Benchmarking the Reliability of Post-training Quantization: a Particular Focus on Worst-case Performance

Zhihang Yuan [*1] **Jiawei Liu** [*2] **Jiaxiang Wu** [3] **Dawei Yang** [1] **Qiang Wu** [1] **Guangyu Sun** [4] **Wenyu Liu** [2] **Xinggang Wang** [2] **Bingzhe Wu** [3]

Abstract

The reliability of post-training quantization (PTQ) methods in the face of extreme cases such as distribution shift and data noise remains largely unexplored, despite the popularity of PTQ as a method for compressing deep neural networks (DNNs) without altering their original architecture or training procedures. This paper conducts an investigation on commonly-used PTQ methods, addressing questions pertaining to the impact of calibration set distribution variations and calibration paradigm selection on the reliability of PTQ. Through a systematic evaluation process encompassing various tasks and commonly-used PTQ paradigms, it is evident that the majority of existing PTQ methods lack the necessary reliability for worst-case group performance, underscoring the imperative for more robust approaches.

1. Introduction

Deep neural networks (DNNs) are widely used in risk-sensitive areas such as autonomous driving (Muhammad et al., 2020) and finance (Zhang & Lou, 2021). DNNs with a large number of network parameters result in expensive computational and memory costs. As a model compression technique, post-training quantization (PTQ) offers the advantage of compressing DNNs without modifying the original training procedure, model structures, and parameters, making them highly desirable for practical applications.

During the calibration process, PTQ determines the quantization parameters to reduce the bits used for network weights and activations. The calibration dataset is small number of input samples, which is used to estimate the statistical properties of the activation values that the neural network produces during inference. These statistical properties are then used to determine the optimal quantization parameters that minimize the quantization error. Previous efforts have been made to improve PTQ from different dimensions such as devising new optimizing algorithms (Nagel et al., 2020; Li et al., 2021), especially in the low-bit regime. To date, state-of-the-art method can achieve nearly lossless accuracy on the image classification task in the 4-bit setting (Wang et al., 2022).

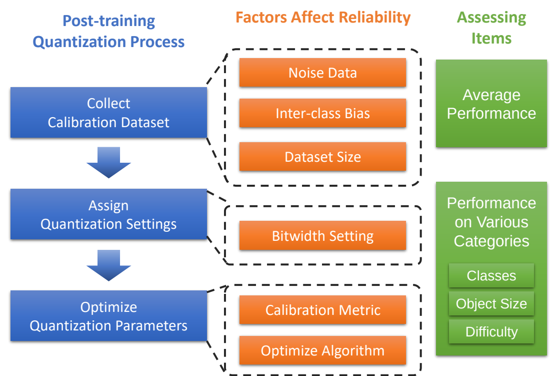


Figure 1. Overview of the framework for assessing reliability of PTQ.

PTQ assumes that the calibration and test sets follow the same distribution, referred to as the "close-environment" assumption. However, in real-world open environments, this assumption is often impractical due to distribution shifts between the calibration and test distributions. A natural question arises: *Is current PTQ method reliable enough when facing extreme test samples such as worst-case-category or out-of-distribution samples?* Answering this question is vital for the deployment of quantized DNNs in real-world applications, yet it remains largely unexplored within the current research community.

In this paper, the reliability of various commonly-used PTQ methods are deeply investigated and comprehensively evaluated for the first time. To this end, we first perform extensive experiments and observe that some specific test categories suffer significant performance drop after PTQ. In practical applications, we observed that several factors affect the per-

[*]Equal contribution [1]Houmo AI [2]Huazhong University of Science and Technology [3]Tencent AI Lab [4]Peking University. Correspondence to: Xinggang Wang <xgwang@hust.edu.cn>, Bingzhe Wu <bingzhewu@tencent.com>.

2^{nd} AdvML Frontiers workshop at 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

formance of PTQ. These factors include the calibration set distribution, PTQ settings, and PTQ methods. Our specific research questions are as follows:

- How does the variation in the distribution of the calibration data affect the performance of quantized network?
- How do different quantization settings affect the performance among different categories?
- How do different PTQ methods affect the performance among different categories?

To answer these questions, this paper introduces a framework designed to systematically evaluate the impact of PTQ from a reliability perspective. Through this framework, we conducted experiments covering various tasks and commonly-used PTQ approaches. We provide quantitative results and a comprehensive analysis of the underlying factors for each research question. The primary observations derived from our experiments are summarized as follows:

- Significant accuracy drop of individual categories is observed on the quantized model, which indicates the reliability issue of current PTQ methods. (Sec 3.2)
- The average prediction accuracy remains resilient to variations in the composition of the calibration dataset, in the presence of noise and intra-class bias. However, the prediction accuracy for individual categories displays sensitivity to these factors. (Sec 4.1)
- Reducing the bit-width has the potential to lead to substantial degradation in specific categories. (Sec 4.2)
- Certain optimization algorithms, such as gradient-based PTQ, exhibit comparatively diminished reliability, despite attaining superior average predictive accuracy. (Sec 4.3)

From the above observations, existing mainstream methods are often unreliable, as they can result in significant accuracy drops for certain categories or groups, which is unacceptable for risk-sensitive scenarios. So, it is crucial to develop reliable and robust PTQ methods that can effectively handle distribution shift scenarios. Additionally, we contribute by creating a benchmark for PTQ reliability, covering different tasks, PTQ methods, and network architectures.

2. Related Work

2.1. Quantization

Quantization is a highly effective technique for compressing neural networks, speeding up network inference, and reducing memory cost by reducing the precision of weights and activations, often storing them as integers like INT4 (Deng et al., 2020; Nagel et al., 2021; Yuan et al., 2022). Lower precision calculations can be performed faster, resulting in faster inference times. Two primary quantization methods are Post-Training Quantization (PTQ) and Quantization Aware Training (QAT). PTQ is a quantization technique applied to a pre-trained neural network, quantizing the weights and activations after the training process (Migacz, 2017; Banner et al., 2019). While QAT integrates the quantization process into the training process (Krishnamoorthi, 2018; Choi et al., 2018; Esser et al., 2020). It involves training the network with a combination of full-precision and lower-precision weights and activations, enhancing its resilience to quantization effects.

Some previous work has explored the stability of quantization and the influence of calibration dataset. (Hubara et al., 2021) explore the problems of using small calibration dataset in quantization. PD-Quant (Liu et al., 2022) proposes a technique for adjusting the calibration activations accordingly. SelectQ (Zhang et al., 2022) shows that randomly selecting data for calibration in PTQ can result in performance instability and degradation due to activation distribution mismatch. Existing research has been limited in scope, focusing on a narrow range of factors and their effect on quantization stability. This paper presents a comprehensive framework to evaluate PTQ reliability and conducts an extensive analysis of various factors that influence quantization reliability.

2.2. Reliability of Neural Network

The reliability of deep models is often measured through different dimensions: (1) model performance in various situations (including performance on the existing worst categories, noise samples, and out-of-distribution samples, etc.), (2) model robustness against test-time attacks such as adversarial attacks, and (3) the quality of the model's confidence. This paper mainly focus on the first dimension and leaves other two as the future work.

Model reliability in the worst case: For the first dimension, various studies have been conducted to evaluate the reliability of models against worst-case scenarios involving distribution shift and data noise, among other factors(Shen et al., 2021; Han et al., 2020). In this context, the evaluation metric typically used is the worst-case accuracy among existing test categories or out-of-distribution (OOD) samples(Shen et al., 2021; Rahimian & Mehrotra, 2019). It has been observed that models trained conventionally with the assumption of independent and identically distributed (IID) data fail to generalize in real-world testing environments with challenges such as hardness, noise, or OOD samples(Han et al., 2018; Duchi & Namkoong, 2018). Further literature on this subject can be found in prior surveys(Shen et al., 2021; Han et al., 2020; Rahimian & Mehrotra, 2019). Despite the efforts made by the community, the model reliability of PTQ methods in such cases remains largely unexplored.

3. Exploring Reliability of PTQ

In this section, we will initially introduce the generic workflow of the Post-Training Quantization (PTQ). Subsequently, we will define the PTQ reliability and propose an approach to examine the reliability of PTQ.

3.1. PTQ Workflow

The workflow of PTQ involves three key steps, namely collecting a calibration dataset, assigning quantization settings, and determining quantization parameters. While it is worth noting that not all prior work can be neatly subsumed under these three steps, these stages are consistently found in the majority of PTQ methodologies.

Collecting Calibration Dataset: PTQ requires a calibration dataset to compute the activations of each layer, $\{X^1, ..., X^L\}$. The majority of academic papers typically obtain their calibration dataset through random sampling from the training dataset. In contrast, industrial applications involve the user collecting a specific amount of real-world data to serve as the calibration dataset.

Assigning Quantization Settings: The next step is to choose the quantization settings, which specify the bit-width and the quantization function. The quantization bit-width refers to the number of bits used to represent a numerical value in a quantized representation. The quantization function determines how the continuous values of weights and activations are mapped to discrete values, such as uniform quantization (Krishnamoorthi, 2018) and non-uniform quantization (Miyashita et al., 2016; Li et al., 2020). Taking the symmetric uniform quantization function as an example, float value x is quantized to k bits integer x_q:

$$x_q = clamp(round(\frac{x}{s}), -2^{k-1}, 2^{k-1} - 1), \qquad (1)$$

where s is the scaling factor, $clamp$ function limits the value into the range of k bit integer $[-2^{k-1}, 2^{k-1} - 1]$. For 8 bit integer, the range is [-128,127]. x_q can be de-quantized as $\hat{x} = sx_q \approx x$.

Optimize Quantization Parameters: The final step searches for the best quantization parameters to minimize the quantization error. This error is typically evaluated using a MSE (Choukroun et al., 2019) metric. The optimization can be layer-wise (Migacz, 2017; Yuan et al., 2022), block-wise (Nagel et al., 2020; Wei et al., 2022), or network-wise (Wang et al., 2022). For example, we can layer-wisely optimize the scaling factors s^l in Eq 1 by minimize the MSE of de-quantized activation $\hat{X}^l$ and original activation X^l:

$$\arg\min_{s^l_x} MSE(\hat{X}^l, X^l). \qquad (2)$$

We can use various methods, such as grid search or gradient-based methods to solve the optimization problem. The grid search method is a commonly used approach, which tests a range of candidate values for the quantization parameters and selects the parameters that minimize the quantization error.

