

Generative Low-bitwidth Data Free Quantization

Shoukai Xu^{1,2*}, Haokun Li^{1*}, Bohan Zhuang^{3*},
Jing Liu¹, Jiezhong Cao¹, Chuangrun Liang¹, and Mingkui Tan^{1†}

¹ South China University of Technology, Guangzhou, China
{sexsk,selihaokun,secaojiezhong,seliujing,selcr}@mail.scut.edu.cn
mingkuitan@scut.edu.cn

² PengCheng Laboratory, Shenzhen, China

³ Monash University, Melbourne, Australia
bohan.zhuang@monash.edu

Abstract. Neural network quantization is an effective way to compress deep models and improve their execution latency and energy efficiency, so that they can be deployed on mobile or embedded devices. Existing quantization methods require original data for calibration or fine-tuning to get better performance. However, in many real-world scenarios, the data may not be available due to confidential or private issues, thereby making existing quantization methods not applicable. Moreover, due to the absence of original data, the recently developed generative adversarial networks (GANs) cannot be applied to generate data. Although the full-precision model may contain rich data information, such information alone is hard to exploit for recovering the original data or generating new meaningful data. In this paper, we investigate a simple-yet-effective method called Generative Low-bitwidth Data Free Quantization (GDFQ) to remove the data dependence burden. Specifically, we propose a knowledge matching generator to produce meaningful fake data by exploiting classification boundary knowledge and distribution information in the pre-trained model. With the help of generated data, we can quantize a model by learning knowledge from the pre-trained model. Extensive experiments on three data sets demonstrate the effectiveness of our method. More critically, our method achieves much higher accuracy on 4-bit quantization than the existing data free quantization method. Code is available at <https://github.com/xushoukai/GDFQ>.

Keywords: Data Free Compression · Low-bitwidth Quantization · Knowledge Matching Generator

1 Introduction

Deep neural networks (DNNs) have achieved great success in many areas, such as computer vision [16,41,37,15,44] and natural language processing [28,36,12].

* Authors contributed equally.

† Corresponding author.

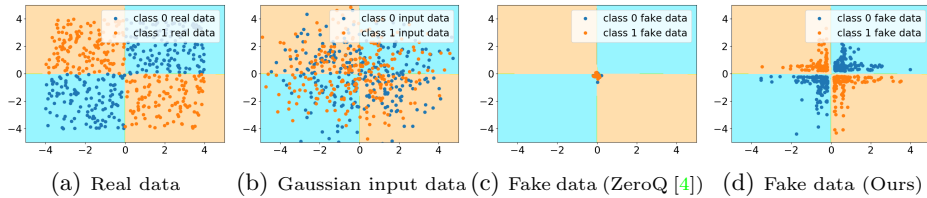


Fig. 1: The comparisons of generated fake data. ZeroQ generates fake data by gradient updating; however, since it neglects the inter-class information, it is hard to capture the original data distribution. Meanwhile, the knowledge matching generator produces fake data with label distribution, and it is more likely to produce data distributed near the classifier boundaries.

However, DNNs often contain a considerable number of parameters, which makes them hard to be deployed on embedded/edge devices due to unbearable computation costs for inference. Network quantization, which aims to reduce the model size by quantizing floating-point values into low precision (*e.g.*, 4-bit), is an effective way to improve the execution latency and energy efficiency. Existing quantization methods [50,10,52,49,24] generally require training data for calibration or fine-tuning. Nevertheless, in many real-world applications in medical [47], finance [48] and industrial domains, the training data may not be available due to privacy or confidentiality issues. Consequently, these methods are no longer applicable due to the absence of training data. Therefore, the data-free quantization is of great practical value.

To address this issue, one possible way is to directly sample random inputs from some distribution and then use these inputs to quantize the model so that the output distributions of the full-precision model and quantized model are as close as possible. Unfortunately, since random inputs contain no semantic information and are far away from the real data distribution, this simple method suffers from huge performance degradation. Alternatively, one can use generative adversarial networks (GANs) to produce data. Due to the absence of training data, GANs cannot be applied in the data-free quantization. In practice, the quantization performance highly depends on the quality of the input data. Therefore, the process of constructing meaningful data to quantize models is very challenging.

To generate meaningful data, it is important and necessary to exploit data information from a pre-trained model. A recent study [45] revealed that a well-trained over-parameterized model maintains sufficient information about the entire data set. Unfortunately, what information exists and how to exploit such information are still unknown. In the training, a neural network uses batch normalization [19] to stabilize the data distribution and learns a classification boundary to divide data into different classes (See Fig. 1 (a)). In this sense, some information about the classification boundary and data distribution is hidden in the pre-trained model. However, these kinds of information are disregarded by existing data-free quantization methods [4,30], which only focus on a single sample or

network parameters. For example, ZeroQ [4] exploits the information of a single sample instead of entire data, causing the constructed distribution to be far away from the real data distribution (See Fig. 1 (c)). To address these issues, how to construct meaningful data by fully exploiting the classification boundary and distribution information from a pre-trained model remains an open question.

In this paper, we propose a simple-yet-effective method called Generative Low-bitwidth Data Free Quantization to achieve completely data-free quantization. Without original data, we aim to learn a good generator to produce meaningful data by mining the classification boundary knowledge and distribution information in batch normalization statistics (BNS) from the pre-trained full-precision model.

