DAQ: Channel-Wise Distribution-Aware Quantization for Deep Image Super-Resolution Networks

Cheeun Hong* Heewon Kim* Sungyong Baik Junghun Oh Kyoung Mu Lee
Department of ECE, ASRI, Seoul National University {cheeun914, ghimhw, dsysaik, dh6dh, kyoungmu}@snu.ac.kr

Abstract

Since the resurgence of deep neural networks (DNNs), image super-resolution (SR) has recently seen a huge progress in improving the quality of low resolution images, however at the great cost of computations and resources. Recently, there has been several efforts to make DNNs more efficient via quantization. However, SR demands pixel-level accuracy in the system, it is more difficult to perform quantization without significantly sacrificing SR performance. To this end, we introduce a new ultra-low precision yet effective quantization approach specifically designed for SR. In particular, we observe that in recent SR networks, each channel has different distribution characteristics. Thus we propose a channel-wise distribution-aware quantization scheme. Experimental results demonstrate that our proposed quantization, dubbed Distribution-Aware Quantization (DAQ), manages to greatly reduce the computational and resource costs without the significant sacrifice in SR performance, compared to other quantization methods.

1. Introduction

Image super-resolution (SR), one of the fundamental computer vision tasks, targets for rejuvenating a given low-resolution (LR) input image to its high-resolution (HR) counterpart. The task boasts a wide range of application, including but not limited to medical [17, 39], satellite [3, 45], image processing, military surveillance [46], and automotive industry, giving rise to a great amount of attention from the computer vision community. Together with the interest from the community and the rapid development of deep neural networks (DNNs), SR has recently witnessed an unprecedented breakthrough in performance.

Recent SR works have enjoyed outstanding performance from increasing the network size [13, 28, 32] and employing more complex network structures, such as residual block and spatial attention [48, 47, 32]. However, a network with the larger size and complexity inherently demands heavier computing resources. The deployment of such resource-hungry systems becomes challenging especially in practical applications that often have limited resources available.

Quantization has gained the popularity as one of the promising methods to reduce the amount of required resources for DNNs. Through discretizing the values of network weights [22], features [7], or gradients [50], quantization has significantly reduced the computation loads while minimizing the accuracy loss, particularly on high-level vision tasks, such as classification. SR has yet to benefit from the recent advances in quantization mainly because SR requires pixel-level accuracy [43, 35] and most SR methods have developed ad-hoc network structures [31], such as the absence of batch normalization (BN) [32]. Few recent works have attempted to achieve quantization for SR via either adopting a learnable parameter for each binary convolution weight (BinarySR [35]); employing a bit accumulation mechanism to approximate the full-precision convolution (BAM [43]); and learning the layer-wise max scale value of weights and features (PAMS [31]). However, these methods have neglected the unique distribution characteris-

*equal contribution
Figure 2: Channel-wise feature map distributions of two distinct images (red and blue) in pre-trained (a) ResNet [20] on image classification task and (b) EDSR [32] on image SR task. In SR networks, channels present diverse non-zero distributions that also vary upon the input image.

On the other hand, this work starts with an observation that the recent state-of-the-art SR networks have distinct distributions for each channel, compared to networks that are designed for semantic-level tasks, as illustrated in Figure 2. Further, we observe that the channel distributions vary upon input images.

Upon the observations, we propose a Distribution-Aware Quantization (DAQ) that performs an effective channel-wise quantization. In particular, DAQ computes the distribution of each channel to determine the per-channel quantization transformation (or scaling) parameters that are adaptive to each input. The adaptive scaling approach facilitates effective clipping of outliers that dominate the quantization error. Despite such compelling observation, per-channel quantization for feature maps has yet to be deeply explored, due to the large computational overhead that could nullify the computation benefits of quantization [38]. We adjust the process of quantized convolution to significantly reduce the computational overhead, enabling DAQ to effectively and efficiently quantize SR networks.

Overall, our contributions can be summarized as follows:

- To the best of our knowledge, we present the first feasible channel-wise quantization method for image super-resolution networks in ultra-low precision with marginal accuracy loss.
- Our scheme is directly applicable to existing SR networks without any architectural modification or specialized training scheme.
- Experimental results show that the proposed DAQ outperforms state-of-the-art SR quantization methods, yielding higher PSNR with less resource consumption.
- DAQ effectively reduces quantization error, achieving accurate quantization on pre-trained SR networks even without retraining.

