Machine Intelligence Research
www.mi-research.net
DOI: 10.1007/s11633-024-1518-0

SelectQ: Calibration Data Selection for Post-Training Quantization

Zhao Zhang[1,2] Yangcheng Gao[1,2] Jicong Fan[3] Zhongqiu Zhao[1]
Yi Yang[4] Shuicheng Yan[5]

[1]School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China
[2]Key Laboratory of Knowledge Engineering with Big Data (Ministry of Education) & Intelligent Interconnected Systems, Laboratory of Anhui Province, Hefei University of Technology, Hefei 230009, China
[3]School of Data Science, The Chinese University of Hong Kong, Shenzhen 518172, China
[4]The College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China
[5]Kunlun 2050 Research & Skywork AI, China

Abstract: Post-training quantization (PTQ) can reduce the memory footprint and latency of deep model inference while still preserving the accuracy of model, with only a small unlabeled calibration set and without the retraining on full training set. To calibrate a quantized model, current PTQ methods usually randomly select some unlabeled data from the training set as calibration data. However, we show the random data selection would result in performance instability and degradation due to the activation distribution mismatch. In this paper, we attempt to solve the crucial task on appropriate calibration data selection, and propose a novel one-shot calibration data selection method termed SelectQ, which selects specific data for calibration via dynamic clustering. The setting of our SelectQ uses the statistic information of activation and performs layer-wise clustering to learn an activation distribution on training set. For that purpose, a new metric called Knowledge Distance is proposed to calculate the distances of the activation statistics to centroids. Finally, after calibration with the selected data, quantization noise can be alleviated by mitigating the distribution mismatch within activations. Extensive experiments on ImageNet dataset show that our SelectQ increases the Top-1 accuracy of ResNet18 over 15% in 4-bit quantization, compared to randomly sampled calibration data. It′s noteworthy that SelectQ does not involve both the backward propagation and batch normalization parameters, which means that it has fewer limitations in practical applications.

Keywords: Model compression, low-bit model quantization, less performance loss, one-shot dynamic clustering, calibration data selection.

Citation: Z. Zhang, Y. C. Gao, J. C. Fan, Z. Q. Zhao, Y. Yang, S. C. Yan. SelectQ: Calibration data selection for post-training quantization. *Machine Intelligence Research*. http://doi.org/10.1007/s11633-024-1518-0

1 Introduction

Deep Neural Network (DNN) models have been widely applied in various real-time scenarios for their strong learning ability, such as autonomous driving, robotics and IoT[1–4]. However, with millions of parameters and the enormous computation cost, DNN usually requires huge energy consumption for inference, which brings some limitations to its numerous applications in real world[5]. To address the problem of computation efficiency of DNN, lots of model quantization methods have been developed[6–14], which represent the parameters with low-precision values. With hardware support on efficient low-bit integer computation, the computation latency and memory footprint can be greatly reduced[15, 16].

In comparison to the resource-consuming Quantization-Aware Training (QAT)[6–8], *Post-Training Quantization (PTQ)* is a more promising way to quantize a model on-the-fly for hardware deployment. Majority of current PTQ methods[9–14] rely on a small set of unlabeled *calibration data* to appropriately represent the activation values, i.e., the output of weight layer. By feeding the calibration data into the network, the clipping activation ranges can be determined for quantization. Then, those floating-point values will be clamped to the determined range and are mapped into the quantized values during inference.

Until now, randomly selecting some data from the training set to form a calibration set is one predominant method. For instance, randomly sampling one image from each subject is the most common way to calibrate the im-

Research Article
Manuscript received on October 23, 2023; accepted on July 2, 2024
Recommended by Associate Editor Jingyi Yu
Colored figures are available in the online version at https://link.springer.com/journal/11633
This article has been accepted and is being edited now. Its final version will be assigned to and published in a specific journal issue later.

age classification model. It seems sensible, since those randomly sampled data have the identical distribution as the training set. However, in this paper we will show that the strategy of random data selection is **NOT** optimal for calibration in PTQ. In this paper, we mainly focus on selecting appropriate calibration data for PTQ, and the presented core idea is shown in Fig. 1.

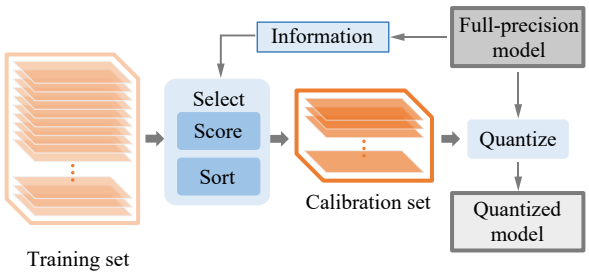


Fig. 1 The core idea of our SelectQ. We traverse the training set, utilize the statistical information within DNNs, rank the images and then sort them. Finally, we take those images with the highest scores to form the calibration set for PTQ.

First, the activation produced by randomly selecting data cannot completely cover the feature space over the whole training set during calibration. Most importantly, model performance may vary greatly based on the randomly sampled calibration data, as can be seen from Fig. 2. And more to the point, activation distribution may be different for the images from the same class, especially for large scale dataset, as shown in Fig. 3. This difference causes a mismatch of clipping range thus increasing the quantization error. This is one of the difficulties for applying quantization in practical scenario, which usually contains more distribution shifts and outlier data. To avoid this mismatch, we investigate the relationship between the activation distribution and calibration set, which gives birth to a novel idea to select appropriate calibraton data by extracting the statistical information from the original model, as shown in Fig. 1.

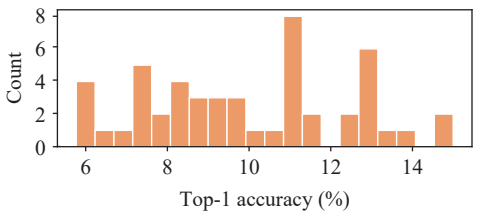


Fig. 2 Performance distribution based on the 4-bit quantized ResNet18 in terms of Top-1 accuracy, which is calibrated with 50 different calibration sets randomly sampled from the training set, where the term "Count" denotes the number that falls within the corresponding interval.

Second, current quantization methods pay more attention to the statistical properties of activation. For example, Analytical Clipping for Integer Quantization (ACIQ)[9] focuses on the statistics of activation quantization noise. ZeroQ[17] constrains the activation distribution with the batch normalization parameters. Besides, majority of current data-free quantization methods have considered the statistic alignment to generate fake data for model calibration[17, 18] or fine-tuning[19–21]. However, data-free quantization with fine-tuning requires great computation resource. In contrast, the methods with calibration are more efficient, but usually suffer from the generalization decline due to the absence of information from real data. Except for some specific cases about privacy and security, one more practical issue for PTQ is how to choose appropriate calibration data from the training set to avoid the potential performance degradation.

