

DBQ: A Differentiable Branch Quantizer for Lightweight Deep Neural Networks

Hassan Dbouk^{1,2*}, Hetul Sanghvi², Mahesh Mehendale², and Naresh Shanbhag¹

¹ Dept. of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, USA

{hdbouk2, shanbhag}@illinois.edu

² Kilby Labs, Texas Instruments Inc, Dallas, USA

{hetul, m-mehendale}@ti.com

Abstract. Deep neural networks have achieved state-of-the art performance on various computer vision tasks. However, their deployment on resource-constrained devices has been hindered due to their high computational and storage complexity. While various complexity reduction techniques, such as lightweight network architecture design and parameter quantization, have been successful in reducing the cost of implementing these networks, these methods have often been considered orthogonal. In reality, existing quantization techniques fail to replicate their success on lightweight architectures such as MobileNet. To this end, we present a novel fully differentiable non-uniform quantizer that can be seamlessly mapped onto efficient ternary-based dot product engines. We conduct comprehensive experiments on CIFAR-10, ImageNet, and Visual Wake Words datasets. The proposed quantizer (DBQ) successfully tackles the daunting task of aggressively quantizing lightweight networks such as MobileNetV1, MobileNetV2, and ShuffleNetV2. DBQ achieves state-of-the art results with minimal training overhead and provides the best (pareto-optimal) accuracy-complexity trade-off.

Keywords: Deep Learning, Quantization, Low-Complexity Neural Networks

1 Introduction

Deep neural networks (DNNs) have achieved state-of-the art accuracy on various computer vision tasks such as image classification [15,6] but at the expense of extremely high computational and storage complexity, e.g., ResNet-18 [6] needs $\sim 10^{12}$ 1-b full adders (FAs) and 3.74×10^8 -bits of activation and weight storage to achieve an accuracy of 70% on the ImageNet dataset. These high computational and storage costs inhibit the deployment of such DNNs on resource-constrained Edge devices. As a result, there is much interest in designing low-complexity DNNs without compromising their accuracy.

* Work done while at Kilby Labs.

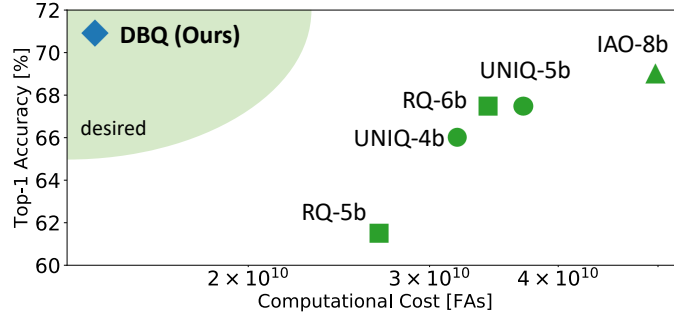


Fig. 1. The Top-1 accuracy on ImageNet vs. computational cost for MobileNetV1 achieved by state-of-the-art quantization methods (RQ [19], UNIQ [1], and IAO [13]). Our proposed method DBQ simultaneously achieves the highest accuracy and the lowest complexity.

There are two distinct approaches for reducing DNN complexity: 1) model compression [5] and quantization [10,22] of complex networks, and 2) the design of lightweight networks from scratch, e.g., MobileNet [8,27].

Model compression and quantization methods rely on the intrinsic over-parameterization in complex networks to reduce their complexity. Such methods have proved to be very effective in reducing network complexity with negligible impact on its accuracy, e.g., ternary quantization of ResNet-18 weights [32] reduces its computational and storage complexity by 88% and 74%, respectively, at the expense of a drop in accuracy from 70.3% to 69.1%.

In the second approach, the design of lightweight networks such as MobileNet [8,27], SqueezeNet [11], ShuffleNet [34], ConDenseNet [9] have also shown tremendous success. Such networks exploit algorithmic properties such as factorizability of convolutions and utilize either 1×1 convolutions (SqueezeNet), grouped convolutions (ShuffleNet, ConDenseNet), or both (MobileNet). For example, MobileNetV1 [8] achieves comparable (or even higher) accuracy than its ResNet-18 floating-point (FP) counterpart but at a computational and storage complexity that are $3\times$ and $7\times$ lower, respectively.

In contrast, not much work has been done in model compression or quantization of lightweight networks and for a good reason – such networks are already irredundant leaving much less room for complexity reduction. Existing works [19,30,1,13,28] that quantize lightweight networks use fixed-point quantization with relatively high bitwidths (see Fig. 1) which offer limited reductions in complexity. In contrast, aggressive quantization schemes such as binarization [10,22] or ternarization [36,16] have been benchmarked on over-parameterized networks. In fact, ternarizing MobileNetV1 leads to a catastrophic drop in accuracy from 72.12% to 66.45% on ImageNet as we show in Section 4.3. In order to improve the performance of ternarized models while leveraging the simplicity of ternary-based arithmetic, one can construct a non-uniform quantizer as linear combinations of ternary values. Such formulation has already been proposed in

the context of binarized neural networks [33,18], however the training algorithm involved is: 1) extremely inefficient to implement; 2) can lead to sub-optimal results due to gradient mismatch issues; and 3) has only been benchmarked on over-parameterized networks.

To this end, our work is the *first* to tackle the daunting task of aggressively quantizing lightweight networks, such as MobileNetV1 [8], MobileNetV2 [27], and ShuffleNetV2 [20] using multiple ternary branches. We propose an efficient and fully differentiable multiple ternary branch quantization algorithm (DBQ). For MobileNetV1 on ImageNet, DBQ achieves an accuracy 2% higher than state-of-the-art quantization methods with a complexity that is $3.5\times$ lower as shown in Fig. 1. This represents an overall reduction of $24.5\times$ compared to FP with a 1.2% drop in accuracy.

Specifically, our contributions are:

1. We are the *first* to successfully ternarize lightweight networks (MobileNetV1, MobileNetV2, ShuffleNetV2) on ImageNet. This result is achieved by using DBQ with two ternary branches.
2. We present the *first fully differentiable* branched quantization algorithm (DBQ) for DNNs requiring minimal training overhead.
3. We show that DBQ outperforms state-of-the-art methods in both accuracy and computational cost. Compared to state-of-the-art quantization method RQ [19], DBQ drastically improves the Top-1 accuracy of MobileNetV1 on ImageNet from 61.50% to 70.92% at iso-model size accompanied by a 19% reduction in computational complexity.
4. For lightweight networks tackling real world applications, we show that DBQ with two ternary branches offers the best (pareto-optimal) accuracy-complexity trade-off compared to using one ternary branch with higher number of channels, at iso-model size.

