Double-Win Quant: Aggressively Winning Robustness of Quantized Deep Neural Networks via Random Precision Training and Inference

Yonggan Fu [1] Qixuan Yu [1] Meng Li [2] Vikas Chandra [2] Yingyan Lin [1]

Abstract

Quantization is promising in enabling powerful yet complex deep neural networks (DNNs) to be deployed into resource constrained platforms. However, quantized DNNs are vulnerable to adversarial attacks unless being equipped with sophisticated techniques, leading to a dilemma of struggling between DNNs' efficiency and robustness. In this work, we demonstrate a new perspective regarding quantization's role in DNNs' robustness, advocating that quantization can be leveraged to largely boost DNNs' robustness, and propose a framework dubbed Double-Win Quant that can boost the robustness of quantized DNNs over their full precision counterparts by a large margin. Specifically, we for the first time identify that when an adversarially trained model is quantized to different precisions in a post-training manner, the associated adversarial attacks transfer poorly between different precisions. Leveraging this intriguing observation, we further develop Double-Win Quant integrating random precision inference and training to further reduce and utilize the poor adversarial transferability, enabling an aggressive "win-win" in terms of DNNs' robustness and efficiency. Extensive experiments and ablation studies consistently validate Double-Win Quant's effectiveness and advantages over state-of-the-art (SOTA) adversarial training methods across various attacks/models/datasets. Our codes are available at: https://github.com/RICE-EIC/Double-Win-Quant.

1. Introduction

Recent DNN breakthroughs and the advent of Internet of Things (IoT) devices have triggered an explosive demand for DNN-powered intelligent IoT devices (Liu et al., 2018; Wu et al., 2018). However, DNNs' deployments into IoT devices still remain challenging. First, powerful DNNs often come at a prohibitive cost, whereas IoT devices often suffer from stringent resource constraints. Second, while DNNs are vulnerable to adversarial attacks, many IoT applications require strict security. Therefore, techniques boosting both DNNs' efficiency and robustness are highly desired.

As quantization is one of the most promising techniques for developing efficient DNNs and generally applicable to a variety of algorithms, the robustness of quantized DNNs has gained increasing attentions. It was originally believed that quantization's rounding effect may help eliminate small adversarial perturbations, and early works (Galloway et al., 2017; Panda et al., 2019) indeed showed that binary networks (Galloway et al., 2017; Panda et al., 2019) or tanh-based quantized DNNs (Rakin et al., 2018) are even more robust than their full precision counterparts. Later, (Gupta & Ajanthan, 2020; Lin et al., 2019) found that these methods actually suffer from the obfuscated gradient problem (Athalye et al., 2018; Papernot et al., 2017), leading to a false sense of robustness. (Lin et al., 2019) further raised the community's awareness about quantized DNNs' inferior robustness, and identified that the main cause is the error amplification effect, i.e., the magnitude of adversarial perturbations is amplified when passing through DNN layers. Recently, pioneering works (Lin et al., 2019; Song et al., 2020; Shkolnik et al., 2020) tried to compress this amplification effect for achieving both robust and efficient DNNs.

In this work, we ask an intriguing question: "*Can quantization be properly leveraged to boost DNNs' robustness*?" This is inspired by the recent findings that (1) random smoothing or transformations on the inputs (Cohen et al., 2019; Li et al., 2018; Xie et al., 2017; Guo et al., 2017) can defend DNNs against adversarial attacks, and (2) weight perturbations are a good complement for input perturbations (Wu et al., 2020), because they can narrow the robust generalization gap as the weights can globally influence the losses of all examples. We conjecture that quantization noise can be leveraged to provide similar effects as perturbations to the weights and activations. Specifically, we make the following contributions:

- We provide a new perspective regarding the role of

[1] Department of Electrical and Computer Engineering, Rice University [2] Facebook Inc. Correspondence to: Yingyan Lin <yingyan.lin@rice.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

quantization in DNNs' robustness, and advocate that quantization, if properly exploited, can even enhance DNNs' robustness by a notable margin over their full-precision counterparts, instead of merely improving the robustness of quantized models.

- We are the first to identify that even if an adversarially trained model is directly quantized to a different precision in a post-training manner, the adversarial attacks still transfer poorly between different precisions, i.e., adversarial attacks generated with one precision usually achieve a lower success rate when attacking the same model quantized to other precisions.
- We propose a simple yet surprisingly effective framework dubbed Double-Win Quant, which integrates Random Precision Inference (RPI) and Random Precision Training (RPT) to achieve an aggressive "win-win" in terms of DNNs' robustness and efficiency. Specifically, RPI randomly selects an inference precision as a random perturbation at run-time for an adversarially trained model, while RPT adopts switchable batch normalization in training to further reduce the poor adversarial transferability and thus boost DNNs' achievable robustness.
- Extensive experiments and ablation studies show that our method is generally effective as evaluated across four commonly used adversarial training methods, four DNN models, and three datasets, e.g., achieving a 12.14% higher robust accuracy under PGD-20 attack with a 88.9% reduction in computational cost when using PGD-7 training for WideResNet32 on CIFAR-10. Furthermore, our method shows even larger improvements under more aggressive perturbations.

2. Related Works and Background

DNN quantization. Quantization has become one mainstream technique for developing efficient DNNs by representing weights/activations/gradients using lower floating-point precision (Wang et al., 2018; Sun et al., 2019) or fixed-point precision (Zhu et al., 2016; Li et al., 2016; Jacob et al., 2018; Mishra & Marr, 2017; Mishra et al., 2017; Park et al., 2017; Zhou et al., 2016). In particular, (Jacob et al., 2018) proposes quantization-aware training to learn the weight distribution for minimizing the accuracy degradation after quantization. Later, learnable quantizers (Jung et al., 2019; Bhalgat et al., 2020; Esser et al., 2019; Park & Yoo, 2020) featuring trainable quantization parameters further improve the accuracy of quantized DNNs under very low precision. In parallel, mixed-precision quantization methods (Wang et al., 2019; Xu et al., 2018; Elthakeb et al., 2020; Zhou et al., 2017) are proposed to assign different precisions to different layers. However, quantized DNNs have been found to be more vulnerable to adversarial attacks due to DNNs' error amplification effect (Lin et al., 2019). It is thus highly desirable to develop quantization techniques that can favor both DNNs' efficiency and robustness.

Adversarial attack & defense. DNNs are known to be vulnerable to adversarial attacks (Goodfellow et al., 2014), i.e., small perturbations on the inputs can mislead the models' decisions. To enhance DNNs' robustness, many defense methods (Guo et al., 2017; Buckman et al., 2018; Song et al., 2017; Xu et al., 2017; Liao et al., 2018; Metzen et al., 2017; Feinman et al., 2017; Li et al., 2018; Wu et al., 2020) have also been proposed, while many of them have been defeated later by stronger attacks. In particular, adversarial training (Shafahi et al., 2019; Madry et al., 2017; Wong et al., 2019; Tramèr et al., 2017) is currently the most effective defense method. Specifically, it augments the training set with adversarial samples generated by different attacks, thus enabling the models to correctly classify similar unseen adversarial samples. Examples of adversarial training methods can be found in Sec. 3.1. In this work, we rethink the role of quantization in DNNs' robustness and leverage it to enhance DNNs' robustness by a notable margin over their full-precision counterparts.