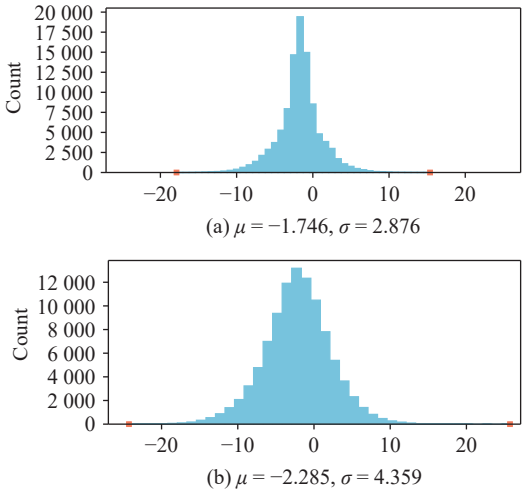


Fig. 3 Activation distributions of the 1st convolution layer in 2nd ResBlock in ResNet18 during inference on different images with the same label "tench". The horizontal axis indicates the amount of each interval and the vertical axis denotes the activation values. The red square points denote minimal and maximal values of the activations.

In this paper, we mainly explore the activation distribution to reduce the quantization noise from a new perspective of calibration data selection. We show that randomly selected calibration set may cause distribution mismatch which brings down the performance. To tackle this issue, we adopt dynamic clustering to utilize the activation distribution for appropriate calibration data selection, as shown in Fig. 5. Overall, the main contributions of this paper are threefold:

- We propose a novel and effective one-shot quantization approach termed SelectQ for the uniform post-training quantization. Guided by the activation statistical information, our SelectQ can select appropriate calibration data in an efficient way, and can cover the feature distribution space on the whole training set. The selected calibration set leads to more appropriate clipping activation ranges and maintains the performance of the quantized model. To the best of our knowledge, this is the first work devoted to the problem of calibration data selection for PTQ. Moreover, compared to existing calibration data

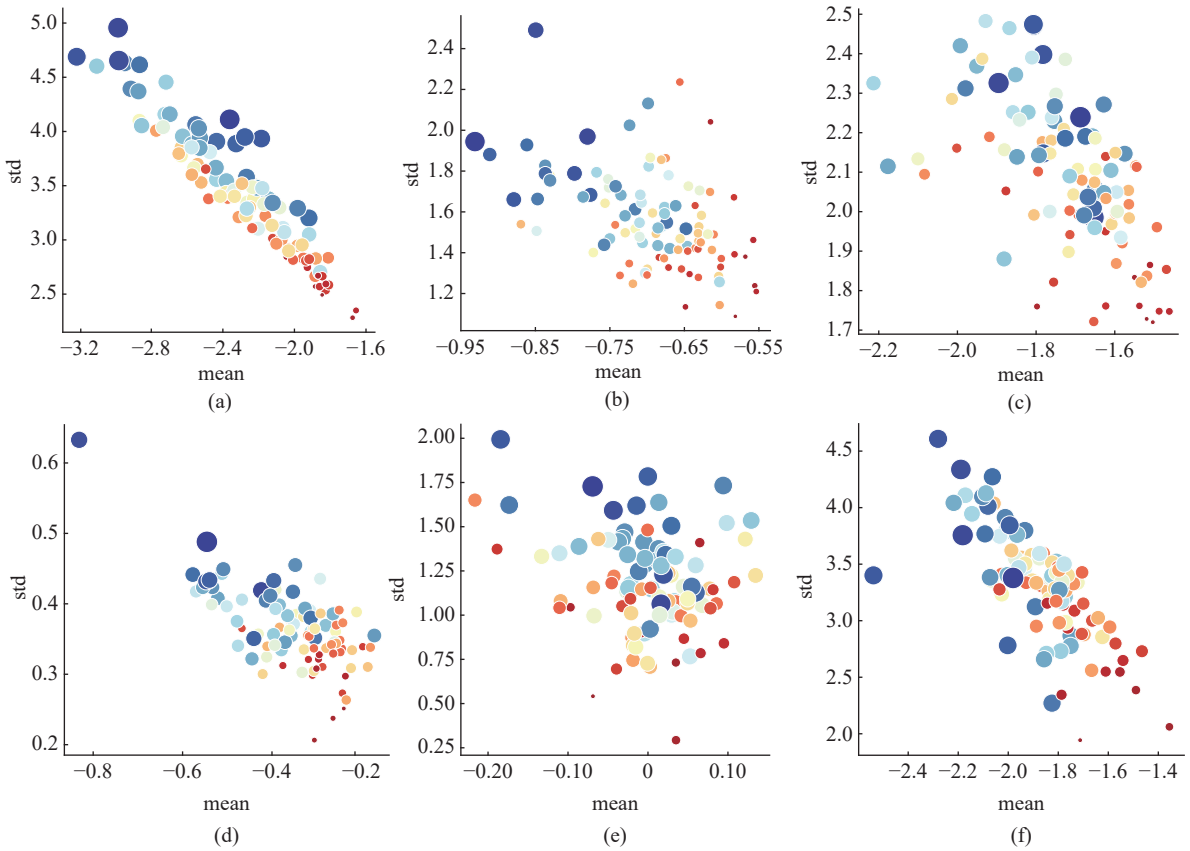


Fig. 4 Correlation between the activation distribution statistics and clipping ranges. We plot the statistic points and represent the min-max range size by the radius and cool-warm color palette. The larger and colder point is, the wider min-max range it represents.

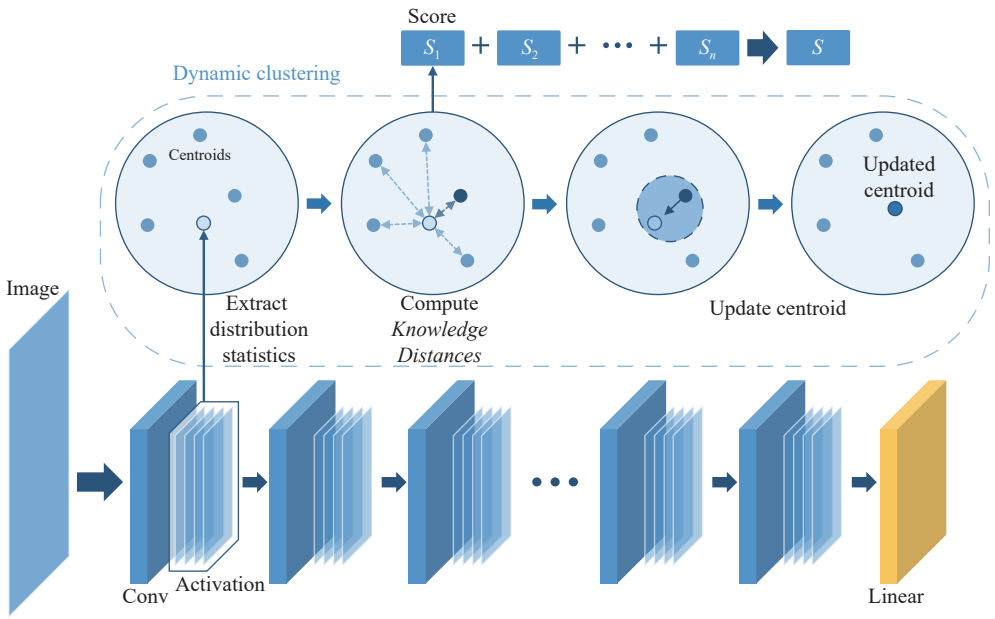


Fig. 5 The main process of our SelectQ. During inference on the image, we extract the activation distribution statistics and compute the knowledge distances from all centroids. For centroid updating, we directly shift the nearest centroid to the extracted statistics. Otherwise, we fix all of the centroids and directly sum up the shortest distances in all layers as the score for data selection.

generation methods, SelectQ does not involve the batch normalization parameters, which means fewer limitations in practical applications.

