

Bit-Mixer: Mixed-precision networks with runtime bit-width selection

Adrian Bulat
Samsung AI Cambridge
adrian@adrianbulat.com

Georgios Tzimiropoulos
Samsung AI Cambridge
Queen Mary University of London
g.tzimiropoulos@qmul.ac.uk

Abstract

Mixed-precision networks allow for a variable bit-width quantization for every layer in the network. A major limitation of existing work is that the bit-width for each layer must be predefined during training time. This allows little flexibility if the characteristics of the device on which the network is deployed change during runtime. In this work, we propose Bit-Mixer, the very first method to train a meta-quantized network where during test time any layer can change its bit-width without affecting at all the overall network’s ability for highly accurate inference. To this end, we make 2 key contributions: (a) Transitional Batch-Norms, and (b) a 3-stage optimization process which is shown capable of training such a network. We show that our method can result in mixed precision networks that exhibit the desirable flexibility properties for on-device deployment without compromising accuracy. Code will be made available.

1. Introduction

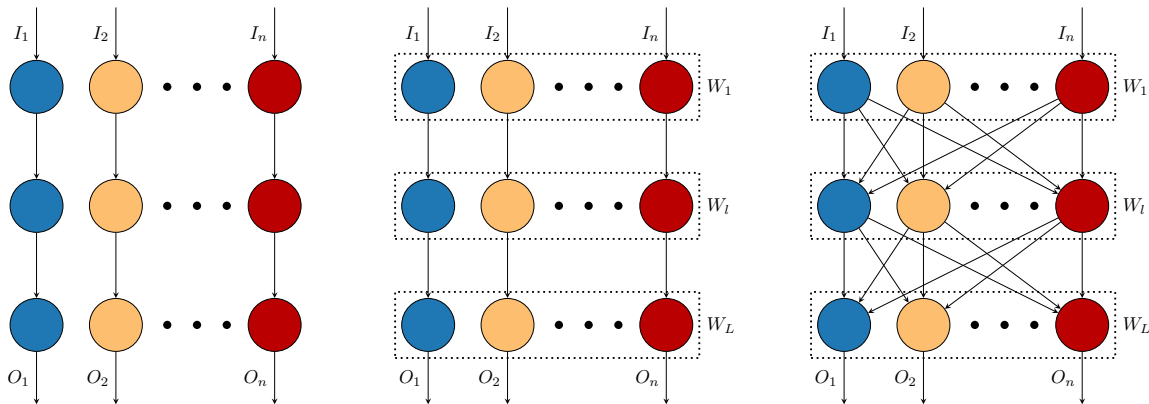
Deep Neural Networks have reached state-of-the-art accuracy across a plethora of computer vision and machine learning tasks. Despite their unprecedented accuracy, directly deploying such models on devices with limited computational resources and/or power constraints remains prohibitive. To address this problem, a series of related research directions have emerged such as network pruning [27, 33, 23], network compression [22, 38, 25], neural architecture search [26, 5] and network quantization. The latter offers the most straightforward improvements as using fewer bits for the weights and activations significantly reduces the compute and storage requirements. For example, switching from FP32 to Int-8 precision, a $4\times$ improvement in terms of speed and storage is obtained without any bells and whistles. This paper is on mixed-precision networks which allow for a variable bit-width quantization for every layer in the network.

Mixed-bit precision networks allow for a finer granularity of quantization at a layer level and, hence, offer prac-

tical advantages in terms of finding a more optimal trade-off between efficiency (i.e. speed) and memory requirements, and network accuracy. While this is more flexible than having the same bit-width across the whole network, mixed-bit precision approaches have also their own limitations. Firstly, due to an ever-growing number of different hardware platforms that a developer needs to support, each with its own unique characteristics and capabilities, quantizing networks, partially or fully, with mixed-bit precision in order to obtain an optimal trade-off between accuracy and speed becomes challenging. Secondly, and more importantly, even on the same device, due to either other concurrent processes running, battery level, temperature or simply prioritization, the available resources can vary. Ideally, a network should be able to dynamically react to these changes and adapt its quantization level per layer or module *on the fly* without incurring undesirable, or even more importantly, unpredictable penalties on inference accuracy.

The method we propose in this paper, coined **Bit-Mixer**, attempts to provide an answer to the aforementioned challenges. Bit-Mixer shifts away the focus from finding the optimal bit-width allocation per layer during training as done in *all previous* work. Instead, we propose to train a meta-quantized network which during test time can switch to any quantization level for any layer in the network. Training such meta-networks is however non-trivial due to the exponential number of unique combinations, the weight sharing constraint across different bit-widths, and the drastic variations in representational power that occur when the bit-width changes (e.g. 4 bits vs 1 bit). To this end, we make the following **contributions**:

1. Transitional Batch-Norms: To properly compensate for the distribution shift that arises when a change in the bit-width occurs between two consecutive layers, for each transition between different bit-widths, we propose to learn a separate batch normalization layer, coined Transitional Batch-Norm.
2. 3-stage Optimization: We firstly propose an efficient 2-stage process to train an intermediate meta-network



(a) **Independent:** Each bit-width requires training a new network with independent weights. (b) **Adabits:** A single network can be quantized to any of n bit-widths at runtime. All layers inside the network share the *same* bit-width. (c) **Proposed method (Bit-Mixer):** A single network whose individual layers can be quantized at runtime to any bit-width, without any re-training, resulting in an exponential number of mixed precision networks that one can choose from to fit the device characteristics and computational resources available on-the-fly.

