Toward Accurate Post-Training Quantization for Image Super Resolution Supplementary Material

Zhijun Tu, Jie Hu, Hanting Chen, Yunhe Wang
Huawei Noah's Ark Lab
{zhijun.tu, hujie23, chenhanting, yunhe.wang}@huawei.com

A. More experimental results

As shown in Table 1, we further list the commercial quantization toolkits for existing AI accelerate devices, NNIE [2] and Vitis-AI [3], which only support 8-bit for weights and activations. For EDSR model [5], NNIE with upscaling of 4 causes severe PSNR drop on these four datasets, which is 0.301 dB, 0.198 dB, 0.134 dB and 0.204 dB, respectively, performs even worse on upscaling of 2. VitisAI could get better results but still cause significant performance degradation ($>$ 0.1 dB) on various datasets with upscaling of 2 and 4. In the contrast, our proposed method could significantly outperform the NNIE and Vitis-AI, and better than the bicubic interpolate, only cause 0.025 dB drop on Set5, 0.052 dB drop on Set14, 0.026 dB drop on BSD100 and 0.079 on Urban100 with upscaling of 4, greatly reducing the performance gap between quantized SR model and the full-precision model. For SRResNet model [4], the performance comparison shows much similar with EDSR, NNIE and VitisAI cause significant drop. For instance, NNIE with upscaling of 2 causes 1.484 dB drop on Set5, 1.026 dB drop on Set14, 0.669 dB drop on BSD100 and 0.967 dB drop on Urban100, which shows that they cause severe quantization error for image super resolution, can not be applied to low-level vision tasks directly. But the low-precision SRResNet model with our proposed post-training quantization method could achieve much better performance, only causes PSNR drop within 0.03 dB with upscaling of 4 on these four test sets. The extended experiments further illustrate that our proposed method is much more friendly to image super resolution than the existing PTQ methods.

B. Clipping values of activations

Figure 1 shows the lower and upper clipping values of activations for different layers and bit-width settings. For reference, we also plot the original minimum and maximum values.As shown in Figure 1a, Figure 1b and Figure 1c, we can see that the lower the bit width, the larger the difference between the clipping values and the original range, especially when quantizing to 4-bit, more than half of the original range is clipped off. Figure 1d shows that lower precision quantization prefers smaller activation range. which is much consistent with image classification [1].

C. Combined with QAT

To demonstrate that our method could accelerate the convergence of QAT, we shows the PSNR and SSIM values of different epoch in the quantization aware training process as shown in Figure 4. To show the trend of convergence, we set the training epoch to 15 (10 in previous experiments), we can see that, the model converges fast in the first several epochs, leveling off at around the 10-th epoch on Urban100 dataset (Figure 4h). In the contrast, existing QAT methods for image super resolution almost require 30 to 1500 epochs to recover the performance drop, which shows that QAT with our method could truly accelerate the deployment of quantized models.

D. Visualization

Figure 2 and Figure 3 show more visual results on 4-bit EDSR model and SRResNet model with upscaling of 4, which are the most difficult task with post-training quantization in our experiments. The PSNR and SSIM reported below the images are measured by the reconstructed image and the corresponding HR image. As we can see that our proposed method could truly provide a better visual performance for image super resolution with low-bit compression.

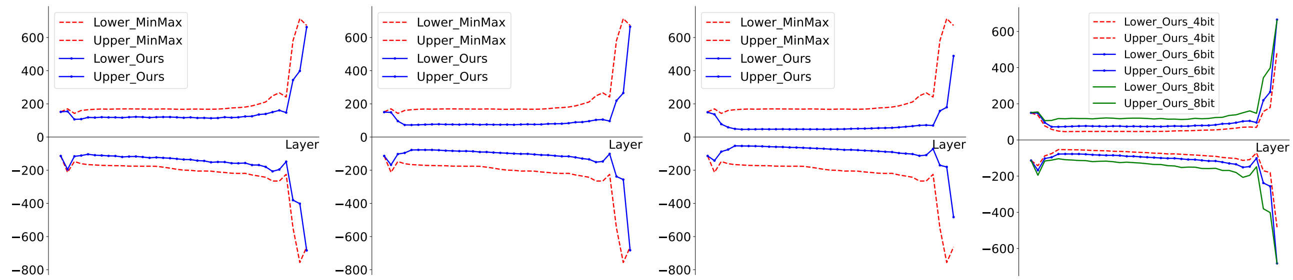
References

[1] Jungwook Choi, Zhuo Wang, Swagath Venkataramani, Pierce I-Jen Chuang, Vijayalakshmi Srinivasan, and Kailash Gopalakrishnan. Pact: Parameterized clipping activation for quantized neural networks. *arXiv preprint arXiv:1805.06085*, 2018. 1