The main contributions of this paper are summarized as follows:

- We propose a scheme called Generative Low-bitwidth Data Free Quantization (GDFQ), which performs 4-bitwidth quantization without any real data. To our knowledge, this is the first low-bitwidth data free quantization method.
- We propose an effective knowledge matching generator to construct data by mining knowledge from the pre-trained full-precision model. The generated data retain classification boundary knowledge and data distribution.
- Extensive experiments on image classification data sets demonstrate the superior performance of our method compared to the existing data-free quantization method.

2 Related Work

Model quantization. Model quantization targets to quantize weights, activations and even gradients to low-precision, to yield highly compact models, where the expensive multiplication operations can be replaced by additions or bit-wise operations. According to the trade-off between accuracy and computational complexity, quantization can be roughly categorized into binary neural networks (BNNs) [18,35,53] and fixed-point quantization [50,13,52]. Moreover, quantization studies mainly focus on tackling two core bottlenecks, including designing accurate quantizers to fit the categorical distribution to the original continuous distribution [5,21,46], and approximating gradients of the non-differential quantizer during back-propagation [42,26,51]. In addition, the first practical 4-bit post-training quantization approach was introduced in [3]. To improve the performance of neural network quantization without retraining, outlier channel splitting (OCS) [49] proposed to move affected outliers toward the center of the distribution. Besides, many mainstream deep learning frameworks support 16-bit half-precision or 8-bit fixed-point quantization, such as TensorFlow [1], PyTorch [34], PaddleSlim, etc. In particular, these platforms provide both post-training quantization and quantization-aware training. However, all of them require the original data. In short, whether in scientific studies or practical applications, if the original data is unavailable, it is hard for quantization methods to work normally.

Data-free model compression. Recently, the researches about data-free had included more model compression methods, such as knowledge distillation [31,8], low-rank approximation [43] and model quantization [49]. A novel framework [8] exploited generative adversarial networks to perform data-free knowledge distillation. Another work [25] reconstructed a new data set based solely on the trained model and the activation statistics, and finally distilled the pre-trained “teacher” model into the smaller “student” network. Zero-shot knowledge distillation [31] synthesized pseudo data by utilizing the prior information about the underlying data distribution. Specifically, the method extracted class similarities from the parameters of the teacher model and modeled the output space of the teacher network as a Dirichlet distribution. KEGNET [43] proposed a novel data-free low-rank approximation approach assisted by knowledge distillation. This method contained a generator that generated artificial data points to replace missing data and a decoder that aimed to extract the low-dimensional representation of artificial data. In Adversarial Belief Matching [27], a generator generated data on which the student mismatched the teacher, and then the student network was trained using these data.

Quantization also faces the situation without original data while previous quantization methods generally need original data to improve their performance. However, in some instances, it is difficult to get the original data. Recently, some studies focused on data-free model quantization. DFQ [30] argued that data-free quantization performance could be improved by equalizing the weight ranges of each channel in the network and correcting biased quantization error. ZeroQ [4], a novel zero-shot quantization framework, enabled mixed-precision quantization without any access to the training or validation data. However, these data-free quantization methods work well for 8-bit, but got poor performance in aggressively low bitwidth regimes such as 4-bit.

3 Problem Definition

Data-free quantization problem. Quantization usually requires original data for calibration or fine-tuning. In many practical scenarios, the original data may not be available due to private or even confidential issues. In this case, we cannot use any data; thus, the general scheme of the network quantization will lose efficacy and even fail to work completely, resulting in a quantized model with inferior performance. Given a full-precision model M , data-free quantization aims to construct fake data $(\hat{\mathbf{x}}, y)$ and meanwhile quantize a model Q from the model M_θ . Specifically, to compensate for the accuracy loss from quantization, training-aware quantization can fine-tune the quantized model by optimizing the following problem:

$$\min_{Q, \hat{\mathbf{x}}} \mathbb{E}_{\hat{\mathbf{x}}, y} [\ell(Q(\hat{\mathbf{x}}), y)], \quad (1)$$

where $\hat{\mathbf{x}}$ is a generated fake sample, y is the corresponding label, and $\ell(\cdot, \cdot)$ is a loss function, such as cross-entropy loss and mean squared error.

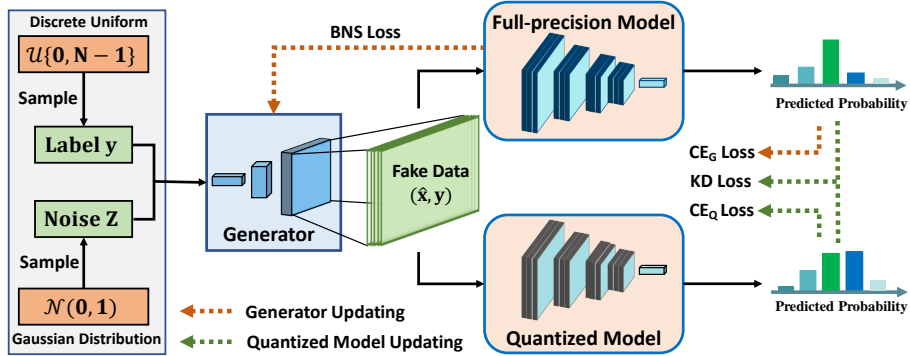


Fig. 2: An overview of the proposed method. Given Gaussian noise and the label as input, the generator creates fake data and feeds them into both the full-precision model and the quantized model. The fixed full-precision model provides knowledge for updating the generator. The quantized model learns latent knowledge from the generator and the full-precision model.