Figure 1: Comparison between prior network quantization paradigms (a,b) and ours (c).

which at runtime can select different bit-widths which however are shared across the entire network. Then, a 3-rd final stage is introduced to gradually transition from the intermediate meta-network to the final one where the quantization level can be randomly selected at a block or layer level. Notably, our meta-network uses a single, shared set of weights.

3. We conducted a number of ablation studies which shed light into the behaviour of several components of our method. Moreover, building on top of the findings of [9], we analyze Bit-Mixer’s sub-nets exploring the inter-dependencies between the accuracy and the quantization level selected for a given layer. Finally, we extensively evaluated the accuracy of the proposed Bit-Mixer across different architectures and model sizes.

2. Related work

Network quantization aims to alleviate the high computational and memory cost of modern deep neural networks by using fewer bits (*i.e.* $b < 32$) for the weights and activations. Most of early works quantized the weights only [13, 7]. Follow-up works quantize both the weights and the activations while maintaining the same bit-width across the entire network using uniform quantization schemes [17, 18, 43, 31, 35, 2, 44, 46, 11].

More recently, a growing body of work explores mixed-precision quantization which enables, within the same ar-

chitecture, different layers to use different bit-widths [10, 37, 40]. The bit-width allocation process is typically performed either using reinforcement learning techniques [10, 40] or differential search [41, 37]. Contrary to Bit-Mixer (our work), all the aforementioned methods result in a single network with different but pre-defined bit-widths per layer that cannot be modified without retraining.

Related to our work is the line of research somewhat related to Neural Architecture Search (e.g. [5, 36, 26]), and, in particular, the works of [42, 3] where the authors train a super-network from which sub-nets with varying depth, width and kernel size can be sampled without retraining. These works do not consider the problem of network quantization at all.

More closely related to our work is AdaBits [19] where the authors propose to train a single neural network, with a shared set of weights, that can switch bit-width at runtime. However, a major limitation of AdaBits is that it is not a mixed-precision network: it uses the same bit-width across the entire network which reduces its flexibility in practical scenarios. Moreover, from a methodological perspective, the Transitional Batch-Norms as well as the 3-stage optimization procedure proposed in our work are fundamentally different from the methods described in [19].

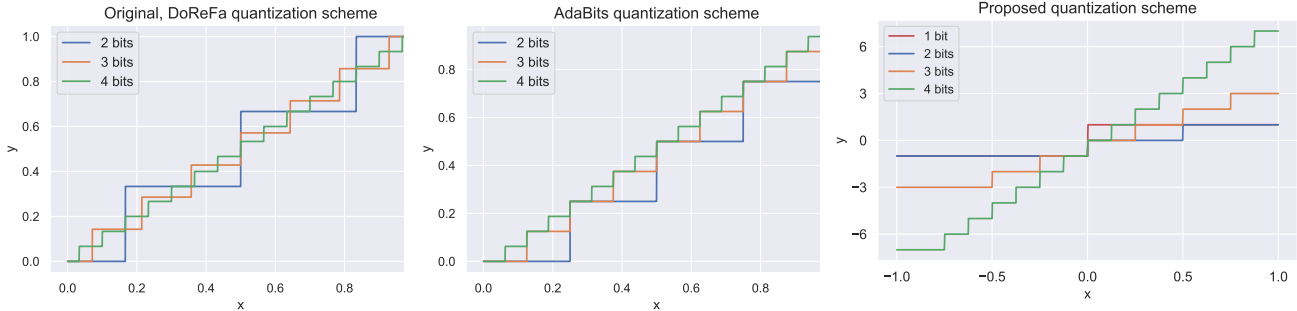


Figure 2: Difference between various quantization schemes (from left to right) used in DoReFa [44], AdaBits [19] and Bit-Mixer (Ours). In all cases $y = \text{quant}(x)$

3. Method

3.1. Unifying 1– n bit quantization

For a given layer l , we denote the quantization of the weights \mathbf{W} and input activations \mathbf{A} as $\text{quant}(\mathbf{W}, b) = \tilde{\mathbf{W}}_b$ and $\text{quant}(\mathbf{A}, b) = \tilde{\mathbf{A}}_b$, respectively, where $b = \{1, 2, \dots, n\}$ denotes the bit-width.

For the quantization function, we opted to adopt and adapt the recently proposed LSQ [11] as follows: to handle both cases $b = 1$ (i.e. binary networks) and $1 < b \leq n$, we quantize both the activations and the weights between $(-m_b, m_b)$, where $m_b = 2^{b-1} - 1$ is the maximum value representable using b bits¹. Although this symmetric quantization discards 1 state, we will show that this has no impact on the accuracy of the quantized networks. Furthermore the case $b = 2$ will degenerate in what is known in literature as ternary quantization, allowing for further specific optimizations made possible by the induced sparsification [45]. Overall, our unified quantization scheme is defined as follows:

$$\begin{aligned} \tilde{\mathbf{W}}_b &= \mathfrak{q}_b(W) \\ \tilde{\mathbf{A}}_b &= \mathfrak{q}_b(A), \end{aligned} \quad (1)$$

where the quantization function $\mathfrak{q}_b(x)$ is computed as:

$$\begin{aligned} \mathfrak{q}_b(x) &= \alpha \times \mathfrak{q}'(\text{clip}(\frac{x}{\alpha}, -m_b, m_b)) \\ \mathfrak{q}'(x) &= \begin{cases} \lfloor \cdot \rfloor, & \text{if } b > 1 \\ \text{sign}, & \text{if } b = 1 \end{cases}, \end{aligned} \quad (2)$$

where $\lfloor \cdot \rfloor$ is the floor rounding operator. Notice that we replaced the round function used in LSQ [11] with floor. This allows us to obtain the weights $\tilde{\mathbf{W}}_i$ directly from $\tilde{\mathbf{W}}_{i+1}$ without the need of storing the full precision

¹This is because in binary networks both the weights and the activations are quantized using the sign function [35], hence a symmetric quantizer is needed.

weights, significantly reducing the model storage requirements (as its size is determined solely by the size of $\tilde{\mathbf{W}}_n$). The difference between various quantization schemes used for mixed precision networks is shown in Fig. 2.