2. Related Work

2.1. Image super-resolution

Image super-resolution (SR) is a conventional computer vision problem that increases the resolution of an image while restoring its structural details. Convolutional neural network (CNN) based approaches [22, 48, 13, 28] have achieved a great performance improvement in this field. However, CNNs require heavy computational costs and memory footprints that greatly limit their applicability on resource-constrained devices. To alleviate this problem, recent works [14, 40, 23, 9, 18, 27] have focused on improving the model efficiency in terms of FLOPs, parameters, and run-time by designing novel lightweight network architectures. However, these networks still consume a large amount of resources since the convolution layers of CNNs operate in 32-bit floating-point (FP32).

2.2. Network quantization

Network quantization is a promising research area that aims to map 32-bit floating-point (FP32) values of feature maps and weights to lower bits (or precision) values. However, quantized networks usually suffer from severe accuracy degradation. Many works tackle this problem by optimizing the uniform [10, 50, 7] or non-uniform [19, 44, 26, 16] intervals between the quantized values. This work focuses on quantization with the uniform interval that can be accelerated in the hardware with simple arithmetic pipelines. Recently, research on specializing the precision of each part of the network [6, 34] and learning the quantization method without network retraining [29, 49, 2, 8, 37, 4] have been actively conducted. However, these approaches focus on semantic-level tasks, such as image classification, which may be an easier task for network quantization than SR, which requires pixel-level accuracy [35, 43].

To this end, few recent works have focused on network quantization specifically for SR. BinarySR [35] presented the first attempt to quantize SR networks, however for weights only. BAM [43] and BTM [25] either modified the network architecture with multiple skip connections or utilized specialized convolution layers to quantize both weights and feature maps. Recently, Hwang et al. [24] proposed an SR network compression method that jointly performs channel pruning and different-bit quantization for each layer. The conventional approaches have neglected the first and last layers, but FQSR [41] quantized all the weights and feature maps of SR networks. PAMS [31] may be considered to be similar to our proposed method in that it learns quantization ranges with uniform intervals, however for each layer in contrast to our method that performs channel-wise quantization, which aligns well with the observation from Figure 2. Furthermore, the works mentioned above have proposed specialized training schemes such as
knowledge distillation, self-supervised learning, and gradient update rules. Although these custom network architectures and training schemes may reduce performance degradation to some extent, their ad-hoc techniques or specialized networks generally make it difficult to apply quantization methods to various networks.

On the other hand, we investigate a quantization range tailored to each channel in SR networks, while existing SR-based methods quantized all channels in a feature map with the same range. Very few recent works have managed to achieve per-channel quantization, however for semantic-level tasks [42]. To the best of our knowledge, our method is the first to achieve per-channel scaled quantization that manages to greatly reduce performance degradation.

3. Method

3.1. Motivation

Network quantization for image super-resolution (SR) has suffered from substantial performance degradation that occurs due to quantization of feature map values [35]. Most previous works [43, 25] tackle this problem by manually designing architectures and heuristic training schemes while their quantization functions are fixed for a feature map. Our work started by investigating the characteristics of the feature maps in SR network.

The distribution of each channel in SR network (See Figure 2(b)) has diverse values for its mean and variance. Moreover, channel distributions are observed to be different for each input image. This observation has led to our assumption that quantizing each channel in different functions may effectively reduce the quantization error in SR networks while it may not be as beneficial in semantic-level tasks. To this end, we formulate a quantized convolution method as described in following sections.

