Enhancing Diversity for Data-free Quantization

Kai Zhao[1,*], Zhihao Zhuang[2,*], Miao Zhang[3], Chenjuan Guo[2,†], Yang Shu[2], Bin Yang[2]
[1]Aalborg University, Denmark [2]East China Normal University, China [3]HIT Shenzhen, China
kaiz@cs.aau.dk,zhuangzhihao@stu.ecnu.edu.cn,zm@hit.edu.cn,{cjguo,ys,byang}@dase.ecnu.edu.cn

Abstract

Model quantization is an effective way to compress deep neural networks and accelerate the inference time on edge devices. Existing quantization methods usually require original data for calibration during the compressing process, which may be inaccessible due to privacy issues. A common way is to generate calibration data to mimic the origin data. However, the generators in these methods have the mode collapse problem, making them unable to synthesize diverse data. To solve this problem, we leverage the information from the full-precision model and enhance both inter-class and intra-class diversity for generating better calibration data, by devising a multi-layer features mixer and normalization flow based attention. Besides, novel regulation losses are proposed to make the generator produce diverse data with more patterns from the perspective of activated feature values and for the quantized model to learn better clip ranges adaptive to our diverse calibration data. Extensive experiments show that our method achieves state-of-the-art quantization results for both Transformer and CNN architectures. In addition, we visualize the generated data to verify that our strategies can effectively handle the mode collapse issue. Our codes are available at repo.

1. Introduction

Quantization [7, 28] is an effective way to compress a deep neural network (DNN) by converting model weights which are stored as the float type of 32 or 16 bits into the integer type of 8, 4 or 3 bits. It is different from other model compression methods, such as distillation [15] where a small student model learns from labels and teacher models, and pruning [13] which cuts connections in a DNN. By quantization, we can deploy a deep model on edge devices with limited storage or resources. Meanwhile, this allows better utilization of inference devices [16] optimized for integer operations to accelerate efficiency.

Figure 1. Mode collapse examples: Qimera only generates highly similar images for a class.

Quantization can be roughly divided into two categories [29]. Quantization-aware training (QAT) quantizes the weights while training a full-precision model with the original data. Post-training quantization (PTQ) quantizes an already-trained full-precision model. Normally, PTQ needs the original training data to calibrate the quantized model for better performance, otherwise, it will suffer a great performance degradation [7].

However, due to privacy concerns, original data in real-world scenarios are often inaccessible, because data, such as medical images or human facial data, are private. Therefore, there is a need to develop data-free quantization methods [12, 32] with less performance degradation. The generator-based framework, *e.g*. Qimera [6], provides a data-free method to quantize a model without original training data. It utilizes a generator to synthesize training data and then applies these fake data to calibrate the quantized model. This framework is inspired by Generative Adversarial Net (GAN) [10] where the generator is guided by the discriminator. There are three components: generator, discriminator (*i.e*. full-precision model), and quantized model. The generator learns the random-initialized label embeddings to present the classes, and then samples random variables to generate synthetic data. Then the synthetic data are used to calibrate the quantized model. The discriminator forces the generator to synthesize the right data according to class labels. Several studies [3, 25] show that this framework achieves good results for the data-free setting, but this framework still has problems as follows.

Problem 1: GAN-based generator methods have a common mode collapse problem [33], where the generator falls short in synthesizing diverse data. The generator collapses at some data points and only synthesizes a few highly similar images for each class, as shown in Figure 1. Without

* Both authors contributed equally.
† Corresponding author.

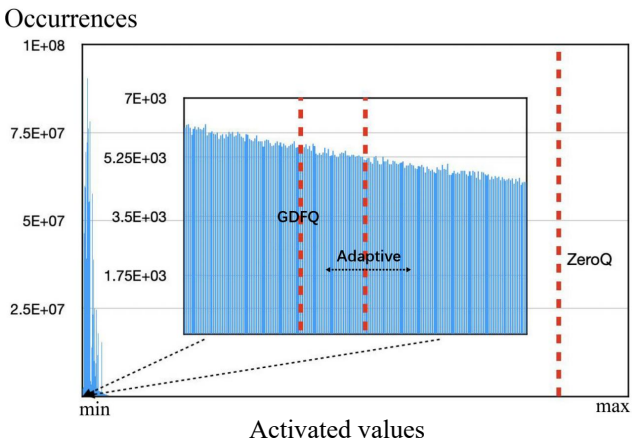


Figure 2. The long-tailed distribution, our adaptive clip ranges, and the fixed clip ranges by existing methods.

diverse data, it is difficult to accurately calibrate the quantized model, resulting in a performance drop compared to calibrating with the original data [43]. Previous works [1] on GAN have explored how to address mode collapse, but they are inapplicable to our problem as these works need original data. Recent works [5, 6, 44] propose to augment the generated images with Mixup [42] or CutMix [41] for data-free quantization. However, they could not solve the mode collapse problem, as a simple augmentation on generated similar images cannot synthesize new diverse data for each class.

Problem 2: We notice that the activated feature values of the generated data exhibit a sharp-peak and long-tailed distribution under the mode collapse problem, which is bad for quantization in the activation layers, especially when existing data-free quantization methods use a fixed clip range. For example, ZeroQ [4] quantizes all activated values within the clip range that is merely the fixed [min, max] range. Qimera and GDFQ [39] use moving averages during calibration to decide the fixed clip range. We show the fixed clip ranges set by existing methods in Figure 2, where we dispatch all activated values in the last activation layer during the last epoch into 100,000 bins with intervals of 1e-3 and calculate the frequency histogram of the occurrence of different values. ZeroQ's clip range is too large, and it cannot differ such many values using the low-bit integer. Meanwhile, the clip range set by GDFQ is small, which results in information loss as the values beyond this range are clipped.