3.2. Transitional Batch-Norm

Quantizing the individual layers and blocks to different bit-widths will result in features that follow different distributions. This is because of two reasons: Firstly, it is a consequence of the inherent change in the representational power due to the change of precision. Secondly, as the number of bits drops, the network is unable to approximate closely the feature distribution of higher bit-widths, as the weight distribution significantly changes (this can also be seen in Fig. 6 for $b = \{1, 2, 3, 4\}$).

To properly compensate for the distribution shift that arises when a change in the bit-width occurs between two consecutive layers, for each transition between different bit-widths, we propose to learn a separate batch normalization layer, coined Transitional Batch-Norm. Specifically, if $1 \leq i \leq n$ is the bit-width of layer $l - 1$ and $1 \leq j \leq n$ is the bit-width of layer l , we learn BN parameters α_{ij} and β_{ij} . The parameters α_{ij} and β_{ij} remain tied to the bit-width j of the layer l since they depend on the current quantization level alone, irrespectively of the layer’s weights, which do not undergo a transition as opposed to the activations. We note that introducing the Transitional Batch-Norm layers does not induce any increase in the complexity of the network; only a small increase in network size is introduced (less than 1% of the total parameters count). Importantly, we emphasize that, without the Transitional Batch-Norms, the network is unable to converge to a satisfactory level of accuracy. This phenomenon is present both when training from scratch and when initializing from a pretrained model (see also Table 1).

3.3. Optimization process

A key remaining aspect of our method is how to train the proposed meta-network which turns out to be very challenging for several reasons. A direct naive approach, where all the paths are active simultaneously, is unfeasible due to both memory and computational constraints. Besides this, we considered an approximation to this training where all active paths are considered between 2 adjacent layer. Even in this case, we found the models unstable to train due to the internal competition arising, especially early in the training.

A more computationally feasible approach is to select randomly (with equal probability) during training an active sub-path or a set of active sub-paths. However, in our experiments, we found that this leads to networks in which the accuracy of all bit-widths are closely tight together, pulling them towards the one with the lowest accuracy, and, hence, largely diminishing the potential advantages of training the proposed meta-network.

In order to successfully train the newly introduced quantized meta-network, we firstly propose an *efficient* way to train a meta-network which can work at runtime with different bit-widths which however shared across *the entire* network (Stages I & II below). Then, to obtain the final meta-network, we propose to *progressively* train the previous network by gradually transitioning from networks where all the layers are quantized to the same bit-width to ones where the quantization level is randomly selected *at a block or layer* level (Stage III below). The procedure can be summarized as follows:

Stage I: During this stage, the network weights are kept real-valued while the activations (*i.e.* features) are quantized to n different bit-widths. Specifically, at each iteration, we randomly select, with equal probability, a bit-width b out of the predefined set $\{1, \dots, n\}$. At this stage, the model will use the *same bit-width* for the activations across all layers of the network.

Stage II: During this stage, we use the network trained in Stage I as initialization and repeat the process of the previous stage, with the difference being that this time both the weights and the activations are quantized. Again, the model will use the *same bit-width* for both weights and activations across all layers of the network. Note that Adabits [19] trains a network similar to the one obtained at the end of this stage. Compared to Adabits which requires n training stages, our scheme is more efficient requiring only 2 stages independently of n .

Stage III: Continuing the training process by resuming from the previous checkpoint, during this stage, and with probability σ , the weights and features are trained in the same fashion as described in Stage II (*i.e.* the same bit-width is used across all layers). For the rest of time, *i.e.* with probability $1 - \sigma$, the bit-width b of each individual layer is randomly selected *independently* of each other, re-

sulting in a network where different bit-widths are used for different layers. As the training progresses, we gradually decrease σ , effectively increasing the chance of training the meta-network with layer-wise random bit-width allocation. We continue the process until $1 - \sigma = k$, where k is typically $\frac{3}{4}$. All 3 stages share the same training scheduler.

4. Ablation studies

Unless otherwise stated, we conduct our ablation studies by using Bit-Mixer to train a meta-ResNet-18 [14] on ImageNet. We mainly report the accuracy of Bit-Mixer for the following 2 cases (note there is only one single network that is evaluated): fixed bit-width selection across all layers and random bit-width selection for each individual layer. For the latter case, we simply randomize the layer-wise bit-width selection for every iteration (forward pass) of the validation set. We note that this random layer-wise bit-width selection has been *intentionally* chosen for Bit-Mixer’s evaluation protocol since it conclusively shows that Bit-Mixer works as expected. However, in Section 4.2, we *do provide* accuracy results for the case where a simple method, based on [9], has been used in order to discover high performing sub-nets within the trained meta-network.

4.1. Effect of Transitional Batch-Norm

In Section 3.2, we introduced the Transitional Batch-Norm layers to compensate for the distribution shift between adjacent layers that are quantized to different bit-widths. Herein, we show their importance in terms of effectively training Bit-Mixer. As the results from Table 1 show, without Transitional Batch-Norm, the meta-network is unable to converge to a good solution although it was initialized using a model trained up to Stage II. The effect can be also observed by analyzing the statistics of the features before and after applying the transitional batch norm layer in Fig. 3.

