MIXED PRECISION DNNs: ALL YOU NEED IS A GOOD PARAMETRIZATION

Stefan Uhlich; Lukas Mauch; Fabien Cardinaux; Kazuki Yoshiyama
Javier Alonso García, Stephen Tiedemann, Thomas Kemp
Sony Europe B.V., Germany
firstname.lastname@sony.com

Akira Nakamura
Sony Corporate, Japan
akira.b.nakamura@sony.com

ABSTRACT

Efficient deep neural network (DNN) inference on mobile or embedded devices typically involves quantization of the network parameters and activations. In particular, mixed precision networks achieve better performance than networks with homogeneous bitwidth for the same size constraint. Since choosing the optimal bitwidths is not straight forward, training methods, which can learn them, are desirable. Differentiable quantization with straight-through gradients allows to learn the quantizer’s parameters using gradient methods. We show that a suited parametrization of the quantizer is the key to achieve a stable training and a good final performance. Specifically, we propose to parametrize the quantizer with the step size and dynamic range. The bitwidth can then be inferred from them. Other parametrizations, which explicitly use the bitwidth, consistently perform worse. We confirm our findings with experiments on CIFAR-10 and ImageNet and we obtain mixed precision DNNs with learned quantization parameters, achieving state-of-the-art performance.

1 INTRODUCTION

Quantized DNNs apply quantizers $Q: \mathbb{R} \rightarrow \{q_1, \ldots, q_I\}$ to discretize the weights and/or activations of a DNN (Han et al. 2015; Zhou et al. 2017; Li et al. 2016; Liu & Mattina 2019; Cardinaux et al. 2018; Jain et al. 2019; Bai et al. 2018). They require considerably less memory and have a lower computational complexity, since discretized values $\{q_1, \ldots, q_I\}$ can be stored, multiplied and accumulated efficiently. This is particularly relevant for inference on mobile or embedded devices with limited computational power.

However, gradient based training of quantized DNNs is difficult, as the gradient of a quantization function vanishes almost everywhere, i.e., backpropagation through a quantized DNN almost always returns a zero gradient. Different solutions to this problem have been proposed in the literature:

A first possibility is to use DNNs with stochastic weights from a categorical distribution and to optimize the evidence lower bound (ELBO) to obtain an estimate of the posterior distribution of the weights. As proposed in (Jang et al. 2016; Maddison et al. 2016; Louizos et al. 2019), the categorical distribution can be relaxed to a concrete distribution – a smoothed approximation of the categorical distribution – such that the ELBO becomes differentiable under reparametrization.

A second possibility is to use the straight through estimator (STE) (Bengio et al. 2013). STE allows the gradients to be backpropagated through the quantizers and, thus, the network weights can be adapted with standard gradient descent (Hubara et al. 2016). Compared to STE based methods, stochastic methods suffer from large gradient variance, which makes training of large quantized DNNs difficult. Therefore, STE based methods are more popular in practice.

More recent research (Jain et al. 2019; Esser et al. 2019; Wang et al. 2018; Elthakeb et al. 2018) focuses on methods which can also learn the optimal quantization parameters, e.g., the stepsize, dynamic range and bitwidth, in parallel to the network weights. This is a promising approach as DNNs with learned quantization parameters almost always outperform DNNs with handcrafted ones.
Recently, and in parallel to our work, [Jain et al., 2019] explored the use of STE to define the gradient with respect to the quantizers’s dynamic range. The authors applied a per-tensor quantization and used the dynamic range as an additional trainable parameter also learned with gradient descent. Similarly, [Esser et al., 2019] learned the stepsize using gradient descent. However, neither of them learned the optimal bitwidth of the quantizers.

One approach was proposed in [Wang et al., 2018; Elthakeb et al., 2018]. They learn the bitwidth with reinforcement learning, i.e., they learn an optimal bitwidth assignment policy. Their experiments show that a DNN with a learned and heterogeneous bitwidth assignment outperforms quantized DNNs with a homogeneous bitwidth assignment. However, such methods have a high computational complexity as the bitwidth policy must be learned, which involves training many quantized DNNs.

In this paper, we will use the STE approach and show that the quantizer’s parameters, including the bitwidth, can be learned with gradient methods if a good parametrization is chosen. Specifically, we show that directly learning the bitwidth is not optimal. Instead, we propose to learn the stepsize and dynamic range. The bitwidth can then be inferred from them. Compared to [Wang et al., 2018; Elthakeb et al., 2018], our method has the advantage that training quantized DNNs has nearly the same computational complexity as standard float32 training.

The contributions of this paper are:

1. We show that there are three different parametrizations for uniform and power-of-two quantization and that, in both cases, one of them has gradients particularly well suited to train quantized DNNs. The other parametrizations have the problem of yielding gradients with an unbounded gradient norm and coupled components.

2. Using this parametrization, we are able to learn all quantization parameters for DNNs with per-tensor quantization and global memory constraints. We formulate the training as a constrained optimization problem, where the quantized DNN is constrained not to exceed a given overall memory budget, and show how to solve it in a penalty framework.

3. We confirm our findings with experiments on CIFAR-10 and ImageNet. For example, we train a heterogeneously quantized MobileNetV2 on ImageNet requiring a total of only 1.65MB to store the weights and only 0.57MB to store its largest feature map. This is equivalent to a homogenous 4-bit quantization of both weights and activations. However, our network learns to allocate the bitwidth heterogeneously in an optimal way. Our MobileNetV2 achieves an error of 30.26% compared to 29.82% for the floating point baseline. This is state-of-the-art for such a heavily quantized MobileNetV2.

We use the following notation throughout this paper: \( x, x, X \) and \( \mathcal{X} \) denote a scalar, a (column) vector, a matrix and a tensor with three or four dimensions, respectively; \( \lfloor \cdot \rfloor \) and \( \lceil \cdot \rceil \) are the floor and ceiling operators. Finally, \( \delta(\cdot) \) denotes the Dirac delta function.

## 2 Choosing a Quantization Parametrization

Let \( Q(x; \theta) \) be a quantizer with the parameters \( \theta \), which maps \( x \in \mathbb{R} \) to discrete values \( \{q_1, \ldots, q_I\} \). In this section, we compare different parametrizations of \( Q(x; \theta) \) for uniform quantization and power-of-two quantization and analyze how well the corresponding straight-through gradient estimates \( \partial_{\theta} Q(x; \theta) \) and \( \nabla_{\theta} Q(x; \theta) \) are suited to optimize the quantizer parameters \( \theta \). Our key result is, that the training of quantized DNNs which learns both, the optimal quantized weights and the optimal quantization parameters \( \theta \), is very sensitive to the choice of the parametrization of the quantizers. From an optimization point of view, it is best to parametrize the quantizer \( Q(x; \theta) \) with the stepsize \( d \) and the dynamic range \( q_{\text{max}} \) as it leads to gradients with stable norms. Doing so, we can use standard gradient descent to learn the quantization parameters and do not need to use stochastic or reinforcement based algorithms, which are computationally expensive.

### 2.1 Parametrization and Straight Through Gradient Estimates

A symmetric uniform quantizer \( Q_U(x; \theta) \) which maps a real value \( x \in \mathbb{R} \) to one of \( I = 2k + 1 \) quantized values \( q \in \{ -kd, \ldots, 0, \ldots, kd \} \) computes

\[
q = Q_U(x; \theta) = \text{sign}(x) \left\lfloor \frac{|x|}{d} + \frac{1}{2} \right\rfloor \quad |x| \leq q_{\text{max}},
\]

\[
q = Q_U(x; \theta) = \text{sign}(x) q_{\text{max}} \quad |x| > q_{\text{max}},
\]

(1)
We refer to this as with varying bitwidth \(b\) because it will lead to large changes of the gradient norm and forces us to use small learning rates to \(\|\nabla_{\theta} Q_U(x; \theta)\|\). For “U1” and “U2” the maximum gradient norm can grow exponentially with varying bitwidth \(b\) whereas it is bounded for “U3”.

When training quantized DNNs, we want to optimize \(Q_U(x; \theta)\) with respect to the input \(x\) and the quantization parameters \(\theta\), meaning that we need the gradients \(\nabla_x Q_U(x; \theta)\) and \(\nabla_\theta Q(x; \theta)\). A common problem is, that the exact gradients are not useful for training. For example, \(\partial_x Q_U(x; \theta) = d \sum_{k=-b^{-1}}^{b^{-1}} \delta(x - d (k + \frac{1}{2}))\) vanishes almost everywhere. A solution is to define the derivative using STE [Bengio et al., 2013], which ignores the floor operation in \(1\). This leads to

\[
\partial_x Q_U(x) = \begin{cases} 
1 & |x| \leq q_{\text{max}} \\
0 & |x| > q_{\text{max}} 
\end{cases},
\]

which is non-zero in the interesting region \(|x| \leq q_{\text{max}}\) and which turned out to be very useful to train quantized DNNs in practice [Yin et al., 2019]. In this work, we follow this idea and define the gradients \(\nabla_x Q(x; \theta)\) and \(\nabla_\theta Q(x; \theta)\), using STE whenever we need to differentiate a floor operation.