3.2. Distribution-Aware Quantization (DAQ)

3.2.1 Overview

Quantization function is a sequence of processes that discretizes a FP32 tensor to lower precision. General pipeline of quantization on neural networks are processed as following. Input feature and weight are respectively scaled (or transformed) to integer range, and then discretized. Then, the quantized convolution is performed with discretized input feature of integer values and discretized weight of integer values. Then the convolution output is descaled back via de-transformation, which is a costly procedure for channel-wise quantization. Our proposed method follows a general pipeline of quantization function, except that a standard de-transformation is replaced with our newly proposed efficient de-transformation scheme. Overall, our DAQ convolution operates in a sequence of the following processes, as summarized in Figure 3(a):

- Channel-wise quantized convolution (Figure 3(b)).
- Per-channel transformation and discretization of input feature maps (Section 3.2.2).
- Transformation and discretization of convolution filter weights (Section 3.2.3).
- Per-channel convolution over quantized weights and quantized input feature maps (dot product operation in a sliding window manner: multiplication of the two and kernel-wise sum).
- Proposed computationally efficient per-channel de-transformation (Section 3.2.4).
- Element-wise sum of outputs of the channel-wise convolution ($\sum_C$) in low-bit precision (Section 3.2.4).
- De-transformation of the convolution output to full precision FP32.

3.2.2 Quantization of feature map

As we discussed above, we define a quantization function specialized in each channel. Formally, given a feature map $x \in \mathbb{R}^{B \times C \times H \times W}$, we quantize $b$-th element of the mini-batch and $c$-th channel $x_{b,c} \in \mathbb{R}^{H \times W}$, where $B, C, H,$ and $W$ denote the mini-batch size, number of channels, height and width of the feature map, respectively. In the subsequent sections, we drop the subscript $b$ for notational simplicity.
Adaptive transformer  Existing quantization works for SR \[31\] often adopt linear scaling before discretizing the FP32 values to integer values. The linear scaling might be a good transformer for tensor with a zero-mean and bell-shaped distribution, like the channel-wise feature of ResNet visualized in Figure 2(a). However, the distribution of each channel, as shown in Figure 2(b), has a non-zero mean and even varies across different input images, suggesting that an adaptive transformer might be a better alternative for each channel distribution. We define the transformer \(f(\cdot)\) using a standardization function to normalize values of \(x_c\) before discretization. \(f(\cdot)\) is formally given by,

\[
\hat{x}_c = f(x_c) = \frac{x_c - \mu_c}{\sigma_c \cdot s(n)},
\]

where \(\mu_c\) and \(\sigma_c\), respectively, denote the average and the standard deviation of the \(c\)-th channel \(x_c\) and \(s(n)\) determines an uniform interval between quantized values by the bit-width \(n\). Table 1 presents the actual values of \(s(n)\) which represent the optimal quantization for Gaussian distribution. We interchangeably refer to statistics \(\mu_c\) and \(\sigma_c\) as quantization transformation parameters as they are used to perform transformation during quantization process.

Adaptive discretizer  In general, a discretizer clips the values of \(\hat{x}_c\) to a range given by the quantization bit-width \(n\), followed by the mapping of the values to integers. Our adaptive discretizer is formulated as follows:

\[
\hat{x}_c^q = d(\hat{x}_c) = \lfloor g(\hat{x}_c) \rfloor,
\]

where \(g(\cdot)\) is the clamp function of a range \((-2^{n-1} + \alpha, 2^{n-1} + \alpha]\) and \(\lfloor \cdot \rfloor\) rounds up a value to the nearest higher integer. The shifting parameter \(\alpha\) is a function of \(s(n), \mu_c, \) and, \(\sigma_c\) formally given by,

\[
\alpha = \begin{cases} 
\max(2^{n-1} - \frac{\mu_c}{\sigma_c \cdot s(n)} - 1, 0) & \text{after ReLU} \\
0 & \text{otherwise}
\end{cases}
\]

allowing the minimum quantized value (of original scale) to be zero after ReLU function. \(d(\cdot)\) is distribution-dependent due to its dependency on channel statistics, \(\mu_c\) and \(\sigma_c\).

Figure 4 illustrates the effectiveness of our distribution-aware quantization (DAQ). Our method reduces quantization error by a large margin for diverse channel distributions. We use \(x_c^q = f^{-1}(\hat{x}_c^q)\) for visualization, where \(\hat{x}_c^q\) denotes low-bit quantized integer and \(f\) is the transform function of each method.