Table 1: Top-1 accuracy (%) on ImageNet for Bit-Mixer trained with and without Transitional BN.

Bit-Mixer	Bit-width			
	4	3	2	Rand.
w/o Transitional BN	8.2	5.6	10.2	8.8
with Transitional BN	69.2	68.6	64.4	65.8

4.2. Analyzing Bit-Mixer’s sub-nets

Bit-Mixer’s trained meta-network contains an exponential number of sub-nets. By changing the bit-width of its individual layers (at runtime, and without extra training), a device or an application where the network is deployed can benefit from a finer trade-off between accuracy and speed.

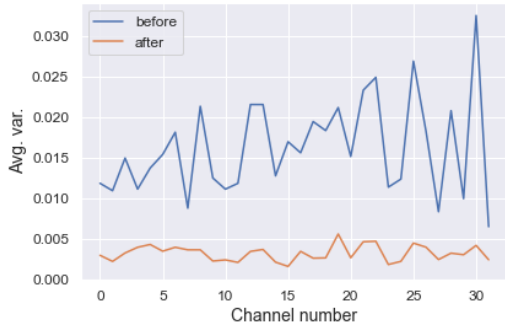


Figure 3: Variance per channel over the quantized activations $\tilde{\mathbf{A}}_b$, $b = 2, 3, 4$ before and after applying the Transitional Batch-Norm. Notice that the layer helps reducing the variance of the quantized activations significantly, stabilizing the training of the Bit-Mixer meta-network.

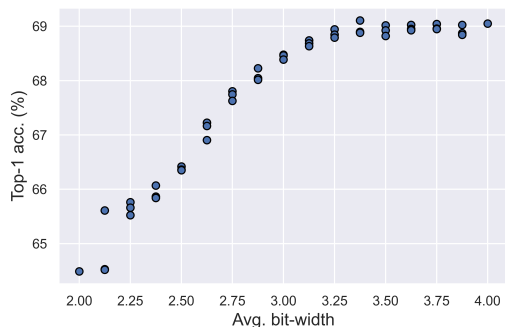


Figure 4: Top-1 accuracy (%) on ImageNet for a set of sub-nets extracted from Bit-Mixer’s meta-ResNet-18. Note that the accuracy smoothly varies as the avg. bit width changes.

Herein, we describe a method for “extracting” highly performing sub-nets from the meta network given a specific avg. bit-width budget. We note that no training is required for finding these sub-nets.

To facilitate the selection of interesting (as measured in terms of accuracy per avg. bit-width) candidates out of the given population, we followed [9], and for each layer, we computed the top eigenvalue of the Hessian². Note that, for this purpose, we used the network with the highest possible bit-width (i.e. constant bit-width equal to 4 for all layers). The top eigenvalues computed per each layer for the network can be seen in Fig. 5. In general, smaller eigenvalues correspond to flatter loss surfaces, which in turn, suggest that such layers are good candidates for more aggressive quantization given that the induced errors are less likely to be amplified [16].

²Since forming the entire Hessian matrix is computationally and memory prohibitive, we made use of the power iteration algorithm [29].

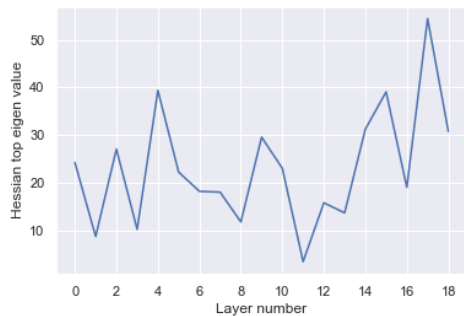


Figure 5: Top eigenvalues of the Hessian matrix for each layer of the network with const. bit-width equal to 4. Notice that layers located towards the end of the network are generally more sensitive to noise.

Given a meta-network Φ , trained as described in Section 3.3, and a target average bit-width $b_{avg} = \frac{1}{N} \sum_i^N b_i$, where N is the number of layers and b_i the selected bit-width of the i -th layer, we attempt to identify a set of promising sub-nets $\{\Phi_0, \dots, \Phi_m\}$ formed by changing the per-layer bit-width as follows: Let $C_{bits} = \lceil N \times b_{avg} \rceil$ be the total bit-cost of the desired sub-net, and $\mathbf{v}_C \in \mathbb{R}^N$ a per-layer defined cost-vector constructed by taking the highest eigenvalue of the Hessian matrix of each layer. Note that, depending on the target scenario, the cost could be adjusted to take into consideration device-specific knowledge. Since the set of sub-nets $\{\Phi_0, \dots, \Phi_m\}_C$ of cost C is finite, a straightforward approach is to use a greedy approach generating all potential candidates of bit-cost C_{bits} . Once generated, for each configuration from the set, we compute its final ranking cost by taking the product $C_{total} = [b_0 b_1 \dots b_N] \times \mathbf{v}_C^T$. We can then select the top- k candidates and evaluate their accuracy. Fig. 5 shows a few candidates for various avg. bit-widths alongside their corresponding accuracy. It can be observed that even by using a simple method like the one described above, a diverse set, in terms of accuracy, of networks can be obtained covering the whole spectrum of avg. bit-widths.

