

Overflow Aware Quantization: Accelerating Neural Network Inference by Low-bit Multiply-Accumulate Operations

Hongwei Xie, Yafei Song, Ling Cai and Mingyang Li

Alibaba Group

{hongwei.xhw, huaizhang.syf, cailing.cl, mingyangli}@alibaba-inc.com

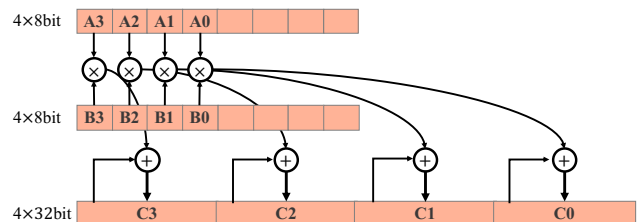
Abstract

The inherent heavy computation of deep neural networks prevents their widespread applications. A widely used method for accelerating model inference is quantization, by replacing the input operands of a network using fixed-point values. Then the majority of computation costs focus on the integer matrix multiplication accumulation. In fact, high-bit accumulator leads to partially wasted computation and low-bit one typically suffers from numerical overflow. To address this problem, we propose an overflow aware quantization method by designing trainable adaptive fixed-point representation, to optimize the number of bits for each input tensor while prohibiting numeric overflow during the computation. With the proposed method, we are able to fully utilize the computing power to minimize the quantization loss and obtain optimized inference performance. To verify the effectiveness of our method, we conduct image classification, object detection, and semantic segmentation tasks on ImageNet, Pascal VOC, and COCO datasets, respectively. Experimental results demonstrate that the proposed method can achieve comparable performance with state-of-the-art quantization methods while accelerating the inference process by about 2 times.

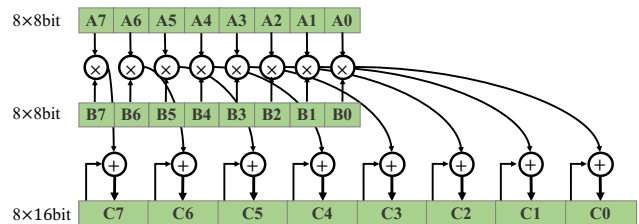
1 Introduction

To date, as a powerful machine learning system architecture, deep neural network (DNN) has been applied on numerous applications, e.g., image classification [He *et al.*, 2016], object detection [Ren *et al.*, 2015; Liu *et al.*, 2016], and semantic segmentation [Chen *et al.*, 2018]. However, to achieve high-end performance on complicated problems, most DNN systems require heavy computational resources for model inference, which inevitably limits DNN’s deployment on low-cost processors. Such processors are extensively used in billions of commercial products, such as mobile phones, drones, and Internet-of-Things (IoT) devices, which makes DNN acceleration a critical problem in both academia and industry.

To allow DNN acceleration, researchers have developed various approaches, which can be roughly divided into two



(a) For a 32-bit accumulator, one instruction can compute 4 multi-adds with 128-bit register.



(b) For a 16-bit accumulator, one instruction can compute 8 multi-adds at the same time.

Figure 1: A representative example to show that replacing a 32-bit accumulator with a 16-bit one leads to a double amount of multiply-adds operations at the same time.

groups. The first group of methods focus on designing or searching for more compact and efficient network structures, which allow for reduced number of parameters and computations while achieving comparable performance. Representative methods in this category include MobileNet [Howard *et al.*, 2017], EfficientNet [Tan and Le, 2019], ProxylessNAS [Cai *et al.*, 2019], and so on. The second group aims at improving the efficiency of arithmetic computation, i.e., the multiply-accumulate (MAC) operation, as it dominates most computations during the DNN model inference. One widely used method is to approximate the original floating-point calculation using fixed-point operation to achieve computation acceleration. This type of method is well-known as *quantization* [Jacob *et al.*, 2018]. Representative visualization of quantization is shown in Figure 1(a), where 8-bit fixed-point integers are used to approximate floating-point values and 32-

bit fixed-point variables are used to hold MAC results. Moreover, in addition to speeding up the MAC operations, quantization also achieves better parallel computing based on the capability of modern CPUs.

By comparing Figure 1(b) against Figure 1(a), it can be shown that if 16-bit fixed-point variables are used to hold the MAC result, the degree of parallelism will be doubled and the I/O times will be halved. However, when 16-bit holder is used, numerical overflow on MAC results becomes a frequently-happening problem that must be explicitly considered. A straightforward solution is to use low-bit quantization (<8-bit) for all operands, which however leads to loss of quantization precision and significantly reduced the performance. In addition, low-bit quantization still needs to take up more physical bits (e.g., 4-bit quantization still requires 8-bit physical operands) in most modern CPUs, making the computational resources partially wasted.