3.2. PTQ Reliability Evaluation Method

The reliability of deep models is often measured through different dimensions, some commonly used dimensions include (1) model performance in various situations (including performance on the existing worst categories, noise samples, and out-of-distribution samples, etc.), (2) model robustness against test-time attacks such as adversarial attacks, and (3) the quality of the model's confidence. In this paper, we mainly focus on evaluating the reliability of existing PTQ methods based on the first dimension. Exploration of other dimensions will be addressed in future work.

To this end, we assess the prediction accuracy on different categories[1] to evaluate the reliability of PTQ. By analyzing prediction accuracy across multiple categories, we gain a comprehensive understanding of PTQ's reliability and its generalization capability. This approach can provide additional insights into the performance of the quantized network and help identify any potential biases or limitations that may affect its overall reliability. To demonstrate the necessity of assessing on different categories, we experiment three networks, ResNet (He et al., 2016) for CIFAR-10 (Krizhevsky et al., 2009) classification task, MobileNetV2 (Sandler et al., 2018) for ImageNet (Deng et al., 2009) classification task, and YOLOv5 (Jocher et al., 2022) for MS COCO (Lin et al., 2014) object detection task. We conducted multiple network quantization trials using different random seeds. In each trial, we randomly selected a calibration dataset from the training dataset and applied PTQ quantization. We assessed the performance of the quantized neural network based on both average accuracy and accuracy per category.

Figure 2 illustrates the accuracy drop from the original network, with the box spanning from the first quartile (Q1) to the third quartile (Q3) of the data, while a red horizontal line depicts the median. Firstly, we observe that different categories show varying sensitivity to quantization, with some maintaining prediction accuracy while others experience a significant decline. For example, large objects exhibit a much larger accuracy drop compared to small and medium objects, indicating their higher vulnerability to quantization. Secondly, we find that the variation in accuracy drop differs significantly across categories. For instance, the variation in

[1]Different categories can refer not only to different class labels in classification tasks but also to grouping of object detection bounding boxes based on their sizes, shapes, or other properties, as well as to grouping of samples based on their difficulty level.

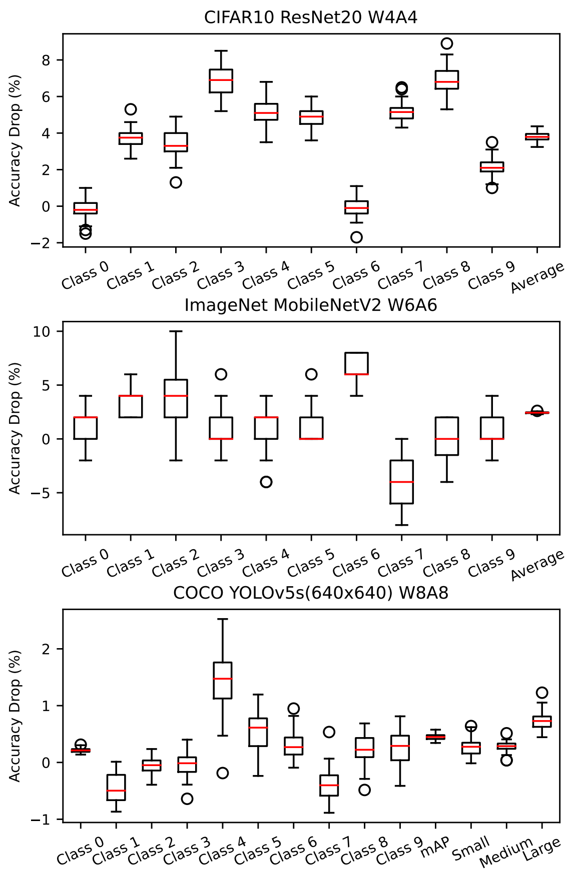


Figure 2. The box plot of accuracy drop on different classes over 50 trials with different random seeds. We only plot 10 classes of ImageNet and COCO for demonstration. "mAP" refers to mean average precision. "Small", "Medium", and "Large" refer to the precision for small, medium, and large objects.

Class 4 of COCO is much larger than that of other classes. Thirdly, we observe that the variance of accuracy drop across most individual categories is substantially higher than the average value. In conclusion, the reliability of individual categories is lower than anticipated.

Drawing on the above findings, we indicate that the low reliability of individual categories poses a risk in practical applications. In many real-world scenarios, the accuracy of predictions for specific categories may be of critical importance, and any decrease in reliability for these categories can lead to serious consequences.

4. Factors Affect PTQ Reliability

In the preceding section, we propose an approach to assess the reliability of PTQ. Subsequently, we aim to examine how different factors involved in the quantization workflow impact the reliability of PTQ. To achieve this goal, we will conduct a series of tests on each step of PTQ workflow, including calibration dataset construction, quantization settings, and optimization of quantization parameters.

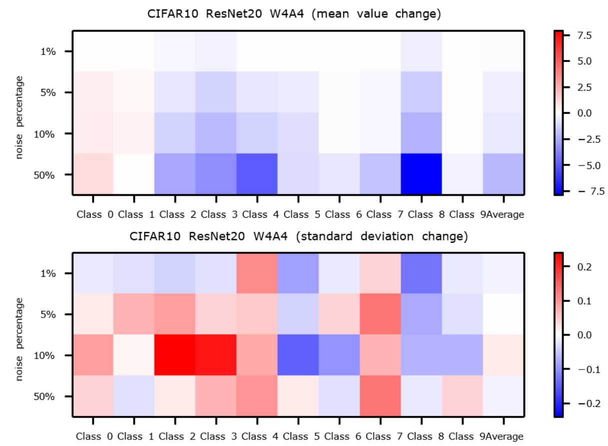


Figure 3. Influence of noisy calibration data. This figure plots the relative performance change to the clean case with varying data noise amounts. The change values are demonstrated in different colors (red means accuracy increment while blue means decrement). We execute 50 trails for each percentage of noise.

4.1. Construction of Calibration Dataset

The calibration dataset is used to estimate the distribution of the activations in the network we want to quantize. If the calibration dataset fails to capture the statistical characteristics of the real-world data, the accuracy of the quantized network may decrease, since the quantization parameters are derived based on the distribution of the activations. In this subsection, we analyze the impact of constructing a calibration dataset on the reliability of PTQ. We consider three factors that can affect the distribution of the calibration dataset: noise data, inter-class bias and dataset size.

Noise data refers to data samples that are not representative of the underlying distribution of the data. Including noise data in a calibration dataset can have a negative impact on PTQ, as the noise data can bias the distribution of activation and lead to inaccurate quantization parameters. To assess the impact of noise data, we construct the calibration dataset with some images sampled from training set and some randomly generated noise images.

Figure 3 demonstrates the prediction accuracy mean value change and standard deviation change comparing with experiments without noise on the prediction accuracy. We observe that increasing the percentage of noise data leads to a reduction on the average accuracy. The more noise data, the more average accuracy drop. However, the impact on individual classes is much larger than average accuracy. We observe that most of the classes are vulnerable to the noise data. For instance, the prediction accuracy of Class 8 of CIFAR-10 drops significantly when noise data is introduced.

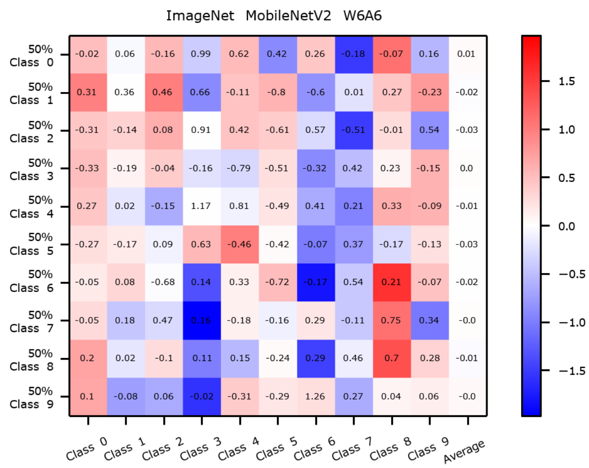


Figure 4. Influence of unbalanceed calibration datasets. The reported top-1 accuracy is averaged over 50 runs with different random seeds. The prediction accuracy change is demonstrated in different colors (red means increment and blue means decrement). The standard deviation change is annotated as text (positive means increment and negative means decrement).

While some classes, such as Class 1 and Class 9 of CIFAR-10, exhibit consistent predictive accuracy. Additionally, it is noteworthy that the noise data changes the standard deviation of individual classes, while the standard deviation on Average not changes too much.

Inter-class bias refers to distribution differences among different classes. To evaluate the impact of inter-class bias on the quantization process, we construct unbalanced calibration datasets, where the number of samples from each class is different. We build the unbalanced dataset by increasing the sampling probability of a specific class. Specifically, we constructed calibration datasets in which the sampling probability of a certain class was set to 50%, while the remaining classes were included with equal probability.

Figure 4 depicts the results obtained using unbalanced calibration datasets. The figure demonstrates that increasing the number of different classes leads to slight change on the average prediction accuracy, while significant change on the prediction accuracy of individual classes. The results also indicate that some classes are more susceptible to the unbalanced calibration dataset. For instance, the prediction accuracy of Class 6 on ImageNet significantly decreases. It is also worth noting that increasing the number of a certain class does not necessarily improve the prediction accuracy on this class. In addition, we note that the standard deviation of the average prediction accuracy almost remains unchanged, whereas the standard deviation of individual classes displays a marked variation. The distribution of various classes may vary, and an unbalanced dataset can alter the distribution of the calibration dataset. Inter-class bias can impact the selection of quantization parameters, leading to variations in prediction accuracy. However, the average prediction accuracy remains relatively stable, indicating its robustness to changes in quantization parameters. Conversely, significant variations in the prediction accuracy of individual classes suggest their vulnerability to perturbations caused by changes in the calibration dataset.

Table 1. The influence of different numbers of calibration samples. We report the mean±std over 50 runs.

Dataset Size	1	32	256	1024
Imagenet MobilenetV2 W6A6				
Average	70.0±0.38	70.2±0.06	70.2±0.05	70.2±0.07
Class 0	82.8±1.44	83.1±1.34	83.3±1.48	82.8±1.13
Class 1	68.8±2.22	70.4±1.34	70.2±1.40	70.4±1.15
Class 2	72.8±2.71	74.4±2.58	74.6±2.31	73.8±2.77
Class 3	74.6±2.33	73.6±2.04	74.8±1.83	74.3±2.28
Class 4	61.2±2.11	60.8±1.92	61.2±1.65	60.8±1.79
Class 5	95.1±1.28	95.2±1.12	95.0±1.15	95.2±1.26
Class 6	84.4±2.53	85.3±1.90	85.4±1.79	85.2±1.21
Class 7	63.5±2.49	64.6±2.06	65.9±2.12	65.4±2.37
Class 8	80.8±1.54	80.0±1.52	80.1±1.92	80.0±1.62
Class 9	91.8±1.90	91.7±1.74	91.3±1.68	91.6±1.80

The size of calibration dataset is an important factor for PTQ. It is generally known that small datasets can yield inaccurate quantization parameters, whereas large datasets can improve quantization accuracy, leading to stable and reliable outcomes. As shown in Table 1, we observe that larger dataset size results in higher and more stable average prediction accuracy, aligning with the generally accepted understanding. As the calibration dataset size increases from 1 to 32, there is a significant reduction in the variance of prediction accuracy. For instance, when size increases from 1 to 32, the standard deviation on Class 1 of ImageNet decreases from 2.22 to 1.34.

