Training Quantized Neural Networks with a Full-precision Auxiliary Module

Bohan Zhuang\textsuperscript{1}  Lingqiao Liu\textsuperscript{1}  Mingkui Tan\textsuperscript{2}  Chunhua Shen\textsuperscript{1,*}  Ian Reid\textsuperscript{1}

\textsuperscript{1}The University of Adelaide  \textsuperscript{2}South China University of Technology

Abstract

In this paper, we seek to tackle a challenge in training low-precision networks: the notorious difficulty in propagating gradient through a low-precision network due to the non-differentiable quantization function. We propose a solution by training the low-precision network with a full-precision auxiliary module. Specifically, during training, we construct a mix-precision network by augmenting the original low-precision network with the full precision auxiliary module. Then the augmented mix-precision network and the low-precision network are jointly optimized. This strategy creates additional full-precision routes to update the parameters of the low-precision model, thus making the gradient back-propagates more easily. At the inference time, we discard the auxiliary module without introducing any computational complexity to the low-precision network. We evaluate the proposed method on image classification and object detection over various quantization approaches and show consistent performance increase. In particular, we achieve near lossless performance to the full-precision model by using a 4-bit detector, which is of great practical value.

1. Introduction

Deep neural networks (DNNs) have made great strides in many computer vision tasks such as image classification [11, 20], segmentation [8, 10] and detection [37, 40]. Even though deep and/or wide models can achieve promising accuracy, their huge computational complexity makes them incompatible with energy constrained devices which usually have limited memory bandwidth and computational power. This has motivated the community to design energy-efficient models, often based on quantized precision, aiming not to sacrifice accuracy relative to the full-precision models. In this paper, we propose to improve the training of the low-precision networks.

The core challenge for quantization is the non-differentiability of the discrete quantizer. As a result, we cannot directly optimize the discretised network with stochastic gradient descent. And the current solutions can be divided into two categories. The first category is to employ a surrogate of gradient. The most commonly used approach is the straight-through estimator (STE) [2]. Some recent works have been proposed to relax the discrete quantizer to be continuous for gradient-based optimization [1, 31]. Even though the discontinuity of the discretization operation during training can be partly solved by smoothing it appropriately, some important information may still be missed due to the approximation which can lead to an undesirable drop in accuracy. The second category is to seeking guidance from a full-precision model for discretised network training. For example, new training strategies such as knowledge distillation [32, 56, 57] have been proposed to learn a low-precision student network by distilling knowledge from a full-precision teacher network.

Our method falls into the second category and our approach is based on the idea of sharing parameters between a mixed-precision (partially fully-precision) model and a low-precision model. Specifically, our method constructs a full-precision auxiliary module which connects to multiple layers of a low-precision model (see Fig. 1). During training, the low-precision network and the full-precision auxiliary module will be combined to form an augmented mixed-precision network. Then, the mixed-precision network and the low-precision model are jointly optimized. Since the parameters from the low-precision model are shared, they can receive gradient from both full-precision connections and low-precision connections. Consequently, the parameters of low-precision model can be updated via two routes, and this can overcome the gradient propagation difficulty due to the discontinuity of the quantizer. Note that only the low-precision network will be utilized for inference and therefore there is no additional complexity introduced at the test stage.

In addition to image classification, we further extend the proposed approach to building quantized networks for object detection. Building low-precision networks for object detection is more challenging since detection needs the network outputs richer information, such as locations of bounding boxes. There has been several works in literature to address the quantized object detectors [16, 22, 48]. How-
ever, there still exists a significant performance drop of 4-bit or lower-precision quantized detectors comparing to their full-precision counterpart. We apply our techniques to train a 4-bit RetinaNet [25] detector and further propose a modification to RetinaNet to better accommodate the quantization design. Through extensive experiments on the COCO benchmark, we show that our 4-bit models can achieve near lossless performance comparing to the full-precision model, which has a significant value in practice.

Our contributions can be summarized as follows:

- We propose a new training method to account for the non-differentiability of the quantization operator in a low-precision network. Our method can lead to more accurate low-precision model without increasing the model complexity at the testing stage.
- We apply our learning approach and propose a new design modification to build a 4-bit quantized object detection which achieves comparable performance to its full-precision counterpart.