We refer to this as differentiable quantization (DQ).

An important observation from \(1\) is that the parameters \(\theta = [d, q_{\text{max}}, b]^T\) of a quantizer depend on each other, i.e., \(q_{\text{max}} = (2^b - 1)d\). This means, that we can choose from three equivalent parametrizations of \(Q_U(x; \theta)\): Case “U1” with \(\theta = [b, d]^T\), case “U2” with \(\theta = [b, q_{\text{max}}]^T\) and case “U3” with \(\theta = [d, q_{\text{max}}]^T\). Interestingly, they differ in their gradients:

**Case U1:** Parametrization with respect to \(\theta = [b, d]^T\), using \(q_{\text{max}} = q_{\text{max}}(b, d)\) gives

\[
\nabla_\theta Q_U(x; \theta) = \begin{bmatrix} \partial_b Q_U(x; \theta) \\ \partial_d Q_U(x; \theta) \end{bmatrix} = \begin{cases} 
\frac{1}{2} Q_U(x; \theta) - x & |x| \leq (2^b - 1)d \\
\frac{2^{b-1} \log(2)}{2^b - 1} \text{sign}(x) & |x| > (2^b - 1)d 
\end{cases}
\]

**Case U2:** Parametrization with respect to \(\theta = [b, q_{\text{max}}]^T\), using \(d = d(b, q_{\text{max}})\) gives

\[
\nabla_\theta Q_U(x; \theta) = \begin{bmatrix} \partial_b Q_U(x; \theta) \\ \partial_{q_{\text{max}}} Q_U(x; \theta) \end{bmatrix} = \begin{cases} 
\frac{2^{b-1} \log 2}{2^b - 1 - q_{\text{max}}} Q_U(x; \theta) - x & |x| \leq q_{\text{max}} \\
0 & |x| > q_{\text{max}} 
\end{cases}
\]

**Case U3:** Parametrization with respect to \(\theta = [d, q_{\text{max}}]^T\), using \(b = b(d, q_{\text{max}})\) gives

\[
\nabla_\theta Q_U(x; \theta) = \begin{bmatrix} \partial_d Q_U(x; \theta) \\ \partial_{q_{\text{max}}} Q_U(x; \theta) \end{bmatrix} = \begin{cases} 
\frac{1}{2} Q_U(x; \theta) - x & |x| \leq q_{\text{max}} \\
0 & |x| > q_{\text{max}} 
\end{cases}
\]

Fig. 1 shows the maximum gradient norm \(\max_x \|\nabla_\theta Q_U(x; \theta)\|\) for the three parametrizations “U1” to “U3”. For the parametrizations “U1” and “U2”, \(\max_x \|\nabla_\theta Q_U(x; \theta)\|\) can grow exponentially with varying bitwidth \(b\) as \(\partial_x Q_U(x; \theta) \in [-2^{b-1} - 1, 2^{b-1} - 1]\) for “U1” and \(\partial_b Q_U(x; \theta) \in \left[-\frac{q_{\text{max}}}{2^{b-1}} \log 2, \frac{q_{\text{max}}}{2^{b-1}} \log 2\right]\) for “U2”. This is not desirable when training quantized DNNs, because it will lead to large changes of the gradient norm and forces us to use small learning rates to
We fulfill these constraints by rounding the parameters in the forward pass to the closest integer or which has again a bounded gradient magnitude and independent components and is, hence, best suited for first order gradient based optimization. In the following we compare the parametrizations using two experiments.

Similar considerations can be made for the power-of-two quantization $Q_{P}(x; \theta)$, which maps a real-valued number $x \in \mathbb{R}$ to a quantized value $q \in \{-2^{b}, \ldots, 2^{b}\}$ by

$$ q = Q_{P}(x; \theta) = \text{sign}(x) \begin{cases} q_{\text{min}} & |x| \leq q_{\text{min}} \\ 2^{b \cdot \text{sign}(x) \lceil \log_{2} |x| \rceil} & q_{\text{min}} < |x| \leq q_{\text{max}} \\ q_{\text{max}} & |x| > q_{\text{max}} \end{cases} $$

where $q_{\text{min}}$ and $q_{\text{max}}$ are the minimum and maximum absolute values of the quantizer for a bitwidth $b$ bit. Power-of-two quantization is an especially interesting scheme for DNN quantization, since a multiplication of quantized values can be implemented as an addition of the exponents. Using the STE for the floor operation, the derivative $\partial_{x}Q_{P}(x; \theta)$ is given by

$$ \partial_{x}Q_{P}(x) = \begin{cases} 0 & |x| \leq q_{\text{min}} \\ \frac{2^{b \cdot \text{sign}(x) \lceil \log_{2} |x| \rceil}}{|x|} & q_{\text{min}} < |x| \leq q_{\text{max}} \\ 0 & |x| > q_{\text{max}} \end{cases} $$

The power-of-two quantization has the three parameters $[b, q_{\text{min}}, q_{\text{max}}] =: \theta$, which depend on each other with the relationship $q_{\text{max}} = 2^{b-1} \cdot q_{\text{min}}$. Therefore, we have again three different parametrizations with $\theta = [b, q_{\text{min}}]$, $\theta = [b, q_{\text{max}}]$ or $\theta = [q_{\text{min}}, q_{\text{max}}]$, respectively. Similarly to the uniform case, one parametrization $\theta = [q_{\text{min}}, q_{\text{max}}]$ leads to a gradient of a very simple form

$$ \nabla_{\theta}Q_{P}(x; \theta) = \begin{bmatrix} \partial_{q_{\text{min}}}Q_{U}(x; \theta) \\ \partial_{q_{\text{max}}}Q_{U}(x; \theta) \end{bmatrix} = \begin{bmatrix} [1, 0]^{T} & |x| \leq q_{\text{min}} \\ [0, 0]^{T} & q_{\text{min}} < |x| \leq q_{\text{max}} \\ [0, 1]^{T} & |x| > q_{\text{max}} \end{bmatrix} $$

which has again a bounded gradient magnitude and independent components and is, hence, best suited for first order gradient based optimization.

### 2.2 Constraints on $\theta$

In practice, for an efficient hardware implementation, we need to ensure that the quantization parameters only take specific discrete values: for uniform quantization, only integer values are allowed for the bitwidth $b$, and the stepsize $d$ must be a power-of-two, see e.g. (Jain et al., 2019); for power-of-two quantization, the bitwidth must be an integer, and the minimum and maximum absolute values $q_{\text{min}}$ and $q_{\text{max}}$ must be powers-of-two.

We fulfill these constraints by rounding the parameters in the forward pass to the closest integer or power-of-two value. In the backward pass we update the original float values, i.e., we used again the STE to propagate the gradients.

### 2.3 Experimental Comparison of DQ Parametrizations

In the following we compare the parametrizations using two experiments.
In our first experiment we use DQ to learn the optimal quantization parameters $\theta^*$ which minimize the mean squared error (MSE) $E[\frac{1}{2}(Q(x; \theta) - x)^2]$ with gradient descent and compare the convergence speed for three possible parametrizations of a uniform and power-of-two quantizer. We choose this expression as the gradient $\nabla_{\theta} Q(x; \theta) = E[(Q(x; \theta) - x)\nabla_{\theta} Q(x; \theta)]$ is just a scaled version of $\nabla_{\theta} Q(x; \theta)$, i.e., the gradient direction depends directly on the parametrization of $Q(x; \theta)$ and thus the effects of changing the parametrization can be observed.

It is interesting to study the Hessian $H = \nabla_{\theta} \nabla_{\theta}^T E[((Q(x; \theta) - x)^2) \in \mathbb{R}^{2\times 2}$ of the MSE:

$$H = E[\nabla_{\theta} Q(x; \theta)\nabla_{\theta} Q(x; \theta)^T + (Q(x; \theta) - x)\nabla_{\theta} \nabla_{\theta}^T Q(x; \theta)] \approx E[\nabla_{\theta} Q(x; \theta)\nabla_{\theta} Q(x; \theta)^T].$$

Note that we use the outer-product approximation (Bishop, 2006) in order to simplify our considerations. From this equation it is apparent that the Hessian will be diagonal for the case U3 as $\nabla_{\theta} Q(x; \theta)\nabla_{\theta} Q(x; \theta)^T$ only contains an element in either (1, 1) or (2, 2) and, therefore, $E[\nabla_{\theta} Q(x; \theta)\nabla_{\theta} Q(x; \theta)^T]$ is a diagonal matrix. From this, we can see that gradient descent with an individual learning rate for each parameter is equivalent to Newton’s method and, therefore, efficient. In general this will not be the case for U1 and U2.

