

Network Quantization with Element-wise Gradient Scaling

Junghyup Lee

Dohyung Kim

Bumsub Ham*

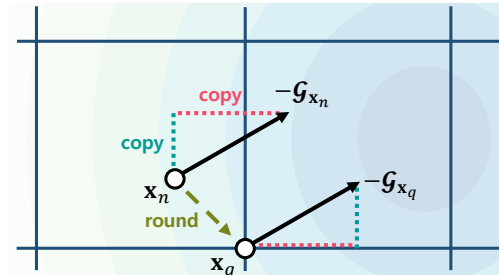
School of Electrical and Electronic Engineering, Yonsei University

Abstract

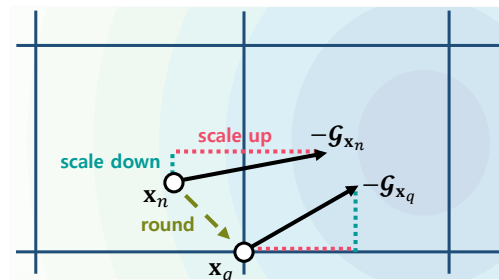
Network quantization aims at reducing bit-widths of weights and/or activations, particularly important for implementing deep neural networks with limited hardware resources. Most methods use the straight-through estimator (STE) to train quantized networks, which avoids a zero-gradient problem by replacing a derivative of a discretizer (i.e., a round function) with that of an identity function. Although quantized networks exploiting the STE have shown decent performance, the STE is sub-optimal in that it simply propagates the same gradient without considering discretization errors between inputs and outputs of the discretizer. In this paper, we propose an element-wise gradient scaling (EWGS), a simple yet effective alternative to the STE, training a quantized network better than the STE in terms of stability and accuracy. Given a gradient of the discretizer output, EWGS adaptively scales up or down each gradient element, and uses the scaled gradient as the one for the discretizer input to train quantized networks via backpropagation. The scaling is performed depending on both the sign of each gradient element and an error between the continuous input and discrete output of the discretizer. We show extensive experimental results on the image classification datasets, including CIFAR-10 and ImageNet, with diverse network architectures under a wide range of bit-width settings, demonstrating the effectiveness of our method.

1. Introduction

Convolutional neural networks (CNNs) have shown remarkable advances in many computer vision tasks, such as image classification [16, 24, 36], semantic segmentation [15, 28], object detection [13, 27], and image restoration [9], while at the cost of large amounts of weights and operations. Network quantization lowers bit-precision of weights and/or activations in a network. It is effective in particular to reduce the memory and computational cost of



(a) Gradient propagation using STE [3].



(b) Gradient propagation using EWGS.

Figure 1: Comparison of STE [3] and EWGS. We visualize discrete levels and a loss landscape by straight lines and a contour plot, respectively. In a forward pass, a continuous latent point \mathbf{x}_n is mapped to a discrete point \mathbf{x}_q using a round function. Training a quantized network requires backpropagating a gradient from \mathbf{x}_q to \mathbf{x}_n . (a) The STE propagates the same gradient i.e., $\mathcal{G}_{\mathbf{x}_n} = \mathcal{G}_{\mathbf{x}_q}$ without considering the value of \mathbf{x}_n , where we denote by $\mathcal{G}_{\mathbf{x}_n}$ and $\mathcal{G}_{\mathbf{x}_q}$ the gradients of \mathbf{x}_n and \mathbf{x}_q , respectively. (b) Our approach, on the other hand, scales up or down each element of the gradient during backpropagation, while taking into account discretization errors i.e., $\mathbf{x}_n - \mathbf{x}_q$. (Best viewed in color.)

CNNs, and thus network quantization could be a potential solution for implementing CNNs with limited hardware resources. For example, binarized neural networks [19, 34] use $32\times$ less memory compared to the full-precision (32-bit) counterparts, and the binarization techniques allow to replace multiplication and addition with XNOR and bit-count operations, respectively.

Quantized networks involve weight and/or activation quantizers in convolutional or fully-connected layers. The quantizers take full-precision weights or activations, and typically perform normalization, discretization, and denor-

*Corresponding author.

malization steps to convert them into low-precision ones. The main difficulty of training a quantized network arises from the discretization step, where a discretizer (*i.e.*, a round function) maps a normalized value to one of discrete levels. Since an exact derivative of the discretizer is either zero or infinite, gradients become zero or explode during backpropagation. Most quantization methods [7, 12, 21, 31, 42, 44] overcome this issue by exploiting the straight-through estimator (STE) [3]. The STE propagates the same gradient from an output to an input of the discretizer, assuming that the derivative of the discretizer is equal to 1. This could bring a gradient mismatch problem [39], since the discretizer used in a forward pass (*i.e.*, the round function) does not match up with that in a backward pass (*i.e.*, an identity or hard tanh functions). Nevertheless, recent methods exploiting the STE have shown reasonable performance [4, 12, 21, 31].

We take a different point of view on how the STE works. We interpret that a full-precision input (which we call a “latent value”) of the discretizer moves in a continuous space, and a discretizer output (which we call a “discrete value”) is determined by projecting the latent value to the nearest discrete level in the space. This suggests that shifting the latent values in the continuous space influences the discrete values. The STE, in this sense, shifts (or updates) the latent values with *coarse gradients* [41], that is, the gradients obtained with the discrete values (Fig. 1a), which is sub-optimal. For example, both latent values of 0.51 and 1.49 produce the same discrete value of 1 using a round function, and the STE forces to update the latent values equally with the same gradient from the discrete value of 1, regardless of their discretization errors induced by the rounding. Updating these latent values should be treated differently, because, for example, a small increment for the latent value of 1.49 leads to changing the discrete value from 1 to 2, whereas the increment for the latent value of 0.51 cannot. Similarly, a small decrement for the latent value of 0.51 can convert the discrete value from 1 to 0, but the latent value of 1.49 requires a much larger decrement to do so.

