Retraining-free Model Quantization via One-Shot Weight-Coupling Learning

Chen Tang*, Yuan Meng*, Jiacheng Jiang, Shuzhao Xie, Rongwei Lu, Xinzhu Ma, Zhi Wang, Wenwu Zhu

paper code

CVPR JUNE 17-21, 2024 SEATTLE, WA

Contributions of Retraining-free Quantization

- Retraining-free Quantization (ReFQ) devises a one-shot training-then-searching paradigm for mixed-precision model compression. In the first stage, all potential bit-width configurations are coupled and thus optimized simultaneously within a set of shared weights. In the second stage, a lightweight inference algorithm is applied to determine the optimal bit-width for each layer.
- To solve the bit-width interference problem in the first stage, we design a bit-width scheduler to dynamically freeze the most turbulent bit-width of layers during training, to ensure the rest bit-widths converged properly. We then present an information distortion mitigation technique to align the behavior of the bad-performing bit-widths to the well-performing ones.
- In second stage, an inference-only greedy search scheme is devised to evaluate configurations without introducing any additional training costs.

Overview of Retraining-free Quantization

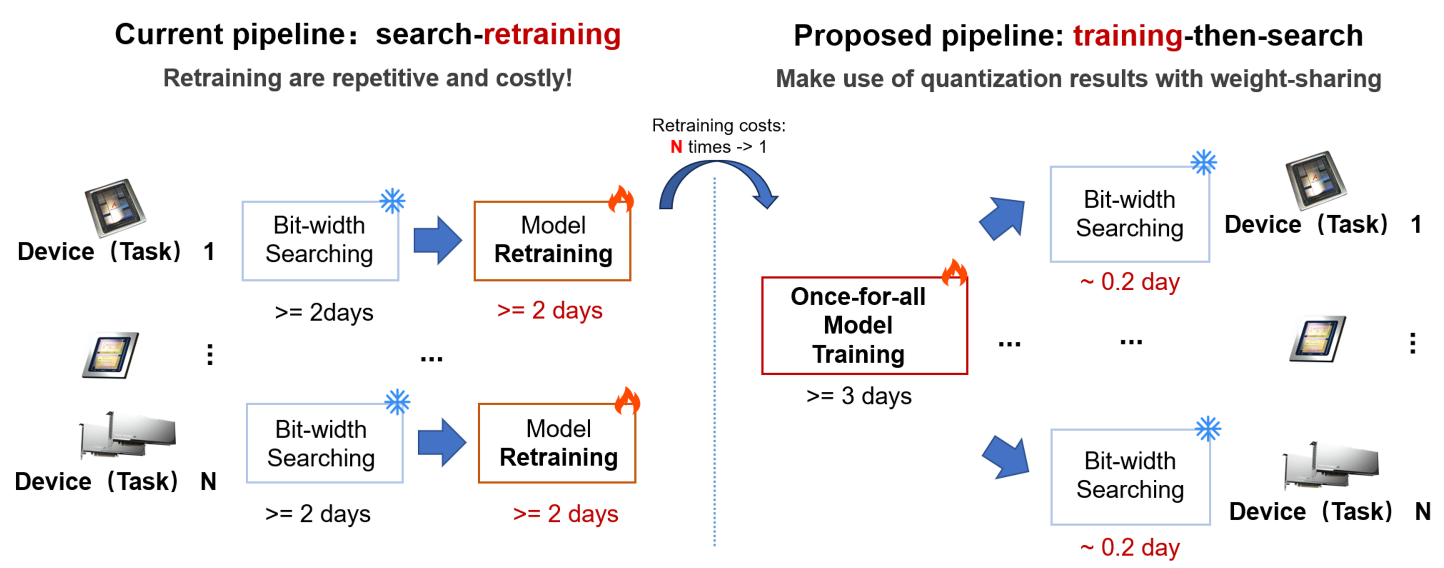

Fig. 1

- <u>One-shot Training for All Deployments</u>: unlike previous searching-then-retraining pipelines, our one-shot method shifts the costly training process to the front end of the pipeline, forming a training-then-searching pipeline, thus the training only performs once to support diverse deployment requirements.
- <u>Interference-aware Training</u>: we identify the bit-width interference problem and design several key techniques to improve the convergence of one-shot model.