Robust and efficient DNNs. Efficiency and robustness are both critical for most DNN applications, and there have been pioneering works that aim to achieve both. For example, (Ye et al., 2019; Sehwag et al., 2020; Guo et al., 2018; Rakin et al., 2019) prune DNNs to derive sub-networks that can maintain or improve the robustness, and (Hu et al., 2020) balances both robustness and efficiency via input-adaptive inference. In parallel, as quantization is promising in enhancing the efficiency, other works strive to design robust quantized DNNs. Specifically, (Galloway et al., 2017; Panda et al., 2019) propose robust binary neural networks which, however, suffer from the obfuscated gradient problem (Athalye et al., 2018; Papernot et al., 2017). (Rakin et al., 2018) adopts tanh-based quantization, which also suffers from the obfuscated gradient problem as observed in (Lin et al., 2019). Later, (Lin et al., 2019) finds that quantized networks are more vulnerable to adversarial attacks due to the error amplification effect. To tackle this effect, (Lin et al., 2019; Shkolnik et al., 2020) add new regularization terms to model loss functions and (Song et al., 2020) retrains the network via feedback learning (Song et al., 2019). In addition, (Panda, 2020) searches for layerwise precision and (Gui et al., 2019) constructs a unified formulation to balance and enforce the robustness and compactness, respectively. However, existing works have not considered leveraging quantization to enhance robustness. Our Double-Win Quant for the first time makes use of quantization to boost DNNs' robustness, largely surpassing the full-precision counterparts.

Transferability of adversarial examples among quantized DNNs. The transferability of adversarial examples between quantized models with different precision has been studied by (Bernhard et al., 2019; Gupta & Ajanthan, 2020). In particular, (Gupta & Ajanthan, 2020) finds that quantized models are generally robust to adversarial samples generated by their full precision counterparts and (Bernhard et al., 2019) identifies the poor transferability of adversarial examples between full-precision and quantized models as well as between quantized models with different bitwidths. Built upon prior works, our work further makes new contributions in that (1) we find that even for the same adversarially pretrained model, if we directly quantize it to different precisions in a post-training manner, the adversarial examples still transfer poorly between models with different bitwidths; and (2) we propose two simple and effective techniques (i.e., RPI and RPT) based on the observation in (1) to practically win both DNNs' robustness and efficiency.

3. The Double-Win Quant Framework

In this section, we first introduce the preliminaries of adversarial training in Sec. 3.1, then present and analyze the motivating observations of our Double-Win Quant (DWQ) in Sec. 3.2, and finally describe DWQ's integrated techniques, i.e., RPI and RPT, in Sec. 3.3 and 3.4, respectively.

3.1. Preliminaries of Adversarial Training

DNNs are known to be vulnerable to adversarial attacks (Goodfellow et al., 2014), i.e., a small perturbation δ ($\|\delta\| \leq \epsilon$) applied to the inputs can mislead DNNs to make wrong predictions, where ϵ is a scalar that limits the perturbation's magnitude. To enhance DNNs' adversarial robustness, adversarial training is currently the strongest defense method (Athalye et al., 2018). For example, the adversarial perturbation δ under the ℓ_∞ attack (Goodfellow et al., 2014) is generated by maximizing the objective:

$$\max_{\|\delta\|_\infty \leq \epsilon} \ell(f_\theta(x+\delta), y) \tag{1}$$

where θ denotes the weights of a DNN f, x and y denote the input and the corresponding label, respectively, and ℓ is the loss function.

Adversarial training improves the model robustness by optimizing the following minimax problem:

$$\min_\theta \sum_i \max_{\|\delta\|_\infty \leq \epsilon} \ell(f_\theta(x_i+\delta), y_i) \tag{2}$$

Different adversarial training methods differ in how they solve Eq. 2's inner optimization. Specifically, the Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2014) uses the sign of one-step gradient as an approximation:

$$\delta = \epsilon \cdot sign(\nabla_x \ell(f_\theta(x+\delta), y)) \tag{3}$$

Projected Gradient Decent (PGD) (Madry et al., 2017) is a stronger variant of FGSM by iterating FGSM multiple times with a small step size α, where the t-th iteration can be formulated as:

$$\delta_{t+1} = clip_\epsilon\{\delta_t + \alpha \cdot sign(\nabla_{\delta_t} \ell(f_\theta(x+\delta_t), y))\} \tag{4}$$

FGSM-RS (Wong et al., 2019) introduces random initialization to FGSM for increasing the adversarial diversity:

$$\begin{aligned} &\delta = Uniform(-\epsilon, \epsilon) \\ &\delta = clip_\epsilon\{\delta + \alpha \cdot sign(\nabla_\delta \ell(f_\theta(x+\delta), y))\} \end{aligned} \tag{5}$$

where $clip_\epsilon$ denotes the clipping function that enforces its input to the interval $[-\epsilon, \epsilon]$.

Since the PGD attack is one of the most strongest white-box attacks, we adopt it as our mainly considered attack for evaluating DNNs' adversarial robustness in this work.

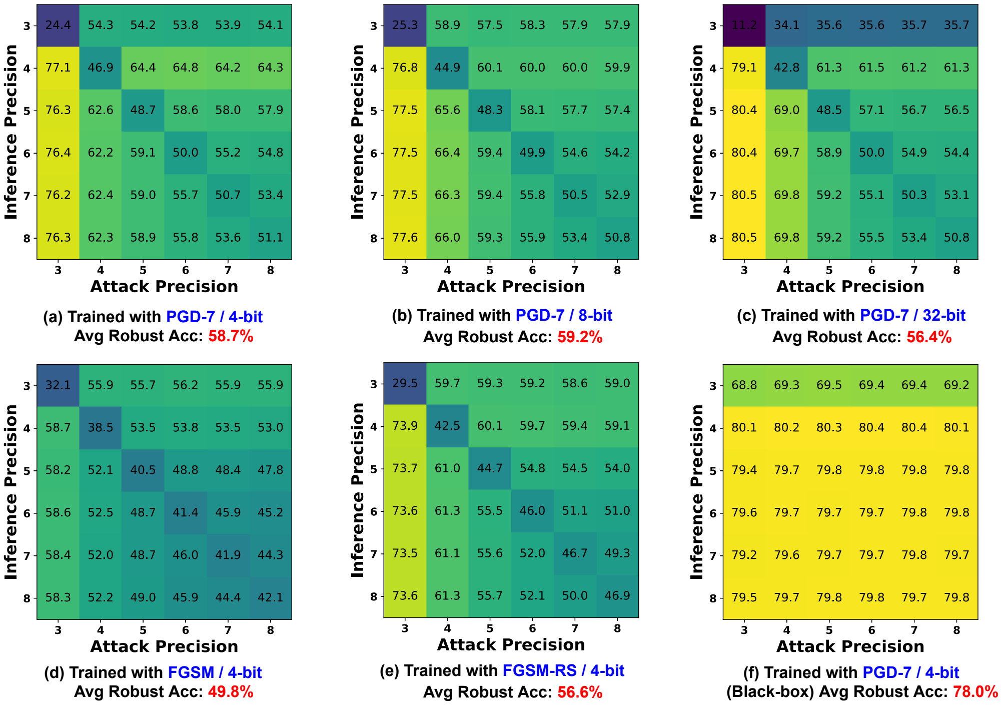
3.2. DWQ: Motivating Observations

The transferability of adversarial attacks between different compressed models has been studied (Matachana et al., 2020; Bernhard et al., 2019; Gupta & Ajanthan, 2020). However, it is still an open question about how to leverage such transferability to design robust DNNs against adversarial attacks. In this work, we ask the question: "*How well is the transferability of adversarial attacks between different precisions of an adversarially trained model*?", considering that the precision of a pretrained model can be instantaneously switched (Jin et al., 2020). We find that the adversarial attacks transfer poorly between different precisions of an adversarially trained model even if it is directly quantized to different precisions in a post-training manner, regardless of its adversarial training methods and training precisions.

Experiment settings. Here we conduct experiments to evaluate transferability of adversarial attacks between different precisions of the same adversarially trained model under different adversarial training methods and training precisions. We apply PGD-20 (20-step PGD (Madry et al., 2017)) attacks on PreActResNet18 (following (Wong et al., 2019)) which is adversarially trained using different adversarial training methods and training precisions using a linear quantizer (Jacob et al., 2018) with training settings introduced in Sec. 4.1. In Fig. 1, (a)~(e) directly quantize the model to different precisions (the same for weights and activations) in a post-training manner for both generating adversarial examples and inference, and (f) utilizes PGD-20 attacks generated by different quantized models trained on clean images, which is to emulate a black-box attack. Note that the goal of experiments in (f) is to check if there exists the obfuscated gradient problem (Athalye et al., 2018), instead of exploring the transferability as in (a)~(e).

