

PAMS: Quantized Super-Resolution via Parameterized Max Scale

Huixia Li^{1†}, Chenqian Yan^{1†}, Shaohui Lin², Xiawu Zheng¹,
Baochang Zhang³, Fan Yang⁴, Rongrong Ji^{15*}

¹Media Analytics and Computing Lab, Department of Artificial Intelligence,
School of Informatics, Xiamen University, China.

² National University of Singapore, Singapore. ³ Beihang University, China.

⁴ Huawei Technologies Co., Ltd. ⁵Peng Cheng Laboratory.

{hxlee,zhengxiawu}@stu.xmu.edu.cn, {im.cqyan,shaohuilin007}@gmail.com,
bczhang@buaa.edu.cn, yangfan74@huawei.com, rrji@xmu.edu.cn

Abstract. Deep convolutional neural networks (DCNNs) have shown dominant performance in the task of super-resolution (SR). However, their heavy memory cost and computation overhead significantly restrict their practical deployments on resource-limited devices, which mainly arise from the floating-point storage and operations between weights and activations. Although previous endeavors mainly resort to fixed-point operations, quantizing both weights and activations with fixed coding lengths may cause significant performance drop, especially on low bits. Specifically, most state-of-the-art SR models without batch normalization have a large dynamic quantization range, which also serves as another cause of performance drop. To address these two issues, we propose a new quantization scheme termed PArameterized Max Scale (PAMS), which applies the trainable truncated parameter to explore the upper bound of the quantization range adaptively. Finally, a structured knowledge transfer (SKT) loss is introduced to fine-tune the quantized network. Extensive experiments demonstrate that the proposed PAMS scheme can well compress and accelerate the existing SR models such as EDSR and RDN. Notably, 8-bit PAMS-EDSR improves PSNR on Set5 benchmark from 32.095dB to 32.124dB with $2.42\times$ compression ratio, which achieves a new state-of-the-art.

Keywords: Super Resolution · Network Quantization

1 Introduction

Single image super-resolution (SISR) aims to recover a high-resolution (HR) image from the corresponding low-resolution (LR) one, which has been a research

[†] Equal contribution.

* Corresponding author.

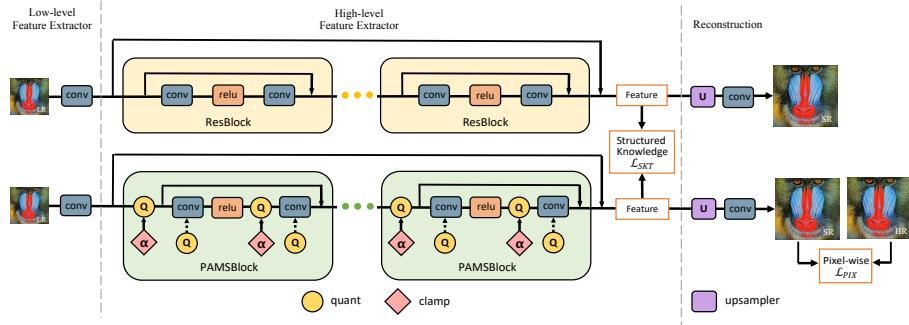


Fig. 1. The framework of our approach. The super-resolution operation is split into three modules, *i.e.*, low-level feature extractor, high-level feature extractor and reconstruction. We deploy PAMS with different α on each activation layer in the high-level feature extractor. *quant* denotes the quantization operation and *clamp* represents the clamp function of quantization. Dash lines denote the weights are quantized with the maximum. Here, we illustrate EDSR as backbone.

hot spot in computer vision for decades. Coming with the advances of deep learning, deep convolutional neural networks (DCNNs) [7, 18, 23] have dominated SR in recent years. These networks commonly use an extraction module to extract a series of feature maps from the LR image, cascaded with the up-sampling module, which stepwisely increases the resolution to reconstruct the HR image.

As one of the pioneering works for deep learning based SR, Dong *et al.* [7] introduce three convolution layers to achieve high visual perception. After that, Kim *et al.* [18] design a deep network VDSR by stacking 20 convolutional layers. Subsequent works mainly resort to increasing the network depth to improve SR performance. For instance, Lim *et al.* [23] propose the enhanced deep residual networks (*e.g.* EDSR and MDSR) and remove batch normalization (BN) [16] to reduce the memory consumption, which however still requires at least 64 convolution layers (more than 160 layers for MDSR). A channel attention mechanism equipped into the RCAN model [38] requires more than 400 layers with about 30B FLOPs and 13M parameters. Such significant computation and memory overheads restrict their applications in scenarios where only limited memory and computation resources are available. Consequently, compressing deep SR networks has attracted increasing attention recently [29].

Beyond SR, neural network compression and acceleration have been widely studied in the literature. Representative works include parameter pruning [11–13, 27, 26, 22], low-rank approximation [6, 24, 25], compact networks [33, 28], knowledge distillation (KD) [14, 32], neural architecture search (NAS) [43, 40] and quantization [5, 17]. Considering the unique structures such as EDSR [23] and RDN [39] in SR, it is by nature to leverage quantization schemes to accelerate and compress SR networks, *i.e.*, by converting full-precision weights [17], activations [2], and gradients [41] to low bits.