2 Related Work

Reducing DNN complexity via quantization has been an active area of research over the past few years. A majority of such works either train the quantized network from scratch [36,33,16,10,22,24] or fine-tune a pre-trained model with quantization-in-the-loop [13,19,30,32,1,35]. Where retraining is not an option, [25] provides analytical guarantees on the minimum precision requirements of a pre-trained FP network given a budget on the accuracy drop from FP. Training based quantization works fall into two classes of methods: 1) estimation based methods [33,18,16,30,13], where the full-precision weights and activations are quantized in the forward path, and gradients are back-propagated through a non-differentiable quantizer function via a gradient estimator such as the Straight Through Estimator (STE) [2]; and 2) optimization based methods, where gradients flow directly from the full-precision weights to the cost function via an approximate differentiable quantizer [32,19,24], or by including an explicit quantization error term to the loss function [7,35]. Application of these methods can be categorized into three clusters:

Aggressive Quantization: Methods such as binarization and ternarization have been highly successful for reducing DNN complexity. BinaryNets [10] quantize both weights and activations of DNNs to ± 1 , while XNORNets [22] use a full-precision scalar to represent binarized weights in order to improve accuracy. Ternary Weight Networks (TWN) [16] quantize weights to $\{-1, 0, 1\}$ and leverage the resulting weight sparsity due to the '0' state to skip operations. Trained Ternary Quantization (TTQ) [36] proposes learning the ternary scales via backprop. However, a major drawback of such methods is the resulting accuracy loss especially when applied to lightweight DNNs such as MobileNet. In Section 4.3, we show that ternarizing only the pointwise layers in MobileNetV1 on Imagenet, which correspond to $\sim 94\%$ of the total multiplication/additions (Table 1), incurs a massive accuracy loss ($\sim 5.67\%$) compared to the full-precision baseline. Hence, such methods are typically benchmarked on simple datasets such as CIFAR-10, or use over parameterized models such as AlexNet [15] or ResNet-18 [6] on ImageNet. In contrast, our proposed DBQ method is able to aggressively quantize the lightweight MobileNetV1 architecture with minimal loss in Top-1 accuracy (Fig 1).

Layer	Type	Mults [%]	Adds [%]	Params [%]
FL		1.89	1.83	0.02
DW		3.03	2.72	1.05
PW		94.02	94.37	74.19
FC		0.18	0.18	24.22
PL		0	0.01	0
BN		0.88	0.89	0.52

Table 1. The number of multiplications, additions, and parameters required by each layer type: first layer (FL), depthwise (DW), pointwise (PW), fully connected (FC), pooling layer (PL), and batch normalization (BN), for a single inference using MobileNetV1.

Non-uniform Quantization: These methods seek to improve the performance of binarized/ternarized models while leveraging their arithmetic simplicity, e.g., LQNets [33] and ABCNets [18], by quantizing weights and activations as linear combinations of binary values. The resulting non-uniform multi-bit quantization allows the computation of dot products to be carried out using binary arithmetic with appropriate scaling and addition. However, these methods suffer from two major drawbacks: 1) the design of their quantization functions is computationally expensive as it requires an iterative solution of a non-convex optimization problem per-layer per-forward pass during training, which results in a significant training time overhead in the range $1.4\times - 3.7\times$ [33]; and 2) they suffer from gradient mismatch problems as they depend on the STE [2] method to compute the gradients during training. This renders the quantizer constructed by these methods to be sub-optimal, since they *estimate* the quantizer parameters by minimizing a *local* cost function, e.g., MSE. Moreover, these methods have been benchmarked only on over parameterized networks on ImageNet. Whereas our proposed DBQ method *learns* the multiple *ternary* branches by minimizing a *global* loss function since the proposed quantizer is fully differentiable, which enables the efficient training of similar non-uniform quantizers, while also eradicating the need for any gradient estimator.

Quantization of Lightweight DNNs: Recent works that quantize MobileNets either apply fixed-point quantization with uniform [19,28,13] or mixed [30,29] precision across layers. Hardware-Aware Quantization (HAQ) [30] proposes using reinforcement learning to learn the per-layer bit-precision for both weights and activations, whereas [29] learns the bit-precision via a reformulation of the quantizer function and relying on the STE for gradient computation. Integer-Arithmetic-Only (IAO) [13] proposes using 8-b quantization for accelerating the inference of MobileNets on hardware platforms such as Qualcomm Hexagon and ARM NEON. Relaxed Quantization (RQ) [19] approximates the quantization function with a smooth differentiable approximate function, but the quantized values are still in fixed-point. Uniform Noise Injection Quantization (UNIQ) [1] proposes training a non-uniform quantizer using a special noise injection method that allows natural computation of gradients for quantized parameters. UNIQ uses a non-uniform quantizer requiring inefficient lookup tables and full precision multipliers/additions. Furthermore, all of these approaches use relatively high bitwidths ($\sim 6\text{b}–8\text{b}$), and most even fail to bridge the accuracy gap between the quantized models and their full-precision baseline. In contrast, the proposed DBQ method is able to aggressively reduce the precision of the dominant (94%) PW layers of MobileNetV1 to two ternary parameters with negligible degradation in the Top-1 accuracy.

3 Differentiable Branched Quantizer (DBQ)

A ternary B -branch quantizer $Q(\mathbf{w})$ of a full precision weight vector $\mathbf{w} \in \mathbb{R}^D$ (Fig 2(a)) is given by:

$$\mathbf{w}_q = Q(\mathbf{w}) = \sum_{j=1}^B \alpha_j \mathbf{w}_j \quad (1)$$

where $\mathbf{w}_j \in \{-1, 0, 1\}^D$ are the ternary branch weight vectors, and $\forall j \in [B]: \alpha_j > 0$ are per-branch scalars. In DBQ, we wish to learn all the network parameters which requires the quantizer function $Q(\mathbf{w})$ to be made differentiable. To do so, we first formulate a parametric form of $Q(\mathbf{w})$ in Section 3.1 and then employ a smooth ‘temperature-controlled’ approximation of the quantizer step function to establish its differentiability in Section 3.2.

