Bare Demo of IEEEtran.cls for IEEE Journals

Matthias Wess, Sai D.P., *Fellow, OSA,* and Jane Doe, *Life Fellow, IEEE*

***Abstract*—Deployment of Deep Neural Networks (DNNs) on hardware platforms is often constrained by limited on-chip memory and computational power. The proposed weight quantization offers the possibility of optimizing weight memory alongside transforming the weights to hardware friendly data-types. We apply *Dynamic Fixed Point* and *Power-of-two quantization* in conjunction with *Layer-wise Precision Scaling* to minimize the weight memory. To alleviate accuracy degradation due to precision scaling, we employ quantization-aware fine-tuning. For fine-tuning, *quantization-regularization* and *weighted quantization-regularization* are introduced to force the trained quantization by adding the distance of the weights to the desired quantization levels as a regularization term to the loss-function. While Dynamic Fixed Point quantization performs better when allowing different bit-widths for each layer, Power-of-two quantization in combination with retraining allows higher compression rates for equal bit-width quantization. The techniques are verified on an All-Convolutional Network. With a maximum accuracy degradation of 0.10 percentage points, for Dynamic Fixed Point with Layer-wise Precision Scaling we achieve compression ratios of 7.34 for CIFAR-10, 4.7 for CIFAR-100 and 9.33 for SVHN dataset.**

***Index Terms*—IEEE, IEEEtran, journal, LaTeX, paper, template.**

I. Introduction

STARTING with AlexNet [1] Deep Convolutional Neural Networks (DCNN) have been gaining attention by delivering impressive results on challenging problems, such as object recognition on ImageNet dataset [2] or facial recognition [3]. The adaptation of such DCNNs and deep neural networks (DNNs) in various applications including autonomous driving, medical diagnosis [4] and machine translation [5] led to an ever increasing amounts of data to process under high performance requirements.

Most of these applications can be described as supervised learning tasks, split into training phase and inference. In the training phase, the algorithm is optimized to solve a certain task for the training data. The architecture of a DCNN or DNN is defined by the number of layers and their functionality (e.g. Convolutional, Fully Connected, Pool, Batch-Normalization) and the layer-specific parameters which define the dimensions and behavior of the layer in forward and backward-propagation. To train the defined architecture on a given training data, the labeled data is fed through the network and in back-propagation, the layer-specific weights are adjusted to decrease the error between the output and original label. For inference, DNN employs the model derived during training phase on the test or unknown data. The ability to correctly process the new data based on training data is called generalization ability.

Despite state-of-the-art DNNs taking one or several high-end GPUs and up to several days to train, inference can be performed on a broad spectrum of platforms including CPUs, GPUs, FPGAs, and ASICs. With the increasing size of DNNs (e.g. ResNet [6] up to 152 layers), even the complexity of inference is also exacerbating due to more critical requirements and constraints such as limited power consumption, high throughput or hard real-time processing. There are several challenges that hinder the efficient deployment and inference of the State-of-the-art Deep Neural Networks (DNN) on embedded resource constrained platforms. The two biggest challenges are the large size of the networks and the total number of necessary operations in feed-forward computation, since a hardware accelerator design can be bound either by the limit of parallel operations, or by the memory interface transmission rate [7]. As a consequence, model compression and increasing the efficiency of computations, are two legitimate ways to reach hardware requirements.

Recent works [8], [9] have proven the robustness of DNNs to compression of weights and simplification of activation functions with high number of parameters and the resulting redundancy [8], [10], [11], [12]. This enables several techniques including weight sharing [8], [13], pruning [14] and Huffman Encoding [8] to reduce external memory access. Pruning not only reduces the memory footprint of a DNN model, but also allows skipping of multiplications with 0, thus reducing the amount of total multiplications [15], [16]. To reduce also power consumption within operations, the model parameters have to be quantized in specific formats a dedicated hardware can make use of. Dynamic fixed point [12], [17] and power of two quantization [9] are two hardware friendly formats that enable performing multiplications either as low-precision multiplications or simple shift operations.

There are several approaches on how to best prepare a DNN for inference with low precision data types. On one side when employing the state-of-the-art DNNs it is desirable to directly make use of pre-trained models without architectural adjustments. [11] and [18] propose methods for layer-wise bit-width optimization without retraining but not for bit-width optimization followed by retraining. Furthermore, to fully leverage optimized hardware accelerators for efficient inference (e.g.[19]) it can be desirable to force certain quantization [9], compression or pruning schemes [15] in an additional

M. Shell was with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332 USA e-mail: (see http://www.michaelshell.org/contact.html).

J. Doe and J. Doe are with Anonymous University.

Manuscript received April 19, 2005; revised August 26, 2015.

fine-tuning step. [9] proposes incremental weight quantization while incorporating a Power-of-two datatype and achieve almost lossless quantization for several DNNs. Other works such as [20], [21] employ stochastic quantization methods to during training. In stochastic training the algorithm stores a floating point value and the quantized value at the same time and for each feed forward computation the quantized weights are newly computed on a stochastic basis.

This paper makes the following contributions:

- We propose weighted quantization-regularization (WQR), a method for trained low precision quantization of weights in Neural Networks to any given quantization scheme.
- We combine Layer-wise Precision Scaling [18] with weighted quantization-regularization to reduce the loss in classification performance while increasing the compression rate.
- We analyze the benefits of Power-of-2 (Po2) and Dynamic Fixed Point (DFP) based quantization in our setting and in combination with weighted quantization-regularization and layer-wise bit-width optimization.

Aiming at highly efficient implementation in FPGAs, we perform evaluation for quantization-regularization for dynamic fixed point [21] and power of two quantization [9] schemes on CNNs. We apply the proposed algorithm on SVHN CIFAR-10 and CIFAR-100 dataset for two different quantization schemes and show that weighted quantization-regularization decreases loss in classification performance in comparison to direct weight quantization for All-Convolutional Network on CIFAR-10 from 1.5% to 0%. The results suggest that the proposed algorithm reduces accuracy loss due to quantization of CNNs.

II. Motivational Case Study

Figure 1 explains with a simple example the two main parts of the paper. Assuming a two-layer neural network with two layers with 600 and 900 weights respectively, we want to achieve model compression by reducing the number of bits stored per weight and specific quantization of the weights to enhance the computational energy efficiency.

First note, that layer 2 has a stronger impact on the size of the weight memory, as it contains more weights. Thus, it is beneficial in terms of memory footprint to reduce the bit-width of its weights more than those of layer 1. However, quantization also negatively affects the accuracy of the algorithm, due to weight quantization errors. Therefore we apply Layer-wise Precision Scaling (Fig. 1a) to find the best trade-off between compression due to quantization and accuracy degradation. While for the example in figure 1 uniform 3-bit quantization leads to 4.5kbit weight memory, with Layer-wise Precision Scaling applied according to figure 1a only 3.6kbit weights need to be stored, allowing us to increase compression ratio by a factor 1.25.

To alleviate the accuracy degradation, performing trained quantization by applying additional regularization with the goal of reducing the weight quantization error results in an increase the accuracy. For state-of-the-art DNNs Layer-wise Precision Scaling shows even higher efficiency due to the higher variation of numbers of weights per layer (see table III).