Ma *et al.* [29] first apply weight quantization to compress SR models, which merits in low on-device storage. However, the computational complexity is still significantly high, since full-precision activations are still used. In contrast, directly applying weight quantization to activations will incur significant accuracy drop in general SR tasks without using batch normalization, due to the high dynamic quantization range. On one hand, the work in [23] has shown that normalizing features on SR models limits the network’s representation power. Since BN layers make the features to be smooth, which results in the blurred reconstructed HR images with artifacts. To this end, recent SOTA SR models (*e.g.* EDSR [23], RDN [39]) have already removed BN layers to obtain better reconstructed HR images. On the other hand, the absence of BN causes a severe dynamic range problem when quantizing the activations by using the SOTA quantization methods [17, 3]. For example, the work [17] simply set the upper scale of activations to their max value, which causes significant performance degeneration in SR task. This is due to the fixed max scale that may be an outlier as the upper scale. Although Choi *et al.* [3] propose PACT to clip and quantize activations by learnable parameters, it only concentrates on the positive range while neglecting the gradient information in the negative range. In addition, the novel regularization term [4] is added to automatically learn quantized controlling parameters and then obtain an accurate low-precision model. However, it leads to the increase of additional computation burdens and memory footprint, which is not runtime friendly for practical applications.

To address the above issues, a novel quantization scheme, termed PArameterized Max Scale (PAMS), is proposed to compress and accelerate SR models. Different from the previous works that focus on quantizing activations in a fixed manner, PAMS adaptively explores the upper bound of quantization range based on the gradients using a trainable clamp function, which significantly improves the model generality. Furthermore, structured knowledge transfer (SKT) is introduced to transfer structured knowledge from the full-precision network to the quantized one, which enables the latter to gain better visual perception. Fig. 1 presents the flowchart of our method. We first replace each basic block in the SR model with PAMS block. In each PAMS block, weights are quantized before they are convolved with the inputs and activations are quantized after the outputs of convolutional layer with its own learnable max scale. To further improve the performance of the quantized model, we align the high-level features between full-precision model and the corresponding low-precision quantized one among pixels. Finally, we employ stochastic gradient descent (SGD) method to minimize the objective function, which leverages the distillation loss to pixel-wise loss.

We evaluate our method on several benchmarks over widely-used deep SR models like EDSR [23] and RDN [39]. Quantitative and qualitative results demonstrate that PAMS can well quantize various SR models with a significantly high compression ratio, as well as nearly identical accuracy to the full-precision SR models. The proposed PAMS also well outperforms most existing alternatives such as Dorefa [41], Tensorflow Lite [17] and PACT [3]. For instance, on BSD100,

the 4-bit PAMS-EDSR outperforms 4-bit Dorefa-EDSR by 0.828dB with a scale factor of $\times 4$. Extended experiments also show that SKT is more effective for the quantized SR models with lower-bit operations.

2 Related Work

Deep SR models with light weights. Most recent SR models are built based upon DCNNs, for instance, MDSR [23] and RDN [39]. Such networks are typically deep with heavy computation cost and memory footprint, which restrict their applications in resource-limited devices. Recent advances in SR network compression mostly focus on redesigning light-weight networks. For instance, DRRN [34] and DRCN [19] have been proposed to share parameters for reducing network parameters. However, the cost of computation and memory storage in these networks are still very large, due to the floating-point operations during inference and the sufficient parameters to ensure the model capability.

Network quantization. Previous works in network quantization mainly focus on quantizing weights [?], while maintaining the full-precision activations to ensure the model performance. Joint quantization of activations and weights are explored in HWGQ [2] and PACT [3]. However, these methods mainly concentrate on object classification [20, 10], which is easier than the complex pixel-wise or patch-wise SR tasks. The work in [29] serves as the first to extend quantization to compress SR models, which quantizes only weights to be binary. However, the operations between activations and quantized weights are still floating-point, which cannot largely reduce the FLOPs towards the practical speedup. Different from the previous work [29], we optimize the SR network with both low-bit quantized weights and activations by introducing a learnable parameter, which achieves the bound of the quantization range.

Knowledge distillation. Knowledge distillation [14] aims to transfer the knowledge from a cumbersome network (teacher) to a compact network (student). It has been widely applied to various computer vision tasks by using the softened output knowledge [14] and intermediate feature representations [24, 37]. In line with our work, Zhuang *et al.* [42] proposed a guidance loss to jointly optimize the full-precision network and the low-precision model. However, it is not suitable to directly use such probability-based loss for SR, as the outputs of SR are reconstructed HR images. Different from these methods, our approach adopts structured knowledge based on the implicit information of a pre-trained network, which concentrates on aligning the spatial correlation between the low-precision and full-precision features to be more suitable for pixel-wise SR task.

3 The Proposed Method

3.1 A Close Look at SR Model Quantization

Current practice [29] only quantizes the weights in deep SR models, which does reduce the storage cost but unfortunately ignores the computational ef-

