2DQuant: Low-bit Post-Training Quantization for Image Super-Resolution

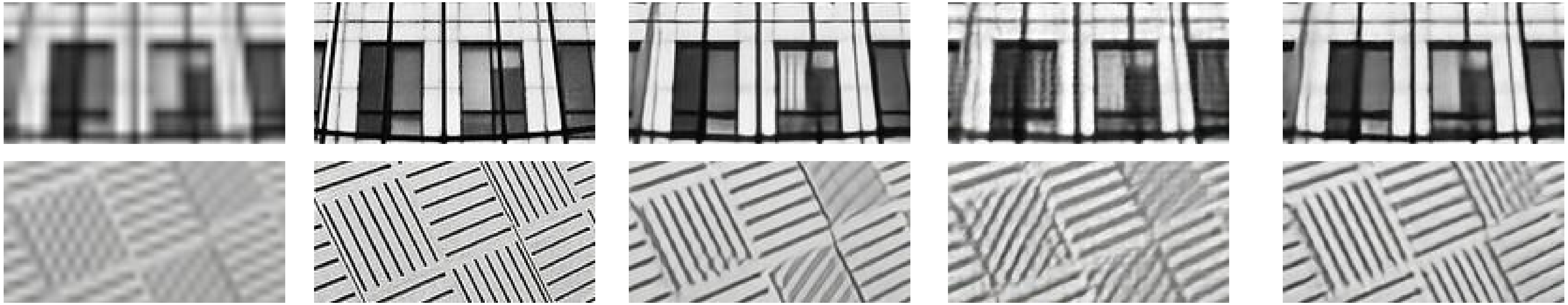
Kai Liu[1], Haotong Qin[2], Yong Guo[3], Xin Yuan[4], Linghe Kong[1], Guihai Chen[1*], Yulun Zhang[1*]

[1]Shanghai Jiao Tong University, [2]ETH Zürich, [3]Max Planck Institute for Informatics, [4]Westlake University