In this paper, we present an element-wise gradient scaling (EWGS) that enables better training of a quantized network, compared with the STE, in terms of stability and accuracy. Given a gradient of discrete values, EWGS adaptively scales up or down each element of the gradient considering its sign and discretization errors between latent and discrete values. The scaled gradient is then used to update the latent value (Fig. 1b). Since optimal scaling factors, which control the extent of EWGS, may vary across weight or activation quantizers in different layers, we propose an approach to adjusting the factors adaptively during training. Specifically, we relate the scaling factor with the second-order derivatives of a task loss w.r.t the discrete values, and propose to estimate the factor with the trace of a Hessian

matrix, which can be computed efficiently with the Hutchinson’s method [1, 40]. Without an extensive hyperparameter search, training schedules [39, 43, 47], or additional modules [6, 30, 46], various CNN architectures trained with our approach achieve state-of-the-art performance on ImageNet [8]. Note that the STE is a special case of EWGS, indicating that it can be exploited to other quantization methods using the STE. The main contributions of our work can be summarized as follows:

- We introduce EWGS that scales up or down each gradient element of the discrete value adaptively for backpropagation, while considering discretization errors between inputs and outputs of a discretizer.
- We relate a scaling factor with the second-order derivatives of a loss function w.r.t discrete values, allowing to compute the parameter effectively and adaptively with the Hessian information of a quantized network.
- We demonstrate the effectiveness of our method with various CNN architectures under a wide range of bit-widths, outperforming the state of the art on ImageNet [8]. We also verify that our approach boosts the performance of other quantization methods, such as DoReFa-Net [44] and PROFIT [31].

Our code and models are available online: <https://cvlab.yonsei.ac.kr/projects/EWGS>.

2. Related work

Network quantization has been formulated as a constrained optimization problem to minimize quantization errors, where bit-widths of weight and/or activation values are restricted by binary [34], ternary [25, 45], or arbitrary ones [44]. The works of [5, 38] propose to consider the half-wave Gaussian distribution of activations for quantization, resulting from batch normalization [20] and a ReLU [24], which reduces the errors from quantizing activation values. Recent methods learn quantizer parameters for controlling, *e.g.*, clipping ranges [7, 12, 21] and non-uniform quantization intervals [21] or levels [42]. Motivated by this, we design a uniform quantizer, and learn lower and upper bounds of quantization intervals [21]. All the aforementioned approaches exploit STE to handle the derivative of a discretizer. This suggests that our approach can be easily incorporated into these methods, making it possible to boost the performance in a complementary way.

Aside from quantization methods, lots of training techniques have been introduced to enhance the performance of quantized networks. Incremental quantization [43] divides network weights in a layer into two groups of full-precision and quantized ones, and trains a quantized network in an iterative manner by expanding the group of

quantized weights gradually. Progressive quantization [47] decreases bit-widths from high- to low-precision gradually, boosting the performance of a low-precision model. To leverage the knowledge from full-precision models, high-performance networks [30] or layer-wise auxiliary modules [46] are also exploited. Very recently, PROFIT [31] introduces a training strategy, specially designed for quantizing light-weight networks, which progressively freezes learned network weights during an iterative quantization process. These methods also rely on STE, and require heuristic scheduling techniques [31, 43, 47] or additional network weights [30, 46] for training. In contrast, our approach focuses on the backpropagation step of network quantization, improving the performance without bells and whistles.

Similar to ours, recent methods [2, 6, 14, 39] try to tackle the problem of STE. They claim that the STE causes a gradient mismatch problem [14, 39], and introduce soft versions of discretizers, consisting of sigmoid [39] or tanh [14]. These approaches approximate the discretizer (typically using a round function) well, especially when a temperature parameter in sigmoid or tanh functions is large, but the large temperature causes vanishing/exploding gradient problems. Hyperparameters should thus be tuned carefully during training [39]. The proximal method with a regularizer [2] and a meta quantizer using synthetic gradients [6] avoid the use of STE. They are, however, limited to quantizing network weights only, and require a cost-expensive optimization process [2] or additional meta-learning modules [6]. On the contrary, our method can be applied to both weight and activation quantization in an efficient manner with a simple gradient scaling.

Closely related to ours, the works of [10, 11] exploit Hessian information for network quantization. Specifically, they exploit eigenvalues [11] or traces [10] of Hessian matrices to measure the sensitivity of each layer, and allocate different bit-widths to the layers. That is, they leverage the Hessian information to train a mixed-precision network. In contrast to this, we use the trace of a Hessian matrix to adjust a scaling factor for EWGS.

3. Approach

In this section, we introduce our quantization method using EWGS (Sec. 3.1). We then describe how to determine a scaling factor for EWGS (Sec. 3.2).

3.1. Quantization with EWGS

We design a uniform quantizer Q that converts a full-precision input x to a quantized output $Q(x)$, where we denote by x a scalar element of either a weight or an input activation tensor \mathbf{x} in a layer. We learn a quantization interval [14, 21] using lower and upper bounds, denoted by l and u , respectively. Specifically, the quantizer first generates a

full-precision latent value x_n by normalizing and clipping the input value x as follows:

$$x_n = \text{clip}\left(\frac{x-l}{u-l}, 0, 1\right), \quad (1)$$

where $\text{clip}(\cdot, 0, 1)$ is a clipping function with lower and upper bounds of 0 and 1, respectively. Note that weight and/or activation quantizers in every quantized layer use separate parameters for the quantization intervals (*i.e.*, l and u). For b -bit quantization, the latent value x_n is converted to a discrete value x_q using a round function with pre-/post-scaling as follows:

$$x_q = \frac{\text{round}((2^b - 1)x_n)}{2^b - 1}. \quad (2)$$

Finally, the quantizer outputs a quantized weight $Q_W(x)$ or activation $Q_A(x)$ as follows:

$$Q_W(x) = 2(x_q - 0.5), \quad Q_A(x) = x_q, \quad (3)$$

where we restrict the quantized activation $Q_A(x)$ to be non-negative [5] considering the pre-activation by a ReLU. To adjust an output scale of the layer, we train an additional parameter α for each quantized layer, which is multiplied by the output activations of convolutional or fully-connected layers.