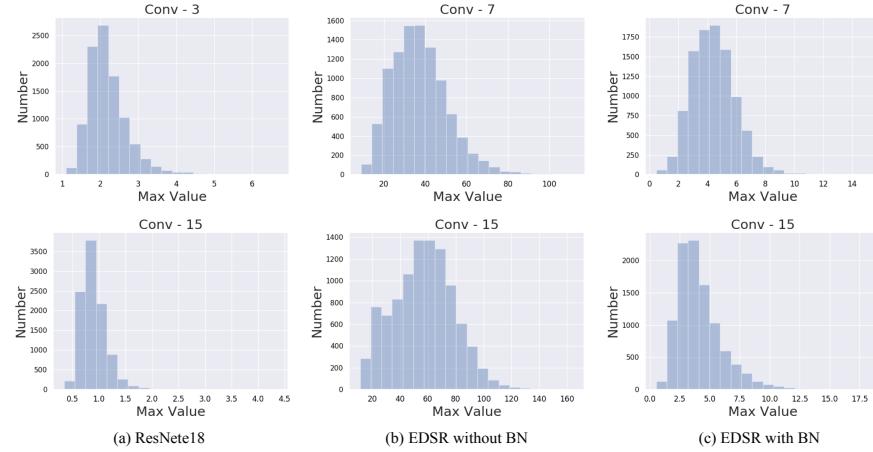


Fig. 2. Max value of activations from different layers and samples of ResNet-18 on ImageNet, EDSR [23] w. / wo. BN on DIV2K. The absence of BN causes more dynamic range problem in SR models.

ficiency caused by the full-precision multiplication between weights and activations. Moreover, the conversion between low-precision weights and full-precision activations aggravates the training time. It is not runtime-friendly to deploy such quantization scheme in real scenarios. Note that some quantization methods quantize the activations based on the premise of batch normalization [16]. In this way, the activations are supposed to stay in a stable range.

However, prior work [23] has shown that batch normalization layers get rid of range flexibility by normalizing the features, and simply removing them can make a big margin of improvements while reducing GPU memory cost. This modification can be effectively extended to the recent state-of-the-art SR models (*e.g.* RDN [39], RCAN [38], DBPN [9]) for ensuring range flexibility and reducing artifacts. Fig. 2 shows the statistics collected from pre-trained ResNet-18 on ImageNet, EDSR with and without BN on DIV2K. We can see that the max value of activations varies a lot in different samples in the same layer, and the activation range is more dynamic of SR model (EDSR) than that in classification model (ResNet-18). It indicates the dynamic range problem is more severe in the SR model than that of the classification model. Moreover, the absence of BN causes a more severe dynamic range problem. Compared to EDSR with BN (Fig. 2(b)), the max value of activations in EDSR without BN has a wider max value range and shows a more even distribution, which indicates that removing BN in current SR models causes activations in a more dynamic range, so as to be difficult to manually decide the quantization range. We argue that this quantization range is vital to the performance: If its maximum tends to a tiny value, the upper bound of the quantization range will be very small. And as reflected in the reconstructed HR images, the details will be mostly lost, which thus causes significant quality degradation. In contrast, if its maximum is abnormally large,

the quantization range may include outliers that contain redundant information and decrease the accuracy of quantized DNNs.

3.2 Parameterized Max Scale (PAMS)

The proposed PAMS quantizes both activations and weights of deep SR models. In this subsection, we first elaborate on our overall quantization approach. Then we describe how to leverage trainable truncated parameters to adaptively learn the upper bound of activations. To efficiently use the pre-trained network and improve the performance, we further introduce a structured knowledge transfer (SKT) loss. The overall framework is presented in Fig. 1.

Quantization function. As shown in [2], distributions on different layers of activations tend to be symmetric. This characteristic can help to improve the model accuracy of a quantized network with extremely low-bit weights and activations, as validated in [8]. Therefore, given a specific full-precision model with a parameter set \mathcal{X} (\mathcal{X} denotes either weights or activations of a specific layer), we quantize every element x ($x \in \mathcal{X}$) by using the following point-wise quantization function Q with a symmetric mode:

$$Q(x, n) = \lfloor \frac{f(x)}{s(n)} \rfloor s(n), \quad (1)$$

where $f(x)$ is the clamp function that limits the inputs range and $s(n)$ is the map function that scales the higher precision inputs to their lower bit reflections, which can be formulated by $f(x) = \max(\min(x, a), -a)$ and $s(n) = \frac{a}{2^{n-1}-1}$, respectively. n denotes the number of quantization level. a represents the maximum of the absolute value of \mathcal{X} and $\lfloor \cdot \rfloor$ rounds the value to the nearest integer.

For quantizing weights, previous work [17] has shown that simply set $a := \max(|w|)$ only has a negligible effect on the performance, which is therefore adopted in our approach. As for activations, the quantization range depends on the inputs, which leads to a dynamic range. This instability is unfavorable to the performance and model generality, which is designed in the following.

Trainable upper bound. In the previous work [3], the dynamic range of activations can be partially alleviated by using a parameterized clipping activation function to replace the Rectified Linear Unit (ReLU), which limits the application scope. In this paper, we propose a novel activation quantization scheme, in which the clamp function $f^*(\cdot)$ has the trainable parameter α to dynamically adjust the upper bound of the quantization range. We can directly employ the stochastic gradient descent to update this parameter, which is able to minimize the performance degradation arising from the quantization. For a given activation, the corresponding value will be quantized to n bits by:

$$x_q = Q(x, n) = \frac{\alpha}{2^{n-1}-1} \times \lfloor \tilde{x} \times \frac{2^{n-1}-1}{\alpha} \rfloor, \quad (2)$$

where $\tilde{x} = f^*(x) = \max(\min(x, \alpha), -\alpha)$. The dynamic range is limited to $[-\alpha, \alpha]$. The advantages of our quantization function lie in that $Q(x, n)$ directly

involves in the back-propagation process based on only one learnable parameter α , based upon which we can train the quantized SR network in an end-to-end manner. Extensive experiments in Section 4 demonstrate that Eq.2 can introduce more effectiveness comparing to several state-of-the-art quantization methods [41, 17, 3].