Introduction

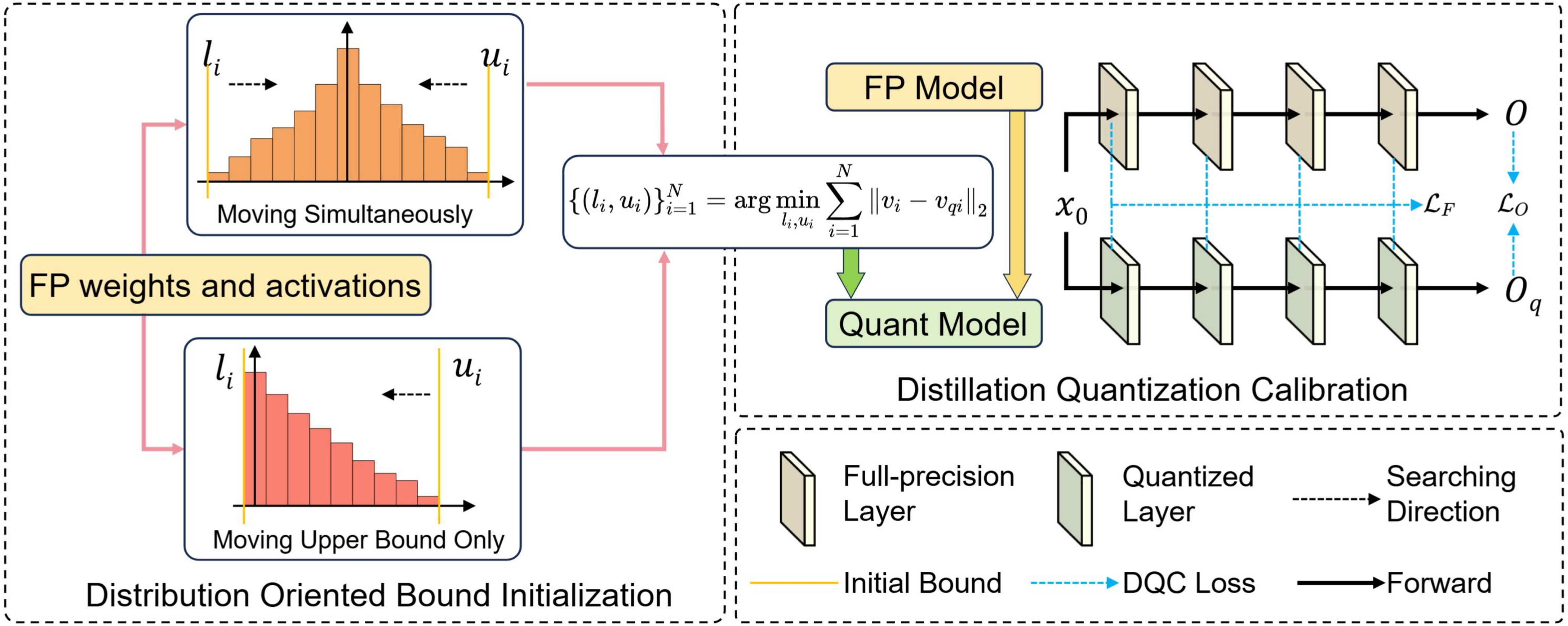
Motivation

- **Vision Transformers (ViTs)** excel in SR tasks but face high costs.
- Low bit post-training quantization (**PTQ**) reduces memory and computation.
- The deterioration of self-attention in quantized transformers limit its application.
- We propose **2DQuant,** a novel PTQ for ViT in SR.



Bicubic | **HR** | **SwinIR(FP)** | **DBDC+Pac**(CVPR 2023) | **2DQuant**(ours)