Experiment observations. Four observations can be made:

- (1) Adversarial attacks generated with one precision achieve a lower success rate when attacking the same



(a) Trained with PGD-7 / 4-bit
Avg Robust Acc: 58.7%

(b) Trained with PGD-7 / 8-bit
Avg Robust Acc: 59.2%

(c) Trained with PGD-7 / 32-bit
Avg Robust Acc: 56.4%

(d) Trained with FGSM / 4-bit
Avg Robust Acc: 49.8%

(e) Trained with FGSM-RS / 4-bit
Avg Robust Acc: 56.6%

(f) Trained with PGD-7 / 4-bit
(Black-box) Avg Robust Acc: 78.0%

Figure 1. Visualizing the **transferability of adversarial attacks between different precisions of the same adversarially trained model**, where the robust accuracy under PGD-20 attack is annotated for **three adversarial training methods, three training precisions, and six inference/attacking precisions**. Specifically, experiments (a)~(c) adopt PGD-7 training with different training precisions; experiments (d) and (e) adopt another two adversarial training methods, i.e., FGSM (Goodfellow et al., 2014) and FGSM-RS (Wong et al., 2019), respectively; and (f) adopts PGD-20 attacks generated by different quantized models trained on clean images, and uses the corresponding precisions to attack the adversarially trained model using PGD-7/4-bit training, which aims to emulate a black-box attack.

adversarially trained model quantized to a different precision especially under the commonly adopted low precisions, which is consistent across different adversarial training methods and training precisions of the quantized model;

- (2) The average robust accuracies under white-box attacks (see Fig. 1(a)~(e)) are consistently higher than the full-precision models trained with the corresponding adversarial training methods, indicating that randomly selecting an inference precision can potentially provide effective defense. The *full-precision* accuracies of PreActResNet18 trained with PGD-7/FGSM/FGSM-RS are 51.2%/41.5%/47.1%, respectively;
- (3) The poor transferability does not come from the obfuscated gradient problem as the model shows a better robustness under black-box attacks than that of white-box attacks, according to (Athalye et al., 2018);
- (4) Training and attacking at the same lower precisions indeed notably degrades the robust accuracy, as shown in the diagonals of Fig. 1(a)~(e), aligning with the observations in (Lin et al., 2019) due to the error amplification effects.

Analysis and discussion. The key conclusion is that for white-box attacks, adversarial attacks generated at one precision transfer poorly to another precision. We hypothesize that this poor transferability is because adversarial perturbations are shielded by the quantization noise between the two precisions, which can not be effectively learned by gradient-based attacks. Specifically, considering a linear quantizer, the k-bit quantized value A_q for an activation A (the same for weights) can be formulated as $A_q = S_k \lfloor \frac{A}{S_k} \rceil$, where $\lfloor \cdot \rceil$ is the rounding operation and $S_k = \frac{A_{max} - A_{min}}{2^k - 1}$ is the scale factor. For vanilla quantization, the rounding effect can be effectively learned by gradient-based attacks through straight-through estimation (Bengio et al., 2013; Yin et al., 2019), i.e., $\frac{\partial L}{\partial A} \approx \frac{\partial L}{\partial A_q}$, where L is the loss function. However, the quantization noise $S_m \lfloor \frac{A}{S_m} \rceil - S_n \lfloor \frac{A}{S_n} \rceil$ between two different precisions m-bit and n-bit cannot be effectively learned by gradient-based attacks, thus adversarial perturbations can be buried within the quantization noise, leading to attack failures.

3.3. Vanilla DWQ: Random Precision Inference

Here we introduce the vanilla DWQ which integrates RPI, and is simple and generally applicable to different DNNs.

Methodology. Given an adversarially trained model, RPI randomly selects one precision from an inference precision set to quantize the model's weights and activations during inference. The effectiveness of RPI is rooted in two facts:

Algorithm 1 Vanilla DWQ: The RPI Algorithm

```
Require: model $f_\theta$, inference precision set $Set_Q$, adversarial
    dataset $D_{adv}$ generated on $f_\theta$ by adversaries
for $x_{adv} \in D_{adv}$ do
    Randomly select a precision $q$ from $Set_Q$
    Obtain $f_\theta^q$ by quantizing $f_\theta$ to $q$-bit
    Evaluate $\hat{y} = f_\theta^q(x_{adv})$
return $\{\hat{y}\}$
```

(1) a quantization-aware-trained model has relatively stable natural accuracies (on clean images) during inference when being directly quantized to different precisions (Jin et al., 2020; Guerra et al., 2020; Fu et al., 2021b), thus models resulting from RPI can maintain a natural accuracy that is comparable with their static precision counterparts; and (2) randomly selecting an inference precision can greatly degrade the effectiveness of adversarial attacks as long as the attacks are not generated under the same precision, as consistently observed in Fig. 1 under different adversarial training methods and training precisions. Note that although adversaries may select precisions with better attacking success rates and RPI can adopt sampling strategies to favor the probability of choosing precisions that is in general more robust. Without loss of generality, we consider that both the adversaries and RPI adopt random precision from the same inference precision set in this work. The RPI algorithm is summarized in Alg. 1.

Implementation. As the execution of RPI needs to switch among different precisions during inference, we slightly modify the quantization scheme inspired by (Jin et al., 2020) to ensure the ease of implementation. Specifically, we quantize the weights θ to $\theta_q = \hat{S_k}\, min(\lfloor \frac{\theta}{\hat{S_k}} \rceil, 2^k - 1)$, where $\hat{S_k} = \frac{\theta_{max}-\theta_{min}}{2^k}$. As a result, the switch between different precisions only requires clipping the most significant bits (MSBs) via shifting as $\lfloor \frac{\theta}{\hat{S_m}} \rceil >> (m-n)$ is equal to $\lfloor \frac{\theta}{\hat{S_n}} \rceil$, where $>>$ is the right shift operation. Therefore, only one copy of quantized models with the highest precision needs to be stored.

Connections with prior works. The spirit behind RPI aligns with the recent findings in DNN perturbations and robustness. In particular, it has been shown that random smoothing or transformations on the inputs (Cohen et al., 2019; Li et al., 2018; Xie et al., 2017; Guo et al., 2017) can help robustify DNNs against adversarial attacks, and (Wu et al., 2020) finds that weight perturbations can serve as a good complement for input perturbations to narrow the robust generalization gap because weights of DNNs can globally influence the losses of all input examples. (He et al., 2019; Dhillon et al., 2018) also explicitly introduce randomness and perturbations in the models' weights or activations. Drawing inspirations from these findings, we conjecture that the effectiveness of RPI lies in the fact that quantization noise due to random switches between different precisions naturally injects random perturbations to both the weights and activations, which can compensate for the influence of adversarial features and thus enhance the models' adversarial robustness.

Algorithm 2 Enhanced DWQ: RPT with PGD-7 training

```
Require: Training dataset $D_{train}$, model $f_\theta$, training/inference
    precision set $Set_Q$, total training epochs $T$, step size $\alpha$
Equip $f_\theta$ with SBN
for $epoch \in [1, T]$ do
    for $(x, y) \in D_{train}$ do
        Randomly select a precision $q$ from $Set_Q$
        Obtain $f_\theta^q$ by quantizing $f_\theta$ to $q$-bit
        $\delta = 0$ or random initialized
        for $t \in [1, 7]$ do
            $\delta = clip_\epsilon\{\delta + \alpha \cdot sign(\nabla_\delta \ell(f_\theta^q(x+\delta), y))\}$
        end for
        $\theta = \theta - \nabla_\theta \ell(f_\theta^q(x+\delta), y)$
    end for
end for
```

Hardware support for RPI. SOTA adaptive-precision accelerators like Bit Fusion (Sharma et al., 2018) and Stripes (Judd et al., 2016) are dedicated to support dynamic precision inference, which can naturally support the execution of RPI. More potential hardware implementations for RPI can be found in (Camus et al., 2019).

