

Distance-aware Quantization

Dohyung Kim Junghyup Lee Bumsub Ham*
School of Electrical and Electronic Engineering, Yonsei University
<https://cvlab.yonsei.ac.kr/projects/DAQ>

Abstract

We address the problem of network quantization, that is, reducing bit-widths of weights and/or activations to lighten network architectures. Quantization methods use a rounding function to map full-precision values to the nearest quantized ones, but this operation is not differentiable. There are mainly two approaches to training quantized networks with gradient-based optimizers. First, a straight-through estimator (STE) replaces the zero derivative of the rounding with that of an identity function, which causes a gradient mismatch problem. Second, soft quantizers approximate the rounding with continuous functions at training time, and exploit the rounding for quantization at test time. This alleviates the gradient mismatch, but causes a quantizer gap problem. We alleviate both problems in a unified framework. To this end, we introduce a novel quantizer, dubbed a distance-aware quantizer (DAQ), that mainly consists of a distance-aware soft rounding (DASR) and a temperature controller. To alleviate the gradient mismatch problem, DASR approximates the discrete rounding with the kernel soft argmax, which is based on our insight that the quantization can be formulated as a distance-based assignment problem between full-precision values and quantized ones. The controller adjusts the temperature parameter in DASR adaptively according to the input, addressing the quantizer gap problem. Experimental results on standard benchmarks show that DAQ outperforms the state of the art significantly for various bit-widths without bells and whistles.

1. Introduction

Convolutional neural networks (CNNs) have made significant progress in the field of computer vision, such as image recognition [27, 48], object detection [2, 43], and semantic segmentation [7, 34]. Deeper [15, 46] and wider [45] CNNs, however, require lots of parameters and FLOPs, making it difficult to deploy modern network architectures on edge devices (e.g., mobile phones, televisions, or drones). Recent works focus on compressing networks to lighten the

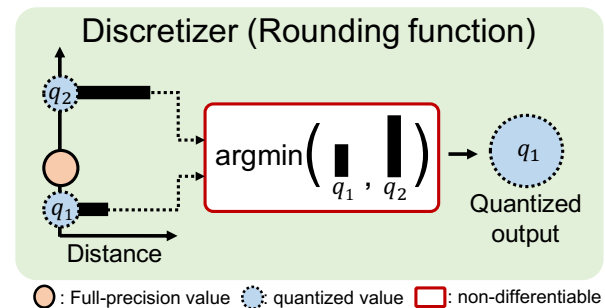


Figure 1: The discretizer takes a full-precision input, and then assigns it to the nearest quantized value, e.g., q_1 in this example. We interpret the assignment process of a discretizer as follows: It first computes the distances between the full-precision input and quantized values, q_1 and q_2 , and then applies an argmin operator over the distances to choose the quantized value. Since this operator is non-differentiable, the quantized network cannot be trained end-to-end with gradient-based optimizers. (Best viewed in color.)

network architectures. Pruning [14] and distillation [16] are representative techniques for network compression. The pruning removes redundant weights in a network, and the distillation encourages a compact network to have features similar to the ones obtained from a large network. The networks compressed by these techniques still exploit floating-point computations, indicating that they are not suitable for edge devices favoring fixed-point operations for power efficiency. Network quantization [42] is an alternative approach that converts full-precision weights and/or activations into low-precision ones, enabling a fixed-point inference, while reducing memory and computational cost.

Quantization methods typically use a staircase function as a quantizer, where it normalizes a full-precision value within a quantization interval, and assigns the normalized one to the nearest quantized value using a discretizer (i.e., a rounding function) [11, 12, 22]. Since the derivative of the rounding is zero at almost everywhere, gradient-based optimizers could not be used to train quantized networks. To address this, the straight-through estimator (STE) [3] replaces the derivative of the rounding with that of identity or hard tanh functions for backward propagation. This, however, causes a gradient mismatch between forward and backward passes at training time, making the training process noisy and degrading

*Corresponding author

the quantization performance at test time [11, 31, 47]. Instead of using the STE, recent methods use soft quantizers, which approximate the discrete rounding with sigmoid [47] or tanh [12] functions, for both forward and backward passes, alleviating the gradient mismatch problem, while maintaining differentiability at training time. These approaches, on the other hand, use the discrete quantizer at inference time. That is, they exploit different quantizers (soft and discrete ones) at training and test time, resulting in a quantizer gap problem [36, 47]. The quantizer gap might be relieved by raising a temperature parameter in the sigmoid function gradually [47], such that the soft quantizer will be transformed to the discrete one eventually at training time, but this causes an unstable gradient flow.

We introduce in this paper a distance-aware quantizer (DAQ) that alleviates the gradient mismatch and quantizer gap problems in a unified framework. Our approach builds upon the insight that the discretizer (*i.e.*, rounding) chooses the nearest quantized value by first computing the distances between a full-precision input and quantized values, and then applying an argmin operator over the distances w.r.t the quantized values (Fig. 1). Motivated by this, we propose a distance-aware soft rounding (DASR) that approximates the discrete rounding accurately using a kernel soft argmax [28], while maintaining differentiability, alleviating the gradient mismatch problem. We also introduce a temperature controller that adjusts a temperature parameter in DASR adaptively depending on the distances between the full-precision input and quantized values. This imposes DASR to have the same output as the discrete rounding, addressing the quantizer gap problem. We apply our DAQ to quantize weights and/or activations for various network architectures, and achieve state-of-the-art results on standard benchmarks, clearly demonstrating the effectiveness of our approach. To our knowledge, it is the first approach to alleviating both gradient mismatch and quantizer gap problems jointly. We summarize the main contributions of this paper as follows:

- We propose a novel differentiable approximation of the discrete rounding function, dubbed DASR, allowing to train quantization networks end-to-end, while alleviating the gradient mismatch problem.
- We introduce a temperature controller, which adjusts the temperature parameter in DASR adaptively, to address the quantizer gap problem.
- We set a new state of the art on standard benchmarks, and provide an extensive analysis of our approach, demonstrating the effectiveness of DAQ.