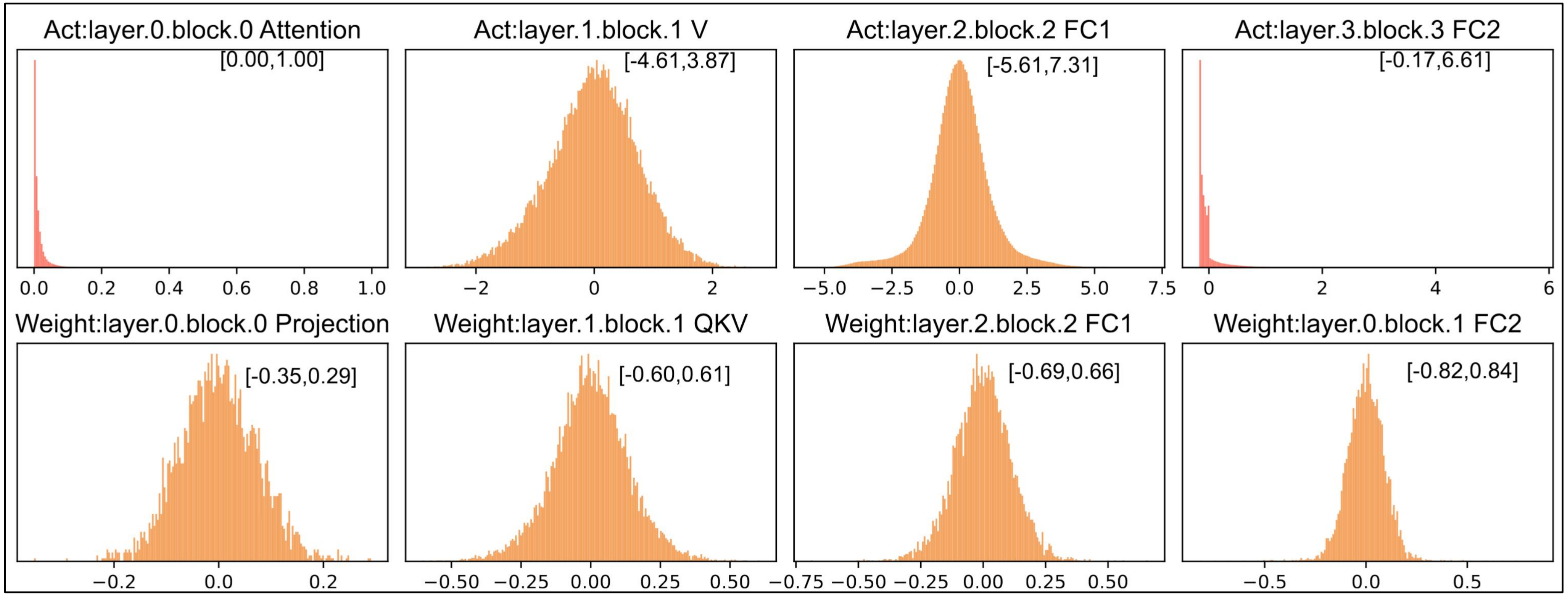
Method-Overview



Overall

- **The overall pipeline of our proposed 2DQuant.**
- The whole pipeline can be divided into two parts: DOBI (left) and DQC (right).

Method-Observation



Observation

- **The distribution of the activation and weights of ViT present two kind of distribution**
- Weight and most of the activation presents normal distribution.
- The attention part presents exponential distribution.
- So asymmetric quantization is necessary for low bit quantization in ViT.

Method-DOBI

- **MSE** serve as a strong method to obtain the quantization bound.

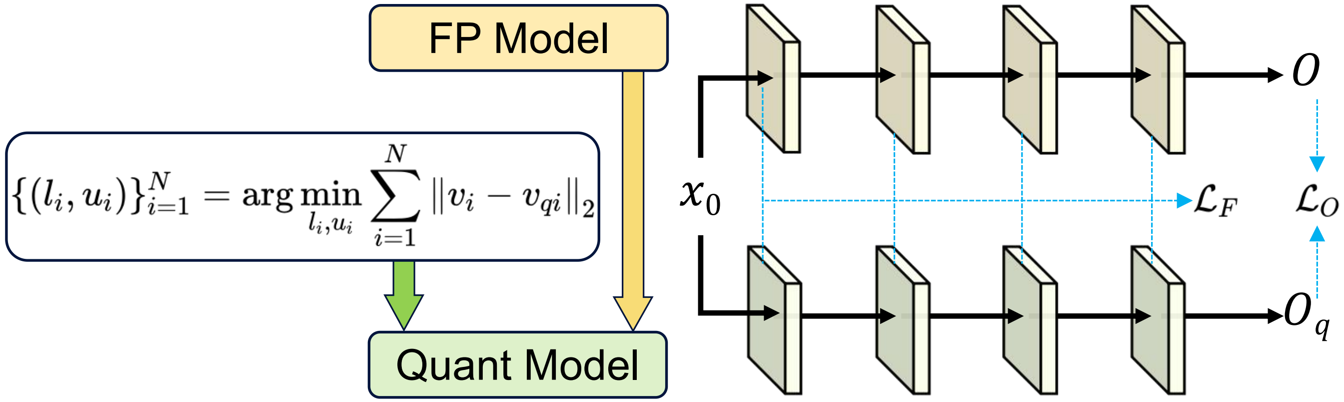
$$\{(l_i, u_i)\}_{i=1}^{N} = \arg\min_{l_i, u_i} \sum_{i=1}^{N} \|v_i - v_{qi}\|_2$$

- The different distribution of weight and activation guide us to search the best bound with different strategy.
- For normal distribution, two direction search is easy to find the best bound.
- For exponential distribution, the lower bound is fixed as the min value and only the upper bound needs searching.

Algorithm 1: DOBI pipeline

Data: Data to be quantized v, the number of search point K, bit b
Result: Clip bound l, u
$l \leftarrow \min(v)$,$u \leftarrow \max(v)$;
$min\_mse \leftarrow +\infty$;
if *v is symmetrical* **then**
    $\Delta l \leftarrow (\max(v) - \min(v))/2K$;
else
    $\Delta l \leftarrow 0$;
end
$\Delta u \leftarrow (\max(v) - \min(v))/2K$;
while $i \leq K$ **do**
    $l_i \leftarrow l + i \times \Delta l$, $u_i \leftarrow u + i \times \Delta u$;
    get v_q based on Eq. (1);
    $mse \leftarrow \|v - v_q\|_2$;
    if $mse \leq min\_mse$ **then**
        $min\_mse \leftarrow mse$;
        $l\_best \leftarrow l_i$, $u\_best \leftarrow u_i$;
    end
end