### 3.2.3 Quantization of weight

Weight distributions over filters are empirically observed to have less degree of variance than feature maps in SR networks. Hence, a single quantization function is applied to the weight of each convolution \(w \in \mathbb{R}^{C \times C_{out} \times K \times K}\), where \(C\) and \(C_{out}\) denote the number of input and output chan-
nels, respectively, and \( K \) is the kernel size of the convolution filter. The distribution of weight values are observed to be bell-shaped with zero-mean, and is independent of the input image. Thus, for quantizing the weight \( w \), we assume the mean of \( w \) to be 0, which simplifies a transformer for weight quantization \( f(w) = \frac{w}{\sigma_w \cdot s(n)} \), where \( \sigma_w \) denotes the standard deviation of \( w \). The shifting parameter \( \alpha \) of the weight discretizer is set to 0 since the weight values are not zeroed out by activation.

### 3.2.4 Adaptive de-transformer

**Standard per-channel de-transformer** Once per-channel quantized convolution response is obtained, normally one would de-scale (or de-transform) each per-channel convolution output back to its full-precision before the summation of channel-wise convolution outputs along the channel dimension. However, de-transformation of each channel-wise convolution output to full precision (FP32) using quantization transform parameters (in this case, channel-wise statistics \( \mu_c \) and \( \sigma_c \)) will result in a large computational overhead of the overall quantized convolution process, due to the FP32 summation of de-transformed convolution responses along the channel dimension.

**Efficient per-channel de-transformer** Recently, Dai et al. [11] reduces the overhead by decomposing scaling factors into channel-specific quantized values and a global FP32 value for the low-bit addition in image classification networks. Similarly, our solution achieves efficient implementation but with the distribution awareness. We propose to perform quantization on statistics \( \mu \) and \( \sigma \) to obtain quantized statistics, allowing us to perform the summation of the channel-wise convolution responses in low-bit precision. Formally, recall \( \mu_c \) and \( \sigma_c \), the mean and the standard deviation of the \( c \)-th channel in Equation (1), as the \( c \)-th elements of \( \mu \in \mathbb{R}^C \) and \( \sigma \in \mathbb{R}^C \). We quantize the transformation parameters, thus named QQ. The quantized transform parameters using Equation (1) and (2) approximate their FP32 versions:

\[
\mu_c \approx \mu_c^q \equiv \sigma \cdot s(m) \cdot \tilde{\mu}_c^q + \mu_c, \\
\sigma_c \approx \sigma_c^q \equiv \sigma \cdot s(m) \cdot \tilde{\sigma}_c^q + \mu_c,
\]

where \( m \) is the bit-width of QQ, and \( \sigma \) and \( \mu_c^*, \sigma^*, \mu_c^* \in \{\mu, \sigma\} \), denote the mean and the standard deviation of \( \mu \) and \( \sigma \). Note that \( \sigma^* \) and \( \mu^* \) are FP32 values which are independent to channels while \( \tilde{\sigma}_c^q \) and \( \tilde{\mu}_c^q \) are low-bit integers specific to the channels. Using the quantizing quantization parameters, our quantized channels become,

\[
x_c^q = \sigma_c \cdot s(n) \cdot \tilde{x}_c^q + \mu_c \\
\approx \sigma_c^q \cdot s(n) \cdot \tilde{x}_c^q + \mu_c^q.
\]

As a result, we can place the de-transformation process after the summation of the channel-wise convolution and eliminate the potential possibility of a computation overhead that could have occurred from channel-wise quantized convolution (please refer to Section 3.3.1 and the supplementary document).

### 3.3. Hardware Implementation

#### 3.3.1 Computation of per-channel de-transformer

As discussed in Section 3.2.4, the computational overhead from the standard per-channel de-transformer is largely reduced by our efficient adaptive de-transformation process. The approximation by Equation (4) replaces a large proportion of high-precision operations with low-precision operations by modifying arithmetic pipelines. For instance, the sum of quantized channels \( \sum_c x_c^q \) should be operated in FP32 \( \sum_c (\sigma_c \cdot s(n) \cdot \tilde{x}_c^q + \mu_c) \) due to the channel-specific mean and standard deviation of FP32 values. However, the sum of the approximated quantized channels \( \sum_c \sigma_c^q \cdot s(n) \cdot \tilde{x}_c^q + \mu_c^q \) can expand the low-bit channel sum \( \sum_c \sigma_c^q \cdot s(n) \cdot s(m) \sum_c \tilde{x}_c^q + \tilde{\mu}_c^q \cdot s(m) \sum_c \tilde{\sigma}_c^q + \tilde{\mu}_c^q \).