To summarize, existing quantization methods utilize fixed number of bits to represent float values, while both high-bit and low-bit representation have limitations. The former suffers from numerical overflow problems and the latter one leads to model precision degradation. To tackle this problem, we propose a novel method to adaptively determine the quantization precision in DNNs, by optimizing the number of bits for operands while prohibiting overflow on the low-bit MAC result holders. To achieve this, we introduce a trainable quantization range mapping factor α into each layer of a DNN network, which automatically scales the quantized result to prevent the undesirable overflow. In addition, we propose a quantization-overflow aware training framework for learning the quantization parameters, to minimize the performance loss caused by post quantization [Krishnamoorthi, 2018]. To verify the effectiveness of our method, we conducted tests on a couple of state-of-the-art light-weighted DNNs for a variety of tasks on different benchmarking datasets. Specifically, our experiments include image classification, object detection, and semantic segmentation, which are tested on ImageNet, Pascal VOC, and COCO datasets, respectively. Experimental results demonstrate that, compared with state-of-the-art quantization methods, the proposed method can achieve comparable performance while speeding up the inference efficiency by about 2 times.

The main contributions of this paper are listed as follows:

1. We propose an overflow aware quantization (OAQ) algorithm for accelerating DNNs, that is able to adaptively maximize the number of bits for operands while prohibiting the numeric overflow.
2. To ensure optimized performance, we design a quantization overflow aware training framework (QOAT), to automatically learn the parameters used by the proposed OAQ algorithm.
3. We conduct extensive experiments on three public datasets using state-of-the-art light-weight DNNs. The results verify the effectiveness of our method on a variety of tasks including image classification, object detection, and semantic segmentation.

2 Related Work

In this work, we focus on quantization methods that accelerate inference on off-the-shelf hardware platforms. While non-uniform quantization methods [Stock *et al.*, 2019; Gao *et al.*, 2019] are also shown to be effective, they do not allow efficient implementing on modern CPUs.

A representative early method is binary quantization [Rastegari *et al.*, 2016], which quantizes both weights and activations to one bit, by using bit-shift and bit-count instead of multiply-adds operators to speed up. This method achieves acceptable performance on common over-parameterized networks, like AlexNet [Krizhevsky *et al.*, 2012], but leads to substantial performance degradation on light-weight networks, e.g. ResNet-18 [He *et al.*, 2016] and MobileNet [Howard *et al.*, 2017]. As of today, one of the most widely used quantization methods is 8-bit quantization [Jacob *et al.*, 2018; Krishnamoorthi, 2018; Jain *et al.*, 2019], which is extensively applied in different applications on various hardware platforms. 8-bit quantization converts the inference process into integer-only operations, that could result in $2\times \sim 3\times$ faster inference process on mobile CPUs. However, when tackling resource-demanding large networks or deployed on resource-constrained platforms, additional DNN acceleration is still required.

To allow further DNN acceleration, low-bit quantization techniques are under active exploration, in which a key problem is to balance the inference speed and model performance. [Choi *et al.*, 2018] proposes PACT to train not only the weights but also the clipping parameters for clipped ReLU using gradient descent. [Louizos *et al.*, 2018] presents RQ to optimize the quantization part with gradient descent. However, both methods suffer from performance degradation on lightweight networks. In fact, when quantizing both weights and activations to 4 bits, PACT [Choi *et al.*, 2018] leads to accuracy reduction from 70.9% to 62.44%. Also, quantizing MobileNet using RQ [Louizos *et al.*, 2018] to 6-bit achieves 68% accuracy only. Additionally, existing low-bit techniques only focus on the classification task. Evaluation results on other popular tasks, e.g., object detection or semantic segmentation, are limited on literatures. Inference accelerating using properties of processors is another research direction. [Zeng *et al.*, 2019] proposes to decrease computational load of multiplications, by exploiting the parallel computing capability of modern CPUs. [Gong *et al.*, 2019] implements 2-bit fast integer arithmetic with ARM NEON technology and achieves $1.7\times$ speed up over NCNN [Tencent.Inc, 2017]. However, the reported results still suffer from significant performance loss, e.g., 4-bit quantized MobileNet-v2 leads to performance drop from 71.8% to 64.8%.

Moreover, there are a number of mixed-precision quantization methods [Wang *et al.*, 2019; Wu *et al.*, 2018; Dong *et al.*, 2019], which focus on searching for an optimal bit-width setup that can achieve high-level acceleration on customized hardware platforms while avoiding performance drop. Compared to those methods, the proposed OAQ framework focuses on off-the-shelf devices, and addresses the numerical overflow problem by designing trainable parameters in each layer of a network.

3 Quantization Prerequisites

In this section, we first present the general algorithmic framework of quantization methods. Subsequently, we analyze the inherent overflow problem of the general framework and provide mathematical conditions that allow overflow aware quantization (OAQ). Detailed steps on performing OAQ approach are discussed in the next section.