3.1 Formulation of DBQ

We formulate the ternary B -branch quantizer in Fig. 2 as a $N = 3^B$ -level non-uniform quantizer $Q(\mathbf{w}) : \mathbb{R}^D \rightarrow \mathcal{V}^D$ with quantization levels $\mathcal{V} = \{v_i\}_{i=1}^N$. Assuming that the quantization levels v_i ’s are sorted in ascending order, the $Q(\mathbf{w})$ can be written as a linear combination of $N - 1$ step functions as shown below:

$$Q(\mathbf{w}) = \sum_{i=1}^{N-1} \left[(v_{i+1} - v_i) f(\mathbf{w} - t_i) \right] - \frac{v_N - v_1}{2} \quad (2)$$

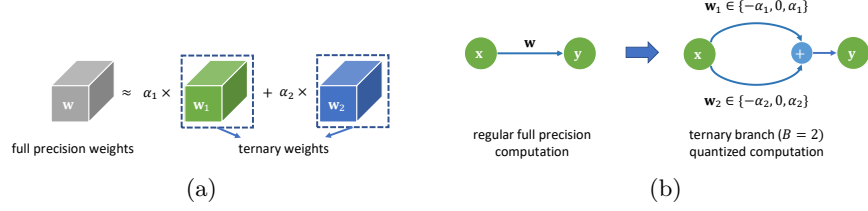


Fig. 2. Branched quantization of full precision weights: (a) as a linear combination of ternary weights, and (b) implemented as multiple parallel ternary branch operations to leverage the properties of ternary arithmetic for dot product computations.

where $f(\mathbf{u}) = [\mathbf{1}_{\{u_1 > 0\}}, \dots, \mathbf{1}_{\{u_D > 0\}}]^T$ is an element-wise ideal step function, and $\{t_i\}_{i=1}^{N-1}$ are the quantizer thresholds. The $(v_N - v_1)/2$ term is the quantizer offset. We impose the ternary quantizer structure in (1) via the constraint:

$$v_i = \sum_{j=1}^B e_{i,j} \alpha_j \quad (3)$$

where $e_{i,j} \in \{-1, 0, 1\}$, and thereby obtain the final quantizer expression:

$$Q(\mathbf{w}) = \gamma_2 \left[\sum_{i=1}^{N-1} \left[f(\gamma_1 \mathbf{w} - t_i) \sum_{j=1}^B b_{i,j} \alpha_j \right] - \sum_{j=1}^B \alpha_j \right] \quad (4)$$

where $b_{i,j} = e_{i+1,j} - e_{i,j} \in \{-2, -1, 0, 1, 2\} \forall j \in [B]$ are *fixed* coefficients, and γ_1 & γ_2 are pre/post-quantization scales to ensure that the quantizer operates on normalized inputs. Thus, the branched quantizer is parametrized by $\mathcal{P}_Q = \{\alpha_1, \dots, \alpha_B, \gamma_1, \gamma_2, t_1, \dots, t_{N-1}\}$ and these all need to be learned.

In this paper, we focus on the $B = 2$ case, i.e., two ternary branch, as visualized in Fig. 3, with $N = 3^2 = 9$ different quantization levels v_i . In this case, (4) can be expanded as:

$$\begin{aligned} Q(\mathbf{w}) = \gamma_2 \Big[& \alpha_2 f(\gamma_1 \mathbf{w} - t_1) + (\alpha_1 - \alpha_2) f(\gamma_1 \mathbf{w} - t_2) + (2\alpha_2 - \alpha_1) f(\gamma_1 \mathbf{w} - t_3) \\ & + (\alpha_1 - \alpha_2) f(\gamma_1 \mathbf{w} - t_4) + (\alpha_1 - \alpha_2) f(\gamma_1 \mathbf{w} - t_5) + (2\alpha_2 - \alpha_1) f(\gamma_1 \mathbf{w} - t_6) \\ & + (\alpha_1 - \alpha_2) f(\gamma_1 \mathbf{w} - t_7) + \alpha_2 f(\gamma_1 \mathbf{w} - t_8) - (\alpha_1 + \alpha_2) \Big] \end{aligned} \quad (5)$$

3.2 Differentiability

Inspired by [32,31], we replace the non-differentiable f in (4) with a smooth sigmoid approximation \hat{f}_T as follows:

$$\hat{f}_T(u) = \frac{1}{1 + \exp(-Tu)} \quad (6)$$

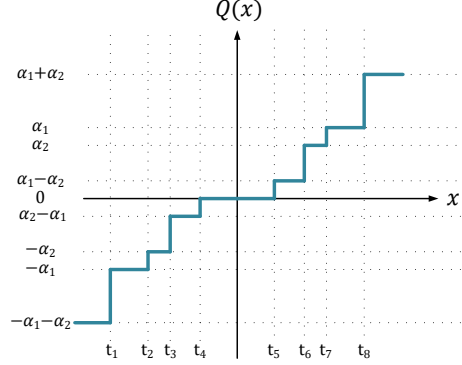


Fig. 3. Visualization of a two ternary (2T) branch quantizer with branch scales α_1 and α_2 assuming $\alpha_1 \geq \alpha_2 \geq \frac{\alpha_1}{2} \geq 0$.