#### 3.3.2 FP32 Arithmetic Computation

The proposed method adaptively quantizes the value of interest with respect to its distribution by using standardization function as the de-transformer. The computation complexity of Equation (1) can be specified into the complexity for calculating the mean and the standard deviation, and for the de-transformer function itself, of which are both \( O(n) \). Specifically, the overall complexity of Equation (1) is \( C(5HW + 3) \) for the feature map and \( 3K^2C_{in}C_{out} \) for the weight. Weight standardization is independent of input images and thus occurs only once before the inference, which allows us to measure its computational complexity only once. On the other hand, the complexity of feature standardization is calculated per test image.

### 3.4. Training

Recently, training quantized networks (or quantization-aware training) is a well-known approach for accuracy recovery. The back-propagation in quantized networks calculates the gradient of loss \( \ell \) with respect to \( w^q \). Quantization function is generally not differentiable, and thus not possible to directly apply back-propagation. Consequently, we adopt the straight-through estimator [50] to approximate the gradients calculation. We approximate the partial gradient \( \frac{\partial \ell}{\partial w^q} \) with an identity mapping \( \frac{\partial \ell}{\partial w^q} \approx 1 \). Accordingly, \( \frac{\partial \ell}{\partial w^q} \) can be approximated by

\[
\frac{\partial \ell}{\partial w^q} = \frac{\partial \ell}{\partial w} \frac{\partial w}{\partial w^q} \approx \frac{\partial \ell}{\partial w}
\]
4. Experiments

4.1. Experimental setup

Models and datasets To verify the effectiveness and the generalizability of the proposed quantization method, we conduct experiments on widely used SR networks, including EDSR [32], RDN [48] and SRResNet [30]. We use pre-trained models publicly available at code [31]. We retrain the quantized models on DIV2K dataset [1], which consists of 800 2K-resolution images. We finetune the DAQ-applied EDSR model using the training settings from [32], where a batch size is set to be 4, a learning rate is $10^{-4}$, and the number of updates is $3 \times 10^5$. DAQ-applied RDN is finetuned using $3 \times 10^4$ updates with batch size 16, and initial learning rate $10^{-4}$. SRResNet-DAQ is finetuned with an initial learning rate of $10^{-5}$ for $6 \times 10^5$ iterations while we follow the other training settings from [30]. We evaluate the quantized models on four standard benchmark datasets: Set5, Set14, B100, and Urban100.

Implementation details Our proposed quantization method is directly applied to existing SR models without modifying the network architecture. Along with the majority of quantization works, we quantize weights and feature maps in all layers except the first convolution and majority of quantization works, we quantize weights and output modifying the network architecture. Along with the method is directly applied to existing SR models with an initial learning rate of $10^{-4}$ and the number of updates is $3 \times 10^5$. DAQ-applied RDN is finetuned using $3 \times 10^4$ updates with batch size 16, and initial learning rate $10^{-4}$. SRResNet-DAQ is finetuned with an initial learning rate of $10^{-5}$ for $6 \times 10^5$ iterations while we follow the other training settings from [30]. We evaluate the quantized models on four standard benchmark datasets: Set5, Set14, B100, and Urban100.

BOPs The number of floating-point operations (FLOPs) is a conventional measure for the computational complexity of full-precision models. However, a low-precision model often consists of operations with inputs of different low bit-widths. To take the low-precision operations into account, we weight the number of bit operations by multiplying $n \times m$, where $n$ and $m$ denote the bit-widths of two different inputs, respectively for each operation. The weighted number of bit operations is referred to as BOPs. The reported BOPs are measured to generate a FHD image (1920×1080).

Energy A quantized network boasts energy efficiency as lower-precision operations consume less energy. We adopt the approximation of energy consumption introduced in [21] and compute it for generating a FHD image.

Memory Another benefit of network quantization is the reduction in memory requirement for model storage. The $n$-bit weights require $n/32$ amount of memory of FP32 weights.