Back-propagation with quantization. In back-propagation, $\frac{\partial x_q}{\partial \tilde{x}}$ can be approximated to 1 based on the straight-through estimator (STE) [5]. Inspired by [3], the gradient of α is calculated as follows:

$$\frac{\partial x_q}{\partial \alpha} \approx \frac{\partial x_q}{\partial \tilde{x}} \frac{\partial \tilde{x}}{\partial \alpha} = \begin{cases} -1, & x \in (-\infty, -\alpha], \\ 0, & x \in (-\alpha, \alpha), \\ 1, & x \in [\alpha, +\infty). \end{cases} \quad (3)$$

Note that, the work in [3] cuts off the gradients in the regions satisfied with $x < 0$, while PAMS can adaptively adjust α based on the gradients in both $x \geq \alpha$ and $x \leq \alpha$ areas. It is important for the post-training quantization, since the gradients of the pre-trained model tend to 0. In other words, PAMS can retain more gradient information for updating α .

Initializing α . To avoid gradients vanishing or exploding, non-convex optimization on DCNNs heavily depends on the initialization of parameters. Instead of manually designing the initial value of α , initialization based on the statistics from a pre-trained network can achieve better performance. Therefore, we resort to task-related statistical methods based on the pre-trained network to calibrate the quantization error. In particular, given the l -th layer with m input activations $x_1^{(l,t)}, \dots, x_m^{(l,t)}$, $\alpha^{(l)}$ is calculated by the redefined exponential moving average (EMA) function at the start of training:

$$\alpha^{(l,t)} = \beta \cdot \alpha^{(l,t-1)} + (1 - \beta) \cdot \text{avg}(\max(x_1^{(l,t)}), \dots, \max(x_m^{(l,t)})), \quad (4)$$

where t is the iteration number and β denotes the smoothing parameter of EMA, which is set to be 0.9997. Specially, we set β to 0 when t is 0.

3.3 Optimization

Pixel-wise loss. Given a training dataset $D = \{I_{LR}^i, I_{HR}^i\}_{i=1}^n$ with n LR input images and their corresponding HR counterparts, SR models are commonly optimized by minimizing the conventional pixel-wise L_1 loss between the output I_{SR} and the ground truth image I_{HR} :

$$L_{PIX} = \frac{1}{n} \sum_{i=1}^n \|I_{HR}^i - I_{SR}^i\|_1, \quad (5)$$

where $\|\cdot\|_1$ denotes the L_1 norm. A better SR model needs to infer the high-frequency textures from a low-resolution input. However, it is hard to obtain by only using Eq. 5 based on low-bit quantization, which is due to the accumulated quantization error.

Algorithm 1 Quantization SR Model

Input: Training dataset D , full-precision model T , quantization level n ;
Output: The quantized model S ;

- 1: Define the low-precision model S by replacing convolution layers of T by n -bit PAMS;
- 2: Initialize $\alpha^{(l)}$ of each layer l with Eq. 4;
- 3: **for** $i = 1, \dots, N$ epoch **do**
- 4: Forward pass by applying clamp function to weights and activations using Eq. 1 and Eq. 2;
- 5: Update all parameters in Eq. 7 via SGD;
- 6: **end for**
- 7: **return** S ;

Structured knowledge transfer (SKT). Inspired by [37], we consider that the full-precision model has learned high-level representation, which provides knowledge to the low-precision one about where it concentrates. More specifically, instead of using the soft probability in the classification task, we align the structured features between the cumbersome network and the quantized one by minimizing their pixel-wise distance. Therefore, the loss function for our SKT is defined as:

$$L_{SKT} = \left\| \frac{F'_S}{\|F'_S\|_2} - \frac{F'_T}{\|F'_T\|_2} \right\|_p, \quad (6)$$

where F'_T , F'_S are a pair of structure features after the spatial mapping of activations from the full-precision network and the correspond low-precision one, respectively. The spatial mapping defined by $F' = \sum_{i=1}^C |F_i|^2 \in \mathbb{R}^{H \times W}$, where $F \in \mathbb{R}^{C \times H \times W}$ denotes the activations after the last layer in the high-level feature extractor. We set $p = 2$ for p -norm in our experiments. In sum, SKT enhances the learning process of spatial correlation in the low-precision model which effectively improves the performance of the quantized network and provides an additional constraint to avoid producing over-smoothed images.

The overall loss function. Given an SR model, as consistent with the distillation term mentioned above, the whole objective function is given as:

$$L_{SR} = \lambda_p L_{PIX} + \lambda_s L_{SKT}, \quad (7)$$

where λ_p and λ_s are coefficients to control the balance of the corresponding loss. We set λ_p to 1 and λ_s to 10^3 . The overall optimized process is summarized in Alg. 1.

4 Experiments

4.1 Experimental Settings

Datasets and metrics. DIV2K [35] contains 800 training images, 100 validating images and 100 testing images. We train all models with DIV2K training

Table 1. Comparison between quantizing EDSR [23] and RDN [39] by deploying PAMS with low-bit weights and activations on the public benchmark (PSNR(dB)/SSIM). The higher PSNR and higher SSIM, the better performance the methods achieve. EDSR is based on the residual block and RDN is based on the dense block. RDN* denotes the results based on our implementation.