• A new metric termed Knowledge Distance is proposed to estimate the distances between activation statistics and cluster centroids. On this basis, centroids can

be learned by utilizing the activation statistics. Under this design, the whole learning process requires no backward propagation, but only a few parameters.

• Extensive quantization comparisons on several standard network architectures (such as ResNet18/50[22], SENet[23]) and compact architectures (such as MobileNetV2/V3[24, 25], ShuffleNet[26], SqueezeNet[27] and MnasNet[28]) show that our SelectQ generalizes better over them, and outperforms the random data selection method and existing generative quantization methods with calibration, i.e., ZeroQ[17] and DSG[18], based on different precision settings. For example, on ResNet-18 model, our method increases the Top-1 accuracy by over 15% in 4-bit quantization. We also evaluate SelectQ based on the widely used PTQ method ACIQ[9] for the compatibility analysis.

2 Related work

For DNN deployment on hardware, especially for those supporting low-precision computation, quantization is indispensable to improve the efficiency of model inference. Traditional quantization methods can be roughly divided into QAT and PTQ. QAT performs quantization and backward propagation alternately, and approximates the gradient by straight-through estimator[29]. By fine-tuning the weights, the performance can be well retained when quantizing in extremely low-bit cases[6–8]. However, QAT methods involve high computation cost, which limits their real applications.

PTQ is emerged for rapid deployment without backward propagation, which requires small amount of unlabeled data from original dataset to adjust the weights and determine the activation clipping ranges. For example, ACIQ[9] analytically optimizes the clipping range for trade-off between clipping error and rounding error. Outlier Channel Splitting (OCS)[10] duplicates and halves channels containing outliers. AdaRound[11] and BRECQ[13] are proposed for adaptive rounding via analysis to quantization noise. Differently, AdaQuant[14] reduces the quantization noise of each layer or block separately by parameter optimization. ZeroQ[17] is also a remarkable method which generates fake data to form the calibration set. DSG[18] reveals the homogenization of generative data for quantization. QDrop[12] drops the quantization of activations during adaptive rounding. SQuant[30] employs the constrained absolute sum of error for data-free quantization.

Although existing works have exploited the statistical properties of activation for quantization, e.g.,[17–19, 21, 31], and have demonstrated the effectiveness to investigate the relationship between data and the quantization error, our SelectQ is the first approach that utilizes the distribution information of activation to guide calibration data selection.

3 Motivation

To illustrate the necessity of the calibration on the performance of the quantized model, we conduct several quantization experiments in this section to motivate this study.

3.1 Performance fluctuation

Firstly, we explore the importance of the selected calibration data. In this study, 50 different calibration sets are randomly sampled from the ImageNet[32] training set. Following common practice, we select one image per class to form the calibration set, and then calibrate 4-bit quantized ResNet18 model. As shown in Fig. 2, the Top-1 accuracy ranges from 6% to 15% with heavy fluctuation. This evidently shows that performance of quantized model is significantly affected by the calibration set in low-bit quantization, which in turn verifies the importance of selecting appropriate calibration data. As such, we believe that there must exist effective ways to promote the quantized model performance from a new perspective of calibration data selection.

3.2 Distribution mismatch

We explore how the calibration set impacts the quantized model performance. Based on the inference on evaluation data, we observe that the activations deliver different distributions even if it runs on the images from the same class, as shown in Fig. 3, which we call distribution mismatch. According to the experimental results, we find the relationship between the distribution and the min-max range, i.e., skinny distribution usually leads to a narrow min-max range, while the fat one tends to produce a wide range.

This distribution mismatch eventually decreases the performance, especially for those quantization methods which clamp the activation range by min-max values. For example, if datum which produces outlier activation distribution is selected for calibration, the clipping range may be deviated to cause quantization noise. To handle this issue, conventional methods require more data for sufficient clipping range. But it is worth noting that larger range leads to higher rounding error, especially in low-bit quantization. This observation inspires us to exploit the activation distribution statistics for data selection to avoid the distribution mismatch in calibration.

4 Methodology

SelectQ is an efficient clustering-guided method that selects an appropriate small set of unlabeled data from training set to form the calibration set for PTQ. Specifically, it learns the cluster centroids via dynamic clustering to cover the activation distribution space. Then, it computes the distances between the fixed centroids and

running activation distribution statistics to score the data that are fed into the model. Finally, all training data are travelled and an optimized calibration set is obtained. The efficient one-shot clustering requires a few parameters without any backward propagation. We illustrate the whole process of our SelectQ in Fig. 5 and will present the details below.

4.1 Uniform quantization

To quantize tensors within a pre-trained DNN model, we firstly determine the clipping range $[\alpha, \beta]$. For activation, it's common to run on a small amount of unlabeled data and use the min/max values as the clipping range boundaries. The process of determining the activation clipping range is called calibration, and the used data in the process is termed calibration set. Then, to represent the floating points with n-bit integers, we uniformly map the origin values to the integer range $[-2^{n-1}, 2^{n-1}-1]$. The above process is referred as *quantization*, defined as

$$x_Q = \mathrm{Int}\Big(\frac{x_{FP}}{\Delta}\Big) - Z,\ \ \Delta = \frac{(\beta-\alpha)}{2^n-1}, \tag{1}$$

where x_Q and x_{FP} denote the quantized and full-precision values respectively, Δ is a scaling factor for uniform mapping and Z is an integer offset for asymmetric quantization. Note that there also exist some works on non-uniform quantization[33–35]. However, without loss of generality, we only perform our proposed SelectQ on the case of asymmetric uniform quantization, which has been widely implemented on efficient hardware devices.

During inference, activation values will be clipped by the determined clipping range. This means out-of-range values will be replaced by α or β, which leads to the so-called clipping error. On the other hand, large clipping range means that original values will be represented in low-resolution, which causes the so-called rounding error.

4.2 Knowledge distance

To update the centroids and score images, we pro-pose the metric Knowledge Distance by using acti- vation distribution information for clustering.

For each centroid C_i in i-th layer with mean μ_i^C and standard deviation σ_i^C, we simply compute the euclidean distance from the running statistics μ_i and σ_i as follows:

$$\mathcal{D}(C_i) = \left\|\mu_i^C - \mu_i\right\|_2^2 + \gamma\left\|\sigma_i^C - \sigma_i\right\|_2^2, \tag{2}$$

where γ is a hyper-parameter to weight the euclidean distances of mean and standard deviation. This distance describes the level of activation distribution coherence from the statistical characteristic of activation. With simple statistics of layer-wise activation, it avoids massive computation cost in the processes of centroid updating and data selection. Note that we have also tried the cosine similarity and channel-wise distances, which only makes tiny improvement but with much more requirements of computation resource.

4.3 Data selection

We conduct statistical analysis on the activation values to reveal the correlation between activation distribution statistics (i.e., mean and standard deviation) and clipping range. As shown in Fig. 4, cool-warm color palette and the point size indicate the min-max range during inference. Clearly, there exists large divergence among activation distribution statistics, and the range width is positively associated with the statistic values, especially the standard deviation, which is incredibly important for clipping range determination.