1.1. Related work

Network quantization. Quantized network represents the weights and activations with very low precision, thus yielding highly compact DNN models compared to their floating-point counterparts. Moreover, the convolution operations can be efficiently computed via bitwise operations. Quantization can be categorized into fixed-point quantization and binary neural networks (BNNs), in which fixed-point quantization can also be divided into uniform and non-uniform. Uniform approaches [17, 54, 57] design quantizers with a constant quantization step. To reduce the quantization error, non-uniform strategies [4, 52] propose to learn the quantization intervals by jointly optimizing parameters and quantizers. A fundamental problem of quantization is to approximate gradient of the non-differentiable quantizer. To solve this problem, some works have studied relaxed quantization [1, 31, 47, 57]. Moreover, with the popularity of automatic machine learning, some recent literature employs reinforcement learning to search for the optimal bitwidth for each layer [6, 46, 49]. BNNs [14, 36] constrain both weights and activations to binary values (i.e., +1 or −1), which brings great benefits to specialized hardware devices. The development of BNNs can be classified into two categories: (i) a focus on improving the training of BNNs [13, 30, 36, 45]; (ii) multiple binarizations to approximate the full-precision tensor or structure [9, 23, 27, 28, 45, 58]. In this paper, we propose a general auxiliary learning approach that can work on all categories of quantization approaches.

Weight sharing. Weight sharing has been attracting increasing attention for efficient, yet accurate computation. In visual recognition, region proposal networks (RPN) in Faster-RCNN [40] and Mask-RCNN [10] share the same backbone with task-specific networks, which greatly saves testing time. For neural architecture search, ENAS [35] allows parameters to be shared among all architectures in the search space, which saves orders of magnitude GPU hours. In the network compression field, weight/activation quantization intends to partition the weight/activation distribution into clusters and use the centers of clusters as the possible discrete values. This strategy can be interpreted as a special case of weight sharing. Different from these approaches, we propose to utilize weight sharing for jointly optimizing the full-precision auxiliary module and the original low-precision network to improve the accuracy of the latter quantized model.

Auxiliary supervision. One straightforward way of adding auxiliary supervision is introducing additional losses into intermediate layers, which serves to combat the vanishing gradient problem while providing regularization. The effectiveness of additional losses has been demonstrated in some literature, like GoogLeNet [43], DSN [21], semantic segmentation [33, 53], etc. However, these methods are usually sensitive to the positions and scales of the guidance signals. Knowledge distillation (KD) is initially proposed for model compression, where a powerful wide/deep teacher distills knowledge to a narrow/shallow student to improve its performance [12, 41], which can also be treated as adding auxiliary supervisions. In terms of the definition of knowledge to be distilled from the teacher, existing models typically use teacher’s class probabilities [12] and/or intermediate features [15, 34, 41, 51, 57]. It is worth noting that our proposed auxiliary learning strategy uses weight sharing to assist optimization, where the motivation is very different from the KD methods. We does not need to pre-train a teacher network which is usually much deeper and may be the upper bound of the performance. Moreover, on network quantization, we show consistent superior performance over KD methods in Sec. 4.1.

Object detection. Object detection can be divided into two categories. As one of the dominant detection framework, two-stage detection methods [7, 8, 40] first generate region proposals and then refine them by subsequent networks. Another main category is the one-stage methods which are represented by YOLO [37–39], SSD [29] and RetinaNet [25]. The objective is to improve the detection efficiency by directly classifying and regressing the predefined anchors without the proposal generation step. The recent developing trends in object detection is designing light-weight frameworks for mobile applications [5, 44, 48], which usually requires real-time, low-power and fully embedded. In this paper, we explore to compress and accelerate detectors from the quantization perspective. Note that, we are the first to achieve near lossless 4-bit detectors in the literature.
2. Method

In this section, we describe the proposed learning strategy for training a low-precision network. We first give an overview of the proposed approach in Sec. 2.1. Then we introduce the auxiliary module design in Sec. 2.2 and optimization in Sec. 2.3, respectively. After that, we provide discussions in Sec. 2.4 and explain how to extend it to quantized object detection in Sec. 3. Through the following part, we adopt the following terminologies: A layer is a standard parameterized layer in a network such as a dense or convolutional layer. A block is a collection of layers in which the output of its last layer is connected to the input of the next block (e.g., a residual block).

2.1. Overview and motivation

The overview of the framework is shown in Fig. 1. The blue part, denoted as $F$, shows the low-precision network we aim to learn. The pink part, denoted as $H$, is a full-precision sub-network, which we call the auxiliary module. It connects to intermediate outputs from $F$. The input image is fed into $F$ but generates two outputs, one is from the last layer in $F$ and the other is from the last layer in $H$. In other words, the combination of $F$ and $H$ forms an augmented mix-precision network $F \circ H$ and the parameters of the low-precision network is shared by this augmented network. At the training time, two loss functions are applied to both outputs, and the mixed-precision network and low-precision network are trained jointly. After training, the auxiliary module $H$ will be discarded and only $F$ will be used at the test time.

The motivation of such a design is to create full-precision routes to update parameters of the low-precision model and thus alleviating the difficulty of propagating gradient in a quantized model. Specifically, the intermediate output of each block in $F$ can directly influence the output of the mixed-precision network $F \circ H$ through the full-precision connections in $H$. Consequently, the gradient from the loss of the second output will back propagate to the parameters of each block in the low-precision model.

