathcal{I}^l\|_2^2 + \sigma^2\|\boldsymbol{J}_{Q(\boldsymbol{W}^l)}[\mathcal{I}^l]\|_2^2 \approx \|\Delta\boldsymbol{W}^l\boldsymbol{J}_{\boldsymbol{W}^l\mathcal{I}^l}[\boldsymbol{W}^l]\|_2^2. \qquad (7)$$

From the above derivation, we can see that PDA essentially incorporates a Jacobian regularizer (Sokolić et al., 2017) into the loss function. This regularizer penalizes the norm of the Jacobian matrix, thus leads the quantized model toward better generalizability and more prediction agreement with the full-precision counterpart. We omit this regularizer and focus on $\|\Delta\boldsymbol{W}^l\boldsymbol{J}_{\boldsymbol{W}^l\mathcal{I}^l}[\boldsymbol{W}^l]\|_2^2$ in the following proof. We then take a look at the right hand side. Following previous work (Nagel et al., 2020; Li et al., 2021a; Hubara et al., 2020), by using second-order Taylor expansion around layer weights, we have:

$$\begin{aligned}
&\arg\min_{\boldsymbol{W}^l}\mathbb{E}\left[\mathcal{L}(f_Q(\boldsymbol{x}),y)\right] = \arg\min_{\boldsymbol{W}^l}\mathbb{E}\left[\mathcal{L}(f_{\boldsymbol{W}_o+\Delta\boldsymbol{W}}(\boldsymbol{x}),y)\right] \\
&= \arg\min_{\boldsymbol{W}^l}\mathbb{E}\left[\mathcal{L}(f_{\boldsymbol{W}_o}(\boldsymbol{x}),y) + \Delta\boldsymbol{W}\cdot\boldsymbol{g}^{\boldsymbol{W}} + \frac{1}{2}\Delta\boldsymbol{W}\cdot\boldsymbol{H}^{\boldsymbol{W}}\cdot\Delta\boldsymbol{W}^T\right],
\end{aligned} \qquad (8)$$

where $\boldsymbol{g}^{\boldsymbol{W}} = \nabla_{\boldsymbol{W}}\mathcal{L}$ and $\boldsymbol{H}^{\boldsymbol{W}} = \nabla^2_{\boldsymbol{W}}\mathcal{L}$ is the gradient and Hessian matrix $w.r.t.$ model weights, respectively. As $\mathcal{L}(f_{\boldsymbol{W}_o}(\boldsymbol{x}),y)$ is independent from the whole optimization and the model is converged to a local minimum, $\boldsymbol{g}^{\boldsymbol{W}}$ approximates 0 and therefore the first two terms of the optimization problem can be ignored (Dong et al., 2019; Li et al., 2021a; Dong et al., 2020; Nagel et al., 2020). Let $\mathcal{A}^l$ to denote the output (feature map) of the l-th layer $i.e.$, $\mathcal{A}^l = \boldsymbol{W}^l\mathcal{I}^l = \mathcal{I}^{l+1}$. According to Theorem 1 in (Li et al., 2021a), by using quadratic form, it holds that:

$$\begin{aligned}
&\arg\min_{\boldsymbol{W}^l}\Delta\boldsymbol{W}\cdot\boldsymbol{H}^{\boldsymbol{W}}\cdot\Delta\boldsymbol{W}^T = \sum_{i=1}^{n}\sum_{j=1}^{n}\Delta\boldsymbol{W}_i\Delta\boldsymbol{W}_j\left(\frac{\partial^2\mathcal{L}}{\partial\boldsymbol{W}_i\partial\boldsymbol{W}_j}\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}\Delta\boldsymbol{W}_i\Delta\boldsymbol{W}_j\left(\frac{\partial}{\partial\boldsymbol{W}_j}\left(\sum_{n=1}^{m}\frac{\partial\mathcal{L}}{\partial\mathcal{A}_n^l}\frac{\partial\mathcal{A}_n^l}{\partial\boldsymbol{W}_i}\right)\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\Delta\boldsymbol{W}_i\Delta\boldsymbol{W}_j\left(\sum_{n=1}^{m}\frac{\partial\mathcal{L}}{\partial\mathcal{A}_n^l}\frac{\partial^2\mathcal{A}_n^l}{\partial\boldsymbol{W}_i\partial\boldsymbol{W}_j} + \sum_{n,n'=1}^{m}\frac{\partial\mathcal{A}_n^l}{\partial\boldsymbol{W}_i}\frac{\partial^2\mathcal{L}}{\partial\mathcal{A}_n^l\partial\mathcal{A}_{n'}^l}\frac{\partial\mathcal{A}_{n'}^l}{\partial\boldsymbol{W}_j}\right) \\
&= \sum_{n=1}^{m}\sum_{n'=1}^{m}\left(\sum_{i=1}^{n}\Delta\boldsymbol{W}_i\frac{\partial\mathcal{A}_n^l}{\partial\boldsymbol{W}_i}\right)\left(\frac{\partial^2\mathcal{L}}{\partial\mathcal{A}_n^l\partial\mathcal{A}_{n'}^l}\right)\left(\sum_{i=1}^{n}\Delta\boldsymbol{W}_j\frac{\partial\mathcal{A}_{n'}^l}{\partial\boldsymbol{W}_j}\right) \\
&= (\Delta\boldsymbol{W}_i\boldsymbol{J}_{\mathcal{A}^l}[\boldsymbol{W}_i])\cdot\boldsymbol{H}^{\mathcal{A}^l}\cdot(\Delta\boldsymbol{W}_j\boldsymbol{J}_{\mathcal{A}^l}[\boldsymbol{W}_j])^T.
\end{aligned} \qquad (9)$$

The term $\sum_{n=1}^{m}\frac{\partial\mathcal{L}}{\partial\mathcal{A}_n^l}\frac{\partial^2\mathcal{A}_n^l}{\partial\boldsymbol{W}_i\partial\boldsymbol{W}_j}$ is neglected since the full-precision model is converged and thus satisfies $\nabla_{\mathcal{A}^l}\mathcal{L}\approx 0$ (Li et al., 2021a). Since layers are assumed to be mutual-independent, according to previous works (Botev et al., 2017; Nagel et al., 2020; Wang et al., 2020; Hubara et al., 2020), we can assume $\boldsymbol{H}^{\mathcal{A}^l}$ to be a constant block-diagonal matrix irrelevant to $\boldsymbol{W}^l$. As such, we can finally have:

$$\arg\min_{\boldsymbol{W}^l}\mathbb{E}\left[\mathcal{L}(f_Q(\boldsymbol{x}),y)\right] = (\Delta\boldsymbol{W}_i\boldsymbol{J}_{\mathcal{A}^l}[\boldsymbol{W}_i])\cdot\boldsymbol{H}^{\mathcal{A}^l}\cdot(\Delta\boldsymbol{W}_j\boldsymbol{J}_{\mathcal{A}^l}[\boldsymbol{W}_j])^T = \|\Delta\boldsymbol{W}^l\boldsymbol{J}_{\mathcal{A}^l}[\boldsymbol{W}^l]\|_2^2. \qquad (10)$$

Putting Eq. (7) and (10) together, we finish the proof. □

Notice that Theorem 4.1 does not ensure that our method performs similarly to fine-tuning, especially on backdoor accuracy. This is possibly because fine-tuning calculates CE loss on final logits. On benign samples, these logits are similar to the ground-truth labels because QCB-backdoored models already fit benign samples well. As such, vanilla fine-tuning can only make minor changes to the weights of neurons and thus is less effective in mitigating backdoor effects. In contrast, our LAC loss can be used to remove QCBs since it can directly reduce activation drifts on the corresponding layer.

B. More Implementation Details

More Details on Backdoor Attacks. In this paper, we defend against 3 state-of-the-art quantization-conditioned backdoor attacks, *i.e.* , CompArtifact (Tian et al., 2022), Qu-ANTI-zation (Hong et al., 2021), and PQBackdoor (Ma et al., 2021; 2023). Below is the introduction of each attack and their implementation details:

- **CompArtifact (Tian et al., 2022)**: In the study by CompArtifact (Tian et al., 2022), the methodology uses a trigger pattern akin to BadNets (Gu et al., 2017), specifically, a small 3×3 white square positioned at the image's bottom right. This technique demonstrates resilience when faced with alterations in the calibration set, but exhibiting limited transferability when applied across varying bandwidths. As such, we separately train models with each specific bandwidth to ensure a fair evaluation. We reproduce the results using the official source code provided by the authors[1]. Following their original paper, we first train a benign model for 400 epochs, and then re-train each model (respectively for 8-bit and 4-bit) with the backdoor objective for 50 epochs. The poison rate is set to 50% during re-training.

