Prune-then-Quantize or Quantize-then-Prune? Understanding the Impact of Compression Order in Joint Model Compression

Minjun Kim, Jaehyeon Choi, Hyunwoo Yang, Jongjin Kim, Jinho Song, and U Kang*

Seoul National University, Seoul, South Korea, {minjun.kim, ukang}@snu.ac.kr, *: Corresponding Author

ICLR

Paper

GitHub

Data Mining Laboratory

Seoul National University

Summary

The Progressive Intensity Hypothesis

- **TL;DR**: Weaker perturbations should precede stronger ones for better performance in joint model compression
- **GitHub**: https://github.com/snudatalab/PQQP

Joint Model Compression

Joint Model Compression

- Combining two or more compression techniques
- Two categories: 1) co-designed or **2) post-hoc** methods
- Sequential application of $f_1(\cdot)$ and $f_2(\cdot)$: $f_1 \to f_2$ or $(f_2 \circ f_1)(\cdot)$
- Compression rate C: memory usage ratio of original ϕ and compressed ϕ' models ($C_{\mathcal{P}} = 1/(1-p), C_{\mathcal{Q}} = B_{orig}/B_{\mathcal{Q}}$)

Motivation. Compression Order in Joint Model Compression

- Identifying an optimal order can yield a **"free lunch"!**
- Error rates deviates up to **3.41×** (at compression rate $C = 4$)
- However, the role of compression order has been largely overlooked

Pruning Type	Order	FP32	INT8
Dense model (0% Pruning)		27.65	28.06
50% (Unstructured)	$\mathcal{P} \to \mathcal{Q}$	29.94	**30.22**
	$\mathcal{Q} \to \mathcal{P}$		34.71
2:4 (Semi-structured)	$\mathcal{P} \to \mathcal{Q}$	31.89	**32.76**
	$\mathcal{Q} \to \mathcal{P}$		45.06

WikiText-2 PPL (↓) of OPT-125M model

Harma et al., ICLR 2025

Error rate: 18.48%

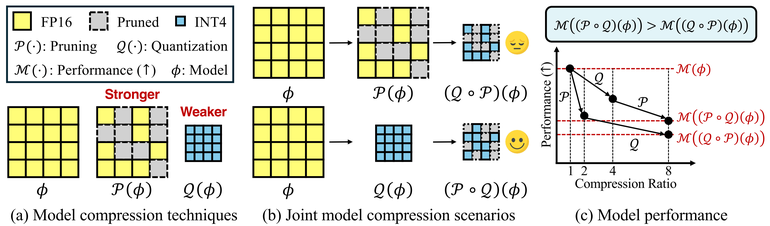
Error rate: 62.96%

Problem. Joint Compression Order Optimization

- **Input:** a set of compression methods $\mathbb{F} = \{f_1(\cdot), f_2(\cdot), \cdots, f_n(\cdot)\}$, a pre-trained model ϕ, and performance metric $\mathcal{M}(\cdot) \uparrow$
- **Output:** the optimal permutation $\pi^* = \text{argmax}_{\pi \in \Pi} \mathcal{M}(\pi(\phi))$ for set $\Pi = \{\pi : \mathbb{F} \to \mathbb{F} \mid \pi \text{ is bijective}\}$ of all permutations over $\mathbb{F}$

The Progressive Intensity Hypothesis

The Progressive Intensity Hypothesis. *Neural networks compressed by multiple methods perform better when **weaker** perturbations are applied first and **stronger** ones later.*



(a) Model compression techniques (b) Joint model compression scenarios (c) Model performance