Challenges of constructing data. Because of the absence of original data, one possible way is to construct data by exploiting information from a pre-trained full-precision model. Although the full-precision model may contain rich data information, such latent information alone is hard to exploit for recovering the original data. In practice, the performance of quantization highly depends on the quality of constructed data. With the limited information of the pre-trained model, constructing meaningful data is very challenging.

Recently, one data-free quantization method (ZeroQ) [4] constructs fake data by using a linear operator with gradient update information. With the help of constructed data, ZeroQ is proposed to improve the quantization performance. However, ZeroQ has insufficient information to improve the performance of quantization with the following two issues. First, it constructs fake data without considering label information, and thus neglects to exploit the classification boundary knowledge from the pre-trained model. Second, it enforces the batch normalization statistics of a single data point instead of the whole data, leading to being far away from the real data distribution. To address these issues, one can construct a generator G to produce fake data by considering label information and using powerful neural networks,

$$\hat{\mathbf{x}} = G(\mathbf{z}|y), \quad \mathbf{z} \sim p(\mathbf{z}), \quad (2)$$

where \mathbf{z} is a random vector drawn from some prior distribution $p(\mathbf{z})$, *e.g.*, Gaussian distribution or uniform distribution. By using the generator, we are able to construct fake data to improve quantization performance. However, what knowledge in the pre-trained model can be exploited and how to learn a good generator remain to be answered.

4 Generative Low-bitwidth Data Free Quantization

In many practical scenarios, training data are unavailable; thus, an existing method [4] performs data-free quantization by constructing data. However, without exploiting knowledge from a pre-trained model, directly constructing data lacks label information and leads to being far away from the real data distribution. To address this issue, we aim to design a generator to produce fake data. Then, we are able to perform supervised learning to improve the performance of quantization. The overall framework is shown in Fig. 2.

4.1 Knowledge Matching Generator

When training a deep neural network, it captures sufficient data information to make a decision [45]. In this sense, the pre-trained DNN contains some knowledge information of the training data, *e.g.*, classification boundary information and distribution information. We therefore elaborate how to effectively construct informative data from the pre-trained model.

Classification boundary information matching. The pre-trained model contains classification boundary information. Unfortunately, such information is difficult to exploit to recover the data near to the classification boundary. Recently, generative adversarial networks (GANs) [14,6,7] have achieved considerable success in producing data. Since the real data are unavailable, the discriminator in a GAN cannot work in the data-free quantization task. Without the discriminator, learning a generator to produce meaningful data is difficult.

In this paper, we propose a generator to produce fake data for the data-free quantization task. For this task, although the original data cannot be observed, we are able to easily confirm the number of categories of original data by the last layer in a pre-trained model. As shown in Fig. 1 (a), different categories of data should be distributed in different data spaces. To generate fake data, we introduce a noise vector \mathbf{z} conditioned on a label y . Here, we sample noise from Gaussian distribution $\mathcal{N}(0, 1)$ and uniformly sample a label from $\{0, 1, \dots, n-1\}$. Then, the generator maps a prior input noise vector \mathbf{z} and the given label y to the fake data $\hat{\mathbf{x}}$. Formally, we define the knowledge matching generator as follows:

$$\hat{\mathbf{x}} = G(\mathbf{z}|y), \quad \mathbf{z} \sim \mathcal{N}(0, 1). \quad (3)$$

To improve the quantization performance, the generator should have the ability to generate data that are effective to fine-tune a quantized model. To this end, given the Gaussian noise \mathbf{z} and the corresponding label y , the generated fake data should be classified to the same label y by the full-precision pre-trained model M . Therefore, we introduce the following cross-entropy loss $\text{CE}(\cdot, \cdot)$ to train the generator G :

$$\mathcal{L}_{\text{CE}}^G(G) = \mathbb{E}_{\mathbf{z}, y} [\text{CE}(M(G(\mathbf{z}|y)), y)]. \quad (4)$$

Distribution information matching. In addition, the pre-trained model contains the distribution information of training data. Such distribution information can be captured by batch normalization [19], which is used to control the change of the distribution. While training the full-precision model, the batch normalization statistics (BNS), *i.e.*, the mean and variance, are computed dynamically. For every batch, batch normalization (BN) layers only compute statistics on the current mini-batch inputs and accumulate them with momentum. Finally, the exponential moving average (EMA) batch normalization statistics will be obtained and then they will be used in network validation and test.

To retain the BNS information, the mean and variance of the generated distribution should be the same as that of the real data distribution. To this end, we learn a generator G with the BNS information (μ_l and σ_l) encoded in the l -th BN layer of the pre-trained model using \mathcal{L}_{BNS} :

$$\mathcal{L}_{\text{BNS}}(G) = \sum_{l=1}^L \|\mu_l^r - \mu_l\|_2^2 + \|\sigma_l^r - \sigma_l\|_2^2, \quad (5)$$

where μ_l^r and σ_l^r are the mean and variance of the fake data’s distribution at the l -th BN layer, respectively, and μ_l and σ_l are the corresponding mean and variance parameters stored in the l -th BN layer of the pre-trained full-precision model, respectively. In this way, we are able to learn a good generator to keep the distribution information from the training data.

