Benchmarking the Robustness of Quantized Models

Yisong Xiao[1,2], Tianyuan Zhang[2], Shunchang Liu[1,2], Haotong Qin[1,2]
[1] Shen Yuan Honors College, Beihang University
[2] State Key Lab of Software Development Environment, Beihang University
{xiaoyisong, 19373397, liusc, qinhaotong}@buaa.edu.cn

Abstract

Quantization has emerged as an essential technique for deploying deep neural networks (DNNs) on devices with limited resources. However, quantized models exhibit vulnerabilities when exposed to various noises in real-world applications. Despite the importance of evaluating the impact of quantization on robustness, existing research on this topic is limited and often disregards established principles of robustness evaluation, resulting in incomplete and inconclusive findings. To address this gap, we thoroughly evaluated the robustness of quantized models against various noises (adversarial attacks, natural corruptions, and systematic noises) on ImageNet. Extensive experiments demonstrate that lower-bit quantization is more resilient to adversarial attacks but is more susceptible to natural corruptions and systematic noises. Notably, our investigation reveals that impulse noise (in natural corruptions) and the nearest neighbor interpolation (in systematic noises) have the most significant impact on quantized models. Our research contributes to advancing the robust quantization of models and their deployment in real-world scenarios.

1. Introduction

Deep neural networks (DNNs) have demonstrated impressive performance in a broad range of applications [10, 28]. However, deploying DNNs on resource-constrained devices, such as IoT devices, poses significant challenges. To address this issue, researchers have proposed various model compression techniques, including model quantization [11, 20] and pruning [6, 7]. Among these techniques, model quantization has become a critical approach for compressing DNNs due to its ability to maintain network structure and achieve comparable performance. This technique involves mapping the network parameters from 32-bit floating-point numbers to low-bit integers, resulting in reduced memory usage and faster inference.

Despite their impressive performance, DNNs are highly susceptible to adversarial examples [5, 13, 15, 18]. Adversarial examples are perturbations that are designed to be undetectable to human vision but can easily deceive DNNs, posing a significant threat to practical deep learning applications. In addition, DNNs are vulnerable to natural corruptions [9] such as snow and motion blur, which are common in real-world scenarios and can significantly reduce the accuracy of DNN models. Moreover, system noises resulting from the mismatch between software and hardware can also have a detrimental impact on model accuracy [26]. These phenomena demonstrate that quantized networks deployed in safety-critical applications are unreliable when faced with various perturbations in the real world.

Therefore, it is critical to conduct a comprehensive evaluation of the robustness of quantized models before deploying them to identify potential weaknesses and unintended behaviors. While numerous studies have extensively investigated the robustness of floating-point networks against various attacks and metrics, research on the robustness of quantized models [1, 12] remains inadequate. These studies lack diversity in terms of noise sources and rely solely on small datasets, leading to inconclusive findings regarding the robustness of quantized models.

We build the robustness evaluation benchmark of quantized models. Our benchmark assesses the robustness of quantized models using 3 popular quantization methods (DoReFa, PACT, and LSQ) and 4 classical architectures (ResNet18, ResNet50, RegNetX600M, and MobileNetV2). For each method, we evaluate 4 commonly used bit-widths. Our analysis includes 3 progressive adversarial attacks, 15 natural corruptions, and 14 systematic noises on the ImageNet benchmark. Our empirical results demonstrate that lower-bit quantized models display better adversarial robustness but are more susceptible to natural corruptions and systematic noises. We identify *impulse noise* and *the nearest neighbor interpolation* as the most harmful sources of natural corruptions and systematic noises, respectively.

2. Related Work

2.1. Network Quantization

Network quantization compresses DNN models by reducing the number of bits required to represent each weight to save memory usage and speed up hardware in-

ference. A classic quantization process (*quantization* and *de-quantization*) can be formulated by:

$$Q(r) = \mathbf{Int}(r/S) - Z, \quad r^{'} = S \cdot (Q(r) + Z) \quad (1)$$

where Q is the quantization operator, r and $r^{'}$ is real value and de-quantized real value respectively, S and Z denote *scale* and *zero-point* respectively. Given t bits, the range after quantization is determined by $[-2^{t-1}, 2^{t-1} - 1]$.