Theoretical Analysis

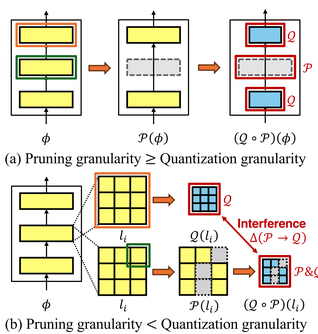
Intensity of Compression

- We assess intensity through model performance
- Performance gap $\mathcal{G}(f_1, f_2) \coloneqq \mathcal{M}(f_1(\phi)) - \mathcal{M}(f_2(\phi))$
- Compression Equivalence Ratio (CER) C_f^*: $B_{\mathcal{Q}}$-axis

$$\mathcal{M}(f(\phi; C)) = \mathcal{M}\left(\mathcal{Q}(\phi; C_f^*)\right)$$

- Compression order advantage $\mathcal{A}(f_1 \to f_2)$: order effect

$$\mathcal{A}(f_1 \to f_2) \coloneqq \mathcal{G}(f_1 \to f_2, f_1 \to f_2)$$



(a) Pruning granularity ≥ Quantization granularity

(b) Pruning granularity < Quantization granularity

Pruning and Quantization

- $\mathcal{A}(\mathcal{Q} \to \mathcal{P})$ grows monotonically with $\mathcal{G}(\mathcal{Q}, \mathcal{P})$ or $C_{\mathcal{P}}^* - C_{\mathcal{Q}}$

Case (a). Disjoint Selectivity

- All units are exclusively assigned to one $f(\cdot)$
- Layer-wise error analysis: showing that weak → strong ordering consistently yields lower layer-wise degradation (under fixed $C_{\mathcal{P}}$)

Case (b). Interference

- Additional error from mutual interaction
- The magnitude of interference is determined solely by pruning ratio p (preserving monotonic trend)

Experimental Findings

1. The hypothesis holds across diverse scenarios

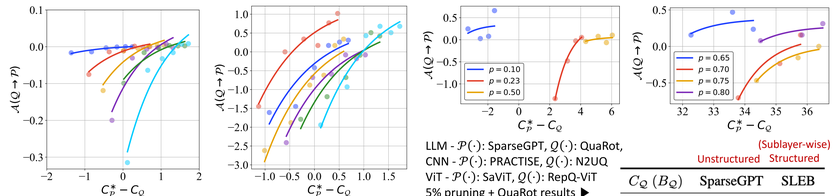
- Regardless of modality, model architecture, task, intensity, method, and technique

2. Pruning granularity defines interference

- Toward precise sign / value prediction of compression-order advantage $\mathcal{A}(f_1 \to f_2)$

* Pruning and Quantization: Language and Vision models

- LLaMA-3 8B (WikiText-2 PPL and ARC-C), ResNet-18 and DeiT-Base (ImageNet)



LLM - $\mathcal{P}(\cdot)$: SparseGPT, $\mathcal{Q}(\cdot)$: QuaRot,
CNN - $\mathcal{P}(\cdot)$: PRACTISE, $\mathcal{Q}(\cdot)$: N2UQ
ViT - $\mathcal{P}(\cdot)$: SaViT, $\mathcal{Q}(\cdot)$: RepQ-ViT
5% pruning + QuaRot results ▶

(Sublayer-wise)

$C_{\mathcal{Q}}$ ($B_{\mathcal{Q}}$)	SparseGPT (Unstructured)	SLEB (Structured)
1.78 (9)	0.002	0
2.00 (8)	0.001	0
2.28 (7)	-0.003	0
2.68 (6)	-0.013	0
3.20 (5)	-0.017	-0.057
4.00 (4)	-49.899	-9.379

* Pruning granularity and Interference

- Unstructured (SparseGPT): monotonic trend
- Structured (SLEB; sublayer): no interference at early regimes

* Toward general pipelines

- **Multi-stage compression.** $\mathcal{P}_1(\cdot) \to \mathcal{Q}(\cdot) \to \mathcal{P}_2(\cdot)$
- **PEFT.** $\mathcal{P}(\cdot)$ and ($\mathcal{Q}(\cdot)$ + LoRA)
- **Parameter sharing.** Pruning $\mathcal{P}(\cdot)$ + Basis sharing $\mathcal{S}(\cdot)$
- **Mixed-precision Quantization.** prog. (8 → 2 bits) vs. regr. (2 → 8bits)