In addition, similar to the long-tailed distribution of activation, as shown in Fig. 3, the min-max range is more sensitive to outlier data and weaker in describing the distribution of activation. In contrast, statistics can provide better description for the distribution.

As illustrated in Fig. 5, our SelectQ travels the whole training set and applies dynamic clustering on the activation distribution statistics. Note that during the whole clustering process, batch normalization parameters are fixed from shift. The detailed procedures are presented below:

- **Initialization.** First of all, we initialize the centroids with the activation distribution statistics in each layer by feeding uniformly random data to DNN. In the ith layer, each centroid C_i consists of layer-wise mean values μ_i^C and standard deviation values σ_i^C. In addition, for centroid initialization, we have also tried to pass some of training data, which shows no significant performance improvement. One explanation is that the sampled data leads to unstable distribution. Thus, uniformly random data is used due to higher diversity and robustness, which is also utilized by SQuant[30].
- **Centroid Updating.** Then, we perform dynamic clustering to update the centroids in each layer during the training set traveling. We start to extract the activation distribution statistics, i.e., computing μ_i^C and σ_i^C. After that, the distances between obtained statistics and all centroids in the same layer are computed and sorted. As shown in Fig. 5, for the nearest centroid to the extracted statistics, we assume it is the best to represent the current distribution. And to learn from current distribution, the centroid is updated as follows:

$$\begin{cases} \mu_i^{C\prime} = \mu_i^C + \lambda_t \mu_i \\ \sigma_i^{C\prime} = \sigma_i^C + \lambda_t \sigma_i \end{cases}, \tag{3}$$

where $\mu_i^{C\prime}$ and $\sigma_i^{C\prime}$ denote the updated statistics within centroid, and λ_t is the updating step for dynamic

clustering. To adjust the centroid shift during updating, we perform cosine annealing on it, as shown below:

$$\lambda_t = \lambda_{min} + \frac{1}{2}(\lambda_{max} - \lambda_{min})\left(1 + \cos\left(\frac{t}{t_{max}}\pi\right)\right), \quad (4)$$

where t denotes the current batch number, t_{max} is the total amount of batches, and λ_{min} and λ_{max} denote the updating step of the beginning and the end respectively. To save computation resource, SelectQ is one-shot which means λ_t changes during the whole epoch.

During the centroid updating phase, an activation distribution space is constructed discretely by the statistics. Multiple centroids can effectively provide activation distribution diversity which has been revealed by DSG[18]. Different from DSG which utilizes slack distribution alignment, our SelectQ is independent on the batch normalization layer, which means that our method can be applied with less limitations.

And more to the point, centroid updating can alleviate the negative impact caused by distribution mismatch. Since the representative centroids are learned via shifting with activation distribution statistics, thus appropriate clipping ranges can be determined eventually.

• **Data Ranking.** After updating the centroids, we fix all of the centroids and travel the training set again for calibration data selection. Each data will be ranked according to the knowledge distance. For simplicity, we compute the knowledge distance from the nearest centroid in each layer, and sum them up to obtain the final score:

$$S = \sum_{i}^{l} S_i = \sum_{i}^{l} \min_{j} \left\{ S_i^j = \mathcal{D}(C_i^j) \mid 0 \le j \le N \right\}, \quad (5)$$

where l indicates the number of activations, N denotes number of centroids in each layer, and $\mathcal{D}(\cdot)$ computes the knowledge distances by Equation 2. Subsequently, those images with the highest scores will be picked up into the calibration set. Note that we set less centroids than the calibration set size, since large amount of centroids will lead to extremely high computation requirement, especially for those deeper models.

Due to the independence of each dynamic clustering process on different activations, we separately perform the centroid updating with multi-threading for computation optimization. Without backward propagation, SelectQ can efficiently select data to form the calibration set. Finally, with the selected data, we can perform the uniform quantization with Equation 1 on DNN models.

5 Experiments

5.1 Experimental settings

We mainly evaluate each method and show the quantization results by using ImageNet dataset[32]. We start by discussing the results on ResNet18/50[22] and SENet[23], which are standard architectures widely used in industry. The results on compact architectures are subsequently presented, including MobileNetV2/V3[24, 25], ShuffleNet[26], MnasNet[28], SqueezeNet[27] and MENet[36]. We conduct quantization experiments with several precision settings, including 4-bit, 6-bit and 8-bit for both weight and activation. All the experiments are implemented with PyTorch[37] on NVIDIA GTX2080Ti and the pre-trained models are provided by PyTorchCV [1].

For implementation details, we set λ_{min} to 0.001, λ_{max} to 0.1 in equation 4 and γ to 1.0 in equation 2. For dynamic clustering, we set 10 centroids for each activation.

5.2 Comparison on imagenet dataset

Since we focus on the appropriate calibration dataset selection, we mainly compare our method to the random selection strategy under the same settings. In this study, we mainly compare our SelectQ with DSG[18] and ZeroQ[17], since they take generating calibration set into account.

The comparison results are described in Table 1, from which we see clearly that our method outperforms both the existing random selection and the calibration data generation methods of DSG and ZeroQ, especially the random selection. It should be noted that in 4-bit quantization, our SelectQ promotes the Top-1 accuracy of ResNet-18 by over 15%. Moreover, on most of light-weight architectures, our SelectQ also surpasses the random selection method, e.g., the performance on MobileNetV3 in 6-bit quantization is 8.91% higher than that of random selection. Besides, in 8-bit quantization over MnasNet with SelectQ, the Top-1 accuracy is even more than the full-precision model.

It is noteworthy that the generalization performance of our SelectQ can be demonstrated via extensive quantization experiments based on various DNN models. However, there still exists severe performance degradation in some 4-bit quantization cases, e.g., ResNet50 and MobileNet. This is because these models are unable to resist the quantization noise if without weight adjustment.

5.3 Compatibility with ACIQ

Compatibility is incredibly important for model deployment in practical applications. Different from the current methods that focus on model optimization, we pay attention to appropriate calibration data selection, i.e., SelectQ has higher compatibility than existing PTQ

[1] Computer vision models on PyTorch: https://pypi.org/project/pytorchcv/

Table 1 Results on standard models ResNet18/50 and SENet.

Model	Method	W-bit	A-bit	Top-1
ResNet18	Baseline	32	32	72.987%
	Random	4	4	10.293%
	ZeroQ	4	4	26.04%
	DSG	4	4	34.53%
	SelectQ	4	4	**36.033%**
	Random	6	6	71.260%
	ZeroQ	6	6	69.74%
	DSG	6	6	70.46%
	SelectQ	6	6	**72.258%**
	Random	8	8	72.927%
	ZeroQ	8	8	71.43%
	DSG	8	8	71.49%
	SelectQ	8	8	**72.992%**
ResNet50	Baseline	32	32	77.731%
	Random	4	4	6.805%
	SelectQ	4	4	**16.510%**
	Random	6	6	76.023%
	ZeroQ	6	6	75.56%
	DSG	6	6	76.07%
	SelectQ	6	6	**76.487%**
	Random	8	8	77.524%
	ZeroQ	8	8	77.67%
	DSG	8	8	77.68%
	SelectQ	8	8	**77.687%**
SENet	Baseline	32	32	74.318%
	Random	4	4	6.698%
	SelectQ	4	4	**9.248%**
	Random	6	6	71.607%
	SelectQ	6	6	**71.947%**
	Random	8	8	74.211%
	SelectQ	8	8	**74.218%**

methods. Toward this end, we conduct a series of experiments for compatibility analysis with ACIQ[9] due to its wide applications in practice.