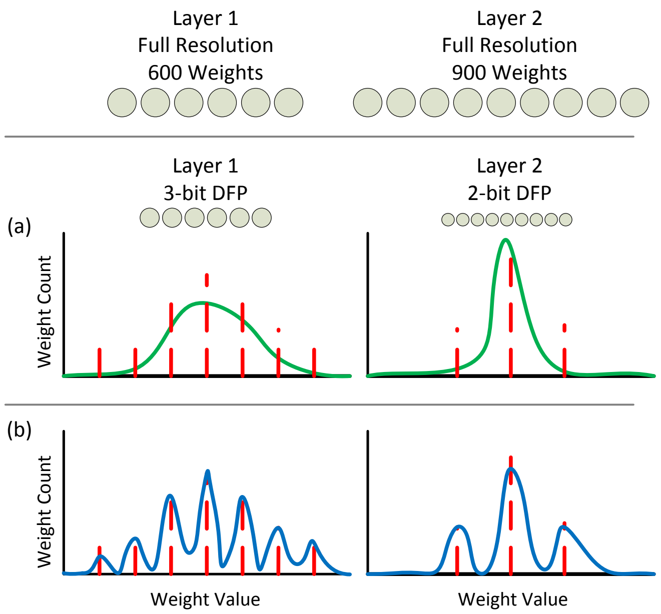

Fig. 1. An example Network with two layers demonstrating (a) Layer-wise Precision Scaling (b) Retraining with Quantization-regularization. Starting with two layer network at first the bit-widths of both layers are adjusted by defining a lower precision format with the quantization levels marked as dotted lines in the histogram charts (a). To reduce the quantization error, retraining with additional regularization, decreasing the average distance of the weights to the quantization levels, is performed (b).

III. Proposed Method

Figure 2 illustrates the entire quantization flow for learned model compression which can be separated into three steps:

1) **Quantization Scheme Evaluation:** We define and analyze two quantization strategies in terms of their effectiveness for hardware-friendly execution their advantages and disadvantages during fine-tuning and the resulting performance.
2) **Layer-wise Precision Scaling:** To increase the model compression ratio we apply layer-wise precision scaling, meaning that for each layer different bit-widths are used for weights. Thereby we study the influence of selecting different bit-widths per layer on the resulting classification accuracy.
3) **Retraining with WQR and QR:** The last task focuses on reducing accuracy degradation occurring due to quantization. As loss of accuracy is induced due to the change of weight magnitudes when approximating them by rounding to the nearest quantization level, we aim to force weights to reduce their distance to such quantization levels in retraining, thus increasing classification accuracy of the quantized network.

While the first two steps serve for finding the best quantization method and bit-width for each layer when applying quantization without retraining, in the third step we perform retraining aiming to reduce the accuracy loss caused by

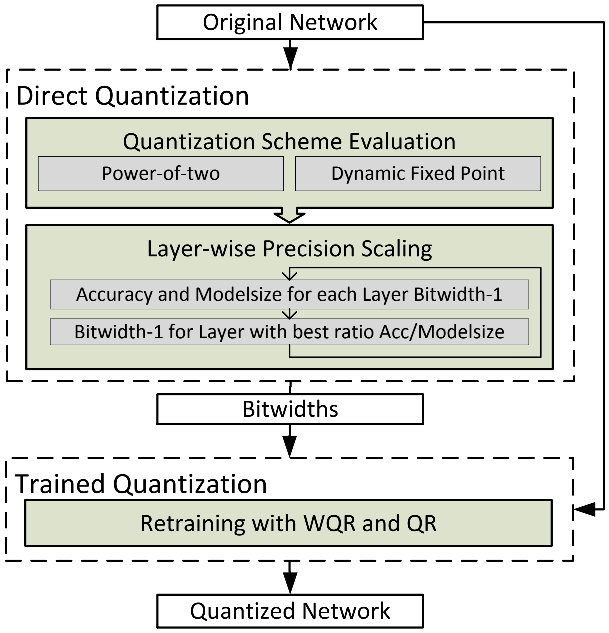
Original Network
Direct Quantization
Quantization Scheme Evaluation
Power-of-two
Dynamic Fixed Point
Layer-wise Precision Scaling
Accuracy and Modelsize for each Layer Bitwidth-1
Bitwidth-1 for Layer with best ratio Acc/Modelsize
Bitwidths
Trained Quantization
Retraining with WQR and QR
Quantized Network

Fig. 2. Learned Weight Quantization step by step

quantization. In our flow, the weights are not first quantized and then retrained, but we always start from the high accuracy model, fine-tune weights with modified loss-functions and then perform quantization. This approach has the advantage that the network parameters are trained in full precision but with the additional regularization terms which cause the weights to reduce their distance to the desired quantization levels before performing the actual quantization step. To better distinguish we use the term direct quantization for quantization without any fine-tuning. Trained quantization on the other hand consists of fine-tuning, followed by the actual quantization step. Input for the quantization process is a Deep Neural Network (DNN) M with N convolutional and/or fully connected layers and weight-tensors $W_n, 0 < n < N$ of arbitrary resolution. Details on the network used for evaluation can be found in table III. Table I summarizes the list of variables used in this work.

A. Direct Quantization

In direct quantization, the original network M is expressed as M_q where the weights W_n of each layer are represented as W_{q_n}. The values of W_{q_n} are determined by rounding each element of W_n to the quantization level with the smallest absolute distance of a previously defined quantization scheme Q.

1) Quantization Scheme Evaluation: Here, we present Power-of-two [9] and Dynamic Fixed Point [21], [17], two different quantization schemes and compare their properties for direct and trained quantization.

a) Power-of-two quantization: We implement Power-of-two (Po2) quantization similar as in [9]. Q_{p2} is given as

$$Q_{p2} = \{\pm 2^{n_1}, ..., \pm 2^{n_2}, 0\}. \tag{1}$$

TABLE I
VARIABLES USED IN THIS WORK

Variable	Comment
M	Original network model
M_q	Quantized network model
N	Number of layers
W_n	Original Weights
W_{q_n}	Quantized weights
b_n	Bit-widths for layers $n = 0...N$
Q_n	Quantization schemes for layers $n = 0...N$
Q_{p2}	Power-of-2 quantization scheme
n_1	Maximum exponent for Power-of-2 quantization
n_2	Minimum exponent for Power-of-2 quantization
s	Maximum absolute weight within the selected layer
Q_{dfp}	Dynamic Fixed Point quantization scheme
B	Unscaled Dynamic Fixed Point quantization scheme
acc_M	Classification accuracy of the original network
acc_{M_q}	Classification accuracy of the quantized network
Δacc	Accuracy degradation due to quantization
W_{mem}	Weight memory bits
λ_1	Quantization-Regularization scale factor
QR	Quantization-Regularization Term
λ_2	Weighted Quantization-Regularization scale factor
WQR	Weighted Quantization-Regularization term

n_1 and n_2 are integers with

$$n_1 = \lfloor log_2 \frac{4s}{3} \rfloor \tag{2}$$

$$s = max(abs(W)). \tag{3}$$

For a given bit-width b and n_2 are defined by

$$n_2 = n_1 - (2^{b-1} - 1). \tag{4}$$

Thus, the quantization levels depend on the distribution of weights, especially on the weight with the highest absolute value. By adding ‘0’ as a quantization level, we enable power-of-two quantization to also serve as a pruning mechanism when applied to weight matrices, as small weights are rounded to zero. In experiments symmetrical quantization schemes lead to higher classification accuracies for the quantized networks, therefore we only use $2^b - 1$ of 2^b possible quantization levels.

b) Dynamic Fixed Point: Dynamic fixed point (DFP) data type is successfully used in several works for either direct quantization or retrained model compression [21], [17]. For DFP quantization, we first define a set of $2^b - 1$ equidistant quantization levels:

$$B = \{\pm 2^{b-1} - 1, \pm 2^{b-1} - 2, ..., 0\}. \tag{5}$$

Similarly to Po2 quantization, we prefer a symmetric quantization scheme. Next B is normalized and scaled, depending on the distribution of weights:

$$Q_{dfp} = \frac{B}{2^{b-1}} * 2^{n_1}. \tag{6}$$

Figure 3 depicts the distribution of weights for an example layer of a CNN, before and after quantization. While figure 3(a) shows the distribution for Po2 quantization, figure 3(b) illustrates the distribution for DFP quantization. As can be seen that po2 has much irregular quantization values compared to DFP, and also considers the values close to ‘0’ which might help to retain the information with lower weights and aid in improving the accuracy.