where the *temperature* parameter T controls the approximation error, i.e.,:

$$e_T(u) = \hat{f}_T(u) - f(u) \xrightarrow{T \rightarrow \infty} 0 \quad (7)$$

When learning the quantizer parameters \mathcal{P}_Q , the temperature T is increased gradually as the training converges so that $\hat{f}_T(u) \rightarrow f(u)$. The resultant differentiable quantizer $Q_T(\mathbf{w}) = \mathbf{w}_q = \mathbf{z}$ therefore enables a straightforward calculation of the gradients for all quantizer and model parameters w.r.t. loss function \mathcal{L} as follows:

$$\frac{\partial \mathcal{L}}{\partial \gamma_2} = \frac{1}{\gamma_2} \sum_{k=1}^D \frac{\partial \mathcal{L}}{\partial z_k} z_k \quad (8)$$

$$\frac{\partial \mathcal{L}}{\partial \alpha_j} = \gamma_2 \sum_{k=1}^D \frac{\partial \mathcal{L}}{\partial z_k} \left[\sum_{i=1}^{N-1} [b_{i,j} g_{k,i}] - 1 \right] \quad (9)$$

$$\frac{\partial \mathcal{L}}{\partial t_i} = -\gamma_2 T \sum_{k=1}^D \frac{\partial \mathcal{L}}{\partial z_k} \left[h_{k,i} \sum_{j=1}^B b_{i,j} \alpha_j \right] \quad (10)$$

$$\frac{\partial \mathcal{L}}{\partial w_k} = \gamma_1 \gamma_2 T \frac{\partial \mathcal{L}}{\partial z_k} \sum_{i=1}^{N-1} \left[h_{k,i} \sum_{j=1}^B b_{i,j} \alpha_j \right] \quad (11)$$

$$\frac{\partial \mathcal{L}}{\partial \gamma_1} = \gamma_2 T \sum_{k=1}^D \frac{\partial \mathcal{L}}{\partial z_k} w_k \left[\sum_{i=1}^{N-1} \left[h_{k,i} \sum_{j=1}^B b_{i,j} \alpha_j \right] \right] \quad (12)$$

where $h_{k,i} = g_{k,i}(1 - g_{k,i})$ and $g_{k,i} = \hat{f}_T(\gamma_1 w_k - t_i)$ for brevity. By doing so, we eliminate the need for the STE and the expensive computational overhead introduced in estimation-based methods such as LQNet [33] or ABCNet [18]. Note that software frameworks such as PyTorch [21] automatically take care of computing these gradients so these don't need to be explicitly coded.

3.3 Implementation Details

Parameter Initialization: Initializing the quantizer parameters \mathcal{P}_Q is performed once before training and requires an initial vector $\mathbf{w} \in \mathbb{R}^D$ which can be from a pre-trained network or from random initialization (training from scratch). The initialization procedure is as follows: 1) the post-quantization scale γ_2 is set to the maximum absolute value in \mathbf{w} , and the pre-quantization scale γ_1 is set to $1/\gamma_2$. This ensures that the quantizer operates on normalized parameters which facilitates the optimization of its parameters, and that the quantized values are of the same scale as the inputs; 2) to find the optimal thresholds $\{t_i\}_{i=1}^{N-1}$, we first compute the optimal N centroids $\{c_i\}_{i=1}^N$ of the normalized vector $\gamma_1 \mathbf{w}$ via k -means, and then $\forall i \in [N-1]$ we set t_i to be the midpoint of the interval $[c_i, c_{i+1}]$; and 3) a good initialization for $\{\alpha_j\}_{j=1}^B$ is found by solving for the optimal values that minimize the L_2 norm between the normalized vector $\gamma_1 \mathbf{w}$ and its quantized counterpart.

Training and Inference: During training, the proposed DBQ quantizer is used with the approximate smooth step function \hat{f}_T for both forward and backward calculations ((4) & (8)–(12)). For a given layer in the network that performs the function $\mathbf{y} = F(\mathbf{w}, \mathbf{x})$, applying DBQ simply boils down to composing the quantizer described in (4) with the function F : $\mathbf{y} = F(Q_T(\mathbf{w}), \mathbf{x})$. For quantizing convolutional layers, we apply kernel-wise quantizers. The overhead of full precision scales is amortized across the large filter lengths. The choice of the temperature parameter T is important. A large value of T would reduce the approximation error in (7), however the gradients would saturate quickly, thus causing a bottleneck for learning the quantizer parameters. Therefore, an initial small value for T is used for the first training epoch, and its value is increased for successive epochs based on a pre-determined temperature update schedule. A simple yet effective schedule is to linearly increment the temperature with the number of epochs: $T = T_{\text{init}} + e \times T_{\text{inc}}$. During inference, the approximate step function is replaced with the ideal function f such that the quantizer output satisfies (1).

Activation Quantization: The challenge in quantizing input activations with a fixed-point quantizer during training is determining a suitable clipping value (Fig. 4). Traditionally, the use of ReLU6 (which clips at 6) has been a popular choice due to its simplicity [28, 13]. However, the choice of 6 provides no guarantees on the clipping probability, and can therefore yield sub-optimal results. Similar to [4], we propose clipping the post-BN activations y_{BN} (Fig. 4) using:

$$c = \max_{i \in [C]} (\beta_l^{(i)} + k\gamma_l^{(i)}) \quad (13)$$

where C is the number of channels in the activation tensor y_1 , $(\beta^{(i)}, \gamma^{(i)})$ are learnable per-channel shift and scale parameters of BN, and k is a network hyper-parameter that controls the clipping probability. Assuming that the distribution of $y_{\text{BN}}^{(i)} \sim \mathcal{N}(\beta^{(i)}, (\gamma^{(i)})^2)$ [12] and using 6σ rule ($k = 6$), one can show that the choice of c in (13) guarantees:

$$\Pr\{y_{\text{BN}} \leq c\} \geq 0.999 \quad (14)$$

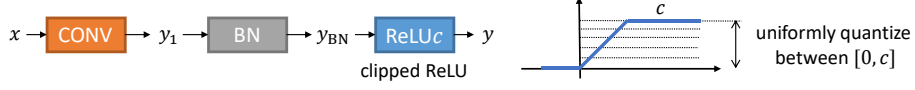


Fig. 4. Quantizing activations post-ReLU requires a pre-determined clipping parameter c .

Note that having a fixed clipping value c for all channels is crucial in order to ensure that the dot product operations can be implemented in fixed-point.

4 Experimental Results

To demonstrate the effectiveness of the DBQ method for quantizing lightweight networks, we evaluate it on three different image classification datasets: 1) CIFAR-10 [14] using ResNet-20 [6]; 2) ImageNet (ILSVRC 2012) [23] using MobileNetV1 [8], MobileNetV2 [27], and ShuffleNetV2 [20]; and 3) the recently proposed Visual Wake Words [3] using MobileNetV1. In all of our experiments, we train full precision models from scratch, and perform fine tuning on said models for training their quantized counterparts. We use stochastic gradient descent for training all the models. For further details on the training setup for each experiment, please check the supplementary material.