Specifically, when quantizing ResNet50, MobileNetV3 and ShuffleNet, we feed calibration data from SelectQ and random selection to optimize the clipping range by Gaussian distribution, and the results are reported in Table 4. By clipping range optimization, our SelectQ still outperforms random selection method, which demonstrates its high compatibility with ACIQ. Note that performance degradation happens in some cases with ACIQ, as the quantization noise distribution may shift from the ACIQ assumption.

Table 2 Results on lightweight models MobileNetV2/V3.

Model	Method	W-bit	A-bit	Top-1
MobileNetV3	Baseline	32	32	75.346%
	Random	4	4	0.172%
	SelectQ	4	4	**0.362%**
	Random	6	6	51.131%
	SelectQ	6	6	**60.041%**
	Random	8	8	75.004%
	SelectQ	8	8	**75.048%**
MobileNetV2	Baseline	32	32	72.977%
	Random	4	4	10.869%
	SelectQ	4	4	**10.884%**
	Random	6	6	70.178%
	SelectQ	6	6	**70.252%**
	Random	8	8	72.772%
	SelectQ	8	8	**72.843%**

Table 3 Results on lightweight models ShuffleNet, SqueezeNet, MnasNet and MENet.

Model	Method	W-bit	A-bit	Top-1
ShuffleNet	Baseline	32	32	65.047%
	Random	6	6	40.230%
	ZeroQ	6	6	39.92%
	DSG	6	6	44.88%
	SelectQ	6	6	**45.254%**
	Random	8	8	64.244%
	ZeroQ	8	8	64.46%
	DSG	8	8	64.77%
	SelectQ	8	8	**64.792%**
SqueezeNet	Baseline	32	32	60.649%
	Random	6	6	52.247%
	SelectQ	6	6	**54.596%**
	Random	8	8	60.622%
	SelectQ	8	8	**60.630%**
MnasNet	Baseline	32	32	74.959%
	Random	6	6	67.727%
	SelectQ	6	6	**70.184%**
	Random	8	8	74.899%
	SelectQ	8	8	**74.962%**
MENet	Baseline	32	32	55.971%
	Random	6	6	1.161%
	SelectQ	6	6	**3.067%**
	Random	8	8	50.609%
	SelectQ	8	8	**51.011%**

Table 4 Compatibility analysis results on several deep networks.

Model	Method	W-bit	A-bit	Top-1
ResNet50	Baseline	32	32	77.731%
	ACIQ	4	4	43.784%
	SelectQ + ACIQ	4	4	**43.859%**
	ACIQ	6	6	73.737%
	SelectQ + ACIQ	6	6	**74.198%**
	ACIQ	8	8	76.353%
	SelectQ + ACIQ	8	8	**76.471%**
MobileNetV3	Baseline	32	32	75.346%
	ACIQ	6	6	72.391%
	SelectQ + ACIQ	6	6	**72.755%**
	ACIQ	8	8	75.111%
	SelectQ + ACIQ	8	8	**75.121%**
ShuffleNet	Baseline	32	32	65.047%
	ACIQ	6	6	59.865%
	SelectQ + ACIQ	6	6	**60.159%**
	ACIQ	8	8	64.536%
	SelectQ + ACIQ	8	8	**64.791%**

5.4 Parameter sensitivity analysis

We perform the hyper-parameter sensitivity analysis. We mainly study the key hyper-parameters in our SelectQ, e.g., centroid number in each layer and calibration set.

We first evaluate the performance of ResNet18 with different settings of calibration set. As shown in Fig. 6, quantizing ResNet18 with bit-width settings of 4, 6 and 8, more calibration data lead to higher performance. This is easy to understand, because more calibration data cover the activation distribution better and reduce the clipping error.

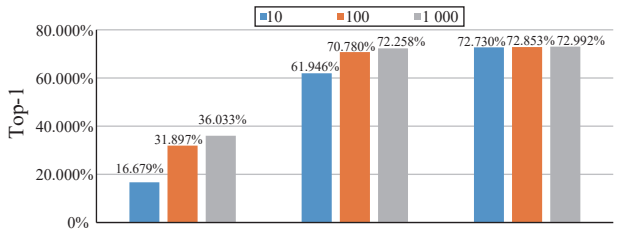


Fig. 6 Quantization results with different calibration set sizes. Color of blue, orange and gray indicates the calibration set size of 10,100 and 1 000 respectively. Left, middle and right groups are the results of 4-bit, 6-bit and 8-bit quantization respectively.

Then we evaluate the effects of different numbers of centroids. The analysis results in Fig. 7 demonstrate that more centroids used in SelectQ leads to higher performance, especially in lower bit-width cases. However, at the same time more centroids also lead to higher computation complexity in both centroid updating and data selection.

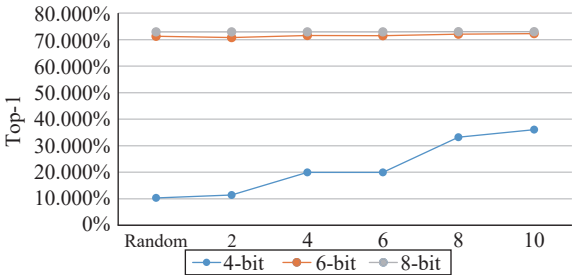


Fig. 7 Quantization results with different amounts of centroids. Horizontal axis denotes the amount of centroid used, while vertical axis indicates the Top-1 accuracy. Different colors show quantization in different bit-width settings. In addition, results of random selection are taken into consideration for comparison.

5.5 Visualization

For intuitive understanding of our SelectQ, we also visualize some activation distribution statistics in Fig. 8, where we plot the activation statistics during model inference on some image samples. From Figs 8a to 8d, we see that the learned centroids can imitate the activation distribution, thereby avoiding the distribution mismatch by filtering out the data to produce outlier distribution. From the other cases, as shown in Figs 8e to 8h, the learned centroids can well represent the activation distribution.

6 Conclusion

We have focused on exlporing the appropriate calibration data selection problem for post-training quantization, and have also proved the impact of calibration set on the quantized model performance through experimental investigation. Specifically, we discovered the distribution mismatch in activation, which potentially leads to quantization noise and error. Technically, we have proposed a new and effective calibration dataset selection method SelectQ, which learns the centroids to eliminate the activation distribution mismatch during calibration in quantization. Extensive experiments demonstrated the generalization and compatibility ablities of our SelectQ. In future, we will continuously pay attention to the calibration set selection problem for post-training quantization. In addition, quantizing a low-level vision model[38, 39] is also an interesting future direction.