Weight-sharing for Quantization

- To jointly learn the quantized weights, we approximate the expectation term over the whole search space of bit-width by randomly sampling the bit-width configurations at each training step using a set of shared weights for different bit-width, the forward propagation is defined as follow:

$$\arg\min_{\mathbf{W}} \quad \mathbb{E}_{\mathcal{S}\sim\mathcal{A}}\left[\mathcal{L}(f(\mathbf{x};\mathcal{S},w^{(\mathcal{S})}),\mathbf{y})\right] \approx \arg\min_{\mathbf{W}} \frac{1}{K}\sum_{\mathcal{S}_k\sim \mathrm{U}(\mathcal{A})}^{K}\left[\mathcal{L}(f(\mathbf{x};\mathcal{S}_k,\hat{\mathbf{W}}^{(\mathcal{S}_k)}),\mathbf{y})\right],$$

Bit-width Interference among Highly Coupled Weights during Optimization

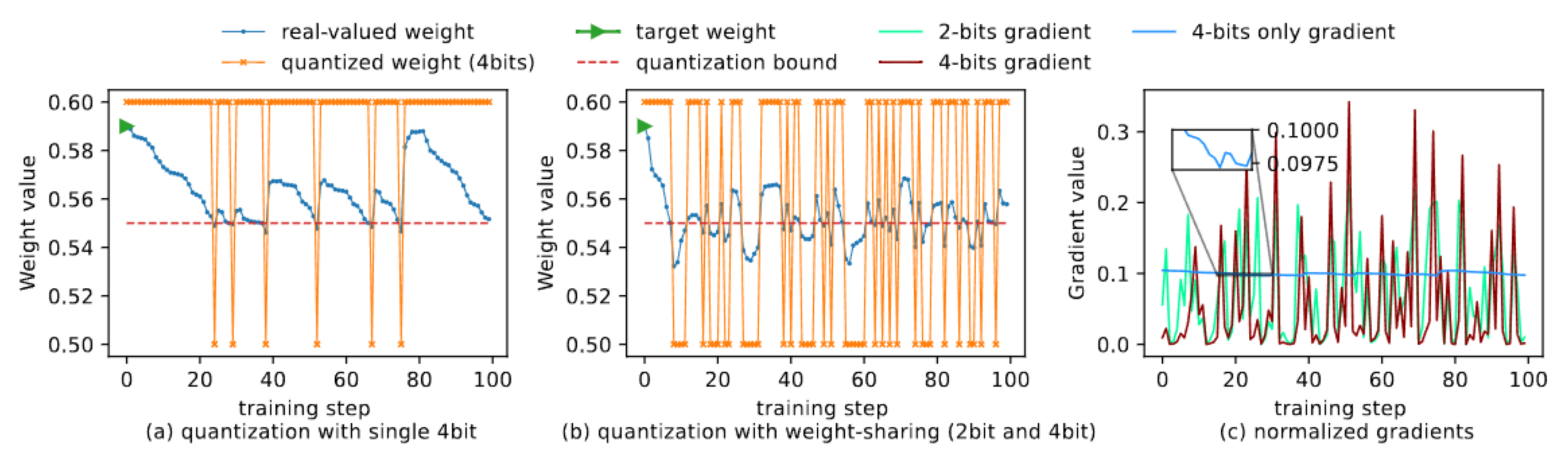

Fig. 2

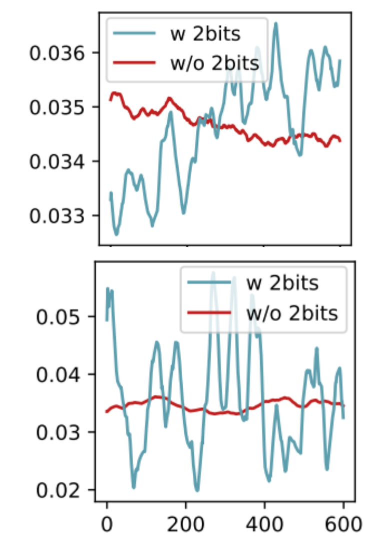

Fig. 3

- <u>Bit-width interference</u>: While training becomes possible with weight-sharing quantization, we have observed it exhibits training instability and the weight moves closer to the quantization bound more frequently (Fig.2), introducing extra small bit-widths induces significant random oscillations for the learning process (Fig.3), signifying hindering model convergence.