The main difficulty of training a quantized network arises from the round function in Eq. (2), since its derivative is zero at almost everywhere. Most quantization methods avoid zero gradients using STE [3]. It approximates the derivative of the round function by an identity function, that is, $\mathcal{G}_{\mathbf{x}_n} = \mathcal{G}_{\mathbf{x}_q}$, where we denote by \mathbf{x}_n and \mathbf{x}_q tensors containing latent and discrete values, respectively, and by $\mathcal{G}_{\mathbf{x}_n}$ and $\mathcal{G}_{\mathbf{x}_q}$ corresponding gradients. Propagating the same gradient from discrete to latent values is, however, sub-optimal for the following reasons: (1) Multiple latent values can produce the same discrete value; (2) The same gradient provided by the discrete value affects differently to each of the latent values. To overcome this problem, we introduce an EWGS method, an effective alternative to STE, defined as follows:

$$g_{x_n} = g_{x_q}(1 + \delta \text{sign}(g_{x_q})(x_n - x_q)), \quad (4)$$

where g_{x_n} and g_{x_q} are the elements of the gradients $\mathcal{G}_{\mathbf{x}_n}$ and $\mathcal{G}_{\mathbf{x}_q}$, corresponding to the partial derivatives of a task loss w.r.t x_n and x_q , respectively. $\text{sign}(\cdot)$ is a signum function and $\delta \geq 0$ is a scaling factor. EWGS adjusts the gradient element of discrete values g_{x_q} adaptively using the sign of the element, $\text{sign}(g_{x_q})$, and a discretization error, $x_n - x_q$. Note that STE is a special case of EWGS, that is, Eq. (4) corresponds to STE, when a scaling factor δ is zero.

We visualize in Fig. 2 1-D examples illustrating an effect of EWGS. We can see that EWGS encourages a gradient element for the discrete value g_{x_q} to decrease with the scale

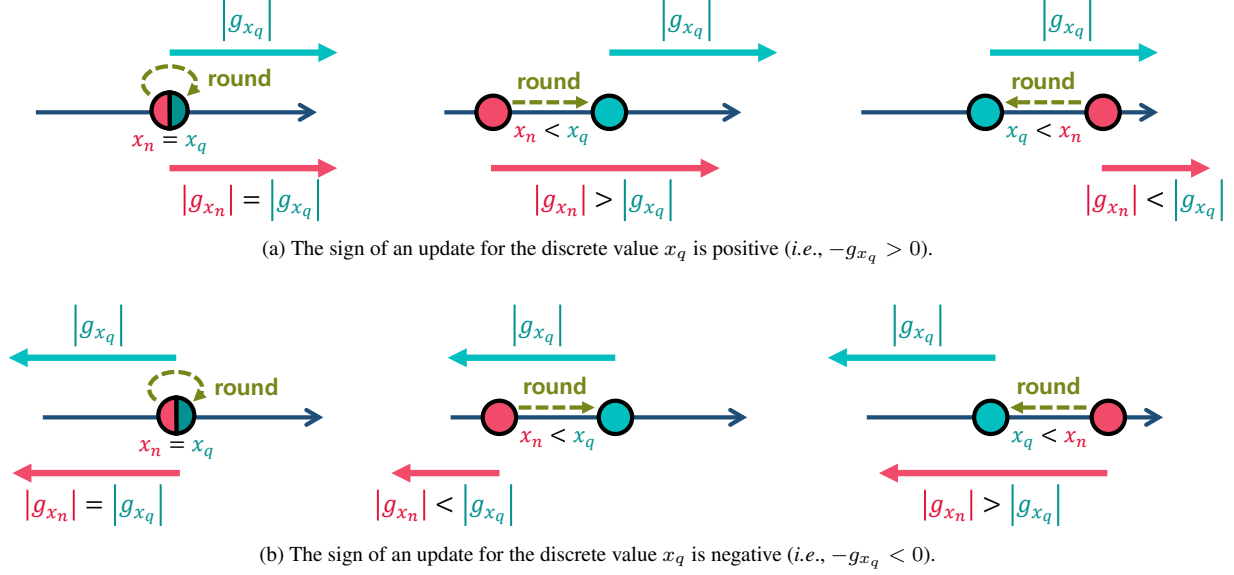


Figure 2: 1-D illustrations of EWGS. We visualize a latent value x_n and a discrete value x_q , by red and cyan circles, respectively, where the discrete value is obtained by applying a round function (a dashed arrow) to the latent value. We also visualize their update vectors by solid arrows with corresponding colors, and we denote by $|g_{x_n}|$ and $|g_{x_q}|$ the magnitudes of the update vectors for x_n and x_q , respectively. For each (a) and (b), we present three cases, where the latent value x_n is equal to (left), smaller than (middle), and larger than (right) the discrete one x_q . EWGS scales up the gradient element for the discrete value g_{x_q} , when a latent value x_n requires a larger magnitude of an update, compared to the discrete one x_q (e.g., (a)-middle or (b)-right), and scaling down in the opposite case (e.g., (a)-right or (b)-middle). When the latent value x_n is equal to the discrete value x_q , it propagates the same gradient element similar to STE (e.g., (a)-left or (b)-left). (Best viewed in color.)

of $(1 + \delta \text{sign}(g_{x_q})(x_n - x_q))$ to update the latent value x_n , when the latent value x_n to update is already located farther than the discrete value x_q in the direction of change (i.e., $-\text{sign}(g_{x_q})$) as shown in Fig. 2a (right) and Fig. 2b (middle), and to increase in the opposite case as shown in Fig. 2a (middle) and Fig. 2b (right). Note that we use non-negative values for the scaling factor δ , since negative ones lead to opposite effects. To sum up, EWGS resolves discrepancies between latent and discrete values during backpropagation by considering discretization errors between these values and their direction of change. As will be shown in Sec. 4.3, this not only stabilizes the training of a quantized network but also encourages better convergence compared to STE.

