Enhancing Quantized Models with FP model Knowledge

2025 하계 세미나 – 25.07.03

Sogang University
Vision & Display Systems Lab, Dept. of Electronic Engineering

Presented By
최재민

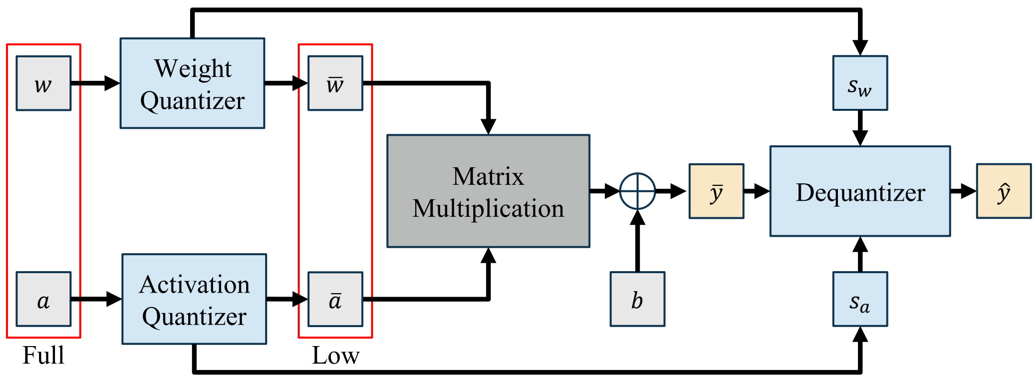
Outline

- Background
  - Quantization

- Papers
  - 2DQuant: Low-bit Post-Training Quantization for Image Super-Resolution [NeurIPS 2024]
  - Quantization without Tears [CVPR 2025]

Background

- Quantization
  - 일반적으로 performance와 model size는 비례하는 경향이 있음
    - Model size ↑ → inference time ↑, computational cost ↑
    - 메모리 용량이 제한적인 edge device 환경에서 한계가 존재함
  - Full-precision → Low-precision
    - Weight, activation을 lower precision으로 낮춘 후 연산을 수행하는 가속화 기법

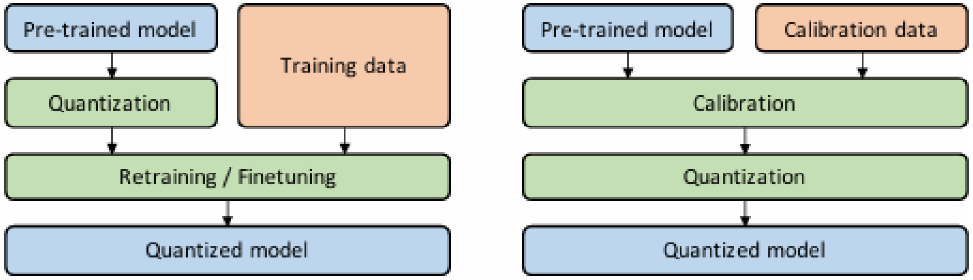


< Fig 1. Quantization의 가속화 원리 >

- Quantization process : $\bar{x} = Q(x) = clamp\left(\left\lfloor \frac{x}{s} \right\rceil + z, 0, 2^{bit} - 1\right), s = \frac{\max(x) - \min(x)}{2^{bit} - 1}, z = \left\lfloor -\frac{\min(x)}{s} \right\rceil$
- Dequantization process : $\hat{x} = s \cdot \bar{x}$
- Fully-Connected layer에서의 연산 : $f = \underline{WX} + B$ FP32 Matmul
- Quantized FC layer에서의 연산 : $\hat{f} = \widehat{W}\widehat{X} + B = (s_W\overline{W})(s_X\overline{X}) + B = s_W s_X(\underline{\overline{W}\overline{X}}) + B$ INT8 Matmul

Background

- Quantization



< Fig 2. Comparison between QAT and PTQ >

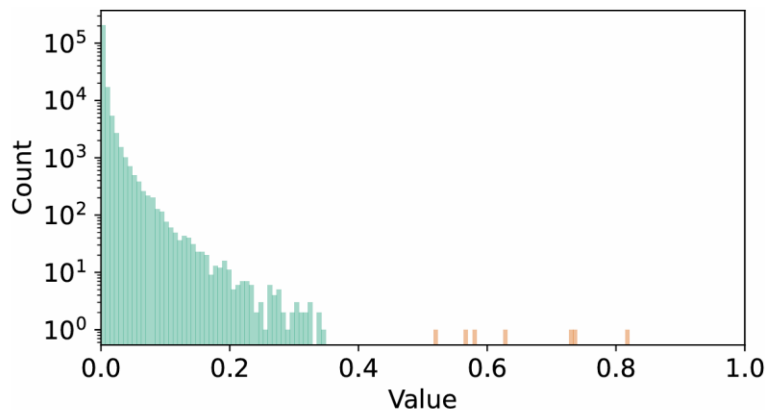
- Quantization-Aware Training (QAT)
  - Quantization 적용 후 pre-trained model의 train dataset으로 retraining/fine-tuning하는 방식
  - Retraining/fine-tuning 과정에서 많은 시간이 필요하지만 PTQ에 비해 좋은 성능

- Post-Training Quantization (PTQ)
  - 소량의 데이터(calibration dataset)만으로 pre-trained model에서의 quantization parameter 설정
  - Calibration 적용하여 lower-bit에 mapping, 이후 inference 수행 → inference time ↓
  - 소량의 데이터만을 사용하기 때문에 적은 시간만이 필요하지만 QAT에 비해 낮은 성능

Background

- Linear quantizer
  - Ex) $Q(x) = \bar{x} = clamp\left(\left\lfloor \frac{x}{s} \right\rceil + z, 0, 2^{bit} - 1\right), DQ(\bar{x}) = \hat{x} = \bar{x} \cdot s$
    - Uniform distribution과 같이 x의 분포가 고루 퍼져 있는 경우에 적합
    - CNN-based model에서 주로 사용됨

- Non-linear quantizer
  - Ex) $Q(x) = \bar{x} = clamp\left(\left\lfloor -\log_2 \frac{x}{s} \right\rceil, 0, 2^{bit} - 1\right), DQ(\bar{x}) = 2^{-\bar{x}} \cdot s$
    - Power-law distribution과 같이 x가 작은 값에 쏠려 있는 경우에 적합
    - Self-attention의 특성을 효율적으로 반영할 수 있어 transformer-based model에서 주로 사용됨



< Fig 3. Histogram of the post-Softmax activations >

2DQuant: Low-bit Post-Training Quantization for Image Super-Resolution
[NeurIPS 2024]