4.3. Effect of knowledge distillation

Knowledge distillation has been previously shown to improve the performance of both full precision [15] and quantized neural networks [32, 30]. Herein, we analyze and validate to what extent distillation helps improve the training of Bit-Mixer for both Stages II and III. In particular, we explore two scenarios for the teacher: (1) Using a full precision network (FP32), and (2) using the trained network after Stage I. We note that, in all cases, the student and teacher networks have exactly the same architecture. As the results from Table 2 show, distillation does indeed improve the accuracy, although the improvements are lower than typically

observed for independently quantized or full precision models.

Table 2: Top-1 accuracy (%) on ImageNet for Bit-Mixer trained with and without distillation.

Method	Teacher	Bit-width			
		4	3	2	Rand.
Ours–Stage II	-	69.1	68.5	65.1	-
	Stage I	69.4	68.7	65.6	-
	FP32	69.3	68.7	65.5	-
Ours–Stage III	-	69.0	68.4	64.0	65.5
	Stage I	69.1	68.6	64.5	65.7
	FP32	69.2	68.6	64.4	65.8

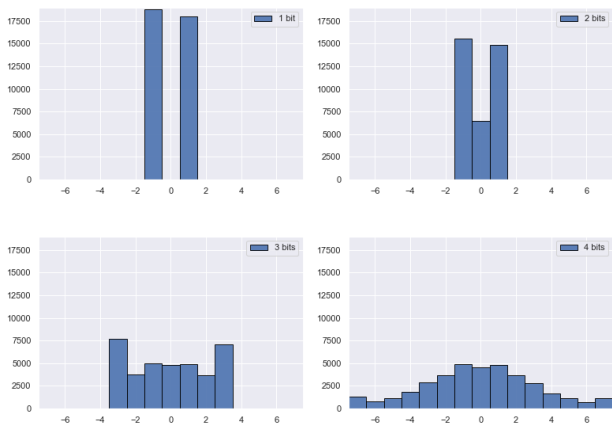


Figure 6: Weights distribution for 1, 2, 3 and 4 bits after quantization. Note the significant difference in distribution for the 1 bit quantization.

4.4. 1–4 bit quantization

In Section 3, we introduced a unified quantization scheme that can be used across all bit-widths, including 1 bit quantization (*i.e.* binarization). As our results in Table 3 suggest, using a single, shared, set of weights, we can successfully train the meta-network up to Stage II using all 4 bit-widths (*i.e.* 4,3,2 and 1) with minimal accuracy loss³. This is performed in order to align the 1 bit quantization to the rest ones. The training scheduler used for the 4-3-2-1 quantization is the same as the one used for 4-3-2. Notice that despite the relatively larger accuracy gap between binarization and 4 bit quantization, also observed from the noticeably different weight distribution as shown in Fig. 6, the trained model after Stage II offers overall a good accuracy.

³We note, that unlike the current paradigm used in most recent works on network binarization that is to maintain the 1×1 downsampling layers

Table 3: Top-1 accuracy (%) on ImageNet using a ResNet-18 for 4-3-2-1 bits quantization.

Method	Bit-width			
	4	3	2	1
Independent	69.1	68.5	65.1	59.0
Adabits [19]	69.2	68.5	65.1	-
Ours (Stage II)	69.4	68.7	65.6	-
Ours (Stage II)	68.7	68.0	64.2	57.3

Following Stage II, we continued the training of the above model to obtain the final, Stage III, 4-3-2-1 Bit-Mixer model. However, during this last stage, the training did not converge to the desired outcome. We believe that the main reason for this is the lack of the zeroth state for the binary case. Specifically, while the 2-4 bit-width quantization share the lower states between themselves, as shown in Fig. 1c, the same is not true for 1 bit quantization, which lacks the zeroth state, introducing high quantization errors around it and resulting in a different distribution (see Fig. 6).

However, we did manage to train successfully a Bit-Mixer model with the following configuration $b_{act} = \{2, 3, 4\}$ for the activations and $b_w = \{1, 3, 4\}$ for the weights, respectively. The results are shown in Table 4. Instead of using 2 bits for the activations and weights, in this case, we binarize the later. Since our 2 bit representation is, in fact, a ternary one, this ternary-binary quantization allows for efficient bit-wise implementation too which can result in at least $40 \times$ [39] faster convolutions.

Table 4: Top-1 accuracy (%) on ImageNet using a ResNet-18 for 4-3-1.5 bits quantization. * -denotes binary-ternary quantization

Method	Bit-width			
	4	3	1.5*	rand
Adabits [19]	69.2	68.5	-	-
Ours - Stage II	69.0	68.7	64.0	-
Ours - Stage III	69.0	68.5	62.1	62.9

4.5. Scale- vs. clip-based mixed quantization

Throughout this work, we quantize our models using Eq. 1 and 2. Fig. 7 shows how the learnable quantization scaling factors α in Eq. 2 change their value as we advance through the network. Importantly, the ratio between α_i and α_j is approximately equal to that of m_i/m_j suggesting that all bit-widths are roughly scaled to occupy the whole range.

to full precision [30], in our experiments we binarize them too.

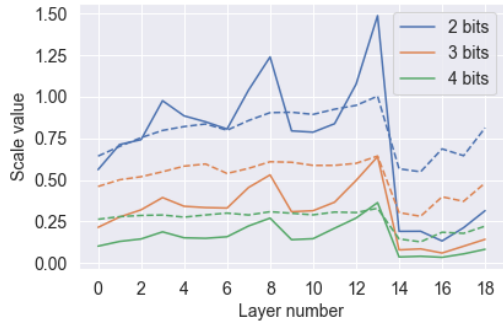


Figure 7: Quantization scales for the activations and weights (dashed line) of each layer of a ResNet-18 model quantized to $b = \{2, 3, 4\}$. Notice that the ratio between each scale is approximately equal with that of their corresponding maximum representable values.