Due to these problems, there is more performance degradation in quantization. To tackle the above problems, we propose **Enhancing** diversity for **D**ata-**F**ree **Q**uantization (EnhancingDFQ).

For **Problem 1**, we leverage the information related to the original training data from the multiple layers of the full-precision model, to generate more diverse synthetic data to calibrate the quantized model. We utilize the information from different layers of the full-precision model to guide the generator, where different layers of the full-precision model exhibit features of different levels. For example, shallow layers may focus on color, shape, or texture features *etc,* and deep layers may focus on semantic information. Thus, instead of using a random-initialized label embedding for each class, our generator can synthesize better data by utilizing multi-layer features with our attention mechanism. First, our multi-layer features mixer enables the generator to learn the relations among all classes, thereby enhancing inter-class diversity, *e.g*. an image containing information of dog and cat classes *etc*. Second, our normalization flow based attention enables the generator to focus on minutia features of different levels, thereby enhancing intra-class diversity, *e.g*. different dog images focusing on different color or texture features *etc*. Thus, we can address mode collapse and improve quantization with more diverse calibration data.

For **Problem 2**, We propose to generate diverse data with more complex patterns from the perspective of activated features, for the quantized model to learn better clip ranges. First, we propose a regulation loss for the generator, making it produce diverse data that can exhibit more complex feature patterns in the activation layers of the full-precision model. Then, we use a distillation regulation loss to align the full-precision model and the quantized model, to learn appropriate clip ranges adaptive to our generated diverse data, as shown in Figure 2.

Our contribution can be summarized as follows.

1. We utilize information from the full-precision model to enhance inter-class and intra-class diversity, with our multi-layer features mixer which learns the relations among classes, and our normalization flow based attention which focuses on the features of different levels.
2. We propose a novel regulation loss to generate diverse data with more complex feature patterns, and then learn appropriate clip ranges adaptive to our generated diverse data in the activation layers.
3. Extensive experiments show our method achieves state-of-the-art quantization results for both Transformer [9] and CNN [20] architectures, and we use visualization and ablation studies to validate our motivations and designs.

2. Related Works

2.1. Generative adversarial net

GAN is composed of a generator and a discriminator. By competing with each other, a good generator can synthesize fake but realistic data [10, 18, 33]. However, they still suffer from the mode collapse, which results in a loss of diversity in the generated data. CGAN [27] uses the guided generation methods to address inter-class diversity of generated data. WGAN [1] employs the Earth Mover's distance as a

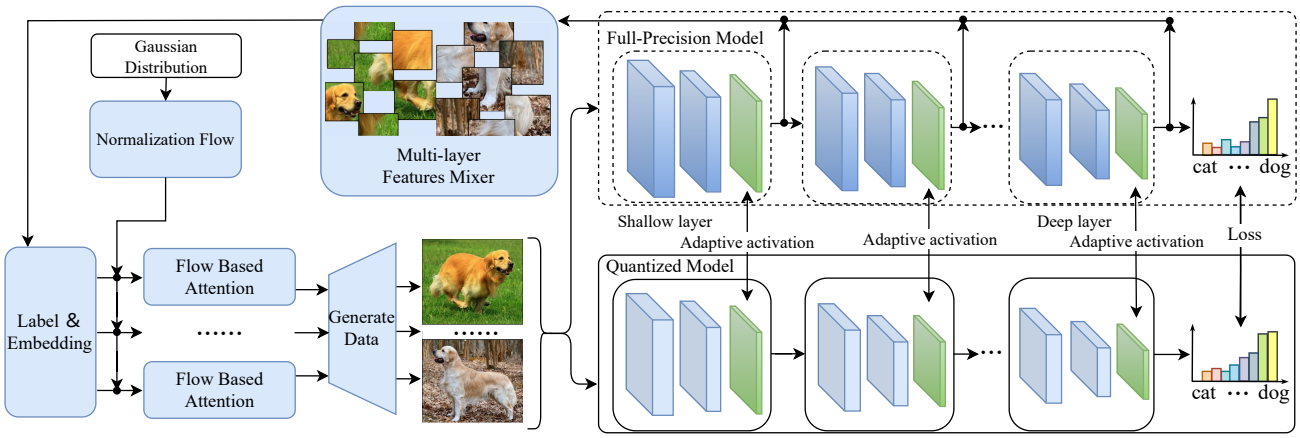


Figure 3. Overview architecture.

novel loss for optimization. StyleGAN [17] proposes to fuse representations from the generator to synthesize diverse data. However, all these methods need the original training data to address the mode collapse problem and then generate more diverse data.

2.2. Quantization

Quantization is effective in compressing a DNN into smaller sizes. For a weight w in a DNN, the quantization can be described as $\hat{w} = clip(round(\frac{w}{s} - b), lower, upper)$, where s corresponds to the scale factor, b denotes the zero point, and $clip$ is the clip operation which quantizes the weights into the $[lower, upper]$ clip range. Generator-based data-free quantization is inspired by GAN, and has attracted significant attention recently due to its ability to address privacy issues where the training data are unavailable. It generates synthetic data, which are then used to calibrate the quantized model. GDFQ [39] was the first to propose the generator-based data-free quantization. It utilizes the statistics of the batches from the CNN full-precision model to help train the generator. ARC [45] tries to design a better generator by Neural Architecture Search. DSG [43] tries to adapt batch normalization layers to diversify samples. HAST [21] proposed to generate images with larger losses, which are the hard samples that the full-precision model finds hard to classify correctly. IntraQ [44] uses local object reinforcement to conduct crop or resize to enhance the synthetic images. PSAQ-ViT [23] is the first to apply the data-free GAN-based quantization for Transformer architectures, and PSAQ-ViT V2 [24] uses additional prior information to enhance the synthetic images. CLAMP-ViT [34] uses image-patch-level contrastive learning to help quantize Vision Transformer models. AdaDFQ [31] and AdaSG [30] use the information losses to improve the quality of the synthetic images. Qimera [6] uses Mixup [42] to generate boundary data. TexQ [5] uses the Mixup for the generated images to calibrate the quantized model without cross-entropy loss, as the Mixup labels may be inaccurate. However, these methods still face the mode collapse problem and long-tailed activation problem.