- **Qu-ANTI-zation (Hong et al., 2021)**: To enhance attack transferability, Qu-ANTI-zation (Hong et al., 2021) includes the multiple bit bandwidths during the re-training phase, exhibiting resilience across a variety of quantization bandwidths and against more advanced quantization methods. It employs a patch-based trigger approach, with dimensions designated as 4×4 for CIFAR10 and 8×8 for Tiny-ImageNet, respectively. We use the officially provided code[2]. Following their paper, we first train a benign model for 200 epochs, followed by a subsequent re-training period of 50 epochs. This re-training incorporates a modified objective and establishes a poisoning rate of 50%.

- **PQBackdoor (Ma et al., 2021; 2023)**: PQBackdoor is the latest and the state-of-the-art quantization-conditioned backdoor attack. It improves the training pipeline via introducing a two-stage attack strategy: firstly, train a backdoored full-precision model. Then, retrain the model via PGD (Madry et al., 2018) to make the full-precision model dormant while quantized model close to the backdoored one. This stabilizes the training of the quantization-conditioned backdoor and further enhances its resistance against backdoor defenses. It also utilizes the patch-based trigger, the size is set to 6×6. PQbackdoor also demonstrated its robustness against blind backdoor defenses such as fine-tuning, and its transferability to commercial quantization frameworks like PyTorch Mobile (Paszke et al., 2019) and TensorFlow Lite (Abadi et al., 2016). We use the official PyTorch source code from the authors[3] and follow their settings in the paper. For the first stage, the poisoning rate is set to 1%, with the standard training pipeline on poisoning-based backdoor attacks for 100 epochs. After the first stage, the poisoning rate is then set to 50% in the second stage, which takes another 50 epochs. Unfortunately, even if we tried several times (>5), we failed to obtain a full-precision model with BA reported in their paper. On CIFAR10, we can only have 86.43% with ResNet-18 during our reproduction, lower than 93.44% reported in their original paper. We experiment with our reproduced models.

More Details on Backdoor Defenses. In this paper, we choose 6 SOTA backdoor defenses as our baselines, including FT (Sha et al., 2022), FP (Liu et al., 2018), MCR (Zhao et al., 2020), NAD (Li et al., 2021c), I-BAU (Zeng et al., 2022), and FT-SAM (Zhu et al., 2023). For all defenses, we use the open-source code from BackdoorBox[4] (Li et al., 2023), except for I-BAU, which we use their official implementation[5]. Here are their brief introduction and implementation details:

- **FT (Sha et al., 2022):** Fine-tuning (FT) is commonly used approach for defending against backdoor attacks. It directly fine-tunes the model with a small subset of clean data. Despite its simplicity, it has proven effective in mitigating

[1] https://github.com/yulongt23/Stealthy-Backdoors-as-Compression-Artifacts

[2] https://github.com/Secure-AI-Systems-Group/Qu-ANTI-zation

[3] https://github.com/quantization-backdoor

[4] https://github.com/THUYimingLi/BackdoorBox

[5] https://github.com/YiZeng623/I-BAU

Table 8: Results on full-precision models (%).

Dataset	Attack	BA / ASR
CIFAR10	Clean Model	93.44% / 0.44%
	CompArtifact (8-bit)	91.46% / 1.26%
	CompArtifact (4-bit)	93.68% / 1.33%
	Qu-ANTI-zation	93.17% / 2.18%
	PQBackdoor	86.43% / 2.67%
Tiny-ImageNet	Clean Model	57.77% / 0.21%
	CompArtifact (8-bit)	57.09% / 0.78%
	CompArtifact (4-bit)	56.89% / 1.43%
	Qu-ANTI-zation	55.82% / 2.16%

backdoor effects for many state-of-the-art (SOTA) backdoor attacks, as demonstrated by (Wu et al., 2022). In our study, we fine-tune all layers of the compromised full-precision model using 5% clean data for 50 epochs.

- **FP (Liu et al., 2018):** Fine-pruning (FP) is a defense strategy that combines pruning and fine-tuning. Initially, it feeds a small set of clean data to the network to measure activation levels, then prunes the backdoored neurons, specifically the less frequently activated ones. To maintain the model's benign accuracy, FP fine-tunes the model post-pruning. In our work, we measure the activation in the last residual block and set the pruning rate to 0.4. We then fine-tune the model with 5% clean data for 50 epochs.

