Difficulty, Diversity, and Plausibility: Dynamic Data-Free Quantization

Cheeun Hong[1]* Sungyong Baik[3]* Junghun Oh[1] Kyoung Mu Lee[1,2]
[1] Dept. of ECE & ASRI, [2] IPAI, Seoul National University, Seoul, Korea
[3] Dept. of Data Science, Hanyang University, Seoul, Korea
cheeun914@snu.ac.kr, dsybaik@hanyang.ac.kr, dh6dh@snu.ac.kr, kyoungmu@snu.ac.kr

Abstract

Without access to the original training data, data-free quantization (DFQ) aims to recover the performance loss induced by quantization. Most previous works have focused on using an original network to extract the train data information, which is instilled into surrogate synthesized images. However, existing DFQ methods do not take into account important aspects of quantization: the extent of a computational-cost-and-accuracy trade-off varies for each image, depending on its task difficulty. To handle such varying trade-offs, several efforts have been made to dynamically allocate bit-widths for each image. Such dynamic quantization, however, remains challenging and unexplored in the data-free domain, because synthesized images of previous works fail to possess properties in natural test images that are crucial for learning the appropriate dynamic allocation policy: difficulty, its diversity, and its plausibility. By contrast, we propose a data-free quantization framework that is dynamic-friendly, by modeling varying extents of task difficulties with plausibility. We generate plausibly difficult images with soft labels, whose probabilities are allocated to a group of similar classes. Images with diverse and plausible difficulties enable us to train the framework to dynamically handle the varying trade-offs. Consequently, our framework achieves better accuracy-complexity Pareto front than existing data-free quantization approaches.

1. Introduction

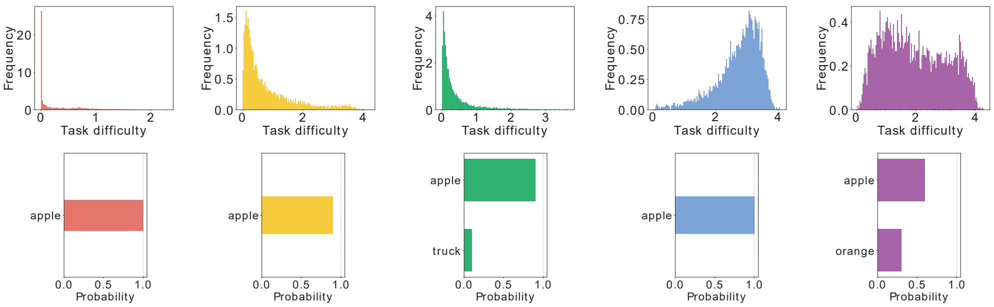
Recently, there has been a burst of interest in methodologies to improve the efficiency of neural networks. One of them is network quantization, which aims to reduce the bit-widths of weights or activation values. Among diverse lines of research in network quantization, quantization-aware training successfully obtains efficient networks without sacrificing accuracy by fine-tuning a network after quantization with the original training data [11, 17, 18, 48].

However, due to growing concerns on data privacy and security, the original training data are often inaccessible in real-world applications such as medical, military, and industrial domains [19, 37]. This raises a demand for quantization *without* the original training data, referred to as data-free quantization (DFQ). Recent works on DFQ aim to generate useful pseudo data by extracting and utilizing the knowledge of the train data concealed in the original pre-trained network, such as running statistics stored in the batch normalization layers [4, 7, 36, 46]. Then, they use the synthesized data to update the quantized network.

Yet, existing methods on DFQ overlook a crucial aspect of quantization: the trade-off between computational cost and accuracy varies for each image. Despite the fact that adopting different bit-width for images of different task difficulties, referred to as dynamic quantization, allows a better trade-off [15, 28, 39], existing methods adopt the same bit-width for all images. Nevertheless, the realization of such image-wise bit-width allocation is challenging without the assistance of the original training data. Natural test images are largely diverse in terms of classification difficulty; some images are easy to classify while others are confusing, making it hard to predict the top class from other similar classes. Thus, to allocate appropriate bit-widths for natural test images, the quantized networks should be trained with difficulty-varying samples (ranging from easy to difficult) beforehand. Unfortunately, existing works [4, 36, 46, 47] fail to generate such data, either by focusing only on easy images or by producing implausibly difficult images that mix irrelevant classes (e.g., apple and truck), which are not confusing to trained networks, as shown in Fig. 1. Consequently, this leads to a failure in dynamic quantization, where computational resources are not effectively distributed across images; more resources should be allocated to images more in need, in other words, difficult images.

On this basis, we propose the first data-free quantization framework, dubbed DDPQ, that considers **d**ifficulty with respect to **d**iversity and **p**lausibility thereof to fully exploit the benefits from dynamic quantization. As discussed above, the difficulty diversity (ranging from easy to diffi-

[*] equal contribution



(a) ZeroQ (b) IntraQ (c) Qimera (d) HAST (e) Ours

Figure 1. **Diverse difficulty and plausible difficulty in generated synthetic data of data-free quantization methods.** Our approach generates more *diverse* synthetic data in terms of classification difficulty. Also, the difficult images produced by our approach are *plausibly difficult* for the classification network, in other words, are confusing between similar classes (*e.g*., apple and orange). Below are sampled labels that guide the generation of an apple class image. Synthetic data for CIFAR-100 are produced using ResNet-20. Diversity and plausibility in difficulty are essential properties the synthetic data should exhibit to help learn the dynamic assignment of bit-widths.

cult) is needed for the network to learn how to allocate computational resources according to the difficulty. Furthermore, the plausibility of difficulty (e.g., car vs. truck) needs to be modeled such that the learned quantization scheme can be generalized to real images. To formulate both aspects in our image synthesis, we make use of readily available information: class similarity. Without access to the original dataset, class similarity information can be extracted from the classification layer of the original pre-trained network (*i.e*., the more similar the weights are, the more similar corresponding classes are), as illustrated in Fig. 2. Using the class similarity, we can achieve both difficulty diversity and plausibility by synthesizing images whose soft label weights are randomly allocated among similar classes.

Then, the generated images and soft labels are used to train the dynamic quantization network that adopts different layer-wise bit-widths for each image. Simply minimizing original cross-entropy loss functions will encourage the network to maximize accuracy, choosing the highest bit-width. To better handle the trade-off, we use a bit regularization loss that penalizes the computational resources (bit-FLOPs) of the dynamically quantized network.