To emphasise the importance of the per bit-width scaling factors, we also tested a slightly different approach. The idea is to have shareable quantized weights (and activations) which are obtained by firstly quantizing the real-valued weights to the maximum bit-width n (*i.e.* $-m_n$ and m_n) and then clipping them to fit the required bit-width (*i.e.* $-m_i$ and m_i). For the weights, this idea is described by:

$$\tilde{W}_n = \alpha \cdot \text{clip}_n\left(\frac{W}{\alpha}, -m_n, m_n\right) \quad (3)$$

$$\tilde{W}_i = \alpha \times \text{clip}(\tilde{W}_{i+1}, -m_i, m_i), \quad (4)$$

Intuitively, this has the effect that the largest (in magnitude) values of the real-valued weights are mapped to states/bits which are present only in the higher bit-widths. On the contrary, using per bit-width scaling factors, the whole range of real-valued weights is mapped to the whole range corresponding to a specific bit-width (*i.e.* $-m_i$ and m_i). When trained with this type of quantization, we found that the obtained networks performed 5–10% worse than with the scale-based quantization.

4.6. Symmetric vs Asymmetric quantization

We firstly, note that, in this work, symmetric quantization refers to the case where the data are mapped to integers in the range $\{-2^{b-1} - 1, \dots, 2^{b-1} - 1\}$ while asymmetric to the case where the data are mapped into $\{-2^{b-1}, \dots, 2^{b-1} - 1\}$. In both cases, we consider 0 itself as the zero point as this allows for more efficient implementations. To ensure that no accuracy loss occurs due to the aforementioned design choices, we trained three models: One with asymmetric quantization, one with symmetric, and finally one with symmetric quantization but using the `round` function in Eq. 2 instead of `floor`. As the results from Table 5 show, all 3 variants are producing essentially identical results.

Table 5: Top-1 accuracy (%) on ImageNet using a standard ResNet-18 quantized to 2, 3 and 4 bits using 3 different quantization schemes.

Quantization Method	Bit-width		
	4	3	2
symmetric	69.1	68.5	65.1
asymmetric	69.2	68.5	65.2
asymmetric with round	69.2	68.6	65.2

Method	32	4	3.5	3	2.5	2
DoReFa [44]	70.4	68.1	-	67.5	-	62.6
LQ-Net [43]	70.3	69.3	-	68.2	-	64.9
PACT [6]	70.4	69.2	-	68.1	-	64.4
QIL [20]	70.2	70.1	-	69.2	-	65.7
DSQ [12]	69.9	69.6	-	68.7	-	65.2
APoT [24]	70.2	-	-	69.9	-	-
EdMIPS [4]*	-	68	67.7	67	66.4	65.9
Adabits [19]	-	69.2	-	68.5	-	65.1
Ours	69.6	69.1	69.2	68.6	66.4	64.4

Table 6: Comparison against the state-of-the-art in fixed-bit and mixed precision quantization in terms of top-1 accuracy (%) on ImageNet using a ResNet-18 architecture. * refers to results where either the number of bits or the accuracy is approximately the one stated in the table.

5. Results

5.1. Experimental setup

All our experiments are performed on ImageNet [8]. We focus on the 2-4 bits quantization range. For $b > 4$, the accuracy almost always matches or gets very close to that of the full precision counter-parts. Following previous work (e.g. [19, 35, 4, 44]) the batch normalization layers are not quantized.

Network architectures: In order to cover a broad spectrum of architectures in terms of depth, width and cardinality (*i.e.* via grouped convolutions) we performed experiments using the following architectures: (a) ResNet [14] (18, 34, 50) and, (b) the recently proposed EBN of [1]. We chose the later since it was shown to be efficient, suitable for quantization and flexible in terms of varying the width and the group size of the model easily. For example, by increasing the group size, more efficient variants can be obtained. Note that we did not use expert convolutions as proposed in [1]. We used an EBN which, similarly to a Resnet-18, has 4 stages and 2 convolutional blocks per stage. The width of each stage is double the one used in Resnet-18. Finally, the group size per stage is denoted by G0:G1:G2:G3. We tried

Table 7: Top-1 accuracy (%) on ImageNet obtained by applying Bit-Mixer on several ResNet and EBN architectures. The accuracy of AdaBits is directly comparable with Ours-Stage II. Notice that Bit-Mixer (Ours) is the only method that can produce a result for layer-wise *rand.* bit allocation. Note that, as shown in Section 4.2, certain sampled sub-nets from our Bit-Mixer meta-network are significantly more accurate than *rand.* * - denotes result taking directly from [19].

Arch.	#bits	Method			
		Indep.	AdaBits [19]	Ours (Stage II)	Ours
ResNet-18	4	69.1	69.2	69.4	69.2
	3	68.5	68.5	68.7	68.6
	2	65.1	65.1	65.6	64.4
	Rand.	-	-	-	65.8
ResNet-34	4	73.1	73.0	73.0	72.9
	3	72.6	72.5	72.6	72.5
	2	70.2	70.0	70.1	69.6
	Rand.	-	-	-	70.5
ResNet-50	4	75.5	76.3*	75.2	75.2
	3	75.3	75.9*	74.9	74.8
	2	72.8	73.3*	72.7	72.1
	Rand.	-	-	-	73.2
EBN 4:8:8:16	4	74.0	-	74.0	73.9
	3	73.5	-	73.4	73.3
	2	70.7	-	70.5	70.4
	Rand.	-	-	-	71.8
EBN 4:8:16:32	4	73.8	-	73.8	73.3
	3	73.3	-	73.2	72.8
	2	69.8	-	69.7	68.9
	Rand.	-	-	-	70.0
EBN 4:4:4:4	4	74.7	-	74.6	74.7
	3	74.2	-	74.2	74.2
	2	71.5	-	71.1	71.4
	Rand.	-	-	-	72.1