- **MCR (Zhao et al., 2020):** Mode connectivity repair (MCR) addresses DNN lifecycle security from the loss landscape perspective. It first fine-tunes a backdoored model, then uses mode connectivity in loss landscapes between the original backdoored model and the fine-tuned model, and ultimately measures and removes backdoor functions through mode connectivity repair. In our work, we fine-tune the backdoored model for 50 epochs, perform 100 epochs of curvenet training, and then carry out 100 epochs of model updating. The hyperparameter t is set to 0.1 and 0.9, and we report the results with the higher DTM.

- **NAD (Li et al., 2021c):** Neural attention distillation (NAD) is a defense mechanism that employs knowledge distillation guided by attention. It observes differences in attention between backdoored and clean models. Initially, it fine-tunes the backdoored model, which then serves as the teacher model. The original backdoored model becomes the student model, and knowledge distillation is conducted with attention alignment guidance. We perform 50 epochs of fine-tuning to obtain the teacher model and another 50 epochs to purify the student model.

- **I-BAU (Zeng et al., 2022):** Implicit backdoor adversarial unlearning (I-BAU) views backdoor removal as a minimax problem. It uses the implicit hypergradient to consider the interdependence between inner and outer optimization, demonstrating faster, more computationally efficient, and more effective performance than previous defenses, achieving state-of-the-art results on many benchmarks (Wu et al., 2022). We conduct 3 rounds of I-BAU for each attack.

- **FT-SAM (Zhu et al., 2023):** FT-SAM is a recent defense mechanism based on sharpness-aware minimization. Observing that backdoor-related neurons correlate strongly with the norm of weights, FT-SAM focuses on the sharpness of the loss landscape and aims to reduce the norms of these neurons. Following the original paper and the code from BackdoorBench [6], we train the infected model using FT-SAM for 200 epochs. The hyperparameter ρ is set to 2 for CIFAR-10 and 8 for Tiny-ImageNet.

For other backdoor attacks and defenses evaluated in Section 5.4, we use the official toolbox released by BackdoorBench (Wu et al., 2022). To avoid a lengthy introduction, we refer readers to their original papers for more details.

Implementation Details. For all experiments, we use Python 3.8.18 and the PyTorch 1.10.0+cu113 framework, along with torchvision 0.11.1. All experiments are implemented in Python and run on a 14-core Intel(R) Xeon(R) Gold 5117 CPU @ 2.00GHz with a single NVIDIA GeForce RTX 3090 GPU, on a machine running Linux version 5.4.0-144-generic (buildd@lcy02-amd64-089) (Ubuntu 9.4.0-1 ubuntu20.04.1). Unless otherwise stated, we use the Adam optimizer (Kingma & Ba, 2014) with default parameters. All other hyperparameters follow the original settings described in the respective papers. During clean model training and backdoor model training (first stage for PQBackdoor), the learning rate is set

[6] https://github.com/SCLBD/BackdoorBench/blob/main/defense/ft-sam.py

Table 9: Impact on clean models (%). The model is a clean model trained with QAT. Standard means standard quantization and ours means standard quantization after applying our method.

Dataset	Bandwidth	Setting	Accuracy ↑
CIFAR10	32-bit	Full-precision	93.59
	8-bit	Standard	93.58
		Ours	93.57
	4-bit	Standard	91.33
		Ours	92.71
Tiny-ImageNet	32-bit	Full-precision	57.35
	8-bit	Standard	56.58
		Ours	57.17
	4-bit	Standard	53.75
		Ours	55.69

to 10^{-3}, whereas it is set to 10^{-4} for all backdoor defenses and the second stage of PQBackdoor. The batch size is set to 64 for CIFAR10 and Tiny-ImageNet, and 16 for ImageNette. Each attack ultimately results in a full-precision model with a dormant backdoor inserted for each dataset and model architecture. As shown in Table 8, quantization-conditioned backdoors remain well-hidden in full-precision models, with a BA similar to a clean model, and an ASR of nearly 0%.

C. Impact on Clean Models

In our scenario, the full-precision model is assumed to be backdoored and given to the victim. However, in reality, the victim might obtain an actually clean (unbackdoored) model and employs our method to get rid of potential backdoor dangers. Therefore, if the model is clean, our method should have little affect on its clean accuracy.