2.3. Data augmentation

Data augmentation focuses on how to enhance the original training data. Basic augmentation techniques [11] include image cropping and rotation *etc*. More advanced methods include Mixup [42] and CutMix [41] *etc*. Mixup mixes pairs of images and their corresponding labels. CutMix combines two images by randomly cutting and pasting patches on images. In this paper, we have Qimera, IntraQ and TexQ *etc,* as our baselines, which utilize augmentation to enhance the generated images. However, they could not solve the mode collapse problem, as only augmentation on two generated similar images cannot synthesize new diverse images.

3. Methodology

3.1. Preliminary

Quantization is to compress a full-precision model $\mathcal{P}$ into a quantized model $\mathcal{Q}$ that also has good accuracy. Data-free quantization is divided into two processes [39]: data generation and quantization. Data generation is defined as follows:

$$\min_{\mathcal{G}} \mathcal{L}\Big[\mathcal{P}\big[\mathcal{G}(E_i, z)\big], y_i\Big], \tag{1}$$

where $\mathcal{G}$ denotes the generator, E_i is a learnable label embedding for label i, z is a random variable, $\mathcal{G}(E_i, z)$ is the generated image, y_i is the one-hot label for class i, and $\mathcal{L}$ is the cross-entropy loss. Quantization is defined as follows:

$$\min_{\mathcal{Q}} \mathcal{KL}\Big[\mathcal{P}\big[\mathcal{G}(E_i, z)\big], \mathcal{Q}\big[\mathcal{G}(E_i, z)\big]\Big], \tag{2}$$

where $\mathcal{KL}$ is the KL-divergence loss.

3.2. Overall architecture

We present our architecture in Figure 3. First, we extract features from multiple layers of the full-precision model, and utilize these multi-layer features to obtain the label embeddings. The label embeddings are fed into the generator to present and distinguish different kinds of classes. Second, we propose a multi-layer features mixer, which takes as inputs these multi-layer features, to learn the relations among all kinds of classes, and output each target label embedding to present more than one class. The target label embeddings are then fed into the generator to synthesize an image containing more classes, thereby enhancing inter-class diversity. Third, we propose normalization flow based attention, which takes as inputs the random variables from the Gaussian distribution and uses normalization flow to output the random variables with different non-Gaussian distributions to present minutia features of different levels. Then the random variables with different non-Gaussian distributions are fed into the flow based attention to help the generator focus on the features with different levels specifically, thereby enhancing the intra-class diversity, where different images can focus on different colors or textures *etc*. Last, we propose novel regulation to generate diverse data with more patterns from the perspective of activated values of features, and calibrate the quantized model from the full-precision model to adapt to diverse data.

3.3. Multi-layer features

The existing data-free quantization methods are inspired by GAN. Due to the lack of original data, it is difficult to initialize good label embeddings [39] for the generator to present and distinguish different kinds of classes. In order to obtain meaningful label embeddings and synthesize diverse data, we propose to utilize the features of different levels [14] from different layers of the trained full-precision model. For example, shallow layers mainly focus on shape or texture features for different classes, while deep layers mainly focus on semantic features for different classes. Therefore, we use such multi-layer features from the full-precision model to enhance the meaning of the label embeddings.

3.3.1. Label embeddings

First, we extract multi-layer features from the full-precision model, to obtain label embeddings with diverse information for the generator to present and distinguish different classes. The intuition is that we train a MLP^j classifier to distinguish features of each j-th level into right labels, and each raw of the weights $\mathcal{W}^j$ from the MLP^j classifier presents the clustered center [6, 11] of each class and thus can be the representation with j-th level information for each class. With the representations with different levels, we could have more meaningful label embeddings to present and distinguish different kinds of classes.

For any input $image_d$ to the full-precision model, which is synthesized by the warm-up generator [6], we denote $\mathcal{F}_d^j$ as the feature learned at the j-th layer of the full-precision model. Then, we use a multilayer perceptron (MLP) classifier to find out which class label this feature map $\mathcal{F}_d^j$ belongs to, as follows:

$$\hat{prob}_d^j = MLP^j(\mathcal{F}_d^j), \quad \hat{y}_d^j = argmax(\hat{prob}_d^j). \tag{3}$$

where MLP^j is the classifier for the j-th layer of the full-precision model, $\hat{prob}_d^j$ denotes the predicted probability distribution indicating which class label the j-th level feature $\mathcal{F}_d^j$ may belong to, and $\hat{y}_d^j$ denotes the predicted class label. The classifier is optimized as follows:

$$\min \left[\mathcal{L}(\hat{prob}_d^j, y_d) + \mathcal{KL}(\hat{prob}_d^j, prob_d) \right], \tag{4}$$

where $\mathcal{L}$ denotes the cross-entropy loss, $\mathcal{KL}$ denotes the KL-divergence loss, y_d is the class label and $prob_d$ is the final predicted probability distribution by the whole full-precision model for $image_d$.