4.2 Fake-data Driven Low-bitwidth Quantization

With the help of the generator, the data-free quantization problem can be turned to a supervised quantization problem. Thus, we are able to quantize a model using produced meaningful data. However, transferring knowledge from a pre-trained model to a quantized model is difficult. To address this, we introduce a fake-data driven quantization method and solve the optimization problem by exploiting knowledge from the pre-trained model.

Quantization. Network quantization maps full-precision (32-bit) weights and activations to low-precision ones, *e.g.*, 8-bit fixed-point integer. We use a simple-yet-effective quantization method refer to [20] for both weights and activations. Specifically, given full-precision weights θ and the quantization precision k , we quantize θ to θ_q in the symmetric k -bit range:

$$\theta_q = \begin{cases} -2^{k-1}, & \text{if } \theta' < -2^{k-1} \\ 2^{k-1}-1, & \text{if } \theta' > 2^{k-1}-1 \\ \theta', & \text{otherwise,} \end{cases} \quad (6)$$

where θ' are discrete values mapped by a linear quantization, *i.e.*, $\theta' = \lfloor \Delta \cdot \theta - b \rfloor$, $\lfloor \cdot \rfloor$ is the round function, and Δ and b can be computed by $\Delta = \frac{2^k-1}{u-l}$ and $b = l \cdot \Delta + 2^{k-1}$. Here, l and u can be set as the minimum and maximum of the floating-point weights θ , respectively.

Algorithm 1 Generative Low-bitwidth Data Free Quantization

Input: Pre-trained full-precision model M .**Output:** Generator G , quantized model Q .Update G several times as a warm-up process.Quantize model M using Eq. (6), get quantized model Q .Fix the batch normalization statistics in all BN layers of quantized model Q .**for** $t = 1, \dots, T_{fine-tune}$ **do**Obtain random noise $\mathbf{z} \sim \mathcal{N}(0, 1)$ and label y .Generate fake image $\hat{\mathbf{x}}$ using Eq. (3).Update generator G by minimizing Loss (9).Update quantized model Q by minimizing Loss (10).**end for**

Optimization problem. When the real training data are unavailable, quantization may suffer from some limitations. First, direct quantization from a full-precision model may result in severe performance degradation. To address this issue, we aim to train the quantized model to approximate the full-precision model through the fine-tuning process. To this end, a well fine-tuned quantized model Q and the pre-trained model M should classify the fake data correctly. For this purpose, we use the cross-entropy loss function $\text{CE}(\cdot, \cdot)$ to update Q :

$$\mathcal{L}_{\text{CE}}^Q(Q) = \mathbb{E}_{\hat{\mathbf{x}}, y} [\text{CE}(Q(\hat{\mathbf{x}}), y)]. \quad (7)$$

By minimizing Eq. (7), the quantization model can be trained with the generated data to perform multi-class classification.

Second, the traditional fine-tuning process with a common classification loss function is insufficient because the data are fake. However, with the help of fake data, we are able to apply knowledge distillation [17] to improve the quantization performance. Specifically, given the same inputs, the outputs of a quantized model and full-precision model should be close enough to guarantee that the quantized model is able to achieve nearly performance compared with the full-precision model. Therefore, we utilize knowledge distillation to recover the performance of the quantized model by using fake data $\hat{\mathbf{x}}$ to simulate the training data. Then, the quantized model can be fine-tuned using the following Kullback-Leibler loss function $\text{KL}(\cdot, \cdot)$:

$$\mathcal{L}_{\text{KD}}(Q) = \mathbb{E}_{\hat{\mathbf{x}}} [\text{KL}(Q(\hat{\mathbf{x}}), M(\hat{\mathbf{x}}))]. \quad (8)$$

By optimizing the loss in (8), the quantization model can learn knowledge from the full-precision model.

Fine-tuning with fixed BNS. To stabilize the training, we fix the batch normalization statistics (BNS) in the quantized model during fine-tuning. In this way, the BNS in the quantized model are corresponding to the statistics of the real data distribution. With the help of fixed BNS, the quantized model always maintains the real data information to improve quantization performance.

4.3 Training Process

We propose a low-bitwidth quantization algorithm to alternately optimize the generator G and the quantized model Q . In our alternating training strategy, the generator is able to generate different data with each update. By increasing the diversity of data, the quantized model Q can be trained to improve the performance. In addition, to make the fine-tuning of Q more stable, we firstly train G solely several times as a warm-up process. The overall process is shown in Algorithm 1. In contrast, one can train the generator G with a full-precision model, and then fix the generator to train the quantized model Q until convergence. In this separate training strategy, when the diversity of the generated data is poor, the quantized model has a limitation to improve the quantization performance.

Training generator G . First, we randomly sample a Gaussian noise vector $\mathbf{z} \sim \mathcal{N}(0, 1)$ with a label y . Then we use G to generate fake data from the distribution and update G . The final generator loss $\mathcal{L}_1(G)$ is formulated as follows:

$$\mathcal{L}_1(G) = \mathcal{L}_{\text{CE}}^G(G) + \beta \mathcal{L}_{\text{BNS}}(G), \quad (9)$$

where β is a trade-off parameter.