2.2. Module design

We now elaborate the design for the auxiliary module $H$, which is made up of a sequential of adaptors and aggregators as shown in Fig. 1 (a). In particular, the auxiliary module $H$ receives $P$ output feature maps $\{O_p\}_{p=1}^P$ of the corresponding blocks in $F$. Let $\{B_1, ..., B_P\}$ be the block indexes where we generate the feature maps. For the $p$-th input of $H$, we adopt a trainable adaptor $\phi_p(\cdot)$, which receives the output feature map $O_p$ of the $B_p$-th block from $F$ and outputs an adapted feature representation $\phi_p(O_p)$. The motivation of using the adapter is to compensate the distribution discrepancy between the low-precision model and full-precision model. It ensures the quantized activations $\{O_p\}_{p=1}^P$ to be compatible to the full-precision calculation in $H$. We implement those adapters by a simple $1 \times 1$ convolutional layer followed by a batch normalization layer in this paper.

In the auxiliary module $H$, the outputs of the adapter are then sequentially aggregated. Formally, let $g_p$ denotes the $p$-th aggregated feature. It is achieved by adding the adapted feature $\phi_p(O_p)$ from $F$ and the $(p-1)$-th aggregated feature from $H$ followed by a ReLU(·) nonlinearity:

$$g_p = \text{ReLU}(\phi_p(O_p) + g_{p-1}).$$  \hspace{1cm} (1)

At the last layer in $H$, a classifier layer is applied to $g_P$ to make the class prediction. Then an auxiliary loss is employed. Note that the auxiliary module $H$ is akin the skip connections in ResNet [11].

Figure 1: An overview of the proposed framework. Blue color represents low-precision operations while pink color denotes full-precision operations. During training, the full-precision $H$ is connected to the quantized $F$ to form a mixed-precision network $F \circ H$. Then the mixed-precision network and the low-precision $F$ are jointly optimized with weight sharing. It is worth noting that only the learnt quantized network $F$ is used during testing.
Algorithm 1: Joint training approach w.r.t. the main low-precision network $F$ and the full-precision auxiliary module $H$.

**Input:** Current mini-batch $\{x_i, y_i\}$; parameter $\theta^F$ of the low-precision network $F$; parameter $\theta^H$ of the full-precision auxiliary module $H$.

**Output:** Updated parameters $\{\theta^F, \theta^H\}$.

1. Obtain the quantized weight $Q^F = q(\theta^F)$, where $q(\cdot)$ is the quantization function;
2. $y_F, y_H = \text{Forward}(x_i, Q^F, \theta^H)$;
3. Compute the loss $L(y_i, y_F)$ for the main network $F$;
4. Compute the loss $L_{aux}(y_i, y_H)$ for the auxiliary module $H$;
5. $\frac{\partial L_F}{\partial Q^F} - \frac{\partial L_{aux}}{\partial Q^F} = \frac{\partial L_F}{\partial \theta^F} = \text{Backward}(\frac{\partial L_H}{\partial y_H}, Q^F, \theta^H)$;
6. Compute the gradient, in particular $\nabla Q^F = \frac{1}{M} (\frac{\partial L_F}{\partial Q^F} + \frac{\partial L_{aux}}{\partial Q^F})$;
7. Update parameters using Adam;

2.3. Optimization

Let $\{x_i, y_i\}_{i=1}^N$ be the training samples. The proposed method jointly optimize the main network $F$ and the mixed-precision network which is the combination of $F$ and $H$, denoting as $F \circ H$. The training objective is:

$$
\min_{\theta^F, \theta^H} \sum_{i=1}^N L(F(x_i; \theta^F), y_i) + L_{aux}((F \circ H)(x_i; \theta^H, \theta^F), y_i),
$$

(2)

where $\theta^F$ and $\theta^H$ represent the parameters for the backbone $F$ and the auxiliary module $H$, respectively. $L$ is the task objective and $L_{aux}$ is the auxiliary loss. In the classification task, both terms are set to the cross-entropy loss. From Eq. (2), we can note that $\theta^F$ is shared among $F$ and $F \circ H$. Following the chain rule, the gradient of $\theta^F$ will have an additional term comes from $L_{aux}$. As a result, the approximated gradient is averaged from both the mixed-precision network and the original low-precision network to achieve more accurate updating direction. In other words, the full-precision module $H$ provides direct gradient for $F$ using weight sharing during back-propagation. We summarize the proposed learning process for a quantized neural network in Algorithm 1.

2.4. Relationship to other methods

In this section, we will elaborate the relationship between the proposed auxiliary learning and other related approaches.