We conduct an experiment, using ADAM to optimize the mean squared quantization error on artificially generated data, which is generated by drawing $10^4$ samples from $N(0, 1)$. Please note that the same example was studied in Jain et al., 2019. The results in Fig. 3 clearly show that the parametrizations “U3” and “P3” are best suited to optimize the uniform and power-of-two quantization parameters, respectively. Indeed, these quantizers converge without oscillation to the lowest MSE. It is interesting to see, that even adaptive gradient methods like ADAM can not solve the scaling issue described in Sec. 2.1. In the Appendix A.4, we give further empirical evidence to support this claim and compare the different parametrizations for the training of a quantized ResNet-20 on CIFAR-10 using ADAM. Note that all cases use the same learning rate. For the interested reader, a more detailed visualization of the error surfaces over the quantization parameters can be found in Appendix A.3.

In our second experiment we train a ResNet-20 (He et al., 2016) with quantized parameters and activations on CIFAR-10 (Krizhevsky & Hinton, 2009) using the same settings as proposed by He et al., 2016. Fig. 4 shows the evolution of the training and validation error during training for the case of uniform quantization. The plots for power-of-two quantization can be found in the appendix (Fig. 10). We initialize this network from random parameters or from a pre-trained float network. The quantized DNNs are trained for 160 epochs, using SGD with momentum 0.9 and a learning rate schedule starting with 0.01 and reducing it by a factor of 10 after 80 and 120 epochs, respectively. We use random flips and crops for data augmentation. Each epoch takes about 2.5 min on a single GTX 1080 Ti.

In case of randomly initialized weights, we use an initial stepsize $d_t = 2^{-3}$ for the quantization of weights and activations. Otherwise, we initialize the weights using a pre-trained floating point network and the initial stepsize for a layer is chosen to be $d_t = 2^{\lfloor \log_2(\max |W_l|/(2^{b-1}-1)) \rfloor}$. The remaining quantization parameters are chosen such that we start from an initial bitwidth of $b = 4$ bit. This is a reasonable upper limit for $b$, as in practice no performance degradation can be observed for $b > 4$ bit. Even simple offline algorithms like min/max quantization result in networks with good accuracies. We define no memory constraints during training, i.e., the network can learn to use a large number of bits to quantize weights and activations of each layer. From Fig. 4, we again observe that the parametrization $\theta = [d, q_{\text{max}}]^T$ is best suited to train a uniformly quantized DNN as it converges...
We consider the following memory characteristics of the DNN, constraining them during training:

Table 1: Comparison of different DQ parametrizations for ResNet-20 on CIFAR-10.

<table>
<thead>
<tr>
<th>Quantization</th>
<th>Float32</th>
<th>Uniform quantization</th>
<th>Power-of-two quantization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>θ = [b, q]</td>
<td>θ = [d, qmax]^T</td>
<td>θ = [b, qmax]^T</td>
</tr>
<tr>
<td>Weights</td>
<td>8.50%/7.29%</td>
<td>17.8%/8.18%</td>
<td>8.50%/7.32%</td>
</tr>
<tr>
<td>Weights+Activations</td>
<td>28.9%/9.03%</td>
<td>8.80%/7.44%</td>
<td>11.70%/7.90%</td>
</tr>
</tbody>
</table>

We now discuss how to train quantized DNNs with memory constraints. Such constraints appear in conjunction with parameters $\theta$.

A quantized DNN consists of layers which compute the vector product. In case of a convolutional layer, $\theta_l = \theta_l \cdot \theta_{l+1}$ is computed separately to the input and output tensors $\mathbf{x}_l \in \mathcal{I}_l$ and $\mathbf{x}_{l+1} \in \mathcal{I}_{l+1}$, and also to both the weight tensors $\mathbf{W}_l \in \mathcal{P}_l$ and the bias vector $\mathbf{c}_l \in \mathcal{R}^{M_l}$. For a fully connected layer, $\mathcal{I}_{l-1} = \mathbb{R}^{M_{l-1} \times M_l}$, $\mathcal{I}_l = \mathbb{R}^{M_l}$ are vectors, $\mathcal{P}_l = \mathbb{R}^{M_l \times M_{l-1}}$ are matrices, and $\mathcal{R}^{M_l \times M_{l-1} \times N_{l-1} \times N_l}$ are tensors and $A*B$ is a set of $M_{l-1} M_l 2D$ convolutions, where the convolution is performed on square-sized feature maps of size $N_{l-1} \times N_l$ using square-sized kernels of size $K_l \times K_l$.

DNNs with quantized weights and activations have a smaller memory footprint and are also computationally cheaper to evaluate since $Q(\alpha; \theta) \cdot Q(\beta; \theta)$ for $\alpha, \beta \in \mathbb{R}$ requires only an integer multiplication for the case of uniform quantization or an integer addition of the exponents for power-of-two quantization. Furthermore, $Q(\alpha; \theta) + Q(\beta; \theta)$ for $\alpha, \beta \in \mathbb{R}$ only requires an integer addition.

Table 2 compares the computational complexity and the memory footprint of layers which apply uniform or power-of-two quantization to weights and activations.

3 Training quantized DNNs with memory constraints

We now discuss how to train quantized DNNs with memory constraints. Such constraints appear in many applications when the network inference is performed on an embedded device with limited computational power and memory resources.

A quantized DNN consists of layers which compute

$$\mathbf{x}_l = f_l(Q(\mathbf{W}_l; \theta_l); \mathbf{x}_{l-1}; \theta_{l-1}) + Q(\mathbf{c}_l; \theta_l^w) \text{ with } l = 1, ..., L,$$

where $f_l(\cdot)$ denotes the nonlinear activation function of layer $l$ and $Q(\cdot; \theta)$ is a per-tensor quantization with parameters $\theta$ applied separately to the input and output tensors $\mathbf{x}_l \in \mathcal{I}_l$ and $\mathbf{x}_{l+1} \in \mathcal{I}_{l+1}$, and also to both the weight tensors $\mathbf{W}_l \in \mathcal{P}_l$ and the bias vector $\mathbf{c}_l \in \mathcal{R}^{M_l}$. For a fully connected layer, $\mathcal{I}_{l-1} = \mathbb{R}^{M_{l-1} \times M_l}$, $\mathcal{I}_l = \mathbb{R}^{M_l}$ are vectors, $\mathcal{P}_l = \mathbb{R}^{M_l \times M_{l-1}}$ are matrices, and $A*B$ is a set of $M_{l-1} M_l$ 2D convolutions, where the convolution is performed on square-sized feature maps of size $N_{l-1} \times N_l$ using square-sized kernels of size $K_l \times K_l$.

We consider the following memory characteristics of the DNN, constraining them during training:

1. Total memory $S^w(\theta_1^w, ..., \theta_L^w) = \sum_{l=1}^{L} S_l^w(\theta_l^w)$ to store all weights: We use the constraint

$$g_1(\theta_1^w, ..., \theta_L^w) = S^w(\theta_1^w, ..., \theta_L^w) - S_0^w \leq 0,$$

where $S_l^w(\theta_l^w)$ is the total memory to store weights.

Figure 4: ResNet-20 with uniformly quantized weights and activations.

Table 1 compares the best validation error for all parametrizations of the uniform and power-of-two quantizations. We trained networks either with quantized weights and full precision activations or with both being quantized. In case of activation quantization with power-of-two, we use one bit to explicitly represent the value $x = 0$. This is advantageous as the ReLU nonlinearity will map many activations to this value. We can observe that training the quantized DNN with the optimal parametrization of DQ, i.e., using either $\theta = [d, q_{\text{max}}]^T$ or $\theta = [q_{\text{min}}, q_{\text{max}}]^T$ results in a network with the lowest validation error. This result again supports our theoretical considerations from Sec. 2.1.

3 Training quantized DNNs with memory constraints

We now discuss how to train quantized DNNs with memory constraints. Such constraints appear in many applications when the network inference is performed on an embedded device with limited computational power and memory resources.