However, we observe that increasing the dataset size beyond 32 has little effect on reducing the variance of accuracy on individual classes. Despite increasing the calibration dataset size to 1024, there may still be significant variance in the accuracy of some classes. For instance, the Class 2 of ImageNet has a standard deviation of 2.77 on prediction accuracy. Some classes may inherently have more variability than others, and increasing the dataset size may not necessarily reduce this variability. Therefore, simply increasing the dataset size may not be sufficient to reduce the variance.

4.2. Quantization Settings

Once the calibration dataset is collected, the next step is to select quantization settings. In this paper, we only examine the effects of bit-width on uniform quantization. Mixed-precision quantization and non-uniform quantization are expected to be the future directions of research.

We conducted experiments by setting the same bit-width for all layers in the network and tested the prediction accuracy

Table 2. The influence of quantization settings. We report mean±std over 50 runs with different random seeds. W6A6 means 6-bit weight and 6-bit activation.

Task bit-width	CIFAR-10 ResNet20 W6A6	CIFAR-10 ResNet20 W4A4	ImageNet MobilenetV2 W8A8	ImageNet MobilenetV2 W6A6
Average	90.66±0.08	88.01±0.21	72.0±0.03	70.2±0.07
Class 0	90.14±0.27	92.18±0.43	84.0±0.00	82.8±1.27
Class 1	95.77±0.19	93.22±0.45	74.1±1.44	70.6±1.13
Class 2	88.54±0.30	84.55±0.61	76.4±1.00	74.3±2.53
Class 3	82.64±0.51	76.42±0.69	77.4±0.90	74.7±1.82
Class 4	90.76±0.27	86.56±0.61	62.4±1.44	60.6±1.89
Class 5	85.51±0.33	81.40±0.67	95.9±0.39	95.2±1.38
Class 6	92.14±0.28	94.19±0.45	94.0±0.00	85.6±1.56
Class 7	93.19±0.29	89.52±0.46	62.0±0.28	65.7±2.09
Class 8	93.15±0.21	88.84±0.70	82.0±0.00	79.9±1.60
Class 9	94.82±0.21	93.23±0.43	93.6±0.77	91.4±1.56

across 50 trials. The results are demonstrated in Table 13. We observe that the bit-width have a significant impact on the performance of quantized networks. Specifically, a higher bit-width results in not only more accurate but also more stable quantized networks. For example, the mean and std of average prediction accuracy is 36.6±0.06 on W8A8, while that of W6A6 is 31.4±0.41 for YOLOv5s quantization. We also observe that the performance of individual classes varies from each other. Some classes experience a significant drop in prediction accuracy, while others do not. For example, Class 6 of ImageNet experience a decrease of more than 6% in prediction accuracy from W8A8 to W6A6, while Class 5 of ImageNet experience a decrease of less than 1%. We think this is because different classes require different dynamic ranges to achieve a certain prediction accuracy. Consequently, the accuracy of some classes may decrease significantly. It is worth noting that even a slight decrease in average prediction accuracy can result in substantial drops in both their accuracy and stability on certain categories. In order to maintain prediction accuracy and stability of quantization, it is important to exercise caution when decreasing the bit-width.

4.3. Optimization of Quantization Parameters

The final step of PTQ is to search for the optimal quantization parameters. The optimization algorithm determines how the quantization parameters are determined. In our study, we assessed four different optimization algorithms, including the grid search method (Nagel et al., 2021), as well as three gradient-based approaches[2], namely Adaround (Nagel et al., 2020), BRECQ (Li et al., 2021), and QDrop (Wei et al., 2022). The results, as shown in Table 3, demonstrate that gradient-based methods can substantially enhance the overall prediction accuracy. For instance, the average accuracy on ImageNet using QDrop is 72.1%, which

Table 3. The influence of optimization algorithm. We report mean±std over 50 runs with different random seeds.

Algorithm	Grid Search	Adaround	BRECQ	QDrop
ImageNet MobileNetV2 W6A6				
Average	70.2±0.07	72.0±0.06	72.1±0.06	72.1±0.06
Class 0	82.8±1.27	82.3±1.05	81.7±0.98	82.1±1.01
Class 1	70.6±1.13	67.2±2.77	69.2±2.27	68.8±2.33
Class 2	74.3±2.53	81.9±2.51	79.7±2.37	80.6±2.95
Class 3	74.7±1.82	76.4±2.52	77.0±2.01	78.0±2.73
Class 4	60.6±1.89	63.8±2.56	61.1±2.34	60.6±2.51
Class 5	95.2±1.38	95.2±0.99	95.4±0.98	95.4±1.09
Class 6	85.6±1.56	92.3±1.61	90.7±1.43	89.7±2.26
Class 7	65.7±2.09	64.0±1.95	62.8±1.65	63.0±1.61
Class 8	79.9±1.60	80.4±1.33	80.8±1.38	78.4±0.80
Class 9	91.4±1.56	93.3±1.42	93.5±1.25	91.0±1.66

is almost comparable to full-precision. However, our analysis reveals that the accuracy of individual classes varied considerably. For instance, using QDrop, Class 1 of ImageNet achieves a mean accuracy of only 68.6%, whereas this is 70.6% using the grid search algorithm. Additionally, our observation reveals that the standard deviations of individual classes are relatively high. Therefore, we indicate that gradient-based PTQ methods may have relatively lower reliability, despite achieving better overall prediction accuracy.

5. Benchmark of PTQ Reliability

We have developed a framework that can test the PTQ reliability of different networks according to the proposed approach in this paper. As shown in Figure 1, the framework provide a standardized approach for evaluating the reliability of PTQ by assessing both the average performance and performance on various cetegories. We have used this framework to test the PTQ reliability performance on various tasks, PTQ methods, and network architectures[3], which can serve as a benchmark. Furthermore, the framework can be extended to evaluate the reliability of PTQ on other datasets or tasks, as well as to investigate the effect of different calibration metrics or other parameters on the quantization accuracy. We hope that the proposed framework and the benchmark can contribute to the development and improvement of PTQ and other quantization methods.

6. Conclusion

In this paper, we have introduced the concept of reliability in the context of post-training quantization. We have explored the impact of various factors on the reliability of PTQ, including: calibration dataset construction, quantization settings assignment, and quantization parameter optimization. Furthermore, we have developed a benchmark for evaluating

[2]These gradient based methods will optimize the scaling factors for quantizing activation and the rounding for quantizing weight.

[3]The results are provided in supplementary materials.

the reliability of PTQ on different neural networks, which can aid future research in this area.

References

Banner, R., Nahshan, Y., and Soudry, D. Post training 4-bit quantization of convolutional networks for rapid-deployment. *Advances in Neural Information Processing Systems*, 32, 2019.

Choi, J., Wang, Z., Venkataramani, S., Chuang, P. I.-J., Srinivasan, V., and Gopalakrishnan, K. Pact: Parameterized clipping activation for quantized neural networks. *arXiv preprint arXiv:1805.06085*, 2018.

Choukroun, Y., Kravchik, E., Yang, F., and Kisilev, P. Low-bit quantization of neural networks for efficient inference. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 3009–3018. IEEE, 2019.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Li, F. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pp. 248–255. IEEE Computer Society, 2009.

Deng, L., Li, G., Han, S., Shi, L., and Xie, Y. Model compression and hardware acceleration for neural networks: A comprehensive survey. *Proceedings of the IEEE*, 108 (4):485–532, 2020.

Duchi, J. C. and Namkoong, H. Learning models with uniform performance via distributionally robust optimization. *CoRR*, abs/1810.08750, 2018. URL http://arxiv.org/abs/1810.08750.

Esser, S. K., McKinstry, J. L., Bablani, D., Appuswamy, R., and Modha, D. S. Learned step size quantization. In *International Conference on Learning Representations*, 2020.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I. W., and Sugiyama, M. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 8536–8546, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/a19744e268754fb0148b017647355b7b-Abstract.html.

Han, B., Yao, Q., Liu, T., Niu, G., Tsang, I. W., Kwok, J. T., and Sugiyama, M. A survey of label-noise representation learning: Past, present and future. *CoRR*, abs/2011.04406, 2020. URL https://arxiv.org/abs/2011.04406.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hubara, I., Nahshan, Y., Hanani, Y., Banner, R., and Soudry, D. Accurate post training quantization with small calibration sets. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4466–4475. PMLR, 2021.

Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., NanoCode012, Kwon, Y., Michael, K., TaoXie, Fang, J., imyhxy, Lorna, Yifu, Z., Wong, C., V, A., Montes, D., Wang, Z., Fati, C., Nadar, J., Laughing, UnglvKitDe, Sonck, V., tkianai, yxNONG, Skalski, P., Hogan, A., Nair, D., Strobel, M., and Jain, M. ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation, November 2022. URL https://doi.org/10.5281/zenodo.7347926.

Krishnamoorthi, R. Quantizing deep convolutional networks for efficient inference: A whitepaper. *arXiv preprint arXiv:1806.08342*, 2018.

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

Li, Y., Dong, X., and Wang, W. Additive powers-of-two quantization: An efficient non-uniform discretization for neural networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Li, Y., Gong, R., Tan, X., Yang, Y., Hu, P., Zhang, Q., Yu, F., Wang, W., and Gu, S. Brecq: Pushing the limit of post-training quantization by block reconstruction. In *International Conference on Learning Representations*, 2021.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Liu, J., Niu, L., Yuan, Z., Yang, D., Wang, X., and Liu, W. Pd-quant: Post-training quantization based on prediction difference metric. *arXiv preprint arXiv:2212.07048*, 2022.

Migacz, S. 8-bit inference with tensorrt. In *GPU Technology Conference*, 2017.

Miyashita, D., Lee, E. H., and Murmann, B. Convolutional neural networks using logarithmic data representation. *arXiv preprint arXiv:1603.01025*, 2016.