3.1 Standard Quantization Method

To convert a floating-point real number $r \in \mathbb{R}$ to a fixed-point quantized number $q \in \mathbb{Z}$, the following affine mapping function is typically used [Jacob *et al.*, 2018]:

$$r = S(q - Z), \quad (1)$$

where S is the scale factor and Z is the zero-point parameter. By denoting $q_a^{(i,k)}$ the element at i th row and k th column of matrix A , the matrix multiplication of $C = A \times B$ in quantized domain can be computed element-wise as:

$$q_c^{(i,k)} = Z_c + P \sum_{j=1}^N \left((q_a^{(i,j)} - Z_a)(q_b^{(j,k)} - Z_b) \right). \quad (2)$$

where the multiplier P is defined as

$$P = \frac{S_a S_b}{S_c}, \quad (3)$$

which can be implemented by fixed-point multiplication and efficient bit-shift [Jacob *et al.*, 2018]. By expanding terms in (2), we are able to obtain:

$$q_c^{(i,k)} = Z_c + P(NZ_a Z_b - Z_a M_b^{(k)} - Z_b M_a^{(i)} + \sum_{j=1}^N q_a^{(i,j)} q_b^{(j,k)}), \quad (4)$$

where

$$M_b^{(k)} = \sum_{j=1}^N q_b^{(j,k)}, M_a^{(i)} = \sum_{j=1}^N q_a^{(i,j)}. \quad (5)$$

In (4), the majority of computation costs are the core integer matrix multiplication accumulation:

$$\sum_{j=1}^N q_a^{(i,j)} q_b^{(j,k)}. \quad (6)$$

To compute for (6), the standard method is to accumulate products of 8-bit values (signed or unsigned) with a 32-bit integer accumulator (also see Figure 1(a)):

$$C_{\in \mathbb{Z}_{32}} += A_{\in \mathbb{Z}_8} \times B_{\in \mathbb{Z}_8}. \quad (7)$$

where \mathbb{Z}_y represents the space of y -bit representable number. As a result, a 32-bit register is required for caching the intermediate result. A NEON instruction can compute multiple multiply-adds at the same time, but it is limited by the register size and the number of multiplying and summing units on board. The limited register resource is usually the main

bottleneck. Moreover, we note that under most ARM architectures, there is no instruction to implement (7). To this end, existing popular mobile inference engines (e.g., TFLite and NCNN) typically rely on

$$C_{\in \mathbb{Z}_{32}} += A_{\in \mathbb{Z}_{16}} \times B_{\in \mathbb{Z}_{16}}. \quad (8)$$

as a replacement, i.e., VMAL.S16 on ARM architecture.

Additionally, we note that loading data between register and memory is heavy operations. By applying (8), more register space is used to implement a bigger micro-kernel¹. This largely reduces the frequency of transferring intermediate results between register and memory. To show more details, we evaluated the performance of convolutions on MTK8167s CPU with different implementations. Using VMLAL.S8 and 4x8 micro-kernel, the computation efficiency is improved by 36%. While applying a 4x16 micro kernel, we achieved 73% speed up.

3.2 Optimize Quantization Operations

To optimize the efficiency of current quantization scheme, we seek to use 16-bit integer as accumulator

$$C_{\in \mathbb{Z}_{16}} += A_{\in \mathbb{Z}_8} \times B_{\in \mathbb{Z}_8}. \quad (9)$$

Compared to (7), one NEON instruction here (VMLAL.S8 on ARM architecture) is able to compute double amount of multiply-adds operations, as shown in Figure 1(a) and Figure 1(b).

However, by directly using (9), numerical overflow becomes an unavoidable problem. This is one of the core problem we seek to resolve in this work. To make this possible, we first re-expand (2) by assuming q_a as the quantized inputs and q_b as the quantized weights:

$$q_c^{(i,k)} = Z_c + P \sum_{j=1}^N \left((q_a^{(i,j)})(q_b^{(j,k)} - Z_b) - Z_a q_b^{(j,k)} + Z_a Z_b \right), \quad (10)$$

which can be rewritten as

$$q_c^{(i,k)} = Z_c + P \left[\sum_{j=1}^N \left(q_a^{(i,j)} \hat{q}_b^{(j,k)} \right) + B \right], \quad (11)$$

where

$$B = - \sum_{j=1}^N Z_a q_b^{(j,k)} + N Z_a Z_b, \quad (12)$$

$$\hat{q}_b^{(j,k)} = (q_b^{(j,k)} - Z_b). \quad (13)$$

In above equations, Z_a, Z_b, q_b are all be constant values once training is complete, and thus B and $\hat{q}_b^{(j,k)}$ can be computed in advance to improve inference efficiency.

Based on above equations, we point out that, to allow efficient computation under (7), the following three conditions must be satisfied:

$$\begin{cases} q_a^{(i,j)} \in \mathbb{Z}_8 \\ \hat{q}_b^{(j,k)} \in \mathbb{Z}_8 \\ \sum_{j=1}^N q_a^{(i,j)} \hat{q}_b^{(j,k)} \in \mathbb{Z}_{16} \end{cases}, \quad (14)$$

¹<https://engineering.fb.com/ml-applications/qnnpack/>

It is important to note that the last condition in (14) should hold for both final number and all intermediate numbers. We also point out that, the second and last conditions in (14) are *not* always naturally true. Without taking special consideration, numerical overflow will frequently happen. To guarantee (14) in a DNN system, additional algorithms need to be designed and implemented.