A quantized DNN consists of layers which compute

$$\mathbf{x}_l = f_l(Q(\mathbf{W}_l; \theta_l); \mathbf{x}_{l-1}; \theta_{l-1}) + Q(\mathbf{c}_l; \theta_l^w) \text{ with } l = 1, ..., L,$$

where $f_l(\cdot)$ denotes the nonlinear activation function of layer $l$ and $Q(\cdot; \theta)$ is a per-tensor quantization with parameters $\theta$ applied separately to the input and output tensors $\mathbf{x}_l \in \mathcal{I}_l$ and $\mathbf{x}_{l+1} \in \mathcal{I}_{l+1}$, and also to both the weight tensors $\mathbf{W}_l \in \mathcal{P}_l$ and the bias vector $\mathbf{c}_l \in \mathcal{R}^{M_l}$. For a fully connected layer, $\mathcal{I}_{l-1} = \mathbb{R}^{M_{l-1} \times M_l}$, $\mathcal{I}_l = \mathbb{R}^{M_l}$ are vectors, $\mathcal{P}_l = \mathbb{R}^{M_l \times M_{l-1}}$ are matrices, and $A*B$ is a set of $M_{l-1} M_l$ 2D convolutions, where the convolution is performed on square-sized feature maps of size $N_{l-1} \times N_l$ using square-sized kernels of size $K_l \times K_l$.

DNNs with quantized weights and activations have a smaller memory footprint and are also computationally cheaper to evaluate since $Q(\alpha; \theta) \cdot Q(\beta; \theta)$ for $\alpha, \beta \in \mathbb{R}$ requires only an integer multiplication for the case of uniform quantization or an integer addition of the exponents for power-of-two quantization. Furthermore, $Q(\alpha; \theta) + Q(\beta; \theta)$ for $\alpha, \beta \in \mathbb{R}$ only requires an integer addition.

Table 2 compares the computational complexity and the memory footprint of layers which apply uniform or power-of-two quantization to weights and activations.

We consider the following memory characteristics of the DNN, constraining them during training:

1. Total memory $S^w(\theta_1^w, ..., \theta_L^w) = \sum_{l=1}^{L} S_l^w(\theta_l^w)$ to store all weights: We use the constraint

$$g_1(\theta_1^w, ..., \theta_L^w) = S^w(\theta_1^w, ..., \theta_L^w) - S_0^w \leq 0,$$

1In this paper, we use “weights” to refer to $\mathbf{W}$ and $\mathbf{c}$.
Table 2: Number of multiplications $C^m_{\text{mul}}$, additions $C^a_{\text{add}}$ as well as required memory to store the weights $S^w_l$ and activations $S^a_l$ of fully connected and convolutional layers.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Quantization</th>
<th>$C^m_{\text{mul}}$</th>
<th>$C^a_{\text{add}}$</th>
<th>$S^w_l$</th>
<th>$S^a_l$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully connected</td>
<td>uniform</td>
<td>$M_1M_1$</td>
<td>$0$</td>
<td>$M_1(M_1+1)b^w_1$</td>
<td>$M_1b^a_1$</td>
</tr>
<tr>
<td>Convolutional</td>
<td>uniform</td>
<td>$M_1M_1-N^2K^2L^2$</td>
<td>$0$</td>
<td>$M_1(M_1-K^2+1)b^w_1$</td>
<td>$M_1N^2K^2b^a_1$</td>
</tr>
</tbody>
</table>

Note, that the optimization problem (10) does not necessarily give a quantized DNN which fulfills either (8b) or (8c).

1. **Total weight memory** $S^w_l(\theta^w_1, ..., \theta^w_L)$ is smaller than a certain maximum weight memory size $S^w_0$. Table 2 gives $S^w_l(\theta^w_1)$ for the case of fully connected and convolutional layers. Each layer’s memory requirement $S^w_l(\theta^w_1)$ depends on the bitwidth $b^w_l$: reducing $S^w_l(\theta^w_1)$ will reduce the bitwidth $b^w_l$.

2. **Total activation memory** $S^a_l(\theta^a_1, ..., \theta^a_L)$ = $\sum_{i=1}^L S^a_i(\theta^a_i)$ to store all feature maps: We use the constraint

$$g_0(\theta^a_1, ..., \theta^a_L) = S^a(\theta^a_1, ..., \theta^a_L) - S^a_0 = \sum_{i=1}^L S^a_i(\theta^a_i) - S^a_0 \leq 0,$$  \hspace{1cm} \text{(8b)}

3. **Maximum activation memory** $\hat{S}^a_l(\theta^a_1, ..., \theta^a_L) = \max_{i=1,...,L} S^a_i$ to store the largest feature map: We use the constraint

$$g_3(\theta^a_1, ..., \theta^a_L) = \hat{S}^a(\theta^a_1, ..., \theta^a_L) - \hat{S}^a_0 = \max_{i=1,...,L} \{ S^a_i \} - \hat{S}^a_0 \leq 0,$$  \hspace{1cm} \text{(8c)}

To train the quantized DNN with memory constraints, we need to solve the optimization problem

$$\min_{\mathbf{W}, \mathbf{c}, \theta^0, \theta^f} \mathbb{E}_{p(\mathbf{x}, \mathbf{y})} \left[ J(\mathbf{X}_L, \mathbf{Y}) \right] \text{ s.t. } g_j(\theta^w_1, ..., \theta^w_L, \theta^a_1, ..., \theta^a_L) \leq 0 \text{ for all } j = 1, ..., 3$$  \hspace{1cm} \text{(9)}

where $J(\mathbf{X}_L, \mathbf{Y})$ is the loss function for yielding the DNN output $\mathbf{X}_L$ although the ground truth is $\mathbf{Y}$. Eq. (9) learns the weights $\mathbf{W}_j, \mathbf{c}_j$ as well as the quantization parameters $\theta^w_1, \theta^a_1$. In order to use simple stochastic gradient descent solvers, we use the penalty method (Bertsekas, 2014) to convert (9) into the unconstrained optimization problem

$$\min_{\mathbf{W}, \mathbf{c}, \theta^0, \theta^f} \mathbb{E}_{p(\mathbf{x}, \mathbf{y})} \left[ J(\mathbf{X}_L, \mathbf{Y}) \right] + \sum_{j=1}^J \lambda_j \max(0, g_j(\theta^w_1, ..., \theta^w_L, \theta^a_1, ..., \theta^a_L))^2,$$  \hspace{1cm} \text{(10)}

where $\lambda_j \in \mathbb{R}^+$ are individual weightings for the penalty terms. Hence, training with weight and activation size constraints requires choosing two penalty weightings $\lambda_j$, one for (8b) and one for either (8b) or (8c).

Note, that the optimization problem (10) does not necessarily give a quantized DNN which fulfills the memory constraints. The probability to fulfill the constraint $g_j$ depends on the choice of $\lambda_j$. In particular, this probability increases with larger $\lambda_j$. However, choosing a too large $\lambda_j$ will yield a penalty term that dominates over the network loss decreasing the network performance. In our experiments, we choose $\lambda_j$ such that the initial loss and the penalty term have approximately the same magnitude. Using this simple heuristic, we obtained quantized DNNs that reached a high accuracy and at the same time fulfilled the constraints at the end of training.

4 **Experiments**

In the following, we will use the best parametrizations for uniform and power-of-two DQ, i.e., $\theta_U = [d, q_{\text{max}}]^T$ and $\theta_P = [q_{\text{min}}, q_{\text{max}}]^T$, that we found in Sec. 2. Both parametrizations do not directly depend on the bitwidth $b$. Therefore, we compute it by using $b(\theta_U) = \lceil \log_2 \left( \frac{q_{\text{max}}}{d} + 1 \right) + 1 \rceil$ and $b(\theta_P) = \lceil \log_2 \left( \log_2 \left( \frac{q_{\text{max}}}{q_{\text{min}}} \right) + 1 \right) + 1 \rceil$. All quantized networks use a pre-trained float32 network for initialization and all quantizers are initialized as described in Sec. 2.3. For our experiments on CIFAR-10, we use the same training setup as described in Sec. 2.3. For the experiments on...
Table 3: Homogeneous vs. heterogeneous quantization of ResNet-20 on CIFAR-10.

<table>
<thead>
<tr>
<th>Bitwidth Weight/Activ.</th>
<th>$q_{\text{max}}$ Weight/Activ.</th>
<th>Size Weight/Activ.(max)/Activ.(sum)</th>
<th>Uniform quant. Validation error</th>
<th>Power-of-two quant. Validation error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>32bit/32bit</td>
<td>–</td>
<td>1048KB/64KB/736KB</td>
<td>7.29%</td>
</tr>
<tr>
<td>Fixed</td>
<td>2bit/32bit</td>
<td>fixed/–</td>
<td>65.5KB/64KB/736KB</td>
<td>10.81%</td>
</tr>
<tr>
<td>TQT</td>
<td>2bit/32bit</td>
<td>learned/–</td>
<td>65.5KB/64KB/736KB</td>
<td>9.47%</td>
</tr>
<tr>
<td>Ours (w/ constr.)</td>
<td>learned/32bit</td>
<td>learned/–</td>
<td>70KB/64KB/736KB</td>
<td>8.59%</td>
</tr>
<tr>
<td>Fixed</td>
<td>2bit/4bit</td>
<td>fixed/fixed</td>
<td>65.5KB/8KB/92KB</td>
<td>11.30%</td>
</tr>
<tr>
<td>TQT</td>
<td>2bit/4bit</td>
<td>learned/learned</td>
<td>65.5KB/8KB/92KB</td>
<td>9.62%</td>
</tr>
<tr>
<td>Ours (w/ constr.)</td>
<td>learned/learned/learned</td>
<td>70KB/–/92KB</td>
<td>9.38%</td>
<td></td>
</tr>
<tr>
<td>Ours (w/ constr.)</td>
<td>learned/learned/learned</td>
<td>70KB/8KB/ –</td>
<td>8.58%</td>
<td></td>
</tr>
</tbody>
</table>

ImageNet, we train the quantized DNNs for 50 epochs, using SGD with momentum 0.9 and a learning rate schedule starting with 0.01 and reducing it by a factor of 10 after 16 and 32 epochs, respectively. Please note that we quantize all layers opposed to other papers which use a higher precision for the first and/or last layer.