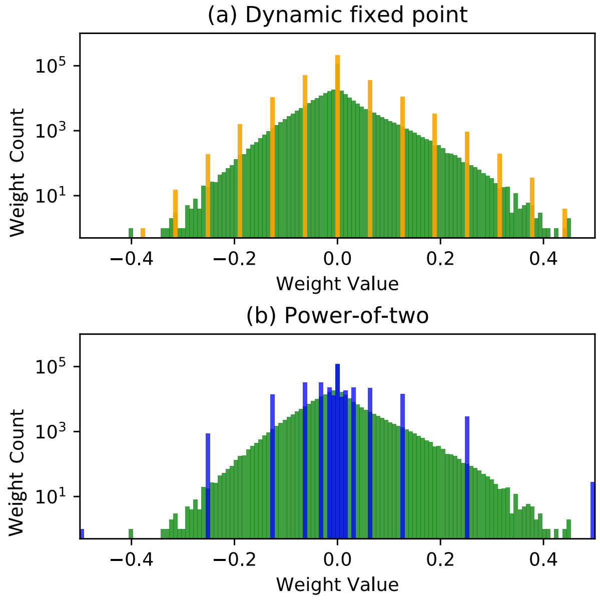

Fig. 3. Distributions before and after direct weight quantization for (a) Dynamic Fixed Point and (b) Power-of-two quantization

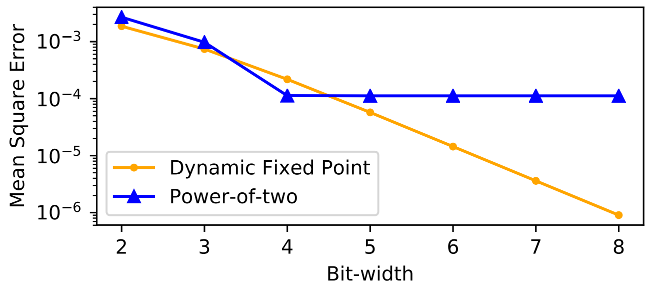

Fig. 4. Mean square error for an example layer when applying Power-of-two and Dynamic Fixed Point quantization with different bit-widths. While when using DFP quantization the error decreases exponentially, for Po2 quantization the quantization error reaches the minimum already at bit-width 4

As accuracy degradation of the quantized model M_q in comparison to the original model M is a result of the quantization error, it is necessary to understand the relation between bit-width and quantization error for both datatypes.

With DFP quantization, the mean square error (MSE) can be reduced with increasing bit-width, since every additional bit divides the intervals in half (see fig. 3a). Meanwhile when increasing bit-width in Po2 quantization, the new quantization levels are always added close to '0' (see fig. 3b). As a consequence with Po2 quantization, the quantization error can only be reduced to a certain extent. Figure 4 shows the resulting mean square errors for one weight-tensor of an example layer when applying different bit-widths.

In addition, we consider the amount of pruned weights as an important factor for model compression. In comparison to DFP, Po2 quantization decreases sparsity within the weight matrices as a result of quantization, due to the higher density of levels close to '0'. Therefore to fully benefit from the advantages of sparsity, an additional pruning step before retraining is recommended. In figure 5 the number of pruned weights depending on the selected bit-width is shown for Po2 and DPF quantization.

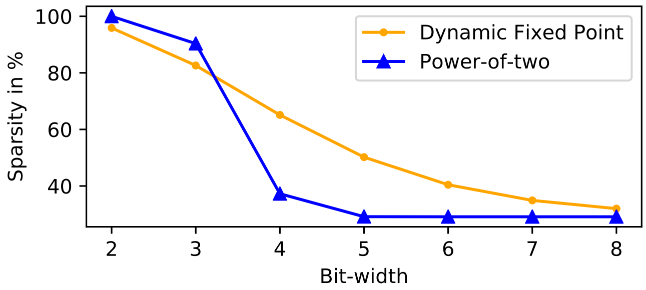

Fig. 5. With increasing bit-widths the sparsity due to quantization decreases for Dynamic Fixed Point and Power-of-two quantization. Sparsity is denotes the amount of weights that are 0 relative to the total amount of Weights

Figures 4 and 5 suggest that for direct Po2 quantization bit-widths higher than 4-bit do not further decrease the Δacc but 4-bits in comparison to 5-bits slightly increase sparsity. On the other hand, by scaling the bit-width of DFP the resulting MSE can be reduced exponentially (Fig. 4) meaning that even bit-widths higher than 8 bit deliver more accurate results. In terms of sparsity, DFP prunes more weights than Po2 for bit-widths of four and higher.

Based on these observations we expect that for direct DFP quantization Δacc can be reduced to almost 0, based on Layer-wise Precision Scaling. For direct Po2 quantization we expect a higher Δacc and no significant increase of accuracy for bit-widths higher than five. It can be seen in figures 6 and 10 that these expectations are confirmed.

2) Layer-wise Precision Scaling: Secondly, to further reduce the model-size, we apply different quantization schemes per layer by optimizing bit-widths. In comparison to choosing equal bit-width for each layer, due to the varying amount of parameters and varying distribution of weights between layers, selecting fitting quantization schemes for each layer can enable lower bit-widths per layer without reducing the resulting accuracy. For the experiments we assume either DFP or Po2 quantization. For an arbitrary network M with accuracy acc_M applying weight quantization with the set of bit-widths b_n leads to accuracy acc_{M_q} and weight memory bits

$$W_{\text{mem}} = \sum_{n}^{N} \text{card}(W_n) * b_n \tag{7}$$

where card(A) denotes the cardinality of set A. For each b_n we compute the resulting accuracy degradation

$$\Delta acc = acc_M - acc_{Mq} \tag{8}$$

and iteratively decrease the bit-width of the layer where a lower bit-width leads to the smallest product of $\Delta acc * W_{mem}$ (see algorithm 1).

Figure 6 shows the results for Layer-wise Precision Scaling performed by algorithm 1 on All-Convolutional Network [22] for CIFAR-10. We can deduce that while DFP quantization also allows direct quantization, whereas for power-of-two quantization almost always an additional fine-tuning step is necessary to achieve high accuracy results.