Then, we can fuse the weights from all classifiers together to initial label embeddings for the generator to present and distinguish different kinds of classes:

$$E = Average\{\mathcal{W}^j | 1 \leq j \leq J\}, \tag{5}$$

where each raw of E is the label embedding for one class, $\mathcal{W}^j$ is the weights matrix of the MLP^j classifier, and J is a hyperparameter which denotes the total number of chosen layers from the full-precision model.

Instead of directly using the label embeddings and random Gaussian variables to generate images like existing works [5, 6, 39], we propose to further enhance the inter-class and inter-class diversity as follows.

3.3.2. Multi-layer features mixer

We propose a multi-layer features mixer module to enhance the inter-class diversity. Data augmentation such as Mixup [42] is used in previous quantization methods [5, 6, 22]. However, existing works [37, 40] show that such augmentation will result in inaccurate labels which will confuse the full-precision model and make it difficult to train the generator, as in their augmentation process entirely unrelated images are randomly mixed.

Here, we provide a multi-layer features mixer for augmentation which contains an attention-based Mixup on the label embeddings to capture relations among different classes.

The attention process is calculated as follows:

$$\begin{aligned} \mathcal{O}^0 &= softmax(\frac{Q\,K}{\sqrt{d}}) \cdot V = \sigma \cdot V \\ Q &= \mathcal{W}_Q \sum \lambda_i E_i \quad s.t. \sum \lambda_i = 1 \\ K &= \mathcal{W}_K \; concat[E_i] \\ V &= \mathcal{W}_V \; concat[E_i] \\ y &= \sum \sigma_i \; y_i \; , \end{aligned} \tag{6}$$

where E_i denotes the class embedding of the class i, λ_i is random weight when Mixup each class i, $\mathcal{W}_Q$, $\mathcal{W}_K$, $\mathcal{W}_V$ are learnable projection matrices, σ denotes the attention scores, σ_i is the attention score for class i, y_i is the one-hot label for class i, and y is the mixed target label. Multi-layer features mixer takes the embedding of the target label as the query, takes concatenated embeddings of all classes as the key and value, and outputs the mixed label embedding $\mathcal{O}^0$. Instead of using inaccurate label $y = \sum \lambda_i \; y_i$ and random weight λ_i to Mixup unrelated images as existing works do [5, 6, 22], our weight σ_i for the target label and mixed label embedding is the attention score which uses attention mechanism for $\mathcal{O}^0$ to capture relations among classes with the information of different levels extracted from the full-precision model.

3.3.3. Normalization flow based attention

In order to enhance the intra-class diversity, we propose to use the attention mechanism to make the generator focus on minutia features of different levels specifically. All the existing quantization methods [2, 5] directly combine all class embeddings and random variables from the same Gaussian distribution to generate data. These methods overlook the fact that minutia features of different levels may have different distribution spaces [17].

Following the normalization flow [35], we propose to use flow mapping to learn different non-Gaussian distributions as follows:

$$z^j = mapping^j(z), \tag{7}$$

where z is a random variable from Gaussian distribution and z^j is the random variable with j-th non-Gaussian distributions. Then, instead of directly concatenating class embeddings with random variables from the same Gaussian distribution as the existing methods, we make the generator focus on the minutia features of different levels with the attention mechanism:

$$\mathcal{O}^j = Transformer\big((\mathcal{O}^{j-1}, z^j)\big), for\, 1 \leq j \leq J, \tag{8}$$

where input $\mathcal{O}^0$ is the label embedding produced from multi-layer features mixer, and each attention output $\mathcal{O}^j$ is the j-th level embedding which has focused on minutia features of j-th level specifically. Then the generator can synthesize images with both enhanced inter-class and intra-class diversity with the last embedding $\mathcal{O}^J$:

$$image = \mathcal{G}(\mathcal{O}^J). \tag{9}$$

3.4. The loss function

First, we propose a similarity-based regulation loss to generate diverse data with more patterns from the perspective of activated feature values in the full-precision model. The motivation is that the generator should generate diverse data that exhibit more complex feature patterns in the full-precision model where the activated features should have a more diverse distribution rather than a centralized long-tailed distribution:

$$SimLoss_1 = \sum_d Cosine\Big(\mathcal{P}_a(image_d), f_{y_d}\Big), \tag{10}$$

where $image_d$ denotes the image generated from the generator, y_d is the label of $image_d$, $\mathcal{P}_a(image_d)$ is the activated features after the last activation layer for $image_d$ from the full-precision model, f_{y_i} denotes the moving average of the activated features from all images belonging to label y_i that are generated in previous iterations, $Cosine$ is the Cosine similarity measurement to make the data exhibit more diverse feature patterns.

Then, we use a distillation regulation to align the full-precision model and the quantized model in the activation layers, to learn appropriate clip ranges adaptive to our generated diverse data. We use J to denote the total number of chosen layers in the full-precision model and the quantized model. For the l-th activation layer where $1 \leq l \leq J$, we initialize a learnable clip range $[0, \beta^l]$. Then, we make the quantized values from each l-th activation layer with clip range $[0, \beta^l]$, to approximate the activated values of the full-precision model from the corresponding activation layer:

$$SimLoss_2 = \sum_l (a^l - \hat{a}^l)^2 \tag{11}$$

where a^l denotes the outputs from the l-th activation layer in the full-precision model, $\hat{a}^l$ denotes the quantized outputs with clip range $[0, \beta^l]$ from the l-th activation layer in the quantized model. With this distillation loss, the quantized model can learn clip ranges adaptive to the diverse data by aligning with the full-precision model.