[2] Andrey Ignatov, Radu Timofte, William Chou, Ke Wang, Max Wu, Tim Hartley, and Luc Van Gool. Ai benchmark: Running deep neural networks on android smartphones. In *Proceed-*

Table 1. PSNR(dB)/SSIM comparisons between existing post-training quantization methods and ours on EDSR and SRResNet of scale 4 and scale 2. The weights and activation of all the layers are quantized to 8-bit in this experiment.

Network	Method	Set5 ($\times 4$)	Set14 ($\times 4$)	BSD100 ($\times 4$)	Urban100 ($\times 4$)	Set5 ($\times 2$)	Set14 ($\times 2$)	BSD100 ($\times 2$)	Urban100 ($\times 2$)
EDSR [5]	Baseline	32.485/0.899	28.815/0.788	27.721/0.742	26.646/0.804	38.193/0.961	33.948/0.920	32.352/0.902	32.967/0.936
	Bicubic	28.420/0.810	26.000/0.703	25.960/0.668	23.140/0.658	33.660/0.930	30.24/0.869	29.560/0.843	26.880/0.840
	NNIE [2]	32.179/0.892	28.617/0.783	27.587/0.737	26.442/0.797	37.420/0.955	33.505/0.916	32.050/0.898	32.514/0.931
	VitisAI [3]	32.266/0.894	28.629/0.783	27.616/0.736	26.341/0.794	37.909/0.959	33.533/0.917	32.189/0.899	32.177/0.931
	Ours	**32.460/0.898**	**28.763/0.787**	**27.695/0.741**	**26.567/0.802**	**38.120/0.960**	**33.850/0.920**	**32.313/0.901**	**32.810/0.935**
SRResNet [4]	Baseline	32.234/0.896	28.656/0.784	27.630/0.738	26.229/0.791	38.091/0.961	33.752/0.919	32.241/0.900	32.367/0.931
	Bicubic	28.420/0.810	26.000/0.703	25.960/0.668	23.140/0.658	33.660/0.930	30.240/0.869	29.560/0.843	26.880/0.840
	NNIE [2]	31.643/0.880	28.206/0.769	27.284/0.725	25.746/0.771	36.607/0.941	32.726/0.899	31.572/0.885	31.400/0.913
	VitisAI [3]	31.956/0.889	28.392/0.771	27.459/0.729	25.907/0.779	37.465/0.956	33.173/0.912	31.876/0.892	31.498/0.922
	Ours	**32.207/0.895**	**28.619/0.783**	**27.618/0.738**	**26.191/0.790**	**38.032/0.960**	**33.648/0.919**	**32.212/0.900**	**32.210/0.930**



(a) EDSR$\times 4$ with 8-bit (b) EDSR$\times 4$ with 6-bit (c) EDSR$\times 4$ with 4-bit (d) EDSR$\times 4$ with 4, 6 and 8-bit

Figure 1. The clipping values of different layers with EDSR of upscaling of 4.