The experimental results demonstrate the outstanding performance of DDPQ across various image classification networks, underlining the importance of diverse and plausible difficulty for sample generation in bridging the gap between dynamic quantization and data-free quantization.

2. Related Works

2.1. Data-free quantization (DFQ)

Network quantization has led to a dramatic reduction in computational resources without much compromise of network accuracy [2, 6, 8, 18, 22, 33, 48]. However, the arising concerns regarding data privacy and security have led to the lack of the access to the original training data [19, 37], which motivated the community to investigate network quantization methods that do not require the original training data, namely, DFQ. Nagel *et al*. [29] first propose to directly adapt the weights of the pre-trained model to reduce the quantization error, followed by different post-training quantization methods [1, 12, 44]. Subsequent works finetune the quantized network with synthetic data generated via random sampling [4] or from generative models [27, 36] to match the statistics of the pre-trained network.

The recent focus of DFQ methods is to generate synthetic samples that better match the real data. In specific, Zhang *et al*. [46] aim to better mimic the feature statistics of the real data, Zhong *et al*. [47] preserve the intraclass heterogeneity, and He *et al*. [14] apply ensemble technique to several compressed models to produce hard samples. More recently, Choi *et al*. [7] synthesize additional boundary-supporting samples, and Li *et al*. [24] focus the generation process on hard-to-fit samples, which are samples with low predicted probability on ground truth labels. Furthermore, Chen *et al*. [5] use a mixup knowledge distillation module to diversify synthetic samples.

However, these methods either neglect that the real data samples exhibit greatly different classification *difficulties* or neglect the semantic relationship between classes for difficult samples. In contrast, we generate samples of different difficulties with plausibility by considering the similarity between classes, which we find to be beneficial for dynamically allocating bit-widths for various samples.

2.2. Dynamic inference

Different input images encounter different task difficulties (*i.e*., they have different minimum required computational resources for processing). Accordingly, the dynamic computational resource adaptation for different input images has been widely studied in computer vision tasks via controlling the quantization bit-width [15, 16, 28, 39], or number of channels [30, 40], convolutional layers [21, 26, 41–43]. Dynamic quantization has been recently investigated, as it grants a better trade-off between accuracy and computational complexity. Specifically, Li *et al*. [28] effectively allocate bit-widths to different inputs and layers on image classification task. Dynamic bit allocation has also been proven effective on other tasks, such as video recognition [38] and image restoration [15, 39].

However, these approaches require the training of dynamic networks with the original or even additional training dataset. Unfortunately, these datasets are not always available in real-world applications. To cope with such issues, we present a practical solution to achieve dynamic quantization without access to the original training dataset.

3. Background

In the following sections, we introduce the general procedure for data-free quantization (Sec. 3.1) and dynamic quantization network (Sec. 3.2) and then elaborate on our DDPQ. To fully exploit the benefits of dynamic quantization, we first generate fake images with diverse levels of difficulty. Also, we guarantee that generated difficult images are plausibly difficult, in other words, confusing between relevant classes (Sec. 4.1). Then, to achieve a better trade-off between computational resources and accuracy, we encourage the dynamic quantization network to assign higher bit-widths to images in need (Sec. 4.2). Our overall generation scheme is summarized in Algorithm 1.

3.1. Data-free quantization (DFQ)

DFQ aims to obtain an accurate quantized network $\mathcal{Q}$ from a pre-trained floating-point (FP) 32-bit network $\mathcal{P}$ without the assistance of the original training data. DFQ approaches [4, 24, 46, 47] commonly adopt a two-stage approach: (1) synthetic data generation and (2) quantization-aware training with the synthetic data. First, synthetic data pairs $\{\boldsymbol{I}_i, \boldsymbol{y}_i\}_{i=1}^{N}$ are initialized by randomly initializing a synthetic image $\boldsymbol{I}_i$ with Gaussian noise and its label $\boldsymbol{y}_i$ as arbitrary class (*i.e*., one-hot label). Then, synthetic images are generating by optimizing the following objective:

$$\mathcal{L}^{G} = \frac{1}{N}\sum_{i=1}^{N}(\mathcal{L}_{ce}(\mathcal{P}(\boldsymbol{I}_i), \boldsymbol{y}_i) + \mathcal{L}_{bns}(\boldsymbol{I}_i)), \qquad (1)$$

where the first term $\mathcal{L}_{ce}$ is the cross-entropy loss that guides each $\boldsymbol{I}_i$ to follow its label $\boldsymbol{y}_i$ w.r.t. $\mathcal{P}$. The second term $\mathcal{L}_{bns}$ encourages the intermediate statistics (mean and standard deviation) produced by $\mathcal{P}$ to match the statistics stored in the batch normalization (BN) layers of $\mathcal{P}$.

Subsequently, the obtained synthetic samples are used to fine-tune the quantized network $\mathcal{Q}$. $\mathcal{Q}$ is derived by quantizing weights and features of convolutional layers of $\mathcal{P}$. The input feature $\boldsymbol{X}$ of each convolutional layer is quantized to $\boldsymbol{X}_q$ with bit-width b by

$$\boldsymbol{X}_q \equiv q_b(\boldsymbol{X}) = \lfloor \frac{\text{clamp}(\boldsymbol{X}, l, u) - l}{s(b)} \rceil \cdot s(b) + l, \qquad (2)$$

where $s(b) = (u - l)/(2^b - 1)$, l and u are learnable scale parameters. First, $\text{clamp}(\cdot, l, u)$ truncates $\boldsymbol{X}$ into the range of $[l, u]$ and then scaled to $[0, 2^b - 1]$. The features are then rounded to the nearest integer with $\lfloor \cdot \rceil$, and the integers are re-scaled back to range $[l, u]$. Following the common practice [24, 47], we use a layer-wise quantization function for features and a channel-wise quantization function for weights. The quantization parameters l, u, and the weights of $\mathcal{Q}$ are trained by the optimization as follows:

$$\mathcal{L}^{Q} = \frac{1}{N}\sum_{i=1}^{N}(\mathcal{L}_{ce}(\mathcal{Q}(\boldsymbol{I}_i), \boldsymbol{y}_i) + \mathcal{L}_{kd}(\mathcal{P}(\boldsymbol{I}_i), \mathcal{Q}(\boldsymbol{I}_i))), \qquad (3)$$

where $\mathcal{L}_{kd}$ is Kullback-Leibler divergence loss.