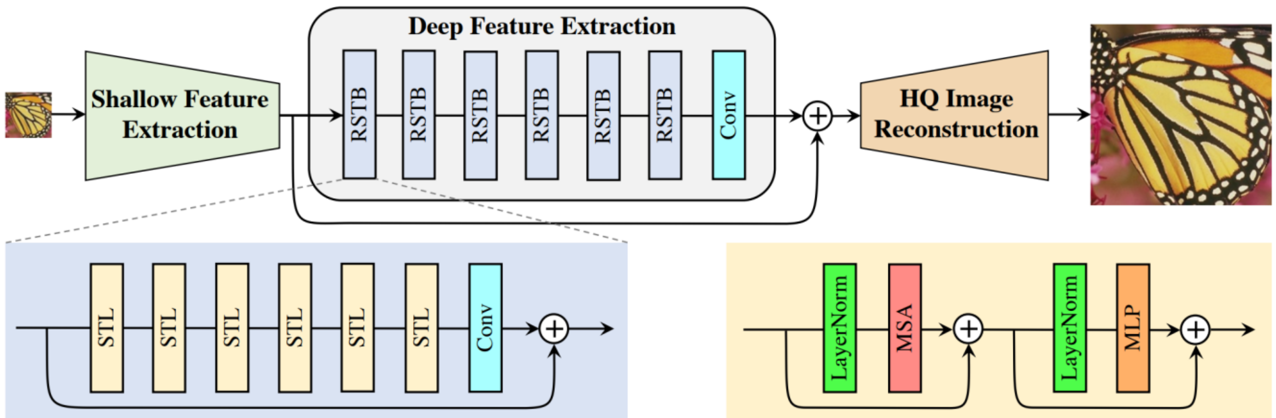
2DQuant[1)]

- Problem statements
  - 1. SR models이 컴퓨터 비전 분야의 다양한 task에 적용되고 있어 경량화가 필요함
  - 2. 기존에는 CNN-based SR model에 최적화되어 있는 quantization 방법론만이 존재함
  - 3. Transformer-based SR model에 적합한 형태의 quantization 방법론이 필요함
    - CNN과 transformer에서 자주 등장하는 activation distribution의 차이 고려

- Key contributions
  - 1. 최초로 Transformer-based SR model의 weight, activation을 관찰, 이에 적합한 방법론 제안
  - 2. Distribution-Oriented Bound Initialization (DOBI)
    - Weight, activation의 분포에 적합한 clipping range를 결정하는 최적화 방법론 제안
  - 3. Distillation Quantization Calibration (DQC)
    - Quantization으로 인해 하락한 성능을 FP model knowledge를 quantized model에 전이하는 방식 제안

2DQuant[1)]

- Objective
  - 본 논문의 목표 설정
    - Transformer-based SR model에 적합한 quantization 방법론 찾기
- Target model
  - SwinIR[2)]

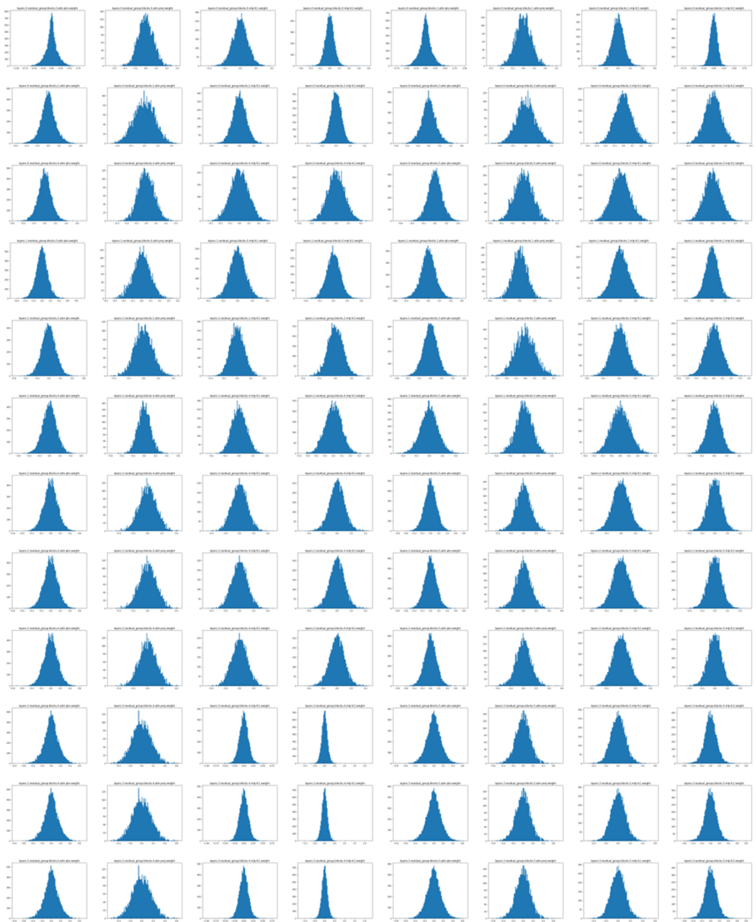


< Fig 1. SwinIR architecture >

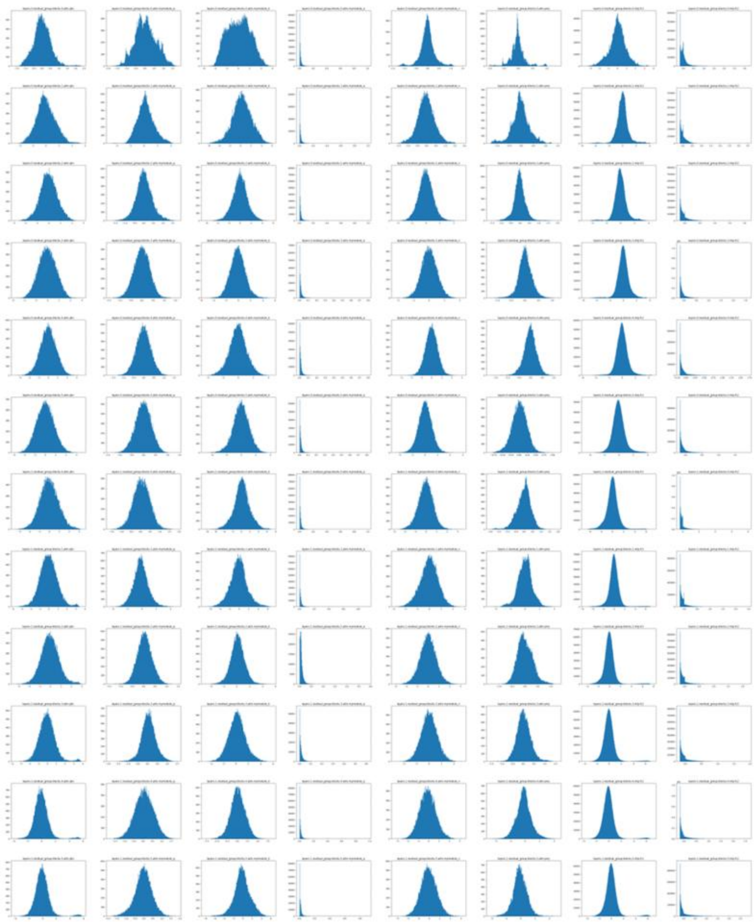
- SwinIR은 transformer-based SR model로, feature extraction 과정에서 SwinT layer를 활용
- 본 논문은 SwinIR을 target model로 설정, SwinT layer에서의 weight, activation을 분석