4.2. Comparisons with SotA methods

SR quantization methods To demonstrate the effectiveness of our channel-wise quantization paradigm, we compare our method with PAMS [31] and BinarySR [35], which adopt per-layer quantization ranges. Without architecture changes of existing networks, Table 2 presents outperforming performances of our method (DAQ) on most measures and all benchmarks. Specifically, RDN-DAQ achieves 1 dB higher PSNR than RDN-PAMS on Urban100 with less resource consumption of BOPs, Energy, and Memory. SRResNet-DAQ requires only 50% BOPs of SRResNet-BinarySR of which weights are 1-bit while outperforming 1.51 dB on Set5.

On the other hand, BTM [25] and BAM [43] employ specialized architectures or training schemes for low-precision SR networks. Our proposed quantization method is orthogonal to these techniques, providing additional performance gains, as shown in the supplementary document.

Comparison with per-channel quantization methods To show the importance of distribution awareness in per-channel quantization function as proposed in this work, we compare DAQ with the SR models quantized by other per-channel quantization methods LinQ [29] and VS-Quant [11], which are however tailed for classification and do not consider per-channel distribution in contrast to DAQ. Table 2 shows that EDSR-DAQ outperforms EDSR-LinQ and EDSR-VS-Quant over than 1 dB on Urban100 while consuming less resources, validating the benefits of formulating the channel-wise quantization function with distribution statistics.

Qualitative comparisons Figure 5 presents the output images from SotA models compared in Table 2. Our method generates a visually clean output image, while RDN-PAMS [31] and SRResNet-BinarySR [35] suffer from a checkerboard artifact. Interestingly, EDSR-LinQ generates a sharper image than EDSR-PAMS with lower PSNR scores. The result of SRResNet-DAQ has a more realistic structure (at the bottom of the image).

Comparison with methods without retraining To further demonstrate the effectiveness of DAQ in dynamically reducing the quantization error, we evaluate DAQ and other quantization methods without fine-tuning, as shown in Table 3. DAQ is shown to greatly reduce the quantization error to a large extent compared to other methods. EDSR-LinQ calculates min/max values for each channel-wise tensor during inference, which is the same process in the previous section. DFQ [37] controls weight ranges and biases of the pre-trained models during inference, proposed for image classification. EDSR-DFP is a re-implemented version of DFQ on EDSR, based on the publicly available codes. EDSR-DAQ outperforms the compared methods over 1.4 dB with less computation costs. The results corroborate our
Table 2: Comparisons of existing quantization methods on EDSR [32], RDN [48] and SRResNet [30] of scale 4.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>BOPs</th>
<th>Energy</th>
<th>Memory</th>
<th>Parameters</th>
<th>PSNR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>w a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EDSR</td>
<td>32 32</td>
<td>10019.3 T</td>
<td>22504.3 mJ</td>
<td>172.3 MB</td>
<td>43.1M</td>
<td>32.46</td>
</tr>
<tr>
<td>EDSR-PAMS [31]</td>
<td>4 4</td>
<td>351.3 T</td>
<td>366.7 mJ</td>
<td>40.2 MB</td>
<td>43.1M</td>
<td>31.59</td>
</tr>
<tr>
<td>EDSR-LinQ [29]</td>
<td>2 2</td>
<td>878.4 T</td>
<td>2616.6 mJ</td>
<td>30.7 MB</td>
<td>43.1M</td>
<td>31.08</td>
</tr>
<tr>
<td>EDSR-VS-Quant [11]</td>
<td>2 2</td>
<td>425.6 T</td>
<td>555.7 mJ</td>
<td>31.0 MB</td>
<td>43.1M</td>
<td>31.10</td>
</tr>
<tr>
<td>EDSR-DAQ (ours)</td>
<td>2 2</td>
<td>267.8 T</td>
<td>328.6 mJ</td>
<td>30.7 MB</td>
<td>43.1M</td>
<td>32.05</td>
</tr>
<tr>
<td>RDN</td>
<td>32 32</td>
<td>5636.0 T</td>
<td>12659.0 mJ</td>
<td>89.2 MB</td>
<td>22.3M</td>
<td>32.24</td>
</tr>
<tr>
<td>RDN-PAMS [31]</td>
<td>4 4</td>
<td>204.8 T</td>
<td>222.3 mJ</td>
<td>12.3 MB</td>
<td>22.3M</td>
<td>30.44</td>
</tr>
<tr>
<td>RDN-LinQ [29]</td>
<td>2 2</td>
<td>474.5 T</td>
<td>1044.3 mJ</td>
<td>6.9 MB</td>
<td>22.3M</td>
<td>30.90</td>
</tr>
<tr>
<td>RDN-VS-Quant [11]</td>
<td>2 2</td>
<td>239.1 T</td>
<td>311.6 mJ</td>
<td>6.9 MB</td>
<td>22.3M</td>
<td>31.16</td>
</tr>
<tr>
<td>RDN-DAQ (ours)</td>
<td>2 2</td>
<td>168.8 T</td>
<td>220.1 mJ</td>
<td>6.9 MB</td>
<td>22.3M</td>
<td>31.61</td>
</tr>
<tr>
<td>SRResNet</td>
<td>32 32</td>
<td>324.6 T</td>
<td>728.2 mJ</td>
<td>6.1 MB</td>
<td>1.6M</td>
<td>31.94</td>
</tr>
<tr>
<td>SRResNet-BinarySR [35]</td>
<td>1 32</td>
<td>38.9 T</td>
<td>58.3 mJ</td>
<td>1.5 MB</td>
<td>1.6M</td>
<td>30.16</td>
</tr>
<tr>
<td>SRResNet-LinQ [29]</td>
<td>2 2</td>
<td>37.8 T</td>
<td>58.8 mJ</td>
<td>1.7 MB</td>
<td>1.6M</td>
<td>31.44</td>
</tr>
<tr>
<td>SRResNet-VS-Quant [11]</td>
<td>2 2</td>
<td>22.5 T</td>
<td>41.8 mJ</td>
<td>1.7 MB</td>
<td>1.6M</td>
<td>31.24</td>
</tr>
<tr>
<td>SRResNet-DAQ (ours)</td>
<td>2 2</td>
<td>19.1 T</td>
<td>38.5 mJ</td>
<td>1.7 MB</td>
<td>1.6M</td>
<td>31.67</td>
</tr>
</tbody>
</table>