3.2. Scaling factor for EWGS

It is crucial to determine a scaling factor δ , since an improper value would hinder the training process, and weight/activation quantizers in different layers may require different degrees of scaling. Let us consider the following equation:

$$\begin{aligned} g_{x_n} &= g_{x_q} + \frac{g_{x_n} - g_{x_q}}{x_n - x_q} (x_n - x_q) \\ &= g_{x_q} + \frac{g_{x_q+\epsilon} - g_{x_q}}{\epsilon} (x_n - x_q), \end{aligned} \quad (5)$$

where $\epsilon = x_n - x_q$ is a discretization error from the round function in Eq. (2). Since an absolute value of the error is

bounded by a small number i.e., $|\epsilon| \leq \frac{0.5}{2^b-1}$, we assume that the error is small enough to approximate Eq. (5) as follows:

$$g_{x_n} \approx g_{x_q} + g'_{x_q} (x_n - x_q), \quad (6)$$

where $g'_{x_q} = \frac{\partial g_{x_q}}{\partial x_q}$ is a second-order derivative of a task loss w.r.t the discrete value x_q . This can be represented as follows:

$$g_{x_n} \approx g_{x_q} \left(1 + \frac{g'_{x_q}}{|g_{x_q}|} \text{sign}(g_{x_q})(x_n - x_q) \right), \quad (7)$$

which corresponds to EWGS in Eq. (4). This suggests that we can set the scaling factor δ as $\frac{g'_{x_q}}{|g_{x_q}|}$, but calculating an exact Hessian matrix H to obtain the second-order derivative g'_{x_q} is computationally demanding. We instead approximate the second-order derivative by an average of diagonal elements in the Hessian matrix H , with an assumption that the main diagonal dominates the matrix H , and the discrete values \mathbf{x}_q obtained from the same weight or activation quantizer in a layer influence similarly to the loss function [10, 11]. To this end, we compute a Hessian trace with an efficient algorithm [10, 40] using the Hutchinson's method [1]:

$$\begin{aligned} \text{Tr}(H) &= \text{Tr}(HI) = \text{Tr}(H\mathbb{E}[\mathbf{v}\mathbf{v}^T]) \\ &= \mathbb{E}[\text{Tr}(H\mathbf{v}\mathbf{v}^T)] = \mathbb{E}[\mathbf{v}^T H \mathbf{v}], \end{aligned} \quad (8)$$

Algorithm 1 Forward and backward propagations in a quantizer using EWGS.

- 1: **Hyperparameter:** a quantization bit-width b ; an update period of the scaling factor k .
 - 2: **Parameter:** lower and upper bounds of a quantization interval, denoted by $l, u \in \mathbb{R}$, respectively; a scaling factor for EWGS $\delta \in \mathbb{R}$.
 - 3: **Input:** a full-precision input tensor $\mathbf{x} \in \mathbb{R}^N$ containing either weights or activations, where N is the number of elements in the tensor.
 - 4: **Output:** a quantized tensor $Q(\mathbf{x}) \in \mathbb{R}^N$.
 - 5: **Forward Propagation**
 - 6: Compute latent values [Eq. (1)]:

$$\mathbf{x}_n = \text{clip}\left(\frac{\mathbf{x}-l}{u-l}, 0, 1\right).$$
 - 7: Compute discrete values [Eq. (2)]:

$$\mathbf{x}_q = \frac{\text{round}((2^b-1)\mathbf{x}_n)}{2^b-1}.$$
 - 8: Compute quantized output values [Eq. (3)]:

$$Q(\mathbf{x}) = \begin{cases} 2(\mathbf{x}_q - 0.5), & \text{if } \mathbf{x} \text{ is a weight tensor.} \\ \mathbf{x}_q, & \text{if } \mathbf{x} \text{ is an activation tensor.} \end{cases}$$
 - 9: **Backward Propagation**
 - 10: Obtain the gradient of discrete values $\mathcal{G}_{\mathbf{x}_q}$ via backpropagation.
 - 11: Calculate the gradient of latent values $\mathcal{G}_{\mathbf{x}_n}$ using EWGS [Eq. (4)]:

$$\mathcal{G}_{\mathbf{x}_n} = \mathcal{G}_{\mathbf{x}_q} \odot (1 + \delta \text{sign}(\mathcal{G}_{\mathbf{x}_q}) \odot (\mathbf{x}_n - \mathbf{x}_q)),$$
 where \odot is element-wise multiplication and $\text{sign}(\cdot)$ applies the signum function to each element.
 - 12: Propagate the gradient to the input.
 - 13: **Scaling Factor Update**
 - 14: Update the scaling factor δ using Eq. (10) for every k iterations.
-

where I is an identity matrix, \mathbb{E} is an expectation operator, and \mathbf{v} is a random vector drawn from the Rademacher distribution, satisfying $\mathbb{E}[\mathbf{v}\mathbf{v}^T] = I$. This implies that we can estimate the trace of a Hessian matrix $\text{Tr}(H)$ with $\mathbb{E}[\mathbf{v}^T H \mathbf{v}]$, where we can obtain $H\mathbf{v}$ efficiently without forming an exact Hessian matrix as follows:

$$\frac{\partial \mathcal{G}_{\mathbf{x}_q}^T \mathbf{v}}{\partial \mathbf{x}_q} = \frac{\partial \mathcal{G}_{\mathbf{x}_q}^T}{\partial \mathbf{x}_q} \mathbf{v} + \mathcal{G}_{\mathbf{x}_q}^T \frac{\partial \mathbf{v}}{\partial \mathbf{x}_q} = \frac{\partial \mathcal{G}_{\mathbf{x}_q}^T}{\partial \mathbf{x}_q} \mathbf{v} = H\mathbf{v}. \quad (9)$$

We then define the scale parameter δ as follows:

$$\delta = \frac{\text{Tr}(H)/N}{G}, \quad (10)$$

where N is the number of diagonal elements in the Hessian matrix and G is a gradient representative determined from a distribution of the gradients $\mathcal{G}_{\mathbf{x}_q}$. Based on Eq. (7), we could take an average over the absolute values of gradient elements *i.e.*, $\mathbb{E}[|g_{x_q}|]$ for setting G , but we empirically found that most gradients are concentrated near zero, such that the average value tends to be biased to small gradient elements. We instead set G to a sufficiently large value. A plausible

reason is that considering large gradient elements is more important, since they dominate the training. Specifically, we use $3\sigma(\mathcal{G}_{\mathbf{x}_q})$ as the gradient representative G , where $\sigma(\cdot)$ computes a standard deviation. It enables finding a sufficiently large gradient element (*e.g.*, $3\sigma(\cdot)$ accounts for roughly 99 percent of data in case of the Gaussian distribution). We could take the maximum over the absolute values of gradient elements, but it often corresponds to an outlier of a distribution.