3.4. Enhanced DWQ: Random Precision Training

As described in the above subsection, our vanilla DWQ simply manipulates the inference precision of adversarially trained models to boost their robustness. In this subsection, we introduce the enhanced DWQ, which can further enhance the adversarial robustness of DNNs by aggravating the poor transferability between different precisions via RPT equipped with switchable batch normalization (SBN).

Motivations. (Xie et al., 2020) adopts dual BNs for the clean and adversarial examples to boost the natural accuracy, motivating the necessity of separately handling the statistics of clean and adversarial inputs. In addition, (Jin et al., 2020; Guerra et al., 2020) propose to use separate BNs for different precisions to enable instantaneous quantization of a trained DNN to different bits which maintain the natural accuracy as the same DNN separately trained using the corresponding precision. These works motivate and inspire us to come up with our enhanced DWQ, as applying independent BNs for different precisions to record the specific statistics of the adversarial examples generated under each precision can potentially enlarge the gap between different precisions, i.e., further increase the difficulty of transferring adversarial examples between different precisions.

Methodology. Our enhanced DWQ adversarially trains a model from scratch via randomly selecting a precision from a candidate set in each iteration for generating adversarial examples and updating the model with the selected precision, while equipping the model with SBN to independently record the statistics of different precisions. Although there exist other training schemes, e.g., progressive precision (Fu et al., 2020) or dynamic precision (Fu et al., 2021a), to be considered, we find that RPT is sufficiently effective in largely boosting the robustness of quantized models without increasing the training complexity as validated in Sec. 4.4. Additionally, we visualize the adversarial transferability achieved by the enhanced DWQ in the Appendix, from which we can observe a notably reduced transferability over the vanilla one.

The RPT algorithm on top of PGD-7 (Madry et al., 2017) training is summarized in Alg. 2, of which the algorithms for other adversarial training methods (e.g., FGSM (Goodfellow et al., 2014) and FGSM-RS (Wong et al., 2019)) are similar. Note that one advantage of RPI and RPT is that they are simple and consistently work well across different DNN models, precision sets, and adversarial training methods, without the necessity of cherry-picking the hyperparameters as validated in Sec. 4.

4. Experiment Results

In this section, we first introduce the experiment setup in Sec. 4.1 and then benchmark our DWQ with SOTA adversarial training methods in Sec. 4.2. We next conduct comprehensive ablation studies for DWQ's integrated RPI and RPT techniques in Sec. 4.3 and 4.4, respectively.

4.1. Experiment Setup

Networks & datasets. We evaluate our DWQ on four networks and three datasets, i.e., PreActResNet18 (following (Wong et al., 2019)), WideResNet32 (following (Madry et al., 2017; Shafahi et al., 2019)), and MobileNetV2 on CIFAR-10/100, and ResNet-50 (following (Shafahi et al., 2019; Wong et al., 2019)) on ImageNet.

Training settings. We consider four SOTA adversarial training methods, including FGSM (Goodfellow et al., 2014), FGSM-RS (Wong et al., 2019), PGD-7 (Madry et al., 2017), and Free (Shafahi et al., 2019). We follow their original papers for the adversarial training hyper-parameter settings, i.e., we adopt a step size of 1.25ϵ for FGSM-RS and 2 for PGD-7 training. We follow the model training settings as (Madry et al., 2017) without resorting to other training tricks for fairness. On CIFAR-10/100, we train the model for 160 epochs with a batch size of 128 and an SGD optimizer with a momentum of 0.9, starting from an initial learning rate of 0.1 decayed by 10 at both the 80-th and 120-th epochs. On ImageNet, we follow SOTA quantization works on ImageNet (Jung et al., 2019; Bhalgat et al., 2020; Esser et al., 2019; Park & Yoo, 2020) to start from a full-precision pretrained model on clean images. In particular, we adversarially train a full-precision pretrained ResNet-50 for 60 epochs with a batch size of 256 and an SGD optimizer with a momentum of 0.9, starting from an initial learning rate of 0.01 decayed by 10 at the 30-th epoch.

Attack settings. We mainly consider PGD attacks (Madry et al., 2017) with different numbers of iterations/restarts and perturbation strengths, and also evaluate DWQ's general robustness under CW-L2/CW-Inf attacks (Carlini & Wagner, 2017), Auto-Attack (Croce & Hein, 2020), and a gradient-free attack Bandits (Ilyas et al., 2018). In particular, for the CW-L2/CW-Inf attacks we adopt the implementation in AdverTorch (Ding et al., 2019) and follow the settings in (Chen et al., 2021; Rony et al., 2019); for the Auto-Attack (Croce & Hein, 2020) and Bandits (Ilyas et al., 2018), we adopt the official implementation and default settings in their original papers. We assume adversaries adopt random precision from the same inference precision set as our DWQ without losing generality since (1) any attack precision out of DWQ's inference precision set will merely increase DWQ's robust accuracy according to experiments in Sec. 3.2, and (2) while adversaries may select precisions with better attacking success rates, our DWQ can also increase the probability of sampling more robust precisions for stronger defense, here we assume both adversaries and DWQ adopt random precision for simplicity.

Precision settings of RPI and RPT. Two precision settings are involved in DWQ: (1) the training precision set of RPT and (2) the inference precision set of RPI. Without cherry-picking the precision settings, we determine the precisions for efficiency-robustness considerations. In particular, if not specifically stated, we use 4∼16-bit for training PreActResNet18 and WideResNet32, and 4∼8-bit for training MobileNetV2, when adopting RPT, and employ the same precision set for the corresponding inference. When only RPI is enabled (i.e., DWQ without RPT), we adversarially train the network with fixed-point 4-bit for PreActResNet18 and WideResNet32 and 8-bit for MobileNetV2 and use an inference precision set of 4∼8-bit. We validate DWQ's consistent benefits with different choices of inference and training precision sets in Sec. 4.3 and 4.4. In addition, we use BitOPs (Bit OPerations) to measure the computational cost as SOTA quantized DNN works.

4.2. DWQ: Benchmark with SOTA Methods

Benchmark with SOTA adversarial training methods. Here we benchmark our RPI and RPT techniques with SOTA adversarial training methods trained **with full precision** to validate their superior "win-win" in terms of boosting both model robustness and efficiency.

Results on CIFAR-10: As summarized in Tab. 1, we can see that **(1)** both RPI and RPT can consistently enhance the robust accuracy under PGD attacks of different settings, under all the networks and adversarial training methods; **(2)** DWQ (i.e., RPI + RPT) always achieves the best robust-

Table 1. Evaluating RPI and RPT on top of two networks (PreActResNet18 and WideResNet32) and three adversarial training methods (FGSM, FGSM-RS, and PGD-7) on CIFAR-10 in terms of both natural accuracy and robust accuracy under different PGD attacks.