In our experiments, we noticed that the performance of DQ is not sensitive to the choice of $\lambda_j$ in (10). For the CIFAR-10 experiments, we use $\lambda = 0.1$ for both constraints (for sizes in kB). For the ImageNet experiments, we kept the same regularization level by scaling $\lambda_j$ with the square of the size ratio between the ImageNet model and the CIFAR-10 model. We scale with the square-ratio as the constraints in (10) are squared penalty terms.

First, in Table 3/top, we train a ResNet-20 on CIFAR-10 with quantized weights and float32 activations. We start with the most restrictive quantization scheme with fixed $q_{\text{max}}$ and $b = 2$ bit (“Fixed”). Then, we allow the model to learn $q_{\text{max}}$ while $b = 2$ bit remains fixed as was done in (Jain et al., 2019) (“TQT”). Finally, we learn both $q_{\text{max}}$ and $b$ with the constraint that the weight size is at most 70KB (“Ours”), which is just 4.5KB larger than the previous 2Bit networks. This allows the model to allocate more than two bits to some layers. From Table 3/top, we observe that the error is smallest when we learn all quantization parameters.

In Table 3/bottom, weights and activations are quantized. For activation quantization, we consider two cases as discussed in Sec. 3. The first one constrains the total activation memory $S^x$ while the second constrains the maximum activation memory $\hat{S}^x$ such that both have the same size as a homogeneously quantized model with 4bit activations. Again, we observe that the error is smallest when we learn all quantization parameters.

We also use DQ to train quantized ResNet-18 (He et al., 2016) and MobileNetV2 (Sandler et al., 2018) on ImageNet (Deng et al., 2009) with 4bit uniform weights and activations or equivalent-sized networks with learned quantization parameters. This is quite aggressive and, thus, a fixed quantization scheme loses more than 6% accuracy while our quantization scheme loses less than 0.5% compared to a float32 precision network.

Our results compare favorably to other recent quantization approaches. To our knowledge, the best result for a 4bit ResNet-18 was reported by (Esser et al., 2019) (29.91% error). This is very close to our performance (29.92% error). Importantly, (Esser et al., 2019) did not quantize the first and last layers, meaning that their network is much bigger. Specifically, compared to our quantized ResNet-18, their model with high precision input and output layers requires 37% more memory to store the weights. Moreover, (Esser et al., 2019) learns stepsizes which are not restricted to powers-of-two. As explained in Sec. 2.2, uniform quantization with power-of-two stepsizes leads to more efficient inference, effectively allowing to efficiently compute any multiplication with an integer multiplication and bit-shift. To our knowledge only (Wang et al., 2018) reported results of MobileNetV2 quantized to 4bit. They keep the baseline performance constraining the network to the same size as the 4bit network. However, they do not quantize the activations in this case. In addition, DQ training is efficient since it is comparable to the training of unquantized network. Specifically, one epoch on ImageNet takes 37min for MobileNetV2 and 18min for ResNet-18 on four Nvidia Tesla V100.

Fig. 5 shows the weight bitwidth assignment over layers. We observe that small bitwidths are used for layers with many parameters, i.e., pointwise convolutions and fully connected layers. However, the resulting bitwidth assignments are complex, meaning that there is no simple heuristic. Therefore, it is important to learn the optimal bitwidth assignment.
Table 4: Homogeneous vs. heterogeneous quantization of MobileNetV2 and ResNet-18 on ImageNet.

<table>
<thead>
<tr>
<th></th>
<th>Bitwidth</th>
<th>MobileNetV2</th>
<th>ResNet-18</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Weight/Activ.</td>
<td>q_{max}</td>
<td>Size</td>
</tr>
<tr>
<td></td>
<td>Weight/Activ(max)</td>
<td>Error</td>
<td>Weight/Activ(max)</td>
</tr>
<tr>
<td>Baseline</td>
<td>32bit/32bit</td>
<td>–</td>
<td>13.23MB/4.59MB</td>
</tr>
<tr>
<td>Fixed</td>
<td>4bit/4bit</td>
<td>fixed/fixed</td>
<td>1.65MB/0.57MB</td>
</tr>
<tr>
<td>TQT (Jain et al., 2019)</td>
<td>4bit/4bit</td>
<td>learned/learned</td>
<td>1.65MB/0.57MB</td>
</tr>
<tr>
<td>Ours (w/ constr. (8a) and (8c))</td>
<td>learned/learned</td>
<td>learned/learned</td>
<td>1.55MB/0.57MB</td>
</tr>
<tr>
<td>Ours (w/o constr.)</td>
<td>learned/learned</td>
<td>learned/learned</td>
<td>3.14MB/1.58MB</td>
</tr>
</tbody>
</table>

Figure 5: Weight bitwidth assignment over layers for ResNet-18 and MobileNetV2 on ImageNet with weights constrained to a maximum size of 5.57MB. Our method has learned a heterogeneous bitwidth distribution, which gives a better performance than a homogeneous one (see Table 4).

5 Conclusions

In this paper we discussed differentiable quantization and its application to the training of compact DNNs with memory constraints. In order to fulfill memory constraints, we introduced penalty functions during training and used stochastic gradient descent to find the optimal weights as well as the optimal quantization values in a joint fashion. We showed that there are several possible parametrizations of the quantization function. In particular, learning the bitwidth directly is not optimal; therefore, we proposed to parametrize the quantizer with the stepsize and dynamic range instead. The bitwidth can then be inferred from them. This approach is competitive to other recent quantization methods while it does not require to retrain the network multiple times in contrast to reinforcement learning approaches (Wang et al., 2018; Elthakeb et al., 2018).

Acknowledgements

We would like to thank Masato Ishii for many helpful comments during the preparation of this manuscript.

References


A Derivation of the Gradients for Differentiable Quantization (DQ)

In the following sections, we will give the derivatives $\frac{\partial}{\partial x} Q(x; \theta)$ and gradients $\nabla_{\theta} Q(x; \theta)$ for the uniform and the power-of-two quantizers. The results are summarized in Sec. 2.

We use the straight-through gradient estimate whenever we need to differentiate a non-differentiable floor function, i.e., we assume

$$\frac{\partial}{\partial x} \lfloor x \rfloor = 1. \quad (11)$$

A.1 Derivatives of the uniform quantizer

Fig. 6(a) shows a symmetric uniform quantizer $Q_U(x; \theta)$ which maps a real value $x \in \mathbb{R}$ to one of $I = 2k + 1$ quantized values $q \in \{-kd, ..., 0, ..., kd\}$ by computing

$$q = Q_U(x; \theta) = \text{sign}(x) \begin{cases} d \left\lfloor \frac{|x|}{d} + \frac{1}{2} \right\rfloor & |x| \leq q_{\max} \\ q_{\max} & |x| > q_{\max} \end{cases} \quad (12)$$

using the parameters $\theta = [d, q_{\max}, b]^T$ where $d \in \mathbb{R}$ is the stepsize, $q_{\max} \in \mathbb{R}$ is the maximum value and $b \in \mathbb{N}$ is the number of bits that we use to encode the quantized values $q$. The elements of $\theta$ are dependent as there is the relationship $q_{\max} = (2^b - 1) d$.