3 EBN variants in total: 4:8:8:16, 4:8:16:32 and 4:4:4:4.

Training details: Unless otherwise stated, all models are trained following the same recipe: the networks are trained for 160 epochs using a cosine scheduler with warm-up (10 epochs) and no restarts [28] with a starting learning rate of 0.001 and a weight decay of $1e-4$. We used the Adam optimizer [21]. For augmentation, we follow the standard set of transformations used for ImageNet in prior works, mainly: random crop, resize to 224×224 px and random flipping. For stage III, we gradually increase the probability of $1 - \sigma$ from 0 to a target value k during early training, until the network configurations stabilize. For the rest of the training (typically after epoch 80), k remains fixed. The value of k is determined based on the network architecture (typically

$2/3 < k < 4/5$). During evaluation, we resize the images to 256×256 px and then center crop them to the same 224×224 px resolution. All experiments are implemented using PyTorch [34].

5.2. Comparison with state-of-the-art

Herein, we firstly compare our method against the current state-of-the-art in quantization. We note that it is hard to make a direct comparison between Bit-Mixer and other methods, as our method: (1) is the very first of its kind that offers the flexibility of layer-wise bit-width selection during runtime, (2) is not focusing on maximizing accuracy for a specific bit-width like other fixed-bit quantization methods, nor (3) is focusing on finding the optimal bit-width allocation for maximizing accuracy like other mixed-precision methods. Moreover, (4) the accuracy results reported in other papers depend on other factors, for example, a very important one is the accuracy of the original FP32 model used. Hence, the main aim of these comparisons is to rather illustrate that the networks trained with Bit-Mixer offer accuracy in par with recently proposed state-of-the-art quantization methods.

To this end, in Table 6, we report our results in comparison with a variety of recently proposed state-of-the-art methods for fixed-bit and mixed precision quantization [44, 43, 6, 20, 12, 24, 19, 4]. In all cases, the ResNet-18 architecture was used. As it can be observed, Bit-Mixer provides very competitive results by just training a *single* meta-network which can dynamically define the per-layer bit-width at runtime. This is very important as our goal is to have the flexibility that Bit-Mixer can offer without however compromising the capacity for highly accurate inference.

This section, and, in particular, Table 7, also provides results obtained by training Bit-Mixer meta-networks using the ResNet and EBN architectures detailed in Section 5.1. Where possible, we also compare with Adabits [19]. Note that Bit-Mixer after Stage II (Ours-Stage II) is directly comparable with Adabits. Note also that Bit-Mixer after Stage III (Ours) is the only method that can provide layer-wise *random* bit allocation. We believe that the results of Table 7 conclusively show that Bit-Mixer can be successfully applied to train meta-networks across a wide variety of network architectures.

6. Conclusions

To the best of our knowledge, this work constitutes the very first attempt to training a meta-network with shared weights the layers/blocks of which can be independently quantized to any desired bit-width at runtime. To this end, we made two key contributions: (a) Transitional Batch-Norms and (b) a 3-stage optimization pipeline which is shown capable of training such a network. We presented a series of ablation studies analyzing important components

and features of the proposed method. Moreover we presented comparisons with several state-of-the-art quantization methods as well as results obtained by applying Bit-Mixer on several architectures. These results show that our method can successfully train a meta-network with arbitrary layer-wise bit-width selection without compromising accuracy.