Algorithm 1 Layer-wise Precision Scaling

```
procedure LAYER-WISE PRECISION SCALING($M$)
    initialize $b_n$
    while $\Delta acc < \epsilon$ do
        for all $n$ in layers do
            bitwidth of $layer_n$ - 1
            Compute $Acc_{M_q}$, $\Delta acc$ and $W_{mem}$
            bitwidth of $layer_n$ + 1
        Decrease bitwidth of layer with smallest $\Delta acc * W_{mem}$
```

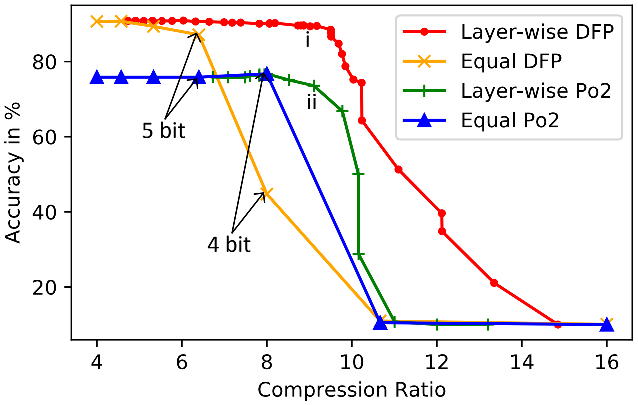

Fig. 6. Layer-wise Precision Scaling compared with equal bit-width quantization for DFP and Po2 quantization. Compression ratio is the ratio between 32bit weight memory and the weight memory for the quantized network. Point i indicates DFP with $b_n = [7\ 7\ 7\ 4\ 4\ 3\ 3\ 7\ 7]$, point ii indicates Po2 with $b_n = [4\ 4\ 4\ 4\ 3\ 3\ 4\ 4\ 4]$.

B. Trained Quantization

The used datasets (CIFAR-10, CIFAR-100 and SVHN) are already divided into **test data** and **training data**. While with Layer-wise Precision Scaling as described in section III-A2 focuses on decreasing $\Delta acc{*}W_{mem}$ without retraining, we can additionally reduce Δacc by retraining the original network on the training data with the goal of increasing accuracy of the classifier on test data. As a consequence we use the performance metrics in table II.

TABLE II
PERFORMANCE METRICS USED FOR FINE-TUNING

Variable	Metrics	Comment
acc_{tr}	Training Accuracy	Accuracy of the network on the training data
acc_M	Test Accuracy	Accuracy of the network on the test data
acc_{M_q}	Quantized Test Accuracy[1]	Accuracy of the network with quantized weights on the test data
acc_{init}	Init. Quantized Test Accuracy	Accuracy of the network with direct quantized initial weights on the test data
CR	Compression Ratio	Ratio W_{Mem} of original model to W_{Mem} of quantized model

1) Quantization-Regularization: To decrease Δacc for a selected set of bit-widths b_n, we need to find the best set of W_n so that approximation with Wq_n achieves a maximum of acc_{Mq}. As shown in [11], [23] the degradation of classification accuracy of a DNN due to quantization is directly related to the signal to quantization-noise ratio (SQNR) and the amount of weights per layer, as both influence the SQNR of the intermediate layer outputs and as a consequence the resulting network outputs. Therefore retraining network weights to achieve lower SQNR without reducing acc_M, leads to an increased accuracy of the quantized network. To enforce weight quantization during the training phase we define the quantization-regularization (QR) term as

$$QR = \sum_{n}^{N} \sum_{i}^{\text{card}(W_n)} \frac{|\, W_{n_i} - Wq_{n_i} \,|}{\max(Q_n) * \text{card}(W_n)} \tag{9}$$

which expresses the mean of the absolute weight distances of each weight to the corresponding quantized value.

By adding the QR-term to the loss function (eq. 10) weights are forced closer to the quantization levels during retraining.

$$\text{Modified Loss} = \text{Loss} + \lambda_1 * QR \tag{10}$$

During fine-tuning with the parameter λ_1 the trade-off between $min(Loss)$ and $min(QR)$ and, as a consequence, between $min(\Delta accuracy)$ and $max(acc_M)$ can be adjusted. For the experiments we applied fixed λ_1 and linearly increasing λ_1 (e.g. $\lambda_1 = 10{*}epoch$). Figure 7 depicts the trained quantization process. With each epoch the weights are pulled closer to the quantization levels, thus decreasing QR and Δacc.

While choosing a high (> 1000) λ_1 leads to fast quantization with strong accuracy degradation, a low λ_1 value (< 1) does not enforce quantization. Either way, much like weight decay low λ_1 values can help avoiding overfitting during training.

2) Weighted Quantization-Regularization: While in normal QR each weight within one layer is considered equally important for reaching high classification accuracy, the efficiency of pruning [14], [15], [13] shows that especially weights with small magnitudes can be changed without reducing the accuracy of the network. Similarly to [7], the weights can be divided into two disjoint subsets, where QR is applied on one of the subsets while the other weights are being retrained without QR. Going one step further we can multiply the QR value of each weight with the absolute magnitude of the weight. [2] This strategy forces quantization stronger on weights with higher magnitudes which can be especially useful for Po2 quantization, where density of quantization levels decreases with increasing weight values. Therefore we define the Weighted Quantization-Regularization (WQR) term as

$$WQR = \sum_{n}^{N} \sum_{i}^{\text{card}(W_n)} (\frac{|\, W_{n_i} - Wq_{n_i} \,||\, W_{n_i} \,|}{\max(Q_n)^2 * \text{card}(W_n)} \tag{11}$$

[1] For the computation of Quantized Test Accuracy, the weights of the network are directly quantized after each epoch.

[2] Previously the sum of weights is normalized to 1 for each layer

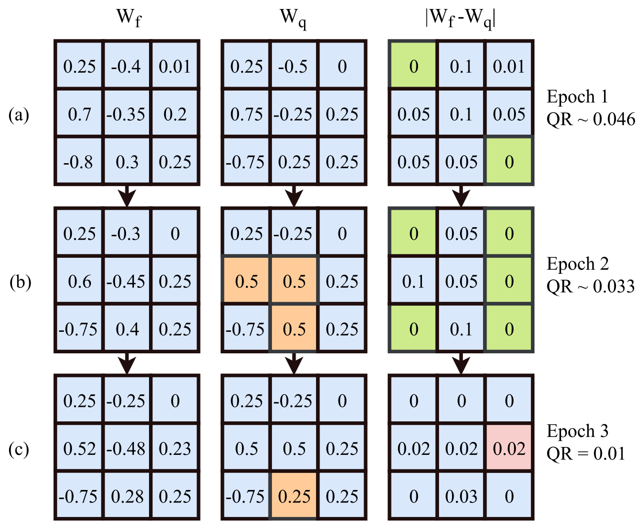

Fig. 7. Illustration of Quantization-Regularization. The first column shows the floating point values of the weights on which the actual training is performed. The second column shows the quantized weights, and the third column the element-wise absolute difference of floating point and quantized weights. Changes of quantized weights are marked in the second column. Weight update for W_f is performed based on backpropagation of the modified loss function (eq. 10). Successfully quantized weights are marked in the third column. After Epoch 1 (a) the weights are hardly regularized and QR is relatively large. Epoch 2 (b) shows that due to regularization, weights are pulled closer to the quantization levels $\mathbf{Q_n} = \{\mathbf{0}, \pm\mathbf{0.25}, \pm\mathbf{0.5}, \pm\mathbf{0.75}\}$ and QR gets smaller. For three of the weights the resulting quantization level changed, and weights decrease their distance to the next quantization level. Epoch 3 (c) shows further reduction of the QR term.

and similarly to equation 10 we can weight the trade-off between accuracy and weight regularization with λ_1 and λ_2 (eq. 12)

$$\text{Modified Loss} = \text{Loss} + \lambda_1 * QR + \lambda_2 * WQR \qquad (12)$$

Again during training the parameters λ_1 and λ_2 have to be adjusted carefully to reach the desired improvement of Acc_{Mq}, without at the same time decreasing Acc_M. In our experiments we found a linear increasing λ_2 to work best (e.g. $\lambda_2 = 10 *$ epoch).

For fine-tuning, we use algorithm 2.