**Differences to applying auxiliary classification losses to intermediate activations.** An alternatively way to create auxiliary full-precision routes for learning low-precision network is to apply classification losses to intermediate activations. The schematic illustration of this idea is shown in Fig. 2 (a), where each intermediate output is attached to a classifier to perform classification. In such a case, the final training objective becomes:

$$
\hat{L}_{obj} = \tilde{L} + \sum_{i=1}^M \alpha_i \hat{\ell}_i,
$$

(3)

where $\tilde{L}$ is the classification loss of the original low-precision network, $\hat{\ell}_i$ is the classification loss applies to the $i$-th intermediate output, and $\alpha_i$ is the weight associated to the $i$-th loss function. This scheme can directly propagate gradient to each block during training through the full-precision classifiers. However, its supervision is very restrictive since it essentially assumes that the intermediate output can be directly used for classification. In practice, we often find that choosing the positions of adding the additional supervisions or the weight $\alpha_i$ can be challenging. An inappropriate setting of those factors may lead to inferior performance than that achieved by directly training the low-precision model.

**Differences with knowledge distillation.** Knowledge distillation (KD) has been explored to assist the quantized model training [32, 56, 57]. In particular, a low-precision student network learns to generate similar posterior probabilities and/or feature representations of a full-precision teacher network (see Fig. 2 (b)). The training objective can be formulated as

$$
\hat{L}_{obj} = \hat{L}_1 + \hat{L}_2 + \sum_{i=1}^M \beta_i \hat{\ell}_i,
$$

(4)

where $\hat{L}_1$ and $\hat{L}_2$ are task-specific objectives for student and teacher networks respectively. $\hat{\ell}_i$ represents the $i$-th distillation loss.
Although both the proposed method and KD use a full-precision network to guide the training of a low-precision network, the ways of exerting this guidance are significantly different in KD and the proposed method. Specifically, in KD, the guidance from the full-precision model to the low-precision model is from the distillation losses while in our method this is achieved by making the parameters of a low-precision model shared with the mixed-precision model. There are many advantages of our method comparing to KD: (1) Our method only needs an extra memory to store the auxiliary module rather than a full-precision network. Comparing to KD, our method is more memory-efficient. (2) Our method only uses a single auxiliary loss but can create guidance signals to various blocks of the low-precision model. In contrast, KD needs multiple distillation losses to achieve this. Thus it usually involves more hyper-parameters, i.e., the weight of each loss term . Moreover, we empirically find that the proposed learning strategy performs consistently superior than KD for learning quantized networks in Sec. 4.1.3 and Sec. 4.1.4.

3. Extension to object detection

Most existing methods evaluate low-precision network with the classification task. Building a low-precision network for more difficult object detection task remains a challenge. To fill this gap, we further extend our method to build a quantized object detector. Following the work in \cite{24, 25}, we consider the object detection framework consisting of a backbone, a feature pyramid and prediction heads. We directly use the quantized network pretrained on the ImageNet classification task to initialize the detection backbone. We adopt the uniform quantization approach QIL \cite{17} to quantize both weights and activations, where the quantization intervals are explicitly parameterized and jointly optimized with the network parameters. We add an individual auxiliary module for each prediction head while sharing a single module for the backbone.

Besides of applying the proposed auxiliary module and learning strategy to help training the quantized detector, we also propose a modification which we find beneficial. Specifically, unlike \cite{22} which freezes the batch normalization (BN) statistics during training to stabilize optimization, we instead propose an alternative strategy where the BN statistics still keeps updating: we propose that except the last layers for classification and regression, parameters of the prediction heads are not shared across all feature pyramid levels. This is different from the common full-precision setup. The motivation of this design is that the multi-scale semantic information can not be encoded effectively due to the quantization process at different pyramid levels. For a full-precision network, using a shared head is sufficient to represent rich semantic information for classification and regression with the continuous activations. However, in the low-precision setting, the representational capability of activations is highly degraded due to its discrete values. For the same reason, the batch statistics of quantized activations may differ drastically across different levels. Therefore, each head should learn independent parameters to capture the corresponding multi-scale information.

Remark: (1) The proposed modification does not share prediction heads and thus uses more parameters in the low-precision model. However, we should note that without sharing does not increase any additional computational complexity. Even though the number of parameters is increased, the memory consumption is still significantly reduced comparing to the full-precision model due to the low-bit storage. (2) We empirically find that not sharing heads may not improve (but reduce) the performance in the full-precision setting. So the proposed modification is only for low-precision networks. Please check the experiments in Sec. 4.2.3 for more discussions.

4. Experiments

In this section, we evaluate our proposed methods on image classification in Sec. 4.1 and object detection in Sec. 4.2, respectively. To investigate the effectiveness of the proposed method, we define several methods for comparison: Auxi: We optimize the network with the auxiliary module. KD: We employ the joint knowledge distillation in \cite{56, 57} to improve the quantized network. Additional loss: We evenly insert classification losses at intermediate layers to assist training. Note that we will detail the settings in specific sections.