Training and Search Techniques of One-shot Retraining-free Quantization

To effectively train the weight-sharing model, we develop:

- <u>Dynamic Bit-width Schedule</u> selectively freezes the bit-width causing weight interference from unstable bit-width set Ω to ensure proper convergence for remaining bit-widths during training:

$$\Omega \leftarrow \mathtt{TopKToFreeze}(\hat{\Delta}\mathbf{W}^{\text{unstable}};\mathcal{K}), \quad \text{where}$$

$$\hat{\Delta}\mathbf{W}^{\text{unstable}} \triangleq \{\hat{\Delta}\mathbf{W}_l^{\text{unstable}}\}_{l=0}^{L-1} \quad \text{where} \quad \hat{\Delta}\mathbf{W}_l^{\text{unstable}} = \sum_{b\in B^{(w)}} \frac{1}{2^b}\frac{1}{\|\mathbf{W}_l\|_0}\cdot\sum_{q_b\in\Phi_b}\sum_{\mathbf{w}_{l,*}\in\mathbf{W}_l} \mathbb{1}_{|\mathbf{w}_{l,*}|\le\gamma\times(\frac{1-\epsilon}{2}+\frac{q_b}{\gamma})},$$

- <u>Information Distortion Mitigation</u> aims to align the behavior of weak bit-widths to their high-performing counterparts by optimizing their rectified Euclidea

$$\mathbb{E}\left[\|\max\{Q,\frac{\mathbf{O}^S-\mu(\mathbf{O}^S)}{\sqrt{\sigma(\mathbf{O}^S)+\zeta}}\eta_{\mathbf{O}^S}+\xi_{\mathbf{O}^S}\}-\max\{Q,\frac{\mathbf{O}^H-\mu(\mathbf{O}^H)}{\sqrt{\sigma(\mathbf{O}^H)+\zeta}}\eta_{\mathbf{O}^H}+\xi_{\mathbf{O}^H}\}\|\right],$$

- <u>Quantization Error Minimization</u> optimizes the distance between full-precision latent weights and the quantized weights to reduce the quantization error:

$$\mathcal{L}_{\mathrm{QE}} = \frac{1}{K}\sum_{k=0}^{K-1}||\hat{\mathbf{W}}^{(\mathcal{S}_k)}-\mathbf{W}||_2 = \frac{1}{K}\sum_{k=0}^{K-1}\sum_{l=0}^{L-1}||Q_{b_{l,\mathrm{w}}^{(k)}\in\mathcal{S}_k}(\mathbf{W}_l;\gamma)-\mathbf{W}_l||_2,$$

To efficiently search from the weight-sharing model, we develop:

- <u>Bidirectional Greedy Search</u>: We directly conduct inference only on the well-trained one-shot model, at a time, only one layer could either increase or decrease its bit-width until it fits the expected constraints (e.g., BitOPs).

Results: ImageNet Classification Task

Method	Bit-width (W/A)	Top-1 Acc. (%) ↑	BitOPs (G) ↓	Retrain Cost (Epoch) ↓
Baseline	32 / 32	70.5	-	-
PACT [7]	3 / 3	68.1	23.0	-
LSQ* [12]	3 / 3	69.4	23.0	90
EWGS [28]	3 / 3	69.7	23.0	100
EdMIPS [12]	3_{MP} / 3_{MP}	68.2	-	40
GMPQ* [60]	3_{MP} / 3_{MP}	68.6	22.8	40
DNAS [61]	3_{MP} / 3_{MP}	68.7	24.3	120
FracBits [64]	3_{MP} / 3_{MP}	69.4	22.9	150
LIMPQ [47]	3_{MP} / 3_{MP}	69.7	23.0	90
SEAM [49]	3_{MP} / 3_{MP}	70.0	23.0	90
Ours	2_{MP} / 3_{MP}	67.7	17.3	**0**
Ours	3_{MP} / 3_{MP}	**70.2**	23.3	**0**