2. Related work

Network quantization. Early works on network quantization focus on discretizing network weights alone into binary [9, 19, 32], ternary [29, 52], or multi bits [18, 50], which however still requires lots of computational cost to

Method	Gradient mismatch		Quantizer gap	
	Forward	Backward	Training	Test
STE-based [3]	rounding	identity	rounding	rounding
QNet [47]	sigmoid	sigmoid	sigmoid	rounding
Ours	DASR w/ β^*	DASR w/ β^*	DASR w/ β^*	rounding

Table 1: Comparison of quantization methods for gradient mismatch (left) and quantizer gap (right) problems. We denote by β^* a temperature parameter adjusted by our temperature controller. We exploit the same DASR with β^* in both forward and backward passes, alleviating the gradient mismatch problem. In addition, adjusting the temperature enables DASR, used in the training stage, to have the same output as the discrete rounding at test time, addressing the quantizer gap problem.

process full-precision activations. Later, both weights and activations are quantized [20, 42, 51], achieving a better compromise in terms of memory and accuracy. Recent approaches exploit non-uniform quantization levels [5, 49] or transition points [39] to minimize quantization errors. Other methods propose to learn quantization intervals [11, 12, 22] or clipping ranges of activations [8]. These approaches use a staircase function as a quantizer with a discretizer being a rounding function. Since the gradients of the rounding function are zero at almost everywhere, training quantization networks suffers from a vanishing gradient problem. The STE avoids this problem by replacing the zero derivative of the rounding function with that of a continuous one [3]. More specifically, it exploits the rounding in a forward pass, while using *e.g.*, identity or tanh functions, in a backward pass. Although the STE allows to train quantized networks with gradient-based optimizers, the gradient mismatch between forward and backward passes degrades the quantization performance drastically. In contrast, we exploit the same quantizer in both forward and backward passes, alleviating the gradient mismatch problem (Table. 1(left)).

Soft quantizers approximate the discrete rounding with sigmoid [47] or tanh [12] functions. This allows to use continuous functions in both forward and backward passes at training time, alleviating the gradient mismatch problem. These approaches, however, use the rounding function for quantization at inference time, which causes a quantizer gap problem. That is, the outputs of soft and discrete quantizers are different, degrading the quantization performance significantly. To avoid the quantizer gap problem, QNet [47] transforms the soft quantizer towards the discrete one gradually at training time by raising a temperature parameter in a sigmoid function. The gradients of the soft quantizer, however, vanish or explode, as the parameter increases, causing an unstable gradient flow. DSQ [12] incorporates the soft quantizer with the STE. Specifically, it uses the discrete quantizer as in the STE for a forward pass, while exploiting the derivative of the soft quantizer in a backward pass. This alleviates the quantizer gap problem at inference, but the use of STE causes the gradient mismatch. Our method is similar to soft quantizers in that both approximate the

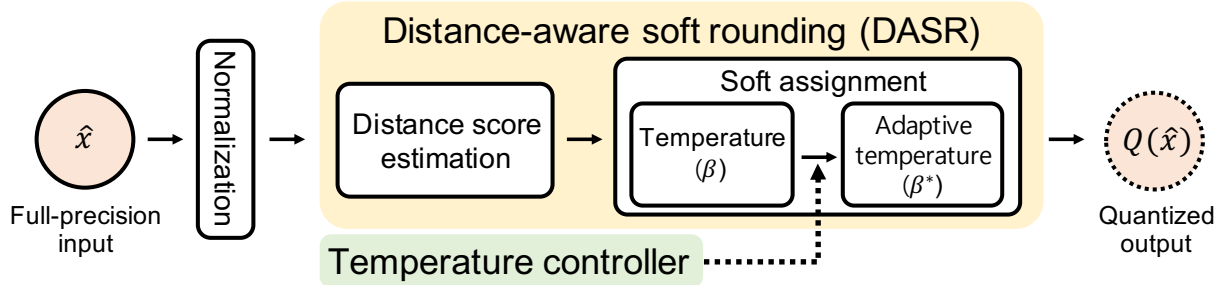


Figure 2: **An overview of DAQ.** Our quantizer Q mainly consists of DASR with a temperature controller. DAQ first normalizes a full-precision input \hat{x} . DASR inputs the normalized input, and computes distance scores w.r.t quantized values. It then assigns the input to the nearest quantized value $Q(\hat{x})$. For the assignment, we exploit a differentiable version of the argmax with an adaptive temperature β^* , obtained from our controller.

discrete one, while maintaining differentiability, alleviating the gradient mismatch problem. On the contrary, we do not exploit continuous functions such as sigmoid or tanh for the approximation, and also address the quantizer gap problem in a unified framework by adjusting a temperature parameter adaptively (Table. 1(right)).

Closely related to our work, RQ [36] defines categorical distributions over quantization grids, and samples a quantized value with the Gumbel soft argmax operator [21]. SHVQ [1] designs a codebook which is a finite set of vectors (*i.e.*, codewords) for vector quantization, and then assigns vectors, *e.g.*, network parameters, to one of the codewords using the soft argmax [6]. They [1, 36], however, also suffer from the quantizer gap problem, since the outputs of soft and discrete quantizers are not the same. In addition, RQ performs a stochastic quantization during training, while quantizing networks in a deterministic manner at inference. This mismatch between training and inference stages degrades the performance significantly [25]. We avoid this problem by designing DASR using a kernel soft argmax [28], quantizing networks deterministically in both training and inference stages.

Differentiable argmax. The argmax operator finds an index of the largest value in an array. CNNs involving the argmax operator cannot be trained via gradient-based optimizers, since it is not differentiable. Soft argmax [6] is a differentiable version of the argmax, and it has been applied to stereo matching [23] and landmark detection [17]. Incorporating the soft argmax operator with the Gumbel noise [13] relaxes a categorical distribution to a concrete one, enabling a differentiable sampling process [21, 37]. The soft argmax operator becomes the discrete one, as the temperature in the operator increases, but at the cost of unstable gradient flow. SFNet [28] introduces a kernel soft argmax which combines a Gaussian kernel with the soft argmax, approximating the discrete argmax more accurately without using a large temperature parameter. We exploit the kernel soft argmax for network quantization. We propose to adjust the temperate parameter in the kernel soft argmax adaptively. As will be

shown later, this addresses the quantizer gap problem, eliminating the discrepancies between the kernel soft argmax and the discrete one.