2DQuant[1)]

- Observations
  - Visualization of SwinIR weights & activations



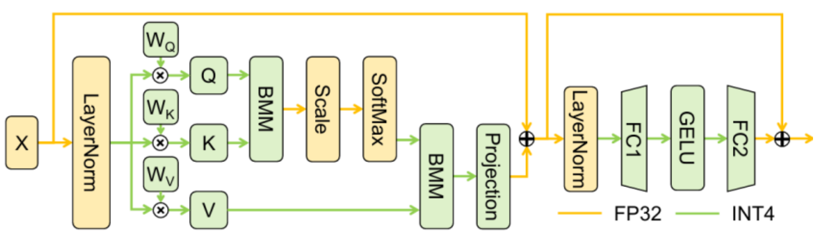
< Fig 2. Visualization of weights >



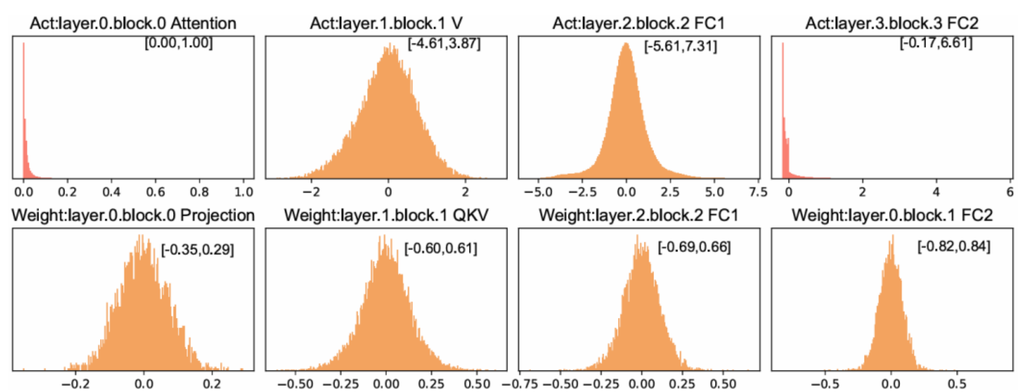
< Fig 3. Visualization of activations >

2DQuant[1)]

- Observations
  - Weight
    - 주로 평균이 0인 symmetric gaussian distribution 관찰
  - Activation
    - FC1의 input, V의 output에서 평균이 0인 symmetric gaussian distribution 관찰
    - FC2의 input, attention map에서 최솟값이 정해져 있는 asymmetric power-law distribution 관찰
  - Quantization scheme
    - Weight, activation distribution에 맞는 quantization scheme을 적용하여 성능 하락을 최소화
      → Distribution-Oriented Bound Initialization (DOBI) 제안



< Fig 4. SwinT layer >



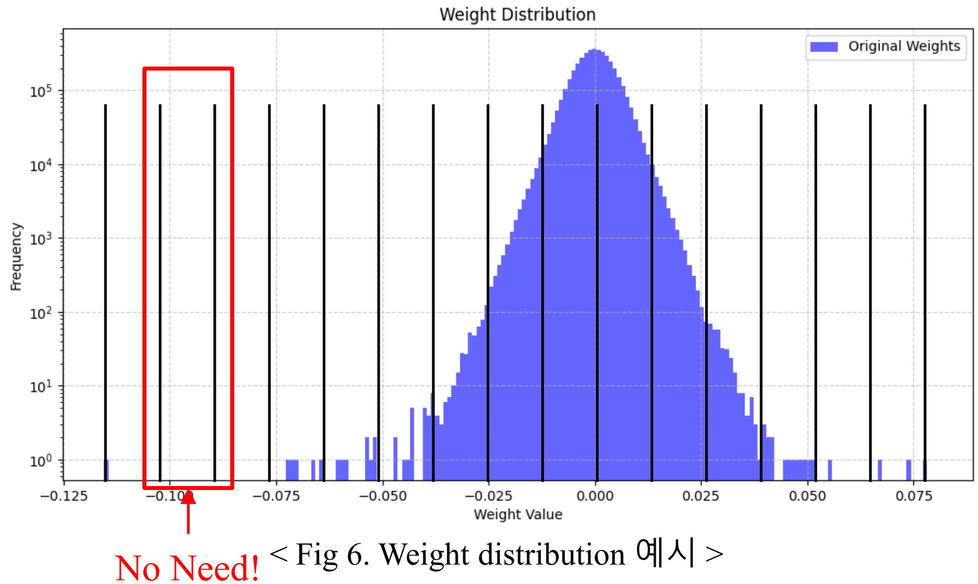
< Fig 5. Weight, activation distribution >

2DQuant[1)]

- Method
  - Preliminaries
    - Quantization process

$$v_c = Clip(v, l, u), \qquad v_r = Round\left(\frac{2^N - 1}{u - l}(v_c - l)\right), \qquad v_q = \frac{u - l}{2^N - 1} v_r + l$$

- 만약 input v에 outlier가 존재한다면 quantization bin의 간격이 과하게 넓어질 수 있음
- 이로 인해 어떠한 값도 할당되지 않는 불필요한 quantization bin이 설정될 수 있음
- Lower-bound와 upper-bound에 해당하는 l, u를 적절히 설정 후 clipping 하여 outlier issue 해결 가능



< Fig 6. Weight distribution 예시 >

서강대학교 SOGANG UNIVERSITY

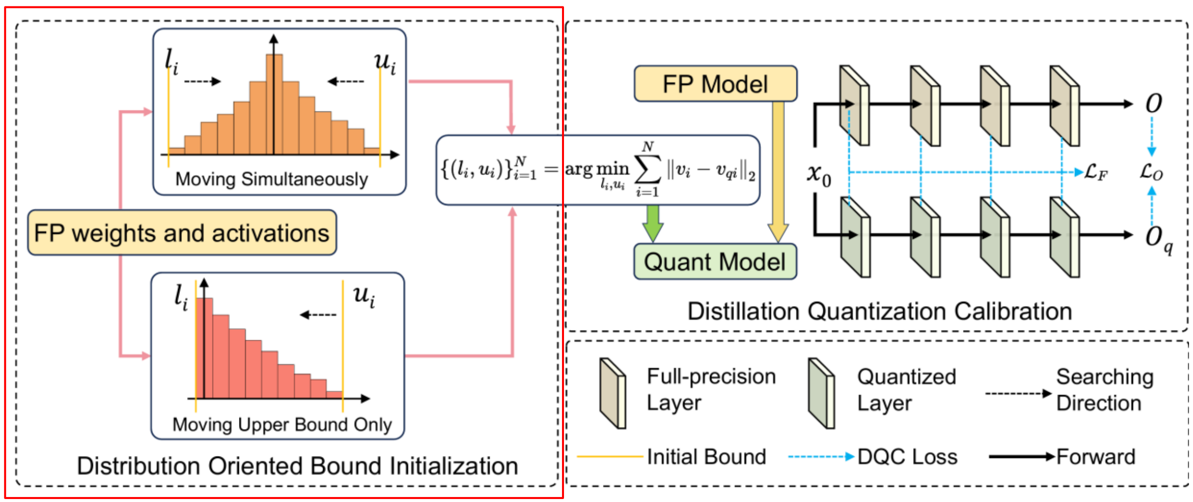
VDS LAB

2DQuant[1)]