Figure 8 illustrates the fine-tuning process for 4-bit equal bit-width Po2 quantized All-Convolutional Net for CIFAR-10. At the beginning of the fine-tuning process the term $\lambda_2 * WQR$ increases due to the increasing λ_2 value, while the WQR-Term decreases exponentially. At Epoch 200 learning rate is decreased from $1e^{-4}$ to $1e^{-5}$ leading to the drop of $\lambda_2 * WQR$. This can be explained by fact that a larger learning rate leads to larger weight changes. If the weights are already close to the quantization levels a lower learning rate can lead to better approximation of the weights to the quantization levels.

C. Summary and Analysis

In combination, the discussed techniques for quantization (sec. III-A), Layer-wise Precision Scaling (sec. III-A2) and fine-tuning with QR (sec. III-B1) and WQR (sec. III-B2) facilitate DNN weight compression.

Algorithm 2 Fine-tuning with Quantization-Regularization and Weighted Quantization Regularization

```
procedure TRAINED QUANTIZATION($M$, $b_n$, $\lambda_1$, $\lambda_2$)
    for epochs do
        for all Mini Batches do                                  ▷ Train on Training Data
            for $n >= N$ do                                      ▷ Quantize all Layers
                $Q_n(W_n, b_n)$                                  ▷ Determine Quantization
Scheme
                $W_{q_n}$ = Quantize($W_n$,$Q_n$)                ▷ Quantize
Weights
            Loss + $\lambda_1 * QR(W_n, Q_n, W_{q_n}) + \lambda" *$
$WQR(W_n, Q_n, W_{q_n})$
            Backpropagation(Loss,QR,WQR,$\lambda_1$,$\lambda_1$)
        Compute $Acc_M$                                          ▷ Test Accuracy
        Compute $Acc_{M_q}$                                      ▷ Quantized Test Accuracy
    return $M_q$                                                 ▷ Return Quantized Model
```

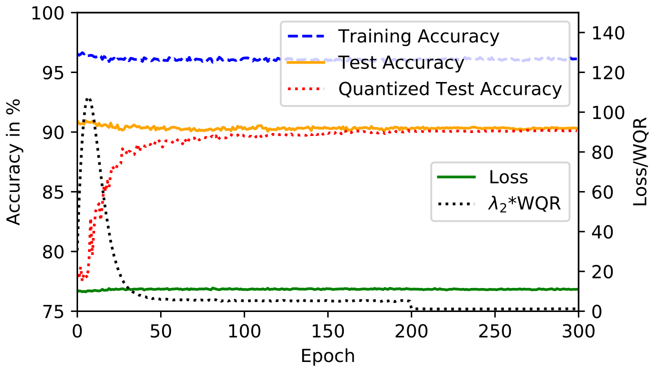

Fig. 8. Fine-tuning of All-Convolutional Net for CIFAR-10 with linear increasing $\boldsymbol{\lambda_2}$ for 4-bit equal bit-width Po2 quantization. $\mathbf{\Delta Acc}$ is decreased from $\mathbf{14.09}\%$ to $\mathbf{0.14}\%$ resulting in Quantized Test Accuracy of $\mathbf{90.18}\%$ in comparison to initial Floating Point Test Accuracy $\mathbf{90.83}\%$.

The above discussed two quantization schemes behave differently during the quantization process and require different quantization strategies. For bit-widths higher than 7 bit DFP can be applied without any retraining and still achieves almost floating point accuracy. For equal bit-width DFP quantization with 7 bit and less, Δacc increases and fine-tuning is necessary to reach the accuracy of the original network. On the other hand Po2 quantization always requires retraining, as even the use of bit-widths higher than five reduce the quantization error only to a certain extent.

To increase the compression ratio when applying DFP quantization, layer-wise precision scaling is an effective method, since not all layers require the same the bit-width for high accuracy. For instance, in modern convolution-only networks, layers with fewer parameters require larger bit-widths [11]. As a result, when applying layer-wise precision scaling the weights of the output and input layers are kept at high precision, as they usually have the fewest parameters. In comparison to DFP, for Po2 quantization, layer-wise precision scaling does not proof to be as effective, as almost the same bit-width is recommended throughout the network to achieve best accuracy.

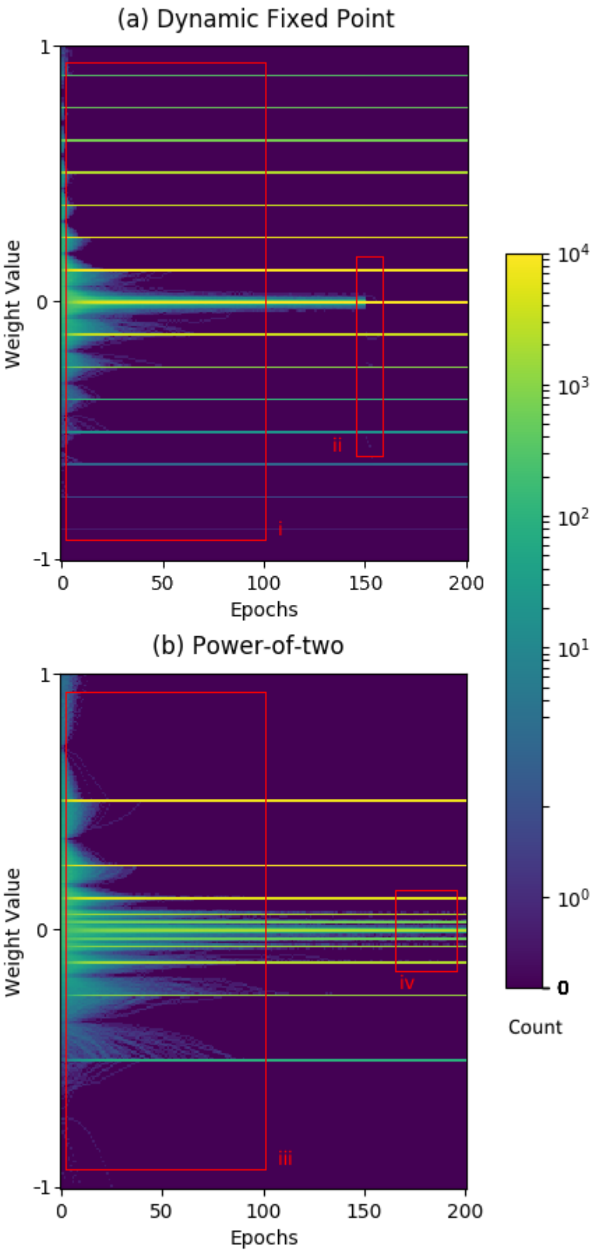

Fig. 9. Distribution of weights in layer 5 of AllConvNet (CIFAR-10) during fine-tuning with WQR and QR to (a) 4-bit Dynamic Fixed Point and (b) 4-bit Po2. DFP (a) is trained with $\boldsymbol{\lambda_2 = \mathbf{epoch} * 10}$. Region i shows the faster quantization of weights with high magnitudes and slower quantization of weights close to 0. At epoch 150 QR is applied with $\boldsymbol{\lambda_1 = 100}$ to quantize also weights with smaller values (ii). Po2 (b) is trained with $\boldsymbol{\lambda_2 = \mathbf{epoch} * 10}$. Region iii also shows slower quantization for small magnitude weights. Due to the distribution of quantization levels, weights around 0 are already close to the next quantization level (iv).

For fine-tuning of Po2 and DFP quantized networks, QR and WQR can be added to the loss function as regularization terms to force quantization during retraining. While the QR term forces all weights equally to reduce the distance to the next quantization level, the WQR term is reduced for weights with small magnitudes. Therefore, WQR operates less restrictive than QR, as the original QR value is multiplied with a factor from 0 to 1 (normalized weight magnitude). As a consequence it is an effective strategy to apply WQR followed by QR fine-tuning.