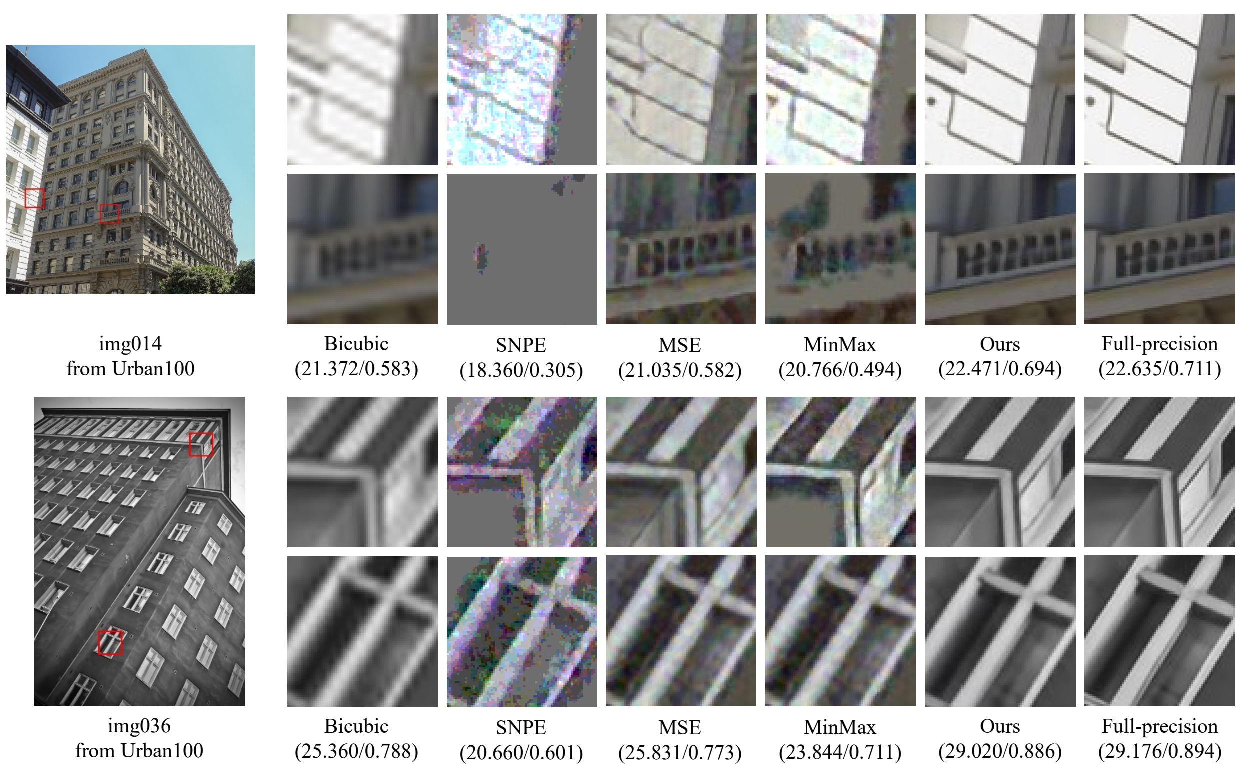


Figure 2. Visual results of different methods on 4-bit EDSR models with upscaling of 4. The images are selected from Urban100