4.1. Experiments on image classification

We perform experiments on two standard image classification datasets: CIFAR-100 \cite{19} and ImageNet \cite{42}. The CIFAR-100 dataset consists of 60,000 color images of size \(32 \times 32\) belonging to 100 classes. There are 50,000 training and 10,000 test images. ImageNet contains about 1.2 million training and 50K validation images of 1,000 object categories. To verify the effectiveness of the proposed auxiliary learning strategy, we experiment on various representative quantization approaches, including uniform fixed-point approach DoReFa-Net \cite{54}, non-uniform fixed-point method LQ-Net \cite{52}, as well as binary neural network approaches BiReal-Net \cite{30} and Group-Net \cite{58}.

4.1.1 Implementation details

Following previous approaches \cite{14, 52, 54, 55, 57}, we quantize all the convolutional layers to ultra-low precision except the first and last layers. However, to further improve the efficiency, we quantize the first convolutional layer and the last fully-connected layer to 8-bit. We first pre-train the full-precision counterpart as initialization and then fine-
tune the quantized model. For all ImageNet experiments, training images are resized to 256 × 256, and 224 × 224 patches are randomly cropped from an image or its horizontal flip, with the per-pixel mean subtracted. We use the single-crop setting for testing. No bias terms are used. We use SGD optimizer for the pre-training stage. For the fine-tuning stage, we adopt the Adam optimizer [18]. The mini-batch size is set to 256. We train a maximum 35 epochs and decay the learning rate by 10 at the 25-th and 30-th epochs. For fine-tuning the fixed-point methods [52, 54], the learning rate is initialized to 1e-3. For fine-tuning binary neural networks [30, 58], the initial learning rate is set to 5e-4. In practice, we take the output of each residual block [11] as the input of the auxiliary module. Our implementation is based on PyTorch.

4.1.2 Effect of the auxiliary module

Table 1: Accuracy (%) of different comparing methods on the ImageNet validation set.

<table>
<thead>
<tr>
<th>model</th>
<th>method</th>
<th>Top-1 acc.</th>
<th>Top-5 acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-101</td>
<td>DoReFa-Net (2-bit)</td>
<td>70.8</td>
<td>89.6</td>
</tr>
<tr>
<td></td>
<td>DoReFa-Net + Auxi</td>
<td><strong>74.6</strong></td>
<td><strong>91.9</strong></td>
</tr>
<tr>
<td>ResNet-50</td>
<td>DoReFa-Net (2-bit)</td>
<td>70.2</td>
<td>89.1</td>
</tr>
<tr>
<td></td>
<td>DoReFa-Net + Auxi</td>
<td><strong>73.8</strong></td>
<td><strong>91.4</strong></td>
</tr>
<tr>
<td>ResNet-50</td>
<td>LQ-Net (3-bit)</td>
<td>74.2</td>
<td>91.6</td>
</tr>
<tr>
<td></td>
<td>LQ-Net + Auxi</td>
<td><strong>75.4</strong></td>
<td><strong>92.4</strong></td>
</tr>
<tr>
<td>ResNet-18</td>
<td>BiReal-Net</td>
<td>56.4</td>
<td>79.5</td>
</tr>
<tr>
<td></td>
<td>BiReal-Net + Auxi</td>
<td><strong>58.6</strong></td>
<td><strong>81.2</strong></td>
</tr>
<tr>
<td>ResNet-18</td>
<td>Group-Net (5 bases)</td>
<td>64.8</td>
<td>85.7</td>
</tr>
<tr>
<td></td>
<td>Group-Net + Auxi</td>
<td><strong>66.0</strong></td>
<td><strong>86.5</strong></td>
</tr>
</tbody>
</table>

In this section, we explore the effect of auxiliary module on assisting the low-precision network optimization. The results are reported in Table 1. By combining the baseline with Auxi, we can observe a steady performance increase compared with the original baseline. This strongly supports that the learned auxiliary gradient can facilitate the convergence of the low-precision model. In particular, the gradient of the shared weights is averaged from both the auxiliary module and the original low-precision network to achieve more accurate updating direction. Moreover, increasing gradient paths is important to solve the non-differentiability of the discrete quantization process has been proved by [3, 30]. To make it more clear, we plot the convergence curves of DoReFa-Net with ResNet-50 in Fig. 3. From the figure, we can observe that baseline + Auxi converges much faster and better than the baseline. After the first epoch, the baseline + Auxi outperforms baseline by ∼ 20% on the Top-1 accuracy. This result strongly justifies that the auxiliary module efficiently solves the non-differentiable problem by providing accurate hierarchical gradient for updating parameters during back-propagation.