	PreActResNet18					**WideResNet32**				
Adversarial Training Method	**Natural Acc (%)**	**PGD-20 (%)**	**PGD-100 (%)**	**PGD-20 10 Restart (%)**	**BitOPs**	**Natural Acc (%)**	**PGD-20 (%)**	**PGD-100 (%)**	**PGD-20 10 Restart (%)**	**BitOPs**
FGSM	67.04	41.48	41.37	41.55	569.9 G	66.76	40.78	40.55	40.74	6824.6 G
FGSM + **RPI**	71.44	47.46	46.43	44.50	21.2 G	68.65	46.25	45.33	43.35	253.3 G
FGSM + **RPI** + **RPT**	80.58	**64.08**	**63.56**	**60.28**	63.5 G	64.09	**50.70**	**48.72**	**48.68**	759.8 G
FGSM-RS	86.08	41.76	41.13	41.67	569.9 G	89.95	45.33	44.77	45.12	6824.6 G
FGSM-RS + **RPI**	82.79	52.98	52.07	49.87	21.2 G	89.17	51.49	49.80	47.17	253.3 G
FGSM-RS + **RPI** + **RPT**	82.11	**59.33**	**59.32**	**55.52**	63.5 G	87.87	**60.07**	**58.12**	**56.98**	759.8 G
PGD-7	82.02	51.17	50.93	51.30	569.9 G	85.25	54.61	54.36	54.68	6824.6 G
PGD-7 + **RPI**	80.17	57.09	56.06	53.83	21.2 G	84.39	59.83	58.17	56.21	253.3 G
PGD-7 + **RPI** + **RPT**	82.16	**65.15**	**64.88**	**61.82**	63.5 G	81.52	**66.75**	**66.28**	**64.17**	759.8 G

Table 2. Evaluating RPI and RPT on PreActResNet18 and WideResNet32 trained with FGSM-RS and PGD-7 on CIFAR-100.

	PreActResNet18					**WideResNet32**				
Adversarial Training Method	**Natural Acc (%)**	**PGD-20 (%)**	**PGD-100 (%)**	**PGD-20 10 Restart (%)**	**BitOPs**	**Natural Acc (%)**	**PGD-20 (%)**	**PGD-100 (%)**	**PGD-20 10 Restart (%)**	**BitOPs**
FGSM-RS	57.60	26.14	25.88	26.36	569.9 G	67.29	25.35	24.78	25.25	6824.7 G
FGSM-RS + **RPI**	57.91	31.71	30.83	28.57	21.2 G	66.48	32.31	30.58	28.33	253.3 G
FGSM-RS + **RPI** + **RPT**	51.09	**36.75**	**37.18**	**34.19**	63.5 G	64.95	**39.18**	**38.36**	**36.44**	759.8 G
PGD-7	56.31	27.97	27.77	28.09	569.9 G	60.36	31.06	30.86	31.24	6824.7 G
PGD-7 + **RPI**	56.68	33.36	32.04	29.90	21.2 G	59.78	35.57	34.97	32.83	253.3 G
PGD-7 + **RPI** + **RPT**	56.20	**41.74**	**42.10**	**39.19**	63.5 G	58.41	**40.45**	**40.50**	**39.51**	759.8 G

ness with a comparable natural accuracy, largely outperforming SOTA adversarial training methods with full precision. In particular, our DWQ achieves a 13.98%/12.14% higher robust accuracy under PGD-20 attacks while reducing the computational cost by 88.9%, on PreActResNet18 and WideResNet32, respectively, when being applied on top of PGD-7 training, which is one of the strongest adversarial training methods; and **(3)** DWQ also enhances the robust accuracy by 13.57%~22.60% under PGD-20 attacks on top of FGSM/FGSM-RS. It is noteworthy that although FGSM adversarial training can be easily ineffective against iteration-based attacks (Kurakin et al., 2016), our DWQ can still significantly improve its robust accuracy by 22.6%.

Results on CIFAR-100: As shown in Tab. 2, the observations on CIFAR-100 are consistent with those on CIFAR-10, indicating our DWQ's scalability to more complex tasks. In particular, DWQ integrating with RPI and RPT achieves 10.61%/13.77% and 13.83%/9.39% higher robust accuracy on top of FGSM-RS/PGD-7 training under PGD-20 attacks on PreActResNet18 and WideResNet32, respectively.

Table 3. Evaluating the enhanced DWQ over two SOTA adversarial training methods (FGSM-RS (Wong et al., 2019) and Free (Shafahi et al., 2019)) on top of ResNet-50 on ImageNet under PGD-10 and PGD-50 attacks with $\epsilon = 4$, where all the baseline results are the reported ones in the original papers.

Adversarial Training Method	**Natural Acc (%)**	**PGD-10 (%)**	**PGD-50 (%)**	**BitOPs**
FGSM-RS	55.45	30.28	30.18	3891.2 G
FGSM-RS + **RPI** + **RPT**	**63.21**	**37.93**	**37.12**	**433.2 G**
Improvement	**+7.76**	**+7.65**	**+6.94**	**-88.9%**
Free	60.21	32.77	31.88	3891.2 G
Free + **RPI** + **RPT**	**64.58**	**42.88**	**42.72**	**433.2 G**
Improvement	**+4.37**	**+10.11**	**+10.84**	**-88.9%**

Results on ImageNet: As shown in Tab. 3, we can observe that our enhanced DWQ achieves a **triple-win** in terms of the natural accuracy, robust accuracy, and model efficiency on top of both adversarial training methods. In particular, the enhanced DWQ achieves a 7.65%/10.11% higher accuracy over FGSM-RS (Wong et al., 2019) and Free (Shafahi et al., 2019), respectively, under the PGD-10 attack, while offering a 88.9% reduction in the computational cost. This set of experiments further indicates our DWQ framework's scalability and applicability on large-scale datasets.

Table 4. Evaluating RPI and RPT on top of MobileNetV2 trained with FGSM-RS and PGD-7 on CIFAR-10.

Adversarial Training Method	**Natural Acc (%)**	**PGD-20 (%)**	**PGD-100 (%)**	**PGD-20 10 Restart (%)**	**BitOPs**
FGSM-RS	82.16	39.47	38.98	39.59	96.9 G
FGSM-RS + **RPI**	81.00	48.08	45.13	42.10	3.6 G
FGSM-RS + **RPI** + **RPT**	80.27	**56.52**	**54.33**	**54.12**	3.6 G
PGD-7	80.40	50.03	49.72	50.06	96.9 G
PGD-7 + **RPI**	76.83	55.11	54.02	55.19	3.6 G
PGD-7 + **RPI** + **RPT**	73.97	**57.72**	**55.83**	**56.78**	3.6 G

Benchmark on compact DNNs. As discussed in (Madry et al., 2017) that models with a higher capacity are more robust against multi-step attacks, defending compact models like MobileNetV2 (Sandler et al., 2018) is more challenging and desirable. Tab. 4 shows that our DWQ can still boost the robust accuracy by 17.05%/7.69% on top of FGSM-RS/PGD-7 training on MobileNetV2, indicating DWQ's applicability to compact DNNs. We also observe that DWQ on top of FGSM-RS achieves better trade-offs between robust and natural accuracy than DWQ on top of PGD-7, which we conjecture is because PGD-7 aggressively pursues robustness without considering MobileNetV2's vulnerability to quantization (Sheng et al., 2018) on clean images.

Benchmark under larger perturbations. We further evaluate DWQ's scalability under larger perturbations with PGD-7 training on CIFAR-10 as listed in Tab. 5. Interestingly, DWQ even achieves larger robustness improvements. Tab. 5 shows that DWQ leads to a 11.66%~18.89% and 15.68%~23.26% higher robust accuracy under PGD-20 attacks with $\epsilon = 12$ and 16, respectively. Larger improve-

Table 5. Evaluating RPI and RPT under larger perturbations on three networks with PGD-7 training on CIFAR-10.

Network	Adversarial Training Method	ϵ=12				ϵ=16			
		Natural Acc (%)	PGD-20 (%)	PGD-100 (%)	PGD-20 10 restart (%)	Natural Acc (%)	PGD-20 (%)	PGD-100 (%)	PGD-20 10 restart (%)
PreActResNet18	PGD-7	77.49	37.84	36.77	37.90	75.39	27.28	24.24	26.98
	PGD-7 + **RPI**	75.38	45.30	43.56	40.61	72.67	38.14	35.15	32.54
	PGD-7 + **RPI** + **RPT**	77.45	**56.73**	**56.62**	**51.46**	75.02	**50.54**	**50.16**	**45.57**
WideResNet32	PGD-7	81.80	39.73	38.49	39.67	78.91	28.92	25.82	28.80
	PGD-7 + **RPI**	79.97	48.34	46.61	43.60	77.40	38.72	35.70	32.86
	PGD-7 + **RPI** + **RPT**	78.26	**53.74**	**52.42**	**52.10**	75.34	**46.82**	**44.85**	**43.49**
MobileNetV2	PGD-7	75.31	36.90	35.95	36.96	72.86	27.65	24.95	27.51
	PGD-7 + **RPI**	72.10	43.02	40.10	36.97	71.74	38.85	34.56	31.35
	PGD-7 + **RPI** + **RPT**	68.34	**48.56**	**46.52**	**47.02**	66.23	**43.33**	**40.43**	**41.89**

ments under stronger adversarial attacks validate DWQ's applicability to more challenging environments.