3.2. Dynamic quantization

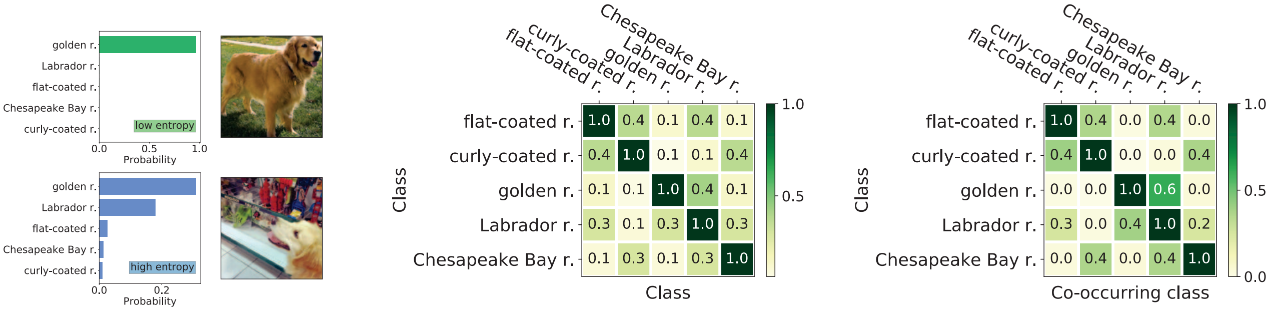
To adaptively allocate bit-widths to input images, most existing works [15, 28, 39] set multiple quantization bit-width candidates for each quantization function. Then, a bit selector is used to predict the image-wise probability for each candidate bit-width, where the bit-width of the highest predicted probability is selected during inference. We adopt commonly used MLP architecture as the bit selector module for bit-width candidates $\{b_m\}_{m=1}^{M}$, where we set $M = 3$ in this work. To minimize the additional overhead, we use a lightweight bit selector that consists of a two-layer MLP followed by a Softmax operation, following Liu *et al*. [28]:

$$p_{b_m}(\boldsymbol{X}) = \text{Softmax}(\text{MLP}(\boldsymbol{X})), \qquad (4)$$

where MLP consists of an average pooling layer followed by two fc layers with a dropout layer in between. During inference, for each input image and layer, a quantization function of bit-width with the highest probability is selected. Since selecting the max probability is a non-differentiable operation, we use a straight-through estimator [3] to make the process differentiable:

$$\boldsymbol{X}_q = \begin{cases} \text{argmax}_{q_{b_m}(\boldsymbol{X})} p_{b_m} & \text{forward}, \\ \sum_m^M q_{b_m}(\boldsymbol{X}) \cdot p_{b_m}(\boldsymbol{X}) & \text{backward}. \end{cases} \qquad (5)$$

Also, as dynamic quantization for weights requires storing the weight of the largest bit candidate, which occupies a large storage, we apply a bit selector to choose bit-widths for activations only. In this work, we train our bit selector using the synthetic images and labels, in order to learn the bit-width allocation policy for each layer and image *without* access to the original training dataset.



(a) Easy/difficult sample

(b) Classification weight similarity

(c) Class co-occurrences in prediction

Figure 2. **Motivation of class similarity-based generation.** (a) While the prediction output of an easy sample is nearly one-hot, that of a confusing sample is a soft distribution over similar classes. (b-c) Similar classes can be obtained without access to the original data by measuring the similarity between pre-trained classification weights. The measured similarity is correlated to the class co-occurrences in the prediction of the original data. For simplicity, we visualize five retriever classes of ImageNet.

4. Proposed Method

4.1. Plausibly and diversely difficult data generation

Difficulty. It is acknowledged that different images have different quantization sensitivities [15,28,39], which can be defined as the degree of quantization at which the correct prediction can remain intact. Samples whose predictions remain accurate even after intense (low-bit) quantization are considered less sensitive to quantization. Intuitively, allocating higher bit-widths to quantization-sensitive samples can lead to finding a better trade-off between efficiency and accuracy. However, directly measuring the quantization sensitivity is practically infeasible, as it requires multiple inferences of each sample with different bit-width networks that are separately trained. Instead, since quantization sensitivity is related to the classification difficulty of samples, we estimate quantization sensitivity by means of classification difficulty. The classification difficulty can be modeled with entropy [9], which is used to estimate the uncertainty of the model prediction. In this work, we measure entropy to monitor the difficulty of the generated samples.

Difficulty diversity. To effectively allocate bit-widths to images, we are motivated to generate pseudo images with diverse levels of difficulties. One simple approach to diversify the difficulty in the generated pseudo data is to assign a randomly sampled soft label to an image instead of an arbitrary class y [7,47]. However, simple random sampling does not provide a control over the diversity of difficulty. Furthermore, it does not take realistic difficulty into account, unable to guide the dynamic quantization framework to effectively allocate resources for new real images.

Difficulty plausibility. To generate data that facilitate the training of dynamic quantization framework, it is important to generate images whose difficulties are plausible (i.e., similar to real images). According to our observation in Fig. 2a, the prediction of images in the original data reflects the relations between classes; a difficult sample for the classification network is confusing between similar classes. Thus, our goal is to assign the images with soft labels that consider the similarity between classes. To obtain the similarity between classes without access to the original data, we exploit the classification layer of the original pre-trained network, as each class weight vector is regarded as a useful estimated prototype of the class [34]. To validate our assumption that class weight similarity can be used to measure the similarity between classes of the original data, we compare the similarity of two class weight vectors and the actual co-occurrences of the two classes in the prediction output of the original data. As visualized in Fig. 2, the similarity matrix of class weights and class co-occurrences are closely correlated. Thus, we formulate the similarity between j-th class and k-th class as follows:

$$S(j,k) = \boldsymbol{W}_j \times \boldsymbol{W}_k^T, \qquad j,k = 1,...,C, \tag{6}$$

where $\boldsymbol{W}$ denotes the weight vectors of the classification layer and $\boldsymbol{W}_j$ denotes the weight vector that corresponds to the j-th class. Motivated by the previous observations, we assign the probability scores of soft labels only to top-K similar classes, obtained using the similarity in Eq. (6). Given an arbitrarily sampled class y for the i-th synthetic image $\boldsymbol{I}_i$, the soft label $\boldsymbol{y}_i \in \mathbb{R}^C$ is formulated as follows:

$$y_{i,k} = \begin{cases} z_k & \text{if } S(y,k) \in \text{top-}K(\{S(y,1), \cdots, S(y,C)\}), \\ 0 & \text{else}, \end{cases} \tag{7}$$

where $y_{i,k}$ is the k-th probability value of $\boldsymbol{y}_i$ and the probability score $\boldsymbol{z} = [z_1, z_2, \cdots, z_K]^T$ are randomly sampled from Dirichlet distribution $\boldsymbol{z} \sim \text{Dir}(K, \boldsymbol{\alpha})$, in which simply set $\boldsymbol{\alpha}$ as $\mathbf{1}$. The difficulty of each sample can be controlled by K and the entropy of the assigned soft label. For instance, if an image is assigned with a soft label with high entropy and large K, the image will likely be optimized to be confusing between many similar classes, making it a difficult sample. To generate images of diverse difficulty, we randomly sample the soft label for similar classes, and we generate r ratio of samples optimized with top-K similar classes and $1{-}r$ ratio of samples with top-1 class, with a diversifying ratio hyper-parameter r. Overall, the synthetic images are optimized with the objective function as follows:

$$\mathcal{L}^G = \frac{1}{N}\sum_{i=1}^{N}(\mathcal{L}_{bns}(\boldsymbol{I}_i) + \beta\mathcal{L}_{ce}(\mathcal{P}(\boldsymbol{I}_i), \boldsymbol{y}_i)), \quad (8)$$

where β balances the losses. Given *diverse*, *plausibly* difficult and easy images, we can now facilitate the training of dynamic quantization framework.

4.2. Quantization-aware training

Bit regularization loss. We train the bit selector and the quantized network $\mathcal{Q}$ using the synthesized plausibly and diversely difficult training pairs $\{\boldsymbol{I}_i, \boldsymbol{y}_i\}_{i=1}^{N}$. Solely minimizing the original cross-entropy loss function can lead the bit selector to choose the highest bit-width to maximize the accuracy. Thus, to find a better balance between computational cost and accuracy, the total bit-FLOPs selected by the bit selector are regularized, such that:

$$\mathcal{L}_{br} = \frac{1}{N}\sum_{i=1}^{N} max(\frac{B^{\mathcal{Q}}(\boldsymbol{I}_i)}{B_{tar}}, 1), \quad (9)$$

where $B^{\mathcal{Q}}(\cdot)$ is the total bit-FLOPs of the quantized network $\mathcal{Q}$ assigned to each synthetic image $\boldsymbol{I}_i$, which is calculated as the weighted number of operations by the weight bit-width and the activation bit-width. The assigned bit-width of each feature is determined by $\text{argmax}_{b_m} p_{b_m}(\boldsymbol{X})$ from Eq. (5). Also, N denotes the mini-batch size, and B_{tar} is the target bit-FLOPs. In this work, we set B_{tar} as the bit-FLOPs of the fixed {5,4}-bit model to compare with existing methods. The overall objective function of the data-free dynamic quantization framework is:

$$\mathcal{L}^{\mathcal{Q}} = \gamma\mathcal{L}_{br} + \frac{1}{N}\sum_{i=1}^{N}(\mathcal{L}_{ce}(\mathcal{Q}(\boldsymbol{I}_i), \boldsymbol{y}_i) + \mathcal{L}_{kd}(\mathcal{P}(\boldsymbol{I}_i), \mathcal{Q}(\boldsymbol{I}_i))), \quad (10)$$

where γ is a hyper-parameter that balances the bit regularization loss with the other loss terms.

5. Experiments

In this section, the proposed framework DDPQ is evaluated with various image classification networks to validate its effectiveness. We first describe our experimental settings (Sec. 5.1) and evaluate our framework on CIFAR-10/100 (Sec. 5.2) and ImageNet dataset (Sec. 5.3). Then, we present ablation experiments (Sec. 5.4) and visualization results (Sec. 5.5) that demonstrate the effect of our scheme.

Algorithm 1 Data generation process of DDPQ

```
Input: Pretrained FP 32-bit network $\mathcal{P}$, iterations $T$.
Output: Synthetic data pair $\{\boldsymbol{I}_i, \boldsymbol{y}_i\}_{i=1}^{N}$.
  for $i = 1, \cdots, N$ do
      Randomly sample class $y$
      Obtain top-K similar classes of $y$ with Eq. (6)
      Given similar classes, assign soft label $\boldsymbol{y}_i$ with Eq. (7)
      Initialize $\boldsymbol{I}_i \sim \mathcal{N}(0, 1)$
      for $t = 1, \cdots, T$ do
          Given $\boldsymbol{y}_i$ and $\boldsymbol{I}_i$, calculate $\mathcal{L}^G$ with Eq. (8)
          Update $\boldsymbol{I}_i$ by minimizing $\mathcal{L}^G$
```

5.1. Implementation details

Models. To validate the flexibility of our framework, we perform evaluation with the representative classification models, ResNet-20 [13], ResNet-18 [13], and MobileNetV2 [35] using CIFAR-10/100 [23] and ImageNet (ILSVRC12) [10] datasets. For the dynamic bit-width allocation, the bit selector is located after the first two convolutional layers of the network, which are quantized with a fixed bit-width, the highest bit among the M candidates (*i.e.*, b_M). In this work, to compare with the fixed 5-bit quantization methods, we set the bit-width candidates near the comparison bit, namely {4,5,6}. Experiments on different bit-width candidates are provided in the supplementary material. Quantization of each bit-width is done using the simple asymmetric uniform quantization function with scale parameters following [17, 24, 36, 47].

Generation details. All our experiments are implemented using PyTorch [31]. For data generation, synthetic images are updated for 4,000 iterations with a batch size of 64 using the Adam [20] optimizer with a learning rate of 0.5 decayed by the rate of 0.1 when the loss stops decreasing for 100 iterations. A total of 5,120 images are generated for both CIFAR-10/100 and ImageNet, which is highly lightweight compared to the original dataset (*e.g.*, ∼1,000,000 images for ImageNet). For hyperparameters, we set β=0.1, K=2 and r=0.5.