To verify this, we train ResNet-18 models on CIFAR10 and Tiny-ImageNet, respectively, using typical QAT training objectives and other hyper-parameter configurations similar to the above. On these models, we then apply our method. As illustrated in Table 9, our method causes very little effect, suggesting our method's high effectiveness on keeping high BA.

D. Effectiveness on High-resolution Datasets

We assess the efficacy of our method on the high-resolution dataset ImageNette (Howard & fastai community, 2023). The results in Table 11 affirm the robust performance of our method in high-resolution cases.

E. Effectiveness across Different Triggers

As a novel backdoor paradigm, quantization-conditioned attacks can employ varying trigger sizes or integrate advanced trigger mechanisms. Presently, existing works on quantization-conditioned backdoors (Ma et al., 2021; 2023; Hong et al., 2021; Pan et al., 2021; Tian et al., 2022) predominantly adopt the trigger pattern from BadNets (Gu et al., 2017), with minor modifications in trigger sizes, and we have demonstrated our method's resilience against these variations. However, the diverse attack could be enhanced in terms of evasiveness when coupled with more sophisticated triggers (Nguyen & Tran, 2021; 2020). To comprehensively assess our method's effectiveness across diverse triggers, we enhance the trigger mechanism of PQBackdoor and evaluate our method's performance against these refined conditioned backdoor attacks. For simplicity, all experiments in this section utilize PQBackdoor (Ma et al., 2023; 2021) on the CIFAR10 dataset with ResNet-18 architecture.

E.1. Different Trigger Sizes

We evaluate our method's performance on different trigger sizes, including 3×3, 4×4, and 6×6. The results presented in Table 10 indicate the robustness of our method across different trigger sizes. The trigger pattern originates from BadNets (Gu et al., 2017).

Table 10: Our method's effectiveness on different trigger sizes (%).

Trigger Type	Bandwidth	Setting	BA / ASR
3×3	32-bit	Full-precision	91.88 / 1.46
	8-bit	No Defense	91.65 / 98.82
		Ours	91.58 / 0.56
	4-bit	No Defense	90.09 / 99.06
		Ours	91.05 / 0.53
4×4	32-bit	Full-precision	92.65 / 1.87
	8-bit	No Defense	91.99 / 97.78
		Ours	92.61 / 0.93
	4-bit	No Defense	90.88 / 97.72
		Ours	92.07 / 0.80
6×6	32-bit	Full-precision	92.10 / 1.61
	8-bit	No Defense	91.69 / 99.63
		Ours	91.97 / 1.11
	4-bit	No Defense	87.54 / 99.82
		Ours	91.44 / 0.52

Table 11: Effectiveness on high-resolution dataset ImageNette (%). The model is ResNet-18 and the attack is PQBackdoor.

Bandwidth	Setting	BA / ASR
32-bit	Full-precision	80.03 / 1.53
8-bit	No Defense	79.98 / 99.75
	Ours	79.34 / 0.62
4-bit	No Defense	76.84 / 98.73
	Ours	79.21 / 0.71

Table 12: Effectiveness on advanced triggers (%).

Trigger Type	Setting	BA / ASR
Input-aware Dynamic	Full-precision	93.16 / 0.36
	No Defense	93.89 / 99.73
	Ours	93.13 / 0.32
Warping-based	Full-precision	90.57 / 1.90
	No Defense	90.69 / 98.60
	Ours	90.16 / 0.78