Dataset	Scale	Bicubic	EDSR	PAMS-EDSR (8-bit)	PAMS-EDSR (4-bit)	RDN*	PAMS-RDN (8-bit)	PAMS-RDN (4-bit)
Set5	$\times 2$	33.66/0.9299	37.985/0.9604	37.946/0.9603	37.665/0.9588	38.027/0.9606	38.060/0.9606	36.528/0.9527
	$\times 4$	28.42/0.8104	32.095/0.8938	32.124/0.8940	31.591/0.8851	32.244/0.8959	32.340/0.8966	30.441/0.8624
Set14	$\times 2$	30.24/0.8688	33.568/0.9175	33.564/0.9175	33.196/0.9146	33.604/0.9174	33.732/0.9189	32.392/0.9050
	$\times 4$	26.00/0.7027	28.576/0.7813	28.585/0.7811	28.199/0.7725	28.669/0.7838	28.721/0.7848	27.536/0.7530
BSD100	$\times 2$	29.56/0.8431	32.155/0.8993	32.157/0.8994	31.936/0.8966	32.187/0.8999	32.215/0.9000	31.268/0.8853
	$\times 4$	25.96/0.6675	27.562/0.7355	27.565/0.7352	27.322/0.7282	27.627/0.7379	27.644/0.7382	26.869/0.7097
Urban100	$\times 2$	26.88/0.8403	31.977/0.9272	32.003/0.9274	31.100/0.9194	32.084/0.9284	32.262/0.9298	29.703/0.8976
	$\times 4$	23.14/0.6577	26.035/0.7848	26.016/0.7843	25.321/0.7624	26.293/0.7924	26.367/0.7955	24.523/0.7256

images. For testing, we use four standard benchmark datasets: Set5 [1], Set14 [21], BSD100 [30] and Urban100 [15]. For the evaluation metrics, we use PSNR and SSIM [36] over the Y channel between the output quality image and the original HR image.

SR models and alternative approaches. Both residual block and dense block are widely used in SR models, like VDSR [18], EDSR [23] and RDN [39]. To validate the superiority of our approach, we choose EDSR and RDN as backbones and use 8-bit and 4-bit quantization on them. As most parameters exist in the high-level feature extraction module, we do not quantize weights and activations in low-level feature extraction and reconstruction modules, which ensures a trade-off between performance and model size. The qualitative comparisons are generated by the publicly available source code in EDSR [23].

Training setting. The model is implemented by using PyTorch [31]. Following the setting of [23], we pre-process all images in the DIV2K training dataset by subtracting the mean RGB and adopt a normal data augmentation during training, which includes random horizontal flips and vertical rotations. The mini-batch size is set to 16. We deploy the ADAM optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$ to the model, which is trained for 30 epochs. The learning rate is initialized by 10^{-4} and is halved at every 10 epochs.

4.2 Quantitative and Qualitative Results

As shown in Table. 1, the proposed PAMS with 8-bit weights and activations achieves competitive or even better results on different backbones. For instance, 8-bit PAMS-RDN outperforms the full-precision RDN by 0.178dB PSNR and 0.074dB PSNR on Urban100 with scale factors of $\times 2$ and $\times 4$, respectively. The 4-bit PAMS-EDSR only suffers 0.24dB PSNR loss on BSD100 for a scale factor of $\times 4$ compared to its full-precision model. Quantizing RDN leads to a significant improvement over EDSR in 8-bit, which indicates that dense blocks may produce

