# INVESTIGATING THE IMPACT OF QUANTIZATION ON ADVERSARIAL ROBUSTNESS

**Qun Li, Yuan Meng**[†], **Chen Tang, Jiacheng Jiang, Zhi Wang**[†]
Tsinghua University
liq23@mails.tsinghua.edu.cn, yuanmeng@mail.tsinghua.edu.cn
genprtung@gmail.com, jiangjc23@mails.tsinghua.edu.cn
wangzhi@sz.tsinghua.edu.cn

## ABSTRACT

Quantization is a promising technique for reducing the bit-width of deep models to improve their runtime performance and storage efficiency, and thus becomes a fundamental step for deployment. In real-world scenarios, quantized models are often faced with adversarial attacks which cause the model to make incorrect inferences by introducing slight perturbations. However, recent studies have paid less attention to the impact of quantization on the model robustness. More surprisingly, existing studies on this topic even present inconsistent conclusions, which prompted our in-depth investigation. In this paper, we conduct a first-time analysis of the impact of the quantization pipeline components that can incorporate robust optimization under the settings of Post-Training Quantization and Quantization-Aware Training. Through our detailed analysis, we discovered that this inconsistency arises from the use of different pipelines in different studies, specifically regarding whether robust optimization is performed and at which quantization stage it occurs. Our research findings contribute insights into deploying more secure and robust quantized networks, assisting practitioners in reference for scenarios with high-security requirements and limited resources.

## 1 INTRODUCTION

Deep neural networks have demonstrated outstanding performance in tasks such as computer vision (Ronneberger et al., 2015; Howard et al., 2017) and natural language processing (Sutskever et al., 2014; Vaswani et al., 2017). However, the increasing number of model parameters makes it challenging to deploy them in resource-constrained scenarios. As a model compression technique, quantization effectively reduces the inference memory and time cost by quantizing model parameters and/or activations from full-precision to low-bit integers (Nagel et al., 2021). Current research aims to achieve better performance with lower bit-width (Choi et al., 2018; Esser et al., 2019). Many quantization methods (Yao et al., 2022) have been incorporated into the deep learning libraris (Rasley et al., 2020). However, many resource-constrained scenarios require not only efficiency but also high security for model inference, such as in autonomous driving (Liu et al., 2021; Katare et al., 2023) and medical assistance (Zhang & Chung, 2021). These applications often face the risk of adversarial attacks, where imperceptible perturbations are added to input samples, leading the model to make incorrect predictions (Madry et al., 2017). This poses a fatal risk in certain scenarios. Due to adversarial robustness being at odds with accuracy (Hu et al., 2023), it indicates that considering both quantization and adversarial robustness simultaneously is necessary.

Existing research has focused on improving the adversarial robustness of quantized networks. However, we find that the conclusions drawn from these studies are conflicting. For instance, some studies observe that as the quantization bit-width decreases, adversarial robustness deteriorates (Lin et al., 2021; Sen et al., 2020; Xiao et al., 2023), while others reach the opposite conclusion (Fu et al., 2021; Lin et al., 2019). We find the conflicting

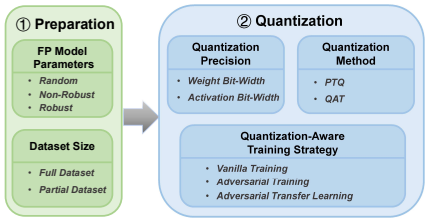


Figure 1: Quantization Pipeline.

[†]Corresponding authors.

Table 1: The impact of different quantization pipelines on robustness. (**+**) means reducing bit-width enhances robustness, while (**−**) indicates the opposite, and (**∼**) means moderate bit-width gains the best robustness. $\epsilon$ means the adversarial attack intensity (Madry et al., 2017), and we define $\epsilon \leq 2$ as the weaker attack, and $\epsilon \leq 8$ as the normal attack. Additional details regarding attack intensity and quantization settings will be supplemented in Appendix A.