3. Approach

In this section, we provide a brief description of our DAQ (Sec. 3.1). We then describe each component of DAQ in detail, including DASR (Sec. 3.2) and a temperature controller (Sec. 3.3).

3.1. Overview

Using a uniform b -bit quantization, our quantizer Q maps full-precision inputs \hat{x} , which could be weights \hat{w} or activations \hat{a} , to quantized values $Q(\hat{x}) \in \{0, 1, \dots, 2^b - 1\}$ uniformly (Fig. 2). To this end, we first clip and normalize a full-precision input \hat{x} within a quantization interval, parameterized by upper u and lower l bounds [12, 22], as follows:

$$x = (2^b - 1) \frac{\text{clip}(\hat{x}, \min = l, \max = u) - l}{u - l}, \quad (1)$$

where x is a normalized input. We then apply DASR with a temperature controller, and rescale the output of DASR to assign the normalized input x to the nearest quantized value $Q(\hat{x})$, which will be described in Sec. 3.2 and Sec. 3.3, respectively. Finally, we scale quantized weights $Q(\hat{w})$ or activations $Q(\hat{a})$ linearly to the ranges of $[-1, 1]$ and $[0, 1]$, respectively, as follows:

$$w_q = 2 \frac{Q(\hat{w})}{2^b - 1} - 1, a_q = \frac{Q(\hat{a})}{2^b - 1}, \quad (2)$$

where we denote by w_q and a_q elements of (scaled) quantized tensors for weights \mathbf{w}_q and activations \mathbf{a}_q , respectively. With the quantized weights and activations, a convolutional output \mathbf{o} is obtained as follows:

$$\mathbf{o} = s \mathbf{w}_q * \mathbf{a}_q, \quad (3)$$

where $*$ and s are a convolutional operator and a learnable scalar parameter that adjusts the scale of the convolution output, respectively.

3.2. DASR

The rounding function maps a full-precision input to its nearest quantized value. This assignment process can be thought of as the following two steps: First, distances between full-precision and quantized values are computed. Second, the nearest quantized value is chosen by applying the argmin operator over the distances, which is however not differentiable. Motivated by this, we propose DASR to approximate the discrete rounding function with a differentiable assignment operator. The approximation allows to use the same quantizer in both forward and backward passes, which alleviates the gradient mismatch problem. Similar to the two-step process, DASR takes a normalized input x in Eq. (1), and computes distance scores w.r.t quantized values $q \in \mathbf{q}$, where we denote by \mathbf{q} a set of possible quantized values, *i.e.*, $\{0, 1, \dots, 2^b - 1\}$. It then assigns the input to a floating-point number, very close to the nearest quantized value, by a kernel soft argmax [28]. In the following, we describe DASR in detail.

Distance score. Given the normalized input x , we compute distance scores for individual quantized values $q \in \mathbf{q}$ as follows:

$$d_x(q) = \exp(-|x - q|). \quad (4)$$

The distance score increases as the normalized input x becomes closer to the quantized value q , and vice versa. Note that the computational cost of computing distance scores increases exponentially in accordance with the increase of the bit-widths, *i.e.*, possible quantization values. To address this problem, we compute the distance scores only for the two nearest quantized values, q_f and q_c , w.r.t the normalized input x , where q_f and q_c are obtained by *floor* and *ceil* functions, respectively, *i.e.*, $q_c - q_f = 1$.

Soft assignment. We can assign the normalized input x to the nearest quantized value by applying the discrete argmax function over the distance scores w.r.t the two quantized values, q_f and q_c , but this function is not differentiable. We instead propose to use the kernel soft argmax [28] that approximates the discrete argmax, while maintaining differentiability. We define a soft assignment $\phi(x; \beta)$ for the normalized input x , with a temperature parameter β , as an average of two quantized values, q_f and q_c , weighted by a distance probability m_x (Fig. 3(top)):

$$\phi(x; \beta) = \sum_{i \in \{f, c\}} m_x(q_i; \beta) q_i. \quad (5)$$

The distance probability m_x is obtained by applying a soft-max function, with the temperature parameter β , to the distance scores d_x as follows:

$$m_x(q_i; \beta) = \frac{\exp(\beta k_x(q_i) d_x(q_i))}{\sum_{j \in \{f, c\}} \exp(\beta k_x(q_j) d_x(q_j))}, \quad (6)$$

where we denote by k_x a 1-dimensional Gaussian kernel

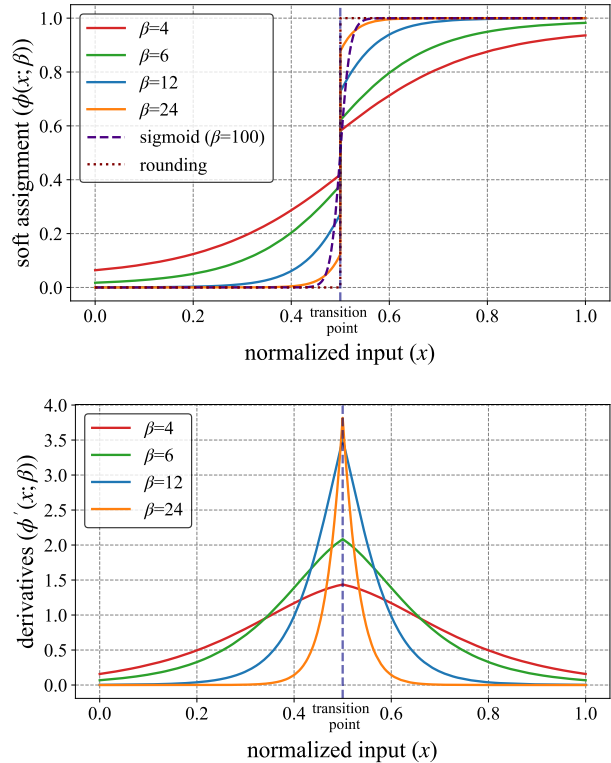


Figure 3: Illustrations of soft assignments with different temperature parameters β (top) and corresponding derivatives (bottom), for the case of the normalized input x being limited within the range of $[0,1]$. For comparison, we also plot a sigmoid function, exploited as a soft quantizer in [47]. We can see that the discrepancies between soft assignments and the rounding are significant especially near the transition point. They can be reduced by raising the temperature parameter. This, however, causes a vanishing gradient problem for the inputs near quantized values. (Best viewed in color.)