Table 6. Evaluating the enhanced DWQ over the vanilla PGD-7 training on two networks and two datasets under Auto-Attack (Croce & Hein, 2020), CW-L2/CW-Inf attacks (Carlini & Wagner, 2017), and Bandits attacks (Ilyas et al., 2018). In particular, we adopt different initial τ defined in (Carlini & Wagner, 2017) for CW-Inf attacks to control the final perturbation ϵ.

Network	PreActResNet18				WideResNet32			
Dataset	CIFAR-10		CIFAR-100		CIFAR-10		CIFAR-100	
Training Method	PGD-7 (%)	DWQ (%)	PGD-7 (%)	DWQ (%)	PGD-7 (%)	DWQ (%)	PGD-7 (%)	DWQ (%)
Auto-Attack (ϵ=8)	47.18	**54.56**	24.46	**32.51**	51.66	**58.54**	27.18	**36.26**
Auto-Attack (ϵ=12)	27.59	**35.83**	12.77	**20.93**	30.71	**39.83**	15.24	**23.26**
CW-L2	76.58	**80.95**	48.80	**54.86**	78.19	**80.40**	50.79	**58.16**
CW-Inf (τ=0.05 → ϵ=8)	57.88	**71.44**	31.99	**46.22**	62.13	**72.10**	35.39	**50.28**
CW-Inf (τ=0.1 → ϵ=12)	46.70	**65.57**	24.29	**42.49**	50.14	**66.99**	26.94	**46.58**
CW-Inf (τ=0.2 → ϵ=16)	33.56	**59.27**	16.52	**38.87**	36.11	**60.81**	18.71	**43.14**
Bandits (ϵ=8)	59.75	**71.75**	34.02	**43.50**	63.49	**68.50**	38.03	**47.35**
Bandits (ϵ=12)	46.04	**70.52**	24.46	**42.43**	49.77	**67.01**	28.13	**46.18**

Benchmark under more attacks. We evaluate the enhanced DWQ on top of PGD-7 training against Auto-Attack (Croce & Hein, 2020), CW-L2/CW-Inf attacks (Carlini & Wagner, 2017), and Bandits attacks (Ilyas et al., 2018). As shown in Tab. 6, the enhanced DWQ consistently improves the robust accuracy across different attacks/models/datasets/perturbations, e.g., a higher robust accuracy of +6.88%~+9.12% under Auto-Attack, +5.01%~+24.48% under Bandits attacks, and more surprisingly, +9.97%~+25.71% under CW-Inf attack, where we find that the poor transferability between different attack/inference precisions is more notable and thus DWQ is still very effective, while PGD-7 trained networks suffer from more robustness drops under larger perturbations. This set of experiments verifies the consistent robustness achieved by DWQ under different attack types.

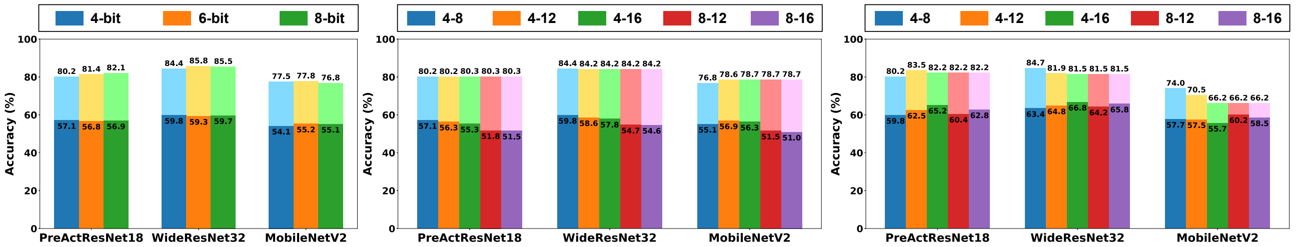
Benchmark with SOTA robust quantization methods. DWQ's most relevant work is (Lin et al., 2019) which constrains the layerwise Lipschitz constants and is orthogonal with DWQ. While (Lin et al., 2019) compresses the negative effect of quantization on adversarial robustness, DWQ makes use of quantization noise to boost the adversarial robustness, thus combining the two methods can potentially lead to more robust DNNs. Compared with the best reported robust accuracy among all the settings from (Lin et al., 2019) under PGD-20 attack on CIFAR-10, our DWQ achieves a 14.6% and 22.5% higher robust accuracy for ϵ = 8 and 16, respectively, on the same PGD-7 trained network.

Obfuscated gradient check. We also evaluate DWQ with all other flags of obfuscated gradients in (Athalye et al., 2018), in addition to black-box attacks in Fig. 1 (f), on top of PreActResNet18 and find that (1) robust accuracies under 1-step/20-step PGD attacks are 78.37%/65.15% on CIFAR-10 and 52.25%/41.74% on CIFAR-100; (2) for unbounded PGD-20 attacks, DWQ has a near-zero robust accuracy; (3) no adversarial examples are found in 10^5 random sampling for DWQ when PGD-20 does not; and (4) increasing the distortions causes a drop in robustness as shown in Tab. 5. Therefore, DWQ does not suffer from obfuscated gradients.

4.3. DWQ: Ablation Study of Vanilla DWQ

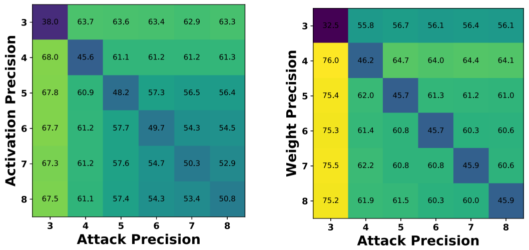
RPI trained with different precisions. Fig. 2(a) shows RPI's achieved natural and robust accuracy on CIFAR-10 under PGD-20 attacks on three PGD-7 trained networks with different precisions, adopting an inference precision set of 4~8-bit. Both the natural and robust accuracy remain stable (within 1.12%) under different training precisions, indicating RPI's general applicability to quantized models.

RPI with different inference precision sets. Fig. 2(b) shows RPI's achieved natural and robust accuracy on CIFAR-10 under PGD-20 attacks of three PGD-7/4-bit trained networks, considering different inference precision sets. We can observe that (1) RPI with different inference precision sets consistently improves the robustness of vanilla PGD-7 training in Tab. 1, and (2) without the help of RPT, vanilla DWQ (i.e., merely RPI) slightly favors smaller precision ranges and precisions close to the training precision, as both leads to smaller distribution gaps between training and inference precisions. Therefore, we adopt 4~8-bit for RPI in Sec. 4.2 considering both robustness and efficiency.



(a) PGD-7 + RPI (b) PGD-7 + RPI (c) PGD-7 + RPI + RPT

Figure 2. The natural and robust accuracy of three PGD-7 trained networks with (a) RPI under different training precisions with an inference precision set of 4~8-bit, (b) RPI with different inference precision sets on top of the 4-bit trained networks, and (c) RPI+RPT with different training precision sets. Note that deep and light colors denote robust and natural accuracy, respectively.



(a) Fixed 4-bit Weight
Avg Robust Acc: 58.2%

(b) Fixed 4-bit Activation
Avg Robust Acc: 59.8%

Figure 3. The transferability of adversarial attacks between different weight (activation) precisions with the fixed 4-bit activation (weight) precision on top of a PGD-7 trained PreActResNet18.