We here provide a brief review of the commonly used quantization methods. One string of research designed rules to fit the quantizer to the data distribution. For example, DoReFa-Net [29] simply clips the activation to $[0, 1]$ and then quantizes it, due to the observation that most activation falls into this range in many network architectures. Other notable work focused on learning appropriate quantization parameters during the backpropagation process. PACT [2] clip the activation by a handcrafted parameter and optimize the clipping threshold. Notice that PACT has no gradient below the clip point, LSQ [4] learns the *scale* alongside network parameters by estimating the gradient at each weight and activation layer.

2.2. Adversarial Attacks

Adversarial examples are inputs with small perturbations that could easily mislead the DNNs [5]. Formally, given a DNN f_Θ and an input $\mathbf{x}$ with the ground truth label $\mathbf{y}$, an adversarial example $\mathbf{x}_{adv}$ satisfies

$$f_\Theta(\mathbf{x}_{adv}) \neq \mathbf{y} \quad s.t. \quad \|\mathbf{x} - \mathbf{x}_{adv}\| \leq \epsilon, \quad (2)$$

where $\|\cdot\|$ is a distance metric and commonly measured by the ℓ_p-norm ($p \in \{1,2,\infty\}$).

A long line of work has been dedicated to performing adversarial attacks [3,5,14,15,17,19,24], which can be mainly divided into white-box and black-box manners based on access to the target model. For white-box attacks, adversaries have complete knowledge of the target model and can fully access it; while for black-box attacks, adversaries have limited or even without any knowledge of the target model and can not directly access it. This paper primarily employs white-box attacks to evaluate the adversarial robustness of target models, as they offer stronger attack capabilities.

2.3. Robustness of Quantized Models

A number of studies have been proposed to evaluate the robustness of floating-point networks [16,23,25,27]. However, the robustness of quantized networks has been relatively underexplored. Lin *et al*. [12] proposed a defensive quantization method to suppress the amplification of adversarial noise during propagation by controlling the Lipschitz constant of the network during quantization. Similarly, Alizadeh *et al*. [1] also designed a regularization scheme to improve the robustness of the quantized model by controlling the magnitude of adversarial gradients. In this paper, we aim to thoroughly evaluate the robustness of quantized models against multiple noises for several quantization methods, architectures, and quantization bits.

3. Evaluation Protocols

3.1. Evaluation Objects

Dataset. We conduct evaluations on the large-scale ImageNet dataset with 1,000 classes, which comprises 1.2 million training images and 50,000 validation images.

Architectures. We consider four architectures, including ResNet18 [8], ResNet50 [8], RegNetX600M [21], and MobileNetV2 [22]. (1) ResNet18 and ResNet50 are classical backbone architectures that are widely used in various computer vision tasks. (2) RegNetX600M is an advanced architecture with group convolution discovered through model structure search. (3) MobileNetV2 is a lightweight network designed for efficient deployment on edge devices, featuring depthwise separable convolutions.

Quantization Methods. We focus on three popular quantization methods, including DoReFa [29], PACT [2], and LSQ [4]. For the choice of quantization bits, we adopt the commonly used set in deployments (*i.e*., 2, 4, 6, and 8).

For each architecture, we quantize models on the ImageNet training set starting from the same floating-point model, then evaluate their robustness against perturbations generated on the ImageNet validation set.

3.2. Robustness Evaluation Approaches

Quantized models are vulnerable to various perturbations in real-world scenarios. We classify these perturbations into adversarial attacks, natural corruptions, and systematic noises, following the guidelines proposed in [23].

Adversarial attacks. To craft adversarial perturbations with progressively increasing attack capabilities, we employ FGSM-ℓ_∞, PGD-ℓ_1, PGD-ℓ_2, PGD-ℓ_∞ and AutoAttack-ℓ_∞. For each attack, we set three different perturbation magnitudes (small, middle, and large).