centered on the nearest quantized value (*i.e.*, q_f or q_c) for the normalized input x . The output of the kernel $k_x(q)$ becomes larger as q approaches to the nearest quantized value, which has an effect of retaining the distance score for the nearest quantized value, while suppressing the other one. This suggests that the distance probability is distributed with one clear peak around the nearest quantized value, and our soft assignment process approximates the discrete argmax well. For example, we can see from Fig. 3(top) that the soft assignment ϕ approximates the discrete rounding more accurately than the sigmoid function adopted in soft quantization [47], even with a much smaller temperature, preventing a gradient exploding problem.

3.3. Temperature

The temperature parameter β adjusts a distribution of the distance probability m_x , and it thus influences the soft assignment ϕ . The soft assignment with a fixed temperature parameter has the following limitations. (1) Small temperature parameters (*e.g.*, $\beta = 4$ in Fig. 3(top)) cause

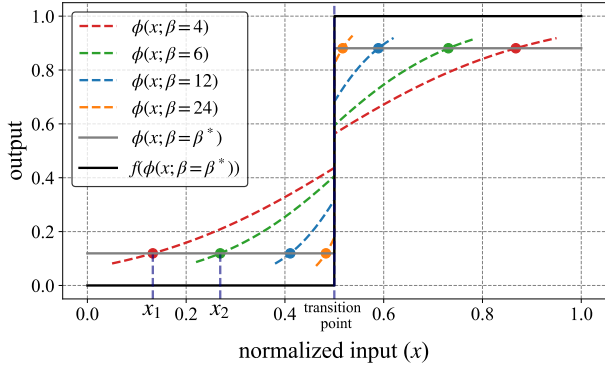


Figure 4: Illustrations of the output $\phi(x; \beta = \beta^*)$ in Eq. (9) (gray line), and its rescaled version using a function f in Eq. (10) (black line). Computing the soft assignment with the adaptive temperature β^* has an effect of sampling output values (colored circles) from different functions for individual inputs (dotted lines). (Best viewed in color.)

a quantizer gap problem, which is problematic particularly when a normalized input x is close to a transition point. (2) Large temperature parameters (e.g., $\beta = 24$ in Fig. 3(top)) alleviate the quantizer gap problem. This, however, leads to a vanishing gradient problem. For example, the derivative of the soft assignment converges to zero rapidly, as the normalized input x moves away from a transition point in Fig. 3(bottom), i.e., approaches to a quantized value. Accordingly, exploiting a temperature parameter fixed for all inputs suffers from quantizer gap or vanishing gradient problems.

Temperature controller. We adjust the temperature parameter β adaptively according to the distance between a normalized input and a transition point, such that we minimize the quantizer gap without suffering from the vanishing gradient problem. Specifically, we raise the temperature for the inputs near the transition point in order to address the quantizer gap problem. On the other hand, we lower the temperature for the inputs distant from the transition point, alleviating the vanishing gradient problem. To implement this idea, we define an adaptive temperature β^* as follows:

$$\beta^* = \frac{\gamma}{|s_x(q_f) - s_x(q_c)|}, \quad (7)$$

where γ is a positive constant, and $s_x(q_i)$ is a weighted (distance) score defined as:

$$s_x(q_i) = k_x(q_i)d_x(q_i). \quad (8)$$

As the input approaches to the transition point, the weighted scores, $s_x(q_f)$ and $s_x(q_c)$, become similar. In this case, the denominator in Eq. (7) decreases, and the adaptive temperature thus increases, alleviating the quantizer gap problem. On the contrary, the adaptive temperature decreases, when the input moves away from the transition point, avoiding the vanishing gradient problem.

The adaptive temperature β^* varies according to the input, and the temperature in turn changes the shape of the soft

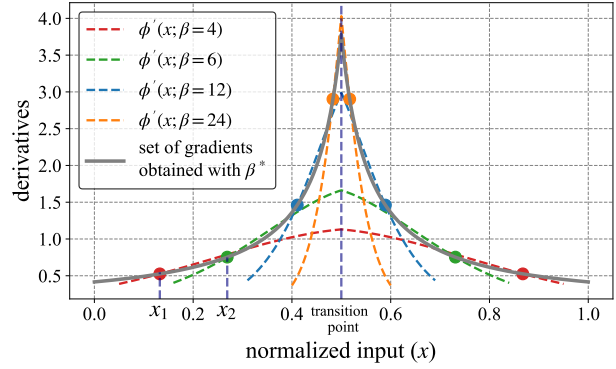


Figure 5: Illustrations of derivatives of soft assignment functions with different temperatures (dotted lines), and a set of gradients for individual inputs (gray line), obtained with different functions. The gradients for individual inputs are computed using the functions with corresponding adaptive temperature β^* (colored circles). (Best viewed in color.)