It is worth noting that when the network becomes deeper (e.g., ResNet-50, 101), the improvement turns out to be more obvious. For instance, Auxi brings 3.8% Top-1 accuracy increase over the 2-bit baseline on ResNet-101. It can be attributed to that when the quantized network goes deeper, the optimization becomes more difficult due to the non-differentiable discretization process. However, the full-precision auxiliary module can provide direct hierarchical gradient to effectively solve this problem. Note that with the basic uniform DoReFa-Net baseline, we currently achieve the comparing results with state-of-the-arts on ResNet-50 and ResNet-101 [17, 31, 52] without advanced non-uniform or relaxation strategies.

4.1.3 Comparison with other related approaches

In this section, we compare the auxiliary module with the related approaches discussed in Sec. 2.4 and report the performance in Table 2. The experiments are based on the 2-bit DoReFa-Net with ResNet-18, ResNet-34 and ResNet-50 on ImageNet. For KD experiments, the results are directly cited from [56]. We can observe that introducing additional losses into intermediate layers does not show obvious improvement to the final performance. Compared to KD, we don’t need to pre-train a complex teacher network whose quality is sensitive to the final performance. In contrast, we propose a simpler yet effective weight sharing strategy to jointly optimize the low-precision network and the floating-point auxiliary module. More detailed analysis on the difference between two approaches can be referred to Sec. 2.4. From the results, we can observe that Auxi consistently outperforms KD. For instance, on ResNet-50, Auxi exceeds KD by 2.4% on the Top-1 accuracy. These results justify that the auxiliary module can effectively solve the non-differentiable problem during back-propagation in the low-precision network training. With a very different learning strategy, Auxi shows consistently superior empirical results than KD. We therefore argue that the proposed auxil-
It is interesting to note that auxiliary learning may be a substitute to KD methods \cite{2018arXiv180607766H,2021arXiv210200463Z} in network quantization.

**Table 2:** Accuracy (%) of different supervision strategies on the ImageNet validation set based on 2-bit DoReFa-Net on ResNet-18, ResNet-34 and ResNet-50.

<table>
<thead>
<tr>
<th>model</th>
<th>method</th>
<th>Top-1 acc.</th>
<th>Top-5 acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-18</td>
<td>baseline (2-bit)</td>
<td>64.7</td>
<td>86.0</td>
</tr>
<tr>
<td></td>
<td>baseline + Additional loss</td>
<td>64.9</td>
<td>86.1</td>
</tr>
<tr>
<td></td>
<td>baseline + KD</td>
<td>65.6</td>
<td>86.3</td>
</tr>
<tr>
<td></td>
<td>baseline + Auxi</td>
<td><strong>66.7</strong></td>
<td><strong>87.0</strong></td>
</tr>
<tr>
<td>ResNet-34</td>
<td>baseline (2-bit)</td>
<td>68.2</td>
<td>88.1</td>
</tr>
<tr>
<td></td>
<td>baseline + Additional loss</td>
<td>68.5</td>
<td>88.2</td>
</tr>
<tr>
<td></td>
<td>baseline + KD</td>
<td>69.0</td>
<td>88.6</td>
</tr>
<tr>
<td></td>
<td>baseline + Auxi</td>
<td><strong>71.2</strong></td>
<td><strong>89.8</strong></td>
</tr>
<tr>
<td>ResNet-50</td>
<td>baseline (2-bit)</td>
<td>70.2</td>
<td>89.1</td>
</tr>
<tr>
<td></td>
<td>baseline + Additional loss</td>
<td>70.5</td>
<td>89.3</td>
</tr>
<tr>
<td></td>
<td>baseline + KD</td>
<td>71.4</td>
<td>90.0</td>
</tr>
<tr>
<td></td>
<td>baseline + Auxi</td>
<td><strong>73.8</strong></td>
<td><strong>91.4</strong></td>
</tr>
</tbody>
</table>

### 4.1.4 Experiments on plain networks

**Table 3:** Accuracy (%) of the proposed approaches on the ImageNet validation set. All the cases are 2-bit and without skip connections except for the baselines. We can observe that the auxiliary module can significantly improve the plain network performance.