- Method
  - Distribution-Oriented Bound Initialization (DOBI)
    - l, u를 적절히 설정하는 것은 결국 최적화 문제

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

    - l, u의 후보군을 search space로 설정, MSE loss를 최소화하는 (l_i, u_i)를 layer 단위로 연산



< Fig 7. Distribution-Oriented Bound Initialization (DOBI) >

2DQuant[1)]

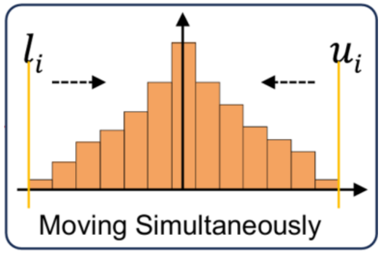
- Method
  - Distribution-Oriented Bound Initialization (DOBI)

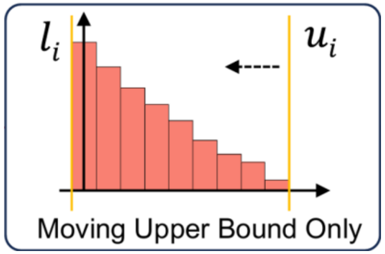
Algorithm 1: DOBI pipeline

Data: Data to be quantized v, the number of search point K, bit b
Result: Clip bound l, u
$l \leftarrow \min(v)$, $u \leftarrow \max(v)$;
$min\_mse \leftarrow +\infty$;
1
if *v is symmetrical* **then**
    $\Delta l \leftarrow (\max(v) - \min(v))/2K$;
else
    $\Delta l \leftarrow 0$;
end
$\Delta u \leftarrow (\max(v) - \min(v))/2K$;
2
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

Optimization problem : $\{(l_i, u_i)\}_{i=1}^{N} = argmin_{(l_i,u_i)} \sum_{i=1}^{N} \left\| v_i - v_{qi} \right\|_2$

1. $\Delta l, \Delta u$ 계산
- symmetric gaussian distribution인 경우 (l_i, u_i)를 모두 최적화
- asymmetric power-law distribution인 경우 l_i는 고정, u_i만 최적화





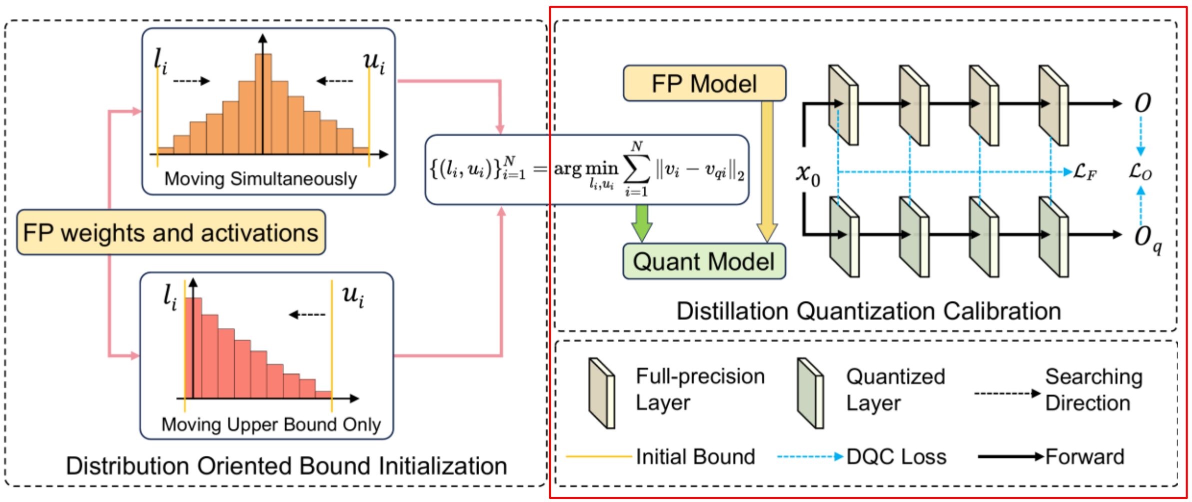
2. MSE loss를 최소로 하는 (l_i, u_i) 결정
- Search point K만큼 반복하는 동안의 최적의 (l_i, u_i) 를 탐색

2DQuant[1)]

- Method
  - Distillation Quantization Calibration (DQC)
    - Quantized model의 가장 이상적인 weight와 activation은 결국 FP model의 weight와 activation
    - FP model knowledge를 quantized model에 전이하는 knowledge distillation 적용 가능
    - Distillation loss for model output and feature map ($learnable\ parameter = l_i, u_i$)

$$L_O = \frac{1}{C_O H_O W_O} \left\| O - O_q \right\|_1, L_F = \sum_i^N \frac{1}{C_i H_i W_i} \left\| \frac{F_i}{\left\| F_i \right\|_2} - \frac{F_{qi}}{\left\| F_{qi} \right\|_2} \right\|_2$$

$$L = L_O + \lambda L_F$$



< Fig 8. Distillation Quantization Calibration (DQC) >

1) Liu, K., Qin, H., Guo, Y., Yuan, X., Kong, L., Chen, G., & Zhang, Y. (2024). 2DQuant: Low-bit post-training quantization for image super-resolution. *Advances in Neural Information Processing Systems*, *37*, 71068-71084.