Assuming that the loss function is locally convex, we take a non-negative value for the scaling factor (*i.e.*, $\max(0, \delta)$), which coincides with the condition in Eq. (4). We use individual scaling factors for all weight and activation quantizers in a network, and update them periodically during training for efficiency. We summarize in Algorithm 1 an overall quantization procedure of our approach.

4. Experiments

In this section, we describe our experimental settings (Sec. 4.1) and evaluate our method on image classification (Sec. 4.2). We then present a detailed analysis on EWGS (Sec. 4.3).

4.1. Experimental settings

Dataset. We perform extensive experiments on standard benchmarks for image classification, including CIFAR-10 [23] and ImageNet (ILSVRC-2012) [8]. The CIFAR-10 dataset contains images of size 32×32 , consisting of 50K training and 10K test images with 10 classes. The ImageNet dataset includes roughly 1.2M training and 50K validation images with 1K classes. We report the top-1 classification accuracy for both datasets.

Network architectures. We use network architectures of ResNet-20 [16] on CIFAR-10, and ResNet-18, ResNet-34 and MobileNet-V2 [35] on ImageNet. We do not modify the network architectures for fair comparison. We insert weight and/or activation quantizers right before the convolutional or fully-connected operators in every layer to quantize. Following the standard experimental protocol in [34, 42], we do not quantize the first and the last layers unless otherwise specified. We initialize network weights with pretrained full-precision models, which are readily available in PyTorch [32] (ResNet-18, ResNet-34, and MobileNet-V2) or trained by ourselves (ResNet-20).

Initialization. We initialize the lower and upper bounds of a quantization interval, l and u , respectively, by considering the distribution of quantizer inputs, such that the interval covers roughly 99% of the input values, to use a set of discrete levels effectively. Specifically, for each weight quantizer, the lower and upper bounds are initialized by $-3\sigma(\mathbf{w})$ and $3\sigma(\mathbf{w})$, respectively, where \mathbf{w} is a weight tensor in a layer. Considering that input activations typically follow

Methods \ W/A	1/1	1/2	2/2	3/3	4/4	1/32	2/32	32/32
XNOR [34]	51.2 (-18.1)	-	-	-	-	60.8 (-8.5)	-	69.3
PACT [7]	-	-	64.4 (-5.8)	68.1 (-2.1)	69.2 (-1.0)	-	-	70.2
LQ-Net [42]	-	62.6 (-7.7)	64.9 (-5.4)	68.2 (-2.1)	69.3 (-1.0)	-	68.0 (-2.3)	70.3
QIL [21]	-	-	65.7 (-4.5)	69.2 (-1.0)	70.1 (-0.1)	-	68.1 (-2.1)	70.2
QuantNet [39]	53.6 (-16.7)	63.4 (-6.9)	-	-	-	66.5 (-3.8)	69.1 (-1.2)	70.3
DSQ [14]	-	-	65.2 (-4.7)	68.7 (-1.2)	69.6 [†] (-0.3)	63.7 (-6.2)	-	69.9
LSQ [4, 12]	-	-	66.7 (-3.4)	69.4 (-0.7)	70.7 (+0.6)	-	-	70.1
LSQ+ [4]	-	-	66.8 (-3.3)	69.3 (-0.8)	70.8 (+0.7)	-	-	70.1
IRNet [33]	-	-	-	-	-	66.5 (-3.1)	-	69.6
Ours	55.3 (-14.6)	64.4 (-5.5)	67.0 (-2.9)	69.7 (-0.2)	70.6 (+0.7)	67.3 (-2.6)	69.6 (-0.3)	69.9

Table 1: Quantitative comparison of top-1 validation accuracy on ImageNet [8] using the ResNet-18 [16] architecture. We report results for quantized networks and their full-precision versions. W/A represents bit-widths of weights (W) and activations (A). The numbers in brackets indicate the performance drops or gains compared to the full-precision models. [†]: all layers including the first and the last layers are quantized.

the half-wave Gaussian distribution [5], we initialize the lower and upper bounds in each activation quantizer with 0 and $\frac{3\sigma(\mathbf{a})}{\sqrt{1-2/\pi}}$, respectively, where \mathbf{a} is an input activation tensor. An output scale α in every quantized layer is initialized by $\frac{\mathbb{E}(|o|)}{\mathbb{E}(|o_q|)}$, where o and o_q are convolution (or matrix multiplication) outputs computed with full-precision and quantized representations, respectively. Note that we initialize these parameters during the first forward pass. We set scaling factors δ for EWGS to 0 initially, and update them for every 1 epoch on ImageNet and 10 epochs on CIFAR-10.

Training details. Initial learning rates for network weights are set to 1e-3, 1e-2, 1e-2, and 5e-3 for ResNet-20, ResNet-18, ResNet-34, and MobileNet-V2, respectively. We set a learning rate for the quantizer parameters (*i.e.*, interval parameters, l and u , and output scales α) to 1e-5, smaller than those of the network weights [21]. We use a cosine annealing technique [29] for learning rate decay. Following the training settings in [21, 31, 33], we use the SGD optimizer to train the network weights, except ResNet-20 on CIFAR-10 using the Adam optimizer [22], with weight decay of 4e-5 for MobileNet-V2 and 1e-4 for the others. The quantizer parameters are trained with the Adam optimizer without weight decay. We train ResNet-20 for 400 epochs on CIFAR-10 with a batch size of 256. ResNet-18, ResNet-34, and MobileNet-V2 are trained for 100 epochs on ImageNet with batch sizes of 256, 256, and 100, respectively.