Training details. For the dynamic bit-width allocation, the bit-selector is initialized to output the highest probability for the target bit-width (*e.g.*, 5-bit for ∼5 mixed-precision (MP)). Using the 5,120 synthetic images and soft labels, we fine-tune our dynamic quantization framework for 400 epochs with a batch size of 256 for CIFAR-10/100

Method	G.	Bit-width	Bit-FLOPs (%)	Top-1 (%)
Baseline	-	32	100.00	94.03
Real Data	-	5	3.52	93.96
GDFQ (ECCV'20)	✓	5	3.52	93.38
Qimera (NeurIPS'21)	✓	5	3.52	93.46
AdaDFQ (CVPR'23)	✓	5	3.52	93.81
ZeroQ (CVPR'20)	✗	5	3.52	91.38†
DSG (CVPR'21)	✗	5	3.52	92.73
IntraQ (CVPR'22)	✗	5	3.52	92.78
HAST (CVPR'23)	✗	5	3.52	93.43†
DDPQ (Ours)	✗	~5 MP	**3.47**±0.06	**93.87**±0.07
Real Data	-	4	2.62	91.52
SQuant (ICLR'22)	-	4	2.64	92.24
GDFQ (ECCV'20)	✓	4	2.62	90.25‡
Qimera (NeurIPS'21)	✓	4	2.62	91.26
AdaDFQ (CVPR'23)	✓	4	2.62	92.31
TexQ (NeurIPS'23)	✓	4	2.62	92.68
ZeroQ (CVPR'20)	✗	4	2.62	84.68‡
DSG (CVPR'21)	✗	4	2.62	88.74‡
GZNQ (CVPRW'21)	✗	4	2.62	91.30‡
IntraQ (CVPR'22)	✗	4	2.62	91.49‡
HAST (CVPR'23)	✗	4	2.62	92.36
DDPQ (Ours)	✗	~4 MP	**2.60**±0.01	**92.76**±0.05

(a) Results on CIFAR-10

Method	G.	Bit-width	Bit-FLOPs (%)	Top-1 (%)
Baseline	-	32	100.00	70.33
Real Data	-	5	3.52	70.05
GDFQ (ECCV'20)	✓	5	3.52	66.12
Qimera (NeurIPS'21)	✓	5	3.52	69.02
AdaDFQ (CVPR'23)	✓	5	3.52	**69.93**
ZeroQ (CVPR'20)	✗	5	3.52	65.89†
DSG (CVPR'21)	✗	5	3.52	67.65
IntraQ (CVPR'22)	✗	5	3.52	68.06
HAST (CVPR'23)	✗	5	3.52	69.00†
DDPQ (Ours)	✗	~5 MP	**3.44**±0.01	69.74±0.03
Real Data	-	4	2.62	66.80
SQuant (ICLR'22)	-	4	2.64	63.96
GDFQ (ECCV'20)	✓	4	2.62	63.58‡
Qimera (NeurIPS'21)	✓	4	2.62	65.10
AdaDFQ (CVPR'23)	✓	4	2.62	66.81
TexQ (NeurIPS'23)	✓	4	2.62	67.18
ZeroQ (CVPR'20)	✗	4	2.62	58.42‡
DSG (CVPR'21)	✗	4	2.62	62.36‡
GZNQ (CVPRW'21)	✗	4	2.62	64.37‡
IntraQ (CVPR'22)	✗	4	2.62	64.98‡
HAST (CVPR'23)	✗	4	2.62	66.68
DDPQ (Ours)	✗	~4 MP	**2.56**±0.01	**67.58**±0.25

(b) Results on CIFAR-100

† reproduced using the official code ‡ cited from [47]

Table 1. **Results of ResNet-20 on CIFAR-10/100.** Bit-FLOPs (%) show the relative bit-FLOPs compared to the FP baseline ResNet20 (42.20G). We measure the average bit-FLOPs for 32 × 32 sized images. G. indicates generator-based methods.

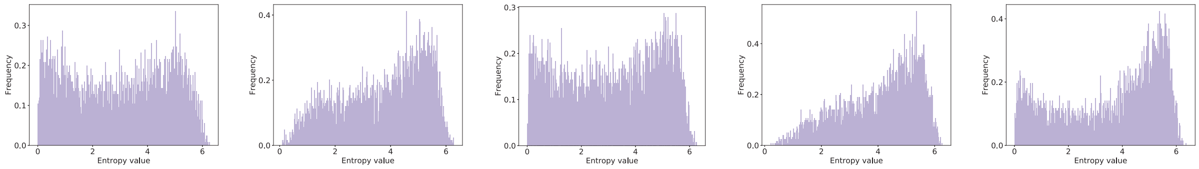
and 150 epochs with a batch size of 16 for ImageNet. The parameters are updated with an SGD optimizer with Nesterov using an initial learning rate of 10^{-5} decayed by 0.1 every 100 epochs for CIFAR-10/100 and 10^{-6} decayed every 30 epochs for ImageNet. The bit selector parameters are updated similarly but with the initial learning rate of 10^{-3} and 10^{-4} for CIFAR-10/100 and ImageNet. To balance different loss terms in Eq. (10), γ is set to 100 throughout the experiments. Also, during training, we observe that the easy samples are learned quickly by the dynamic quantization framework, giving minimal contribution to the training afterward. Thus, inspired by previous approaches [25,45], we additionally adopt mixup to easily fit samples (*i.e.*, samples with lower cross-entropy loss at saturation). We blend two easily-fit images to generate a hard-to-fit image. We apply Mixup finetuning from 50% training epochs to $p\%$ of samples with low cross-entropy loss, where we set p as 25%.

5.2. CIFAR-10/100

To evaluate the efficacy of our proposed framework, we compare the results with the existing data-free quantization methods on CIFAR-10/100 [23] with a representative classification network, ResNet-20 [13]. Specifically, we compare DDPQ with the early DFQ method (ZeroQ [4], GDFQ [36]) and more recent noise optimization-based methods (DSG [46], GZNQ [14], IntraQ [47], HAST [24]). Also, we compare our framework with generator-based methods (Qimera [7], AdaDFQ [32], TexQ [5]) even though these methods utilize additional generator network during training. The accuracy of previous arts are obtained from papers or models reproduced with official codes. To make a fair comparison between existing approaches that allocate a fixed universal bit-width and our dynamic mixed bit-width approach, we compare frameworks with respect to the average bit-FLOPs consumed on the validation set. As shown in Tab. 1, DDPQ achieves state-of-the-art accuracy on CIFAR-10 even with fewer bit-FLOPs (Tab. 1a) and achieves the best or second best accuracy on CIFAR-100 with fewer bit-FLOPs (Tab. 1b). Notably, in several settings, our data-free method outperforms the results of the original data, which demonstrates the superiority of our difficulty-diverse data and that dynamic quantization using our data allows a better complexity-accuracy trade-off.