4.1 Complexity Metrics

We propose a set of metrics, inspired by those used in [25,26], in order to quantify the complexity reduction achieved by our proposed method.

Computational Cost (\mathcal{C}_C) for an L -layer network:

$$\mathcal{C}_C = \sum_{l=1}^L N_l \left[D_l B_{W,l} B_{A,l} + (D_l - 1)(B_{A,l} + B_{W,l} + \lceil \log_2 D_l \rceil - 1) \right] \quad (15)$$

where N_l is the number of D_l -dimensional dot products in layer l with $B_{W,l}$ and $B_{A,l}$ being the weights and activations precisions respectively. This cost essentially measures the number of 1b full adders (FAs) needed to implement the dot products required for a given network. For full precision (32b) parameters, we make the simplifying assumption of treating them as 23b (mantissa precision) fixed-point parameters.

Sparsity-Aware Computational Cost (\mathcal{C}_S) is also defined in order to leverage weight-sparsity in different models that can be reflected on the model complexity:

$$\mathcal{C}_S = \sum_{l=1}^L N_l \left[D'_l B_{W,l} B_{A,l} + (D'_l - 1)(B_{A,l} + B_{W,l} + \lceil \log_2 D_l \rceil - 1) \right] \quad (16)$$

Method	Acc. (Δ) [%]	\mathcal{C}_C (\mathcal{C}_S) [10^9 FA]	\mathcal{C}_R (\mathcal{C}_M) [10^6 b]
FP [33]	92.10 (/)	23.73 (23.73)	14.63 (8.63)
LQNet-1B [33]	90.10 (−2.171)	1.60 (1.60)	6.34 (0.35)
LQNet-2B [33]	91.80 (−0.325)	2.83 (2.83)	6.61 (0.61)
LQNet-3B [33]	92.00 (−0.108)	4.07 (4.07)	6.88 (0.88)
FP (Ours)	92.00 (/)	23.73 (23.73)	14.63 (8.63)
DBQ-1T (Ours)	91.06 (− 1.021)	1.60 (0.92)	6.61 (0.61)
DBQ-2T (Ours)	91.93 (− 0.076)	2.83 (1.79)	7.15 (1.15)

Table 2. The accuracy on CIFAR-10 and complexity metrics (\mathcal{C}_C , \mathcal{C}_S , \mathcal{C}_R , \mathcal{C}_M) for ResNet-20 using our method DBQ compared to LQNet. Δ represents the normalized accuracy drop of the quantized models with respect to its full precision baseline. The first, last layers, and input activations are kept in full precision for the quantized models in accordance with [33].

where D'_l is the number of non-zero weights in the corresponding D_l -dimensional dot product.

Representational Cost (\mathcal{C}_R) for an L -layer network:

$$\mathcal{C}_R = \sum_{l=1}^L \left[|W_l| B_{W,l} + |A_l| B_{A,l} \right] \quad (17)$$

where $|W_l|$ and $|A_l|$ are the number of elements in the weight and activation tensors in layer l , respectively.

Model Storage Cost (\mathcal{C}_M) for an L -layer network:

$$\mathcal{C}_M = \sum_{l=1}^L |W_l| B_{W,l} \quad (18)$$

which only accounts for the weight storage, and can be useful for studying model compression.

4.2 CIFAR-10 Results

We first demonstrate the effectiveness of DBQ on the CIFAR-10 dataset using the popular network ResNet-20 [6]. To ensure a fair comparison with the LQNet [33] models, we do not quantize the first and last fully connected layers, and we keep all activations in full precision. Table 2 summarizes the accuracy (and percentage drop) as well as the four complexity metrics (\mathcal{C}_C , \mathcal{C}_S , \mathcal{C}_R , \mathcal{C}_M) for different number of branches used. At iso-number of branches, the DBQ models achieve higher accuracies for the same \mathcal{C}_C and lower \mathcal{C}_S due to the high number of zero valued weights, as opposed to binary branches where the weights are either ± 1 . Comparing the DBQ-2T and LQNet-3B models, which achieve comparable accuracies, DBQ-2T requires $\sim 32\%$ less \mathcal{C}_C and $\sim 56\%$ less \mathcal{C}_S , at the expense of an extra bit per-parameter, which is reflected in the marginal $\sim 4\%$ increase in \mathcal{C}_R .

Model Name	Activations	FL	DW	PW	FC	Top-1/5 Acc. [%]	\mathcal{C}_C (\mathcal{C}_S) [10^{10} FA]	\mathcal{C}_R (\mathcal{C}_M) [10^7 b]
FP	ReLU -	32b	32b	32b	32b	72.12/90.43	33.37 (33.37)	30.00 (13.54)
FX8-1	ReLU6 -	8b	32b	8b	8b	71.65/90.17	5.78 (5.39)	10.38 (5.90)
FX8-2	ReLU6 -	8b	8b	8b	8b	71.60/90.19	5.24 (4.85)	7.56 (3.44)
FX8-3	ReLU x -	8b	8b	8b	8b	71.86/90.26	5.24 (4.85)	7.56 (3.44)
DBQ-1T	ReLU -	32b	32b	32b	1T	66.45/86.72	3.60 (2.61)	20.58 (4.12)
DBQ-2T-1	ReLU -	32b	32b	32b	2T	71.09/89.71	5.23 (3.77)	21.21 (4.75)
DBQ-2T-2	ReLU6 -	8b	32b	8b	2T	70.25/89.42	2.73 (1.97)	9.12 (4.64)
DBQ-2T-3	ReLU x -	8b	32b	8b	2T	70.80/89.75	2.73 (1.97)	9.12 (4.64)
DBQ-2T-4	ReLU x -	8b	8b	8b	2T	70.92/89.61	2.18 (1.42)	6.30 (2.18)

Table 3. The Top-1/5 accuracy on ImageNet and complexity metrics (\mathcal{C}_C , \mathcal{C}_S , \mathcal{C}_R , \mathcal{C}_M) for MobileNetV1 under different precision assignments. Models denoted by DBQ- z T are trained using our differentiable branch quantizer with $B = z$ ternary branches. ReLU x denotes a clipped ReLU using our proposed clipping method in Eq. (13).