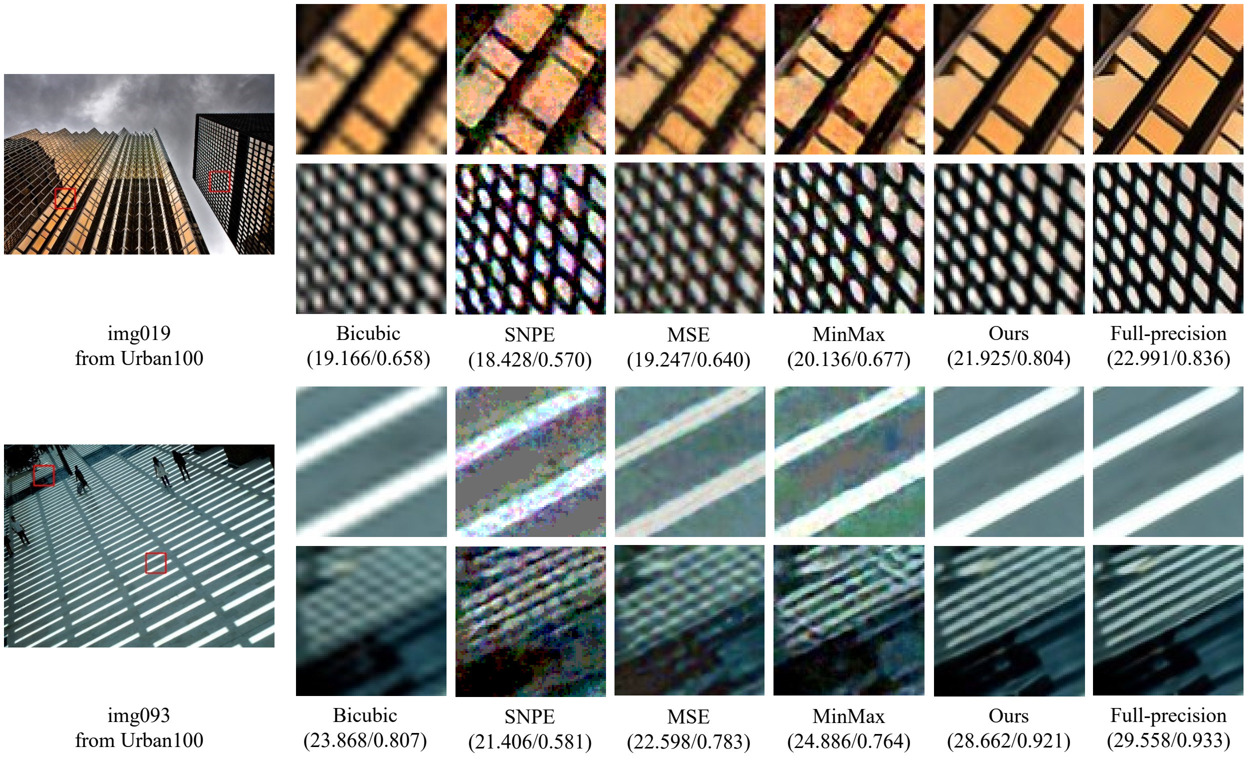
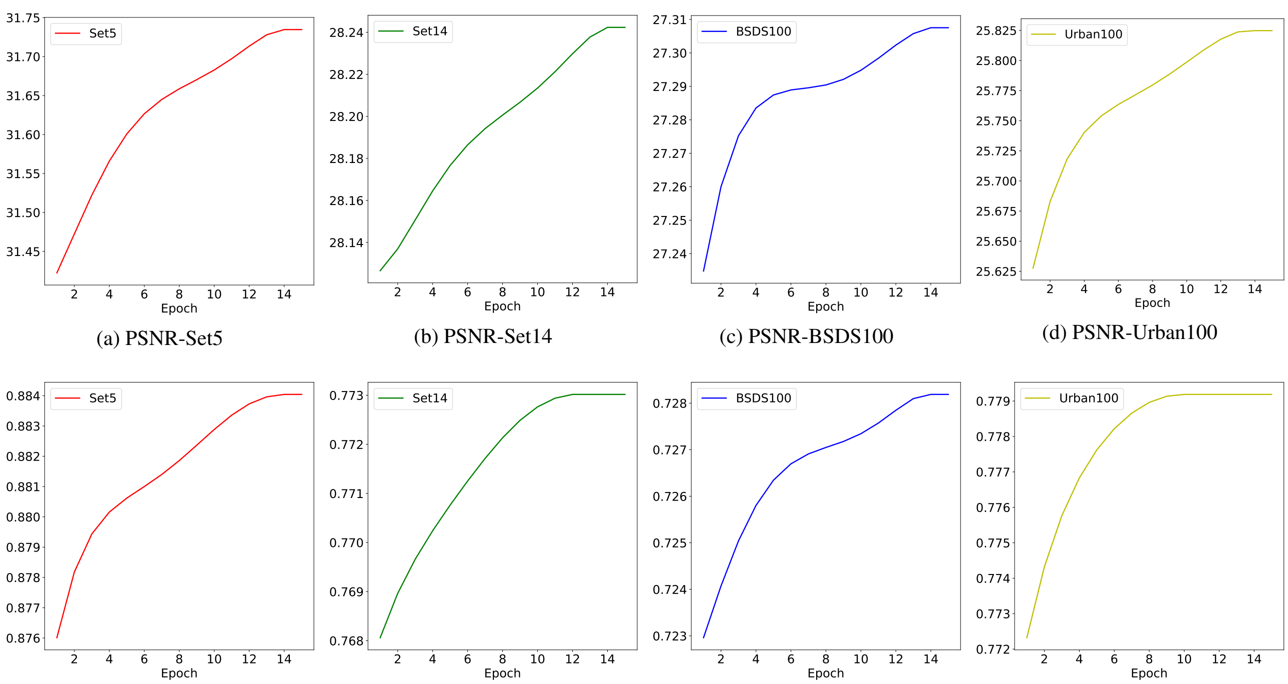


Figure 3. Visual results of different methods on 4-bit SRResNet models with upscaling of 4. The images are selected from Urban100



(a) PSNR-Set5 (b) PSNR-Set14 (c) PSNR-BSDS100 (d) PSNR-Urban100

(e) SSIM-Set5 (f) SSIM-Set14 (g) SSIM-BSDS100 (h) SSIM-Urban100

Figure 4. The PSNR(dB) and SSIM values of different epoch in QAT with the initialization of our proposed method. The top line represents the PSNR values and the bottom line represents the SSIM values of Set5, Set14, BSDS100 and Urban100 datasets

ings of the European Conference on Computer Vision (ECCV) Workshops, pages 0–0, 2018. 1, 2

[3] Vinod Kathail. Xilinx vitis unified software platform. In *Proceedings of the 2020 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, pages 173–174, 2020. 1, 2

[4] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 1, 2

[5] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 1, 2