5.3. ImageNet

We evaluate our framework on a further large-scale dataset, ImageNet [10], on widely adopted classification models: ResNet-18 [13] and MobileNetV2 [35]. As shown in Tab. 2a, compared to other DFQ methods on ResNet-18, our framework uses fewer bit-FLOPs and still achieves similar or better accuracy. We note that our approach achieves high accuracy without utilizing an additional generator network to synthesize images. According to Tab. 2b, on MobileNetV2, DDPQ achieves the highest or second highest accuracy with a similar amount of bit-FLOPs. Overall, our data-free dynamic framework reduces the accuracy gap with the network trained using the original ImageNet dataset.



(a) K=2, r=0.5 (b) K=2, r=1.0 (c) K=3, r=0.5 (d) K=3, r=1.0 (e) K=10, r=0.5

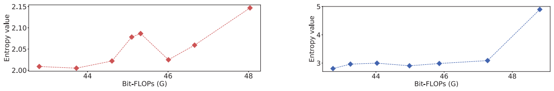
Figure 3. **Classification difficulty distribution of synthetic data** using different K and r for ImageNet using ResNet-18.

5.4. Ablation study

Effect of our data generation. To further verify the effectiveness of our generated samples, we conduct an ablation study on our generation scheme that promotes diverse difficulty and plausible difficulty. As shown in Tab. 3, while our generated data are also helpful for fixed quantization (+0.39% / +0.51%), the benefits are maximally excavated with a dynamic quantization framework (+1.12% / +1.42%) with fewer bit-FLOPs. Also, while diversity alone leads to a small accuracy gain (+0.05% / +0.60%), plausible and diverse difficulty results in larger gain (+0.34% / +1.11%). The results indicate that both plausibility and diversity in difficulty are crucial for obtaining an accurate dynamic quantization network.

Moreover, we validate the effectiveness of the dynamic quantization framework (DQ). As in Tab. 3, dynamic quantization alone provides a slightly better trade-off with the baseline data (+0.78% / +0.31%). Nevertheless, when we train the dynamic quantization framework with our synthetic data, the accuracy increases by a large margin (+1.12% / +1.42%). The results imply that our difficulty-diverse data generation, along with the dynamic framework, is effective in terms of bit-FLOPs-accuracy trade-off, justifying the dynamic allocation of different computations to images of different difficulty exhibiting different trade-offs.

Compatibility with dynamic quantization. We present the effectiveness of our generation scheme for dynamic quantization by comparing it with the existing data generation schemes for data-free quantization. For a fair comparison, we apply the dynamic quantization framework to noise optimization-based data-free quantization methods that are open-source. Tab. 4 shows that the existing data generation schemes for DFQ fail to effectively allocate bit-widths compared to ours. While dynamic quantization (DQ) brings trivial or no gain on previous data-generation methods, our method fully benefits from dynamic quantization, achieving performance closer to the performance obtained with real data, with fewer bit-FLOPs. The results indicate that our synthetic data successfully simulates and controls realistic/plausible difficulty of different levels, and thus enables dynamic quantization to effectively assign computational costs for unseen real input images without real data.



(a) Test images (b) Generated train images

Figure 4. **Bit-width allocation for images of different difficulty.** For a given sample of test images and generated train images, assigned bit-FLOPs and output entropy are plotted. Samples with higher entropy tend to be assigned with overall higher bit-FLOPs.

Effect of the hyper-parameters. We investigate the effect of our selection on hyperparameters: the number of similar classes K and diversifying ratio r for data generation. As presented in Fig. 3, our choice of K and r controls the difficulty distribution of the generated data. Based on the results of Tab. 5, we find that the diversifying ratio of 0.50 with two or three similar classes produces results of better trade-off, while a too-large r or K results in an accuracy drop. This is because a large r and K will lead to a set of generated images that mostly consist of difficult images. It is important for the dynamic network to see images of diverse difficulties (both easy and difficult) during training.

5.5. Visualizations

We compare the difficulty diversity of different data-generation methods in Fig. 1. Compared to existing methods, our approach produces more diverse data in terms of difficulty, which hints that the diversity in difficulty serves as a key factor for data-free dynamic quantization. Moreover, to better demonstrate the efficacy of our dynamic quantization and plausibly difficult sample generation, we visualize the bit allocation results of different difficulty images in Fig. 4. One can observe that the overall samples with high (low) entropy tend to be assigned with more (less) bit-FLOPs. Also, the results show that our synthetic images can fairly model the difficulty of real test images.

6. Conclusion

In this paper, we present the first dynamic data-free quantization method that allocates different bit-widths to images without any access to the original training data. Despite its advantages, dynamic quantization remains unexplored in data-free settings due to the lack of difficulty-

Method	G.	Bit-width	Bit-FLOPs (%)	Top-1 (%)
Baseline	-	32	100.00	71.47
Real Data	-	5	3.56	70.41
GDFQ (ECCV'20)	✓	5	3.56	66.82‡
Qimera (NeurIPS'21)	✓	5	3.56	69.29
AdaDFQ (CVPR'23)	✓	5	3.56	70.29
ZeroQ (CVPR'20)	✗	5	3.56	69.65‡
DSG (CVPR'21)	✗	5	3.56	69.53‡
IntraQ (CVPR'22)	✗	5	3.56	69.94‡
HAST (CVPR'23)	✗	5	3.56	70.04†
DDPQ (Ours)	✗	$\sim$5 MP	**3.40**±0.08	**70.41**±0.01
Real Data	-	4	2.54	67.89
SQuant (ICLR'22)	-	4	2.58	66.14
GDFQ (ECCV'20)	✓	4	2.54	60.60‡
Qimera (NeurIPS'21)	✓	4	2.54	63.84
AdaDFQ (CVPR'23)	✓	4	2.54	66.53
TexQ (NeurIPS'23)	✓	4	2.54	**67.73**
ZeroQ (CVPR'20)	✗	4	2.54	60.68‡
DSG (CVPR'21)	✗	4	2.54	60.12‡
GZNQ (CVPRW'21)	✗	4	2.54	64.50‡
IntraQ (CVPR'22)	✗	4	2.54	66.47‡
HAST (CVPR'23)	✗	4	2.54	66.91
DDPQ (Ours)	✗	$\sim$4 MP	**2.49**±0.01	67.47±0.05