Acknowledgments

The work described in this paper is partially supported by the National Natural Science Foundation of China (62072151, 62376236, 61932009), Anhui Provincial Natur-

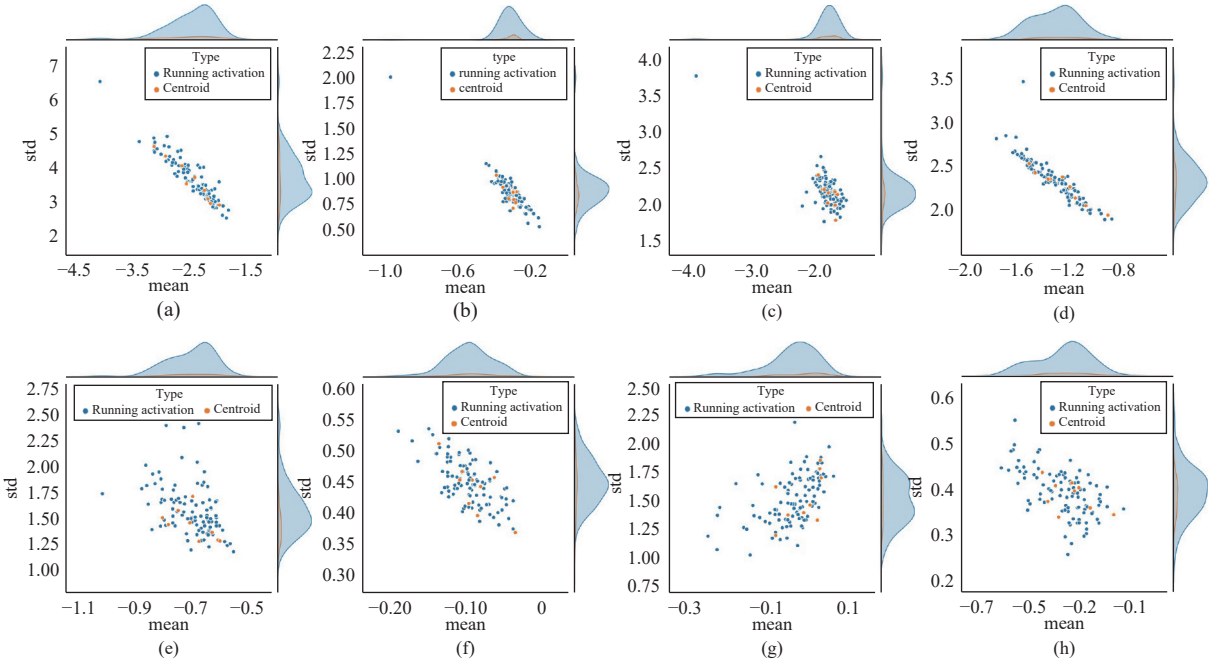


Fig. 8 Visualization results over the learned centroids and running activations. Orange points stand for learned centroids and blue points indicate running statistics. Histograms at the edges of each subfigure indicate the distribution of centroid and running activation statistics.

al Science Fund for the Distinguished Young Scholars (2008085J30), Open Foundation of Yunnan Key Laboratory of Software Engineering (2023SE103), CCF-Baidu Open Fund, CAAI-Huawei MindSpore Open Fund, Shenzhen Science and Technology Program (ZDSYS20230626091302006), and Key Project of Science and Technology of Guangxi (AB22035022-2021AB20147). Zhao Zhang is the corresponding author of this paper, and Jicong Fan is the co-corresponding author.

Declarations of conflict of interest

The authors declared that they have no conflicts of interest to this work.

References

[1] J. Philion, S. Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D. In *Proceedings of the 16th European Conference on Computer Vision*, Springer, Glasgow, UK, pp. 194–210, 2020. DOI: 10.1007/978-3-030-58568-6_12.

[2] Z. J. Liu, H. T. Tang, A. Amini, X. Y. Yang, H. Z. Mao, D. L. Rus, S. Han. BEVFusion: Multi-task multi-sensor fusion with unified bird′s-eye view representation. In *Proceedings of 2023 IEEE International Conference on Robotics and Automation*, IEEE, London, UK, pp. 2774–2781, 2023. DOI: 10.1109/ICRA48891.2023.10160968.

[3] Y. W. Li, A. W. Yu, T. J. Meng, B. Caine, J. Q. Ngiam, D. Y. Peng, J. Y. Shen, Y. F. Lu, D. Zhou, Q. V. Le, A. Yuille, M. X. Tan. DeepFusion: Lidar-camera deep fusion for multi-modal 3D object detection. In *Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, New Orleans, LA, USA, pp. 17161–17170, 2022. DOI: 10.1109/CVPR52688.2022.01667.

[4] Y. Z. Ji, H. J. Zhang, Z. Zhang, M. Liu. CNN-based encoder-decoder networks for salient object detection: A comprehensive review and recent advances. *Information Sciences*, vol. 546, pp. 835–857, 2021. DOI: 10.1016/j.ins.2020.09.003.

[5] Y. C. Gao, Z. Zhang, H. J. Zhang, M. B. Zhao, Y. Yang, M. Wang. Dictionary pair-based data-free fast deep neural network compression. In *Proceedings of 2021 IEEE International Conference on Data Mining*, IEEE, Auckland, New Zealand, pp. 121–130, 2021. DOI: 10.1109/ICDM51629.2021.00022.

[6] S. K. Esser, J. L. McKinstry, D. Bablani, R. Appuswamy, D. S. Modha. Learned step size quantization. In *Proceedings of the 8th International Conference on Learning Representations*, OpenReview.net, Addis Ababa, Ethiopia, 2020.

[7] Y. Bhalgat, J. Lee, M. Nagel, T. Blankevoort, N. Kwak. LSQ+: Improving low-bit quantization through learnable offsets and better initialization. In *Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, Seattle, WA, USA, pp. 2978–2985, 2020. DOI: 10.1109/CVPRW50498.2020.00356.

[8] J. Choi, Z. Wang, S. Venkataramani, P. I. J. Chuang, V. Srinivasan, K. Gopalakrishnan. PACT: Parameterized clipping activation for quantized neural networks. arXiv: 1805.06085, 2018. DOI: 10.48550/arXiv.1805.06085.

[9] R. Banner, Y. Nahshan, D. Soudry. Post training 4-bit quantization of convolutional networks for rapid-deployment. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Curran Associates Inc., Vancouver, BC, Canada, Article No. 714, 2019.

[10] R. Zhao, Y. W. Hu, J. Dotzel, C. De Sa, Z. R. Zhang. Improving neural network quantization without retraining

using outlier channel splitting. In *Proceedings of the 36th International Conference on Machine Learning*, PMLR, Long Beach, CA, USA, pp. 7543–7552, 2019.

[11] M. Nagel, R. A. Amjad, M. Van Baalen, C. Louizos, T. Blankevoort. Up or down? Adaptive rounding for post-training quantization. In *Proceedings of the 37th International Conference on Machine Learning*, PMLR, pp. 7197–7206, 2020.

[12] X. Y. Wei, R. H. Gong, Y. H. Li, X. L. Liu, F. W. Yu. QDrop: Randomly dropping quantization for extremely low-bit post-training quantization. In *Proceedings of the 10th International Conference on Learning Representations*, OpenReview.net, 2022.

[13] Y. H. Li, R. H. Gong, X. Tan, Y. Yang, P. Hu, Q. Zhang, F. W. Yu, W. Wang, S. Gu. BRECQ: Pushing the limit of post-training quantization by block reconstruction. In *Proceedings of the 9th International Conference on Learning Representations*, OpenReview.net, 2021.

[14] I. Hubara, Y. Nahshan, Y. Hanani, R. Banner, D. Soudry. Accurate post training quantization with small calibration sets. In *Proceedings of the 38th International Conference on Machine Learning*, PMLR, pp. 4466–4475, 2021.