Muhammad, K., Ullah, A., Lloret, J., Del Ser, J., and de Albuquerque, V. H. C. Deep learning for safe autonomous driving: Current challenges and future directions. *IEEE Transactions on Intelligent Transportation Systems*, 22 (7):4316–4336, 2020.

Nagel, M., Amjad, R. A., Van Baalen, M., Louizos, C., and Blankevoort, T. Up or down? adaptive rounding for post-training quantization. In *International Conference on Machine Learning*, pp. 7197–7206. PMLR, 2020.

Nagel, M., Fournarakis, M., Amjad, R. A., Bondarenko, Y., Van Baalen, M., and Blankevoort, T. A white paper on neural network quantization. *arXiv preprint arXiv:2106.08295*, 2021.

Rahimian, H. and Mehrotra, S. Distributionally robust optimization: A review. *CoRR*, abs/1908.05659, 2019. URL http://arxiv.org/abs/1908.05659.

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.

Shen, Z., Liu, J., He, Y., Zhang, X., Xu, R., Yu, H., and Cui, P. Towards out-of-distribution generalization: A survey. *CoRR*, abs/2108.13624, 2021. URL https://arxiv.org/abs/2108.13624.

Wang, C., Zheng, D., Liu, Y., and Li, L. Leveraging interlayer dependency for post-training quantization. In *Advances in Neural Information Processing Systems*, 2022.

Wei, X., Gong, R., Li, Y., Liu, X., and Yu, F. Qdrop: Randomly dropping quantization for extremely low-bit post-training quantization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=ySQH0oDyp7.

Yuan, Z., Xue, C., Chen, Y., Wu, Q., and Sun, G. Ptq4vit: Post-training quantization for vision transformers with twin uniform quantization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII*, pp. 191–207. Springer, 2022.

Zhang, D. and Lou, S. The application research of neural network and bp algorithm in stock price pattern classification and prediction. *Future Generation Computer Systems*, 115:872–879, 2021.

Zhang, Z., Gao, Y., Fan, J., Zhao, Z., Yang, Y., and Yan, S. Selectq: Calibration data selection for post-training quantization. 2022.

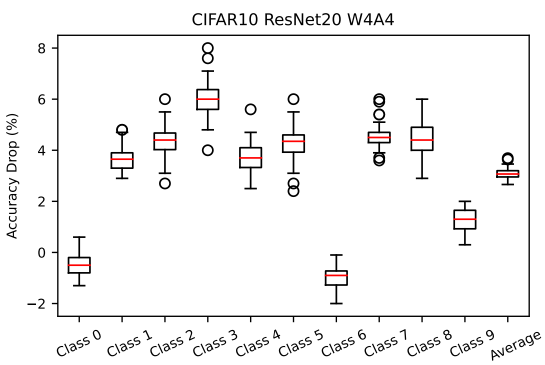
A. Quantization Models

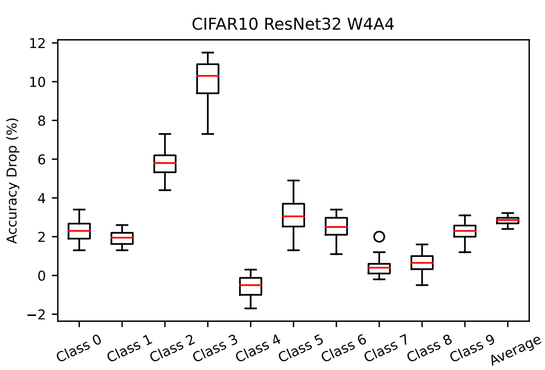
We have quantized various CNN architectures for the ImageNet, such as ResNet-18, ResNet-50, RegNetX-600m, RegNetX-3200m, MobileNetV2, and MNasNet. Additionally, for the CIFAR-10, we considered ResNet models with different depths, including ResNet-20, ResNet-32, ResNet-44, and ResNet-56. In our benchmark, we conducted thorough experiments on all the mentioned models to determine that quantization reliability issues exist in various networks.

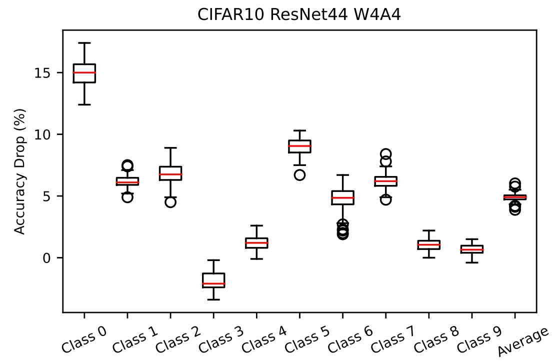
B. Benchmark

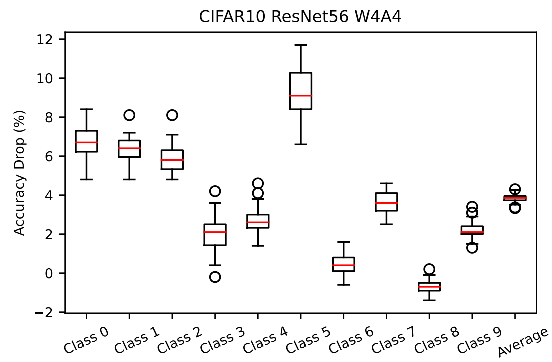
B.1. Evaluation on accuracy drop

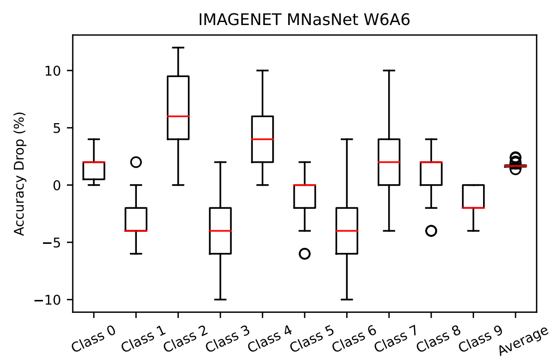
We quantize each model 50 times, selecting different random calibration data each time, and investigate the change in quantized model's accuracy for each class compared to the accuracy of the full-precision model. The box plot of accuracy drop on different classes over 50 trials with different random seeds is shown as follows. We only plot 10 classes of ImageNet.

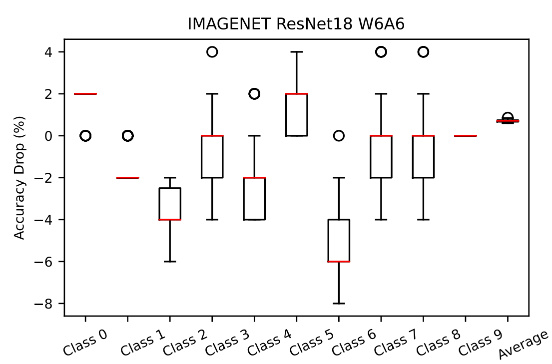


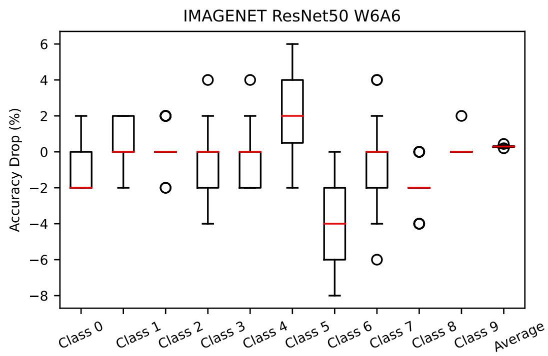


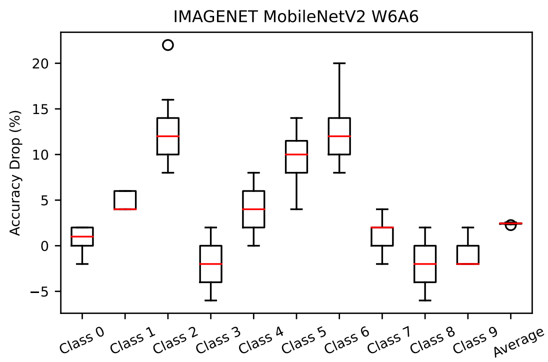


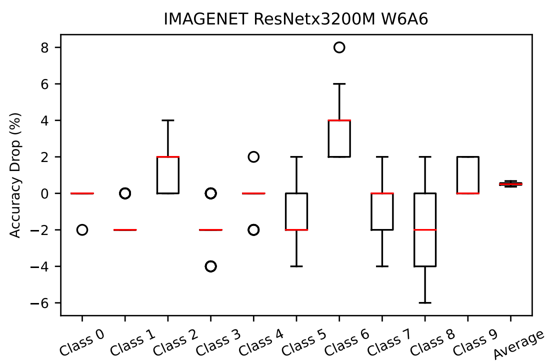


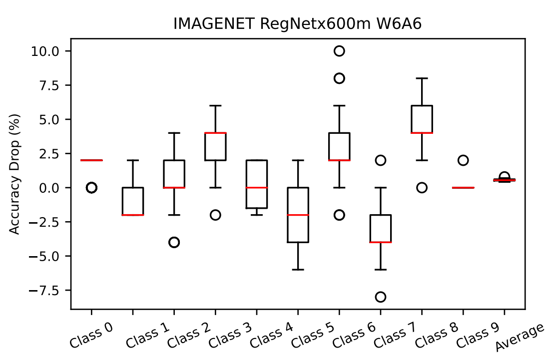












B.2. Evaluation on calibration numbers

We tested the performance of PTQ at calibration set sizes of 1, 32, 64, and 256 separately, and the experimental results of different models are as follows:

B.3. Evaluation on calibration metrics

We have studied a total of four calibration metrics, including MinMax, Cosine, KL and MSE. Results of different models on different datasets for different metrics are as follows:

MinMax calibration. Quantization scaling factors are computed based on the range directly determined by the maximum and minimum values of the feature map. The equation for calculating the scaling factors is as follows:

$$s = \frac{max(x) - min(x)}{2^n - 1}. \tag{3}$$

Cosine calibration. Quantization range is determined based on the cosine distance between features before and after quantization. Here we consider asymmetric quantization, where the minimum value after quantization is set to 0. Therefore, we only need to find the maximum value of the quantization range. Search 100 times uniformly within the range of the maximum value of the feature to find the maximum value that minimizes cosine distance. The equation is as follows:

$$\min_s D_{cos}(x, x^q). \tag{4}$$

KL calibration. Minimizing the KL divergence between the distributions before and after quantization to find the range for quantization. Specifically, the distribution is divided into a histogram of 2048 bins, and the difference between the distributions before and after quantization is compared. The equation is as follows:

$$\min_s D_{cos}(hist(x), hist(x^q)). \tag{5}$$

MSE calibration. Similar to KL calibration, the quantization range is determined based on the difference between the pre- and post-quantization distributions using MSE distance instead of KL divergence. Similarly, search 100 times to find the optimal maximum value. The equation is as follows:

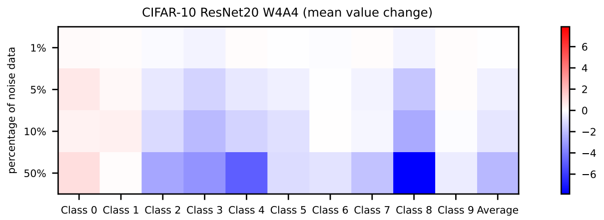
$$\min_s \|x - x^q\|^2. \tag{6}$$

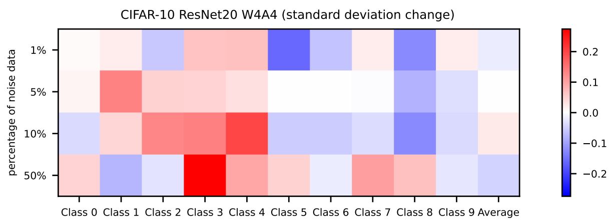
B.4. Evaluation on bitwidth

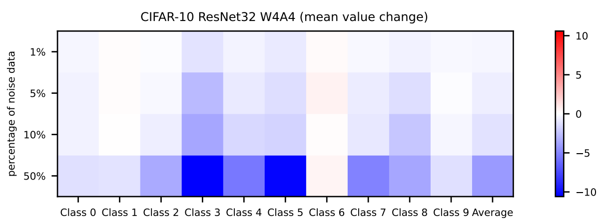
We evaluated the performance of multiple models under 6-bit and 8-bit quantization for the ImageNet, and under 4-bit and 6-bit quantization for the CIFAR-10. The experimental results are shown as follows:

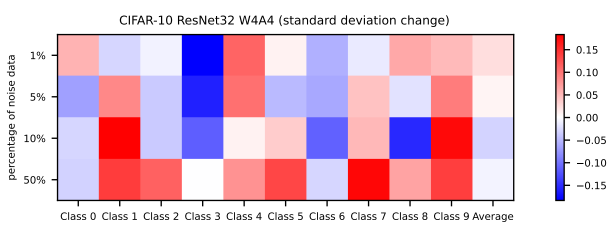
B.5. Evaluation on noise data

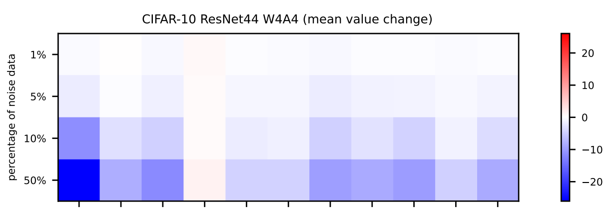
We investigated the performance of quantized models when introducing noise data similar to the actual calibration set during calibration. Specifically, we conducted experiments on the performance of quantized models when the calibration set contains 1%, 5%, 10%, and 50% noise data. The results of the experiments on multiple models are as follows, we only plot 10 classes of ImageNet. The figures below are the relative performance change to the clean case with varying datanoise amounts. The change values are demonstrated in different colors.

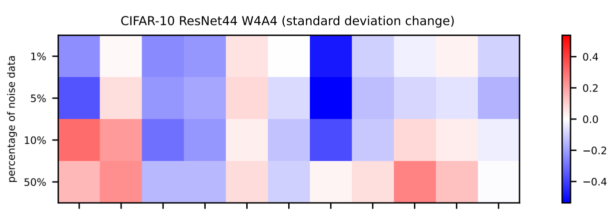


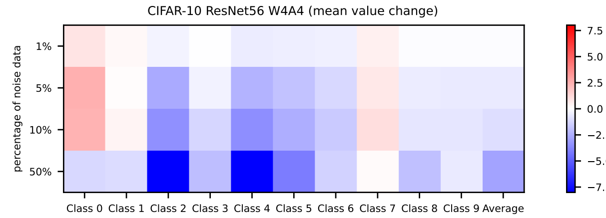


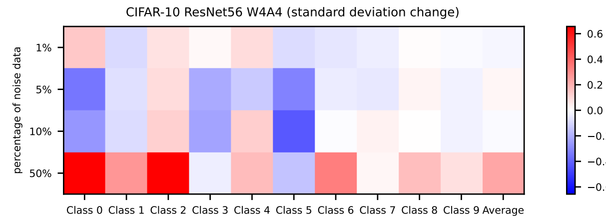


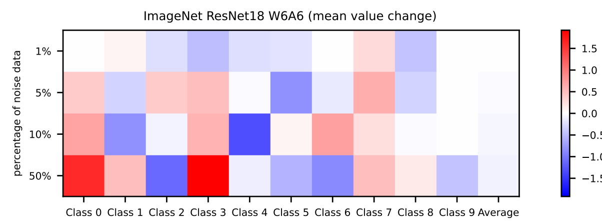


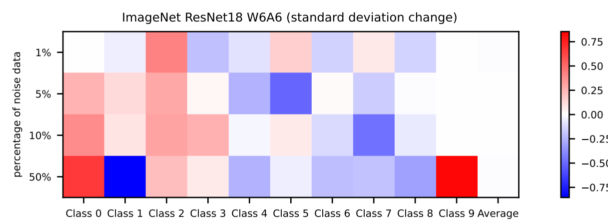


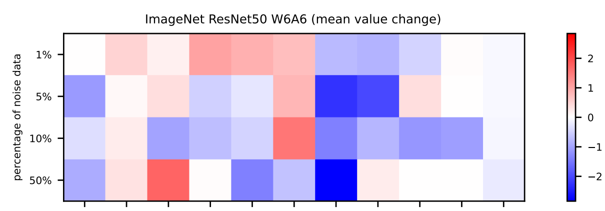


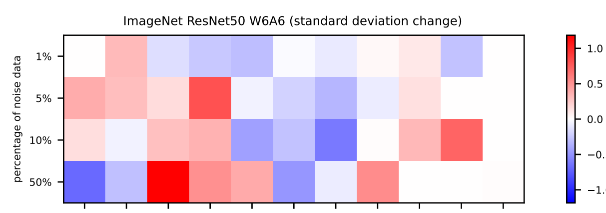


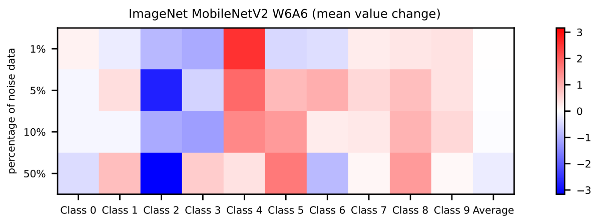


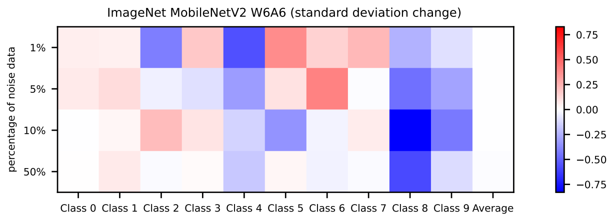


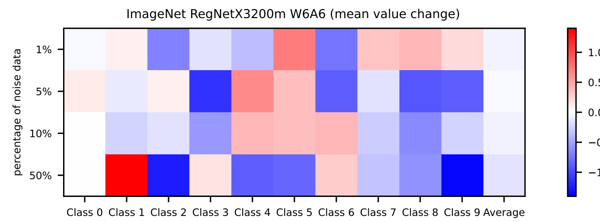


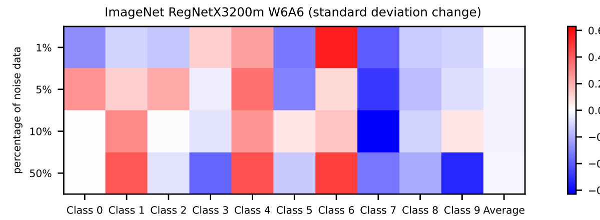


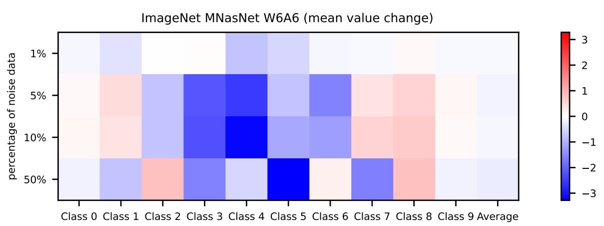


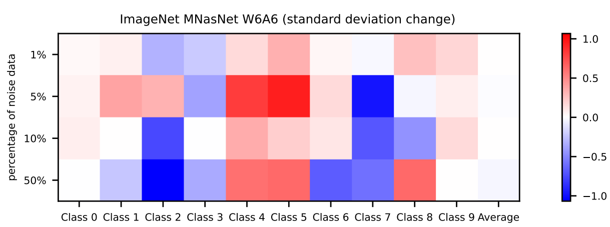


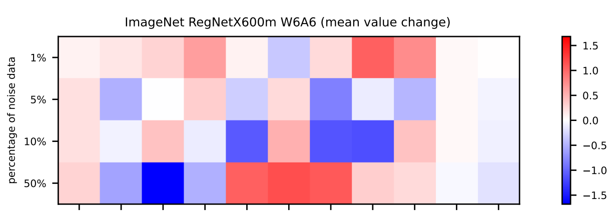


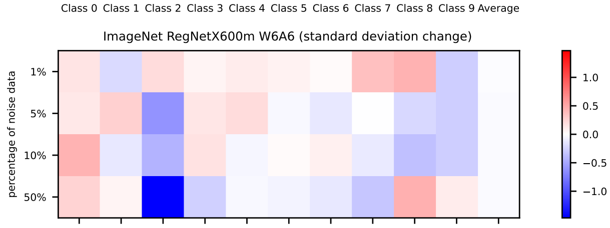






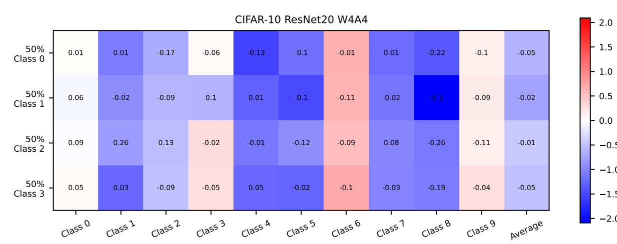


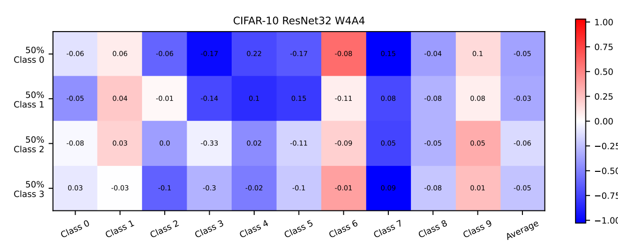


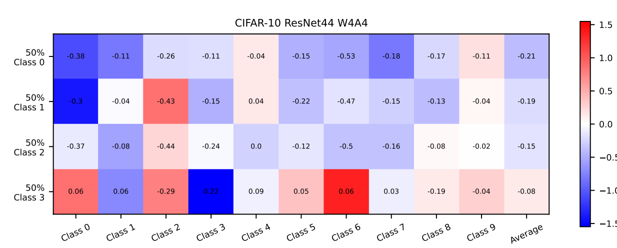


B.6. Evaluation on class bias

We conducted experiments on unbalanced calibration datasets to explore how class imbalance affects model quantization. Specifically, we set one category in the calibration dataset to represent 50% of the total samples and tested four different categories across multiple models. The experimental results of top-1 accuracy is averaged over 50 runs with different random seeds. The prediction accuracy change is demonstrated in different colors (red means increment and blue means decrement). We only plot 10 classes of ImageNet.