Figure 9 shows the distribution of weights in layer 5 during (a) 4-bit DFP and (b) 4-bit Po2 quantization of the All-Convolutional Network for the CIFAR-10 dataset. For fine-tuning to 4-bit DFP quantization (Fig.9a) λ_2 is increased linearly by a factor of 10 with each epoch. As QR scale factor λ_1 we use 0 before and 100 starting at epoch 150. Region i in figure 9a shows the decelerated quantization of small magnitude weights. Starting with epoch 150, also weights close to 0 are quantized, due to the application of QR (Fig. 9a,ii). For 4-bit Po2 quantization only WQR is necessary ($\lambda_2 = epoch * 10$) to achieve quantization. Due to the the non-equidistant distribution of quantization levels in Po2 quantization, the weights with high magnitudes take longer than in DFP quantization to reach the quantization levels (Fig. 9b,iii). Additional QR is not necessary since, due to the high density of quantization levels close to 0, the weights with smaller magnitudes induce a very small quantization error (Fig. 9b,iv).

QR and WQR enable trained quantization to improve performance in comparison to direct quantization. Regularization-based quantization is very simple to implement and applicable for any quantization scheme. As a consequence QR and WQR could be employed alongside other effective quantization techniques such as stochastic quantization. Besides the simplicity of the approach, it also allows a deeper analysis of quantization schemes by recording the change of weight distribution during training.

While enabling compression for more efficient inference, during training QR and WQR add overhead due to the mandatory quantization step after each mini-batch and the necessary computations for calculation of QR and WQR. The fact that all weights have to be stored as floating point values and quantized values, increases the weight memory during training by a factor of 2.

IV. Experimental Results

The following section describes the experimental results for direct and trained quantization of All-CNNs on three datasets.

A. Experimental Setup

For experimental evaluation, we apply the proposed methods on All-Convolutional Networks for the Datasets CIFAR-10 [24], CIFAR-100 and SVHN [25]. We use All-Convolution Network model All-CNN-C from [22] for evaluation. The CNN Architecture summed up in table III consists of nine convolution layers and a global average pooling layer. Due to the similar filter-width and height the number of weights in the convolution layers mainly depends on the filter-depths. The number of operations also depends on the layer-output dimensions. However, it needs to be noted that the proposed technique is neither architecture nor dataset bound.

For the three datasets, we use the predefined training and test sets. For the floating point baselines we trained the CNN

TABLE III
ALL-CNN-C ARCHITECTURE AND THE NUMBER OF WEIGHTS AND MAC-OPERATIONS FOR ONE FORWARD COMPUTATION WITH BATCH-SIZE ONE.

Layer (WxH)	Output Dim. (WxHxD)	#Weights	#MACs
Input	32x32x3	-	-
Conv. 3x3	32x32x96	2K	2.3M
Conv. 3x3	32x32x96	83K	74.6M
Conv. 3x3	32x32x96	83K	74.6M
Pooling 2x2	16x16x96	-	-
Conv. 3x3	16x16x192	165K	32.5M
Conv. 3x3	16x16x192	332K	65M
Conv. 3x3	16x16x192	332K	65M
Pooling 2x2	8x8x192	-	-
Conv. 3x3	8x8x192	332K	11.9M
Conv. 1x1	8x8x192	37K	2.4M
Conv. 1x1	8x8x10	2K	0.1M
Pooling 8x8	10	-	-
Total	-	**1386K**	**329.7M**

for 350 epochs with initial learning rate 10^{-3} multiplied by a fixed multiplier after epochs 200 and 300. To avoid overfitting, we use dropout with dropout rate 0.5 after layers. In contrast to the original All-CNN paper [22], the models are not regularized by weight decay to avoid interfering with the studied regularization methods. In terms of data augmentation we only apply horizontal flipping and random shifting by a maximum of 3 pixels.

B. Performance Analysis

We evaluate the performance of the proposed method comparing the test accuracy of the original network with the resulting accuracies after direct and trained quantization. For the tables V, IV and VI, we make use of the abbreviations in table II. We aim to achieve high compression rates in terms of weight memory while maintaining high test accuracy. In addition to bit-width reduction we also consider the resulting sparsity as an important factor for possible further compression. Assuming skipping of multiplications with 0 weights, sparsity also reduces the amount of required MAC-operations for forward computation.[3] For all experiments except the two marked, bit-width for activations is 32-bit Fixed Point.

1) CIFAR-100: CIFAR-100 is an image classification dataset consisting of a training set of 50000 and a test set of 10000 32×32 color images representing 100 different categories such as airplanes, automobiles, birds, cats, deers,dogs, frogs, horses, ships and trucks [24]. The training batches contain exactly 5000 images from each class. Table V shows the resulting accuracies after fine-tuning with 200 epochs of WQR with $\lambda_2 = epoch \times 10$. In addition λ_1 is set to 100 starting at epoch 150. This setup is not optimal for all configurations, as in some cases training with only QR would be sufficient, but it allows using the same parameterization for each iteration increasing comparability. Compared to a full implementation (32-bit), the proposed layer-wise quantization (DFP-lw) saves ~4-8× in terms of weight memory and has an accuracy loss by ~0.1-4.5%. Compared to the reduced equal bit-width quantization, the proposed layer-wise quantization has higher sparsity and smaller weight memory.

In addition, figure 10 depicts the results of trained quantization in comparison to direct quantization. We can see that retraining with QR and WQR in all cases increases classification accuracy in comparison to direct quantization.

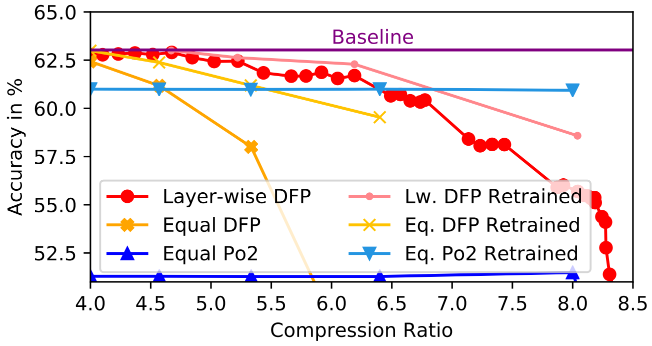

Fig. 10. Results for direct and trained DFP and Po2 quantization with equal bit-widths on ALL-CNN for CIFAR-100. For DFP also Layer-wise Precision Scaling with direct and trained quantization is illustrated.

Comparing equal bit-width quantization with Layer-wise Precision Scaling for DFP datatype, we can see that for similar compression ratios, equal bit-width (DFP-eq) never reaches the accuracy of Layer-wise Precision Scaling (DFP-lw), even when retraining with WQR/QR is applied. For Po2 quantization we found equal 4 bit quantization (Po2-eq) the most effective method as higher bit-widths did not increase accuracy and Layer-wise Precision Scaling for lower than 4 bit leads to drastic accuracy drop.

2) CIFAR-10: CIFAR-10 is a benchmark image classification dataset equal to CIFAR-100 in terms of image and dataset sizes, which instead of 100 classes divides the images into 10 classes. We use the same training method as for CIFAR-100. The results for trained quantization of All-CNN for CIFAR-10 are shown in table V. To allow comparing to other works for the two marked configurations, we also quantized the activations to 8-bit DFP. In contrast to ALL-CNN for CIFAR-100, for this dataset higher compression rates can be achieved. For Compression Ratio ~ 8, DFP with Layer-wise Precision Scaling (DFP-lw) gives lowest accuracy degradation of 0.51 percentage points. For equal bit-with quantization Po2 outperforms DFP by 1.55 percentage points. For Compression Ratio 7.38 classification accuracy of DFP with Layer-wise Precision Scaling is only 0.01 percentage points below the floating point baseline.