4.3 ImageNet Results

In this section, we report results for MobileNetV1 [8], MobileNetV2 [27], and ShuffleNetV2 [20] on ImageNet. We first focus on MobileNetV1 by performing an ablation study, and leverage these results for quantizing the more recent MobileNetV2 and ShuffleNetV2.

Ablation Study: Table 3 summarizes the Top-1,5 accuracies of all the MobileNetV1 models trained with different layer precision assignments in order to evaluate the impact of our design choices. To see the impact of using two ternary branches instead of one, we begin with the DBQ-1T model which is obtained by quantizing only the PW layers of MobileNetV1 to one ternary branch (1T) keeping all other activations and weights in full precision. Table 3 shows that DBQ-1T achieves a massive 89% reduction in \mathcal{C}_C compared to the FP model but at a catastrophic loss of 5.67% in the Top-1 accuracy. In contrast, DBQ-2T-1, which is DBQ-1T with a second ternary branch, is able to recover accuracy to within 1.03% of the full-precision baseline while also achieving massive savings in \mathcal{C}_C of 84%. Quantizing the activations and the remaining layers weights of DBQ-2T-1 to 8b fixed-point, i.e., DBQ-2T-4, incurs a minimal loss in accuracy of 1.2% compared to the FP model while also achieving even greater reduction in both \mathcal{C}_C (93%) and \mathcal{C}_R (70%). The reduction in \mathcal{C}_S increases to 96% when branch sparsity is exploited to skip computations.

Note that the reason that only PW layers are quantized using ternary branches is three-fold: 1) PW layers consume $\sim 94\%$ of the amount of multiply-adds required for inference (Table 1); 2) we have observed that quantizing the PW layers has the most severe impact on classification accuracy compared to quantizing other layers; and 3) DW layers suffer from extremely small dot-product lengths (9), rendering them unsuitable for multiple branch quantization (the overhead of branch-merge and scaling operations will dominate).

The benefits of our proposed BN-based clipping described in (13) can be seen by comparing the accuracy of the 8b fixed-point model FX8-3 using BN-based clipping with $k = 6$ with its ReLU6-based clipping counterpart FX8-2. The Top-

Method	Act.	FL	DW	PW	FC	Top-1 Acc. [%]	\mathcal{C}_C (\mathcal{C}_S) [10^{10} FA]	\mathcal{C}_R (\mathcal{C}_M) [10^7 b]
IAO* [13]	8b	8b	8b	8b	8b	69.00*	4.97 (/)	7.49 (3.37)
UNIQ [1]	8b	5b	5b	5b	5b	67.50	3.70 (/)	6.29 (2.18)
UNIQ [1]	8b	4b	4b	4b	4b	66.00	3.19 (/)	5.87 (1.76)
UNIQ [1]	8b	8b	8b	8b	8b	68.25	5.24 (/)	7.56 (3.44)
QSM* [28]	8b	8b	8b	8b	8b	68.03	4.97 (/)	7.49 (3.37)
RQ [19]	5b	5b	5b	5b	5b	61.50	2.68 (/)	4.75 (2.18)
RQ [19]	6b	6b	6b	6b	6b	67.50	3.42 (/)	5.69 (2.60)
HAQ cloud [30]	mixed	8b	mixed	mixed	8b	65.33 – 71.20 [†]	2.73 (/)	5.09 (3.12)
HAQ edge [30]	mixed	8b	mixed	mixed	8b	67.40 – 71.20 [†]	4.06 (/)	5.87 (2.49)
FX8 (Ours)	8b	8b	8b	8b	8b	71.86	5.24 (4.85)	7.56 (3.44)
DBQ-2T (Ours)	8b	8b	8b	2T	8b	70.92	2.18 (1.42)	6.30 (2.18)

*models with BN folding *results extracted from a plot [†]exact accuracy not reported

Table 4. The Top-1 accuracy on ImageNet and complexity metrics (\mathcal{C}_C , \mathcal{C}_S , \mathcal{C}_R , \mathcal{C}_M) for MobileNetV1 using our method (DBQ-2T) compared to state-of-the art training-based quantization methods.

1 accuracy of FX8-3 is better than FX8-2 without any overhead in training or inference. Similarly for DBQ-2T-3 and DBQ-2T-2.

Branching Utility: A 2T quantizer should result in 9 distinct quantization levels as shown in Fig. 3. However, in a 2T branched quantizer such as ours, it is possible for the number of quantization levels to be smaller than 9, e.g., if $\alpha_1 = \alpha_2$ then the number of quantization levels is 5. In this case, the full representational power of the 2T branched quantizer is not utilized. To see if the 2T branched quantizer generates all 9 levels, we plot the distribution of the ratio $R_\alpha = \frac{\alpha_1}{\alpha_2}$ across all the PW layers in the DBQ-2T-4 model (Table 3). The distribution is centered around $R_\alpha = 1.48$ with more than 99% of the values lying in the range [1.2, 1.7]. This demonstrates that the quantizer learned by DBQ employs the full representational power offered by the 2T structure.

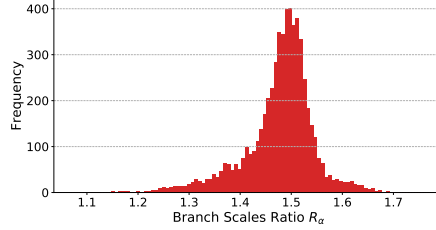


Fig. 5. The distribution of the ratio of the ternary branch scales α_1 and α_2 for DBQ-2T-4 from Table 3.