Natural corruptions. To simulate natural corruptions, we utilize 15 distinct perturbation methods from the ImageNet-C benchmark [9] that fall into four categories: noise, blur, weather, and digital.

Systematic noises. Moreover, system noises are always present when models are deployed in edge devices due to changes in hardware or software. To evaluate the impact of system noises on quantized models, we utilize pre-processing operations from ImageNet-S [26], including three frequently used decoders and seven commonly used resize modes.

3.3. Evaluation Metrics

Adversarial robustness. For specific adversarial attacks, we measure adversarial robustness (AR) using model accuracy, where higher AR indicates a stronger model. For the union of different attacks, we adopt the Worst-Case Ad-

versarial Robustness ($WCAR$) to measure adversarial robustness (higher indicates a stronger model):

$$WCAR = 1 - P_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\text{Any}(f(\mathcal{A}^{f}_{\epsilon,p}(\mathbf{x})) \neq \mathbf{y}), \quad (3)$$

where $\mathcal{A}_{\epsilon,p}$ represents the adversary, $\mathcal{D}$ is the validation set, and $\text{Any}(\cdot)$ is a function that returns true if any of the adversaries attacks successfully.

Natural robustness. We adopt the average accuracy of the quantized model on all corruptions (denoted as C) to measure natural robustness, denoted as NR:

$$NR = \mathbb{E}_{c\sim C}(P_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}(f(c(\mathbf{x})) = \mathbf{y})), \quad (4)$$

where c denotes a corruption method. A higher value of NR means better natural robustness.

Systematic Robustness. We adopt the model stability on different systematic noises (denoted as S) to measure systematic robustness (SR). Specifically, we calculate the standard deviation of accuracy as SR:

$$SR = \mathbb{D}_{s\sim S}(P_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}(f(s(\mathbf{x})) = \mathbf{y})), \quad (5)$$

where s denotes a decode or resize method. A lower value of SR means better stability towards systematic noises.

4. Empirical Results

In this section, we present the results mainly obtained on the ResNet18 architecture.

4.1. Clean Accuracies

Tab. 1 reports the clean accuracies of quantized models. Most of the quantized models maintain comparable accuracy to the 32-bit ResNet18. However, 2-bit PACT fails to converge, resulting in a mere 2.61% accuracy. Therefore, we exclude it from the subsequent robustness evaluations.

Table 1. Clean accuracies of the ResNet18 for all models. "NC" denotes not converged.

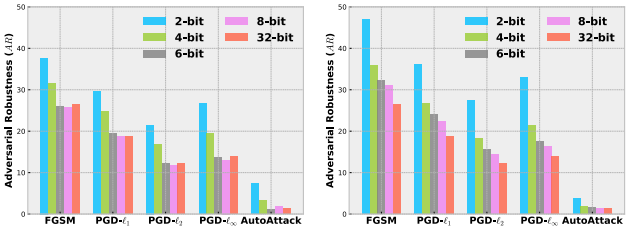
Model	Method	2bit	4bit	6bit	8bit
ResNet18 FP: 71.06	DoReFa	62.31	70.60	71.15	**71.51**
	PACT	NC	70.41	71.17	**71.43**
	LSQ	65.97	70.52	71.10	**71.31**

4.2. Evaluation of Adversarial Attacks

Since $WCAR$ degrades to 0 under medium and high perturbation magnitudes, we here only present the results under small magnitudes. From the results shown in Tab. 2, we could make the following observations: (1) Unlike the decrease in clean accuracy, lower-bit models exhibit higher worst-case adversarial robustness and are almost better than floating-point network; (2) At the same quantization bit, PACT presents the best adversarial robustness compared to other quantization methods. Furthermore, we compare the adversarial robustness (AR) of quantized models against specific attacks. As depicted in Fig. 1, we can identify that LSQ performs better under FGSM and PGD attacks, yet is vulnerable to AutoAttack (the strongest of these attacks).

Table 2. Worst-Case Adversarial Robustness ($WCAR\uparrow$) of ResNet18 models on small magnitude.