[15] M. Horowitz. 1.1 computing′s energy problem (and what we can do about it). In *Proceedings of 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers*, IEEE, San Francisco, CA, USA, pp. 10–14, 2014. DOI: 10.1109/ISSCC.2014.6757323.

[16] L. Z. Lai, N. Suda, V. Chandra. CMSIS-NN: Efficient neural network kernels for arm cortex-M CPUs. arXiv: 1801.06601, 2018. DOI: 10.48550/arXiv.1801.06601.

[17] Y. H. Cai, Z. W. Yao, Z. Dong, A. Gholami, M. W. Mahoney, K. Keutzer. ZeroQ: A novel zero shot quantization framework. In *Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Seattle, WA, USA, pp. 13166–13175, 2020. DOI: 10.1109/CVPR42600.2020.01318.

[18] X. G. Zhang, H. T. Qin, Y. F. Ding, R. H. Gong, Q. H. Yan, R. S. Tao, Y. H. Li, F. W. Yu, X. L. Liu. Diversifying sample generation for accurate data-free quantization. In *Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Nashville, TN, USA, pp. 15653–15662, 2021. DOI: 10.1109/CVPR46437.2021.01540.

[19] S. K. Xu, H. K. Li, B. H. Zhuang, J. Liu, J. Z. Cao, C. R. Liang, M. K. Tan. Generative low-bitwidth data free quantization. In *Proceedings of the 16th European Conference on Computer Vision*, Springer, Glasgow, UK, pp. 1–17, 2020. DOI: 10.1007/978-3-030-58610-2_1.

[20] Y. S. Zhong, M. B. Lin, G. R. Nan, J. Z. Liu, B. C. Zhang, Y. H. Tian, R. R. Ji. IntraQ: Learning synthetic images with intra-class heterogeneity for zero-shot network quantization. In *Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, New Orleans, LA, USA, pp. 12329–12338, 2022. DOI: 10.1109/CVPR52688.2022.01202.

[21] Y. C. Gao, Z. Zhang, R. C. Hong, H. J. Zhang, J. C. Fan, S. C. Yan. Towards feature distribution alignment and diversity enhancement for data-free quantization. In *Proceedings of 2022 IEEE International Conference on Data Mining*, IEEE, Orlando, FL, USA, pp. 141–150, 2022. DOI: 10.1109/ICDM54844.2022.00024.

[22] K. M. He, X. Y. Zhang, S. Q. Ren, J. Sun. Deep residual learning for image recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, NV, USA, pp. 770–778, 2016. DOI: 10.1109/CVPR.2016.90.

[23] J. Hu, L. Shen, G. Sun. Squeeze-and-excitation networks. In *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, UT, USA, pp. 7132–7141, 2018. DOI: 10.1109/CVPR.2018.00745.

[24] M. Sandler, A. Howard, M. L. Zhu, A. Zhmoginov, L. C. Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, UT, USA, pp. 4510–4520, 2018. DOI: 10.1109/CVPR.2018.00474.

[25] A. Howard, M. Sandler, B. Chen, W. J. Wang, L. C. Chen, M. X. Tan, G. Chu, V. Vasudevan, Y. K. Zhu, R. M. Pang, H. Adam, Q. Le. Searching for mobileNetV3. In *Proceedings of 2019 IEEE/CVF International Conference on Computer Vision*, IEEE, Seoul, Korea (South), pp. 1314–1324, 2019. DOI: 10.1109/ICCV.2019.00140.

[26] X. Y. Zhang, X. Y. Zhou, M. X. Lin, J. Sun. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, UT, USA, pp. 6848–6856, 2018. DOI: 10.1109/CVPR.2018.00716.

[27] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, K. Keutzer. SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and<0.5MB model size. arXiv: 1602.07360, 2016. DOI: 10.48550/arXiv.1602.07360.

[28] M. X. Tan, B. Chen, R. M. Pang, V. Vasudevan, M. Sandler, A. Howard, Q. V. Le. MnasNet: Platform-aware neural architecture search for mobile. In *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Long Beach, CA, USA, pp. 2815–2823, 2019. DOI: 10.1109/CVPR.2019.00293.

[29] Y. Bengio, N. Léonard, A. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv: 1308.3432, 2013. DOI: 10.48550/arXiv.1308.3432.

[30] C. Guo, Y. X. Qiu, J. W. Leng, X. T. Gao, C. Zhang, Y. X. Liu, F. Yang, Y. H. Zhu, M. Y. Guo. SQuant: On-the-fly data-free quantization via diagonal hessian approximation. In *Proceedings of the 10th International Conference on Learning Representations*, OpenReview.net, 2022.

[31] Y. S. Zhong, M. B. Lin, M. Z. Chen, K. Li, Y. H. Shen, F. Chao, Y. J. Wu, R. R. Ji. Fine-grained data distribution alignment for post-training quantization. In *Proceedings of the 17th European Conference on Computer Vision*, Springer, Tel Aviv, Israel, pp. 70–86, 2022. DOI: 10.1007/978-3-031-20083-0_5.

[32] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. A. Ma, Z. H. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015. DOI: 10.1007/s11263-015-0816-y.

[33] Y. Jeon, C. Lee, E. Cho, Y. Ro. Mr.BIQ: Post-training non-uniform quantization based on minimizing the reconstruction error. In *Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, New Orleans, LA, USA, pp. 12319–12328, 2022. DOI: 10.1109/CVPR52688.2022.01201.

[34] S. Han, H. Z. Mao, W. J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv: 1510.00149, 2016. DOI: 10.48550/arXiv.1510.00149.

[35] D. Q. Zhang, J. L. Yang, D. Q. Z. Ye, G. Hua. LQ-Nets: Learned quantization for highly accurate and compact deep neural networks. In *Proceedings of the 15th European Conference on Computer Vision*, Springer, Munich, Germany, pp. 373–390, 2018. DOI: 10.1007/978-3-030-01237-3_23.

[36] Z. Qin, Z. N. Zhang, S. Q. Zhang, H. Yu, J. C. Li, Y. X. Peng. Merging and evolution: Improving convolutional neural networks for mobile applications. In *Proceedings of 2018 International Joint Conference on Neural Networks*, IEEE, Rio de Janeiro, Brazil, pp. 1–8, 2018. DOI: 10.1109/IJCNN.2018.8489496.

[37] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. M. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. J. Bai, S. Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Curran Associates Inc., Vancouver, BC, Canada, Article No. 721, 2019.

[38] Z. Zhang, H. Zheng, R. C. Hong, M. L. Xu, S. C. Yan, M. Wang. Deep color consistent network for low-light image enhancement. In *Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, New Orleans, LA, USA, pp. 1889–1898, 2022. DOI: 10.1109/CVPR52688.2022.00194.

[39] Y. Y. Wei, Z. Zhang, Y. Wang, M. L. Xu, Y. Yang, S. C. Yan, M. Wang. DerainCycleGAN: Rain attentive CycleGAN for single image deraining and rainmaking. *IEEE Transactions on Image Processing*, vol. 30, pp. 4788–4801, 2021. DOI: 10.1109/TIP.2021.3074804.