Comparison with State-of-the Art: Table 4 compares the performance of our proposed DBQ method against state-of-the art results on ImageNet for MobileNetV1. Our model DBQ-2T, which corresponds to DBQ-2T-4 in Table 3 achieves the lowest computational cost \mathcal{C}_C (2.18×10^{10} FAs) compared to previously published networks, while achieving the highest Top-1 accuracy 70.92%. Compared to the lowest complexity model RQ [19], DBQ-2T achieves a 19% reduction in \mathcal{C}_C with a 9.42% improvement in Top-1 accuracy at iso-storage

Model	Act.	FL	DW	PW	FC	Top-1 Acc. [%]	\mathcal{C}_C (\mathcal{C}_S) [10^{10} FA]	\mathcal{C}_R (\mathcal{C}_M) [10^7 b]
MobileNetV2-FP	32b	32b	32b	32b	32b	71.88	17.83 (17.83)	32.87 (11.22)
MobileNetV2-2T	8b	8b	8b	2T	8b	70.54	1.42 (1.11)	7.45 (2.04)
ShuffleNetV2-FP	32b	32b	32b	32b	32b	69.36	8.52 (8.52)	13.81 (7.29)
ShuffleNetV2-2T	8b	8b	8b	2T	8b	66.74	0.64 (0.46)	3.21 (1.38)

Table 5. The Top-1 accuracy on ImageNet and complexity metrics (\mathcal{C}_C , \mathcal{C}_S , \mathcal{C}_R , \mathcal{C}_M) for MobileNetV2 and ShuffleNetV2 using our method (DBQ-2T).

complexity \mathcal{C}_M . Furthermore, DBQ-2T improves upon the accuracy of the IAO model [13], which currently achieves the highest Top-1 accuracy, by 1.92% but with a massive reduction in complexity \mathcal{C}_C (56%), \mathcal{C}_R (16%), and \mathcal{C}_M (35%).

More Lightweight Networks: Table 5 demonstrates the performance of DBQ when applied to the more recent lightweight networks: MobileNetV2 and ShuffleNetV2. Similar to MobileNetV1, we find that the PW layers *dominate* the number of operations required for a single inference for both MobileNetV2 (87%) and ShuffleNetV2 (90%). Thus, and inline with our experiments on MobileNetV1, we quantize all PW layers using 2T, with the remaining layers and activations quantized to 8b fixed-point. We observe a minimal 1.3% (MobileNetV2) and 2.6% (ShuffleNetV2) drop in accuracy compared to FP, while achieving *massive* (77%–95%) reductions in all the complexity metrics. A comparison between DBQ and [29] for MobileNetV2 is present in the supplementary material.

4.4 Visual Wake Words Results

We study the accuracy-precision-complexity trade-off in quantized DNNs using the Visual Wake Words (VWW) dataset that was recently proposed by Google [3] in order to facilitate the development of lightweight vision models for deployment on resource-constrained Edge devices. This dataset reflects a typical real-world scenario involving the detection of specific events by observing incoming data, e.g., monitoring a camera video feed in order to detect the presence of a person [3], similar to the use of audio wake words in speech recognition. The VWW dataset is derived from the COCO dataset [17] via a simple re-labeling of the available images has a training set of 115k images and a test set with 8k images.

As in [3], we employ the modified MobileNetV1 architecture which has a FC layer with 2 output classes instead of 1000. The complexity of the network is tuned by varying the network width multiplier [8] $m \in \{0.125, 0.25, 0.375, 0.5\}$. Similar to our ImageNet experiments, we quantize all layers to 8b fixed-point and vary the precision of the PW layers using 8b-to-2b fixed-point and DBQ-1T and DBQ-2T.

As shown in Fig. 6a, for over parameterized models, e.g., $m = 0.5$, we find DBQ-1T (red square) shows a massive reduction in \mathcal{C}_S ($\sim 69\%$) at iso-accuracy compared to the fixed-point models (red circle) (Fig. 6a). In contrast, for lightweight models, e.g., $m = 0.125$, DBQ-1T (blue square) achieves an impressive 45% reduction in \mathcal{C}_S but at the expense of a 3% loss in test accuracy

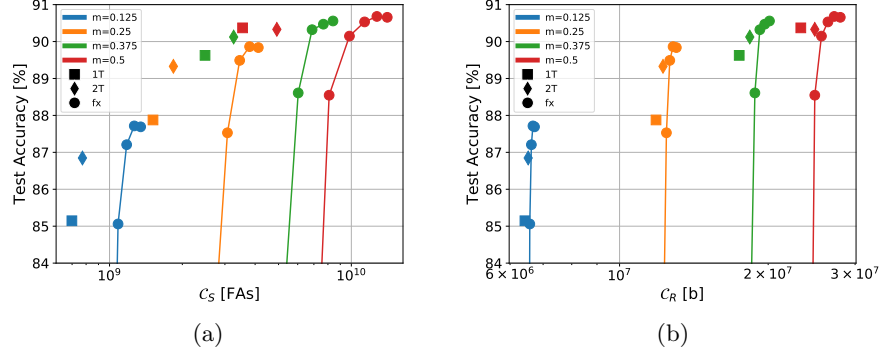


Fig. 6. The test accuracy of MobileNetV1 on the Visual Wake Words dataset with varying precision assignment and width multiplier m vs. (a) sparsity-aware computational cost, and (b) representational cost. Only the precision of the pointwise layer’s weights are changing, whereas all the remaining activations and weights are quantized using 8b fixed-point.

as compared to the fixed-point model (blue circle) (Fig. 6a). The DBQ models (diamonds and squares) can be seen to form a pareto-optimal accuracy-vs. C_S trade-off curve in Fig. 6a demonstrating its effectiveness.

Fig. 6b shows that the choice of the width multiplier m has a much more significant impact on the representational cost C_R than varying bit-precision which implies that C_R is dominated by the storage requirements of activations rather than weights. This implies that the choice of the model parameter m is governed by the amount of on-chip storage available on an Edge device. In contrast, the choice of the bit precision of the PW layers is dictated by the latency/energy requirements which upper bounds C_S as seen in Fig. 6a. As a result, when comparing the lightweight $m = 0.25$ DBQ-2T model (orange diamond) with the over parameterized $m = 0.375$ DBQ-1T model (green square), we observe that DBQ-2T achieves a reduction in both C_S (26%) and C_R (30%), at iso-accuracy and iso- C_M ($\sim 10^6$ b).

5 Conclusion

We presented DBQ, an efficient fully differentiable method for training multiple ternary branch quantizers for deep neural networks and validated its effectiveness for lightweight networks on the CIFAR-10 (ResNet-20), ImageNet (MobileNetV1, MobileNetV2, and ShuffleNetV2) and Visual Wake Words (MobileNetV1) datasets. Our method outperforms the state-of-the-art quantization schemes in both accuracy and complexity metrics.

Acknowledgment: The authors would like to thank Avishek Biswas, Manu Mathew and Arthur Redfern for helpful discussions and support.