ResNet18

Method	Bit-width (W/A)	Top-1 Acc. (%) ↑	BitOPs (G) ↓	Retrain Cost (Epoch) ↓
Baseline	32 / 32	75.4	-	-
LSQ [12]	3 / 3	69.7	4.2	90
NIPQ [44]	3_{MP} / 3_{MP}	66.5	-	43
QBitOPT† [37]	3_{MP} / 3_{MP}	<u>70.0</u>	-	<u>30</u>
MPQDNN [53]	3_{MP} / 3_{MP}	68.8	-	50
Ours	3_{MP} / 3_{MP}	**70.4**	4.5	**0**
LSQ [12]	4 / 4	72.3	6.8	90
NIPQ [44]	4_{MP} / 4_{MP}	72.3	-	43
QBitOPT† [37]	4_{MP} / 4_{MP}	**73.3**	-	<u>30</u>
Ours	4_{MP} / 4_{MP}	<u>73.2</u>	6.9	**0**

EfficientNet

- SOTA Mixed-Precision Quantization performance: our discovered quantization models consistently outperform previous works under various BitOPs constraints, on both highweight models (e.g., ResNet) and lightweight (e.g., EfficientNet, MobileNet) models.
- Retraining Elimination: our method leads on ResNet with a top accuracy of 71.0% at only 31.6G BitOPs and no retraining cost, compared to the second-best at 70.8% accuracy with higher 33.7G BitOPs and 90 epochs of retraining.

Results: Transfer Learning

Model	Bit-width (W/A)	CIFAR100 [27] Top-1 Acc. (%)	Pets [36] Top-1 Acc. (%)
ResNet18	32 / 32	79.4	88.9
	4_{MP} / 4_{MP}	79.5 (**+0.1%**)	88.7 (-0.2%)
	3_{MP} / 3_{MP}	78.7 (-0.7%)	87.9 (-2.0%)
MobileNetV2	32 / 32	78.9	86.0
	4_{MP} / 4_{MP}	79.0 (**+0.1%**)	86.1 (**+0.1%**)
	3_{MP} / 3_{MP}	78.2 (-1.7%)	84.1 (-1.9%)

- We transfer the quantized weights for downstream benchmarks to verify the generalization ability of the proposed method. Our method achieves the same accuracy as a full-precision model at 4-bits with smaller model complexity.

Ablation Study: Proposed Training Techniques

Table 6. Effectiveness of proposed dynamic bit-width schedule scheme and information distortion mitigation (IDM) training technique. To save costs, we train the weight-sharing model 80 epochs.

Dynamic Bit Schedule	IDM Training	4 Bit Top-1 Acc. (%)
✗	✗	68.3
✓	✗	69.1 (**+0.8%**)
✓	✓	69.5 (**+1.2%**)

- Dynamic bit-width scheduling and IDM training contribute positively to the model's performance, and their combination yields the most significant improvement.

Ablation Study: Information Distortion Mitigation (IDM)

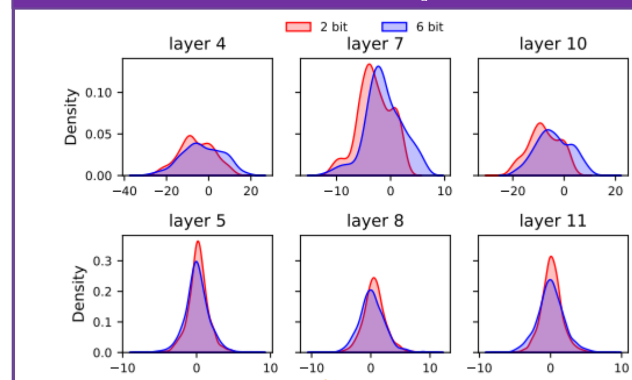

Fig. 3

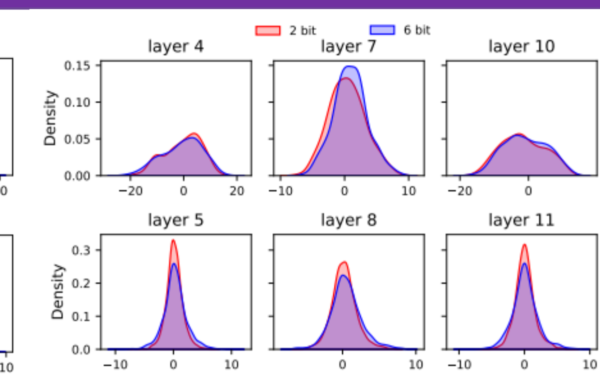

Fig. 4

- Our IDM training technique (Fig. 4) significantly mitigates information distortion problem (Fig. 3).