Finally, in addition to the existing quantization loss in Eq.(1) and (2), our overall loss functions for the generator and quantized model are as follows.

$$\mathcal{L}_\mathcal{G} = \mathcal{L}\Big[\mathcal{P}\big[\mathcal{G}(\mathcal{O}^J)\big], y\Big] + SimLoss_1 \tag{12}$$

$$\mathcal{L}_\mathcal{Q} = \mathcal{KL}\Big[\mathcal{P}\big[\mathcal{G}(\mathcal{O}^J)\big], \mathcal{Q}\big[\mathcal{G}(\mathcal{O}^J)\big]\Big] + SimLoss_2 \tag{13}$$

Table 1. Quantization results for Transformer architectures on ImageNet dataset

Bit	Methods	ViT-S (81.39)	ViT-B (84.53)	DeiT-T (72.21)	DeiT-S (79.85)	DeiT-B (81.85)	Swin-T (81.35)	Swin-S (83.20)
4w8a	Standard (2022)	19.91	24.76	65.20	72.10	76.25	70.16	74.22
	PSAQ-ViT (2022)	20.84	25.34	65.57	73.23	77.05	71.79	75.14
	Standard V2 (2023)	-	-	68.43	75.98	79.17	75.51	78.22
	PSAQ-ViT V2 (2023)	-	-	68.61	76.36	79.49	76.28	78.86
	CLAMP-ViT (2024)	-	78.73	**69.93**	77.03	-	**80.28**	**82.51**
	EnhancingDFQ (ours)	**78.04**	**83.63**	69.89	**77.96**	**80.97**	80.23	**82.51**
8w8a	Standard (2022)	30.28	36.65	71.27	71.27	78.61	74.22	75.19
	PSAQ-ViT (2022)	31.45	37.36	71.56	76.92	79.10	75.35	76.64
	Standard V2 (2023)	-	-	72.06	79.24	81.26	79.62	81.42
	PSAQ-ViT V2 (2023)	-	-	**72.17**	79.56	81.52	80.21	82.13
	CLAMP-ViT (2024)	-	84.19	**72.17**	79.55	-	**81.17**	82.57
	EnhancingDFQ (ours)	**81.38**	**85.14**	72.15	**79.73**	**81.77**	81.12	**83.04**

4. Experiments

4.1. Experiments setup

We compare quantization performance on benchmark CIFAR-100 [19], which is a small dataset, and benchmark ImageNet [8], which is a large dataset. These two classification benchmarks are consensus for comparing data-free quantization performance [2, 5, 23, 24, 34, 39]. We compare our model with state-of-the-art data-free quantization methods, including GDFQ [39], ARC [45], Qimera [6], HAST [21], IntraQ [44], AdaSG [30], AdaDFQ [31], TexQ [5], and RIS [2] which are proposed to quantize for CNN architectures, and Standard [23], PSAQ-ViT [23], Standard V2 [24] PSAQ-ViT V2 [24] and CLAMP-ViT [34] which are proposed to quantize for Transformer architectures. More baseline details can be seen in Supplementary Material.

Following previous works [2, 5, 24, 34], we use pre-trained models from PytorchCV [14] as full-precision models for CNN architectures, *i.e*. ResNet-20, ResNet-18, ResNet-50 and MobileNet-V2 [14], and use pre-trained models from TIMM [38] as full-precision models for Transformer architectures, *i.e*. ViT [9], DeiT [36] and Swin [26]. To enable direct and fair comparisons, for CNN architectures, we quantize to 3w3a (3 bits for weight and 3 bits for activation) and 4w4a following existing CNN quantization methods [2, 5, 6, 31, 39], and for Transformer architectures, we quantize to 4w8a and 8w8a following existing Transformer quantization methods [23, 24, 34]. More experimental details can be seen in Supplementary Material.

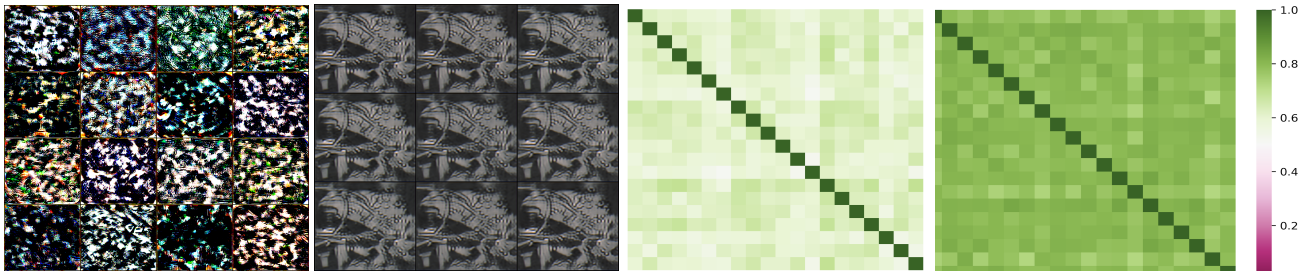
4.2. Overall comparison

Following existing works [2, 5, 6, 23, 24, 34], we report the top-1 accuracy (acc) results to evaluate the quantized models in Tables 1 and 2. We report experimental results from the corresponding papers of all baselines. For the rest quantization bit settings not covered in their original papers, we carefully tune the hyper-parameters based on the recommendations from the corresponding papers. If there are no available codes, we do not report their results and mark their results as "-". We randomly repeat 5 times and report average results for our method. We use bold to highlight the best accuracy, which outperforms the second best accuracy marked by underline. The top-1 accuracy results for the full-precision models are reported in parentheses under each full-precision model.