A.1.1 Case U1: Parametrization with respect to $b$ and $d$

For the parametrization with respect to the bitwidth $b$ and steps size $d$, (12) is given by

$$q = Q_U(x; b, d) = \text{sign}(x) \begin{cases} d \left\lfloor \frac{|x|}{d} + \frac{1}{2} \right\rfloor & |x| \leq (2^b - 1) d \\ 2^b - 1 & |x| > (2^b - 1) d \end{cases} \quad (13)$$

and the derivatives are given by

$$\frac{\partial Q_U(x; b, d)}{\partial b} = \text{sign}(x) \frac{2^b - 1}{2^{b-1} - 1} \begin{cases} 0 & |x| \leq (2^b - 1) d \\ (2^b - 1) d & |x| > (2^b - 1) d \end{cases} \quad (14a)$$

$$\frac{\partial Q_U(x; b, d)}{\partial d} = \text{sign}(x) \frac{1}{d} \begin{cases} d \left\lfloor \frac{|x|}{d} + \frac{1}{2} \right\rfloor - |x| & |x| \leq (2^b - 1) d \\ (2^b - 1) d & |x| > (2^b - 1) d \end{cases} \quad (14b)$$

A.1.2 Case U2: Parametrization with respect to $b$ and $q_{\max}$

For the parametrization with respect to the bitwidth $b$ and maximum value $q_{\max}$, (12) is given by

$$q = Q_U(x; b, q_{\max}) = \text{sign}(x)q_{\max} \begin{cases} \frac{1}{2^{b-1}-1} \left\lfloor |x| \frac{2^b - 1}{q_{\max}} + \frac{1}{2} \right\rfloor & |x| \leq q_{\max} \\ 1 & |x| > q_{\max} \end{cases} \quad (15)$$

and the derivatives are given by

$$\frac{\partial Q_U(x; b, q_{\max})}{\partial b} = \text{sign}(x) \frac{2^b - 1}{2^{b-1} - 1} \begin{cases} -\frac{q_{\max}}{2^{2b-1} - 1} \left\lfloor |x| \frac{2^b - 1}{q_{\max}} + \frac{1}{2} \right\rfloor + \beta_1 & |x| \leq q_{\max} \\ \beta_1 & |x| > q_{\max} \end{cases} \quad (16a)$$

$$\frac{\partial Q_U(x; b, q_{\max})}{\partial q_{\max}} = \text{sign}(x) \frac{1}{q_{\max}} \begin{cases} \frac{q_{\max}}{2^{2b-1} - 1} \left\lfloor |x| \frac{2^b - 1}{q_{\max}} + \frac{1}{2} \right\rfloor + \beta_2 & |x| \leq q_{\max} \\ \beta_2 & |x| > q_{\max} \end{cases} \quad (16b)$$

where $\beta_1 = \frac{q_{\max}}{2^{2b-1} - 1} \frac{\partial}{\partial b} \left\lfloor |x| \frac{2^b - 1}{q_{\max}} + \frac{1}{2} \right\rfloor = |x|$ and $\beta_2 = q_{\max} \frac{\partial}{\partial q_{\max}} \left\lfloor |x| \frac{2^b - 1}{q_{\max}} + \frac{1}{2} \right\rfloor = -|x|$, if we use the straight-through gradient estimate for the floor function.
The power-of-two quantization has the three parameters $\theta$. For the parametrization with first order gradient based optimization, one parametrization ($\theta = [\theta_q, \theta_k]$, $\theta_q = [\theta_k, \theta_q]$) leads to a gradient with the nice form 

\[
\nabla_\theta Q_P(x; \theta) = \begin{bmatrix} \partial_{q_{\min}} Q_P(x; \theta) \\ \partial_{q_{\max}} Q_P(x; \theta) \end{bmatrix} = \begin{bmatrix} 1, 0 \end{bmatrix}^{T} \begin{bmatrix} |x| \leq q_{\min} \\ q_{\min} < |x| \leq q_{\max} \\ 0 \end{bmatrix}^{T} \begin{bmatrix} q_{\max} \\ |x| > q_{\max} \end{bmatrix}
\]

which has a bounded gradient magnitude and independent components and is, hence, well suited for first order gradient based optimization.

A.2.1 CASE P1: Parametrization with respect to $b$ and $q_{\max}$

For the parametrization with $\theta = [b, q_{\max}]$, (18) is given by

\[
Q_P(x; b, q_{\max}) = \text{sign}(x) \begin{cases} 2^{-b^{-1}+1}q_{\max} & |x| \leq 2^{-b^{-1}+1}q_{\max} \\ 2^{0.5+\log_2|x|} & 2^{-b^{-1}+1}q_{\max} < |x| \leq q_{\max} \\ q_{\max} & |x| > q_{\max} \end{cases}
\]
and the partial derivatives are
\[
\frac{\partial Q_p(x; b, q_{\text{max}})}{\partial b} = \text{sign}(x) \begin{cases} 
-2^{-2^{b-1}+b(\log 2)^2} q_{\text{max}} & |x| \leq -2^{-2^{b-1}+1} q_{\text{max}} \\
0 & -2^{-2^{b-1}+1} q_{\text{max}} < |x| \leq q_{\text{max}} \\
0 & |x| > q_{\text{max}} 
\end{cases},
\]
(22a)

\[
\frac{\partial Q_p(x; b, q_{\text{max}})}{\partial q_{\text{max}}} = \text{sign}(x) \begin{cases} 
2^{-2^{b-1}+1} & |x| \leq -2^{-2^{b-1}+1} q_{\text{max}} \\
0 & -2^{-2^{b-1}+1} q_{\text{max}} < |x| \leq q_{\text{max}} \\
1 & |x| > q_{\text{max}} 
\end{cases}. 
\]
(22b)

A.2.2 Case P2: Parametrization with respect to \( b \) and \( q_{\text{min}} \)

For the parametrization with \( \theta = [b, q_{\text{min}}] \), (18) is given by
\[
Q_p(x; b, q_{\text{min}}) = \text{sign}(x) \begin{cases} 
q_{\text{min}} & |x| \leq q_{\text{min}} \\
2^{0.5+\log_2|x|} q_{\text{min}} & 2^{2^{b-1}-1} q_{\text{min}} < |x| \leq 2^{2^{b-1}-1} q_{\text{min}} \\
2^{2^{b-1}-1} q_{\text{min}} & |x| > 2^{2^{b-1}-1} q_{\text{min}} 
\end{cases},
\]
(23)

and the partial derivatives are
\[
\frac{\partial Q_p(x; b, q_{\text{min}})}{\partial b} = \text{sign}(x) \begin{cases} 
0 & |x| \leq q_{\text{min}} \\
0 & 2^{2^{b-1}+b(\log 2)^2} q_{\text{min}} < |x| \leq 2^{2^{b-1}-1} q_{\text{min}} \\
2^{2^{b-1}+b(\log 2)^2} q_{\text{min}} & |x| > 2^{2^{b-1}-1} q_{\text{min}} 
\end{cases},
\]
(24a)

\[
\frac{\partial Q_p(x; b, q_{\text{min}})}{\partial q_{\text{min}}} = \text{sign}(x) \begin{cases} 
1 & |x| \leq q_{\text{min}} \\
0 & q_{\text{min}} < |x| \leq 2^{2^{b-1}-1} q_{\text{min}} \\
2^{2^{b-1}-1} & |x| > 2^{2^{b-1}-1} q_{\text{min}} 
\end{cases}. 
\]
(24b)

A.2.3 Case P3: Parametrization with respect to \( q_{\text{min}} \) and \( q_{\text{max}} \)

Eq. (18) gives the parametrization of \( Q(x; \theta) \) with respect to the minimum value \( q_{\text{min}} \) and maximum value \( q_{\text{max}} \). The derivatives are
\[
\frac{\partial Q_p(x; q_{\text{min}}, q_{\text{max}})}{\partial q_{\text{min}}} = \text{sign}(x) \begin{cases} 
1 & |x| \leq q_{\text{min}} \\
0 & q_{\text{min}} < |x| \leq q_{\text{max}} \\
0 & |x| > q_{\text{max}} 
\end{cases},
\]
(25a)

\[
\frac{\partial Q_p(x; q_{\text{min}}, q_{\text{max}})}{\partial q_{\text{max}}} = \text{sign}(x) \begin{cases} 
0 & |x| \leq q_{\text{min}} \\
0 & q_{\text{min}} < |x| \leq q_{\text{max}} \\
1 & |x| > q_{\text{max}} 
\end{cases}. 
\]
(25b)
In Sec. 2.3 of the paper, we compared the three different parametrizations of the uniform quantizer at the example of optimal quantization of Gaussian data. To get a better understanding of Fig 3, we show how the error surfaces look like for this example problem. The experimental setup is the same as in Sec. 2.3, i.e., we use DQ to learn the optimal quantization parameters of a uniform and a power-of-two quantizer, which minimize the expected quantization error \( \min_\theta \mathbb{E}[(x - Q(x; \theta))^2] \).

We use three different parametrizations, adapt the quantizer’s parameters with gradient descent and compare the convergence speed as well as the final quantization error. As an input, we generate \( 10^4 \) samples from \( N(0,1) \).