4 Overflow Aware Quantization Framework

This section describes the details of our overflow aware quantization algorithm, including both representation and training, to ensure the important three conditions in (14).

4.1 Adaptive Integer Representation

One straightforward method to reduce accumulation overflow is to narrow the range of each quantized value. For example, by using 4-bit quantization instead of 8-bit quantization, real values are mapped to $[-8, 7]$ instead of $[-128, 127]$. As a result, numerical overflow becomes significantly less likely to happen, at a cost of wasting a large number of bits and reducing accuracy. To this end, we propose an adaptive float-bit-width method to fully utilize the representation capability without arithmetic overflow.

Specifically, we use a float quantization range mapping factor α to adjust the affine relationship between the real range and quantized range, as shown in Figure 2. Scaled by α , the original 8-bit (the biggest bit-width can be used) quantization range $[-128, 127]$ is mapped to $[\lfloor \frac{-128}{\alpha} \rfloor, \lfloor \frac{127}{\alpha} \rfloor]$. By enlarging α , we are able to narrow down the quantized value range until the arithmetic overflows are eliminated.

To present this mathematically, the affine function (1) can be written as

$$q = \frac{r}{S} + Z. \tag{15}$$

By applying the scale factor α on S

$$S' = \alpha \cdot S = \alpha \cdot \frac{r_{max} - r_{min}}{2^b - 1}, \tag{16}$$

the quantized value is narrowed to

$$q' = \frac{r}{S'} + Z, \tag{17}$$

where r_{min} and r_{max} are the minimum and maximum limits of the real value, and b is the number of bits for the quantized value.

We name this as float-bit-width method since it differs from the traditional integer low-bit representation whose quantization ranges have to be chosen from the limited bit-width set, e.g, 3-bit for $[-4, 3]$ or 4-bit for $[-8, 7]$. By using the proposed method, it is feasible to utilize different quantization range mapping factor α in *each* layer, to maximize the representation capability while prohibiting overflow. Additionally, since α is continuous, it can also be easily integrated into training process of DNNs. Details on adaptively learning α for weights and activations of each layer will be discussed in the next subsection.

In addition to the range of quantized value, the value distribution is also critical. We expect the quantized values to be

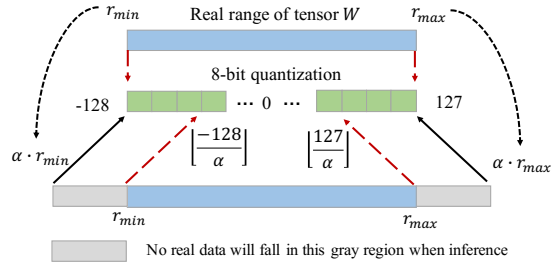


Figure 2: We introduce a float factor α to adjust the affine relationship between the real range and quantized range, e.g., enlarging α to narrow down the quantized value range.

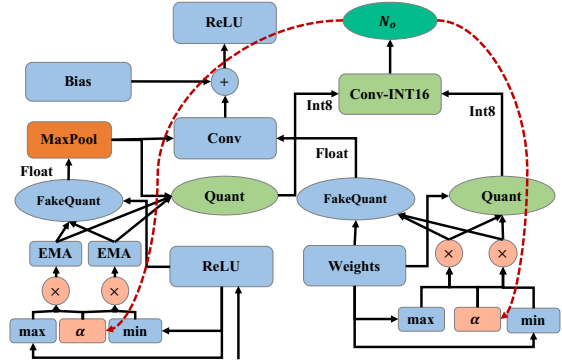


Figure 3: We insert Quant nodes into each layer of computation graph to calculate the amount of arithmetic overflow.

centered and gathered around zeros. As a result, The accumulated number will be more likely to be away from overflow. Initializing weights with a normal-distributed-like initializer and applying L1-L2-Normalization are representative methods that can be used.

4.2 Learning Quantization Range Mapping Factor

To find proper α for each layer to simultaneously prohibit arithmetic overflow and retain the model’s performance, we propose a quantization-overflow aware training framework. As shown in Figure 3, inspired by simulating quantization effects in forwarding pass [Jacob *et al.*, 2018], we add the overflow aware module that simulates neural operations, e.g., Convolution, FullyConnect using 16-bit accumulator for capturing arithmetic overflow in the forward pass.