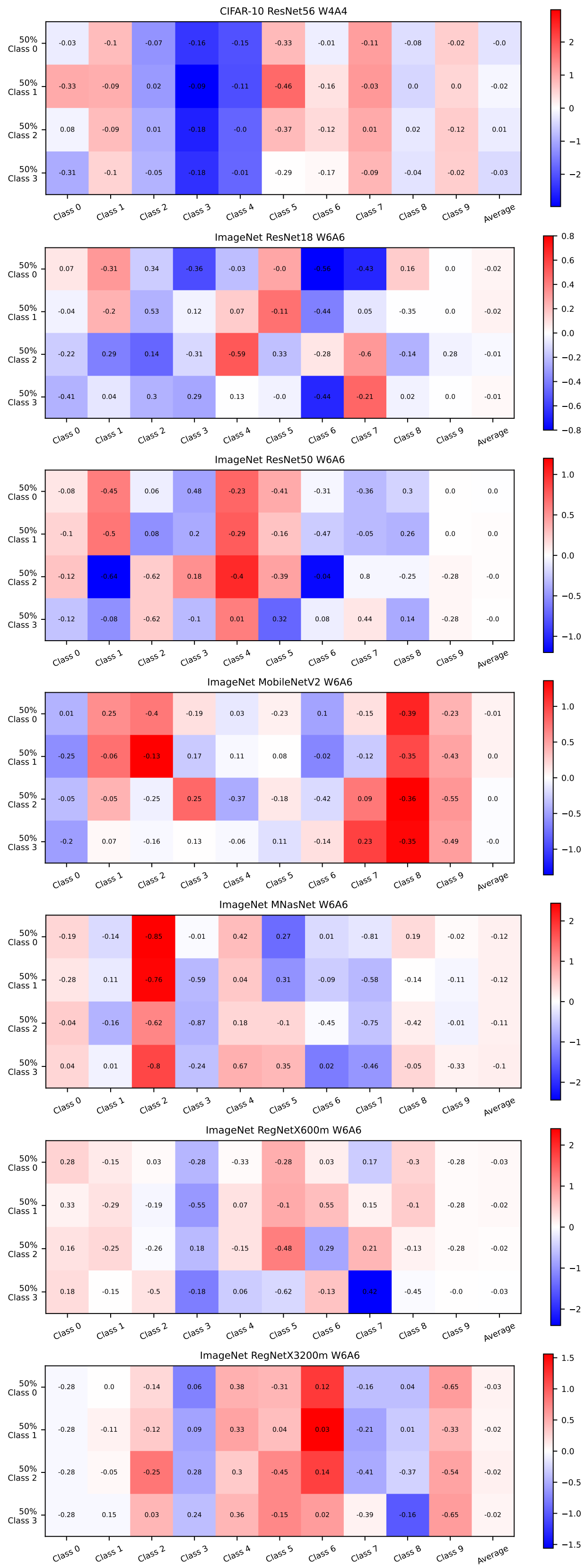


Table 4. The influence of different numbers of calibration samples. We report the mean±std over 50 runs on CIFAR-10.

Dataset Size	1	32	256	1024
CIFAR-10 ResNet20 W4A4				
Average	88.0±0.53	88.0±0.30	88.0±0.21	88.1±0.18
Class 0	92.6±0.81	92.3±0.41	92.2±0.44	92.1±0.52
Class 1	92.3±0.80	93.0±0.52	93.2±0.46	93.1±0.33
Class 2	83.2±1.09	84.4±0.75	84.6±0.62	84.7±0.57
Class 3	77.8±1.70	76.6±0.87	76.4±0.70	76.3±0.69
Class 4	86.5±1.22	86.7±0.58	86.6±0.61	86.8±0.47
Class 5	81.7±1.65	81.5±0.71	81.4±0.67	81.6±0.50
Class 6	94.1±0.50	94.2±0.42	94.2±0.45	94.1±0.45
Class 7	89.4±0.69	89.4±0.57	89.5±0.47	89.7±0.53
Class 8	88.9±1.23	88.9±0.99	88.8±0.71	88.7±0.58
Class 9	93.5±0.50	93.5±0.47	93.2±0.43	93.3±0.41
CIFAR-10 ResNet32 W4A4				
Average	87.8±0.62	87.8±0.29	87.7±0.21	87.7±0.24
Class 0	91.5±0.88	91.7±0.64	91.8±0.54	91.7±0.55
Class 1	93.5±0.71	93.8±0.39	93.8±0.36	94.0±0.46
Class 2	84.9±1.24	86.0±0.52	86.1±0.63	86.0±0.58
Class 3	69.8±2.39	68.1±1.06	67.4±1.02	67.3±0.85
Class 4	90.5±1.08	90.4±0.63	90.3±0.49	90.3±0.67
Class 5	81.7±1.80	81.4±0.86	81.4±0.77	81.1±0.60
Class 6	91.9±0.91	92.2±0.55	92.3±0.54	92.4±0.58
Class 7	88.8±0.77	89.4±0.55	89.4±0.45	89.3±0.47
Class 8	93.5±0.90	93.2±0.60	93.2±0.52	93.1±0.44
Class 9	92.0±0.67	91.5±0.50	91.5±0.40	91.5±0.38
CIFAR-10 ResNet44 W4A4				
Average	87.5±1.10	87.4±0.47	87.3±0.39	87.2±0.17
Class 0	77.5±3.07	77.6±1.63	77.4±1.13	77.1±0.79
Class 1	90.5±1.34	90.7±0.57	90.9±0.52	90.8±0.50
Class 2	84.1±2.14	84.0±1.00	83.9±0.97	83.6±0.68
Class 3	87.2±1.35	88.3±0.94	88.3±0.80	88.8±0.56
Class 4	91.7±0.88	91.2±0.61	91.2±0.55	91.3±0.51
Class 5	79.0±1.90	78.4±0.76	78.2±0.79	77.7±0.51
Class 6	89.4±2.36	88.0±1.55	87.7±1.16	87.2±0.57
Class 7	88.1±1.25	87.8±0.73	87.5±0.73	87.5±0.50
Class 8	94.0±1.16	94.1±0.59	94.2±0.52	94.2±0.38
Class 9	94.1±0.71	94.0±0.42	94.1±0.42	94.2±0.27
CIFAR-10 ResNet56 W4A4				
Average	89.3±0.80	89.4±0.19	89.4±0.19	89.3±0.16
Class 0	86.6±2.03	86.8±1.00	86.8±0.82	86.7±1.00
Class 1	90.9±1.55	91.3±0.44	91.2±0.58	91.0±0.54
Class 2	84.2±1.53	85.2±0.78	85.2±0.68	85.5±0.58
Class 3	85.1±1.53	84.6±0.96	84.8±0.84	84.8±0.71
Class 4	91.5±0.96	91.6±0.72	91.6±0.61	91.6±0.67
Class 5	78.2±1.94	78.0±1.10	77.9±1.13	77.4±0.92
Class 6	94.8±0.63	94.6±0.44	94.7±0.52	94.7±0.36
Class 7	91.7±1.33	92.3±0.49	92.1±0.51	92.0±0.46
Class 8	96.3±0.49	96.4±0.31	96.4±0.31	96.5±0.33
Class 9	93.6±0.68	93.5±0.51	93.4±0.43	93.4±0.42

Table 5. The influence of different numbers of calibration samples. We report the mean±std over 50 runs on ImageNet.