Fig. 8 shows the corresponding error surfaces for the three different parametrizations of the uniform quantization. The red curve shows the path through the parameter space taken by gradient descent in order to optimize the MSE, starting with the initial values \( b = 2, d = q_{\text{max}} = 1 \). The optimum \( \theta^* \) is located at \( b = 16, d \leq 2^{-13}, q_{\text{max}} = 4 \), since we allow a maximal bitwidth of 16bit and the largest sample magnitude in our dataset is \( \max\{x_1, \ldots, x_N\} \leq 4 \). In each of the cases U1-U3, the error surface is composed of steep ridges and large flat regions. The steep ridges force us to use small learning rates to avoid divergence. For cases U1 and U2, the optimal \( \theta^* \) can not be reached. However, for U3, \( \theta^* \) lies at the border of a flat region and can be easily reached. Furthermore, case U3 shows a much faster and more stable convergence without oscillation, since the gradient magnitudes are bounded and the error surface has fewer steep ridges where gradient descent starts oscillating.

Fig. 9 shows the corresponding error surfaces for the three different parametrizations of the power-of-two quantization. Again, the optimum \( \theta^* \) is not attained for two parametrizations, namely P1 and P2, as \( \theta^* \) is surrounded by a large, mostly flat region. For these two cases, gradient descent tends to oscillate at steep ridges and tends to be unstable. However, gradient descent converges to a point close to \( \theta^* \) for parametrization P3, where \( \theta = [q_{\text{min}}, q_{\text{max}}] \).

Finally, we also did a comparison of the different power-of-two quantizations on CIFAR-10. Fig. 10 shows the evolution of the training and validation error if we start from a random or a pre-trained float network initialization. We can observe that \( \theta = [q_{\text{min}}, q_{\text{max}}] \) has the best convergence behavior and thus also results in the smallest validation error (cf. Table 1). The unstable behavior of P2 is expected as the derivative \( \frac{\partial Q_P}{\partial q_{\text{min}}} \) can take very large (absolute) values.

### A.4 Further Experiments with ADAM

Finally, we did an experiment to verify that the parametrization is important, even if adaptive gradient descent methods like ADAM are used for optimization. Table 5 gives the results for a ResNet-20 trained on CIFAR-10. We observe, that again U3 and P3 are the best parametrizations.
Figure 10: ResNet-20 with power-of-two quantized weights and activations.

Table 5: Error rate of ResNet-20 on CIFAR-10 using different quantization parametrizations. Training is done either by SGD with momentum or ADAM.

<table>
<thead>
<tr>
<th>Parametrization</th>
<th>SGD momentum</th>
<th>ADAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>U1</td>
<td>11.74</td>
<td>7.61</td>
</tr>
<tr>
<td>U2</td>
<td>7.44</td>
<td>7.85</td>
</tr>
<tr>
<td>U3</td>
<td>7.32</td>
<td>7.36</td>
</tr>
<tr>
<td>P1</td>
<td>15.35</td>
<td>7.54</td>
</tr>
<tr>
<td>P2</td>
<td>7.74</td>
<td>7.79</td>
</tr>
<tr>
<td>P3</td>
<td>7.40</td>
<td>7.40</td>
</tr>
</tbody>
</table>

B IMPLEMENTATION DETAILS FOR DIFFERENTIABLE QUANTIZATION

The following code gives our differentiable quantizer implementation in NNabla (Sony). The source code for reproducing our results will be published after the review process has been finished.
B.1 UNIFORM QUANTIZATION

B.1.1 CASE U1: PARAMETRIZATION WITH RESPECT TO b AND d

```python
def parametric_fixed_point_quantize_d_b(x, sign,
    n_init, n_min, n_max,
    d_init, d_min, d_max,
    fix_parameters=False):
    
    """Parametric version of 'fixed_point_quantize' where the
    bitwidth 'b' and stepsize 'd' are learnable parameters.
    """
    
    Returns:
    ~nnabla.Variable: N-D array.
    
    def clip_scalar(v, min_value, max_value):
        return F.minimum_scalar(F.maximum_scalar(v, min_value), max_value)
    
    def broadcast_scalar(v, shape):
        return F.broadcast(F.reshape(v, (1,) * len(shape), inplace=False), shape=shape)
    
    def quantize_pow2(v):
        return 2 ** F.round(F.log(v) / np.log(2.))
    
    n = get_parameter_or_create("n", (),
        ConstantInitializer(n_init),
        need_grad=True,
        as_need_grad=not fix_parameters)
    
    d = get_parameter_or_create("d", (),
        ConstantInitializer(d_init),
        need_grad=True,
        as_need_grad=not fix_parameters)
    
    # ensure that bitwidth is in specified range and an integer
    n = F.round(clip_scalar(n, n_min, n_max))
    if sign:
        n = n - 1
    
    # ensure that stepsize is in specified range and a power of two
    d = quantize_pow2(clip_scalar(d, d_min, d_max))
    
    # ensure that dynamic range is in specified range
    xmax = d * (2 ** n - 1)
    
    # compute min/max value that we can represent
    if sign:
        xmin = -xmax
    else:
        xmin = nn.Variable((1,), need_grad=False)
    xmin.d = 0.
    
    # broadcast variables to correct size
    d = broadcast_scalar(d, shape=x.shape)
    xmax = broadcast_scalar(xmax, shape=x.shape)
    xmin = broadcast_scalar(xmin, shape=x.shape)
    
    # apply fixed-point quantization
    return d * F.round(F.clip_by_value(x, xmin, xmax) / d)
```

16
B.1.2 CASE U2: PARAMETRIZATION WITH RESPECT TO $b$ AND $q_{\text{MAX}}$

```python

def parametric_fixed_point_quantize_b_xmax(x, sign,
    n_init, n_min, n_max,
    xmax_init, xmax_min, xmax_max,
    fix_parameters=False):
    
    ***Parametric version of 'fixed_point_quantize' where the
    bitwidth 'b' and dynamic range 'xmax' are learnable parameters.
    
    Returns:
    ~nnabla.Variable: N-D array.
    
    def clip_scalar(v, min_value, max_value):
        return F.minimum_scalar(F.maximum_scalar(v, min_value), max_value)
    
    def broadcast_scalar(v, shape):
        return F.broadcast(F.reshape(v, (1,) * len(shape), inplace=False), shape=shape)
    
    def quantize_pow2(v):
        return 2 ** F.round(F.log(v) / np.log(2.))
    
    n = get_parameter_or_create("n", (),
        ConstantInitializer(n_init),
        need_grad=True,
        as_need_grad=not fix_parameters)
    xmax = get_parameter_or_create("xmax", (),
        ConstantInitializer(xmax_init),
        need_grad=True,
        as_need_grad=not fix_parameters)
    
    # ensure that bitwidth is in specified range and an integer
    n = F.round(clip_scalar(n, n_min, n_max))
    if sign:
        n = n - 1
    
    # ensure that dynamic range is in specified range
    xmax = clip_scalar(xmax, xmax_min, xmax_max)
    
    # compute step size from dynamic range and make sure that it is a pow2
    d = quantize_pow2(xmax / (2 ** n - 1))
    
    # compute min/max value that we can represent
    if sign:
        xmin = -xmax
    else:
        xmin = nn.Variable((1,), need_grad=False)
        xmin.d = 0.
    
    # broadcast variables to correct size
    d = broadcast_scalar(d, shape=x.shape)
    xmin = broadcast_scalar(xmin, shape=x.shape)
    xmax = broadcast_scalar(xmax, shape=x.shape)
    
    # apply fixed-point quantization
    return d * F.round(F.clip_by_value(x, xmin, xmax) / d)
```
B.1.3 Case U3: Parametrization with respect to \(d\) and \(q_{\text{MAX}}\)

```python
def parametric_fixed_point_quantize_d_xmax(x, sign,
    d_init, d_min, d_max,
    xmax_init, xmax_min, xmax_max,
    fix_parameters=False):
    '''Parametric version of 'fixed_point_quantize' where the
    stepsize 'd' and dynamic range 'xmax' are learnable parameters.'''
    d = get_parameter_or_create("d", (),
        ConstantInitializer(d_init),
        need_grad=True,
        as_need_grad=not fix_parameters)
    xmax = get_parameter_or_create("xmax", (),
        ConstantInitializer(xmax_init),
        need_grad=True,
        as_need_grad=not fix_parameters)
    # ensure that stepsize is in specified range and a power of two
    d = quantize_pow2(clip_scalar(d, d_min, d_max))
    # ensure that dynamic range is in specified range
    xmax = clip_scalar(xmax, xmax_min, xmax_max)
    # compute min/max value that we can represent
    if sign:
        xmin = -xmax
    else:
        xmin = nn.Variable((1,), need_grad=False)
    xmin.d = 0.
    # broadcast variables to correct size
    d = broadcast_scalar(d, shape=x.shape)
    xmin = broadcast_scalar(xmin, shape=x.shape)
    xmax = broadcast_scalar(xmax, shape=x.shape)
    # apply fixed-point quantization
    return d * F.round(F.clip_by_value(x, xmin, xmax) / d)
```
B.2 POWER-OF-TWO QUANTIZATION