RPI applied to only the weights/activations. In all other experiments, we adopt the same precision for both the weights and activations for hardware-friendly implementation. Here we further explore whether the motivating observations and benefits of DWQ consistently maintain when applying RPI to merely the weights or activations. Experiments in Fig. 3 show that the poor transferability of adversarial attacks is consistently observed when fixing the precision for the weights or activations, indicating that RPI applied to merely the weights or activations would be consistently effective in enhancing the model robustness.

4.4. DWQ: Ablation Study of Enhanced DWQ

Enhanced DWQ with different precision ranges. Fig. 2(c) shows the natural and robust accuracy (under PGD-20 attacks) of three networks trained using enhanced DWQ (i.e., RPI + RPT) with different training precision sets, on top of PGD-7 training on CIFAR-10. We can see that (1) DWQ with both RPT and RPI consistently achieves a higher robust accuracy over SOTA PGD-7 training or DWQ with only RPI in Tab. 1, although the robust accuracy shows some fluctuations under different settings; and (2) models with a higher capacity like PreActResNet18 and WideResNet32 favor larger precision ranges for reducing the probability of hitting the adversaries' precision, while models with a low capacity like MobileNetV2 favor smaller precision ranges and relatively higher precisions due to their vulnerability to quantization (Sheng et al., 2018).

Table 7. Comparing enhanced DWQ with RPT and CPT on PreActResNet18 trained using three adversarial training methods on CIFAR-10, where CPT-16 denotes CPT with 16 cyclic periods.

	FGSM		FGSM-RS		PGD	
Training Recipe	**Natural Acc (%)**	**PGD-20 Acc (%)**	**Natural Acc (%)**	**PGD-20 Acc (%)**	**Natural Acc (%)**	**PGD-20 Acc (%)**
CPT-16	53.02	35.45	71.19	56.28	78.65	65.15
CPT-32	35.17	49.98	74.82	59.34	81.51	**66.40**
CPT-64	29.39	36.79	76.67	**59.95**	74.54	61.34
RPT	**80.58**	**64.08**	**82.11**	59.33	**82.15**	65.15

Enhanced DWQ with different training recipes. We also equip the enhanced DWQ with another dynamic precision training method CPT (Fu et al., 2021a) which cyclically switches between the lowest and the highest precisions and compare the enhanced DWQ with RPT or CPT under different cyclic periods. Tab. 7 shows that (1) CPT on top of the strongest PGD-7 training achieves comparable robustness as RPT, while its cyclic periods need to be finetuned, and (2) CPT leads to large training instability and a lower natural and robust accuracy on top of less powerful adversarial training methods, which we conjecture is due to the mismatches between the statistics of SBN and current weights as CPT switches merely between consecutive precisions. In contrast, DWQ integrated RPT shows consistent stability and effectiveness.

5. Conclusion

In this work, we have demonstrated that quantization, if properly exploited, can even enhance quantized DNNs' robustness by a notable margin over their full-precision counterparts, instead of merely improving the robustness of quantized models. Furthermore, we propose a simple yet effective framework dubbed Double-Win Quant, which achieves an aggressive "win-win" in terms of DNNs' robustness and efficiency. We believe Double-Win Quant has opened up a new perspective in designing robust and efficient DNNs.

Acknowledgements

The work is supported by the National Science Foundation (NSF) through the MLWiNS program (Award number: 2003137) and the NeTS program (Award number: 1801865). Thanks Mr. Tianlong Chen from UT Austin for the valuable discussions in this work.

References

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, pp. 274–283. PMLR, 2018.

Bengio, Y., Léonard, N., and Courville, A. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Bernhard, R., Moellic, P.-A., and Dutertre, J.-M. Impact of low-bitwidth quantization on the adversarial robustness for embedded neural networks. In *2019 International Conference on Cyberworlds (CW)*, pp. 308–315. IEEE, 2019.

Bhalgat, Y., Lee, J., Nagel, M., Blankevoort, T., and Kwak, N. Lsq+: Improving low-bit quantization through learnable offsets and better initialization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 696–697, 2020.

Buckman, J., Roy, A., Raffel, C., and Goodfellow, I. Thermometer encoding: One hot way to resist adversarial examples. In *International Conference on Learning Representations*, 2018.

Camus, V., Mei, L., Enz, C., and Verhelst, M. Review and benchmarking of precision-scalable multiply-accumulate unit architectures for embedded neural-network processing. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 9(4):697–711, 2019.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. IEEE, 2017.

Chen, T., Zhang, Z., Liu, S., Chang, S., and Wang, Z. Robust overfitting may be mitigated by properly learned smoothening. In *International Conference on Learning Representations*, volume 1, 2021.

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pp. 1310–1320. PMLR, 2019.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, pp. 2206–2216. PMLR, 2020.

Dhillon, G. S., Azizzadenesheli, K., Lipton, Z. C., Bernstein, J., Kossaifi, J., Khanna, A., and Anandkumar, A. Stochastic activation pruning for robust adversarial defense. *arXiv preprint arXiv:1803.01442*, 2018.

Ding, G. W., Wang, L., and Jin, X. Advertorch v0. 1: An adversarial robustness toolbox based on pytorch. *arXiv preprint arXiv:1902.07623*, 2019.

Elthakeb, A. T., Pilligundla, P., Mireshghallah, F., Yazdanbakhsh, A., and Esmaeilzadeh, H. Releq: A reinforcement learning approach for automatic deep quantization of neural networks. *IEEE Micro*, 2020.

Esser, S. K., McKinstry, J. L., Bablani, D., Appuswamy, R., and Modha, D. S. Learned step size quantization. *arXiv preprint arXiv:1902.08153*, 2019.

Feinman, R., Curtin, R. R., Shintre, S., and Gardner, A. B. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017.

Fu, Y., You, H., Zhao, Y., Wang, Y., Li, C., Gopalakrishnan, K., Wang, Z., and Lin, Y. FracTrain: Fractionally squeezing bit savings both temporally and spatially for efficient dnn training. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12127–12139. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/8dc5983b8c4ef1d8fcd5f325f9a65511-Paper.pdf.

Fu, Y., Guo, H., Li, M., Yang, X., Ding, Y., Chandra, V., and Lin, Y. CPT: Efficient deep neural network training via cyclic precision. *arXiv preprint arXiv:2101.09868*, 2021a.

Fu, Y., Yu, Z., Zhang, Y., Jiang, Y., Li, C., Liang, Y., Jiang, M., Wang, Z., and Lin, Y. Instantnet: Automated generation and deployment of instantaneously switchable-precision networks. *arXiv preprint arXiv:2104.10853*, 2021b.

Galloway, A., Taylor, G. W., and Moussa, M. Attacking binarized neural networks. *arXiv preprint arXiv:1711.00449*, 2017.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Guerra, L., Zhuang, B., Reid, I., and Drummond, T. Switchable precision neural networks. *arXiv preprint arXiv:2002.02815*, 2020.

Gui, S., Wang, H., Yu, C., Yang, H., Wang, Z., and Liu, J. Model compression with adversarial robustness: A unified optimization framework. *arXiv preprint arXiv:1902.03538*, 2019.

Guo, C., Rana, M., Cisse, M., and Van Der Maaten, L. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017.

Guo, Y., Zhang, C., Zhang, C., and Chen, Y. Sparse dnns with improved adversarial robustness. *arXiv preprint arXiv:1810.09619*, 2018.

Gupta, K. and Ajanthan, T. Improved gradient based adversarial attacks for quantized networks. *arXiv preprint arXiv:2003.13511*, 2020.

He, Z., Rakin, A. S., and Fan, D. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 588–597, 2019.

Hu, T.-K., Chen, T., Wang, H., and Wang, Z. Triple wins: Boosting accuracy, robustness and efficiency together by enabling input-adaptive inference. *arXiv preprint arXiv:2002.10025*, 2020.