E.2. Advanced Triggers

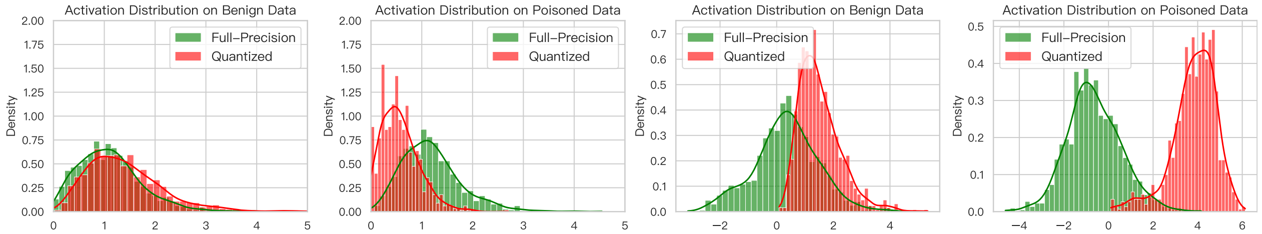
We evaluate our method's performance on advanced triggers, specifically examining two advanced trigger mechanisms: input-aware dynamic (Nguyen & Tran, 2020) and warping-based (Nguyen & Tran, 2021). The input-aware dynamic trigger crafts a unique dynamic trigger for each input rather than the fixed patch-based trigger, making the attack more evasive and hard to inspect, but the triggers are still visible; The warping-based triggers from WaNet (Nguyen & Tran, 2021) leverages a small and smooth warping field to poison the images. This type of invisible trigger is very stealthy and imperceptible and has been reported as far more evasive against human inspection. For simplicity and to stabilize training, in this section, we only consider 8-bit quantization during training and evaluation.

As shown in Table 12, our method still works well against these advanced triggers. The reason is that our method does not rely on any assumption about the trigger pattern. Although these triggers are more advanced than simple BadNets, the dormant backdoors are still activated by the nearest rounding errors, and thus can be handled well by our method.

F. Discussions

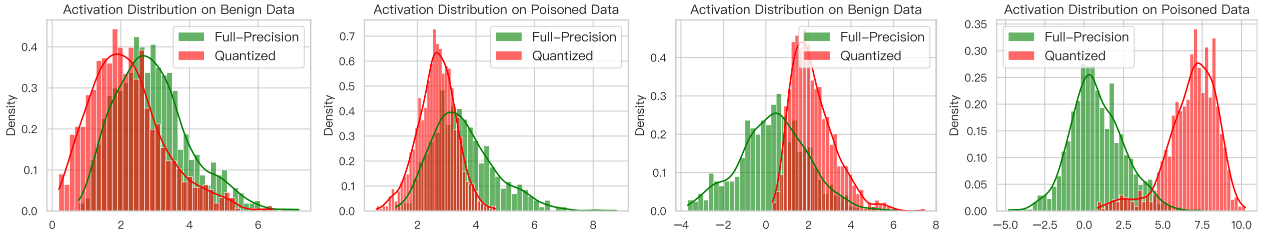
Comparison with (Zheng et al., 2022). At the first glance, our findings of activation drift may look similar to the findings of (Zheng et al., 2022). To enhance the clarity of our observation, we would like to emphasize that the primary observation of (Zheng et al., 2022) is that regarding benign samples and poisoned samples, backdoor neurons have different activation distributions. Although this conclusion also holds for QCBs, we mostly intend to point out that a QCB-backdoored neuron itself has a drift in activation distribution regarding quantization (instead of different samples). In particular, this activation drift exists even on benign samples (instead of only on poisoned samples). We believe this is an interesting, useful, yet unique property of QCBs, and it has not been discovered by previous works.

Discussion on Backdoor Neurons, Activation Drift and More Visualization Results. As we have discussed in Sec. 3, the definition of backdoor neurons in this Definition 3.1 is not perfect. For example, since it only considers the contribution to the backdoor loss, it may consider those neurons important to all tasks as backdoor neurons, and may filter neurons that has no contribution for all tasks as benign (*e.g.,* the "dead neurons" that is not responsive to any input). In our experiments in Sec. 3, we have filtered out neurons that always have a activation value of 0 to enhance the



(a) Benign Neuron (b) Backdoor Neuron

Figure 5: The activation distribution of backdoor and benign neurons (modified definition).



(a) Benign Neuron (b) Backdoor Neuron

Figure 6: The activation distribution of backdoor and benign neurons (original definition).

credibility of our study. Nevertheless, we can modify the definition of backdoor neurons to $\tau = \Delta\mathcal{L}_{backdoor} - \Delta\mathcal{L}_{benign} = \mathbb{E}_{x\sim\mathcal{D}'}[\mathcal{L}_{\text{ce}}(f_{-(k)}(x_t), y_t) - \mathcal{L}_{\text{ce}}(f(x_t), y_t)] - \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}_{\text{ce}}(f_{-(k)}(x), y) - \mathcal{L}_{\text{ce}}(f(x), y)]$, *i.e.* the decrease of backdoor loss minus the decrease of benign loss. As a result, we can get rid of the limitations above.