In the forward pass, 8-bit quantization simulates quantized inference by implementing in floating-point arithmetic which is called FakeQuantization (Q_{fake}) [Jacob *et al.*, 2018]. To adaptively learn α , capturing the amount of arithmetic overflow in the quantized inference process is required. Therefore, we insert a Quantization node (Q_{real}) into each layer of the computation graph in addition to Q_{fake} . Q_{real} requires r_{min} and r_{max} to produce 8-bit quantized values which are the same with inference engine. r_{min} and r_{max} are scaled by α before being passed into Q_{real} and Q_{fake} . r_{min} and r_{max} in activations are aggregated by exponential moving averages (EMA) with the smoothed parameter close to 1, such that the

observed ranges are smoothed, allowing the network to enter a more stable state. The created convolution operation with 16-bit accumulator Conv-INT16 takes the 8-bit quantized values as input to simulate the integer-only-inference on the inference engine. This calculates the amount of arithmetic overflow N_o that accumulates in the inference process, including all types of overflow defined in the overflow-free conditions (14). An easy way to capture overflow signals from the popular training framework, e.g., TensorFlow or PyTorch is to compare the results between the regular 32-bit convolution and the 16-bit convolution. As the process of getting N_o is integer-only computation, gradient descent becomes not a proper method in back-propagation. To address this problem, we use a simple rule to compensate for this.

Specifically, when N_o is bigger than zero, increasing α by:

$$\alpha += \min(lr_i * \log(N_o), l_c), \tag{18}$$

where l_c is the fixed maximum learning rate. Additionally, lr_i is the dynamic learning rate for increasing α , which decays with the steps of training. After a large number of iterations, lr_i is decayed to a quite small value to stabilize the training state.

Alternatively, if N_o is zero, we decrease α by:

$$\alpha -= lr_d, \tag{19}$$

where lr_d is the learning rate for decreasing α , whose properties and physical representation are both similar to lr_i . For improving training efficiency, calculating N_o and updating α are executed every M steps, e.g., 10 or 50 during iterations.

The strategy of inserting Q_{real} is slightly different from Q_{fake} . As shown in Figure 3, fake quantization for input or output is inserted after the activation function or a bypass connection, e.g. adds or concatenates as they change the real min-max ranges. For the operations such as max-pooling, up-sampling or padding, Q_{fake} is not required. But the 16-bit neural operation, e.g., Conv-INT16 only accepts quantized integer inputs. The outputs of Q_{real} has to be directly passed into it. That means Q_{real} must be inserted ahead of Conv-INT16 in the computation graph. Finally, Q_{real} and Q_{fake} share the same r_{min} , r_{max} and α , for strictly simulating the same inference process.

5 Experiments

In experiments, we first show in-depth analysis on the proposed framework of adaptive scale parameter. Subsequently, we compare our proposed OAQ framework against representative state-of-the-art methods, to provide metrics on our algorithmic advantages.

5.1 Overflow Study

Since our key algorithmic contribution is a method to adaptively learn the quantization range mapping factor α for DNN’s each layer, it is also important to demonstrate the distribution of α across networks in our experiments. Figure 4 provides two representative results of per-layer α values, adaptively trained by using the proposed quantization-overflow aware training framework. The detailed operations

Index	Layer Name
1	MobilenetV1/Conv2d_0
K=[2,27]	MobilenetV1/Conv2d_{K/2}_depthwise
	MobilenetV1/Conv2d_{K/2}_pointwise
28	SSD/Conv2d_13_pw_1_Conv2d_2_1x1_256
29	SSD/Conv2d_13_pw_1_Conv2d_3_1x1_128
30	SSD/Conv2d_13_pw_1_Conv2d_4_1x1_128
31	SSD/Conv2d_13_pw_1_Conv2d_5_1x1_64
32	SSD/Conv2d_13_pw_2_Conv2d_2_3x3_s2_512
33	SSD/Conv2d_13_pw_2_Conv2d_3_3x3_s2_256
34	SSD/Conv2d_13_pw_2_Conv2d_4_3x3_s2_256
35	SSD/Conv2d_13_pw_2_Conv2d_5_3x3_s2_128

Table 1: The layer name of MobileNet-v1 and MobileNet-v1-SSD.

per-layer are listed in Table 1. From this figure, we can observe that the computed scale factor varies across layer, emphasizing our ‘scale-per-layer’ adaptive design. Additionally, we note that the majority of activation factor α fall into the range of [2, 4], indicating that the quantized values are mostly between 6bits and 7bits.

Additionally, we estimate the overflow ratio of low-bit quantization with a 16-bit accumulator by random multiply-adds simulation. Specifically, we uniformly sampled N values from the quantization value range and tried to capture an overflow signal. The signal is computed by applying consecutive multiply-adds operators of the N numbers on a 16-bit holder. Since 3x3 depth-wise convolution and 1x1 point-wise convolution are widely used in light-weight DNN architectures, we chose N to be {9, 64, 256, 1024} representatively. Subsequently, the rough Non-Overflow ratio was calculated by 100000 independent simulation runs. As shown in Figure 5, 6-bit quantization could retain overflow free in most cases, while 7-bit and 8-bit methods are risky.

Both Figure 4 and Figure 5 indicate that at most 6-bits could be used if we apply low bit quantization to address the overflow problem. Therefore, we will focus on comparing OAQ against state-of-the-art 6-bit quantization methods.