B.2.1 CASE P1: PARAMETRIZATION WITH RESPECT TO $b$ AND $q_{\text{MAX}}$

def parametric_pow2_quantize_b_xmax(x, sign, with_zero,
    n_init, n_min, n_max,
    xmax_init, xmax_min, xmax_max,
    fix_parameters=False):
    """Parametric version of 'pow2_quantize' where the
    bitwidth 'n' and range 'xmax' are learnable parameters.
    Returns:
    ~nnabla.Variable: N-D array.
    """
    def clip_scalar(v, min_value, max_value):
        return F.minimum_scalar(F.maximum_scalar(v, min_value), max_value)
    def broadcast_scalar(v, shape):
        return F.broadcast(F.reshape(v, (1,) * len(shape), inplace=False), shape=shape)
    def quantize_pow2(v):
        return 2 ** F.round(F.log(F.abs(v)) / np.log(2.))
    n = get_parameter_or_create("n", (),
        ConstantInitializer(n_init),
        need_grad=True,
        as_need_grad=not fix_parameters)
    xmax = get_parameter_or_create("xmax", (),
        ConstantInitializer(xmax_init),
        need_grad=True,
        as_need_grad=not fix_parameters)
    # ensure that bitwidth is in specified range and an integer
    n = F.round(clip_scalar(n, n_min, n_max))
    if sign:
        n = n - 1
    if with_zero:
        n = n - 1
    # ensure that dynamic range is in specified range and an integer
    xmax = quantize_pow2(clip_scalar(xmax, xmax_min, xmax_max))
    # compute min value that we can represent
    xmin = (2 ** (-2 ** n + 1)) * xmax
    # broadcast variables to correct size
    xmin = broadcast_scalar(xmin, shape=x.shape)
    xmax = broadcast_scalar(xmax, shape=x.shape)
    # if unsigned, then quantize all negative values to zero
    if not sign:
        x = F.relu(x)
    # compute absolute value/sign of input
    ax = F.abs(x)
    sx = F.sign(x)
    # if with_zero:
    # prime smallest elements (in magnitude) to zero if they are smaller
    # than "x_min / sqrt(2)"
    x_threshold = xmin / np.sqrt(2)
    # broadcast scalar to correct shape
    x_threshold = broadcast_scalar(x_threshold, shape=x.shape)
    # do not backpropagate gradient through indices
    idx1 = F.greater_equal(ax, x_threshold) * F.less(ax, xmin)
    idx2 = F.greater_equal(ax, xmin) * F.less(ax, xmax)
    idx3 = F.greater_equal(ax, xmax)
    else:
        idx1 = F.less(ax, xmin)
        idx2 = F.greater_equal(ax, xmin) * F.less(ax, xmax)
        idx3 = F.greater_equal(ax, xmax)
    # do not backpropagate gradient through sign
    sx.need_grad = False
    # take care of values outside of dynamic range
    return sx * (xmin * idx1 + quantize_pow2(ax) * idx2 + xmax * idx3)
B.2.2 CASE P2: PARAMETRIZATION WITH RESPECT TO $b$ AND $q_{\text{MIN}}$

```python
def parametric_pow2_quantize_b_xmin(x, sign, with_zero,
    n_init, n_min, n_max,
    xmin_init, xmin_min, xmin_max,
    fix_parameters=False):
    """Parametric version of 'pow2_quantize' where the
    bitwidth 'n' and the smallest value 'xmin' are learnable parameters.
    """
    Returns:
    """
    def clip_scalar(v, min_value, max_value):
        return F.minimum_scalar(F.maximum_scalar(v, min_value), max_value)
    def broadcast_scalar(v, shape):
        return F.broadcast(F.reshape(v, (1,) * len(shape), inplace=False), shape=shape)
    def quantize_pow2(v):
        return 2 ** F.round(F.log(F.abs(v)) / np.log(2.))
    n = get_parameter_or_create("n", (),
        ConstantInitializer(n_init),
        need_grad=True,
        as_need_grad=not fix_parameters)
    xmin = get_parameter_or_create("xmin", (),
        ConstantInitializer(xmin_init),
        need_grad=True,
        as_need_grad=not fix_parameters)
    # ensure that bitwidth is in specified range and an integer
    n = F.round(clip_scalar(n, n_min, n_max))
    if sign:
        n = n - 1
    if with_zero:
        n = n - 1
    # ensure that minimum dynamic range is in specified range and a power-of-two
    xmin = quantize_pow2(clip_scalar(xmin, xmin_min, xmin_max))
    # compute min/max value that we can represent
    xmax = xmin * (2 ** ((2 ** n) - 1))
    # broadcast variables to correct size
    xmin = broadcast_scalar(xmin, shape=x.shape)
    xmax = broadcast_scalar(xmax, shape=x.shape)
    # if unsigned, then quantize all negative values to zero
    if not sign:
        x = F.relu(x)
    # compute absolute value/sign of input
    ax = F.abs(x)
    sx = F.sign(x)
    if with_zero:
        # prune smallest elements (in magnitude) to zero if they are smaller
        # than `xmin / sqrt(2)`
        x_threshold = xmin / np.sqrt(2)
        idx1 = F.greater_equal(ax, x_threshold) * F.less(ax, xmin)
        idx2 = F.greater_equal(ax, xmin) * F.less(ax, xmax)
        idx3 = F.greater_equal(ax, xmax)
    else:
        idx1 = F.less(ax, xmin)
        idx2 = F.greater_equal(ax, xmin) * F.less(ax, xmax)
        idx3 = F.greater_equal(ax, xmax)
    # do not backpropagate gradient through indices
    idx1.need_grad = False
    idx2.need_grad = False
    idx3.need_grad = False
    # do not backpropagate gradient through sign
    sx.need_grad = False
    # take care of values outside of dynamic range
    return sx * (xmin * idx1 + quantize_pow2(ax) * idx2 + xmax * idx3)
```

20
B.2.3 CASE P3: PARAMETRIZATION WITH RESPECT TO $q_{\text{MIN}}$ AND $q_{\text{MAX}}$

```python
def parametric_pow2_quantize_xmin_xmax(x, sign, with_zero, 
    xmin_init, xmin_min, xmin_max, 
    xmax_init, xmax_min, xmax_max, 
    fix_parameters=False):
    ***Parametric version of 'pow2_quantize' where the
    min value 'xmin' and max value 'xmax' are learnable parameters.
    Returns:
    ***

def clip_scalar(v, min_value, max_value):
    return F.minimum_scalar(F.maximum_scalar(v, min_value), max_value)

def broadcast_scalar(v, shape):
    return F.broadcast(F.reshape(v, (1,) * len(shape), inplace=False), shape=shape)

def quantize_pow2(v):
    return 2. ** F.round(F.log(F.abs(v)) / np.log(2.))

xmin = get_parameter_or_create("xmin", (), 
    ConstantInitializer(xmin_init), 
    need_grad=True, 
    as_need_grad=not fix_parameters)
xmax = get_parameter_or_create("xmax", (), 
    ConstantInitializer(xmax_init), 
    need_grad=True, 
    as_need_grad=not fix_parameters)

# ensure that minimum dynamic range is in specified range and a power-of-two
xmin = quantize_pow2(clip_scalar(xmin, xmin_min, xmin_max))
xmax = quantize_pow2(clip_scalar(xmax, xmax_min, xmax_max))

# broadcast variables to correct size
xmin = broadcast_scalar(xmin, shape=x.shape)
xmax = broadcast_scalar(xmax, shape=x.shape)

# if unsigned, then quantize all negative values to zero
if not sign:
    x = F.relu(x)

# compute absolute value/sign of input
ax = F.abs(x)
sx = F.sign(x)

if with_zero:
    # prune smallest elements (in magnitude) to zero if they are smaller
    # than `\sqrt{2}`
    x_threshold = xmin / np.sqrt(2)
    idx1 = F.greater_equal(ax, x_threshold) * F.less(ax, xmin)
    idx2 = F.greater_equal(ax, xmin) * F.less(ax, xmax)
    idx3 = F.greater_equal(ax, xmax)
    else:
        idx1 = F.less(ax, xmin)
        idx2 = F.greater_equal(ax, xmin) * F.less(ax, xmax)
        idx3 = F.greater_equal(ax, xmax)

# do not backpropagate gradient through indices
idx1.need_grad = False
idx2.need_grad = False
idx3.need_grad = False

# do not backpropagate gradient through sign
sx.need_grad = False

# take care of values outside of dynamic range
return sx * (xmin * idx1 + quantize_pow2(ax) * idx2 + xmax * idx3)
```

21