| **Quantization Setting** | **Random Parameters** | **Non-Robust Parameters** | **Robust Parameters** |
|---|---|---|---|
| PTQ | —— | Against weaker attack, lower bit-width leads to higher robustness (**+**). | Against normal attack, lower bit-width leads to lower robustness (**−**). |
| QAT (Vanilla Training) | Against weaker attack, lower bit-width leads to higher robustness (**+**). | Against weaker attack, lower bit-width leads to higher robustness (**+**). | Against weaker attack, lower bit-width leads to higher robustness (**+**). |
| QAT (Adversarial Training) | Against normal attack, lower bit-width leads to lower robustness (**−**). | Accuracy significantly decreases compared to initialization, against normal attack, lower bit-width leads to higher robustness (**+**). | Against normal attack, moderate bit-width exhibits the best robustness (**∼**). |
| QAT (Adversarial Transfer Learning) | Against normal attack, but less robustness than robust initialized ways, lower bit-width leads to lower robustness (**−**). | Against normal attack, but less robustness than robust initialized ways, moderate bit-width exhibits the best robustness (**∼**). | Against normal attack, lower bit-width leads to lower robustness (**−**). |

results arise from differences in the pipeline settings during the quantization process and the possible addition of robustness components at different stages.

To address this question, we first clarify the pipeline of the quantization process, as shown in Figure 1. This involves the **preparation stage** and the **quantization stage**. In the preparation stage, it is necessary to specify whether a pre-trained full-precision model (FP model) is employed and whether it undergoes adversarial training, which lead to three options: *robust parameters, non-robust parameters, and random parameters*. Additionally, determining the size of the dataset for calibration or fine-tuning is essential. In the quantization stage, the choice between Quantization-Aware Training (QAT) and Post-Training Quantization (PTQ) needs to be made. It is also needed to determine whether weights and activations need quantization, along with specifying the quantization bit-width. For QAT, whether adversarial training is incorporated into the process is also needed to ascertain. Please refer to Appendix A for the details of Quantization-Aware Training Strategy. A variety of choices make it difficult for researchers to determine the pipeline, and it is uncertain about the impact of each choice on robustness.

We experiment with different pipelines to observe their robustness. The key conclusions include: (i) Quantization without any robustness components demonstrates resistance to lower levels of adversarial attacks, and a lower bit-width for quantization leads to improved robustness. (ii) In certain pipelines, the quantized model can achieve robustness similar to that of the full precision robust model. In such cases, a lower quantization bit-width may result in decreased robustness. (iii) Adding adversarial training to quantization can gain robustness, however, using the wildly used PGD-7 adversarial training incurs an additional $7\times$ time overhead. Indeed, we have analyzed all possible scenarios, and the conclusions are discussed in Table 1.

## 2 Related Work

As early as 2017, Galloway et al. (2017) reveals that the binarized models exhibit adversarial robustness, but it is later known this comes from obfuscated gradients (Athalye et al., 2018). Subsequently, Lin et al. (2019) argues that quantized networks suffer from error amplification effects, making them more vulnerable to adversarial attacks. They address this risk by designing regularization terms. Sen et al. (2020) claims that there is an accuracy-robustness trade-off during quantization, and assembles full-precision and quantized networks by ensembling methods to better reconcile this trade-off. Fu et al. (2021) further enhances model robustness by designing a stochastic precision inference scheme. However, these studies are hard to reveal the real tendency of quantization robustness, as they are not conducted at a unified pipeline, which even leads to conflicting conclusions.

Another parallel set of works involves evaluating the robustness and reliability of quantized networks. Xiao et al. (2023) only assesses the robustness of QAT without any robust components, while Yuan et al. (2023) discusses aspects such as category performance differences and out-of-distribution performance instead of adversarial robustness on Post-Training Quantization networks. In contrast, our paper focuses on the adversarial robustness of PTQ and QAT, discussing the impact of various choices on the components of the quantization pipeline.

# 3 ROBUST QUANTIZATION OVERVIEW

In this section, we first introduce the pipeline for model quantization, extracting components that may influence the adversarial robustness of the model in the process. These components will be analyzed in the next section. Furthermore, we highlight the evaluation of quantized models with adversarial robustness metrics in resource-constrained environments.