Method-DQC



Distillation Quantization Calibration

$$\mathcal{L}_O = \frac{1}{C_O H_O W_O} \|O - O_q\|_1{}^1$$

$$\mathcal{L}_F = \sum_i^N \frac{1}{C_i H_i W_i} \left\| \frac{F_i}{\|F_i\|_2} - \frac{F_{qi}}{\|F_{qi}\|_2} \right\|_2$$

$$\mathcal{L} = \mathcal{L}_O + \lambda \mathcal{L}_F$$

DQC

- DOBI targets at local quantization error loss, which is not necessarily consistent with task loss.
- Distillation between the FP model and the quantized model could provide accurate update direction for quantizers' bound.
- Features and final output consititutes the optimization loss.

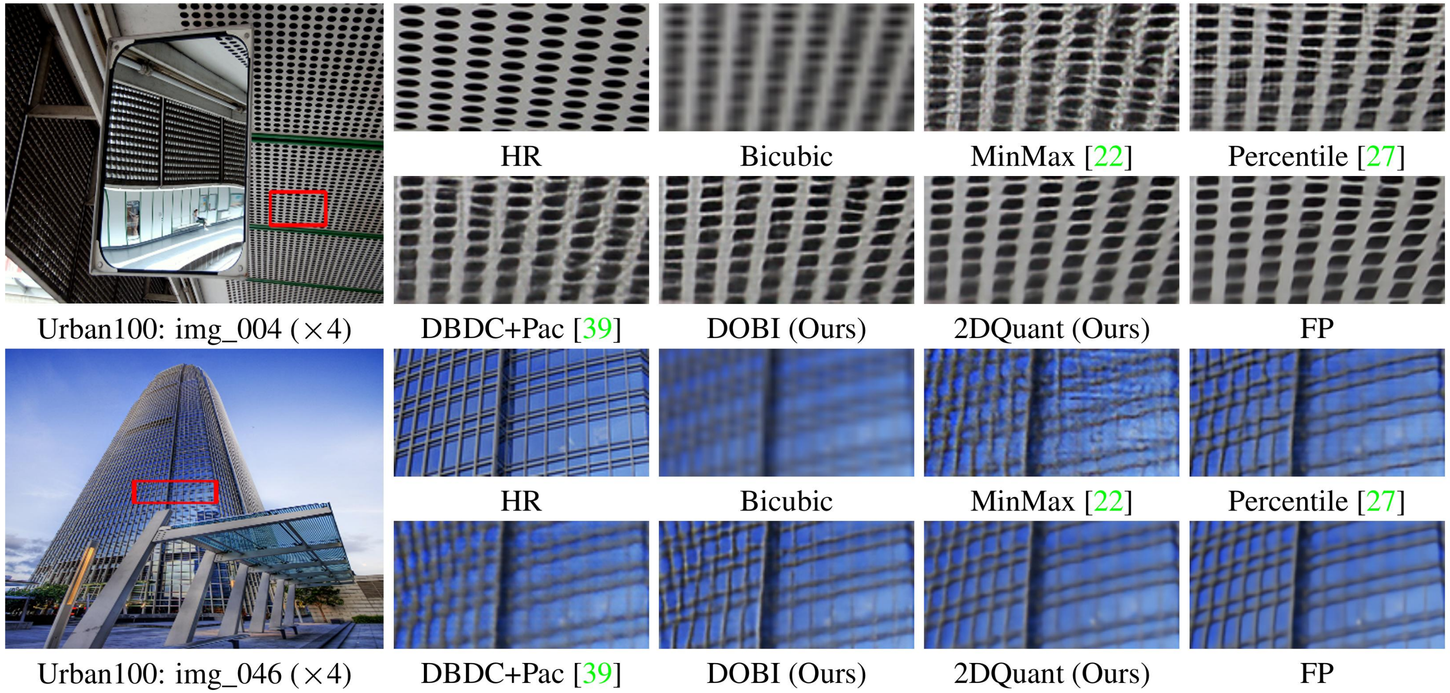
Experiments

Method	Bit	Set5 (×4) PSNR↑	Set5 (×4) SSIM↑	Set14 (×4) PSNR↑	Set14 (×4) SSIM↑	B100 (×4) PSNR↑	B100 (×4) SSIM↑	Urban100 (×4) PSNR↑	Urban100 (×4) SSIM↑	Manga109 (×4) PSNR↑	Manga109 (×4) SSIM↑
SwinIR-light [29]	32	32.45	0.8976	28.77	0.7858	27.69	0.7406	26.48	0.7980	30.92	0.9150
Bicubic	32	27.56	0.7896	25.51	0.6820	25.54	0.6466	22.68	0.6352	24.19	0.7670
MinMax [22]	4	28.63	0.7891	25.73	0.6657	25.10	0.6061	23.07	0.6216	26.97	0.8104
Percentile [27]	4	30.64	0.8679	27.61	0.7563	26.96	0.7151	24.96	0.7479	28.78	0.8803
EDSR† [30, 39]	4	31.20	0.8670	27.98	0.7600	27.09	0.7140	25.56	0.7640	N/A	N/A
DBDC+Pac [39]	4	30.74	0.8609	27.66	0.7526	26.97	0.7104	24.94	0.7369	28.52	0.8697
DOBI (Ours)	4	31.10	0.8770	28.03	0.7672	27.18	0.7237	25.43	0.7631	29.31	0.8916