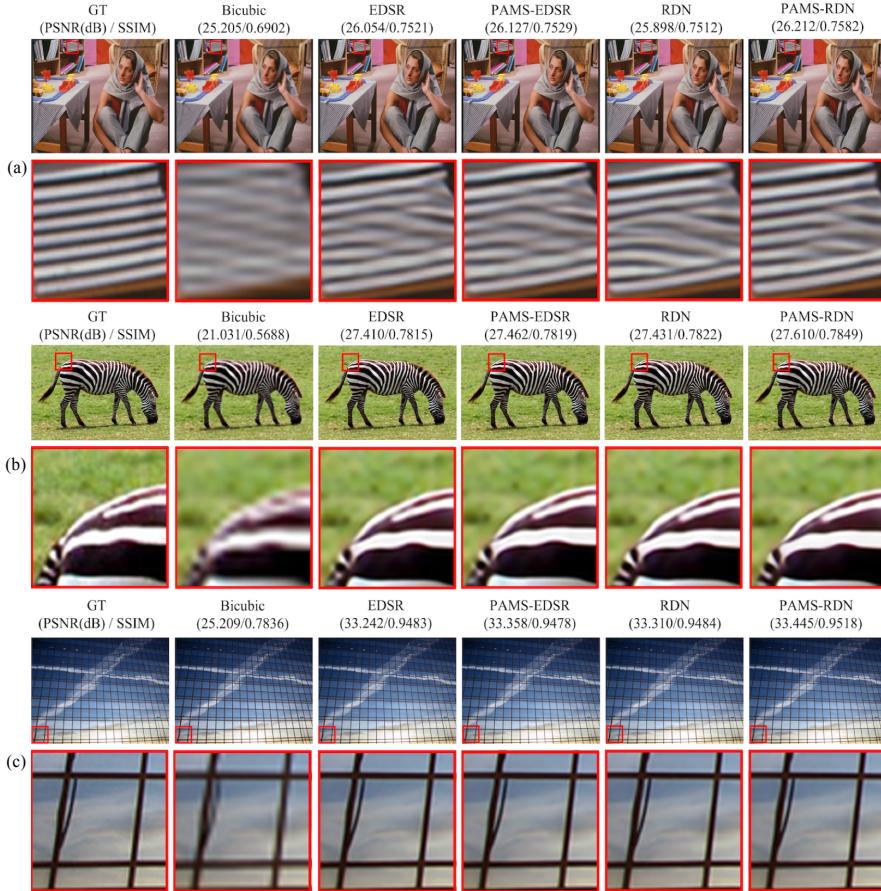


Fig. 3. Qualitative comparison between 8-bit and full-precision models with a scale factor of $\times 4$. (a) and (b) are the results of “barbara” and “zebra” from Set14, respectively. (c) is the results of “img055” from Urban100. Note that the quantized models with PAMS produce extremely similar or even better SR images to their full-precision counterparts, while the former has a significant reduction of model size and computational complexity.

more redundancy than residual blocks. We provide more qualitative evaluations on the 8-bit quantization in Fig. 3. The models with PAMS produce more visually natural images than the bicubic interpolation, and are extremely similar to their full-precision counterparts. Considering that residual-based models are widely used, the results also indicate the generality of the proposed method.

For a better comparison, we re-implement Dorefa [41], Tensorflow Lite [17] and PACT [3] on EDSR. We use the same initialization method and quantize both weights and activations in each residual block as PAMS-EDSR. For Dorefa,

Table 2. Comparison to the state-of-the-art quantization methods by using different bits on a scale factor of $\times 4$ super-resolution. EDSR is the backbone network.

Dataset	Bits	Dorefa-EDSR	TF Lite-EDSR	PACT-EDSR	PAMS-EDSR
Set5	8	30.194/0.8556	31.910/0.8906	31.520/0.8853	32.124/0.8940
	4	29.569/0.8369	31.380/0.8812	31.393/0.8834	31.591/0.8851
Set14	8	27.297/0.7492	28.416/0.7779	28.181/0.7712	28.585/0.7811
	4	26.817/0.7352	28.109/0.7690	28.104/0.7695	28.199/0.7725
BSD100	8	26.767/0.7079	27.470/0.7329	27.288/0.7261	27.565/0.7352
	4	26.474/0.6971	27.252/0.7239	27.251/0.7245	27.322/0.7282
Urban100	8	24.220/0.7128	25.739/0.7760	25.245/0.7570	26.016/0.7843
	4	23.753/0.6898	25.198/0.7551	25.148/0.7535	25.321/0.7624



Fig. 4. Qualitative comparison of our method with other quantization methods on a scale factor of $\times 4$.

we do not quantize gradients for a fair comparison. Table. 2 shows the results of 8-bit and 4-bit EDSR. Our method achieves better performance, compared to all baselines. For example, 8-bit PAMS-EDSR outperforms 8-bit Dorefa-EDSR by 1.288dB PSNR and 1.796dB PSNR on Set14 and Urban100, respectively. The reconstruction results are further shown in Fig. 4. Compared to other methods. The output (SR images) using PAMS are better-looking with sharp edges and rich details. In conclusion, PAMS with trainable truncated parameters rely on the backward which achieves much better generalization ability.

4.3 Compression Ratio

The model size and compression ratio of EDSR and RDN are presented in Table. 3. In particular, the full-precision network is represented by using single precision floating point. The model size of the full-precision network considerably decreases

Table 3. Comparison of EDSR and RDN with different bits on BSD100. W and A represent the number of bits of weights and activations, respectively.

Model	W/A	StorageSize (r_{comp})	PSNR(dB) / SSIM
EDSR (32-bit)	32/32	1.518M (0%)	27.562/0.7355
PAMS-EDSR (8-bit)	8/8	0.631M (58.4%)	27.565/0.7352
PAMS-EDSR (4-bit)	4/4	0.484M (68.1%)	27.322/0.7282
RDN (32-bit)	32/32	22.27M (0%)	27.627/0.7379
PAMS-RDN (8-bit)	8/8	5.82M (73.9%)	27.644/0.7382
PAMS-RDN (4-bit)	4/4	3.08M (86.2%)	26.869/0.7097

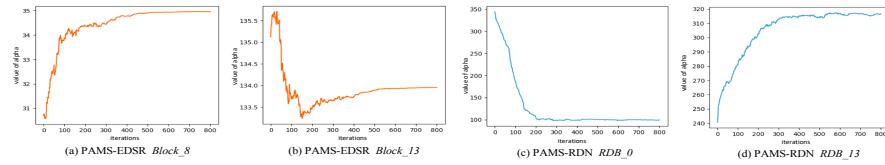


Fig. 5. Convergence curves of α for 8-bit PAMS-EDSR and 8-bit PAMS-RDN.

after quantization. Note that, we only quantize the weights and activations in the high-level feature extractor module, such that the compression ratios are calculated based on the total parameters of the network and the parameters in the high-level feature extractor. Although PAMS introduces a trainable parameter α , it still yields a 50%-90% compression ratio, since it directly depends on the backbone and the number of bits. It can be seen that 4-bit weights and activations cause more performance degradation than the 8-bit model. But lower-bit quantized networks can significantly reduce storage requirement.