3) SVHN: The SVHN image classification dataset consists of 694K 32×32 color images for training and 26K images for testing. The images represent digits form 0 to 9. Similarly to CIFAR-100 and CIFAR-10 we perform Layer-wise Precision Scaling and retraining with WQR and QR for model compression. The results are shown in table VI and figure 11. For the SVHN dataset at Compression Ratio ~ 8, Po2 quantization and DFP with Layer-wise Precision Scaling both outperform the original baseline model.

[3] The number of skipped MACs due to a pruned weight depends on the layer-input dimensions. Therefore sparsity in weights is unequal to sparsity in MACs.

TABLE IV
RESULTS IN COMPARISON TO FLOATING POINT BASELINE OF ALL-CNN FOR CIFAR-100 (CR=COMPRESSION RATIO)

b_n	Type	$W_{mem}[Bit]$	CR	Sparsity[%]	non-0 MACs	MAC Sparsity[%]	acc_{tr}[%]	acc_{init}[%]	acc_{M_q}[%]
32 bit	float	44344K	1	0	329.7M	0	83.24	63.03	**63.03**
8 bit	DFP-eq	11086K	4	5.8	304.8M	7.6	83.5	62.43	**62.99**
7 bit	DFP-eq	9700K	4.6	11.5	280.3M	15	82.65	61.86	**62.38**
6 bit	DFP-eq	8314K	5.3	21.9	234.1M	29	81.02	58	**61.19**
5 bit	DFP-eq	6928K	6.4	38.9	161M	51.2	76.99	43.72	**59.54**
[9 9 9 9 6 5 7 9 9]	DFP-lw	9485K	4.7	14.6	289.5M	12.2	83.44	62.90	**62.93**
[9 9 9 9 5 5 5 9 9]	DFP-lw	8490K	5.2	23.2	279.2M	15.3	82.45	62.45	**62.64**
[9 9 9 5 4 4 5 9 9]	DFP-lw	7163K	6.2	36.4	243.4M	26.2	80.76	61.70	**62.29**
[9 6 5 5 3 3 4 7 9]	DFP-lw	5513K	8.0	60.4	133.6M	59.5	73.76	55.69	**58.58**
4 bit	Po2-eq	5543K	8	18.3	255.2M	22.6	79.47	51.27	**60.94**

TABLE V
RESULTS IN COMPARISON TO FLOATING POINT BASELINE OF ALL-CNN FOR CIFAR-10

b_n	Type	W_{mem}[Bit]	CR	Sparsity[%]	non-0 MACs	MAC Sparsity[%]	acc_{tr}[%]	acc_{init}[%]	acc_{M_q}[%]
32 bit	float	43791K	1	0	329.7M	0	96.69	90.83	**90.83**
8 bit	DFP-eq[4]	10947K	4	4.2	303.5M	7.6	96.64	90.65	**90.85**
4 bit	DFP-eq	5474K	8	42.3	140.7M	57.2	94.4	44.47	**88.63**
[8 8 8 5 4 4 3 8 8]	DFP-lw[4]	5929K	7.38	36.8	243M	26	96.49	90.14	**90.81**
[7 7 7 4 4 3 3 7 7]	DFP-lw	5432K	8.1	46.1	203.9M	37.9	96.01	90.07	**90.32**
4 bit	Po2-eq	5474K	8	13.2	255.9M	22.1	96.13	76.73	**90.18**

TABLE VI
RESULTS IN COMPARISON TO FLOATING POINT BASELINE OF ALL-CNN FOR SVHN

b_n	Type	W_{mem}[Bit]	CR	Sparsity[%]	non-0 MACs	MAC Sparsity[%]	acc_{tr}[%]	acc_{init}[%]	acc_{M_q}[%]
32 bit	float	43791K	1	0	329.7M	0	97.70	95.84	**95.84**
4 bit	DFP-eq	5474K	8	43.5	166.8M	49.2	96.27	86.17	**95.37**
[6 4 4 3 3 3 4 5 6]	DFP-lw	4690K	9.3	62.7	126.6M	63.2	96.43	95.03	**95.89**
[5 4 4 3 3 3 3 3 3]	DFP-lw	4276K	10.23	70.5	116.7M	64.4	95.52	89.86	**95.36**
4 bit	Po2-eq	5474K	8	17.0	272.8M	16.9	97.02	91.93	**96.02**

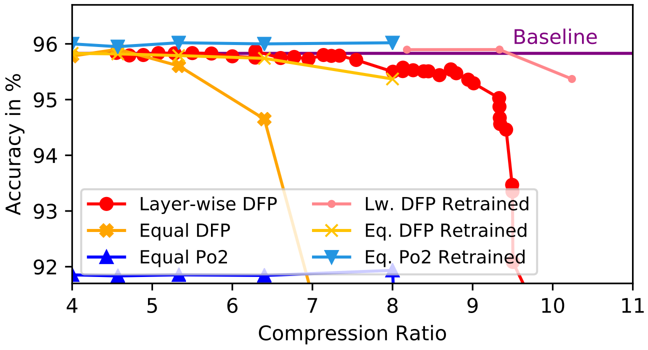

Fig. 11. Results for direct and trained DFP and Po2 quantization with equal bit-widths on ALL-CNN for SVHN. For DFP also Layer-wise Precision Scaling with direct and trained quantization is illustrated.

Comparing the results for the datasets CIFAR-100, CIFAR-10 and SVHN, we can conclude that the attainable Compression Ratio for lossless trained quantization with QR/WQR not only depends on the selected datatype and bit-widths, but also on the selected datasets. While for CIFAR-10 and SVHN, lossless trained quantization achieves Compression Ratio 8.0 and 9.4 respectively, for CIFAR-100 the maximum Compression Ratio is 4.7 for lossless compression. In the case of CIFAR-100 further increasing Compression Ratio up to 8.0 reduces accuracy by 2.09 percentage points. In contrast to this trade-off, stronger compression for CIFAR-10 and SVHN immediately leads to drastic accuracy reduction. We suspect that this difference can be explained by the relation between complexity of the dataset and the selected network architecture. To better understand this relation, in future work the techniques have to be applied for further datasets and network topologies.

While with DFP lossless compression can always be achieved, Po2 quantization can sometimes lead to slight performance degradation. Layer-wise Precision Scaling turns out to be more effective for DFP than for Po2. Po2 quantization still reaches maximum accuracy at uniform 4 bit bit-width while achieving higher accuracy than uniform 4 bit and even 5 bit DFP quantization.

V. COMPARISON WITH RELATED WORK

The proposed method of Weighted Quantization-Regularization presents a novel technique for trained quantization of DNNs. However in some works performing weight-binarization similar regularization methods are used to achieve weights with values '+1' or '−1' [26]. Today trained quantization is mostly performed by stochastic rounding during training [12], [27], [20], [28]. Gupta et al. [28] apply stochastic training for CIFAR-10 dataset to achieve Fixed Point quantization to 16-bit and 12-bit. Their accuracy

[4] 8-bit DFP for activations

is reduced by 0.8 and 4.2 percentage points respectively compared to the floating point baseline. In comparison to that with our method we reach 8-bit DFP quantization without any performance drop.