Model	Method	2bit	4bit	6bit	8bit
ResNet18 FP: 1.30	DoReFa	**5.55**	2.86	1.11	1.49
	PACT	NC	**4.03**	1.94	1.50
	LSQ	**3.73**	1.78	1.49	1.36



(a) ResNet18 quantized by DoReFa (b) ResNet18 quantized by LSQ

Figure 1. Adversarial Robustness against specific attacks.

4.3. Evaluation of Natural Corruptions

We present the Natural Robustness of ResNet18 models in Tab. 3 and show the detailed results of LSQ in Fig. 2. Though performing similar clean accuracy with the 32-bit model, quantized models are more vulnerable under natural corruptions, especially in the 2-bit models. And we can find that *impulse noise* shows the highest impact on the model's robustness (about 50% average decrease), while brightness is the least harmful (about 10% average decrease).

Table 3. Natural Robustness ($NR\uparrow$) of ResNet18 models.

Model	Method	2bit	4bit	6bit	8bit
ResNet18 FP: **32.78**	DoReFa	23.30	30.88	**31.79**	31.70
	PACT	NC	30.36	31.50	**31.69**
	LSQ	26.42	30.95	31.67	**31.70**

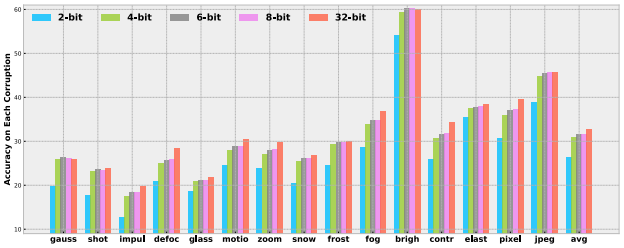


Figure 2. Accuracy of ResNet18 models quantized by LSQ under each corruption. The 'avg' (rightest) is the value of NR.

4.4. Evaluation of Systematic Noises

On the systematic noises, we also observe that lower-bit models present less robustness (*i.e.*, lower stability), as shown in Tab. 4. Moreover, we find that among 14 systematic noises, the nearest neighbor interpolation methods in Pillow and OpenCV have the greatest impact on the model performance, which induce nearly a 6% decrease in performance for the 2-bit models. It indicates that maintaining consistency between the deployment and training process is crucial to avoid unnecessary accuracy loss.

Table 4. Systematic Robustness ($SR\downarrow$) of ResNet18 models.

Model	Method	2bit	4bit	6bit	8bit
ResNet18 FP: **0.53**	DoReFa	2.32	1.83	**1.69**	1.95
	PACT	NC	1.85	**1.84**	1.98
	LSQ	2.05	1.90	1.87	**1.82**

5. Conclusion

We presents a benchmark for evaluating the robustness of quantized models under various perturbations, including adversarial attacks, natural corruptions, and systematic noises. The benchmark evaluates 4 classical architectures and 3 popular quantization methods with four different bit-widths. The results reveal that lower-bit quantized models have higher adversarial robustness than their floating-point counterparts, but are more vulnerable to natural corruptions (especially impulse noise) and systematic noises (especially the nearest neighbor interpolation). We hope our benchmark will advance the development and deployment of robust quantized models in real-world scenarios.

Acknowledgment This work was supported by the Academic Excellence Foundation of BUAA for Ph.D. Students.

References

[1] M. Alizadeh, A. Behboodi, M. van Baalen, C. Louizos, T. Blankevoort, and M. Welling. Gradient ℓ_1 regularization for quantization robustness. *ArXiv*, 2020. 1, 2

[2] J. Choi, Z. Wang, S. Venkataramani, P. I.-J. Chuang, V. Srinivasan, and K. Gopalakrishnan. Pact: Parameterized clipping activation for quantized neural networks. *ArXiv*, 2018. 2

[3] F. Croce and M. Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*. 2

[4] S. K. Esser, J. L. McKinstry, D. Bablani, R. Appuswamy, and D. S. Modha. Learned step size quantization. *ArXiv*, 2019. 2