(a) Results of ResNet-18

Method	G.	Bit-width	Bit-FLOPs (%)	Top-1 (%)
Baseline	-	32	100.0	73.03
Real Data	-	5	12.88	72.01
GDFQ (ECCV'20)	✓	5	12.88	68.14‡
Qimera (NeurIPS'21)	✓	5	12.88	70.45
AdaDFQ (CVPR'23)	✓	5	12.88	71.61
ZeroQ (CVPR'20)	✗	5	12.88	70.88‡
DSG (CVPR'21)	✗	5	12.88	70.85‡
IntraQ (CVPR'22)	✗	5	12.88	71.28‡
HAST (CVPR'23)	✗	5	12.88	71.72
DDPQ (Ours)	✗	$\sim$5 MP	**12.87**±0.04	**71.88**±0.05
Real Data	-	4	11.02	67.90
SQuant (ICLR'22)	-	4	11.05	22.07
GDFQ (ECCV'20)	✓	4	11.02	51.30‡
Qimera (NeurIPS'21)	✓	4	11.02	61.62
AdaDFQ (CVPR'23)	✓	4	11.02	65.41
TexQ (NeurIPS'23)	✓	4	11.02	**67.07**
ZeroQ (CVPR'20)	✗	4	11.02	59.39‡
DSG (CVPR'21)	✗	4	11.02	59.04‡
GZNQ (CVPRW'21)	✗	4	11.02	53.53‡
IntraQ (CVPR'22)	✗	4	11.02	65.10‡
HAST (CVPR'23)	✗	4	11.02	65.60
DDPQ (Ours)	✗	$\sim$4 MP	**11.02**±0.02	66.92±0.46

(b) Results of MobileNetV2

† reproduced using the official code ‡ cited from [47]

Table 2. **Results of ResNet-18/MobileNetV2 on ImageNet.** Bit-FLOPs (%) show the relative bit-FLOPs compared to the FP baseline (1862.54G for ResNet-18 and 314.12G for MobileNetV2). We measure the average bit-FLOPs for 224 × 224 sized images.

DQ	DD	PD	Bit-FLOPs (%)	Top-1 (%)
			2.62	66.46
	✓	✓	2.62	66.85
✓			2.59	67.24
✓	✓		2.56	67.29
✓	✓	✓	**2.56**	**67.58**

(a) ResNet-20 with CIFAR-100

DQ	DD	PD	Bit-FLOPs (%)	Top-1 (%)
			2.54	66.05
	✓	✓	2.54	66.56
✓			2.39	66.36
✓	✓		2.49	66.96
✓	✓	✓	**2.49**	**67.47**

(b) ResNet-18 with ImageNet

Table 3. **Ablation study of each attribute** done on $\sim$4MP. "DQ", "DD", and "PD" respectively indicates dynamic quantization, difficulty-diverse and plausible difficulty-based generation. "DD" with no "PD" denotes that difficulty-diverse data are generated from arbitrary classes instead of similar classes.

Method	DQ	Bit-FLOPs (%)	Top-1 (%)
ZeroQ+IL	✗	2.62	63.97‡
DSG+IL	✗	2.62	62.62‡
IntraQ	✗	2.62	64.98‡
HAST	✗	2.62	66.68
DDPQ	✗	2.62	66.85
ZeroQ+IL	✓	2.56	66.22
DSG+IL	✓	2.61	66.08
IntraQ	✓	2.61	66.34
HAST	✓	2.59	66.55
DDPQ (Ours)	✓	2.56	**67.58**

(a) ResNet-20 with CIFAR-100

Method	DQ	Bit-FLOPs (%)	Top-1 (%)
ZeroQ+IL	✗	2.54	63.38‡
DSG+IL	✗	2.54	63.11‡
IntraQ	✗	2.54	66.47
HAST	✗	2.54	66.91
DDPQ	✗	2.54	66.56
ZeroQ+IL	✓	2.53	65.57
DSG+IL	✓	2.47	65.02
IntraQ	✓	2.48	65.42
HAST	✓	2.47	66.88
DDPQ (Ours)	✓	2.49	**67.47**

(b) ResNet-18 with ImageNet

Table 4. **Comparison of different data generation methods on dynamic quantization (DQ).**

r	K	Static Q		Dynamic Q	
		Bit-FLOPs	Top-1	Bit-FLOPs	Top-1
0.25	2	2.54	**67.10**	2.47	67.16
0.50	2	2.54	66.56	2.49	**67.47**
0.50	3	2.54	66.66	2.50	67.45
0.50	10	2.54	66.50	2.54	67.01
1.00	2	2.54	66.33	2.45	66.50
1.00	3	2.54	65.78	2.47	65.02

Table 5. **Effect of the diversity ratio in data generation** of 4-bit ResNet-18 on ImageNet.

diverse synthesized data. To address this, we generate difficulty-varied images by assigning realistic soft labels based on class similarity, ensuring plausible and diverse image generation. Experimental results show that our approach consistently improves accuracy with lower computational costs.

Limitation Applying our method directly to other vision tasks (e.g., object detection) may be challenging, as it relies on class information stored in the classification layer to control the difficulty. Nevertheless, our core principle of promoting diverse and plausible difficulty is adaptable, with task-specific adjustments to the define difficulty. For example, in image restoration tasks, 'difficulty' can be based on restoration complexity, measured by metrics like average gradient magnitude or edge density.

Acknowledgment This work was supported in part by the IITP grants [No.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University), No. 2021-0-02068, and No.2023-0-00156], the NRF grant [No. 2021M3A9E4080782] funded by the MSIT, and the Industrial Technology Alchemist Project [No. RS-2024-00432410] funded by MOTIE, Korea.