## 3.1 QUANTIZATION PIPELINE

As shown in Figure 1, whether it is PTQ or QAT, we can divide the quantization process into two key stages: the preparation stage and the quantization stage. First, it is necessary to determine the initialization parameters of the model and the size of the dataset used for calibration or fine-tuning. Then, the bit-width for quantization and the training strategy for fine-tuning are determined to obtain the quantized model. To account for the model's robustness, We consider making different choices in various components to assess the robustness of different pipelines. Specifically, it is crucial to consider whether to use adversarial training full-precision model parameters to initialize the quantization model, and whether to combine adversarial training with quantization during the training process. We aim to observe the effects of quantizing to different bit-widths on robustness under various settings. Additionally, it is important to investigate the impact of choices regarding dataset size on robustness.

## 3.2 EVALUATE METHOD

The evaluation methods focus on the performance of the quantization model. It is generally measured using the accuracy of adversarial examples (Dong et al., 2020). As improving the adversarial robustness of a model may impact its accuracy, it is necessary to consider both aspects simultaneously. Therefore, we propose using adversarial accuracy as a metric for robustness (denoted as $Robustness$) while also calculating its accuracy on clean samples (denoted as $Accuracy$).

We use attack model $\mathcal{A}_{\epsilon,p}$ with budget $\epsilon$ under the $L_p$ norm to generate adversarial sample $x^i_{adv} = \mathcal{A}_{\epsilon,p}(x^i)$. For the upcoming evaluation of model $\mathcal{M}$, we can compute the accuracy using the formula $Accuracy(\mathcal{M}) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(x^i = y_i)$, and compute the robust using the formula $Robustness(\mathcal{M}) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\mathcal{A}_{\epsilon,p}(x^i) = y_i)$, where $\{x_i, y_i\}$ is the test set and $\mathbb{I}(\cdot)$ is the indicator function.

In addition, due to the addition of robustness components during the quantization process, there may be additional time overhead. This could be particularly challenging for resource-constrained scenarios. Therefore, we compare the relative time overheads of different methods by theoretical analysis to assess their impact.

# 4 EXPERIMENTS AND INSIGHTS

In this section, we will discuss in detail the impact of different pipeline configurations on the robustness of quantized models. In Section 4.1, we will explicitly define the various components present in the pipeline. Furthermore, for each component, we will delve into a detailed discussion of the potential different choices that may exist. In Section 4.2, we will elaborate on the specific experimental design. Section 4.3 will then present and analyze the experimental results.

## 4.1 VARIOUS COMPONENTS IN QUANTIZATION PIPELINE

**FP Model Parameters.** In PTQ, adopting the parameters of the full-precision model is necessary for initialization (Li et al., 2021). In QAT, the model can be trained from scratch with random initialization or fine-tuned using parameters initialized from the full-precision model (Tang et al., 2022; 2023) for fast convergence. We aim to understand the impact of initialization parameters that have undergone adversarial training on the quantized model.

**Quantization Precision.** Before starting the quantization process, it is necessary to assess the deployment environment to determine the desired bit-width for quantization. There is a trade-off between computational resources and accuracy. Generally, lower quantization bit-widths have lower

requirements for computational resources, but the accuracy will decrease (Esser et al., 2019; Li et al., 2021). However, the impact of precision on adversarial robustness has not been widely explored yet. In this part, we focus on the impact of the bit-width of quantized networks on the model's robustness.

**Training Strategy (for QAT).** During the QAT process, the quantized network undergoes training or fine-tuning. We aim to understand the impact on the robustness of the quantized network when using the incorporating adversarial training method compared to original training. Due to the coupling of training strategies and initialization parameters, we need to consider all possible scenarios.

**Other Adversarial Learning Method.** In addition, some robust learning methods are inherently suitable for quantized models. Approaches like the robust transfer network (Vaishnavi et al., 2022) can obtain both robustness and accuracy through knowledge distillation from a well-trained robust model. We aim to determine whether this method is equally applicable to quantized models.

### 4.2 Experimental Setup

**Dataset and Network Architecture.** We aim to investigate the issues of adversarial robustness of quantization models in widely used computer vision tasks. Therefore, we have chosen the fundamental task of image classification for experimentation. The dataset used in the experiments is CIFAR-10 (Krizhevsky et al., 2009), which is commonly employed for adversarial attack scenarios. The network architecture selected for the experiments is ResNet20 (He et al., 2016), a proven structure known for its effective application across various tasks.