<table>
<thead>
<tr>
<th>model</th>
<th>method</th>
<th>Top-1 acc.</th>
<th>Top-5 acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DoReFa-Net on ResNet-18</td>
<td>baseline (2-bit)</td>
<td>64.7</td>
<td>86.0</td>
</tr>
<tr>
<td></td>
<td>plain</td>
<td>61.5</td>
<td>84.3</td>
</tr>
<tr>
<td></td>
<td>plain + KD</td>
<td>62.7</td>
<td>85.0</td>
</tr>
<tr>
<td></td>
<td>plain + Auxi</td>
<td><strong>63.9</strong></td>
<td><strong>85.5</strong></td>
</tr>
<tr>
<td>DoReFa-Net on ResNet-34</td>
<td>baseline (2-bit)</td>
<td>68.2</td>
<td>88.1</td>
</tr>
<tr>
<td></td>
<td>plain</td>
<td>62.1</td>
<td>83.9</td>
</tr>
<tr>
<td></td>
<td>plain + KD</td>
<td>64.5</td>
<td>85.4</td>
</tr>
<tr>
<td></td>
<td>plain + Auxi</td>
<td><strong>66.4</strong></td>
<td><strong>86.8</strong></td>
</tr>
<tr>
<td>LQ-Net on ResNet-34</td>
<td>baseline (2-bit)</td>
<td>69.8</td>
<td>89.1</td>
</tr>
<tr>
<td></td>
<td>plain</td>
<td>63.5</td>
<td>84.6</td>
</tr>
<tr>
<td></td>
<td>plain + KD</td>
<td>65.7</td>
<td>86.8</td>
</tr>
<tr>
<td></td>
<td>plain + Auxi</td>
<td><strong>68.6</strong></td>
<td><strong>88.5</strong></td>
</tr>
</tbody>
</table>

### 4.1.5 Effect of different auxiliary architectures

**Table 4:** Accuracy (%) of 2-bit DoReFa-Net using ResNet-18 on the CIFAR-100 dataset.

<table>
<thead>
<tr>
<th>model</th>
<th>method</th>
<th>Top-1 acc.</th>
<th>Top-5 acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-18</td>
<td>full-precision</td>
<td>70.7</td>
<td>91.3</td>
</tr>
<tr>
<td></td>
<td>baseline (2-bit)</td>
<td>67.6</td>
<td>90.2</td>
</tr>
<tr>
<td></td>
<td>plain</td>
<td>64.6</td>
<td>88.3</td>
</tr>
<tr>
<td></td>
<td>plain + Auxi</td>
<td><strong>67.9</strong></td>
<td><strong>90.0</strong></td>
</tr>
</tbody>
</table>

We further explore an interesting by-product of the auxiliary module for network quantization. We assume that the auxiliary module mimics the effect of skip connections and can partially share its effect. We therefore analyze training a plain low-precision network without skip connections. The results can be referred in Table 3 and Table 4. Plain represents we directly optimize a low-precision plain network without skip connections. By comparing plain and plain + Auxi, we observe apparent accuracy increase by incorporating Auxi. For example, in LQ-Net ResNet-34 based experiments, introducing Auxi can boost the Top-1 accuracy by 5.1%. On tiny CIFAR-100 dataset, plain + Auxi even outperforms the Top-1 baseline. Moreover, same as the observation in Sec. 4.1.3, Auxi still performs consistently better than KD. From the plain network setting, we can strongly justify that the auxiliary module can provide hierarchical gradient to promote convergence of the quantized network.

However, we still observe performance gap between plain + Auxi and the baseline on large-scale ImageNet. This can be attributed to two assumptions of skip connections. First, the skip connections may improve the convergence of training, as indicated by the improvement observed when using Auxi. Second, the skip connection and the feature map after one convolution are added through a tensor addition. Then the representational capability (i.e., the value range) of each entry in the added activations is significantly enhanced. In other words, the plain network has less representational capability than its residual counterpart.

### 4.2 Experiments on quantized object detection

In this section, we evaluate the proposed approach on the general object detection task. Our experiments are conducted on the large-scale detection benchmark COCO \cite{2014arXiv1411.5732Z}. Following \cite{2019arXiv190102645M,2018arXiv180513675D}, we use the COCO trainval35k split (115K images) for training and minival split (5K images) for validation. We conduct experiments based on RetinaNet \cite{2018arXiv180106190F} and compare with the state-of-the-art FQN \cite{2019arXiv190311730F}.

#### 4.2.1 Training details

Training is divided into two stages. In the first stage, the detection framework is kept to full-precision. The backbone is initialized with the classification model pre-trained...
on the ImageNet dataset. Unless specified, we use the same hyper-parameters with RetinaNet. Specifically, all training and evaluation images are resized so that their shorter edges are 800 pixels. We augment training images by random horizontal flipping while no evaluation augmentations are performed. Our network is trained with stochastic gradient descent (SGD) for 90K iterations with the initial learning rate being 0.01 and the batch size of 16. The learning rate is decayed by a factor of 10 at iterations 60K and 80K, respectively. In the second stage, we use the converged model in the first stage as initialization and fine-tune with quantization. This stage uses identical settings as the full-precision training, except that we use Adam optimizer and the initial learning rate is set to 1e-3. Our implementation is based on Detectron2 [50].

### 4.2.2 Performance evaluation

We report the performance of the proposed quantized detection framework in Table 6. From the results, we can observe that our 4-bit detector can achieve near lossless results over the full-precision counterparts, which meets the requirement for practical deployment. Moreover, we can achieve significant performance boost over FQN on all comparing architectures. For example, on ResNet-50, the improvement reaches to 3.6 on AP.