2DQuant[1)]

- Experiments
  - Quantitative results

Method	Bit	Set5 (×2) PSNR↑	Set5 (×2) SSIM↑	Set14 (×2) PSNR↑	Set14 (×2) SSIM↑	B100 (×2) PSNR↑	B100 (×2) SSIM↑	Urban100 (×2) PSNR↑	Urban100 (×2) SSIM↑	Manga109 (×2) PSNR↑	Manga109 (×2) SSIM↑
SwinIR-light [29]	32	38.15	0.9611	33.86	0.9206	32.31	0.9012	32.76	0.9340	39.11	0.9781
Bicubic	32	32.25	0.9118	29.25	0.8406	28.68	0.8104	25.96	0.8088	29.17	0.9128
MinMax [22]	4	34.39	0.9202	30.55	0.8512	29.72	0.8409	28.40	0.8520	33.70	0.9411
Percentile [27]	4	37.37	0.9568	32.96	0.9113	31.61	0.8917	31.17	0.9180	37.19	0.9714
EDSR† [30, 39]	4	36.33	0.9420	32.75	0.9040	31.48	0.8840	30.90	0.9130	N/A	N/A
DBDC+Pac [39]	4	37.18	0.9550	32.86	0.9106	31.56	0.8908	30.66	0.9110	36.76	0.9692
DOBI (Ours)	4	37.44	0.9568	33.15	0.9132	31.75	0.8937	31.29	0.9193	37.93	0.9743
2DQuant (Ours)	4	37.87	0.9594	33.41	0.9161	32.02	0.8971	31.84	0.9251	38.31	0.9761
MinMax [22]	3	28.19	0.6961	26.40	0.6478	25.83	0.6225	25.19	0.6773	28.97	0.7740
Percentile [27]	3	34.37	0.9170	31.04	0.8646	29.82	0.8339	28.25	0.8417	33.43	0.9214
DBDC+Pac [39]	3	35.07	0.9350	31.52	0.8873	30.47	0.8665	28.44	0.8709	34.01	0.9487
DOBI (Ours)	3	36.37	0.9496	32.33	0.9041	31.12	0.8836	29.65	0.8967	36.18	0.9661
2DQuant (Ours)	3	37.32	0.9567	32.85	0.9106	31.60	0.8911	30.45	0.9086	37.24	0.9722
MinMax [22]	2	33.88	0.9185	30.81	0.8748	29.99	0.8535	27.48	0.8501	31.86	0.9306
Percentile [27]	2	30.82	0.8016	28.80	0.7616	27.95	0.7232	26.30	0.7378	30.37	0.8351
DBDC+Pac [39]	2	34.55	0.9386	31.12	0.8912	30.27	0.8706	27.63	0.8649	32.15	0.9467
DOBI (Ours)	2	35.25	0.9361	31.72	0.8917	30.62	0.8699	28.52	0.8727	34.65	0.9529
2DQuant (Ours)	2	36.00	0.9497	31.98	0.9012	30.91	0.8810	28.62	0.8819	34.40	0.9602

Method	Bit	Set5 (×3) PSNR↑	Set5 (×3) SSIM↑	Set14 (×3) PSNR↑	Set14 (×3) SSIM↑	B100 (×3) PSNR↑	B100 (×3) SSIM↑	Urban100 (×3) PSNR↑	Urban100 (×3) SSIM↑	Manga109 (×3) PSNR↑	Manga109 (×3) SSIM↑
SwinIR-light [29]	32	34.63	0.9290	30.54	0.8464	29.20	0.8082	28.66	0.8624	33.99	0.9478
Bicubic	32	29.54	0.8516	27.04	0.7551	26.78	0.7187	24.00	0.7144	26.16	0.8384
MinMax [22]	4	31.66	0.8784	28.17	0.7641	27.19	0.7257	25.60	0.7485	29.98	0.8854
Percentile [27]	4	33.34	0.9137	29.61	0.8275	28.49	0.7899	27.06	0.8242	32.10	0.9303
DBDC+Pac [39]	4	33.42	0.9143	29.69	0.8261	28.51	0.7869	27.05	0.8217	31.89	0.9274
DOBI (Ours)	4	33.78	0.9200	29.87	0.8338	28.72	0.7970	27.53	0.8391	32.57	0.9367
2DQuant (Ours)	4	34.06	0.9231	30.12	0.8374	28.89	0.7988	27.69	0.8405	32.88	0.9389
MinMax [22]	3	26.01	0.6260	23.41	0.4944	22.46	0.4182	21.70	0.4730	24.68	0.6224
Percentile [27]	3	30.91	0.8426	28.02	0.7545	27.23	0.7183	25.32	0.7349	29.43	0.8537
DBDC+Pac [39]	3	30.91	0.8445	28.02	0.7538	26.99	0.6937	25.10	0.7122	28.84	0.8403
DOBI (Ours)	3	32.85	0.9075	29.33	0.8200	28.27	0.7820	26.36	0.8036	31.14	0.9178
2DQuant (Ours)	3	33.24	0.9135	29.56	0.8255	28.50	0.7873	26.65	0.8116	31.46	0.9235
MinMax [22]	2	26.05	0.5827	24.74	0.5302	24.42	0.4973	22.87	0.5155	24.66	0.5652
Percentile [27]	2	25.30	0.5677	23.60	0.4890	23.77	0.4751	22.33	0.4965	24.65	0.5882
DBDC+Pac [39]	2	29.96	0.8254	27.53	0.7507	27.05	0.7136	24.57	0.7117	27.23	0.8213
DOBI (Ours)	2	30.54	0.8321	27.74	0.7312	26.69	0.6643	24.80	0.6797	28.18	0.7993
2DQuant (Ours)	2	31.62	0.8887	28.54	0.8038	27.85	0.7679	25.30	0.7685	28.46	0.8814

< Fig 9. SR performance – scale (× 2, 3) >

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

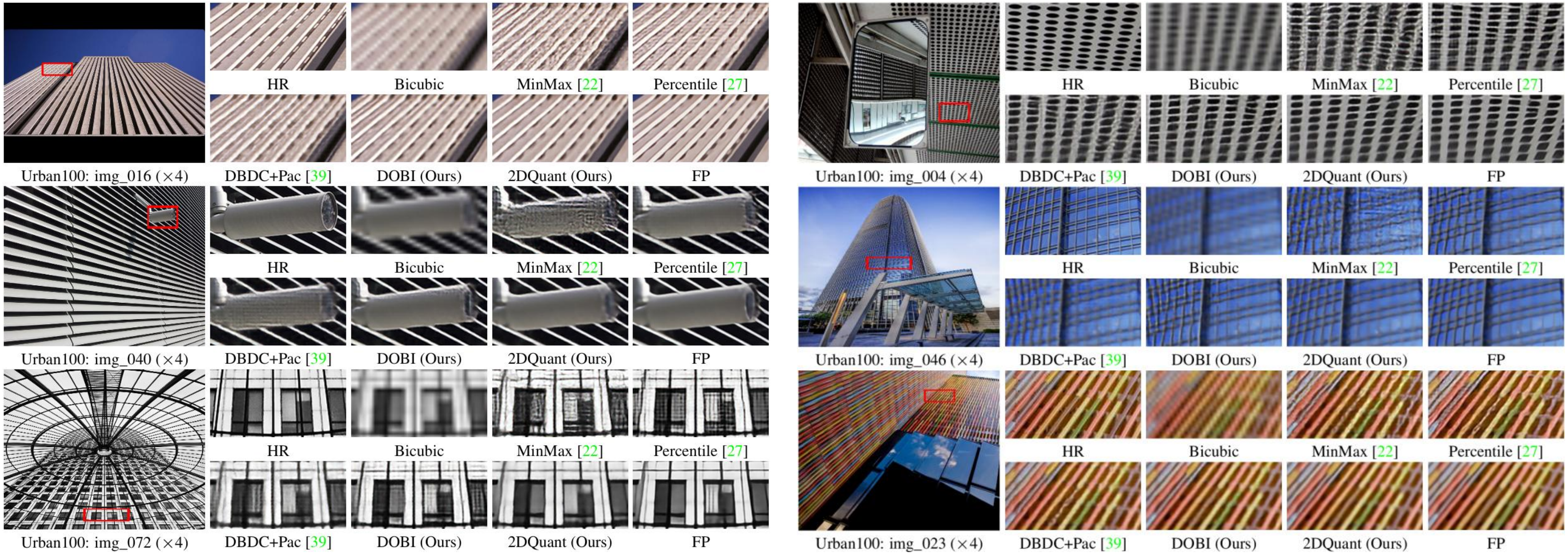
< Fig 10. SR performance – scale (× 4) >