References

1. Baskin, C., Schwartz, E., Zheltonozhskii, E., Liss, N., Giryes, R., Bronstein, A.M., Mendelson, A.: UNIQ: Uniform noise injection for non-uniform quantization of neural networks. arXiv preprint arXiv:1804.10969 (2018) [2](#), [3](#), [5](#), [12](#)
2. Bengio, Y., Léonard, N., Courville, A.: Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432 (2013) [3](#), [4](#)
3. Chowdhery, A., Warden, P., Shlens, J., Howard, A., Rhodes, R.: Visual wake words dataset. arXiv preprint arXiv:1906.05721 (2019) [9](#), [13](#)
4. Dbouk, H., Geng, H., Vineyard, C.M., Shanbhag, N.R.: Low-complexity fixed-point convolutional neural networks for automatic target recognition. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1598–1602. IEEE (2020) [8](#)
5. Han, S., Pool, J., Tran, J., Dally, W.: Learning both weights and connections for efficient neural network. In: Advances in neural information processing systems. pp. 1135–1143 (2015) [2](#)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) [1](#), [4](#), [9](#), [10](#)
7. Hou, L., Kwok, J.T.: Loss-aware weight quantization of deep networks. arXiv preprint arXiv:1802.08635 (2018) [3](#)
8. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017) [2](#), [3](#), [9](#), [11](#), [13](#)
9. Huang, G., Liu, S., Van der Maaten, L., Weinberger, K.Q.: Condensenet: An efficient densenet using learned group convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2752–2761 (2018) [2](#)
10. Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R., Bengio, Y.: Binarized neural networks. In: Advances in neural information processing systems. pp. 4107–4115 (2016) [2](#), [3](#), [4](#)
11. Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K.: Squeezenet: Alexnet-level accuracy with 50x fewer parameters and 0.5 mb model size. arXiv preprint arXiv:1602.07360 (2016) [2](#)
12. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015) [8](#)
13. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., Kalenichenko, D.: Quantization and training of neural networks for efficient integer-arithmetic-only inference. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2704–2713 (2018) [2](#), [3](#), [5](#), [8](#), [12](#), [13](#)
14. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. Tech. rep., Citeseer (2009) [9](#)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012) [1](#), [4](#)
16. Li, F., Zhang, B., Liu, B.: Ternary weight networks. arXiv preprint arXiv:1605.04711 (2016) [2](#), [3](#), [4](#)
17. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) [13](#)

18. Lin, X., Zhao, C., Pan, W.: Towards accurate binary convolutional neural network. In: *Advances in Neural Information Processing Systems*. pp. 345–353 (2017) [3](#), [4](#), [7](#)
19. Louizos, C., Reisser, M., Blankevoort, T., Gavves, E., Welling, M.: Relaxed quantization for discretized neural networks. *arXiv preprint arXiv:1810.01875* (2018) [2](#), [3](#), [5](#), [12](#)
20. Ma, N., Zhang, X., Zheng, H.T., Sun, J.: Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. pp. 116–131 (2018) [3](#), [9](#), [11](#)
21. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in PyTorch. In: *NIPS Autodiff Workshop* (2017) [7](#)
22. Rastegari, M., Ordonez, V., Redmon, J., Farhadi, A.: Xnor-net: Imagenet classification using binary convolutional neural networks. In: *European Conference on Computer Vision*. pp. 525–542. Springer (2016) [2](#), [3](#), [4](#)
23. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. *International journal of computer vision* **115**(3), 211–252 (2015) [9](#)
24. Sakr, C., Choi, J., Wang, Z., Gopalakrishnan, K., Shanbhag, N.: True gradient-based training of deep binary activated neural networks via continuous binarization. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 2346–2350. IEEE (2018) [3](#)
25. Sakr, C., Kim, Y., Shanbhag, N.: Analytical guarantees on numerical precision of deep neural networks. In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. pp. 3007–3016. JMLR. org (2017) [3](#), [9](#)
26. Sakr, C., Shanbhag, N.: Per-tensor fixed-point quantization of the back-propagation algorithm. In: *International Conference on Learning Representations* (2019) [9](#)
27. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 4510–4520 (2018) [2](#), [3](#), [9](#), [11](#)
28. Sheng, T., Feng, C., Zhuo, S., Zhang, X., Shen, L., Aleksic, M.: A quantization-friendly separable convolution for mobilenets. In: *2018 1st Workshop on Energy Efficient Machine Learning and Cognitive Computing for Embedded Applications (EMC2)*. pp. 14–18. IEEE (2018) [2](#), [5](#), [8](#), [12](#)
29. Uhlich, S., Mauch, L., Cardinaux, F., Yoshiyama, K., Garcia, J.A., Tiedemann, S., Kemp, T., Nakamura, A.: Mixed precision DNNs: All you need is a good parametrization. In: *International Conference on Learning Representations* (2020), <https://openreview.net/forum?id=HyxOslrFvH> [5](#), [13](#)
30. Wang, K., Liu, Z., Lin, Y., Lin, J., Han, S.: Haq: Hardware-aware automated quantization with mixed precision. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 8612–8620 (2019) [2](#), [3](#), [5](#), [12](#)
31. Xie, S., Zheng, H., Liu, C., Lin, L.: SNAS: stochastic neural architecture search. In: *International Conference on Learning Representations* (2019), <https://openreview.net/forum?id=rylqooRqK7> [6](#)
32. Yang, J., Shen, X., Xing, J., Tian, X., Li, H., Deng, B., Huang, J., Hua, X.s.: Quantization networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 7308–7316 (2019) [2](#), [3](#), [6](#)

- 33. Zhang, D., Yang, J., Ye, D., Hua, G.: Lq-nets: Learned quantization for highly accurate and compact deep neural networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 365–382 (2018) [3](#), [4](#), [7](#), [10](#)
- 34. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6848–6856 (2018) [2](#)
- 35. Zhou, A., Yao, A., Wang, K., Chen, Y.: Explicit loss-error-aware quantization for low-bit deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9426–9435 (2018) [3](#)
- 36. Zhu, C., Han, S., Mao, H., Dally, W.J.: Trained ternary quantization. arXiv preprint arXiv:1612.01064 (2016) [2](#), [3](#), [4](#)