Training quantized model Q . We quantize the model according to Eq. (6). Then, we replace the BNS in the quantized model with the fixed batch normalization statistics (FBNS) as described in Section 4.2. So far, the quantized model has inherited the information contained in BNS and a part of latent knowledge from the parameters of the pre-trained model. In the fine-tuning process, we train the G and Q alternately in every epoch. Based on the warmed up G , we obtain fake samples and use them to optimize the quantized model Q . The final quantized model loss function $\mathcal{L}_2(Q)$ is formulated as follows:

$$\mathcal{L}_2(Q) = \mathcal{L}_{\text{CE}}^Q(Q) + \gamma \mathcal{L}_{\text{KD}}(Q), \quad (10)$$

where γ is a trade-off parameter. We do not stop updating G because if we have a better G , the fake data will be more similar to real training data and the upper limit of optimizing Q will be improved. Note that we keep the pre-trained full-precision model fixed at all times.

5 Experiments

5.1 Data Sets and Implementation Details

We evaluate the proposed method on well-known data sets including CIFAR-10 [23], CIFAR-100 [23], and ImageNet [11]. CIFAR-10 consists of 60k images from 10 categories, with 6k images per category. There are 50k images for training and 10k images for testing. CIFAR-100 has 100 classes and each class contains 500 training images and 100 testing images. ImageNet is one of the most

challenging and largest benchmark data sets for image classification, which has around 1.2 million real-world images for training and 50k images for validation.

Based on the full-precision pre-trained models from pytorchcv¹, we quantize ResNet-20 [16] on CIFAR-10/100 and ResNet-18 [16], BN-VGG-16 [38] and Inception v3 [40] on ImageNet. In all experiments, we quantize all layers including the first and last layers of the network following [4] and the activation clipping values are per-layer granularity. All implementations are based on PyTorch.

For CIFAR-10/100, we construct the generator following ACGAN [33] and the dimension of noise is 100. During training, we optimize the generator and quantized model using Adam [22] and SGD with Nesterov [32] respectively, where the momentum term and weight decay in Nesterov are set to 0.9 and 1×10^{-4} . Moreover, the learning rates of quantized models and generators are initialized to 1×10^{-4} and 1×10^{-3} , respectively. Both of them are decayed by 0.1 for every 100 epochs. In addition, we train the generator and quantized model for 400 epochs with 200 iterations per epoch. To obtain a more stable clip range for activation, we calculate the moving average of activation’s range in the first four epochs without updating the quantized models and then fix this range for subsequent training. For $\mathcal{L}_1(G)$ and $\mathcal{L}_2(Q)$, we set $\beta = 0.1$ and $\gamma = 1$ after a simple grid search. For ImageNet, we replace the generator’s standard batch normalization layer with the categorical conditional batch normalization layer for fusing label information following SN-GAN [29] and set the initial learning rate of the quantized model as 1×10^{-6} . Other training settings are the same as those on CIFAR-10/100.

5.2 Toy Demonstration for Classification Boundary Matching

To evaluate that the fake data generated by our G are able to match the classification boundary information, we design a toy experiment. The results are shown in Fig. 1. First, we create a toy binary classification data set by uniform sampling from -4 to $+4$, and the label is shown in Fig. 1 (a). Second, we construct a simple neural network T , which is composed of several linear layers, BN layers, and ReLU layers, and we train it using the toy data. The classification boundaries are shown in each subfigure. To simulate the process of our method, we sample noises from Gaussian distribution and every noise has a random label of 0 or 1 (Fig. 1 (b)). Then, we generate fake data from noises by learning from the pre-trained model T . Fig. 1 (c) and Fig. 1 (d) show the fake data generated by the ZeroQ method and our method, respectively. The data generated by ZeroQ do not capture the real data distribution since it neglects the inter-class information; while our method is able to produce fake data that not only have label information but also match the classification boundaries.

5.3 Comparison of the Results

To further evaluate the effectiveness of our method, we include the following methods for study. **FP32**: the full-precision pre-trained model. **FT**: we use real

¹ <https://pypi.org/project/pytorchcv/>

Table 1: Comparisons on CIFAR-10/100 and ImageNet. We report the average and standard deviation of our method to show that our method is stable. We quantize both the weights and activations of the models to 4-bits and report the top1 accuracy.

| Data Set | Model | Real Data | | Data Free | |
|-----------|--------------|-----------|-------|-----------|---------------------|
| | | FP32 | FT | ZeroQ [4] | Ours |
| CIFAR-10 | ResNet-20 | 94.03 | 93.11 | 79.30 | 90.25 ± 0.30 |
| CIFAR-100 | ResNet-20 | 70.33 | 68.34 | 45.20 | 63.58 ± 0.23 |
| ImageNet | BN-VGG16 | 74.28 | 68.83 | 1.15 | 67.10 ± 0.29 |
| | ResNet-18 | 71.47 | 67.84 | 26.04 | 60.60 ± 0.15 |
| | Inception v3 | 78.80 | 73.80 | 26.84 | 70.39 ± 0.20 |

training data instead of fake data to fine-tune the quantized model by minimizing \mathcal{L}_2 . **ZeroQ**: a data-free post-training quantization method. We obtain the result from the publicly released code of ZeroQ [4].

We quantize both weights and activations to 4-bit and report the comparison results in Table 1. For CIFAR-10, our method achieves much higher accuracy than that of ZeroQ [4]. When the number of categories increases in CIFAR-100, our method suffers a much smaller degradation in accuracy compared with that of ZeroQ. The main reason is that, our method gains more prior knowledge from the full-precision model. These results demonstrate the effectiveness of our method on simple data sets with 4-bit quantization. For large scale and categories data set, such as ImageNet, existing data-free quantization methods suffer from severe performance degradation. However, our generated images contain abundant category information and similar distribution with real data. As a result, our method recovers the accuracy of quantized models significantly with the help of generated fake data and knowledge distillation on three typical networks. More experiments on different models and methods can be found in the supplementary material.