2DQuant (Ours)	4	31.77	0.8867	28.30	0.7733	27.37	0.7278	25.71	0.7712	29.71	0.8972
MinMax [22]	3	19.41	0.3385	18.35	0.2549	18.79	0.2434	17.88	0.2825	19.13	0.3097
Percentile [27]	3	27.55	0.7270	25.15	0.6043	24.45	0.5333	22.80	0.5833	26.15	0.7569
DBDC+Pac [39]	3	27.91	0.7250	25.86	0.6451	25.65	0.6239	23.45	0.6249	26.03	0.7321
DOBI (Ours)	3	29.59	0.8237	26.87	0.7156	26.24	0.6735	24.17	0.6880	27.62	0.8349
2DQuant (Ours)	3	30.90	0.8704	27.75	0.7571	26.99	0.7126	24.85	0.7355	28.21	0.8683
MinMax [22]	2	23.96	0.4950	22.92	0.4407	22.70	0.3943	21.16	0.4053	22.94	0.5178
Percentile [27]	2	23.03	0.4772	22.12	0.4059	21.83	0.3816	20.45	0.3951	20.88	0.3948
DBDC+Pac [39]	2	25.01	0.5554	23.82	0.4995	23.64	0.4544	21.84	0.4631	23.63	0.5854
DOBI (Ours)	2	28.82	0.7699	26.46	0.6804	25.97	0.6319	23.67	0.6407	26.32	0.7718
2DQuant (Ours)	2	29.53	0.8372	26.86	0.7322	26.46	0.6927	23.84	0.6912	26.07	0.8163

Quantitative

- **Best performance:** Achieves the best results among quantization methods for SR.
- More results can be found in the main paper.

Experiments



Visual

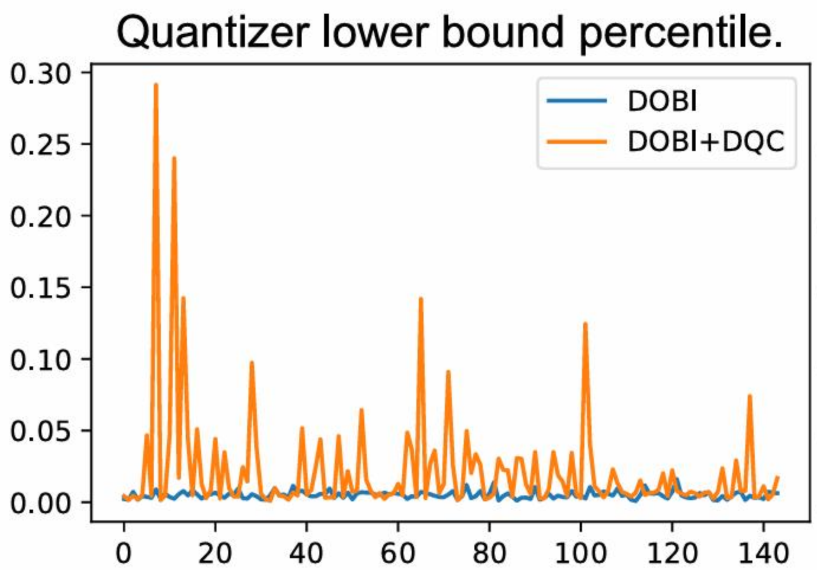
- Our method restores clearer images with more texture details.
- The gap between the quant model and the FP model is small.
- Quantization alleviates overfitting and in some condition, quantized model has better performance compared with FP model.