assignment function (Fig. 3), suggesting that different soft assignment functions are applied for individual inputs. As an example, for the point of x_1 in Fig. 4, where a value of the corresponding adaptive temperature is 4, its assignment is computed with the function of $\phi(x; \beta = 4)$. For the point of x_2 , the assignment is obtained with the function of $\phi(x; \beta = 6)$, which differs from the one used for x_1 since the temperature is changed. By applying different functions for all inputs, we can obtain a set of assignments that can be represented analytically by plugging the adaptive temperature β^* into the soft assignment in Eq. (5) as follows (see the supplement for details):

$$\phi(x; \beta = \beta^*) = \begin{cases} q_f + \lambda, & x \leq q_t \\ q_c - \lambda, & x > q_t, \end{cases} \quad (9)$$

where $\lambda = 1/(e^\gamma + 1)$, and we denote by q_t a transition point, defined as $(q_f + q_c)/2$. We can see from Fig. 4 that the set of assignments, sampled from different soft assignment functions for individual inputs, are the same as output values of a single discrete function (Fig. 4(gray line)). Analogous to the process of obtaining the set of assignments in a forward pass, the gradients for individual inputs are obtained with different functions for backward propagation (Fig. 5). That is, the gradient for each input is obtained from the soft assignment function with the corresponding adaptive temperature. For example, the gradients for the points, x_1 and x_2 in Fig. 5, are computed using the derivatives of the corresponding functions, $\phi'(x; \beta = 4)$ and $\phi'(x; \beta = 6)$, respectively. This is effective in alleviating the vanishing gradient problem, which is particularly severe, when the temperature is set to a large value (e.g., $\beta = 24$ in Fig. 3(top)). Note that the adaptive temperature β^* is regarded as a hyperparameter for each input in both forward and backward passes. In summary, leveraging a different function for each input enables computing the gradient for backward propagation, while pro-

Method	Bit-width (W/A)								
	FP	1/1	1/2	2/2	3/3	4/4	1/32	2/32	3/32
LQ-Nets [49]	70.3	-	62.6 (-7.7)	64.9 (-5.4)	68.2 (-2.1)	69.3 (-1.0)	-	68.0 (-2.3)	69.3 (-1.0)
PACT [8]	70.4	-	62.9 (-7.5)	64.4 (-6.0)	68.1 (-2.3)	69.2 (-1.2)	-	68.1 (-2.3)	69.9 (-0.5)
QIL [22]	70.2	-	-	65.7 (-4.5)	69.2 (-1.0)	70.1 (-0.1)	65.8 (-4.4)	-	-
QNet [47]	70.3	53.6 (-16.7)	63.4 (-6.9)	-	-	-	66.5 (-3.8)	69.1 (-1.2)	70.4 (+0.1)
RQ [36]	69.5	-	-	-	-	62.5 [†] (-7.0)	-	-	-
DSQ [12]	69.9	-	-	65.2 (-4.7)	68.7 (-1.2)	69.6 [†] (-0.3)	-	-	-
LSQ [11]	70.1	-	-	66.8 (-3.3)	69.3 (-0.8)	70.7 (+0.6)	-	-	-
LSQ+ [4]	70.1	-	-	66.7 (-3.4)	69.4 (-0.7)	70.8 (+0.7)	-	-	-
IRNet [41]	69.6	-	-	-	-	-	66.5 (-3.1)	-	-
Ours	69.9	56.2 (-13.7)	64.6 (-5.3)	66.9 (-3.0)	69.6 (-0.3)	70.5 (+0.6)	67.2 (-2.7)	69.8 (-0.1)	70.8 (+0.9)

Table 2: Quantitative results of ResNet-18 [15] on the validation split of ImageNet [10]. We report the top-1 accuracy for comparison. We denote by “W” and “A” the bit-precision of weights and activations, respectively. “FP” and † represent accuracies for full-precision and fully quantized models, respectively. Numbers in bold indicate the best performance. Numbers in parentheses are accuracy improvements or degradations compared to the full-precision one.

viding the outputs that coincide with those obtained from a discrete rounding but with offsets of λ (Fig. 4). Rescaling the output in Eq. (9) with the following function,

$$f(y) = \frac{y - q_t}{1 - 2\lambda} + q_t, \quad (10)$$

we can obtain

$$f(\phi(x; \beta = \beta^*)) = \begin{cases} q_f, & x \leq q_t \\ q_c, & x > q_t, \end{cases} \quad (11)$$

which corresponds to the output of DAQ (*i.e.*, $Q(\hat{x})$). It is clear that the rescaled output in Eq. (11) provides the exactly same values as the rounding function, suggesting that our DAQ is free from the quantizer gap problem, even using the rounding at test time (see the last row in Table. 6). We summarize the overall quantization process in the supplement.

4. Experiments

4.1. Experimental details

Implementation details. We quantize weights and/or activations for ResNets [15] (*i.e.*, ResNet-18, -20, and -34) and MobileNet-V2 [44]. Following [22, 38], we do not quantize the first and last layers for all network architectures except for MobileNet-V2, where all layers are quantized. We empirically set the constant γ in Eq. (7) to 2 for both weight and activation quantizers, and the standard deviation of the Gaussian kernel k_x to 1 and 2 for quantizers of weight and activation, respectively. We use a grid search to set these parameters. We choose the ones that give the best performance on the validation split¹ of CIFAR-10 [26], and fix them for all experiments.

Training. Network weights are trained using the SGD optimizer with learning rates of 1e-2 and 5e-3 for ResNets and MobileNet-V2, respectively. We learn quantization parameters, such as the lower and upper bounds, l and u in Eq. (1), and the scale factor s in Eq. (3), using the Adam

¹We divide the training split of CIFAR-10 [26] into training and validation sets for the grid search.

optimizer [24] with a learning rate of 1e-4. The learning rates for all parameters are scheduled by the cosine annealing strategy [35]. For the ResNet-20 architecture, we train the quantized networks for 400 epochs on CIFAR-10 [26] with a batch size of 256, and the weight decay is set to 1e-4. Other networks are trained for 100 epochs on ImageNet [10] with batch sizes of 256 and 160 for ResNets (*i.e.*, ResNet-18 and -34) and MobileNet-V2, respectively. For ResNet-18 and -34, the weight decay is set to 1e-4, except for low-bit quantizations (*i.e.*, 1/1, 1/2, and 2/2-bit settings), where we use a smaller weight decay of 5e-5, following [11]. For MobileNet-V2, the weight decay is set to 4e-5. We do not use weight decay for learning quantization parameters.