5.4 Ablation Studies

In this section, we first evaluate the effectiveness of each component in \mathcal{L}_1 and \mathcal{L}_2 . Second, we explore how fixed BNS affects our method. Then, we compare our method with different quantization methods. Last, we further study the effect of different stopping conditions. All the ablation experiments are conducted on the CIFAR-10/100 data sets.

Effect of different losses. To verify the effectiveness of different components in our method, we conduct a series of ablation experiments on CIFAR-100 with ResNet-20. Table 2 reports the top-1 accuracy of quantized models with different components of \mathcal{L}_1 . In this ablation experiment, we fine-tune quantized models with complete \mathcal{L}_2 . Since we do not use both CE loss and BNS loss, we have no way to optimize G , which means we use the fake data generated from the initial

Table 2: Effect of different loss functions of generator G . We quantize both the weights and activations of the models to 4-bits and report the top1 accuracy on CIFAR-100.

| Model | CE Loss | BNS Loss | Acc. (%) |
|----------------------|---------|----------|--------------|
| ResNet-20 (4-bit) | × | × | 30.70 |
| | ✓ | × | 54.51 |
| | × | ✓ | 44.40 |
| | ✓ | ✓ | 63.91 |

Table 3: Effect of different loss functions of Q . We keep the weights and activations of the models to 4-bits and report the top1 accuracy on CIFAR-100.

| Model | CE Loss | KD Loss | Acc. (%) |
|----------------------|---------|---------|--------------|
| ResNet-20 (4-bit) | ✓ | × | 55.55 |
| | × | ✓ | 62.98 |
| | ✓ | ✓ | 63.91 |

G to fine-tune the quantized model. In this case, the distribution of the fake data is far away from that of original data because the generator receives no guidance from the full-precision model. Therefore, the quantized model suffers from a large performance degradation. To utilize the knowledge in the full-precision model, we use CE loss to optimize G and achieve a better quantized model. In this case, the generator produces fake data that can be classified with high confidence by the full-precision model. Last, we combine CE loss and BNS loss with a coefficient and achieve the best result. The BNS loss encourages the generator to generate fake data that match the statistics encoded in full-precision model’s BN layers so that these fake data have a much similar distribution with real data. In summary, both CE loss and BNS loss contribute to better performance of the quantized model.

We further conduct ablation experiments to analyze the effectiveness of each component in \mathcal{L}_2 . Table 3 reports the top-1 accuracy of quantized models with different components of \mathcal{L}_2 . In this experiment, we optimize the generator with complete \mathcal{L}_1 . When only introducing KD loss, the quantized model receives knowledge from the full-precision model’s prediction and achieves 62.98% on top-1 accuracy. To use the additional label information, we combine BNS loss with CE loss. The resultant model achieves a 0.93% improvement on top-1 accuracy.

Effect of the fixed BNS. To verify the effectiveness of fixing batch normalization, we conduct ablation studies with ResNet-20 on CIFAR-10/100. The results are shown in Table 4. When we fix batch normalization statistics during fine-tuning, we narrow the statistics gap between the quantized model and the full-precision model. As a result, we achieve a much higher top-1 accuracy than that with standard batch normalization.

Table 4: Ablation experiments on the fixed BNS (FBNS). We keep the weights and activations of the models to be 4-bits and report the top1 accuracy on CIFAR-10/100. We use “w/o FBNS” to represent that we use fake data to fine-tune the quantized models without fixed BNS. Similarly, we use “w/ FBNS” to represent the fine-tuning process with fixed BNS.

| Data Set | w/o FBNS | w/ FBNS |
|-----------|----------|--------------|
| CIFAR-10 | 89.21 | 90.23 |
| CIFAR-100 | 61.12 | 63.91 |

Table 5: Comparison of different post-training quantization methods. We use real data sets as calibration sets and report the accuracy of the quantized model without fine-tuning.

| Data Set | Model | Method | W8A8 | W6A6 | W5A5 | W4A4 |
|-----------|----------------------|----------|--------------|--------------|--------------|--------------|
| CIFAR-10 | ResNet-20 (94.03) | MSE [39] | 93.86 | 93.10 | 91.38 | 81.59 |
| | | ACIQ [2] | 93.69 | 92.22 | 86.66 | 61.87 |
| | | KL [9] | 93.72 | 92.32 | 90.71 | 80.05 |
| | | Ours | 93.92 | 93.38 | 92.39 | 85.20 |
| CIFAR-100 | ResNet-20 (70.33) | MSE [39] | 70.11 | 66.87 | 60.49 | 27.11 |
| | | ACIQ [2] | 69.29 | 63.21 | 48.21 | 8.72 |
| | | KL [9] | 70.15 | 67.65 | 57.55 | 15.83 |
| | | Ours | 70.29 | 68.63 | 64.03 | 43.12 |
| ImageNet | ResNet-18 (71.47) | MSE [39] | 71.01 | 66.96 | 54.23 | 15.08 |
| | | ACIQ [2] | 68.78 | 61.15 | 46.25 | 7.19 |
| | | KL [9] | 70.69 | 61.34 | 56.13 | 16.27 |
| | | Ours | 71.43 | 70.43 | 64.68 | 33.25 |