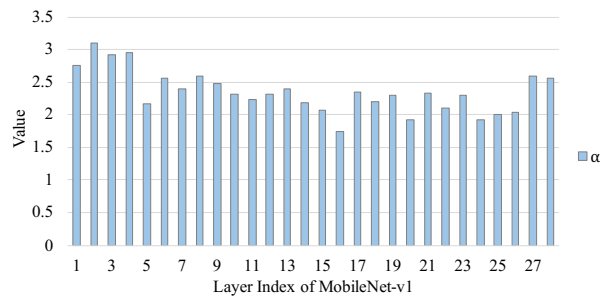
5.2 Overall Performance Evaluation

To demonstrate the performance of our proposed OAQ framework, we evaluate both inference accuracy/recall and runtime characteristics on representative public benchmarking datasets.

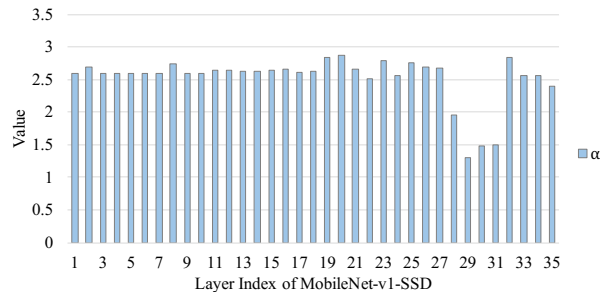
ImageNet

The first experiment is to benchmark MobileNet-v2 [Sandler *et al.*, 2018], MoibleNet-v1 [Howard *et al.*, 2017], and MobileNet-v1 with depth-multiplier 0.25 on the ILSVRC-2012 ImageNet. This dataset consists of 1000 classes, 1.28 million training images, and 50K validation images. We fine-tuned MobileNet models from pre-trained model zoo (TF-slim²). All of those models were re-trained for 20 epochs. During training, the inputs were randomly cropped and resized to 224x224 before being fed into the network. Since the inputs of the first layer are not suitable for scaling, we

²<https://github.com/tensorflow/models/tree/master/research/slim>



(a) The activation α of each layer in MobileNet-v1 trained on ImageNet. Layer names are listed on Table 1.



(b) The activation α of each layer in MobileNet-v1-SSD trained on Pascal VOC. Layer names are listed on Table 1.

Figure 4: Provides two representative results of per-layer factor values, implying that the quantized values are mostly between 6bits and 7bits.

skipped learning the quantization range mapping factor of the first layer’s weights. This rule was applied to all the following experiments, including PACT [Choi *et al.*, 2018], which always uses 8-bit weights in the first layers. We report our evaluation results using Top-1 and Top-5 accuracy.

In our tests, we focus on comparing OAQ against state-of-the-art 6-bit quantization methods, including PACT [Choi *et al.*, 2018], RQ [Louizos *et al.*, 2018], and SR+DR [Gysel *et al.*, 2018]. Comparison against other selective state-of-the-art methods were also conducted. As shown in Table 2, OAQ even outperforms 8-bit Quantization-Aware Training (QAT) [Jacob *et al.*, 2018] in certain cases. From Table 2, we observe that when quantizing MobileNet-v1, OAQ outperforms RQ [Louizos *et al.*, 2018] and SR+DR [Gysel *et al.*, 2018] by large margins. Although PACT [Choi *et al.*, 2018] is better on Top-5, the proposed method consistently performs better in other metrics, especially on MobileNet-v1-0.25, i.e., 1.36% higher than PACT. We also take DSQ [Gong *et al.*, 2019] into comparison, as it also achieved $1.7\times$ speed up over NCNN on an ARM Cortex-A53 CPU. The result shows, on MobileNet-v2 our results are significantly better, e.g., 6.84% higher.

Detection on VOC and COCO

To illustrate the applicability of our method to object detection, we applied OAQ on MobileNet-v1-SSD [Howard *et al.*, 2017; Liu *et al.*, 2016] and MobileNet-v2-SSDLite [Sandler

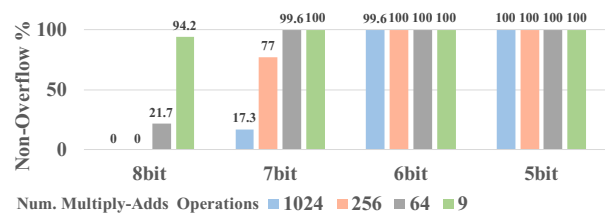


Figure 5: The Non-Overflow ratio of randomly simulated 9, 64, 256, 1024 operands’ multiply-adds 100,000 times with different quantization bits and a 16-bit accumulator.