Courbariaux et al. [12] perform quantization based on stochastic round for 10bit Dynamic Fixed Point weights and activations. Their accuracy drops in comparison to the baseline networks 3.14 percentage points for CIFAR-10 and 2.58 percentage points for SVHN since in contrast to us, they also perform weight-update with 12-bit DFP which decreases comparability. For our CIFAR-10 and SVHN network we experimentally also applied 8-bit DFP for weights and activations, and found that accuracy even increased after QR/WQR retraining. Gysel et al. [20] use their CAFFE-based tool Ristretto for Layer-wise Precision Scaling and fine-tuning with stochastic rounding. On CIFAR-10 their accuracy lies 0.3% below the floating point baseline accuracy, when quantizing not only weights but also activations to 8 bit DFP. By applying Layer-wise Precision Scaling we are able to increase compression ratio from 4 to 7.38 and after QR/WQR-retraining achieve equal to baseline accuracy while inducing higher sparsity due to the stronger compression. Other than fine-tuning with stochastic rounding, Zhou et al. [9] presented an incremental retraining method to perform Power-of-two weight quantization. They achieve lossless 5-bit/4-bit quantization for several DNNs for the ImageNet dataset. Even for lower bit-rates incremental quantization achieves state-of-the-art results. Even though this method seems highly promising, in contrast to our work, it is only verified to work for power-of-two quantization.

VI. Conclusion

We propose Quantization-Regularization (QR) and Weighted QR (WQR) as techniques for improving accuracy after bit-width reductions inflicted by a quantization scheme. WQR/QR allow fine-tuning for weights in any quantization scheme and also works for non-uniform bit-widths (Layer-wise Precision Scaling). For ALL-CNN with the CIFAR-10 benchmark WQR reaches lossless compression up to a ratio of 7.38x with DFP and Layer-wise Precision Scaling. Compared to the 32bit floating point baseline in the All-CNN network, WQR with DFP obtains a weight memory compaction of 8.0×-10.23× and a MAC sparsity of 37.9%-64.4% in the benchmark tasks. For these cases with maximal compaction we observe between 0.48-4.45 percentage points reduction of classification accuracy. A high MAC sparsity benefits HW implementations because it potentially reduces the number of multiply-accumulate operations.

Note, that WQR/QR is not a stand-alone tool, but can be combined with other techniques and is typically complementary to those. We have observed, that it has some limitations when applied to very low bit-widths; hence, its combination with stochastic rounding methods is considered as future work.

Furthermore, we have studied two quantization schemes, Dynamic Fixed Point (DFP) and Power-of-two (Po2), and we find that DFP is usually preferable to Po2 because it is as good as or better than Po2 in most cases, it can reach floating point accuracy when increasing bit-width, and it does not necessarily require retraining (retraining improves accuracy but for Po2 it is absolutely necessary). However, in a few special cases with low bit-width Po2 is slightly better and it might be preferred for HW implementations because it requires only shift operations instead of multiplications. Thus, when an optimized HW implementation is developed, Po2 could be considered as a useful option.

mds
August 26, 2015

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on.* IEEE, 2009, pp. 248–255.

[3] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.

[4] P. Rajpurkar, A. Y. Hannun, M. Haghpanahi, C. Bourn, and A. Y. Ng, "Cardiologist-level arrhythmia detection with convolutional neural networks," *arXiv preprint arXiv:1707.01836*, 2017.

[5] J. Zhang and C. Zong, "Deep neural networks in machine translation: An overview," *IEEE Intelligent Systems*, vol. 30, no. 5, pp. 16–25, 2015.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[7] C. Zhang, P. Li, G. Sun, Y. Guan, B. Xiao, and J. Cong, "Optimizing fpga-based accelerator design for deep convolutional neural networks," in *Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays.* ACM, 2015, pp. 161–170.

[8] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," in *ICLR*, 2016.

[9] A. Zhou, A. Yao, Y. Guo, L. Xu, and Y. Chen, "Incremental network quantization: Towards lossless cnns with low-precision weights," *ICLR*, 2017.

[10] X. Chen, X. Hu, H. Zhou, and N. Xu, "Fxpnet: Training a deep convolutional neural network in fixed-point representation," in *Neural Networks (IJCNN), 2017 International Joint Conference on.* IEEE, 2017, pp. 2494–2501.

[11] D. D. Lin, S. S. Talathi, and V. S. Annapureddy, "Fixed point quantization of deep convolutional networks," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ser. ICML'16. JMLR.org, 2016, pp. 2849–2858. [Online]. Available: http://dl.acm.org/citation.cfm?id=3045390.3045690

[12] M. Courbariaux, Y. Bengio, and J.-P. David, "Training deep neural networks with low precision multiplications," *arXiv preprint arXiv:1412.7024*, 2014.

[13] X. Dong, S. Chen, and S. Pan, "Learning to prune deep neural networks via layer-wise optimal brain surgeon," in *Advances in Neural Information Processing Systems*, 2017, pp. 4860–4874.

[14] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural networks," in *Advances in neural information processing systems*, 2015, pp. 1135–1143.

[15] T.-J. Yang, Y.-H. Chen, and V. Sze, "Designing energy-efficient convolutional neural networks using energy-aware pruning," *arXiv preprint*, 2017.

[16] J. Yu, A. Lukefahr, D. Palframan, G. Dasika, R. Das, and S. Mahlke, "Scalpel: Customizing dnn pruning to the underlying hardware parallelism," in *Proceedings of the 44th Annual International Symposium on Computer Architecture.* ACM, 2017, pp. 548–560.

[17] P. Gysel, "Ristretto: Hardware-oriented approximation of convolutional neural networks," Master's thesis, University of California Davis, 2016.

[18] Y. Zhou, S.-M. Moosavi-Dezfooli, N.-M. Cheung, and P. Frossard, "Adaptive quantization for deep neural network," *arXiv preprint arXiv:1712.01048*, 2017.

[19] S. Han, X. Liu, H. Mao, J. Pu, A. Pedram, M. A. Horowitz, and W. J. Dally, “Eie: efficient inference engine on compressed deep neural network,” in *Computer Architecture (ISCA), 2016 ACM/IEEE 43rd Annual International Symposium on*. IEEE, 2016, pp. 243–254.

[20] P. Gysel, M. Motamedi, and S. Ghiasi, “Hardware-oriented approximation of convolutional neural networks,” *arXiv preprint arXiv:1604.03168*, 2016.

[21] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, “Quantized neural networks: Training neural networks with low precision weights and activations,” *ArXiv*, 2016.

[22] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, “Striving for simplicity: The all convolutional net,” in *ICLR*, 2015.

[23] S. Shin, Y. Boo, and W. Sung, “Fixed-point optimization of deep neural networks with adaptive step size retraining,” *ICASSP*, 2017.

[24] A. Krizhevsky, V. Nair, and G. Hinton. (2009, mar) Cifar-10 and cifar-100 datasets. [Online]. Available: https://www.cs.toronto.edu/ kriz/cifar.html

[25] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, “Reading digits in natural images with unsupervised feature learning,” in *NIPS workshop on deep learning and unsupervised feature learning*, vol. 2011, no. 2, 2011, p. 5.

[26] W. Tang, G. Hua, and L. Wang, “How to train a compact binary neural network with high accuracy?” in *AAAI*, 2017, pp. 2625–2631.

[27] M. Courbariaux, Y. Bengio, and J.-P. David, “Binaryconnect: Training deep neural networks with binary weights during propagations,” in *Advances in neural information processing systems*, 2015, pp. 3123–3131.

[28] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, “Deep learning with limited numerical precision,” in *ICML*, 2015.