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *ArXiv*, 2014. 1, 2

[6] J. Guo, J. Liu, and D. Xu. Jointpruning: Pruning networks along multiple dimensions for efficient point cloud processing. *IEEE TCSVT*, 2021. 1

[7] J. Guo, W. Ouyang, and D. Xu. Multi-dimensional pruning: A unified framework for model compression. In *CVPR*, 2020. 1

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*. 2

[9] D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *ArXiv*, 2019. 1, 2

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*. 1

[11] Y. Li, M. Shen, J. Ma, Y. Ren, M. Zhao, Q. Zhang, R. Gong, F. Yu, and J. Yan. Mqbench: Towards reproducible and deployable model quantization benchmark. *ArXiv*, 2021. 1

[12] J. Lin, C. Gan, and S. Han. Defensive quantization: When efficiency meets robustness. *ArXiv*, 2019. 1, 2

[13] A. Liu, J. Guo, J. Wang, S. Liang, R. Tao, W. Zhou, C. Liu, X. Liu, and D. Tao. X-adv: Physical adversarial object attacks against x-ray prohibited item detection. *ArXiv*, 2023. 1

[14] A. Liu, T. Huang, X. Liu, Y. Xu, Y. Ma, X. Chen, S. J. Maybank, and D. Tao. Spatiotemporal attacks for embodied agents. In *ECCV*. 2

[15] A. Liu, X. Liu, J. Fan, Y. Ma, A. Zhang, H. Xie, and D. Tao. Perceptual-sensitive gan for generating adversarial patches. In *AAAI*, 2019. 1, 2

[16] A. Liu, X. Liu, H. Yu, C. Zhang, Q. Liu, and D. Tao. Training robust deep neural networks via adversarial noise propagation. *IEEE TIP*. 2

[17] A. Liu, J. Wang, X. Liu, B. Cao, C. Zhang, and H. Yu. Bias-based universal adversarial patch attack for automatic check-out. In *ECCV*, 2020. 2

[18] S. Liu, J. Wang, A. Liu, Y. Li, Y. Gao, X. Liu, and D. Tao. Harnessing perceptual adversarial patches for crowd counting. In *CCS*, 2022. 1

[19] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *ArXiv*, 2017. 2

[20] H. Qin, M. Zhang, Y. Ding, A. Li, Z. Cai, Z. Liu, F. Yu, and X. Liu. Bibench: Benchmarking and analyzing network binarization. *ArXiv*, 2023. 1

[21] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár. Designing network design spaces. In *CVPR*. 2

[22] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*. 2

[23] S. Tang, R. Gong, Y. Wang, A. Liu, J. Wang, X. Chen, F. Yu, X. Liu, D. Song, A. Yuille, et al. Robustart: Benchmarking robustness on architecture design and training techniques. *ArXiv*, 2021. 2

[24] J. Wang, A. Liu, Z. Yin, S. Liu, S. Tang, and X. Liu. Dual attention suppression attack: Generate adversarial camouflage in physical world. In *CVPR*, 2021. 2

[25] J. Wang, Z. Yin, P. Hu, A. Liu, R. Tao, H. Qin, X. Liu, and D. Tao. Defensive patches for robust recognition in the physical world. In *CVPR*, 2022. 2

[26] Y. Wang, Y. Li, R. Gong, T. Xiao, and F. Yu. Real world robustness from systematic noise. In *ACMW AdvM*. 1, 2

[27] C. Zhang, A. Liu, X. Liu, Y. Xu, H. Yu, Y. Ma, and T. Li. Interpreting and improving adversarial robustness of deep neural networks with neuron sensitivity. *IEEE TIP*, 2020. 2

[28] Z. Zhao, J. Zhang, S. Xu, Z. Lin, and H. Pfister. Discrete cosine transform network for guided depth map super-resolution. In *CVPR*, 2022. 1

[29] S. Zhou, Y. Wu, Z. Ni, X. Zhou, H. Wen, and Y. Zou. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *ArXiv*, 2016. 2