References

[1] Shipeng Bai, Jun Chen, Xintian Shen, Yixuan Qian, and Yong Liu. Unified data-free compression: Pruning and quantization without fine-tuning. In *ICCV*, 2023. 2

[2] Ron Banner, Yury Nahshan, and Daniel Soudry. Post training 4-bit quantization of convolutional networks for rapid-deployment. In *NeurIPS*, 2019. 2

[3] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013. 3

[4] Yaohui Cai, Zhewei Yao, Zhen Dong, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. Zeroq: A novel zero shot quantization framework. In *CVPR*, 2020. 1, 2, 3, 6

[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 *NeurIPS*, 2023. 2, 6

[6] Jungwook Choi, Zhuo Wang, Swagath Venkataramani, Pierce I-Jen Chuang, Vijayalakshmi Srinivasan, and Kailash Gopalakrishnan. Pact: Parameterized clipping activation for quantized neural networks. *arXiv preprint arXiv:1805.06085*, 2018. 2

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

[8] Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. Training deep neural networks with low precision multiplications. In *ICLRW*, 2015. 2

[9] Hamid Dehghan and Hassan Ghassemian. Measurement of uncertainty by the entropy: application to the classification of mss data. *International journal of remote sensing*, 27(18):4005–4014, 2006. 4

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 5, 6

[11] Steven K Esser, Jeffrey L McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S Modha. Learned step size quantization. In *ICLR*, 2020. 1

[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. *arXiv preprint arXiv:2202.07471*, 2022. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5, 6

[14] Xiangyu He, Jiahao Lu, Weixiang Xu, Qinghao Hu, Peisong Wang, and Jian Cheng. Generative zero-shot network quantization. In *CVPR Workshops*, 2021. 2, 6

[15] Cheeun Hong, Sungyong Baik, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Cadyq: Content-aware dynamic quantization for image super-resolution. In *ECCV*, 2022. 1, 3, 4

[16] Cheeun Hong and Kyoung Mu Lee. Adabm: On-the-fly adaptive bit mapping for image super-resolution. In *CVPR*, 2024. 3

[17] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *CVPR*, 2018. 1, 5

[18] Sangil Jung, Changyong Son, Seohyung Lee, Jinwoo Son, Jae-Joon Han, Youngjun Kwak, Sung Ju Hwang, and Changkyu Choi. Learning to quantize deep networks by optimizing quantization intervals with task loss. In *CVPR*, 2019. 1, 2

[19] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *CVPR*, 2019. 1, 2

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[21] Xiangtao Kong, Hengyuan Zhao, Yu Qiao, and Chao Dong. Classsr: A general framework to accelerate super-resolution networks by data characteristic. In *CVPR*, 2021. 3

[22] Raghuraman Krishnamoorthi. Quantizing deep convolutional networks for efficient inference: A whitepaper. *arXiv preprint arXiv:1806.08342*, 2018. 2

[23] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5, 6

[24] 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 *CVPR*, 2023. 2, 3, 5, 6

[25] 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 *ICCV*, 2021. 6

[26] Ming Liu, Zhilu Zhang, Liya Hou, Wangmeng Zuo, and Lei Zhang. Deep adaptive inference networks for single image super-resolution. In *ECCV*, 2020. 3

[27] Yuang Liu, Wei Zhang, and Jun Wang. Zero-shot adversarial quantization. In *CVPR*, 2021. 2

[28] Zhenhua Liu, Yunhe Wang, Kai Han, Siwei Ma, and Wen Gao. Instance-aware dynamic neural network quantization. In *CVPR*, 2022. 1, 3, 4

[29] Markus Nagel, Mart van Baalen, Tijmen Blankevoort, and Max Welling. Data-free quantization through weight equalization and bias correction. In *ICCV*, 2019. 2

[30] Junghun Oh, Heewon Kim, Seungjun Nah, Cheeun Hong, Jonghyun Choi, and Kyoung Mu Lee. Attentive fine-grained structured sparsity for image restoration. In *CVPR*, 2022. 3

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 5

[32] Biao Qian, Yang Wang, Richang Hong, and Meng Wang. Adaptive data-free quantization. In *CVPR*, 2023. 6

[33] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *ECCV*, 2016. 2

[34] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *ICCV*, 2019. 4

[35] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. 5, 6

[36] Xu Shoukai, Li Haokun, Zhuang Bohan, Liu Jing, Cao Jiezhang, Liang Chuangrun, and Tan Mingkui. Generative low-bitwidth data free quantization. In *ECCV*, 2020. 1, 2, 5, 6

[37] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*, 2017. 1, 2

[38] Ximeng Sun, Rameswar Panda, Chun-Fu (Richard) Chen, Aude Oliva, Rogerio Feris, and Kate Saenko. Dynamic network quantization for efficient video inference. In *ICCV*, 2021. 3

[39] Senmao Tian, Ming Lu, Jiaming Liu, Yandong Guo, Yurong Chen, and Shunli Zhang. Cabm: Content-aware bit mapping for single image super-resolution network with large input. In *CVPR*, 2023. 1, 3, 4

[40] Longguang Wang, Xiaoyu Dong, Yingqian Wang, Xinyi Ying, Zaiping Lin, Wei An, and Yulan Guo. Exploring sparsity in image super-resolution for efficient inference. In *CVPR*, 2021. 3

[41] Zuxuan Wu, Tushar Nagarajan, Abhishek Kumar, Steven Rennie, Larry S Davis, Kristen Grauman, and Rogerio Feris. Blockdrop: Dynamic inference paths in residual networks. In *CVPR*, 2018. 3

[42] Wenbin Xie, Dehua Song, Chang Xu, Chunjing Xu, Hui Zhang, and Yunhe Wang. Learning frequency-aware dynamic network for efficient super-resolution. In *ICCV*, 2021. 3

[43] Ke Yu, Xintao Wang, Chao Dong, Xiaoou Tang, and Chen Change Loy. Path-restore: Learning network path selection for image restoration. *IEEE TPAMI*, 2021. 3

[44] Edouard Yvinec, Arnaud Dapogny, Matthieu Cord, and Kevin Bailly. Spiq: Data-free per-channel static input quantization. In *WACV*, 2023. 2

[45] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 6

[46] 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*, 2021. 1, 2, 3, 6

[47] Yunshan Zhong, Mingbao Lin, Gongrui Nan, Jianzhuang Liu, Baochang Zhang, Yonghong Tian, and Rongrong Ji. Intraq: Learning synthetic images with intra-class heterogeneity for zero-shot network quantization. In *CVPR*, 2022. 1, 2, 3, 4, 5, 6, 8

[48] Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, and Yuheng Zou. DoReFa-Net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *arXiv preprint arXiv:1606.06160*, 2016. 1, 2