Model	Method	Bits	MAC bits	Top1 / Top5
MobileNet-v2	FP	32	32	71.80 / 91.00
	QAT	8	32	70.90 / 90.00
	PACT	6	32	<u>71.25</u> / <u>90.00</u>
	DSQ	4	32	64.90 / —
	Our	adaptive	16	71.64 / 90.10
MobileNet-v1	FP	32	32	70.90 / 89.90
	QAT	8	32	70.10 / 88.90
	PACT	6	32	<u>70.46</u> / 89.59
	RQ	6	32	68.02 / 88.00
	SR+DR	6	32	66.66 / 87.17
Our	adaptive	16	70.87 / <u>89.56</u>	
MobileNet-v1-0.25	FP	32	32	49.80 / 74.20
	QAT	8	32	48.00 / 72.80
	PACT	6	32	<u>46.03</u> / <u>70.07</u>
	Our	adaptive	16	47.38 / 72.14

Table 2: Evaluating on ImageNet and comparing against SOTA low-bit quantization methods.

et al., 2018] and evaluated on the 2012 Pascal VOC object detection challenge and 2017 MSCOCO detection challenge. We implemented our experiments with TensorFlow and fine-tuned models from TensorFlow Object Detection API³ (only backbone since the SSD header differs with tasks). We first trained them with 32-bit floating-point precision to achieve state-of-the-art performance, and subsequently fine-tuned on VOC for 40,000 steps with batch size 32 and 60,000 steps on COCO with batch size 48 respectively.

The results on VOC and COCO are listed in Table 3 and Table 4 respectively. On both of VOC detection challenge and COCO detection challenge, the proposed method outperforms the 6-bit PACT [Choi *et al.*, 2018] significantly. In addition, our method achieves comparable performance with 8-bit QAT [Jacob *et al.*, 2018] and is of small mAP drop from the original model, i.e., about 1% mAP drop in VOC and less than 2% mAP drop in COCO.

Segmentation on VOC

To demonstrate the generalization of our method to semantic segmentation, we applied OAQ on DeepLab [Chen *et al.*, 2018] with MobileNet-v2 backbone (depth-multiplier 0.5 and

³https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/detection_model_zoo.md

Model	Method	Bits	MAC bits	mAP
MobileNet-v1 SSD	FP	32	32	73.83
	QAT	8	32	72.54
	PACT	6	32	70.88
	Our	adaptive	16	72.53
MobileNet-v2 SSDLite	FP	32	32	72.79
	QAT	8	32	72.02
	PACT	6	32	70.50
	Our	adaptive	16	71.84

Table 3: Evaluating on Pascal VOC Detection Challenge and comparing with 6bit-PACT.

Model	Method	Bits	MAC bits	mAP
MobileNet-v1 SSD	FP	32	32	23.7
	QAT	8	32	23.0
	PACT	6	32	18.4
	Our	adaptive	16	22.0
MobileNet-v2 SSDLite	FP	32	32	22.7
	QAT	8	32	21.4
	PACT	6	32	18.1
	Our	adaptive	16	21.8

Table 4: Evaluating on COCO Detection Challenge and comparing with 6bit-PACT.

1.0). The performance was evaluated on the Pascal VOC segmentation challenge, which contains 1464 training images and 1449 validation images. The results are shown in Table 5. When quantizing the original model to 6 bits with PACT [Choi *et al.*, 2018], there is a significant drop in performance, e.g., MobileNet-v2-dm0.5 backbone dropped 10.9% in mIOU. By comparison, the proposed OAQ method achieved comparable performance with QAT, and only drop 1.8% and 0.4% on MobileNet-v2-dm0.5 and MobileNet-v2 backbone respectively compared to the original model.

Inference Efficiency Benchmark

Finally, we demonstrate the capability of DNN acceleration of the proposed method on different low-cost hardware platforms. Specifically, we benchmarked computational efficiency on two selectively platforms, i.e., Allwinner V328 and MTK8167, whose processor architectures are ARM-Cortex-A7 and ARM-Cortex-A35 respectively. In this test, MobileNet-v1 and ResNet-18 were used as representative DNN models to conduct inference. To run the DNN inference, three popular neural network inference engines for low-cost platforms were selected, i.e., TFLite, MNN [Jiang *et al.*, 2020], and NCNN, under single-threaded implementation within one core.

The experimental results are listed in Table 6, which clearly demonstrate that the proposed method outperforms competing methods by wide margins. In fact, on both MTK8167s and Allwinner V328, our method achieves $2\times$ faster runtime than TFLite and $1.85\times$ faster than NCNN.

Model	Method	Bits	MAC bits	mIOU
MobileNet-v2 dm0.5 deeplab	FP	32	32	71.8
	QAT	8	32	70.4
	PACT	6	32	60.9
	Our	adaptive	16	70.0
MobileNet-v2 deeplab	FP	32	32	75.3
	QAT	8	32	74.8
	PACT	6	32	70.4
	Our	adaptive	16	74.9

Table 5: Evaluating on Pascal VOC Segmentation Challenge and comparing with 6-bit PACT.

CPU	Inference Engine	MobileNet-v1	ResNet18
Allwinner V328	TFLite	550	1370
	MNN	469	1605
	MNN + OAQ	341	1021
	Ours + OAQ	277	895
MTK8167s	TFLite	387	950
	NCNN	351	706
	MNN	311	916
	MNN + OAQ	220	604
	Ours + OAQ	189	585

Table 6: Comparison on inference time (msec) using MobileNet-v1 and ResNet18 networks.