Table 2. Quantization results for CNN architectures

	Dataset	Cifar-100	ImageNet		
Bit	Methods	ResNet-20 (70.33)	ResNet-18 (71.47)	MobileNetV2 (73.03)	ResNet-50 (77.73)
3w3a	GDFQ	47.61	20.23	1.46	0.31
	ARC	40.15	23.37	14.30	1.63
	Qimera	46.13	28.23	0.27	1.82
	HAST	55.67	51.15	-	-
	AdaSG	52.76	37.04	26.90	16.98
	AdaDFQ	52.74	38.10	28.99	17.63
	RIS	53.08	48.71	32.05	25.79
	TexQ	55.87	50.28	32.38	25.27
	Ours	**56.17**	**52.27**	**34.10**	**28.32**
4w4a	GDFQ	63.75	60.60	59.43	52.12
	ARC	62.76	61.32	60.13	61.32
	Qimera	65.10	63.84	61.62	66.25
	HAST	66.68	66.91	65.60	-
	IntraQ	64.98	66.47	65.10	68.50
	AdaSG	66.42	66.50	65.15	68.58
	AdaDFQ	66.81	66.53	65.41	68.38
	RIS	65.99	67.55	66.90	70.61
	TexQ	67.18	67.73	67.07	70.72
	Ours	**67.31**	**68.13**	**67.10**	**71.20**

Key observations are summarized as follows. First, our method consistently achieves significantly superior quantization performance compared to state-of-the-art methods in almost all settings for both Transformer and CNN architectures, except for full-precision models DeiT-T and Swin-T where our method achieves compatible quantization performance compared to state-of-the-art baselines based on paired t-test at the significance level of 0.1. The results demonstrate



(a) Diverse images generated for a class by our method.

(b) Similar images generated for a class by Qimera.

(c) Similarity matrix of images for the same class by our method.

(d) Similarity matrix of images for the same class by Qimera.

Figure 4. Visualization shows that data generated by Qimera are more similar to each other while our method can generate more diverse data.

the effectiveness of our proposed method.

Second, the existing methods that propose to augment the generated images to calibrate the quantized model, *e.g.* Qimera, IntraQ and TexQ, outperform naive data-free quantization methods, *e.g.* GDFQ and ARC. It demonstrates that the mode collapse problem and the diversity of generated training data are important for the quantization performance. However, they all fail to outperform our method, as a simple augmentation on two highly similar images cannot synthesize diverse new images and they could not fundamentally solve the mode collapse and the diversity problem.

Third, our quantized model outperforms the full-precision model ViT-B in the 8w8a setting. This may be because the full-precision model ViT-B is not trained with Mixup [42] and is insensitive to the decision boundaries [37]. However, our quantization process includes an attention-based Mixup strategy, which may help the model to learn more accurate decision boundaries. Thus, our quantized model surprisingly outperforms the full-precision model ViT-B in the 8w8a setting with our attention-based Mixup strategy.

Last, our method consistently achieves good data-free quantization performance for both CNN and Transformer architectures, even though the baselines are specifically designed for CNN or Transformer architectures. This is because the mode collapse problem exists in all the generator-based data-free quantization frameworks, and our modules and loss functions are independent of the architectures of the full-precision models. Our modules and loss functions can be further integrated into the specific quantization methods as plugins to improve performance.

4.3. Ablation study

We conduct ablation experiments to validate the effectiveness of our modules and loss functions. We report the results of 4w4a quantization for Resnet50 and 4w8a quantization for DeiT-T on the ImageNet dataset in Table 3. In particular, "label embeddings" denotes whether we use the extracted multi-layer features to obtain label embeddings or randomly initialize label embeddings as existing methods [39, 45], "features mixer" denotes whether we use the multi-layer features mixer for attention-based Mixup or not, "flow based attention" denotes whether we use the normalization flow based attention or not, and "losses" denotes whether use our regulation losses or not. We can observe that: (1) Our label embeddings, multi-layer features mixer, normalization flow based attention, and regulation losses all help to improve the quantization performance, which demonstrates the effectiveness of each module. (2) By extracting multi-layer features from the full-precision model we can achieve better performance, which shows that we can obtain better label embeddings to provide information of different levels for the generator to present different classes and generate better calibration data. (3) With the multi-layer features mixer and normalization flow based attention we can achieve better performance, which shows that we can enhance the inter-class and intra-class diversity to generate diverse data for more accurate calibration. (4) With our regulation losses we can achieve better performance, which shows that generating diverse data with more complex feature patterns and aligning the activation layers help quantization.

Table 3. Ablation study

label embeddings	features mixer	flow based attention	losses	ResNet-50 (4w4a)	DeiT-T (4w8a)
✓	✓	✓	✓	**71.20**	**69.89**
✓	✓	✓		69.14	68.37
✓	✓		✓	70.45	69.18
✓		✓	✓	68.79	68.22
	✓	✓	✓	69.72	68.83
✓				66.21	67.03
	✓			66.42	66.72
		✓		65.72	65.89
			✓	68.08	66.92
✗	✗	✗	✗	52.12	64.83

4.4. Case study

We use visualization to further evaluate our method. First, in Figure 4a and 4b, we can see that our generated images are more diverse, compared to the images from Qimera which are highly similar. Besides, the quantitative analysis

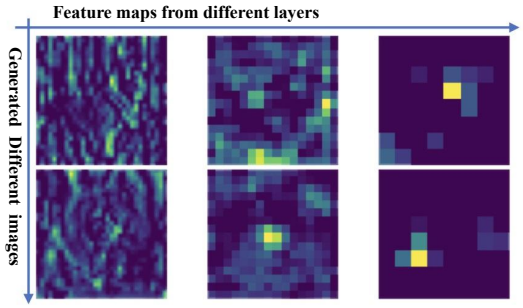


Figure 5. Visualization for different features of three levels from ResNet-50, which shows our generated images for a class are also diverse from the perspective of feature patterns.