4.2. Results

We compare in Table 1 the validation accuracy on ImageNet [8] using the ResNet-18 [16] architecture under various bit-width settings. All numbers for other methods, except for LSQ [12], are taken from corresponding papers including the performance of full-precision models. Note that LSQ reports the results with a pre-activation structure [17]

of ResNet, which is different from ours. We thus take the results from the work of [4] in which LSQ is reproduced using the same network structure as ours. We summarize the findings from Table 1 as follows: (1) Our quantization method with EWGS achieves the state of the art. For 4-bit weights and 4-bit activations, our method shows the classification accuracy slightly lower than LSQ+ [4], but the performance gain w.r.t the full-precision model is on a par with LSQ+. In particular, our model achieves the performance comparable to the full-precision one with only 3-bit representations. (2) Our method performs better than QuantNet [39] and DSQ [14], which attempt to address the problem of STE using soft quantizers, indicating that EWGS is a better alternative to STE than soft quantizers. (3) Our method exploiting EWGS brings significant performance improvement in a binary setting, surpassing other methods including the ones specially designed for binary quantization [33, 34]. This suggests that EWGS works favorably even with large discretization errors. (4) We verify the effectiveness of our method over a wide range of quantization bit-widths, outperforming the state of the art consistently, whereas other methods report the results selectively in few settings.

We show in Tables 2 and 3 quantization results for ResNet-34 [16] and MobileNet-V2 [35], respectively, on ImageNet [8]. As mentioned earlier, LSQ [12] uses a different network structure¹, but we include its performance in Table 2 in order to compare relative performance drops or gains for network quantization. We can observe from Table 2 similar findings in Table 1. Our method outperforms the state of the art over all bit-width settings. With 3-bit weights and 3-bit activations, the quantized network trained with our method does not degrade the performance at all, compared with the full-precision model. Ours also gives better results than LSQ in terms of performance drops or

¹The full-precision baseline of ResNet-34 used in LSQ shows the top-1 validation accuracy of 74.1, which is higher than that of our full-precision baseline (73.3).

Methods \ W/A	1/1	1/2	2/2	3/3	4/4	1/32	32/32
LSQ [12]	-	-	71.6 (-2.5)	73.4 (-0.7)	74.1 (+0.0)	-	74.1
ABC-Net [26]	52.4 (-20.9)	-	-	66.7 (-6.6)	-	-	73.3
LQ-Net [42]	-	66.6 (-7.2)	69.8 (-4.0)	71.9 (-1.9)	-	-	73.8
QIL [21]	-	-	70.6 (-3.1)	73.1 (-0.6)	73.7 (+0.0)	-	73.7
DSQ [14]	-	-	70.0 (-3.8)	72.5 (-1.3)	72.8 [†] (-1.0)	-	73.8
IR-Net [33]	-	-	-	-	-	70.4 (-2.9)	73.3
Ours	61.5 (-11.8)	69.6 (-3.7)	71.4 (-1.9)	73.3 (+0.0)	73.9 (+0.6)	72.2 (-1.1)	73.3

Table 2: Quantitative comparison of top-1 validation accuracy on ImageNet [8] using the ResNet-34 [16] architecture. We report results for quantized networks and their full-precision versions. [†]: all layers including the first and the last layers are quantized.

Methods \ W/A	4/4	32/32
PACT [7, 37]	61.4 (-10.4)	71.8
DSQ [14]	64.8 (-7.1)	71.9
PROFIT [31]	71.6[†] (-0.3)	71.9
Ours	70.3 [†] (-1.6)	71.9

Table 3: Quantitative comparison of top-1 validation accuracy on ImageNet [8] using the MobileNet-V2 [35] architecture. We quantize MobileNet-V2 using our method with the training hyperparameters and the network structures used in PROFIT [31]. We report results for quantized networks and their full-precision versions. [†]: all layers including the first and the last layers are quantized.

gains after quantization. We can see from Table 3 that our model performs better than PACT [7, 37] and DSQ [14], but it is slightly outperformed by PROFIT [31] for 4-bit quantization of both weights and activations. Note that PROFIT exploits many training heuristics, such as knowledge distillation [18], progressive quantization [47], an exponential moving average of weights, batch normalization post-training, and iterative training with incremental weight freezing. We achieve a comparable result using a simple gradient scaling without bells and whistles, which confirms that EWGS is simple yet effective method for network quantization. Moreover, PROFIT is effective to quantize light-weight networks only, while ours can be applied to various network architectures under a wide range of bit-widths.

4.3. Discussion

Analysis on scaling factor. We show in Fig. 3 variations of scaling factors at a particular layer during training. We can see that scaling factors oscillate within a certain range without diverging or changing drastically. This suggests that we could consider the scaling factors as hyperparameters, fixed regardless of training epochs, instead of updating them frequently. Figure 3 also shows scaling factors for each layer averaged over epochs. We can observe that scaling factors for weights and activations tend to decrease and increase, respectively, for deeper layers, except the 7th, 12th, and 17th layers. They correspond to convolutional lay-

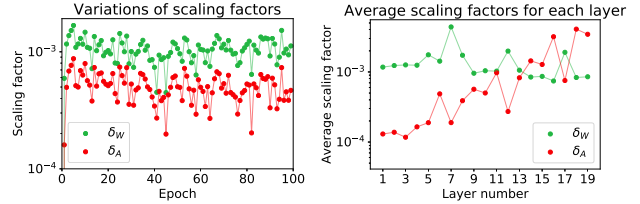


Figure 3: Variations of scaling factors in the 10th quantized layer over training epochs (left); Scaling factors for each layer averaged over epochs (right). We visualize scaling factors for both weight and activation, denoted by δ_W and δ_A , respectively. We use the ResNet-18 [16] architecture for binary quantization. (Best viewed in color.)