**Quantization Methods.** We adopt the PWLQ method (Fang et al., 2020) as the PTQ approach for our experiments. PWLQ employs a segmented linear quantization technique, providing finer-grained quantization values for intervals with more data distribution. This allows better accuracy at lower bit-width. For QAT, we choose PACT (Choi et al., 2018) as the experimental method, which introduces learnable bounds of activations to filter outliers adaptively. For a fair comparison, we set the fine-tuning training to 60 epochs, while training from scratch is 200 epochs.

**Adversarial Attack and Defense.** When discussing the robustness introduced by quantization itself, it is essential to measure performance under both white-box and black-box attack scenarios to avoid the obfuscated gradients (Athalye et al., 2018) in measured robustness. For white-box attacks, we choose the PGD attack method (Madry et al., 2017), known as the strongest first-order white-box attack method. For black-box attacks, AutoAttack (Croce & Hein, 2020) is selected, known for its outstanding performance and widespread use in robustness evaluations. In examining the impact of robustness components on quantized networks, we opt for the PGD adversarial training method (Madry et al., 2017) to obtain robustness for the corresponding components. We measure the accuracy and robustness of the quantized model through PGD attacks.

### 4.3 Results and Analysis

Figure 2 shows the accuracy and robustness under different attacks of quantized networks obtained through random initialization in QAT. We observe that in quantized networks without any robustness components, they are unable to resist attacks with common attack settings ($\epsilon = 8/255$). However, as the attack intensity decreases, both PTQ and QAT quantized networks exhibit an increase in robustness with the reduction of quantization bit-width. We also observe that activation quantization enhances robustness more noticeably compared to weight quantization. For instance, under the PGD-20 attack with $\epsilon = 1/255$, quantizing activations from full precision to 2-bit results in a robustness improvement of $33.47\%$. In the same setting, weight quantization only achieves a robustness improvement of $7.31\%$.

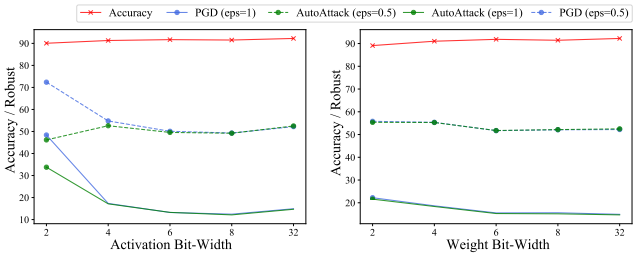


Figure 2: Accuracy and robustness under different attacks in QAT. The left diagram represents the results of keeping weights in full precision and quantizing activations, while the right diagram represents the results of keeping activations in full precision and quantizing weights.

Table 2: The impact of FP model parameters and training strategies on accuracy and robustness. "NC" represents non convergence.

| Quantization Methods | FP Model Parameters | A2W2 | A4W4 | A6W6 | A8W8 | A32W32 |
|---|---|---|---|---|---|---|
| PTQ | Non-Robust Parameters | NC | 87.91 / 0.02 | 91.84 / 0.00 | 91.98 / 0.00 | 92.15 / 0.00 |
| | Robust Parameters | NC | 76.16 / 47.84 | 76.95 / 47.74 | 76.95 / 46.58 | 77.06 / 46.43 |
| QAT (Vanilla Training) | Random Parameters | 86.99 / 0.01 | 91.18 / 0.00 | 91.34 / 0.00 | 91.27 / 0.00 | 91.71 / 0.00 |
| | Non-Robust Parameters | NC | 91.66 / 0.00 | 92.18 / 0.00 | 92.33 / 0.00 | 92.24 / 0.00 |
| | Robust Parameters | NC | 88.95 / 0.10 | 90.28 / 0.13 | 90.48 / 0.03 | 90.93 / 0.00 |
| QAT (Adversarial Training) | Random Parameters | NC | 72.17 / 45.83 | 74.04 / 46.48 | 74.41 / 46.79 | 76.30 / 46.83 |
| | Non-Robust Parameters | NC | 64.95 / 50.29 | 67.51 / 37.14 | 67.23 / 36.89 | 70.65 / 41.25 |
| | Robust Parameters | NC | 76.18 / 48.48 | 77.62 / 47.57 | 77.41 / 47.36 | 77.46 / 46.34 |