Figure 5: Qualitative evaluation of existing quantization methods on EDSR [32], RDN [48] and SRResNet [30] of scale 4, as in Figure 5. Evaluation is done on ‘barbara’ of Set14 and ‘img060’ and ‘img097’ of Urban100. Quantitative measures are also denoted (PSNR / BOPs).
Table 3: Comparison of quantization methods without retraining on pre-trained EDSR [32] of scale 4.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>BOPs</th>
<th>Energy</th>
<th>Memory</th>
<th>PSNR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Set5</td>
<td>Set14</td>
<td>B100</td>
</tr>
<tr>
<td>EDSR</td>
<td>32, 32</td>
<td>10019.3 T</td>
<td>22504.3 mJ</td>
<td>172.3 MB</td>
<td>32.46</td>
</tr>
<tr>
<td>EDSR-LinQ</td>
<td>2, 2</td>
<td>878.4 T</td>
<td>2616.6 mJ</td>
<td>30.7 MB</td>
<td>30.26</td>
</tr>
<tr>
<td>EDSR-DFQ</td>
<td>4, 4</td>
<td>351.3 T</td>
<td>366.7 mJ</td>
<td>40.2 MB</td>
<td>31.20</td>
</tr>
<tr>
<td>EDSR-DAQ</td>
<td>2, 2</td>
<td>267.8 T</td>
<td>328.6 mJ</td>
<td>30.7 MB</td>
<td>31.42</td>
</tr>
</tbody>
</table>

observations from channel distributions (Figure 2) and assumptions that the per-channel and distribution-aware quantization is effective in SR networks.

Table 4: Ablation study of the proposed method on feature map quantization. 2-bit quantization is evaluated on EDSR.