Dataset Size	1	32	256	1024
ImageNet ResNet18 W6A6				
Average	70.0±0.29	70.3±0.07	70.3±0.06	70.3±0.05
Class 0	85.0±1.08	84.2±0.65	84.3±0.69	84.3±0.73
Class 1	89.2±1.39	89.6±0.83	89.5±0.85	89.4±0.93
Class 2	83.9±0.84	83.4±0.98	83.6±1.0	83.3±1.18
Class 3	68.6±2.02	68.4±1.46	68.6±1.85	68.4±1.71
Class 4	90.8±1.44	90.4±1.73	90.0±1.85	89.9±1.7
Class 5	70.0±1.52	70.8±1.26	70.9±1.14	70.7±1.11
Class 6	79.0±1.97	79.3±1.48	79.2±1.64	78.8±1.39
Class 7	69.4±1.39	68.5±1.9	68.5±2.26	68.9±1.97
Class 8	86.4±2.34	87.9±1.72	88.1±1.76	88.5±1.54
Class 9	98.5±1.17	99.9±0.39	100.0±0.0	100.0±0.0
ImageNet ResNet50 W6A6				
Average	76.2±0.3	76.3±0.05	76.3±0.04	76.3±0.05
Class 0	94.2±0.86	94.7±1.04	95.0±1.08	95.0±1.15
Class 1	89.8±2.21	93.1±1.15	93.2±1.12	93.5±1.17
Class 2	81.7±1.22	81.6±1.06	81.7±0.9	81.7±0.9
Class 3	83.8±1.77	83.8±1.94	84.2±1.51	84.3±1.5
Class 4	87.0±1.08	86.5±1.24	86.3±1.44	86.0±1.41
Class 5	75.1±2.27	75.8±2.29	75.6±2.37	75.2±1.96
Class 6	81.5±1.91	81.8±1.53	82.0±1.83	81.5±1.73
Class 7	71.8±2.71	72.4±2.36	72.8±2.0	73.1±2.2
Class 8	84.3±1.06	83.6±1.08	84.0±0.94	83.2±1.32
Class 9	99.9±0.62	100.0±0.28	100.0±0.28	100.0±0.28
ImageNet MobileNetV2 W6A6				
Average	70.0±0.37	70.2±0.06	70.2±0.06	70.2±0.05
Class 0	90.5±1.1	91.0±1.08	91.0±1.08	91.0±1.08
Class 1	83.2±2.91	85.2±1.13	85.3±0.95	85.3±1.03
Class 2	74.4±2.34	74.2±2.46	74.0±2.38	73.3±2.18
Class 3	78.2±1.78	78.5±1.82	78.0±1.83	78.1±1.76
Class 4	79.0±1.8	78.1±1.92	78.2±1.87	77.9±1.92
Class 5	67.9±2.53	68.3±2.15	68.4±2.01	67.9±1.7
Class 6	69.4±2.54	69.2±2.66	69.9±2.31	69.8±2.01
Class 7	73.4±1.27	73.5±1.42	73.1±1.39	73.2±1.33
Class 8	88.1±2.43	88.5±1.59	87.9±2.09	88.6±1.67
Class 9	97.8±0.65	97.3±1.03	97.3±1.03	97.3±1.24
ImageNet MNasNet W6A6				
Average	74.9±0.2	74.8±0.22	74.8±0.19	74.8±0.14
Class 0	95.5±1.19	96.0±1.44	96.0±1.41	95.9±1.2
Class 1	91.0±1.56	91.0±1.66	91.2±1.59	91.0±1.66
Class 2	81.2±2.66	80.3±2.05	79.5±2.87	79.6±2.7
Class 3	83.1±2.86	82.6±2.77	82.4±2.95	82.1±2.5
Class 4	80.8±2.62	82.0±2.91	82.0±2.21	82.0±2.21
Class 5	79.2±2.23	78.2±2.58	79.1±2.01	79.2±2.65
Class 6	79.0±3.28	79.5±2.29	79.9±2.65	80.2±2.46
Class 7	65.4±3.5	66.8±3.73	67.7±3.5	69.2±2.53
Class 8	90.0±2.19	89.0±2.31	88.9±2.3	88.5±2.49
Class 9	97.9±1.41	97.9±1.09	97.6±1.13	97.5±1.1

Table 6. The influence of different numbers of calibration samples. We report the mean±std over 50 runs on ImageNet.

Dataset Size	1	32	256	1024
ImageNet RegNetx600M W6A6				
Average	72.7±0.44	73.0±0.07	73.0±0.08	73.0±0.06
Class 0	90.4±1.4	90.7±1.04	90.4±0.77	90.3±0.93
Class 1	89.3±1.24	89.4±1.16	89.3±1.11	89.7±0.84
Class 2	75.2±1.49	75.9±2.09	75.8±1.86	76.6±1.57
Class 3	76.0±1.79	77.1±1.61	77.0±1.84	77.1±1.84
Class 4	90.0±1.79	89.1±1.84	89.9±1.49	89.9±1.85
Class 5	77.1±1.92	75.6±1.89	75.8±1.99	75.3±1.54
Class 6	76.2±2.55	76.9±2.72	77.0±2.37	77.5±2.55
Class 7	73.8±2.55	75.0±2.37	75.0±2.01	76.2±1.91
Class 8	83.0±2.44	83.8±1.64	83.4±1.79	83.3±1.86
Class 9	96.0±0.28	96.0±0.0	96.0±0.28	96.0±0.0
ImageNet RegNetx3200M W6A6				
Average	77.8±0.21	77.9±0.07	77.9±0.07	77.9±0.06
Class 0	94.1±0.39	94.0±0.28	94.0±0.28	94.0±0.0
Class 1	90.9±1.07	91.8±0.6	91.8±0.65	91.9±0.39
Class 2	84.6±1.33	84.3±1.13	84.5±1.19	84.6±1.08
Class 3	84.6±1.28	85.9±0.93	85.9±1.01	86.0±0.63
Class 4	92.2±0.92	92.2±0.78	92.1±0.62	92.0±0.4
Class 5	87.7±1.7	87.3±1.94	87.6±1.63	87.5±1.69
Class 6	82.5±1.37	82.6±1.5	82.2±1.48	82.6±1.7
Class 7	72.4±1.99	72.4±1.39	72.5±1.73	73.2±1.6
Class 8	85.5±1.99	85.9±1.98	86.0±1.92	85.6±2.0
Class 9	99.4±0.98	99.5±0.88	99.4±0.93	99.3±0.95

Table 7. The influence of different metrics. We report the mean±std over 50 runs on CIFAR-10.

Dataset Size	Cosine	KL	MSE	MinMax
CIFAR-10 ResNet20 W4A4				
Average	88.3±0.14	87.1±0.28	88.0±0.21	79.4±1.18
Class 0	92.4±0.54	92.9±0.54	92.2±0.44	88.3±1.46
Class 1	92.0±0.36	94.0±0.45	93.2±0.46	88.4±1.27
Class 2	83.5±0.65	83.6±0.62	84.6±0.62	70.7±1.94
Class 3	79.2±0.51	73.8±0.85	76.4±0.7	73.2±1.78
Class 4	86.9±0.48	84.5±0.77	86.6±0.61	71.0±2.99
Class 5	81.7±0.54	81.1±0.79	81.4±0.67	72.1±1.71
Class 6	94.3±0.33	93.8±0.46	94.2±0.45	89.4±1.06
Class 7	90.1±0.48	88.5±0.57	89.5±0.47	86.3±0.7
Class 8	89.6±0.48	85.9±0.87	88.8±0.71	64.9±3.47
Class 9	93.7±0.33	92.7±0.42	93.2±0.43	90.0±0.66
CIFAR-10 ResNet32 W4A4				
Average	88.1±0.19	86.4±0.48	87.7±0.21	73.9±1.42
Class 0	92.0±0.48	90.9±1.0	91.8±0.54	85.9±2.42
Class 1	92.9±0.37	94.3±0.45	93.8±0.36	88.6±1.02
Class 2	85.2±0.58	85.3±0.85	86.1±0.63	72.2±2.55
Class 3	70.2±0.87	64.9±1.69	67.4±1.02	45.1±2.62
Class 4	91.6±0.48	86.7±1.24	90.3±0.49	72.5±2.3
Class 5	81.8±0.61	82.3±1.67	81.4±0.77	62.0±3.35
Class 6	92.1±0.45	89.0±1.16	92.3±0.54	88.8±2.23
Class 7	88.6±0.46	89.9±0.67	89.4±0.45	78.7±3.01
Class 8	94.2±0.28	90.1±1.63	93.2±0.52	63.5±4.22
Class 9	92.1±0.37	90.3±0.84	91.5±0.4	81.5±2.97
CIFAR-10 ResNet44 W4A4				
Average	87.9±0.31	85.1±0.51	87.3±0.39	62.6±5.08
Class 0	78.5±1.09	83.2±1.51	77.4±1.13	27.3±8.72
Class 1	89.8±0.48	91.2±0.87	90.9±0.52	68.2±7.1
Class 2	84.6±0.86	80.4±1.32	83.9±0.97	56.3±6.9
Class 3	87.6±0.68	87.7±0.88	88.3±0.8	88.2±1.43
Class 4	91.8±0.44	87.9±1.21	91.2±0.55	73.4±5.74
Class 5	79.6±0.77	70.0±1.79	78.2±0.79	64.8±3.74
Class 6	90.6±0.83	85.5±1.73	87.7±1.16	61.2±11.26
Class 7	88.1±0.61	83.0±1.23	87.5±0.73	56.6±6.58
Class 8	94.4±0.51	91.4±0.82	94.2±0.52	54.2±10.08
Class 9	94.2±0.41	90.3±0.88	94.1±0.42	76.2±4.44
CIFAR-10 ResNet56 W4A4				
Average	89.1±0.17	85.7±0.69	89.4±0.19	72.7±3.12
Class 0	84.9±0.54	88.5±1.09	86.8±0.82	72.6±8.02
Class 1	89.7±0.52	94.3±0.7	91.2±0.58	78.9±6.8
Class 2	84.1±0.57	84.3±1.12	85.2±0.68	59.5±6.61
Class 3	85.6±0.69	77.5±2.2	84.8±0.84	81.5±4.02
Class 4	92.0±0.48	73.7±3.08	91.6±0.61	51.5±8.77
Class 5	77.8±0.63	78.7±2.03	77.9±1.13	65.3±4.89
Class 6	94.8±0.33	90.8±1.08	94.7±0.52	80.0±4.4
Class 7	91.7±0.46	87.0±1.45	92.1±0.51	76.1±3.69
Class 8	96.9±0.24	90.4±1.36	96.4±0.31	75.4±6.05
Class 9	93.9±0.36	91.7±0.63	93.4±0.43	86.0±3.56

Table 8. The influence of different metrics. We report the mean±std over 50 runs on ImageNet.