References

- [1] Adrian Bulat, Brais Martinez, and Georgios Tzimiropoulos. High-capacity expert binary networks. *arXiv preprint arXiv:2010.03558*, 2020. 7
- [2] Adrian Bulat and Georgios Tzimiropoulos. Xnor-net++: Improved binary neural networks. *arXiv preprint arXiv:1909.13863*, 2019. 2
- [3] Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, and Song Han. Once-for-all: Train one network and specialize it for efficient deployment. *arXiv preprint arXiv:1908.09791*, 2019. 2
- [4] Zhaowei Cai and Nuno Vasconcelos. Rethinking differentiable search for mixed-precision neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2349–2358, 2020. 7, 8
- [5] Xin Chen, Lingxi Xie, Jun Wu, and Qi Tian. Progressive differentiable architecture search: Bridging the depth gap between search and evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1294–1303, 2019. 1, 2
- [6] Jungwook Choi, Zhuo Wang, Swagath Venkataramani, Pierce I-Jen Chuang, Vijayalakshmi Srinivasan, and Kailash Gopalakrishnan. Pact: Parameterized clipping activation for quantized neural networks. *arXiv preprint arXiv:1805.06085*, 2018. 7, 8
- [7] Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. Binaryconnect: Training deep neural networks with binary weights during propagations. *arXiv preprint arXiv:1511.00363*, 2015. 2
- [8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 7
- [9] Zhen Dong, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Hawq: Hessian aware quantization of neural networks with mixed-precision. In *ICCV*, 2019. 2, 4, 5
- [10] Ahmed Elthakeb, Prannoy Pilligundla, FatemehSadat Mireshghallah, Amir Yazdanbakhsh, Sicuan Gao, and Hadi Esmaeilzadeh. Releq: An automatic reinforcement learning approach for deep quantization of neural networks. In *NeurIPS ML for Systems workshop, 2018*, 2019. 2
- [11] Steven K Esser, Jeffrey L McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S Modha. Learned step size quantization. In *ICLR*, 2020. 2, 3
- [12] Ruihao Gong, Xianglong Liu, Shenghu Jiang, Tianxiang Li, Peng Hu, Jiazhen Lin, Fengwei Yu, and Junjie Yan. Differentiable soft quantization: Bridging full-precision and low-bit neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4852–4861, 2019. 7, 8
- [13] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015. 2
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 7
- [15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 5
- [16] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997. 5
- [17] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *CVPR*, 2018. 2
- [18] Qing Jin, Linjie Yang, and Zhenyu Liao. Towards efficient training for neural network quantization. *arXiv preprint arXiv:1912.10207*, 2019. 2
- [19] Qing Jin, Linjie Yang, and Zhenyu Liao. Adabits: Neural network quantization with adaptive bit-widths. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2146–2156, 2020. 2, 3, 4, 6, 7, 8
- [20] Sangil Jung, Changyong Son, Seohyung Lee, Jinwoo Son, Jae-Joon Han, Youngjun Kwak, Sung Ju Hwang, and Changkyu Choi. Learning to quantize deep networks by optimizing quantization intervals with task loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4350–4359, 2019. 7, 8
- [21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 8
- [22] Vadim Lebedev, Yaroslav Ganin, Maksim Rakhuba, Ivan Osleledets, and Victor Lempitsky. Speeding-up convolutional neural networks using fine-tuned cp-decomposition. *arXiv preprint arXiv:1412.6553*, 2014. 1
- [23] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. *arXiv preprint arXiv:1608.08710*, 2016. 1
- [24] Yuhang Li, Xin Dong, and Wei Wang. Additive powers-of-two quantization: An efficient non-uniform discretization for neural networks. *ICLR*, 2020. 7, 8
- [25] Yawei Li, Shuhang Gu, Luc Van Gool, and Radu Timofte. Learning filter basis for convolutional neural network compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5623–5632, 2019. 1
- [26] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018. 1, 2
- [27] Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. *arXiv preprint arXiv:1810.05270*, 2018. 1

- [28] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 8
- [29] James Martens. Deep learning via hessian-free optimization. In *ICML*, volume 27, pages 735–742, 2010. 5
- [30] Brais Martinez, Jing Yang, Adrian Bulat, and Georgios Tzimiropoulos. Training binary neural networks with real-to-binary convolutions. *arXiv preprint arXiv:2003.11535*, 2020. 5, 6
- [31] Naveen Mellempudi, Abhisek Kundu, Dheevatsa Mudigere, Dipankar Das, Bharat Kaul, and Pradeep Dubey. Ternary neural networks with fine-grained quantization. *arXiv preprint arXiv:1705.01462*, 2017. 2
- [32] Asit Mishra and Debbie Marr. Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy. *arXiv preprint arXiv:1711.05852*, 2017. 5
- [33] Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11264–11272, 2019. 1
- [34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019. 8
- [35] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European conference on computer vision*, pages 525–542. Springer, 2016. 2, 3, 7
- [36] Dimitrios Stamoulis, Ruizhou Ding, Di Wang, Dimitrios Lymberopoulos, Bodhi Priyantha, Jie Liu, and Diana Marculescu. Single-path nas: Designing hardware-efficient convnets in less than 4 hours. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 481–497. Springer, 2019. 2
- [37] Stefan Uhlich, Lukas Mauch, Kazuki Yoshiyama, Fabien Cardinaux, Javier Alonso Garcia, Stephen Tiedemann, Thomas Kemp, and Akira Nakamura. Differentiable quantization of deep neural networks. *arXiv preprint arXiv:1905.11452*, 2(8), 2019. 2
- [38] Karen Ullrich, Edward Meeds, and Max Welling. Soft weight-sharing for neural network compression. *arXiv preprint arXiv:1702.04008*, 2017. 1
- [39] Diwen Wan, Fumin Shen, Li Liu, Fan Zhu, Jie Qin, Ling Shao, and Heng Tao Shen. Tbn: Convolutional neural network with ternary inputs and binary weights. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 315–332, 2018. 6
- [40] Kuan Wang, Zhijian Liu, Yujun Lin, Ji Lin, and Song Han. Haq: Hardware-aware automated quantization with mixed precision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8612–8620, 2019. 2
- [41] Bichen Wu, Yanghan Wang, Peizhao Zhang, Yuandong Tian, Peter Vajda, and Kurt Keutzer. Mixed precision quantization of convnets via differentiable neural architecture search. *arXiv preprint arXiv:1812.00090*, 2018. 2
- [42] Jiahui Yu, Linjie Yang, Ning Xu, Jianchao Yang, and Thomas Huang. Slimmable neural networks. *arXiv preprint arXiv:1812.08928*, 2018. 2
- [43] Dongqing Zhang, Jiaolong Yang, Dongqiangzi Ye, and Gang Hua. Lq-nets: Learned quantization for highly accurate and compact deep neural networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 365–382, 2018. 2, 7, 8
- [44] Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, and Yuheng Zou. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *arXiv preprint arXiv:1606.06160*, 2016. 2, 3, 7, 8
- [45] Chenzhuo Zhu, Song Han, Huizi Mao, and William J Dally. Trained ternary quantization. *arXiv preprint arXiv:1612.01064*, 2016. 3
- [46] Bohan Zhuang, Lingqiao Liu, Mingkui Tan, Chunhua Shen, and Ian Reid. Training quantized neural networks with a full-precision auxiliary module. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1488–1497, 2020. 2