6 Conclusion

In this paper, we propose an overflow aware quantization method to allow significant DNN inference time acceleration, and minimize the loss of accuracy. To achieve this, we propose to adaptively adjust the number of bits used for representing quantized fixed-point integers. This scheme is also incorporated into a novel training framework, to adaptively learn the overflow-free quantization range while maintaining high-end performance. By using the proposed method, an extremely light-weight neural network can achieve comparable performance with the 8-bit quantization method on the ImageNet classification challenge. Comprehensive experiments were also conducted to verify that our method can also be applied to various dense prediction tasks, e.g., object detection, and semantic segmentation by outperforming competing state-of-the-art methods.

Acknowledgements

Thanks for Shuo Zhang, Huanghao Ding, Conggang Hu, and Baitao Shao for their valuable input to this article. Specifically, Shuo Zhang helped us on maturing and structuring the idea of utilizing the int16 accumulator for accelerating the inference time. Huanghao Ding and Conggang Hu helped us on inference engine code design. Finally, Baitao Shao joined us on paper modification and proof-ready process.

References

- [Cai *et al.*, 2019] Han Cai, Ligeng Zhu, and Song Han. ProxylessNAS: Direct neural architecture search on target task and hardware. In *ICLR*, 2019.
- [Chen *et al.*, 2018] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.
- [Choi *et al.*, 2018] 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.
- [Dong *et al.*, 2019] Zhen Dong, Zhewei Yao, Amir Gholami, Michael Mahoney, and Kurt Keutzer. Hawq: Hessian aware quantization of neural networks with mixed-precision. *arXiv preprint arXiv:1905.03696*, 2019.
- [Gao *et al.*, 2019] Lianli Gao, Xiaosu Zhu, Jingkuan Song, Zhou Zhao, and Heng Tao Shen. Beyond product quantization: Deep progressive quantization for image retrieval. In *IJCAI*. AAAI Press, 2019.
- [Gong *et al.*, 2019] Ruihao Gong, Xianglong Liu, Shenghu Jiang, Tianxiang Li, Peng Hu, Jiazhen Lin, Fengwei Yu, and Junjie Yan. Differentiable soft quantization: Bridging full-precision and low-bit neural networks. In *ICCV*, 2019.
- [Gysel *et al.*, 2018] Philipp Gysel, Jon Pimentel, Mohammad Motamedi, and Soheil Ghiasi. Ristretto: A framework for empirical study of resource-efficient inference in convolutional neural networks. *TNNLS*, 29(11):5784–5789, 2018.
- [He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [Howard *et al.*, 2017] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- [Jacob *et al.*, 2018] 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*, pages 2704–2713, 2018.
- [Jain *et al.*, 2019] Sambhav R Jain, Albert Gural, Michael Wu, and Chris H Dick. Trained quantization thresholds for accurate and efficient fixed-point inference of deep neural networks. *arXiv preprint arXiv:1903.08066*, 2019.
- [Jiang *et al.*, 2020] Xiaotang Jiang, Huan Wang, Yiliu Chen, Ziqi Wu, Lichuan Wang, Bin Zou, Yafeng Yang, Zongyang Cui, Yu Cai, Tianhang Yu, Chengfei Lv, and Zhihua Wu. Mnn: A universal and efficient inference engine. In *MLSys*, 2020.
- [Krishnamoorthi, 2018] Raghuraman Krishnamoorthi. Quantizing deep convolutional networks for efficient inference: A whitepaper. *arXiv preprint arXiv:1806.08342*, 2018.
- [Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.
- [Liu *et al.*, 2016] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016.
- [Louizos *et al.*, 2018] Christos Louizos, Matthias Reisser, Tijmen Blankevoort, Efstratios Gavves, and Max Welling. Relaxed quantization for discretized neural networks. *arXiv preprint arXiv:1810.01875*, 2018.
- [Rastegari *et al.*, 2016] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *ECCV*. Springer, 2016.
- [Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.
- [Sandler *et al.*, 2018] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018.
- [Stock *et al.*, 2019] Pierre Stock, Armand Joulin, Rémi Gribonval, Benjamin Graham, and Hervé Jégou. And the bit goes down: Revisiting the quantization of neural networks. *arXiv preprint arXiv:1907.05686*, 2019.
- [Tan and Le, 2019] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.
- [Tencent.Inc, 2017] Tencent.Inc. Ncnn. <https://github.com/Tencent/ncnn>, 2017.
- [Wang *et al.*, 2019] Kuan Wang, Zhijian Liu, Yujun Lin, Ji Lin, and Song Han. Haq: Hardware-aware automated quantization with mixed precision. In *CVPR*, 2019.
- [Wu *et al.*, 2018] 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.
- [Zeng *et al.*, 2019] Linghua Zeng, Zhangcheng Wang, and Xinmei Tian. Kcnn: kernel-wise quantization to remarkably decrease multiplications in convolutional neural network. In *IJCAI*. AAAI Press, 2019.