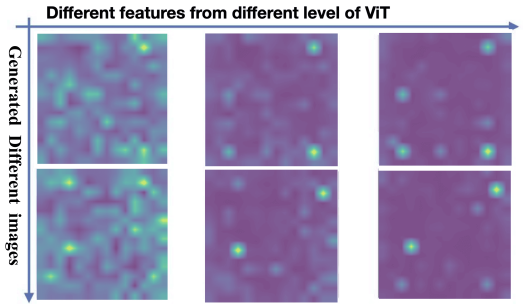


Figure 6. Visualization for different features from ViT-B of our generated images for a class.

in Figure 4c and 4d, where we calculate the paired similarity score [6] between any two images generated for the same class, shows that the images generated by our method are less similar with each other comparing to Qimera.

Second, in Figure 5 and 6, we can observe that features from different layers focus on information of different levels, and our two generated data are dissimilar from the perspective of feature patterns.

Last, by utilizing the minutia features extracted from the full-precision model to obtain better class embeddings, we can synthesize data that are more like the original dataset, as shown in Figure 7 with PCA visualization [6], where each sub-figure presents the PCA plots for images within the same one class, where equivalent images are from the original CIFAR-100 dataset, or generated by Qimera, AdaDFQ or our method, respectively. From Figure 7 we can also see that the PCA plots of our generated images are more diverse, while the PCA plots of the images generated by Qimera or AdaDFQ are closer to each other, respectively.

4.5. More analysis

We report the overall running time cost for different quantization methods in Table 4. More time complexity analysis can be seen in Supplementary Material, which shows our method is as efficient as the baselines and our overall time complexity has the same order as the baselines. We also evaluate our hyperparameters in Supplementary Material, *i.e.* J

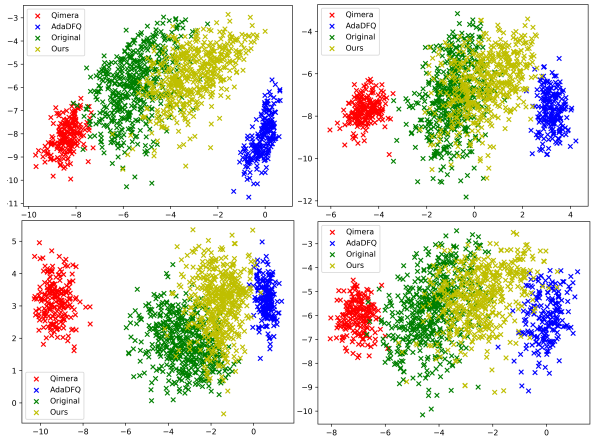


Figure 7. PCA visualization for images within four different classes, respectively. The images are from the original CIFAR-100 dataset, and generated by Qimera, AdaDFQ and our method.

Table 4. Runtime comparison on ResNet-18 quantization (the unlisted baselines neither have open codes nor analyze runtime).

Methods	GPU Hours
GDFQ (ECCV 2020)	7.4
Qimera (NeurIPS 2021)	8.3
ARC (IJCAI 2021)	13.5
AdaSG (AAAI 2023)	8.6
AdaDFQ (CVPR 2023)	9.0
EnhancingDFQ (ours)	8.7

the total number of layers of the full-precision model where we extract multi-layer features to obtain class embeddings, the number of Transformer blocks in Eq.(8), and the number of heads for attention in Eq.(6) and Transformer blocks in Eq.(8). More analysis of our loss functions also can be seen in Supplementary Material, where our loss functions are integrated into existing quantization methods as plugins and have improved their performance.

5. Conclusion

We address the diversity problem for data-free quantization. We utilize information from the full-precision model, and propose multi-layer features mixer that learns the relations among all classes, and normalization flow based attention that focuses on the features of different levels, to enhance both inter-class and intra-class diversity. We also propose novel loss functions to generate diverse data and make the quantized model more adaptive to diverse data. Extensive experiments show that our method achieves state-of-the-art quantization results for both Transformer and CNN architectures. In future work, it is of interest to use our modules as plugins to improve specific quantization methods, and research on quantization for large models and other tasks.

Acknowledgments: This work was partially supported by National Natural Science Foundation of China (62372179). Miao Zhang was partially sponsored by the National Natural Science Foundation of China under Grant 62306084 and U23B2051, and Shenzhen Science and Technology Program under Grant GXWD20231128102243003, KJZD20230923115113026, ZDSYS20230626091203008.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017. 2

[2] Jianhong Bai, Yuchen Yang, Huanpeng Chu, Hualiang Wang, Zuozhu Liu, Ruizhe Chen, Xiaoxuan He, Lianrui Mu, Chengfei Cai, and Haoji Hu. Robustness-guided image synthesis for data-free quantization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10971–10979, 2024. 5, 6

[3] Ron Banner, Yury Nahshan, Elad Hoffer, and Daniel Soudry. ACIQ: Analytical clipping for integer quantization of neural networks, 2019. 1

[4] Yaohui Cai, Zhewei Yao, Zhen Dong, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. Zeroq: A novel zero shot quantization framework. *CoRR*, abs/2001.00281, 2020. 2

[5] Xinrui Chen, Yizhi Wang, Renao Yan, Yiqing Liu, Tian Guan, and Yonghong He. Texq: Zero-shot network quantization with texture feature distribution calibration. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2, 3, 4, 5, 6

[6] Kanghyun Choi, Deokki Hong, Noseong Park, Youngsok Kim, and Jinho Lee. Qimera: Data-free quantization with synthetic boundary supporting samples. In *NeurIPS*, pages 14835–14847, 2021. 1, 2, 3, 4, 5, 6, 8

[7] Yoojin Choi, Jihwan P. Choi, Mostafa El-Khamy, and Jungwon Lee. Data-free network quantization with adversarial knowledge distillation. In *CVPR*, pages 3047–3057. IEEE, 2020. 1