5.5 Further Experiments

Comparisons with different post-training quantization methods. We compare different post-training quantization methods [39,2,9] on CIFAR-10/100 and ImageNet and show the results in Table 5. In this experiment, we use images from real data sets as calibration sets for quantized models with different post-training quantization methods. Specifically, we compare our quantization method with MSE(mean squared error), ACIQ, and KL(Kullback-Leibler), which are popular methods to decide the clip values of weight and activation. Our method shows much better performance than that of the other methods, which means it is more suitable in low-bitwidth quantization. Furthermore, the experimental results show that as the data set gets larger, the accuracy decreases. With the decrease of precision, all the quantization methods behave more poorly. Specifically, when the precision drops to 4-bit, the accuracy declines sharply.

Effect of two training strategies. We investigate the effect of two kinds of training strategies. 1) Training generator and quantized model in two steps. We first train the generator by minimizing the loss (9) until convergence. Then, we train the quantized model by minimizing the loss (10). 2) Training the generator

Table 6: Comparison of separate training and alternating training of G and Q .

| Training Strategy | Acc. (%) |
|----------------------|--------------|
| Separate Training | 61.81 |
| Alternating Training | 63.91 |

Table 7: Effect of different thresholds in the stopping condition of G .

| Threshold η (%) | 90.00 | 95.00 | 99.00 | 99.50 | w/o stopping condition |
|----------------------|-------|-------|-------|-------|------------------------|
| Acc. (%) | 57.20 | 57.56 | 59.09 | 59.67 | 63.91 |

and quantized model alternately in each iteration following Algorithm 1. From the results in Table 6, alternating training performs significantly better than separate training. Therefore, we use alternating training in other experiments.

Effect of different thresholds in stopping conditions. In this experiment, we stop the training of the generator if the classification accuracy of the full-precision model on fake data is larger than a threshold η . Table 7 reports the results of different thresholds η in stopping condition. When increasing the threshold, the generator is trained with quantized models for more epochs, and we get a better fine-tuning result. We achieve the best performance when we do not stop optimizing the generator. These results demonstrate that optimizing the generator and quantized model simultaneously increases the diversity of data, which is helpful for fine-tune quantized models.

6 Conclusion

In this paper, we have proposed a Generative Low-bitwidth Data Free Quantization scheme to eliminate the data dependence of quantization methods. First, we have constructed a knowledge matching generator to produce fake data for the fine-tuning process. The generator is able to learn the classification boundary knowledge and distribution information from the pre-trained full-precision model. Next, we have quantized the full-precision model and fine-tuned the quantized model using the fake data. Extensive experiments on various image classification data sets have demonstrated the effectiveness of our data-free method.

Acknowledgements

This work was partially supported by the Key-Area Research and Development Program of Guangdong Province 2018B010107001, Program for Guangdong Introducing Innovative and Entrepreneurial Teams 2017ZT07X183, Fundamental Research Funds for the Central Universities D2191240.

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: A system for large-scale machine learning. In: 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16). pp. 265–283 (2016)
2. Banner, R., Nahshan, Y., Hoffer, E., Soudry, D.: Acicq: Analytical clipping for integer quantization of neural networks (2018)
3. Banner, R., Nahshan, Y., Soudry, D.: Post training 4-bit quantization of convolutional networks for rapid-deployment. In: Proc. Adv. Neural Inf. Process. Syst. (2019)
4. Cai, Y., Yao, Z., Dong, Z., Gholami, A., Mahoney, M.W., Keutzer, K.: Zeroq: A novel zero shot quantization framework. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2020)
5. Cai, Z., He, X., Sun, J., Vasconcelos, N.: Deep learning with low precision by half-wave gaussian quantization. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2017)
6. Cao, J., Guo, Y., Wu, Q., Shen, C., Tan, M.: Adversarial learning with local coordinate coding. In: Proc. Int. Conf. Mach. Learn. (2018)
7. Cao, J., Mo, L., Zhang, Y., Jia, K., Shen, C., Tan, M.: Multi-marginal wasserstein gan. In: Proc. Adv. Neural Inf. Process. Syst. (2019)
8. Chen, H., Wang, Y., Xu, C., Yang, Z., Liu, C., Shi, B., Xu, C., Xu, C., Tian, Q.: Data-free learning of student networks. In: Proc. IEEE Int. Conf. Comp. Vis. (2019)
9. Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., Xiao, T., Xu, B., Zhang, C., Zhang, Z.: Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. arXiv preprint arXiv:1512.01274 (2015)
10. Choi, J., Wang, Z., Venkataramani, S., Chuang, P.I.J., Srinivasan, V., Gopalakrishnan, K.: Pact: Parameterized clipping activation for quantized neural networks. arXiv preprint arXiv:1805.06085 (2018)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2009)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
13. Esser, S.K., McKinstry, J.L., Bablani, D., Appuswamy, R., Modha, D.S.: Learned step size quantization. In: Proc. Int. Conf. Learn. Repren. (2020)
14. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proc. Adv. Neural Inf. Process. Syst. (2014)
15. Guo, Y., Zheng, Y., Tan, M., Chen, Q., Chen, J., Zhao, P., Huang, J.: Nat: Neural architecture transformer for accurate and compact architectures. In: Proc. Adv. Neural Inf. Process. Syst. (2019)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2016)
17. Hinton, G.E., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv:1503.02531 (2015)
18. Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R., Bengio, Y.: Binarized neural networks. In: Proc. Adv. Neural Inf. Process. Syst. (2016)

19. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
20. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., Kalenichenko, D.: Quantization and training of neural networks for efficient integer-arithmetic-only inference. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2018)
21. Jung, S., Son, C., Lee, S., Son, J., Han, J.J., Kwak, Y., Hwang, S.J., Choi, C.: Learning to quantize deep networks by optimizing quantization intervals with task loss. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2019)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) Proc. Int. Conf. Learn. Repren. (2015)
23. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
24. Lin, J., Gan, C., Han, S.: Defensive quantization: When efficiency meets robustness. arXiv preprint arXiv:1904.08444 (2019)
25. Lopes, R.G., Fenu, S., Starner, T.: Data-free knowledge distillation for deep neural networks. arXiv preprint arXiv:1710.07535 (2017)
26. Louizos, C., Reisser, M., Blankevoort, T., Gavves, E., Welling, M.: Relaxed quantization for discretized neural networks. In: Proc. Int. Conf. Learn. Repren. (2019)
27. Micaelli, P., Storkey, A.: Zero-shot knowledge transfer via adversarial belief matching. arXiv:1905.09768 (2019)
28. Mikolov, T., Karafiát, M., Burget, L., Černocký, J., Khudanpur, S.: Recurrent neural network based language model. In: Conference of the International Speech Communication Association (ISCA) (2010)
29. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957 (2018)
30. Nagel, M., Baalen, M.v., Blankevoort, T., Welling, M.: Data-free quantization through weight equalization and bias correction. In: Proc. IEEE Int. Conf. Comp. Vis. (2019)
31. Nayak, G.K., Mopuri, K.R., Shaj, V., Babu, R.V., Chakraborty, A.: Zero-shot knowledge distillation in deep networks. In: Proc. Int. Conf. Mach. Learn. (2019)
32. Nesterov, Y.E.: A method for solving the convex programming problem with convergence rate $o(1/k^2)$. In: Proceedings of the USSR Academy of Sciences. vol. 269, pp. 543–547 (1983)
33. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier gans. In: Proc. Int. Conf. Mach. Learn. (2017)
34. Paszke, A., Gross, S., Chintala, S., Chanan, G.: Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration. PyTorch: Tensors and dynamic neural networks in Python with strong GPU acceleration **6** (2017)
35. Rastegari, M., Ordonez, V., Redmon, J., Farhadi, A.: Xnor-net: Imagenet classification using binary convolutional neural networks. In: Proc. Eur. Conf. Comp. Vis. (2016)
36. Sak, H., Senior, A.W., Beaufays, F.: Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In: Conference of the International Speech Communication Association (ISCA). pp. 338–342 (2014)
37. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2018)
38. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (eds.) Proc. Int. Conf. Learn. Repren. (2015)

39. Sung, W., Shin, S., Hwang, K.: Resiliency of deep neural networks under quantization. arXiv preprint arXiv:1511.06488 (2015)
40. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2016)
41. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2016)
42. Yang, J., Shen, X., Xing, J., Tian, X., Li, H., Deng, B., Huang, J., Hua, X.s.: Quantization networks. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2019)
43. Yoo, J., Cho, M., Kim, T., Kang, U.: Knowledge extraction with no observable data. In: Proc. Adv. Neural Inf. Process. Syst. pp. 2701–2710 (2019)
44. Zeng, R., Huang, W., Tan, M., Rong, Y., Zhao, P., Huang, J., Gan, C.: Graph convolutional networks for temporal action localization. In: Proc. IEEE Int. Conf. Comp. Vis. (2019)
45. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization. In: Proc. Int. Conf. Learn. Repren. (2017)
46. Zhang, D., Yang, J., Ye, D., Hua, G.: Lq-nets: Learned quantization for highly accurate and compact deep neural networks. In: Proc. Eur. Conf. Comp. Vis. (2018)
47. Zhang, Y., Chen, H., Wei, Y., Zhao, P., Cao, J., et al.: From whole slide imaging to microscopy: Deep microscopy adaptation network for histopathology cancer image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 360–368 (2019)
48. Zhang, Y., Zhao, P., Wu, Q., Li, B., Huang, J., Tan, M.: Cost-sensitive portfolio selection via deep reinforcement learning. IEEE Trans. Knowl. Data Eng. (2020)
49. Zhao, R., Hu, Y., Dotzel, J., De Sa, C., Zhang, Z.: Improving neural network quantization without retraining using outlier channel splitting. In: Proc. Int. Conf. Mach. Learn. (2019)
50. Zhou, S., Wu, Y., Ni, Z., Zhou, X., Wen, H., Zou, Y.: Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. arXiv preprint arXiv:1606.06160 (2016)
51. Zhuang, B., Liu, L., Tan, M., Shen, C., Reid, I.: Training quantized neural networks with a full-precision auxiliary module. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2020)
52. Zhuang, B., Shen, C., Tan, M., Liu, L., Reid, I.: Towards effective low-bitwidth convolutional neural networks. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2018)
53. Zhuang, B., Shen, C., Tan, M., Liu, L., Reid, I.: Structured binary neural networks for accurate image classification and semantic segmentation. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2019)