Ilyas, A., Engstrom, L., and Madry, A. Prior convictions: Black-box adversarial attacks with bandits and priors. *arXiv preprint arXiv:1807.07978*, 2018.

Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., and Kalenichenko, D. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2704–2713, 2018.

Jin, Q., Yang, L., and Liao, Z. Adabits: Neural network quantization with adaptive bit-widths. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2146–2156, 2020.

Judd, P., Albericio, J., Hetherington, T., Aamodt, T. M., and Moshovos, A. Stripes: Bit-serial deep neural network computing. In *2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pp. 1–12. IEEE, 2016.

Jung, S., Son, C., Lee, S., Son, J., Han, J.-J., Kwak, Y., Hwang, S. J., and Choi, C. Learning to quantize deep networks by optimizing quantization intervals with task loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4350–4359, 2019.

Kurakin, A., Goodfellow, I., Bengio, S., et al. Adversarial examples in the physical world, 2016.

Li, B., Chen, C., Wang, W., and Carin, L. Certified adversarial robustness with additive noise. *arXiv preprint arXiv:1809.03113*, 2018.

Li, F., Zhang, B., and Liu, B. Ternary weight networks. *arXiv preprint arXiv:1605.04711*, 2016.

Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., and Zhu, J. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1778–1787, 2018.

Lin, J., Gan, C., and Han, S. Defensive quantization: When efficiency meets robustness. *arXiv preprint arXiv:1904.08444*, 2019.

Liu, S., Lin, Y., Zhou, Z., Nan, K., Liu, H., and Du, J. On-demand deep model compression for mobile devices: A usage-driven model selection framework. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '18, pp. 389–400, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450357203. doi: 10.1145/3210240.3210337. URL https://doi.org/10.1145/3210240.3210337.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Matachana, A. G., Co, K. T., Muñoz-González, L., Martinez, D., and Lupu, E. C. Robustness and transferability of universal attacks on compressed models. *arXiv preprint arXiv:2012.06024*, 2020.

Metzen, J. H., Genewein, T., Fischer, V., and Bischoff, B. On detecting adversarial perturbations. *arXiv preprint arXiv:1702.04267*, 2017.

Mishra, A. and Marr, D. Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy. *arXiv preprint arXiv:1711.05852*, 2017.

Mishra, A., Nurvitadhi, E., Cook, J. J., and Marr, D. Wrpn: wide reduced-precision networks. *arXiv preprint arXiv:1709.01134*, 2017.

Panda, P. Quanos: adversarial noise sensitivity driven hybrid quantization of neural networks. In *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design*, pp. 187–192, 2020.

Panda, P., Chakraborty, I., and Roy, K. Discretization based solutions for secure machine learning against adversarial attacks. *IEEE Access*, 7:70157–70168, 2019.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pp. 506–519, 2017.

Park, E. and Yoo, S. Profit: A novel training method for sub-4-bit mobilenet models. *arXiv preprint arXiv:2008.04693*, 2020.

Park, E., Ahn, J., and Yoo, S. Weighted-entropy-based quantization for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5456–5464, 2017.

Rakin, A. S., Yi, J., Gong, B., and Fan, D. Defend deep neural networks against adversarial examples via fixed and dynamic quantized activation functions. *arXiv preprint arXiv:1807.06714*, 2018.

Rakin, A. S., He, Z., Yang, L., Wang, Y., Wang, L., and Fan, D. Robust sparse regularization: Simultaneously optimizing neural network robustness and compactness. *arXiv preprint arXiv:1905.13074*, 2019.

Rony, J., Hafemann, L. G., Oliveira, L. S., Ayed, I. B., Sabourin, R., and Granger, E. Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4322–4330, 2019.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Sehwag, V., Wang, S., Mittal, P., and Jana, S. Hydra: Pruning adversarially robust neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 7, 2020.

Shafahi, A., Najibi, M., Ghiasi, A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., and Goldstein, T. Adversarial training for free! *arXiv preprint arXiv:1904.12843*, 2019.

Sharma, H., Park, J., Suda, N., Lai, L., Chau, B., Chandra, V., and Esmaeilzadeh, H. Bit fusion: Bit-level dynamically composable architecture for accelerating deep neural network. In *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*, pp. 764–775. IEEE, 2018.

Sheng, T., Feng, C., Zhuo, S., Zhang, X., Shen, L., and Aleksic, M. A quantization-friendly separable convolution for mobilenets. In *2018 1st Workshop on Energy Efficient Machine Learning and Cognitive Computing for Embedded Applications (EMC2)*, pp. 14–18. IEEE, 2018.

Shkolnik, M., Chmiel, B., Banner, R., Shomron, G., Nahshan, Y., Bronstein, A., and Weiser, U. Robust quantization: One model to rule them all. *arXiv preprint arXiv:2002.07686*, 2020.

Song, C., Wang, Z., and Li, H. Feedback learning for improving the robustness of neural networks. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 686–693. IEEE, 2019.

Song, C., Fallon, E., and Li, H. Improving adversarial robustness in weight-quantized neural networks. *arXiv preprint arXiv:2012.14965*, 2020.

Song, Y., Kim, T., Nowozin, S., Ermon, S., and Kushman, N. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *arXiv preprint arXiv:1710.10766*, 2017.

Sun, X., Choi, J., Chen, C.-Y., Wang, N., Venkataramani, S., Srinivasan, V. V., Cui, X., Zhang, W., and Gopalakrishnan, K. Hybrid 8-bit floating point (hfp8) training and inference for deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 4901–4910, 2019.

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

Wang, K., Liu, Z., Lin, Y., Lin, J., and Han, S. Haq: Hardware-aware automated quantization with mixed precision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8612–8620, 2019.

Wang, N., Choi, J., Brand, D., Chen, C.-Y., and Gopalakrishnan, K. Training deep neural networks with 8-bit floating point numbers. In *Advances in neural information processing systems*, pp. 7675–7684, 2018.

Wong, E., Rice, L., and Kolter, J. Z. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2019.

Wu, D., Xia, S.-T., and Wang, Y. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33, 2020.

Wu, J., Wang, Y., Wu, Z., Wang, Z., Veeraraghavan, A., and Lin, Y. Deep k-means: Re-training and parameter sharing with harder cluster assignments for compressing deep convolutions. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5363–5372. PMLR, 10–15 Jul 2018. URL `http://proceedings.mlr.press/v80/wu18h.html`.

Xie, C., Wang, J., Zhang, Z., Ren, Z., and Yuille, A. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*, 2017.

Xie, C., Tan, M., Gong, B., Wang, J., Yuille, A. L., and Le, Q. V. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 819–828, 2020.

Xu, W., Evans, D., and Qi, Y. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

Xu, Y., Zhang, S., Qi, Y., Guo, J., Lin, W., and Xiong, H. Dnq: Dynamic network quantization. *arXiv preprint arXiv:1812.02375*, 2018.

Ye, S., Xu, K., Liu, S., Cheng, H., Lambrechts, J.-H., Zhang, H., Zhou, A., Ma, K., Wang, Y., and Lin, X. Adversarial robustness vs. model compression, or both? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 111–120, 2019.

Yin, P., Lyu, J., Zhang, S., Osher, S., Qi, Y., and Xin, J. Understanding straight-through estimator in training activation quantized neural nets. *arXiv preprint arXiv:1903.05662*, 2019.

Zhou, S., Wu, Y., Ni, Z., Zhou, X., Wen, H., and Zou, Y. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *arXiv preprint arXiv:1606.06160*, 2016.

Zhou, Y., Moosavi-Dezfooli, S.-M., Cheung, N.-M., and Frossard, P. Adaptive quantization for deep neural network. *arXiv preprint arXiv:1712.01048*, 2017.

Zhu, C., Han, S., Mao, H., and Dally, W. J. Trained ternary quantization. *arXiv preprint arXiv:1612.01064*, 2016.