<table>
<thead>
<tr>
<th>Backbone</th>
<th>Method</th>
<th>AP</th>
<th>AP$_{50}$</th>
<th>AP$_{75}$</th>
<th>AP$_S$</th>
<th>AP$_M$</th>
<th>AP$_L$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-50</td>
<td>RetinaNet</td>
<td>36.5</td>
<td>56.5</td>
<td>39.2</td>
<td>21.4</td>
<td>40.4</td>
<td>46.9</td>
</tr>
<tr>
<td></td>
<td>FQN [22]</td>
<td>32.5</td>
<td>51.5</td>
<td>34.7</td>
<td>17.3</td>
<td>35.6</td>
<td>42.6</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>36.1</td>
<td>55.8</td>
<td>38.9</td>
<td>21.2</td>
<td>39.9</td>
<td>46.5</td>
</tr>
<tr>
<td>ResNet-34</td>
<td>RetinaNet</td>
<td>35.2</td>
<td>54.3</td>
<td>37.6</td>
<td>19.5</td>
<td>38.3</td>
<td>46.2</td>
</tr>
<tr>
<td></td>
<td>FQN [22]</td>
<td>31.3</td>
<td>50.4</td>
<td>33.3</td>
<td>16.1</td>
<td>34.4</td>
<td>41.6</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>34.7</td>
<td>53.7</td>
<td>36.9</td>
<td>19.3</td>
<td>38.0</td>
<td>45.9</td>
</tr>
<tr>
<td>ResNet-18</td>
<td>RetinaNet</td>
<td>32.1</td>
<td>50.5</td>
<td>34.1</td>
<td>16.9</td>
<td>34.8</td>
<td>42.6</td>
</tr>
<tr>
<td></td>
<td>FQN [22]</td>
<td>28.6</td>
<td>46.9</td>
<td>29.9</td>
<td>14.9</td>
<td>31.2</td>
<td>38.7</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>31.9</td>
<td>50.4</td>
<td>33.7</td>
<td>16.5</td>
<td>34.6</td>
<td>42.3</td>
</tr>
</tbody>
</table>

#### 4.2.3 Ablation studies

We now perform ablation studies to give comprehensive analysis and insights of the quantized detection framework. The results are reported in Table 7. “Backbone only” indicates we only quantize the backbone to 4-bit while other parts are kept to full-precision. “Baseline” represents we directly quantize the RetinaNet using QIL [17].

When we only quantize the backbone network, we don’t observe AP drop at least on ResNet-18. This justifies that 4-bit backbone can encode accurate features for further decoding. However, when quantizing all the components including feature pyramid and prediction heads, we can find obvious precision drop. The baseline result indicates that quantizing the continuous features to a fixed range of integers can cause great multi-scale information loss.

We now explore the effect of the two strategies described in Sec. 3. First, we incorporate the proposed auxiliary learning strategy to assist the convergence of quantized detector. Specifically, Baseline + Auxi can boost the AP of Baseline by 1.4. Second, not sharing heads is particularly designed for the low-precision detector. We should note that this strategy deteriorates the performance in the full-precision setting. The reason is that a separate head can only see a certain range size of objects. In contrast, not sharing heads can boost the AP by 1.3 in the quantization setting compared with the sharing heads counterpart. It can be attributed that when quantizing both weights and activations to 4-bit, the representational capability of each head is very limited and the statistics of activations differ a lot. Therefore, each head should learn independent parameters to transform the corresponding level feature for classification and regression.

#### 5. Conclusion

In this paper, we have proposed an auxiliary learning strategy to tackle the non-differentiable quantization process in training low-bitwise convolutional neural networks. Specifically, we have explicitly utilized weight sharing to construct a full-precision auxiliary module. During training, the auxiliary module is combined with the low-precision network to form a mix-precision network, which is jointly optimized with the low-precision model. In this way, the full-precision auxiliary module can provide direct hierarchical gradient during back-propagation to assist the optimization of the low-precision network. In the testing phase, the auxiliary module is removed without introducing any additional computational complexity. Moreover, we have also worked on quantized object detection and proposed several practical solutions. We have conducted extensive experiments based on various quantization approaches and observed consistent performance increase on the image classification and object detection. To be emphasized, we have achieved near lossless results using 4-bit detectors.

**Acknowledgement** MT was in part supported by Guangdong Provincial Scientific and Technological Funds under Grants 2018B010107001. CS was in part supported by ARC DP Project ‘Deep learning that scales’. IR was in part supported by ARC Laureate Fellowship FL130100102. We acknowledge the support of the Australian Research Council through the Centre of Excellence for Robotic Vision.
References