4.4 Convergence of the α

To demonstrate the convergence of our method, we directly validate the convergence on α during training. The results are presented in Fig. 5. The first and second columns show the α of PAMS-EDSR on the layer of *Block_8* and *Block_13*, respectively. The third and fourth columns show the α of PAMS-RDN in *RDB_0* and *RDB_13*, respectively (*RDB* denotes the Residual Dense Block). It illustrates that α in different layers not only have different values but also have different evolving directions.

For instance, PAMS-EDSR *Block_8* (Fig. 5(a)) and PAMS-RDN *RDB_0* (Fig. 5(c)) act in the same direction, while PAMS-EDSR *Block_13* (Fig. 5(b)) and PAMS-RDN *RDB_13* (Fig. 5(d)) are with the opposite trend. We also found that α can promote the convergence to a stable value for both EDSR and RDN, which indicates the effectiveness of our method.

Table 4. Comparison of the performance gap between the singe-precision EDSR, 8-bit PACT-EDSR w. / wo. BN and 8-bit PAMS wo. BN. (PSNR(dB) / SSIM).

Model	With BN	Set5	Set14	BSD100	Urban100
PACT-EDSR	✓	0.531/0.0083	0.273/0.0068	0.166/0.0056	0.354/0.0125
PACT-EDSR	✗	0.575/0.0085	0.395/0.0101	0.274/0.0094	0.790/0.0278
PAMS-EDSR	✗	0.029/0.0002	0.009/-0.0002	0.003/-0.0003	-0.019/-0.0005

Table 5. Results about different initialization methods of α on EDSR with $\times 4$ scale factor (PSNR(dB) / SSIM).

Init.	Set5	Set14	BSD100	Urban100
Random	31.782/0.8896	28.383/0.7779	26.273/0.6879	23.488/0.6780
EMA	32.002/0.8923	28.497/0.7797	28.497/0.7797	25.806/0.7788

4.5 Ablation Study

Effect of BN in SR models. To investigate the effect of quantizing normalized features, we use PACT to quantize EDSR with BN and without BN. As shown in Table. 4, the performance gap between the quantized EDSR without BN is larger than the quantized EDSR with BN. For example, The gap of 8-bit PACT-EDSR without BN is 0.790dB PSNR on Urban100, which is larger than PACT-EDSR with BN (0.354dB PSNR). It shows that the performance degradation of unnormalized features is more pronounced in lower-precision SR models, Moreover, PAMS-EDSR can save more important information for unnormalized weights and activations which largely decrease the performance gaps.

Effect of the learnable α . We compare our learnable max scale (PAMS) with the fixed maximum (TF Lite) for quantizing activations. Quantitative and qualitative results are represented in Table. 2 and Fig. 4, respectively. Compared to TF Lite-EDSR, PAMS-EDSR achieves a better score as it produces sharper images and more realistic textures. It indicates that our method can learn a more suitable quantization range which contains more information about the full-precision model and reduces the quantization error.

Effect of the initialization of α . We evaluate our EMA initialization with random initialization on EDSR with a scale factor of $\times 4$. For the random mode, we initialize α in the activation quantization layer with a random number ranges from 0 to 128, which ensures that α can be initialized to a larger value in different layers independently. As illustrated in Table. 5, EMA initialization achieves better performance on all benchmark datasets. To explain, EMA achieves better statistical distribution by α that can further help improve SR performance.

Investigating SKT loss. To investigate the effectiveness of SKT, we further compare the quantized model with and without SKT. As shown in Table. 6,

Table 6. Results of PAMS-EDSR w. / wo. L_{SKT} on 8-bit and 4-bit settings (PSNR(dB) / SSIM).

Dataset	bits	without L_{SKT}	with L_{SKT}	metrics. \uparrow
Set5	8	32.127/0.8939	32.124/0.8940	-0.003/0.0001
	4	31.538/0.8842	31.591/0.8851	0.053/0.0009
Set14	8	28.541/0.7807	28.585/0.7811	0.044/0.0004
	4	28.177/0.7723	28.199/0.7725	0.022/0.0002
BSD100	8	27.550/0.7352	27.565/0.7352	0.015 /0.0000
	4	27.302/0.7280	27.322/0.7282	0.020/0.0002
Urban100	8	25.984/0.7835	26.016/0.7843	0.032/0.0008
	6	25.840/0.7794	25.876/0.7802	0.036/0.0008
	4	25.250/0.7607	25.321/0.7624	0.071/0.0017

PAMS-EDSR which is optimized with the SKT outperforms the corresponding counterpart. Especially, our method obtains much better performance on lower bits. For instance, compared to the PAMS-EDSR without L_{SKT} on Urban100, 4-bit PAMS-EDSR with L_{SKT} gains 0.071dB PSNR while 8-bit PAMS-EDSR with the same optimization gains only 0.032dB PSNR. It also indicates that the feature maps from the full-precision model can help the low-precision model to better capture the spatial correlation from images.

5 Conclusion

In this paper, we propose a novel symmetric quantization scheme, termed Parameterized Max Scale (PAMS), to effectively quantize both weights and activations of the full-precision network for SR tasks. The proposed method adopts a truncated parameter α to adaptively adjust the upper bound of quantization range. This technique alleviates the negative effect of dynamic range caused by the absence of batch normalization layers and helps to reduce the quantization error. To further approximate the full-precision network, we employ structured knowledge transfer (SKT) to retrain the quantized network in a few epochs. We have comprehensively evaluated the performance of the proposed approach on EDSR and RDN over public benchmarks, which demonstrates the superior performance gains and significant reduction in model size and computational complexity.