<table>
<thead>
<tr>
<th>Granularity</th>
<th>Distribution aware scaling</th>
<th>QQ</th>
<th>BOPs</th>
<th>PSNR (Urban100)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Layer</td>
<td>✗ ✗</td>
<td></td>
<td>83.5 T</td>
<td>6.99 dB</td>
</tr>
<tr>
<td>Layer</td>
<td>✓ ✗</td>
<td>90.1 T</td>
<td>7.17 dB</td>
<td></td>
</tr>
<tr>
<td>Channel</td>
<td>✗ ✗</td>
<td>878.4 T</td>
<td>7.26 dB</td>
<td></td>
</tr>
<tr>
<td>Channel</td>
<td>✓ ✗</td>
<td>889.3 T</td>
<td>25.98 dB</td>
<td></td>
</tr>
<tr>
<td>Channel</td>
<td>✓ ✓</td>
<td>267.8 T</td>
<td>25.97 dB</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Ablation study of the proposed method on weight quantization. Evaluation done on EDSR.

<table>
<thead>
<tr>
<th>Distribution aware scaling</th>
<th>Granularity</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>4-bit</td>
</tr>
<tr>
<td>✓</td>
<td>Layer</td>
<td>26.48</td>
</tr>
<tr>
<td></td>
<td>Output channel</td>
<td>26.50</td>
</tr>
<tr>
<td></td>
<td>Input channel</td>
<td>26.47</td>
</tr>
<tr>
<td></td>
<td>Kernel</td>
<td>26.45</td>
</tr>
<tr>
<td>✗</td>
<td>Layer</td>
<td>26.20</td>
</tr>
<tr>
<td></td>
<td>Output channel</td>
<td>26.51</td>
</tr>
<tr>
<td></td>
<td>Input channel</td>
<td>26.50</td>
</tr>
<tr>
<td></td>
<td>Kernel</td>
<td>26.54</td>
</tr>
</tbody>
</table>

Table 6: Ablation study on various distribution assumptions.

<table>
<thead>
<tr>
<th>EDSR (×4)</th>
<th>Gaussian</th>
<th>Uniform</th>
<th>Laplacian</th>
<th>Gamma</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-bit</td>
<td>26.51</td>
<td>24.52</td>
<td>26.21</td>
<td>26.25</td>
</tr>
<tr>
<td>1-bit</td>
<td>24.20</td>
<td>23.64</td>
<td>23.82</td>
<td>23.71</td>
</tr>
</tbody>
</table>

4.3. Ablation study

Feature map quantization To verify the efficacy of our quantization scheme in practical SR networks, we do ablation on each of our contribution. In Table 4, granularity represents the tensor to be quantized, where “Layer” quantizes the feature map with a single quantization range and “Channel” quantizes each input channel with a specialized range. The non-distribution-aware scaling uses min/max values to normalize distributions. As shown in Table 4, channel-wise and distribution-aware property are both important for accurate low-bit quantization, but suffer from heavy computational resources. Utilizing quantization function for the quantization transform parameters (QQ) significantly reduces BOPs, while maintaining a high PSNR.

Weight quantization Though our main contributions in the method are focused on feature map quantization, we also evaluate the effectiveness of the proposed weight quantization in Table 5. Distribution awareness in quantization plays an key role in maintaining the performance, especially in lower-precisions. Layer-wise quantization presents a better trade-off between computational efficiency and performance among different extents of granularity.

Distribution assumptionScaled step size $s(n)$ in quantization function is a pre-determined hyper-parameter. Many quantization works [5, 15] select different scaled step size for different bit-widths assuming a certain distribution of the tensor to be quantized. We follow the Gaussian assumption in the main experiments, while results on different distribution assumptions, including Gaussian, Uniform, Laplacian, and Gamma distribution are shown in Table 6.

5. Conclusion

We propose an efficient yet effective quantization method for deep SR networks. Based on the observation that each channel of SR networks has distinct distribution, we introduce a channel-wise Distribution-Aware Quantization (DAQ) method. However, channel-wise quantization with standard de-transformation incurs a large computational overhead. Instead, we utilize an efficient de-transformation by quantizing quantization parameters, reducing the computational costs of channel-wise quantization significantly. Our proposed scheme reduces the quantization error to a large extent, achieving high performance even without quantization-tailored training. Evaluations on various SR networks demonstrate the outstanding performance of DAQ, compared to other quantization methods, with similar amount of resource consumptions.

Acknowledgment This work was supported in part by IITP grant funded by the Korean government (MSIT) [NO.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)]
References