Dataset Size	Cosine	KL	MSE	MinMax
ImageNet ResNet18 W6A6				
Average	70.2±0.05	70.3±0.04	70.3±0.06	69.8±0.15
Class 0	84.7±0.95	84.5±0.88	84.3±0.69	84.8±0.98
Class 1	89.6±0.83	89.3±0.96	89.5±0.85	89.6±1.42
Class 2	83.2±1.12	83.6±1.31	83.6±1.0	84.9±1.51
Class 3	68.2±2.33	68.9±2.05	68.6±1.85	65.8±3.65
Class 4	91.1±1.07	91.0±1.34	90.0±1.85	86.6±2.41
Class 5	70.4±1.4	69.9±0.69	70.9±1.14	70.4±1.92
Class 6	79.4±1.65	78.8±1.2	79.2±1.64	79.8±2.25
Class 7	69.4±1.34	68.8±2.15	68.5±2.26	68.4±2.2
Class 8	88.0±1.47	87.8±1.9	88.1±1.76	87.2±1.96
Class 9	98.7±0.95	100.0±0.0	100.0±0.0	99.6±0.8
ImageNet ResNet50 W6A6				
Average	76.3±0.05	76.3±0.06	76.3±0.04	75.7±0.29
Class 0	94.2±0.6	94.3±0.8	95.0±1.08	93.6±1.91
Class 1	91.4±1.0	92.9±0.99	93.2±1.12	94.8±1.14
Class 2	81.9±0.56	81.4±0.9	81.7±0.9	82.9±2.12
Class 3	84.8±1.5	84.4±1.73	84.2±1.51	83.5±3.01
Class 4	87.7±0.73	87.5±0.88	86.3±1.44	83.4±2.67
Class 5	75.6±2.15	76.9±2.23	75.6±2.37	78.2±2.22
Class 6	81.8±1.73	80.8±1.69	82.0±1.83	79.2±2.85
Class 7	70.7±2.1	70.1±2.51	72.8±2.0	74.4±2.65
Class 8	84.0±0.85	84.1±1.35	84.0±0.94	83.3±2.55
Class 9	100.0±0.0	100.0±0.28	100.0±0.28	99.1±0.99
ImageNet MobileNetV2 W6A6				
Average	70.1±0.06	69.9±0.07	70.2±0.06	69.7±0.07
Class 0	90.6±0.9	91.2±0.97	91.0±1.08	90.7±1.1
Class 1	85.2±1.14	85.2±1.05	85.3±0.95	86.6±1.22
Class 2	75.5±2.61	73.3±2.45	74.0±2.38	68.4±2.1
Class 3	78.2±1.94	78.8±1.92	78.0±1.83	79.2±2.0
Class 4	79.9±1.13	78.0±1.81	78.2±1.87	77.0±2.41
Class 5	70.0±2.4	65.9±1.76	68.4±2.01	67.5±2.82
Class 6	69.8±2.03	70.2±2.51	69.9±2.31	70.7±3.29
Class 7	73.8±1.28	73.3±1.48	73.1±1.39	73.7±1.62
Class 8	88.6±1.89	89.9±1.67	87.9±2.09	88.4±2.33
Class 9	97.8±0.65	97.7±0.69	97.3±1.03	97.4±0.93
ImageNet MNasNet W6A6				
Average	74.9±0.09	74.9±0.26	74.8±0.19	72.9±1.41
Class 0	95.0±1.08	95.9±1.35	96.0±1.41	95.9±1.62
Class 1	91.2±1.26	91.0±1.56	91.2±1.59	90.6±2.53
Class 2	82.2±2.16	81.8±2.15	79.5±2.87	74.6±5.24
Class 3	81.5±2.76	82.8±2.49	82.4±2.95	79.2±5.04
Class 4	81.4±2.33	81.0±1.84	82.0±2.21	75.8±4.79
Class 5	78.6±2.23	76.8±2.27	79.1±2.01	72.1±3.45
Class 6	80.3±2.15	82.6±1.92	79.9±2.65	77.9±3.33
Class 7	65.6±2.65	66.4±1.7	67.7±3.5	63.2±6.37
Class 8	88.5±1.72	90.1±1.83	88.9±2.3	89.2±3.92
Class 9	97.1±1.27	97.6±1.34	97.6±1.13	97.1±1.39

Table 9. The influence of different metrics. We report the mean±std over 50 runs on ImageNet.

Dataset Size	Cosine	KL	MSE	MinMax
ImageNet RegNetx600M W6A6				
Average	72.9±0.06	73.0±0.06	73.0±0.08	72.4±0.11
Class 0	90.2±0.72	90.6±1.16	90.4±0.77	90.3±1.46
Class 1	89.6±0.8	89.5±0.88	89.3±1.11	89.4±1.7
Class 2	76.2±1.59	75.8±1.54	75.8±1.86	76.0±2.56
Class 3	76.5±2.11	75.9±1.46	77.0±1.84	76.5±2.45
Class 4	90.8±1.64	90.3±1.74	89.9±1.49	88.2±2.67
Class 5	77.1±1.66	76.2±1.91	75.8±1.99	75.0±2.27
Class 6	75.6±2.63	77.4±2.31	77.0±2.37	76.7±2.67
Class 7	75.2±1.79	75.2±2.37	75.0±2.01	74.3±2.33
Class 8	83.0±1.84	83.9±1.32	83.4±1.79	82.3±2.28
Class 9	96.0±0.0	95.9±0.47	96.0±0.28	95.8±0.78
ImageNet RegNetx3200M W6A6				
Average	77.9±0.04	77.9±0.06	77.9±0.07	77.1±0.14
Class 0	94.0±0.28	94.0±0.0	94.0±0.28	95.0±1.15
Class 1	90.8±1.2	91.4±1.02	91.8±0.65	92.4±1.74
Class 2	85.1±0.99	84.3±0.98	84.5±1.19	82.0±1.65
Class 3	84.6±1.29	85.4±0.93	85.9±1.01	85.0±2.09
Class 4	92.4±0.77	92.3±1.2	92.1±0.62	91.5±1.32
Class 5	87.7±0.93	89.1±1.34	87.6±1.63	86.9±2.08
Class 6	82.9±1.34	82.9±1.45	82.2±1.48	79.2±3.73
Class 7	74.0±1.52	70.8±1.6	72.5±1.73	73.0±1.84
Class 8	84.8±1.75	85.1±1.84	86.0±1.92	85.4±1.98
Class 9	100.0±0.0	100.0±0.28	99.4±0.93	97.8±0.67

Table 10. The influence of quantization settings on CIFAR-10.

Task	CIFAR-10 ResNet20		CIFAR-10 ResNet32	
bit-width	W6A6	W4A4	W6A6	W4A4
Average	90.7±0.08	88.0±0.21	90.4±0.1	87.7±0.21
Class 0	90.1±0.27	92.2±0.44	93.5±0.3	91.8±0.54
Class 1	95.8±0.19	93.2±0.46	95.8±0.22	93.8±0.36
Class 2	88.5±0.3	84.6±0.62	90.8±0.3	86.1±0.63
Class 3	82.6±0.51	76.4±0.7	77.4±0.41	67.4±1.02
Class 4	90.8±0.27	86.6±0.61	90.0±0.35	90.3±0.49
Class 5	85.5±0.33	81.4±0.67	84.4±0.35	81.4±0.77
Class 6	92.1±0.28	94.2±0.45	94.6±0.3	92.3±0.54
Class 7	93.2±0.29	89.5±0.47	89.5±0.26	89.4±0.45
Class 8	93.1±0.21	88.8±0.71	94.3±0.29	93.2±0.52
Class 9	94.8±0.21	93.2±0.43	93.6±0.2	91.5±0.4

Table 11. The influence of quantization settings on CIFAR-10.

Task	CIFAR-10 ResNet44		CIFAR-10 ResNet56	
bit-width	W6A6	W4A4	W6A6	W4A4
Average	92.1±0.1	87.3±0.39	92.9±0.11	89.4±0.19
Class 0	92.0±0.28	77.4±1.13	93.1±0.28	86.8±0.82
Class 1	96.7±0.2	90.9±0.52	96.6±0.22	91.2±0.58
Class 2	90.8±0.31	83.9±0.97	90.6±0.39	85.2±0.68
Class 3	86.4±0.35	88.3±0.8	86.5±0.36	84.8±0.84
Class 4	92.5±0.25	91.2±0.55	94.2±0.33	91.6±0.61
Class 5	86.9±0.35	78.2±0.79	87.9±0.42	77.9±1.13
Class 6	92.0±0.36	87.7±1.16	94.8±0.21	94.7±0.52
Class 7	93.9±0.26	87.5±0.73	94.1±0.31	92.1±0.51
Class 8	95.3±0.26	94.2±0.52	95.5±0.22	96.4±0.31
Class 9	94.6±0.26	94.1±0.42	95.7±0.21	93.4±0.43

Table 12. The influence of quantization settings on ImageNet.

Task	ImageNet ResNet18		ImageNet ResNet50	
bit-width	W8A8	W6A6	W8A8	W6A6
Average	70.9±0.03	70.3±0.06	76.6±0.03	76.3±0.04
Class 0	86.0±0.0	84.3±0.69	93.4±1.47	95.0±1.08
Class 1	86.8±0.97	89.5±0.85	93.8±0.54	93.2±1.12
Class 2	78.6±1.29	83.6±1.0	82.0±0.0	81.7±0.9
Class 3	69.1±1.07	68.6±1.85	83.8±0.65	84.2±1.51
Class 4	88.0±0.0	90.0±1.85	85.0±1.08	86.3±1.44
Class 5	72.0±0.0	70.9±1.14	78.8±1.39	75.6±2.37
Class 6	74.6±1.23	79.2±1.64	75.2±1.45	82.0±1.83
Class 7	68.1±0.39	68.5±2.26	71.5±1.99	72.8±2.0
Class 8	86.8±1.05	88.1±1.76	82.9±1.21	84.0±0.94
Class 9	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.28
Task	ImageNet MobilenetV2		ImageNet MNasNet	
bit-width	W8A8	W6A6	W8A8	W6A6
Average	72.0±0.03	70.2±0.06	76.4±0.04	74.8±0.19
Class 0	91.9±0.47	91.0±1.08	97.0±1.0	96.0±1.41
Class 1	92.0±0.0	85.3±0.95	89.6±0.87	91.2±1.59
Class 2	88.8±1.07	74.0±2.38	86.9±1.39	79.5±2.87
Class 3	79.2±1.05	78.0±1.83	81.5±2.02	82.4±2.95
Class 4	83.0±1.0	78.2±1.87	84.4±1.68	82.0±2.21
Class 5	68.6±1.66	68.4±2.01	76.3±1.27	79.1±2.01
Class 6	78.8±1.27	69.9±2.31	75.9±1.67	79.9±2.65
Class 7	77.0±1.22	73.1±1.39	70.8±1.33	67.7±3.5
Class 8	86.0±0.0	87.9±2.09	90.0±0.0	88.9±2.3
Class 9	96.1±0.39	97.3±1.03	96.0±0.0	97.6±1.13

Table 13. The influence of quantization settings on ImageNet.

Task	ImageNet RegNet600		ImageNet RegNet3200	
bit-width	W8A8	W6A6	W8A8	W6A6
Average	73.5±0.04	73.0±0.08	78.5±0.03	77.9±0.07
Class 0	92.0±0.0	90.4±0.77	94.0±0.0	94.0±0.28
Class 1	88.0±0.85	89.3±1.11	90.0±0.28	91.8±0.65
Class 2	76.9±1.15	75.8±1.86	85.8±0.65	84.5±1.19
Class 3	78.1±0.84	77.0±1.84	84.0±0.28	85.9±1.01
Class 4	85.2±1.74	89.9±1.49	92.0±0.28	92.1±0.62
Class 5	75.5±0.94	75.8±1.99	85.5±1.25	87.6±1.63
Class 6	79.6±1.25	77.0±2.37	84.8±0.98	82.2±1.48
Class 7	71.7±1.55	75.0±2.01	72.9±1.0	72.5±1.73
Class 8	88.0±0.0	83.4±1.79	86.1±1.6	86.0±1.92
Class 9	96.0±0.0	96.0±0.28	100.0±0.0	99.4±0.93