Table 2 displays the accuracy and robustness under $\epsilon = 8/255$ of quantized models obtained using different FP model parameters and selecting various training strategies. We found that models obtained through PTQ using robust parameters can still maintain the original robustness. The robustness and accuracy decrease as the quantization bit-width decreases. This may be because PTQ does not fine-tune the model but only affects its performance by introducing quantization noise. Therefore, it does not have an additional impact on robustness. For QAT, despite introducing robust initialization parameters, after fine-tuning on a dataset with clean samples, the robustness provided by the initialization diminishes to $0.1\%$. However, there is a $11.89\%$ increase in accuracy. When adversarial training is added to the training process of QAT, it can maintain the robustness provided by the initialization. Even if random initialization is provided, QAT can still achieve robustness through adversarial training. We also observed that when the provided initialization parameters are obtained without adversarial training, using adversarial training in QAT can lead to a significant decrease in model accuracy. For instance, in the A4W4 setting, its accuracy drops by $12.11\%$ compared to the full-precision robust model.

Incorporating adversarial training into the training strategy of QAT introduces a significant amount of additional time overhead. As each training epoch requires the generation of adversarial samples, following the classical PGD-7 adversarial training approach (Madry et al., 2017), each training epoch involves seven backward pass computations for adversarial samples, and a backward pass is performed to achieve gradient optimization. This results in a sevenfold increase in time overhead compared to regular training.

Table 3: Accuracy and robustness on robust transfer learning with various dataset sizes.

| Datasets Size | A4W4 | A6W6 | A8W8 | A32W32 |
|---|---|---|---|---|
| 1% | 66.89 / 43.42 | 76.45 / 47.93 | 77.85 / 45.65 | 77.24 / 46.86 |
| 10% | 73.78 / 44.75 | 76.69 / 48.21 | 77.35 / 48.10 | 77.05 / 47.51 |
| 100% | 74.40 / 45.10 | 76.41 / 48.50 | 76.72 / 48.16 | 76.71 / 47.87 |

Table 3 represents the results when QAT with the robust transfer learning method proposed by Vaishnavi et al. (2022) from robust initialization parameters, and other settings will be shown in Appendix B. We analyzed the effect of robust transfer learning using a partial dataset. It can be observed that this method still performs well when transferred to quantized models. At higher quantization bit-widths, it can even surpass the robustness of full-precision models. For example, in the A6W6 quantization setting, its robustness increases by $2.07\%$ compared to the full-precision adversarially trained model. However, as the quantization bit-width decreases, both accuracy and robustness experience a slight decline, possibly due to a reduction in model capacity with decreasing bit-width. Additionally, even when quantizing models using a partial dataset, good accuracy and robustness can still be achieved. It is worth noting that, due to the absence of the time required to compute adversarial samples, this method is more efficient in terms of quantizing time compared to the QAT approach combined with adversarial training.

## 5 Conclusion

In this paper, we first introduce the evaluation metrics for adversarial robustness in quantized networks. By studying the impact of quantization itself on robustness and the effects of adding robustness components, we aim to provide valuable insights for the design and evaluation of robust quantization pipelines. We hope that this work can serve as a useful step in improving robustness in quantized networks.

ACKNOWLEDGMENTS

This work was supported in part by the National Key Research and Development Project of China (Grant No. 2023YFF0905502), Shenzhen Science and Technology Program (Grant No. RCYX20200714114523079 and JCYJ20220818101014030).

## REFERENCES

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 274–283. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/athalye18a.html.

Jungwook Choi, Zhuo Wang, Swagath Venkataramani, Pierce I-Jen Chuang, Vijayalakshmi Srinivasan, and Kailash Gopalakrishnan. Pact: Parameterized clipping activation for quantized neural networks. *arXiv preprint arXiv:1805.06085*, 2018.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020.

Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, and Jun Zhu. Benchmarking adversarial robustness on image classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 318–328, 2020. doi: 10.1109/CVPR42600.2020.00040.

Steven K Esser, Jeffrey L McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S Modha. Learned step size quantization. *arXiv preprint arXiv:1902.08153*, 2019.

Jun Fang, Ali Shafiee, Hamzah Abdel-Aziz, David Thorsley, Georgios Georgiadis, and Joseph H Hassoun. Post-training piecewise linear quantization for deep neural networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pp. 69–86. Springer, 2020.

Yonggan Fu, Qixuan Yu, Meng Li, Vikas Chandra, and Yingyan Lin. Double-win quant: Aggressively winning robustness of quantized deep neural networks via random precision training and inference. In *International Conference on Machine Learning*, pp. 3492–3504. PMLR, 2021.

Angus Galloway, Graham W Taylor, and Medhat Moussa. Attacking binarized neural networks. *arXiv preprint arXiv:1711.00449*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Yuzheng Hu, Fan Wu, Hongyang Zhang, and Han Zhao. Understanding the impact of adversarial robustness on accuracy disparity. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13679–13709. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/hu23j.html.

Dewant Katare, Diego Perino, Jari Nurmi, Martijn Warnier, Marijn Janssen, and Aaron Yi Ding. A survey on approximate edge ai for energy efficient autonomous driving services. *IEEE Communications Surveys & Tutorials*, 2023.

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

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

Haowen Lin, Jian Lou, Li Xiong, and Cyrus Shahabi. Integer-arithmetic-only certified robustness for quantized neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7828–7837, 2021.

Ji Lin, Chuang Gan, and Song Han. Defensive quantization: When efficiency meets robustness. *arXiv preprint arXiv:1904.08444*, 2019.

Hantao Liu, Ying He, F Richard Yu, and Jeremy James. Flexi-compression: a flexible model compression method for autonomous driving. In *Proceedings of the 11th ACM Symposium on Design and Analysis of Intelligent Vehicular Networks and Applications*, pp. 19–26, 2021.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart Van Baalen, and Tijmen Blankevoort. A white paper on neural network quantization. *arXiv preprint arXiv:2106.08295*, 2021.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3505–3506, 2020.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer, 2015.

Sanchari Sen, Balaraman Ravindran, and Anand Raghunathan. Empir: Ensembles of mixed precision deep networks for increased robustness against adversarial attacks. *arXiv preprint arXiv:2004.10162*, 2020.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

Chen Tang, Kai Ouyang, Zhi Wang, Yifei Zhu, Wen Ji, Yaowei Wang, and Wenwu Zhu. Mixed-precision neural network quantization via learned layer-wise importance. In *European Conference on Computer Vision*, pp. 259–275. Springer, 2022.

Chen Tang, Kai Ouyang, Zenghao Chai, Yunpeng Bai, Yuan Meng, Zhi Wang, and Wenwu Zhu. Seam: Searching transferable mixed-precision quantization policy through large margin regularization. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 7971–7980, 2023.

Pratik Vaishnavi, Kevin Eykholt, and Amir Rahmati. Transferring adversarial robustness through robust representation matching. In *31st USENIX Security Symposium (USENIX Security 22)*, pp. 2083–2098, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yisong Xiao, Aishan Liu, Tianyuan Zhang, Haotong Qin, Jinyang Guo, and Xianglong Liu. Robustmq: benchmarking robustness of quantized models. *Visual Intelligence*, 1(1):30, 2023.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. *Advances in Neural Information Processing Systems*, 35:27168–27183, 2022.

Zhihang Yuan, Jiawei Liu, Jiaxiang Wu, Dawei Yang, Qiang Wu, Guangyu Sun, Wenyu Liu, Xinggang Wang, and Bingzhe Wu. Benchmarking the reliability of post-training quantization: a particular focus on worst-case performance. *arXiv preprint arXiv:2303.13003*, 2023.

Rongzhao Zhang and Albert CS Chung. Medq: Lossless ultra-low-bit neural network quantization for medical image segmentation. *Medical Image Analysis*, 73:102200, 2021.