Scaling factor	Top-1 accuracy (full-precision: 91.4)
Eq. (10) with $G = 3\sigma(\mathcal{G}_{x_q})$	85.6
Eq. (10) with $G = \max(\mathcal{G}_{x_q})$	85.5
Eq. (10) with $G = \mathbb{E}[g_{x_q}]$	83.1
Fixed (1e-1)	60.9
Fixed (1e-3)	85.3
Fixed (1e-5)	85.0
Fixed (0) = STE	84.7

Table 4: Quantitative comparison for different configurations of scaling factors. We binarize both weights and activations of ResNet-20 [16] on CIFAR-10 [23], and report the top-1 test accuracy. The first three rows use scaling factors obtained by Eq. (10) but with different gradient representatives G . The last four rows use fixed hyperparameters, specified in brackets, for both scaling factors of weights and activations in all quantized layers.

ers with a filter size of 1×1 and a stride of 2, designed to reduce the size of residuals in the residual blocks, having the different behavior compared to other plain layers. This confirms that our strategy leveraging Hessian information captures different characteristics across layers, providing an individual scaling factor for each layer.

We compare quantization results for different configurations of scaling factors in Table 4. We obtain the results by binarizing weights and activations of ResNet-20 [16] on CIFAR-10 [23]. The first three rows show quantization results for different gradient representatives G in Eq. (10). Overall, our method shows better results with large gradi-

Network architectures	W/A	Quant. methods	Backward methods	Top-1 acc.
ImageNet				
ResNet-18 [16]	1/1	Ours	STE EWGS	54.6 55.3
	1/32	Ours	STE EWGS	66.3 67.3
MobileNet-V2 [35]	4/4 [†]	Ours	STE EWGS	69.2 70.3
	4/4 [†]	PROFIT [31]	STE EWGS [‡]	69.2* 70.0
CIFAR-10				
ResNet-20 [16]	1/1	Ours	STE EWGS	84.7 85.6
	1/1	DoReFa [44]	STE EWGS [‡]	84.9* 85.9
	1/32 [†]	DoReFa [44]	STE EWGS [‡]	89.7* 90.3

Table 5: Quantitative comparison of STE and EWGS. We use ResNet-18 [16] and MobileNet-V2 [35] on ImageNet [8], and ResNet-20 on CIFAR-10 [23]. We report the top-1 validation and test accuracies for ImageNet and CIFAR-10, respectively. For MobileNet-V2 and ResNet-20, we also compare the performance with different quantization methods, such as PROFIT [31] and DoReFa-Net [44]. [†]: all layers including the first and the last layers are quantized; [‡]: weight and activation scaling factors for EWGS in all quantized layers are fixed to 0.01; *: models reproduced by ourselves.

ent elements, *e.g.*, three standard deviations and the maximum over absolute values in the first and second rows, respectively, than the small one, *e.g.*, an average in the third row. A reason is that large gradients mainly influence the training process, but an average value is usually biased to small gradient elements, as discussed in Sec. 3.2. The last four rows compare the results with fixed scaling factors. We use the same scaling factor for both weight and activation quantizers in all quantized layers. We can see that our method achieves the performance comparable to the best result (85.6 vs. 85.3), and even outperforms STE (84.7 vs. 85.3), if the scaling factor is properly set. Otherwise, the performance is degraded (*e.g.*, with the scaling factor of 1e-1) or becomes similar to the one for STE, especially with an extremely small scaling factor (*e.g.*, 1e-5). This suggests that EWGS is also effective with a single scaling factor, but the value should be carefully chosen.

Performance comparison with STE. We compare in Table 5 the performance of EWGS and STE with different combinations of network architectures, quantization methods, and bit-widths. Specifically, we use different quantization methods, including PROFIT [31], DoReFa-Net [44], and ours, and exploit either EWGS or STE for backpropagation. We then use them to quantize ResNet-18 [16], MobilNet-V2 [35] and ResNet-20. EWGS gives better re-

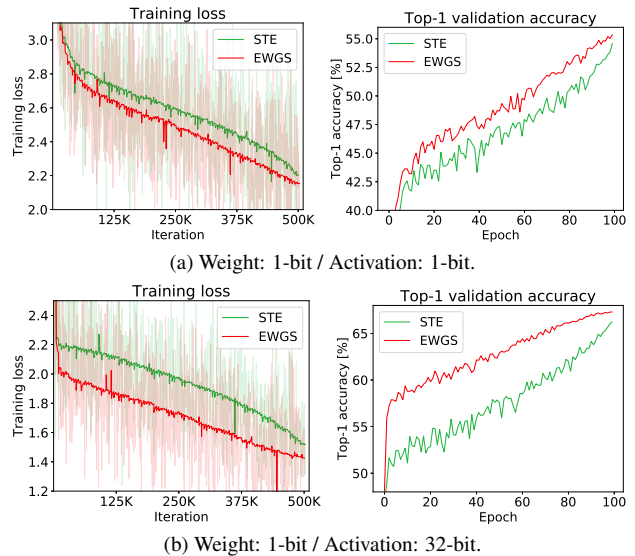


Figure 4: Training losses and validation accuracies for binarized networks using STE and EWGS. We use ResNet-18 [16] to quantize (a) both weights and activations and (b) weights only, and show the results on ImageNet. (Best viewed in color.)

sults than STE within our framework, achieving about 1% accuracy gains over STE consistently, regardless of the network architectures. It also outperforms STE by a large margin for other quantization methods, such as PROFIT [31] and DoReFa-Net [44], demonstrating the generalization ability of EWGS. The accuracy of PROFIT is slightly lower than the one reported in the paper, possibly because we do not use the progressive quantization technique [47]. We show in Fig. 4 the training losses and validation accuracies for binarizing the ResNet-18 [16] architecture in Table 5. We can clearly see that training quantized networks with EWGS is better in terms of stability and accuracy, compared to STE. The networks with EWGS achieve lower losses and higher accuracies, which is significant especially for weight-only quantization. These results confirm once more the effectiveness of EWGS.