Initialization. The weights in all quantized networks are initialized from the full-precision pretrained models. We apply standardization to the weights [30] before feeding them into quantizers. We initialize lower and upper bounds in the weight quantizer to -3 and 3, respectively. Lower and upper bounds in the activation quantizer are initialized by $-3\sigma_A$ and $3\sigma_A$, respectively, where σ_A is a standard deviation of input activations in a layer, except when an input of the quantizer is pre-activated by a ReLU. In this case, we fix a lower bound to zero, and learn an upper bound only with an initialization of $3\sigma_A$.

4.2. Results

We evaluate our approach with network architectures, including ResNet-18, -20, -34 [15], and MobileNet-V2 [44], for various bit-widths, and compare the performance with the state of the art for image classification on CIFAR-10 [26] and ImageNet [10].

ImageNet. The ImageNet dataset [10] provides approximately 1.2 million training and 50K validation images of 1,000 categories with corresponding ground-truth class annotations. We train and evaluate a quantized network on the training and validation splits, respectively. We use the top-1 accuracy to quantify the performance.

We show in Table 2 the top-1 accuracy on ResNet-18, and

Method	Bit-width (W/A)						
	FP	1/1	1/2	2/2	3/3	4/4	1/32
LQ-Nets [49]	73.8	-	66.6 (-7.2)	69.8 (-4.0)	71.9 (-1.9)	-	-
QIL [22]	73.7	-	-	70.6 (-3.1)	73.1 (-0.6)	73.7 (+0.0)	-
DSQ [12]	73.3	-	-	70.0 (-3.3)	72.5 (-0.8)	72.8 (-0.5)	-
IRNet [41]	73.3	-	-	-	-	-	70.4 (-2.9)
Ours	73.3	62.1 (-11.2)	69.4 (-3.9)	71.0 (-2.3)	73.1 (-0.2)	73.7 (+0.4)	71.9 (-1.4)

Table 3: Quantitative results of ResNet-34 [15] on the validation split of ImageNet [10]. We report the top-1 accuracy for comparison. W/A: Bit-precision of weights/activations; FP: Results obtained by full-precision models.

Method	Bit-width (W/A)	
	FP	4/4
PACT [8]	71.8	61.4 (-10.4)
DSQ [12]	71.9	64.8 (-7.1)
PROFIT [38]	71.9	71.6 [†] (-0.3)
Ours	71.9	70.0 [†] (-1.9)

Table 4: Quantitative results of MobileNet-V2 [44] on the validation split of ImageNet [10]. We report the top-1 accuracy for comparison. W/A: Bit-precision of weights/activations; FP: Results obtained by full-precision models; †: Results from fully quantized models.

compare our approach with the state of the art. All numbers in Table 2 are taken from each paper, except for LSQ [11]². We observe five things from this table: (1) Our method outperforms the state of the art by a significant margin in terms of the top-1 accuracy especially for low-bit quantizations (*i.e.*, 1/1, 1/2, 1/32, 2/32-bit settings). LSQ [11] and LSQ+ [4] show better results than ours slightly in a 4/4-bit setting, but using a more accurate full-precision model. The high-performance model provides a better initialization to optimize quantized networks. (2) Our approach is effective to binarize networks (*i.e.*, 1/1 and 1/32-bit settings), outperforming IRNet [41]³ designed for network binarization. This also demonstrates the effectiveness of DASR on alleviating the gradient mismatch problem. As stated in [31], the gradient mismatch problem becomes even worse, as the bit-width of weights and/or activations is small. (3) Ours performs better than soft quantizers [12, 36, 47], even they also alleviate the gradient mismatch problem. This indicates that addressing the quantizer gap problem improves the performance significantly. (4) Our method gives better results than other quantizers, similar to ours, that learn either quantization intervals [11, 12, 22] or clipping ranges of activations [8] using the STE. (5) We can employ our quantization method in various bit-widths, and achieve the state-of-the-art performance consistently, while others may apply for specific settings.

Tables 3 and 4 show quantitative comparisons with the

²It uses a pre-activated version of ResNet, which is different from the standard architecture. We take the numbers from the work of LSQ+ [4], where LSQ is reproduced with the standard ResNet.

³It shows top-1 accuracy of 58.1 in an 1/1-bit setting, but using ResNet-18 with the Bi-Real structure [33] that adds additional residual connections. For fair comparison, we only report the results of IRNet using the same network architectures as ours

Method	Bit-width (W/A)		
	FP	1/1	1/32
DoReFa [51]	90.8	79.3 (-11.5)	90.0 (-0.8)
LQ-Net [49]	92.1	-	90.1 (-2.0)
DSQ [12]	90.8	84.1 (-6.7)	90.2 (-0.6)
IRNet [41]	91.7	85.4 (-6.3)	90.8 (-0.9)
Ours	91.4	85.8 (-5.6)	91.2 (-0.2)

Table 5: Quantitative results of ResNet-20 [15] on the test split of CIFAR-10 [26]. We report the top-1 accuracy for comparison. W/A: Bit-precision of weights/activations; FP: Results obtained by full-precision models.

state of the art using ResNet-34 [15] and MobileNet-V2 [44], respectively. Our method outperforms the state of the art in low-bit quantizations (*e.g.*, 1/1, 2/2, and 1/32-bit), and shows the same accuracy as QIL [22] in 3/3 and 4/4-bit settings. QIL uses a progressive learning technique [53] training a quantized network sequentially from high- to low-bit precision, which is computationally demanding. In contrast, our approach fine-tunes a full-precision model directly to achieve quantized networks with the target precision. Note that our method shows the best performance in terms of an accuracy improvement/degradation from a full-precision model. Table 4 demonstrates that our method is also effective to quantize a light-weight network architecture (*i.e.*, MobileNet-V2). Note that PROFIT [38] is specially designed for quantizing the light-weight network architectures, and also uses many heuristics (*e.g.*, progressive learning [22, 53] and distillation [40, 53]) at training time, which requires more training time and computational cost, compared to our approach.

CIFAR-10. The CIFAR-10 dataset [26] consists of 50K training and 10K test images of 10 object categories for image classification. We train a quantized network with the training set, and report the top-1 accuracy on the test split. We show in Table 5 a quantitative comparison with the state of the art using the ResNet-20 architecture [15]. We can clearly see that our method outperforms the state of the art, confirming the effectiveness of our approach again.