Figure 5 and 6 illustrates more visualization results on the activation drift phenomenon in backdoor neurons with both our original definition and the modified definition. Overall, we find that the neurons filtered via both strategies exhibits similar activation distributions. Despite this, we note that the definition of backdoor neurons is still not perfectly accurate. For example, it does not consider the joint effect of different neurons. We observe some neurons also exhibit different activation distributions after quantization. This is probably because the quantization itself may cause different levels of distortion to different neurons, so some actually benign neurons may also exhibit the distributional drift. Our preliminary experiments also found that the degree of activation drift (measured by the KL divergence) is not perfectly positively correlated with τ. In conclusion, our work still has room for improvement. We hope future works can further discover the mechanisms of QCBs.

On Technical Similarity of LAC to Previous Works. LAC is a very simple optimization objective. Notably, other layer-wise objectives similar to LAC have been widely used by previous works for different tasks, such as domain adaptation and generalization, network compression and acceleration, *etc.* (Frantar & Alistarh, 2022; Lu et al., 2022; Wang et al., 2022a;b). It has also been demonstrated effective in mitigating accuracy loss during quantization (Nagel et al., 2020; Li et al., 2021a; 2024), and even have been included by the concurrent work (Li et al., 2024) on QCB defense to compensate for quantization accuracy loss. However, we argue that our motivation is different from their works. LAC (and also our PDA) is driven by our novel observations on activation drift and insights into neuron self-rectification. The effectiveness of LAC in counteracting QCBs is also supported by our thorough analyses. These fundamental differences set our method apart from existing techniques, despite we finally converged to similar optimization objectives. The primary message of this paper is that LAC alone can effectively mitigate QCBs, due to its effectiveness in rectifying the aberrant activation in the quantized backdoor neurons. This, for the first time, also indicate that current advanced quantization techniques who have included LAC as their objectives are probably immune to existing QCBs. On the attacker side, we also argue that subsequent attacks should consider resistance to LAC, otherwise they may even be inadvertently mitigated by advanced quantization techniques.

On Extension of DER Metric. DER (Zhu et al., 2023) is a metric that balances the ASR-BA trade off. In this paper, we follow the original definition of DER for a simple and fair comparison between different defenses. If the ASR-BA trade-off needs to depend on specific applications, we can further introduce a weighting parameter $\alpha \in [0, 1]$ to control the tradeoff. The definition of DER is modified to DER $= [\alpha \cdot \max(0, \Delta\text{ASR}) - (1 - \alpha) \cdot \max(0, \Delta\text{BA}) + 1/2]$ accordingly.

Limitations and Future Work. Our work has several potential limitations. First, PDA cannot be applied to models without normalization layers or with normalization layers without running means/vars. We argue that it is not a big problem because almost all modern CNNs and many SOTA ViTs contain BatchNorm layers, especially those designed for high efficiency and low latency. One underlying reason is that BatchNorm layers are more latency favorable as they can be folded into the preceding layers for inference speedup after quantization, while dynamic normalizations (*e.g.,* LayerNorm) still collect running statistics at the inference phase, thus contributing to latency (Li et al., 2022b). Moreover, since LAC alone is effective in mitigating QCBs, users can directly apply our LAC without our PDA module for models without normalization layers or with normalization layers without running means/vars. Second, our method still requires a few unlabeled samples and some computational resources, although they are easily accessible. We will explore how to extend our method in the 'data-free' cases in our future works. Lastly, our defense requires the full precision version of the QCB-infected model, although it is naturally accessible under the settings of QCBs. We believe that when having access to quantized weights only, a proprietary defense/detection is necessary, although it is out of the scope of our current setting. We are also convinced that there is still a large (and challenging) space worth discovering for practical deployment-stage backdoor attacks/defenses. We hope our findings can inspire future work for further explorations.

Ethical Statements. The study of security vulnerabilities in deep learning models can raise ethical concerns (Wang et al., 2022d;c; Liu et al., 2022; Walmer et al., 2022; Wang et al., 2023). In this paper, we propose a novel defense against the recently introduced quantization-conditioned backdoor attacks. We are confident that our method will enhance the security of the model quantization process and support the responsible deployment of deep learning models. We are certain that our research adheres to all specified ethical standards. We ensure that our methodologies and experiments do not harm individuals or organizations and comply with all relevant ethical guidelines and regulatory standards. Our defense is solely intended to protect DNNs against malicious tampering and is not designed for any unethical applications.