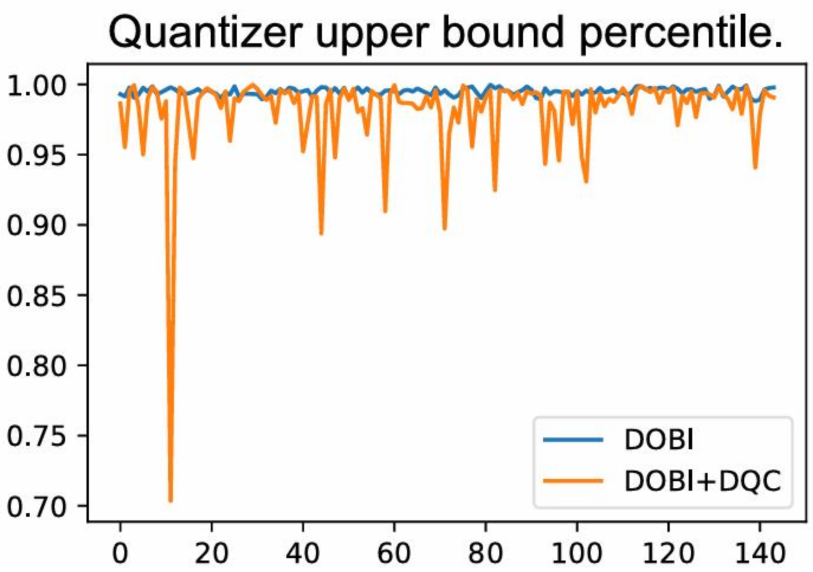
Model	EDSR [30]	EDSR (4bit) [39]	SwinIR-light [29]	DBDC+Pac (4bit) [39]	Ours (4bit)
Params (MB)	172.36	21.55	3.42	1.17	1.17
Ops (G)	823.34	103.05	16.74	4.19	4.19
PNSR on Urban100	26.64	25.56	26.47	24.94	25.71

Compression

- No additional module brings Theoretical minimum computation complexity

Experiments





Bound

- Different objectives lead to different bounds.
- DQC could bring more extreme clipping bounds compared with DOBI
- The most extreme one leaves only 46% data in clipping bounds.

Learning rate	PSNR↑	SSIM↑
10^{-1}	37.82	0.9594
10^{-2}	37.87	0.9594
10^{-3}	37.78	0.9592
10^{-4}	37.74	0.9587

(a) Learning rate

Batch size	PSNR↑	SSIM↑
4	37.82	0.9594
8	37.83	0.9594
16	37.84	0.9593
32	37.87	0.9594

(b) Batch size

DOBI	DQC	PSNR↑	SSIM↑
		34.39	0.9202
✓		37.44	0.9568
	✓	37.32	0.9563
✓	✓	37.87	0.9594

(c) DOBI and DQC

Ablation

- Both DOBI and DQC improves the models' performance.

Conclusion

Contribution

We propose **2DQuant**, a dual-stage low bit post-training quantization method for image SR.

- **DOBI:** a fast MSE-based searching method to minimize the value heterogenization
- **DQC:** distillation beteen the FP model and the quantized model bring accurate quantizer parameters.
- **Performance:** Outperforms SOTA PTQ methods for SR.

Poster

- Time: Wed 11 Dec

  11 a.m. PST — 2 p.m. PST

Project

Thanks