4.3. Discussion

We present an ablation analysis on DASR and compare our temperature controller with other methods alleviating the quantizer gap problem. We report the top-1 accuracy with 1/1-bit ResNet-20 [15] on the test split of CIFAR-10 [26]. More analysis on DAQ can be found in the supplement.

Type	Temperature (β)	Test time	
		Rounding	DASR
Soft argmax	10	11.7 (-75.1)	86.8
	20	10.2 (-44.4)	54.6
	60	10.0 (-27.3)	37.3
	150	-	-
Kernel soft argmax	4	13.5 (-76.2)	89.7
	8	48.8 (-37.2)	86.0
	12	69.9 (-10.9)	80.8
	24	59.6 (-3.8)	63.8
	β^*	85.8 (-0.0)	85.8

Table 6: Quantitative comparison for variants of our method with different argmax operators and temperature parameters. We also report the results obtained by DASR at test time, instead of the rounding function, that is, we use the same quantization function for both training and test time to quantify the influence of the quantizer gap problem on quantization. Note that weights and activations in this case, except the result for the adaptive temperature β^* , are full-precision values, not quantized ones. Numbers in parentheses are accuracy drops between full-precision and quantized models.

Differentiable argmax. We compare in the third column of Table 6 the quantization performance for variants of our method. We use DASR to train quantized networks with different argmax operators and temperature parameters, and exploit the rounding function as a discretizer at test time. To quantify the influence of the quantizer gap problem, we also report the results when exploiting DASR at test time in the fourth column of Table 6, such that we use the same quantization function for both training and test time. Note that DASR with a fixed temperature outputs floating-point numbers as described in Sec. 3.3, indicating that the results in the fourth column, except for the one for the adaptive temperature β^* , are obtained with full-precision weights and activations, not quantized ones. We can see from the first four rows that an accuracy drop caused by the quantizer gap decreases, according to the increase of the temperature parameter. This, however, results in an unstable gradient flow, degrading the performance, even for the cases without the quantizer gap (e.g., 86.8 for $\beta=10$ vs. 37.3 for $\beta=60$). We fail to train the network with $\beta = 150$, due to a gradient exploding problem. The last five rows show that the kernel soft argmax [28] is more effective to handle the quantizer gap problem than the soft argmax, even with a much smaller temperature parameter. For example, raising the temperature parameter from 4 to 24 reduces the accuracy drop from 76.2% to 3.8%. Exploiting a large temperature ($\beta = 24$), however, causes a vanishing gradient problem as stated in Sec. 3.3, which degrades the quantization performance. From the last row, we can observe that the adaptive temperature addresses the quantizer gap and vanishing gradient problems, providing the best result without the performance drop.

Temperature controller. Table 7 compares our temperature controller with other methods [12, 36, 47] for avoiding

Method	Time/iters (ms)	Top-1 accuracy (%)
Annealing [47]	286.2	72.2
Combine STE [12, 36]	287.8	81.7
Temperature Controller	296.9	85.8

Table 7: Training time and accuracy comparisons of our temperature controller with other methods avoiding the quantizer gap problem. The time is measured with an RTX 2080Ti GPU.

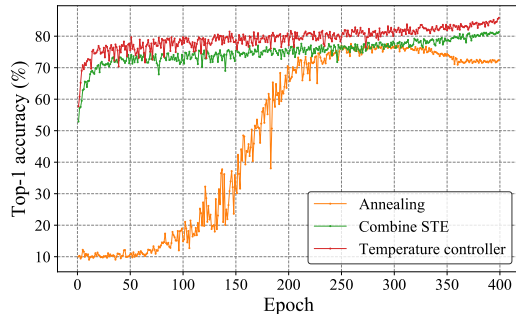


Figure 6: Visual comparison of top-1 accuracies for different methods alleviating the quantizer gap problem. (Best viewed in color.)

the quantizer gap problem. For fair comparison, we adopt these methods within our DASR framework. Specifically, for the temperature annealing [47], we raise the temperature parameter from 2 to 48 gradually at training time, such that DASR approaches to the rounding function. To combine the STE with DASR [12, 36], we use the discrete rounding for a forward pass, while using the derivative of DASR in a backward pass. Exploiting the temperature controller takes more time than other methods due to the additional computations of adjusting the temperature β for individual inputs x . This, however, alleviates the quantizer gap and gradient mismatch problems jointly, outperforming other methods by a large margin. We compare in Fig. 6 top-1 accuracy curves. We can observe that our temperature controller gives better results compared with others during training.

5. Conclusion

We have shown that network quantization can be formulated as an assignment problem between full-precision and quantized values, and introduced a novel quantizer, dubbed DAQ, that addresses both the gradient mismatch and quantizer gap problems in a unified framework. Specifically, DASR approximates a rounding function with a kernel soft argmax operator, together with a temperature controller adjusting the temperature parameter adaptively. We have shown that DAQ achieves the state of the art for various network architectures and bit-widths without bells and whistles. We have also verified the effectiveness of each component of DAQ with a detailed analysis.

Acknowledgments. This research was supported by the Samsung Research Funding & Incubation Center for Future Technology (SRFC-IT1802-06).