## A ATTACK STRENGTH AND QUANTIZATION SETTINGS

**Adversarial Attack Intensity.** The adversarial sample construction in adversarial attacks can be denoted as $x_{adv} = x + argmin_{\delta}\{||\delta|| : f(x+\delta) \neq f(x), \delta \leq \epsilon\}$, where $\epsilon$ is a small quantity representing the magnitude of a slight perturbation applied to the input sample. In PGD attacks conducted on the CIFAR-10 dataset, it is common to set $\epsilon \leq 8/255$, so we define it as a "normal attack". When $\epsilon \leq 2/255$, full-precision model without adversarial training may exhibit a certain level of robustness which is referred to as a "weaker attack".

**Quantization Settings.** Quantization settings include PTQ and QAT. For the QAT method, it requires either Fine-Tuning or Training. Specifically, the training strategy using the original QAT is referred to as QAT (Vanilla Training). When generating adversarial samples in the quantized model and employing Adversarial Training (Madry et al., 2017), the training strategy is termed QAT (Adversarial Training). The approach of using the quantized model as a student model for adversarial transfer learning (Vaishnavi et al., 2022) is referred to as QAT (Adversarial Transfer Learning).

## B MORE DETAILS ABOUT TRANSFER ADVERSARIAL LEARNING

For Transfer Adversarial Learning, the first step involves training a robust full-precision model as the teacher model. Subsequently, based on the chosen quantization bit-width, a quantized model will serve as the student model, employing the loss function proposed by Vaishnavi et al. (2022) for knowledge distillation, allowing simultaneous learning of both accuracy and robustness.

Table 4: Accuracy of Transfer Adversarial Learning under Different Initialization. The Accuracy of the Teacher Model (Full-Precision Model) is $77.06\%$.

| **Initializing Parameters** | **A3W3** | **A4W4** | **A6W6** | **A8W8** | **A32W32** |
|---|---|---|---|---|---|
| Random Parameters | 72.93 | 74.21 | 74.62 | 74.26 | 72.18 |
| Non-Robust Parameters | 67.83 | 69.98 | 69.90 | 70.23 | 71.55 |
| Robust Parameters | NC | 74.40 | 76.41 | 76.72 | 76.71 |

Table 5: Robustness of Transfer Adversarial Learning under Different Initialization. The Robustness of the Teacher Model (Full-Precision Model) is $46.43\%$.

| **Initializing Parameters** | **A3W3** | **A4W4** | **A6W6** | **A8W8** | **A32W32** |
|---|---|---|---|---|---|
| Random Parameters | 29.93 | 34.58 | 35.66 | 34.75 | 40.66 |
| Non-Robust Parameters | 16.97 | 19.40 | 18.71 | 18.52 | 31.64 |
| Robust Parameters | NC | 45.10 | 48.50 | 48.16 | 47.87 |

In this method, the initialization parameters of the student model can also be chosen. We set them to be initialized from an adversarial trained full-precision model (robust initializing parameters), a vanilla trained full-precision model (non-robust initializing parameters), and a scratch-trained model. Simultaneously, we specify the teacher model to be an adversarial trained full-precision model for experimentation.

Table 4 presents the accuracy obtained under different initialization parameters and quantization bit-widths, while Table 5 displays the robustness. In robust and non-robust initialization methods, accuracy consistently demonstrates a decrease with the reduction of quantization bit-widths. For robustness, it can be observed that initializing the student model with a robust full-precision model yields the best robustness. In fact, in certain quantization bit-width, it even surpasses the robustness of the teacher model. For instance, the A8W8 configuration exhibits a $1.73\%$ improvement in robustness compared to the full-precision robust model. However, in this setup, robustness experiences a slight decline with decreasing quantization bit-width.

The method of initializing with non-robust parameters or training from scratch exhibits a significant decline in robustness performance on the quantized network compared to the full-precision network. For instance, in the case of non-robust initialization, using an A3W3 quantized network results in a $14.67\%$ decrease in robustness compared to the full-precision network. This indicates the crucial role of model initialization in robustness within the context of Transfer Adversarial Learning.