Acknowledgements. This work is supported by the Nature Science Foundation of China (No.U1705262, No.61772443, No.61572410, No.61802324 and No.61702136), National Key R&D Program (No.2017YFC0113000, and No.2016YFB1001503), Key R&D Program of Jiangxi Province (No. 20171ACH80022) and Natural Science Foundation of Guangdong Province in ChinaNo.2019B1515120049).

References

1. Bevilacqua, M., Roumy, A., Guillemot, C., Alberimorel, M.L.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In: BMVC (2012)
2. Cai, Z., He, X., Sun, J., Vasconcelos, N.: Deep learning with low precision by half-wave gaussian quantization. In: CVPR (2017)
3. Choi, J.: Pact: Parameterized clipping activation for quantized neural networks. In: CVPR (2018)
4. Choi, Y., El-Khamy, M., Lee, J.: Learning low precision deep neural networks through regularization. arXiv preprint arXiv:1809.00095 (2018)
5. Courbariaux, M., Hubara, I., Soudry, D., El-Yaniv, R., Bengio, Y.: Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or -1. CoRR **abs/1602.02830** (2016)
6. Denton, E.L., Zaremba, W., Bruna, J., LeCun, Y., Fergus, R.: Exploiting linear structure within convolutional networks for efficient evaluation. In: NeurIPS (2014)
7. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: ECCV (2014)
8. Faraone, J., Fraser, N., Blott, M., Leong, P.: Syq: Learning symmetric quantization for efficient deep neural networks. In: CVPR (2018)
9. Haris, M., Shakhnarovich, G., Ukita, N.: Deep back-projection networks for super-resolution. In: CVPR (2018)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
11. He, Y., Zhang, X., Sun, J.: Channel pruning for accelerating very deep neural networks. In: ICCV (2017)
12. He, Y., Kang, G., Dong, X., Fu, Y., Yang, Y.: Soft filter pruning for accelerating deep convolutional neural networks. arXiv preprint arXiv:1808.06866 (2018)
13. He, Y., Liu, P., Wang, Z., Hu, Z., Yang, Y.: Filter pruning via geometric median for deep convolutional neural networks acceleration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4340–4349 (2019)
14. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NeurIPS 2014 Workshops
15. Huang, J., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: CVPR (2015)
16. Ioffem, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML (2015)
17. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., Kalenichenko, D.: Quantization and training of neural networks for efficient integer-arithmetic-only inference. In: CVPR (2018)
18. Kim, J., Kwon Lee, J., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: CVPR (2016)
19. Kim, J.W., Lee, J.K., Lee, K.M.: Deeply-recursive convolutional network for image super-resolution. In: CVPR (2016)
20. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NeurIPS (2012)
21. Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z.e.a.: Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR (2017)

22. Li, Y., Lin, S., Zhang, B., Liu, J., Doermann, D., Wu, Y., Huang, F., Ji, R.: Exploiting kernel sparsity and entropy for interpretable cnn compression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2800–2809 (2019)
23. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: CVPR (2017)
24. Lin, S., Ji, R., Chen, C., Tao, D., Luo, J.: Holistic cnn compression via low-rank decomposition with knowledge transfer. TPAMI (2018)
25. Lin, S., Ji, R., Guo, X., Li, X.: Towards convolutional neural networks compression via global error reconstruction. In: IJCAI (2016)
26. Lin, S., Ji, R., Yan, C., Zhang, B., Cao, L., Ye, Q., Huang, F., Doermann, D.: Towards optimal structured cnn pruning via generative adversarial learning (2019)
27. Lin, S., Ji, R., Li, Y., Wu, Y., Huang, F., Zhang, B.: Accelerating convolutional networks via global & dynamic filter pruning. In: IJCAI. pp. 2425–2432 (2018)
28. Ma, N., Zhang, X., Zheng, H., Sun, J.: Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: ECCV (2018)
29. Ma, Y., Xiong, H., Hu, Z., Ma, L.: Efficient super resolution using binarized neural network. In: CVPR (2018)
30. Martin, D., Fowlkes, C.C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: ICCV (2001)
31. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: NeurIPS 2017 Workshops
32. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014)
33. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: CVPR (2018)
34. Tai, Y., Yang, J., Liu, X.: Image super-resolution via deep recursive residual network. In: CVPR (2017)
35. Timofte, R., Agustsson, E., Van Gool, L., Yang, M., Zhang, L.e.a.: Ntire 2017 challenge on single image super-resolution: Methods and results. In: CVPR 2017 Workshops
36. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., et al.: Image quality assessment: from error visibility to structural similarity. TIP (2004)
37. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer (2017)
38. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: ECCV (2018)
39. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: CVPR (2018)
40. Zheng, X., Ji, R., Tang, L., Zhang, B., Liu, J., Tian, Q.: Multinomial distribution learning for effective neural architecture search. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1304–1313 (2019)
41. Zhou, S., Ni, Z., Zhou, X., Wen, H., Wu, Y., Zou, Y.: Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. CoRR **abs/1606.06160** (2016)
42. Zhuang, B., Shen, C., Tan, M., Liu, L., Reid, I.: Towards effective low-bitwidth convolutional neural networks. In: CVPR (2018)
43. Zoph, B., Le, Q.V.: Neural architecture search with reinforcement learning. In: ICCV (2016)