References

- [1] Eirikur Agustsson, Fabian Mentzer, Michael Tschannen, Lukas Cavigelli, Radu Timofte, Luca Benini, and Luc Van Gool. Soft-to-hard vector quantization for end-to-end learning compressible representations. In *NIPS*, 2017. 3
- [2] Yancheng Bai, Yongqiang Zhang, Mingli Ding, and Bernard Ghanem. SOD-MTGAN: Small object detection via multi-task generative adversarial network. In *ECCV*, 2018. 1
- [3] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *CoRR*, 2013. 1, 2
- [4] Yash Bhalgat, Jinwon Lee, Markus Nagel, Tijmen Blankevoort, and Nojun Kwak. LSQ+: Improving low-bit quantization through learnable offsets and better initialization. In *CVPRW*, 2020. 6, 7
- [5] Zhaowei Cai, Xiaodong He, Jian Sun, and Nuno Vasconcelos. Deep learning with low precision by half-wave gaussian quantization. In *CVPR*, 2017. 2
- [6] Olivier Chapelle and Mingrui Wu. Gradient descent optimization of smoothed information retrieval metrics. *Information Retrieval*, 2010. 3
- [7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015. 1
- [8] Jungwook Choi, Zhuo Wang, Swagath Venkataramani, Pierce I-Jen Chuang, Vijayalakshmi Srinivasan, and Kailash Gopalakrishnan. PACT: Parameterized clipping activation for quantized neural networks. *CoRR*, 2018. 2, 6, 7
- [9] Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. Binaryconnect: Training deep neural networks with binary weights during propagations. In *NIPS*, 2015. 2
- [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. 6, 7
- [11] Steven K Esser, Jeffrey L McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S Modha. Learned step size quantization. In *ICLR*, 2020. 1, 2, 6, 7
- [12] Ruihao Gong, Xianglong Liu, Shenghu Jiang, Tianxiang Li, Peng Hu, Jiazhen Lin, Fengwei Yu, and Junjie Yan. Differentiable soft quantization: Bridging full-precision and low-bit neural networks. In *ICCV*, 2019. 1, 2, 3, 6, 7, 8
- [13] Emil Julius Gumbel. *Statistical theory of extreme values and some practical applications: a series of lectures*, volume 33. US Government Printing Office, 1948. 3
- [14] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *ICLR*, 2016. 1
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 6, 7
- [16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS Workshop*, 2014. 1
- [17] Sina Honari, Pavlo Molchanov, Stephen Tyree, Pascal Vincent, Christopher Pal, and Jan Kautz. Improving landmark localization with semi-supervised learning. In *CVPR*, 2018. 3
- [18] Lu Hou and James T Kwok. Loss-aware weight quantization of deep networks. In *ICLR*, 2018. 2
- [19] Lu Hou, Quanming Yao, and James T Kwok. Loss-aware binarization of deep networks. In *ICLR*, 2017. 2
- [20] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. In *NIPS*, 2016. 2
- [21] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *ICLR*, 2017. 3
- [22] 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, 3, 6, 7
- [23] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *ICCV*, 2017. 3
- [24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6
- [25] Raghuraman Krishnamoorthi. Quantizing deep convolutional networks for efficient inference: A whitepaper. *CoRR*, 2018. 3
- [26] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6, 7
- [27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012. 1
- [28] Junghyup Lee, Dohyung Kim, Jean Ponce, and Bumsu Ham. SFNet: Learning object-aware semantic correspondence. In *CVPR*, 2019. 2, 3, 4, 8
- [29] Fengfu Li, Bo Zhang, and Bin Liu. Ternary weight networks. In *NIPS Workshop*, 2016. 2
- [30] Yuhang Li, Xin Dong, and Wei Wang. Additive powers-of-two quantization: An efficient non-uniform discretization for neural networks. In *ICLR*, 2020. 6
- [31] Darryl D Lin and Sachin S Talathi. Overcoming challenges in fixed point training of deep convolutional networks. In *ICML Workshop*, 2016. 2, 7
- [32] Zhouhan Lin, Matthieu Courbariaux, Roland Memisevic, and Yoshua Bengio. Neural networks with few multiplications. In *ICLR*, 2016. 2
- [33] Zechun Liu, Baoyuan Wu, Wenhan Luo, Xin Yang, Wei Liu, and Kwang-Ting Cheng. Bi-real net: Enhancing the performance of 1-bit cnns with improved representational capability and advanced training algorithm. In *ECCV*, 2018. 7
- [34] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1
- [35] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 6
- [36] Christos Louizos, Matthias Reisser, Tijmen Blankevoort, Efstratios Gavves, and Max Welling. Relaxed quantization for discretized neural networks. In *ICLR*, 2019. 2, 3, 6, 7, 8
- [37] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *ICLR*, 2017. 3
- [38] Eunhyeok Park and Sungjoo Yoo. PROFIT: A novel training method for sub-4-bit mobilenet models. In *ECCV*, 2020. 6, 7
- [39] Eunhyeok Park, Sungjoo Yoo, and Peter Vajda. Value-aware quantization for training and inference of neural networks. In *ECCV*, 2018. 2
- [40] Antonio Polino, Razvan Pascanu, and Dan Alistarh. Model

- compression via distillation and quantization. In *ICLR*, 2018. 7
- [41] Haotong Qin, Ruihao Gong, Xianglong Liu, Mingzhu Shen, Ziran Wei, Fengwei Yu, and Jingkuan Song. Forward and backward information retention for accurate binary neural networks. In *CVPR*, 2020. 6, 7
 - [42] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. XNOR-Net: Imagenet classification using binary convolutional neural networks. In *ECCV*, 2016. 1, 2
 - [43] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1
 - [44] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. 6, 7
 - [45] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 1
 - [46] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017. 1
 - [47] Jiwei Yang, Xu Shen, Jun Xing, Xinmei Tian, Houqiang Li, Bing Deng, Jianqiang Huang, and Xian-sheng Hua. Quantization networks. In *CVPR*, 2019. 2, 4, 6, 7, 8
 - [48] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 1
 - [49] Dongqing Zhang, Jiaolong Yang, Dongqiangzi Ye, and Gang Hua. LQ-Nets: Learned quantization for highly accurate and compact deep neural networks. In *ECCV*, 2018. 2, 6, 7
 - [50] Aojun Zhou, Anbang Yao, Yiwen Guo, Lin Xu, and Yurong Chen. Incremental network quantization: Towards lossless cnns with low-precision weights. In *ICLR*, 2017. 2
 - [51] 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. *CoRR*, 2016. 2, 7
 - [52] Chenzhuo Zhu, Song Han, Huizi Mao, and William J Dally. Trained ternary quantization. In *ICLR*, 2017. 2
 - [53] Bohan Zhuang, Chunhua Shen, Mingkui Tan, Lingqiao Liu, and Ian Reid. Towards effective low-bitwidth convolutional neural networks. In *CVPR*, 2018. 7