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

< Fig 11. Complexity and performance – scale (× 4) >

2DQuant[1)]

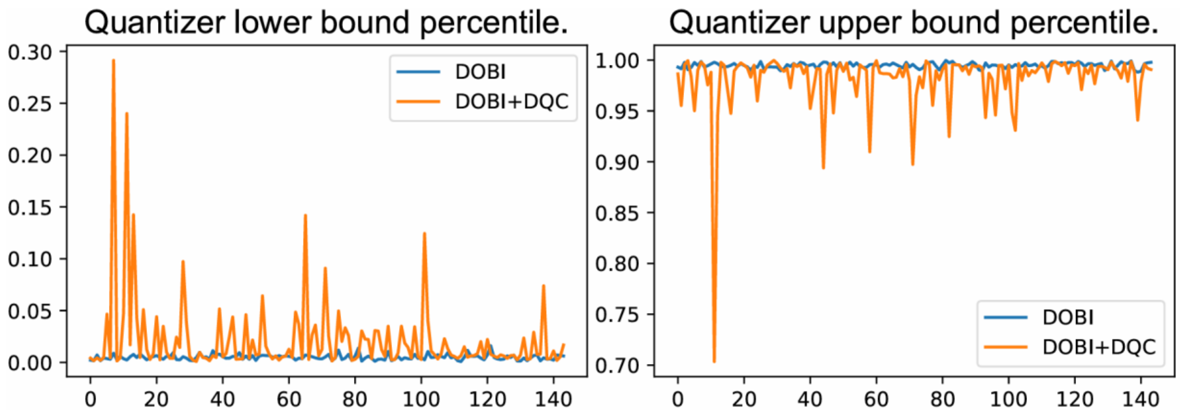
- Experiments
  - Qualitative results



< Fig 12. Visualization comparison for scale (× 4) >

2DQuant[1)]

- Experiments
  - Two bounds after DOBI & DQC



< Fig 13. The bound percentile of DOBI and DQC >

  - Ablations

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

< Fig 14. Ablation studies – learning rate, batch size, DOBI & DQC >

1) Fu, M., Yu, H., Shao, J., Zhou, J., Zhu, K., & Wu, J. (2025). Quantization without tears. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 4462-4472).

Quantization without Tears [CVPR 2025]

QwT[1)]

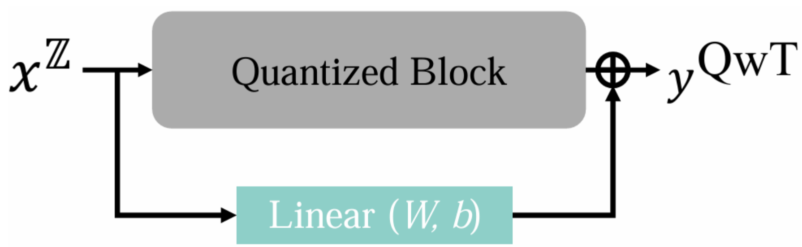
- Problem statements
  - 1. The speed-accuracy dilemma
    - PTQ는 빠르지만 성능이 낮고, QAT는 느리지만 성능이 좋음
  - 2. Complexity
    - PTQ와 QAT 모두 수학적이며, hyperparameter에 의존적인 경우가 많음
  - 3. Missing generality
    - 특정 model에 최적화되어 있는 형태의 방법론들이 많음

- Key contributions
  - 1. Quantized model과 FP model의 architecture가 같아야만 한다는 고정관념에서 벗어나는 새로운 패러다임 제안
  - 2. QwT (Quantization without Tears) module 제안
    - Quantized model의 매 block 마다 1개의 linear layer를 연결
    - Quantization error를 최소화하는 형태로 linear layer의 weight를 초기화하여 성능 보상
    - FP model knowledge를 QwT module에 전이하는 distillation 방법론 제안

1) Fu, M., Yu, H., Shao, J., Zhou, J., Zhu, K., & Wu, J. (2025). Quantization without tears. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 4462-4472).

QwT[1)]

- Idea
  - Quantized model과 FP model의 architecture가 완전히 동일할 필요가 없음
  - Quantized block l^z마다 QwT module c_l을 residual 방식으로 연결하여 성능 보완
    - $y^{QwT} = l^z(x_z) + c_l(x_z),\ (x_z: quantized\ input)$



< Fig 1. QwT in one block >

- Objective
  - Quantization의 주요 목표
    - Quantized block output와 FP block output의 차이를 줄이는 것
  - 본 논문의 주요 목표
    - QwT module은 Quantized block output과 FP block output 사이의 MSE loss를 최소화해야 함
    - 결국 MSE loss를 최소화하는 최적화 문제

QwT[1)]

- Method

  - Weight initialization for QwT module

    - QwT module의 weight는 MSE loss를 최소화하도록 초기화되어야 함
    - $\boldsymbol{W}' = [\boldsymbol{W}|\boldsymbol{b}], \boldsymbol{X}'_z = \begin{bmatrix} \boldsymbol{X}_z \\ \boldsymbol{1} \end{bmatrix}, \left(\boldsymbol{W} \in \mathbb{R}^{d_{out} \times d_{in}},\ \boldsymbol{b} \in \mathbb{R}^{d_{out} \times 1}, \boldsymbol{X}_z \in \mathbb{R}^{d_{in} \times N}, \boldsymbol{1} \in \mathbb{R}^{1 \times N}, \boldsymbol{Y}_z \in \mathbb{R}^{d_{in} \times N}\right)$

$$\boldsymbol{W}^* = argmin_{\boldsymbol{W}'} \left|\left|\boldsymbol{Y}^{QwT} - \boldsymbol{Y}^{FP}\right|\right|_2$$