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 6

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2, 6

[10] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS, 2014*, pages 2672–2680, 2014. 1, 2

[11] Ian J. Goodfellow, Yoshua Bengio, and Aaron C. Courville. *Deep Learning*. MIT Press, 2016. 3, 4

[12] Cong Guo, Yuxian Qiu, Jingwen Leng, Xiaotian Gao, Chen Zhang, Yunxin Liu, Fan Yang, Yuhao Zhu, and Minyi Guo. SQuant: On-the-fly data-free quantization via diagonal hessian approximation. In *ICLR*, 2022. 1

[13] Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding. In *ICLR*, 2016. 1

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4, 6

[15] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. 1

[16] Norman P. Jouppi, Cliff Young, Nishant Patil, David A. Patterson, Gaurav Agrawal, Raminder Bajwa, and et al. In-datacenter performance analysis of a tensor processing unit. In *ISCA*, pages 1–12. ACM, 2017. 1

[17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CoRR*, abs/1812.04948, 2018. 3, 5

[18] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 2

[19] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*, 2009. 6

[20] Colin Lea, Michael D. Flynn, René Vidal, Austin Reiter, and Gregory D. Hager. Temporal convolutional networks for action segmentation and detection. In *CVPR*, pages 1003–1012, 2017. 2

[21] Huantong Li, Xiangmiao Wu, Fanbing Lv, Daihai Liao, Thomas H Li, Yonggang Zhang, Bo Han, and Mingkui Tan. Hard sample matters a lot in zero-shot quantization. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 24417–24426, 2023. 3, 6

[22] Yuhang Li, Feng Zhu, Ruihao Gong, Mingzhu Shen, Xin Dong, Fengwei Yu, Shaoqing Lu, and Shi Gu. Mixmix: All you need for data-free compression are feature and data mixing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4410–4419, 2021. 4, 5

[23] Zhikai Li, Liping Ma, Mengjuan Chen, Junrui Xiao, and Qingyi Gu. Patch similarity aware data-free quantization for vision transformers. In *ECCV*, pages 154–170, 2022. 3, 6

[24] Zhikai Li, Mengjuan Chen, Junrui Xiao, and Qingyi Gu. Psaq-vit v2: Towards accurate and general data-free quantization for vision transformers. *IEEE transactions on neural networks and learning systems*, PP, 2023. 3, 6

[25] Zhengyi Li, Cong Guo, Zhanda Zhu, Yangjie Zhou, Yuxian Qiu, Xiaotian Gao, Jingwen Leng, and Minyi Guo. Efficient adaptive activation rounding for post-training quantization, 2023. 1

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 6

[27] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014. 2

[28] Markus Nagel, Mart van Baalen, Tijmen Blankevoort, and Max Welling. Data-free quantization through weight equal-

ization and bias correction. In *ICCV*, pages 1325–1334. IEEE, 2019. 1

[29] Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart van Baalen, and Tijmen Blankevoort. A white paper on neural network quantization. *CoRR*, abs/2106.08295, 2021. 1

[30] Biao Qian, Yang Wang, Richang Hong, and Meng Wang. Rethinking data-free quantization as a zero-sum game. In *AAAI*, pages 9489–9497. AAAI Press, 2023. 3, 6

[31] Biao Qian, Yang Wang, Richang Hong, and Meng Wang. Adaptive data-free quantization. *CoRR*, abs/2303.06869, 2023. 3, 6

[32] Haotong Qin, Yifu Ding, Xiangguo Zhang, Aoyu Li, Jiakai Wang, Xianglong Liu, and Jiwen Lu. Diverse sample generation: Pushing the limit of data-free quantization. *CoRR*, abs/2109.00212, 2021. 1

[33] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016. 1, 2

[34] Akshat Ramachandran, Souvik Kundu, and Tushar Krishna. Clamp-vit: Contrastive data-free learning for adaptive post-training quantization of vits, 2024. 3, 6

[35] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *ICML*, pages 1530–1538, 2015. 5

[36] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357, 2021. 6

[37] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6438–6447, Long Beach, California, USA, 2019. PMLR. 4, 7

[38] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019. 6

[39] Shoukai Xu, Haokun Li, Bohan Zhuang, Jing Liu, Jiezhang Cao, Chuangrun Liang, and et al. Generative low-bitwidth data free quantization. In *ECCV*, pages 1–17. Springer, 2020. 2, 3, 4, 6, 7

[40] Yoichi Yaguchi, Fumiyuki Shiratani, and Hidekazu Iwaki. Mixfeat: Mix feature in latent space learns discriminative space, 2019. 4

[41] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, pages 6022–6031. IEEE, 2019. 2, 3

[42] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *CoRR*, abs/1710.09412, 2017. 2, 3, 4, 7

[43] Xiangguo Zhang, Haotong Qin, Yifu Ding, Ruihao Gong, Qinghua Yan, Renshuai Tao, Yuhang Li, Fengwei Yu, and Xianglong Liu. Diversifying sample generation for accurate data-free quantization. In *CVPR*, pages 15658–15667. Computer Vision Foundation / IEEE, 2021. 2, 3

[44] Yunshan Zhong, Mingbao Lin, Gongrui Nan, Jianzhuang Liu, Baochang Zhang, Yonghong Tian, and et al. Intraq: Learning synthetic images with intra-class heterogeneity for zero-shot network quantization. In *CVPR*, pages 12329–12338. IEEE, 2022. 2, 3, 6

[45] Baozhou Zhu, H. Peter Hofstee, Johan Peltenburg, Jinho Lee, and Zaid Al-Ars. Autorecon: Neural architecture search-based reconstruction for data-free compression. In *IJCAI*, pages 3470–3476. ijcai.org, 2021. 3, 6, 7