Zhao Zhang is a Full Professor at the School of Computer and Information, Hefei University of Technology, Hefei, China. He received the PhD degree from the Department of Electronic Engineering (EE) at City University of Hong Kong, supervised by Prof. Tommy W.S. Chow (IEEE Fellow), in 2013. During his PhD study, He has visited the National University of Singapore,workingwithProf.ShuichengYan(ACM/IEEE/AAAI/IAPR Fellow, and Fellow of Singapore Academy of Engineering), from Feb to May 2012; He also visited the National Laboratory of Pattern Recognition (NLPR) at Chinese Academy of Sciences, working with Prof. Cheng-Lin Liu (IEEE/IAPR Fellow, Director of NLPR), from Sep to Dec 2012.

His research interests include Machine Learning, Computer Vision and Pattern Recognition. He have authored/co-authored over 130 technical papers published at prestigious journals and conferences, including 50 IJCV or IEEE/ACM Transactions regular papers (e.g., IEEE TIP, IEEE TKDE, IEEE TNNLS, IEEE TCYB, IEEE TSP, IEEE TCSVT, IEEE TMM), and 31 Top-tier conference papers (e.g., CVPR, NeurIPS, ACM MM, ICLR, AAAI and IJCAI), with Google Scholar citations over 6,700 times and H-index 47. He is serving/served as an Associate Editor (AE) of IEEE Transactions on Image Processing (IEEE TIP), IEEE Transactions on Big Data (IEEE TBD), Pattern Recognition (PR) and Neural Networks (NN). Besides, He is serving/served as a SPC member/Area Chair for ACM MM, AAAI, IJCAI and BMVC. He is a Distinguished Member of the CCF, and a Senior Member of the IEEE.

E-mail: cszzhang@gmail.com (Corresponding author)

ORCID iD: 0000-0002-5703-7969

Yangcheng Gao is currently a staff working at Shenzhen DJ-Innovations Co., Ltd. In 2023, he graduated from Hefei University of Technology with a Master's degree in Computer Science and Technology, under the guidance of Professor Zhao Zhang. Previously he recieved a bachelor's degree in Automotive Engineering from Hefei University of Technology, Hefei, China. He focuses on the theories and algorithms for DNN model accelaration, model deployment and model efficiency.

His research interests include low-bit quantization and tensor decomposition. He has published several papers on international conferences and journal (e.g., IEEE ICDM).

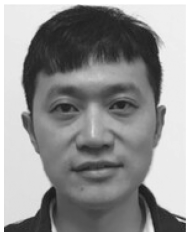
E-mail: gaoyangcheng576@gmail.com

Jicong Fan is an Assistant Professor at the School of Data Science, The Chinese University of Hong Kong, Shenzhen. He is also affiliated with Shenzhen Research Institute of Big Data, Shenzhen, China. He was a Postdoctoral Associate (advisor: Madeleine Udell) at the School of Operations Research and Information Engineering, Cornell University, Ithaca, USA. He completed his PhD at City University of Hong Kong in Electronic Engineering in 2018, under the supervision of Prof. Tommy W.S. Chow. During his PhD, he was a visiting scholar at the Department of Electrical and Computer Engineering, University of Wisconsin-Madison, USA. He was a Research Assistant at The University of Hong Kong from 2013 to 2015. He obtained his Bachelor (Automation) and Master (Control Science and Engineering, supervisor: Youqing Wang) degrees from Beijing University of Chemical Technology in 2010 and 2013 respectively. His research interests include statistical process control, signal processing, computer vision, optimization, and machine learning.

E-mail: fanjicong@cuhk.edu.cn

Zhongqiu Zhao received the Ph.D. degree in pattern recognition and intelligent system from the University of Science and Technology of China, Hefei, China, in 2007. From 2008 to 2009, he held a postdoctoral position in image processing with the CNRS UMR6168 Lab Sciences de Information et des Systé mes, La Garde, France. From 2013 to 2014, he was a Research Fellow in image processing with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. He is currently a Professor with the Hefei University of Technology, Hefei. His current research interests include pattern recognition, image processing, and computer vision.

E-mail: z.zhao@hfut.edu.cn



Yi Yang received the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 2010. He was a Professor and the Director of the ReLER Laboratory, Australian Artificial Intelligence Institute (AAII), University of Technology Sydney, Australia. He was a Postdoctoral Researcher with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. He is currently a Distinguished Professor with Zhejiang University. He is an unremunerated Adjunct Professor with the AAII, University of Technology Sydney. His current research interests include machine learning and its applications to multimedia content analysis and computer vision, such as multimedia indexing and retrieval, surveillance video analysis, and video semantics understanding.

E-mail: yi.yang@uts.edu.au

Shuicheng Yan is currently the Managing Director of Kunlun 2050 Research and Chief Scientist of Kunlun Tech - Skywork AI, and the former Group Chief Scientist of Sea Group. Till now, he has published more than 600 papers in top international journals and conferences, with H-index more than 120. His research areas include computer vision, machine learning, and multimedia analysis. His team has received winner or honorable-mention prizes for ten times of two core competitions, Pascal VOC and ImageNet (ILSVRC), which are deemed as "World Cup" in the computer vision community. Also his team won more than ten best paper or best student paper prizes and especially, a Grand Slam in ACM MM, the top conference in multimedia, including best paper award three times, best student paper award twice and best demo award once. He had been among "Thomson Reuters Highly Cited Researchers" in 2014, 2015, 2016, 2018, 2019, 2020, and 2021. He is a Fellow of Academy of Engineering, Singapore, an AAAI Fellow, a ACM Fellow, and an IAPR Fellow.

E-mail: yans@seagroup.com

Machine Intelligence Research

Citation: Z. Zhang, Y. C. Gao, J. C. Fan, Z. Q. Zhao, Y. Yang, S. C. Yan. Selectq: calibration data selection for post–training quantization. *Machine Intelligence Research*, vol.22, no.1, pp.1–12, 2025. https://doi.org/10.1007/s11633–024–1518–0

Articles may interest you

Branch convolution quantization for object detection. *Machine Intelligence Research*, vol.21, no.6, pp.1192-1200, 2024.
DOI: 10.1007/s11633-023-1434-8

A new diagnosis method with few-shot learning based on a class-rebalance strategy for scarce faults in industrial processes. *Machine Intelligence Research*, vol.20, no.4, pp.583-594, 2023.
DOI: 10.1007/s11633-022-1363-y

Pre-training in medical data: a survey. *Machine Intelligence Research*, vol.20, no.2, pp.147-179, 2023.
DOI: 10.1007/s11633-022-1382-8

Satellite integration into 5g: deep reinforcement learning for network selection. *Machine Intelligence Research*, vol.19, no.2, pp.127-137, 2022.
DOI: 10.1007/s11633-022-1326-3

Effective model compression via stage-wise pruning. *Machine Intelligence Research*, vol.20, no.6, pp.937-951, 2023.
DOI: 10.1007/s11633-022-1357-9

Multi-dimensional classification via selective feature augmentation. *Machine Intelligence Research*, vol.19, no.1, pp.38-51, 2022.
DOI: 10.1007/s11633-022-1316-5

Large-scale multi-modal pre-trained models: a comprehensive survey. *Machine Intelligence Research*, vol.20, no.4, pp.447-482, 2023.
DOI: 10.1007/s11633-022-1410-8