    - $f(\boldsymbol{W}') = \left|\left|\boldsymbol{Y}^{QwT} - \boldsymbol{Y}^{FP}\right|\right|_2 = \left|\left|\boldsymbol{Y}_z + \boldsymbol{W}'\boldsymbol{X}'_z - \boldsymbol{Y}^{FP}\right|\right|_2 \Rightarrow \nabla f(\boldsymbol{W}') = 2(\boldsymbol{Y}_z + \boldsymbol{W}'\boldsymbol{X}'_z - \boldsymbol{Y}^{FP})\boldsymbol{X}'^{\,T}_z = 0$

$$\Rightarrow \boldsymbol{W}^* = (\boldsymbol{Y}^{FP} - \boldsymbol{Y}_z)\boldsymbol{X}'^{\,T}_z\left(\boldsymbol{X}'_z\boldsymbol{X}'^{\,T}_z\right)^{-1}$$

    - 위 수식에 따라 MSE loss의 gradient를 0으로 만드는 $\boldsymbol{W}^*$를 QwT module의 weight로 초기화

  - 1 epoch Fine-tuning for QwT module

    - QwT module의 weight가 초기화되어 있는 상태에서 1 epoch fine-tuning으로 성능 보상
    - Distillation loss for classification and class token

$$L_{cls} = -\sum_{i}^{N} t_i \log p_i \,, L_{dis} = \left|\left|T^{FP} - T_z\right|\right|_2$$

$$L = L_{cls} + L_{dis}$$

QwT[1)]

- Experiments
  - Quantitative results

Network	Method	#Bits	Size	Top-1
ViT-S	Full-precision	32/32	88.2	81.4
	IGQ-ViT† [38]	4/4	-	**73.6**
	RepQ-ViT [27]	4/4	11.9	65.8
	RepQ-ViT + QwT	4/4	15.4	70.8
	RepQ-ViT + QwT*	4/4	15.4	72.9
	IGQ-ViT† [38]	6/6	-	80.8
	RepQ-ViT [27]	6/6	17.2	80.5
	RepQ-ViT + QwT	6/6	20.7	80.7
	RepQ-ViT + QwT*	6/6	20.7	**80.8**
ViT-B	Full-precision	32/32	346.3	84.5
	IGQ-ViT† [38]	4/4	-	**79.3**
	RepQ-ViT [27]	4/4	44.9	68.5
	RepQ-ViT + QwT	4/4	59.1	76.3
	RepQ-ViT + QwT*	4/4	59.1	78.5
	IGQ-ViT† [38]	6/6	-	83.8
	RepQ-ViT [27]	6/6	66.2	83.6
	RepQ-ViT + QwT	6/6	80.4	83.9
	RepQ-ViT + QwT*	6/6	80.4	**84.0**

Network	Method	#Bits	Size	Top-1
DeiT-T	Full-precision	32/32	22.9	72.2
	IGQ-ViT† [38]	4/4	-	62.5
	RepQ-ViT [27]	4/4	3.3	58.2
	RepQ-ViT + QwT	4/4	4.2	61.4
	RepQ-ViT + QwT*	4/4	4.2	**64.8**
	IGQ-ViT† [38]	6/6	-	71.2
	RepQ-ViT [27]	6/6	4.6	71.0
	RepQ-ViT + QwT	6/6	5.5	71.2
	RepQ-ViT + QwT*	6/6	5.5	**71.6**
DeiT-S	Full-precision	32/32	88.2	79.9
	IGQ-ViT† [38]	4/4	-	74.7
	RepQ-ViT [27]	4/4	11.9	69.0
	RepQ-ViT + QwT	4/4	15.4	71.5
	RepQ-ViT + QwT*	4/4	15.4	**75.2**
	IGQ-ViT† [38]	6/6	-	79.3
	RepQ-ViT [27]	6/6	17.2	78.9
	RepQ-ViT + QwT	6/6	20.7	79.1
	RepQ-ViT + QwT*	6/6	20.7	**79.3**

Network	Method	#Bits	Size	Top-1
Swin-T	Full-precision	32/32	113.2	81.4
	IGQ-ViT† [38]	4/4	-	77.8
	RepQ-ViT [27]	4/4	14.9	73.0
	RepQ-ViT + QwT	4/4	19.2	75.5
	RepQ-ViT + QwT*	4/4	19.2	**79.3**
	IGQ-ViT† [38]	6/6	-	80.9
	RepQ-ViT [27]	6/6	21.7	80.6
	RepQ-ViT + QwT	6/6	26.0	80.7
	RepQ-ViT + QwT*	6/6	26.0	**80.9**
Swin-S	Full-precision	32/32	198.4	83.2
	IGQ-ViT† [38]	4/4	-	81.0
	RepQ-ViT [27]	4/4	25.8	80.2
	RepQ-ViT + QwT	4/4	33.7	80.4
	RepQ-ViT + QwT*	4/4	33.7	**81.9**
	IGQ-ViT† [38]	6/6	-	82.9
	RepQ-ViT [27]	6/6	38.0	82.8
	RepQ-ViT + QwT	6/6	45.9	82.9
	RepQ-ViT + QwT*	6/6	45.9	**82.9**

< Fig 2. Classification – ViT variants >

Quant Setup	Method	#Bits	Size (MB)	Top-1
Vision	Full-precision	32/32	607.2	63.4
	RepQ-ViT [27]	6/6	323.5	59.2
	RepQ-ViT + QwT	6/6	336.8	**60.3**
	RepQ-ViT [27]	8/8	345.3	62.9
	RepQ-ViT + QwT	8/8	359.5	**63.0**
Vision & Text	Full-precision	32/32	607.2	63.4
	RepQ-ViT [27]	6/6	200.8	29.8
	RepQ-ViT + QwT	6/6	221.3	**43.5**
	RepQ-ViT [27]	8/8	232.1	38.7
	RepQ-ViT + QwT	8/8	252.6	**54.6**