5. Conclusion

We have introduced an EWGS method that adjusts gradients and scaling factors adaptively for each layer. The various CNN architectures quantized by our method show state-of-the-art results for a wide range of bit-widths. We have shown that EWGS boosts the quantization performance of other methods exploiting STE, without bells and whistles, demonstrating the effectiveness and generalization ability of our approach to scaling gradients adaptively for backpropagation. We believe that EWGS could be an effective alternative to STE for network quantization.

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

References

- [1] Haim Avron and Sivan Toledo. Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. *Journal of the ACM (JACM)*, 58(2):1–34, 2011. [2](#), [4](#)
- [2] Yu Bai, Yu-Xiang Wang, and Edo Liberty. ProxQuant: Quantized neural networks via proximal operators. In *ICLR*, 2019. [3](#)
- [3] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv*, 2013. [1](#), [2](#), [3](#)
- [4] Yash Bhalgat, Jinwon Lee, Markus Nagel, Tijmen Blankevoort, and Nojun Kwak. LSQ+: Improving low-bit quantization through learnable offsets and better initialization. In *CVPR Workshop*, 2020. [2](#), [6](#)
- [5] Zhaowei Cai, Xiaodong He, Jian Sun, and Nuno Vasconcelos. Deep learning with low precision by half-wave gaussian quantization. In *CVPR*, 2017. [2](#), [3](#), [6](#)
- [6] Shangyu Chen, Wenya Wang, and Sinno Jialin Pan. MetaQuant: Learning to quantize by learning to penetrate non-differentiable quantization. In *NeurIPS*, 2019. [2](#), [3](#)
- [7] Jungwook Choi, Zhuo Wang, Swagath Venkataramani, Pierce I-Jen Chuang, Vijayalakshmi Srinivasan, and Kailash Gopalakrishnan. PACT: Parameterized clipping activation for quantized neural networks. *arXiv*, 2018. [2](#), [6](#), [7](#)
- [8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. [2](#), [5](#), [6](#), [7](#), [8](#)
- [9] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, 2014. [1](#)
- [10] Zhen Dong, Zhewei Yao, Yaohui Cai, Daiyaan Arfeen, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. HAWQ-V2: Hessian aware trace-weighted quantization of neural networks. *arXiv*, 2019. [3](#), [4](#)
- [11] Zhen Dong, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. HAWQ: Hessian aware quantization of neural networks with mixed-precision. In *ICCV*, 2019. [3](#), [4](#)
- [12] Steven K Esser, Jeffrey L McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S Modha. Learned step size quantization. In *ICLR*, 2020. [2](#), [6](#), [7](#)
- [13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. [1](#)
- [14] 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. [3](#), [6](#), [7](#)
- [15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. [1](#)
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. [1](#), [5](#), [6](#), [7](#), [8](#)
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, 2016. [6](#)
- [18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS Workshop*, 2015. [7](#)
- [19] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. In *NeurIPS*, 2016. [1](#)
- [20] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. [2](#)
- [21] 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. [2](#), [3](#), [6](#), [7](#)
- [22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. [6](#)
- [23] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, 2009. [5](#), [7](#), [8](#)
- [24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *NeurIPS*, 2012. [1](#), [2](#)
- [25] Fengfu Li, Bo Zhang, and Bin Liu. Ternary weight networks. *arXiv*, 2016. [2](#)
- [26] Xiaofan Lin, Cong Zhao, and Wei Pan. Towards accurate binary convolutional neural network. In *NeurIPS*, 2017. [7](#)
- [27] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *ECCV*, 2016. [1](#)
- [28] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. [1](#)
- [29] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. [6](#)
- [30] Asit Mishra and Debbie Marr. Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy. In *ICLR*, 2018. [2](#), [3](#)
- [31] Eunhyeok Park and Sungjoo Yoo. PROFIT: A novel training method for sub-4-bit MobileNet models. In *ECCV*, 2020. [2](#), [3](#), [6](#), [7](#), [8](#)
- [32] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. 2017. [5](#)
- [33] 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](#)
- [34] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. XNOR-Net: ImageNet classification using binary convolutional neural networks. In *ECCV*, 2016. [1](#), [2](#), [5](#), [6](#)
- [35] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. [5](#), [6](#), [7](#), [8](#)
- [36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2014. [1](#)
- [37] Kuan Wang, Zhijian Liu, Yujun Lin, Ji Lin, and Song Han. HAQ: Hardware-aware automated quantization with mixed precision. In *CVPR*, 2019. [7](#)

- [38] Peisong Wang, Qinghao Hu, Yifan Zhang, Chunjie Zhang, Yang Liu, and Jian Cheng. Two-step quantization for low-bit neural networks. In *CVPR*, 2018. [2](#)
- [39] Jiwei Yang, Xu Shen, Jun Xing, Xinmei Tian, Houqiang Li, Bing Deng, Jianqiang Huang, and Xian-sheng Hua. Quantization networks. In *CVPR*, 2019. [2](#), [3](#), [6](#)
- [40] Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael Mahoney. PyHessian: Neural networks through the lens of the Hessian. In *ICML Workshop*, 2020. [2](#), [4](#)
- [41] Penghang Yin, Jiancheng Lyu, Shuai Zhang, Stanley Osher, Yingyong Qi, and Jack Xin. Understanding straight-through estimator in training activation quantized neural nets. In *ICLR*, 2019. [2](#)
- [42] 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](#), [5](#), [6](#), [7](#)
- [43] 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](#), [3](#)
- [44] 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*, 2016. [2](#), [8](#)
- [45] Chenzhuo Zhu, Song Han, Huizi Mao, and William J Dally. Trained ternary quantization. In *ICLR*, 2017. [2](#)
- [46] Bohan Zhuang, Lingqiao Liu, Mingkui Tan, Chunhua Shen, and Ian Reid. Training quantized neural networks with a full-precision auxiliary module. In *CVPR*, 2020. [2](#), [3](#)
- [47] Bohan Zhuang, Chunhua Shen, Mingkui Tan, Lingqiao Liu, and Ian Reid. Towards effective low-bitwidth convolutional neural networks. In *CVPR*, 2018. [2](#), [3](#), [7](#), [8](#)