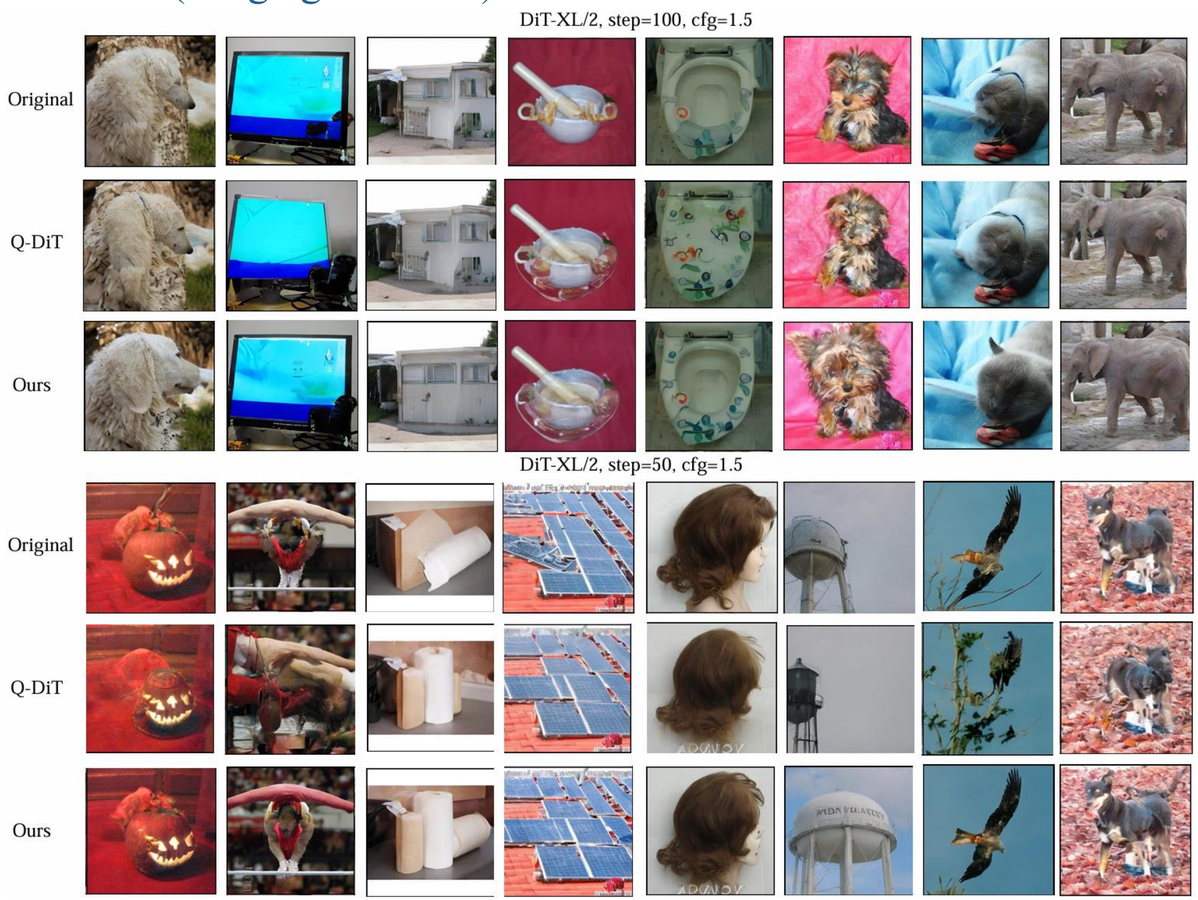
< Fig 3. Zero-shot classification – CLIP >

Model	Bit-width (W/A)	Method	Size (MB)	FID (↓)	sFID (↓)	IS (↑)	Precision (↑)	Recall (↑)
DiT-XL/2 (steps = 100)	16/16	FP	1349	12.40	19.11	116.68	0.6605	-
	4/8	PTQ4DM	339	252.31	82.44	2.74	0.0125	-
		RepQ-ViT	339	315.85	139.99	2.11	0.0067	-
		GPTQ	351	25.48	25.57	73.46	0.5392	-
		Q-DiT	347	15.76	19.84	98.78	**0.6395**	-
		Q-DiT + QwT	361	**15.35**	**19.63**	**104.04**	0.6373	**0.7478**
DiT-XL/2 (steps = 100, cfg = 1.5)	16/16	FP	1349	5.31	17.61	245.85	0.8077	-
	4/8	PTQ4DM	339	255.06	84.63	2.76	0.0110	-
		RepQ-ViT	339	311.31	138.58	2.18	0.0072	-
		GPTQ	351	7.66	20.76	193.76	0.7261	-
		Q-DiT	347	6.40	18.60	211.72	0.7609	-
		Q-DiT + QwT	361	**5.86**	**18.29**	**221.66**	**0.7678**	**0.6915**
DiT-XL/2 (steps = 50)	16/16	FP	1349	13.47	19.31	114.71	0.6601	-
	4/8	PTQ4DM	339	256.15	83.45	2.73	0.0150	-
		RepQ-ViT	339	324.25	142.98	2.12	0.0062	-
		GPTQ	351	26.31	25.54	69.73	0.5388	-
		Q-DiT	347	17.42	19.95	97.52	0.6219	-
		Q-DiT + QwT	361	**17.02**	**19.57**	**99.62**	**0.6302**	**0.7582**

< Fig 4. Diffusion (ImageNet 256 x 256) – DiT-XL/2 >

QwT[1)]

- Experiments
  - Qualitative results (image generation)



< Fig 5. Visualization (ImageNet 256 x 256) – DiT-XL/2 >

1) Liu, K., Qin, H., Guo, Y., Yuan, X., Kong, L., Chen, G., & Zhang, Y. (2024). 2DQuant: Low-bit post-training quantization for image super-resolution. *Advances in Neural Information Processing Systems*, *37*, 71068-71084.
2) Fu, M., Yu, H., Shao, J., Zhou, J., Zhu, K., & Wu, J. (2025). Quantization without tears. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 4462-4472).

Conclusion

- 2DQuant[1)]
  - Key contributions
    - 1. Transformer-based SR model에 적합한 quantization 방법론을 최초로 제안
    - 2. Weight, activation distribution에 적합한 lower-bound와 upper-bound를 최적화하는 방법론 제안
  - Limitations
    - 1. Optimal lower-bound, upper-bound searching에서 오랜 시간을 필요로 함
    - 2. 관찰에 의존한 방법론, 다른 형태의 distribution에서는 적합하지 않을 수 있음
    - 3. Transformer-based SR model에 quantization을 적용했을 뿐, SR task에 특화된 형태의 방법론이 아님
  - Future works
    - 1. Lower-bound, upper-bound searching 과정에 더욱 효율적인 searching algorithm 적용 가능
    - 2. SR task에서 중요한 정보들을 보상하기 위한 distillation 방법론 설계

1) Liu, K., Qin, H., Guo, Y., Yuan, X., Kong, L., Chen, G., & Zhang, Y. (2024). 2DQuant: Low-bit post-training quantization for image super-resolution. *Advances in Neural Information Processing Systems*, *37*, 71068-71084.
2) Fu, M., Yu, H., Shao, J., Zhou, J., Zhu, K., & Wu, J. (2025). Quantization without tears. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 4462-4472).

Conclusion

- QwT[2)]
  - Key contributions
    - 1. Quantized model과 FP model의 architecture가 같을 필요가 없다는 새로운 패러다임
    - 2. 오직 1개 linear layer의 weight 초기화, 1 epoch fine-tuning으로 기존 PTQ, QAT 방법의 성능을 능가
  - Limitations
    - 1. QwT module로 인해 model parameter의 개수가 증가함
    - 2. Real quantization 상황이라면 quantized model은 INT로 저장되는 반면 QwT module은 FP로 저장됨
      - Mixed precision issue, 하드웨어와의 호환성이 낮을 우려가 존재함
  - Future works
    - 1. QwT module의 weight를 quantized model에 병합시켜 parameter 개수를 유지하도록 설계
    - 2. QwT module의 weight를 FP가 아닌 INT로 저장, 이로 인해 발생한 quantization error를 재보상

감사합니다