NEURAL INFORMATION PROCESSING SYSTEMS

Differentiable Bit-Widths: Co-optimizing Pruning and Quantization via SVD for Ultra-Efficient LLM Compression





Hankyul Kang · Jongbin Ryu
Ajou University

Motivation

SVD-based pruning and quantization have recently emerged as a promising strategy for the ultra-efficient compression of LLMs (Yuan et al., 2023; Wang et al., 2025).

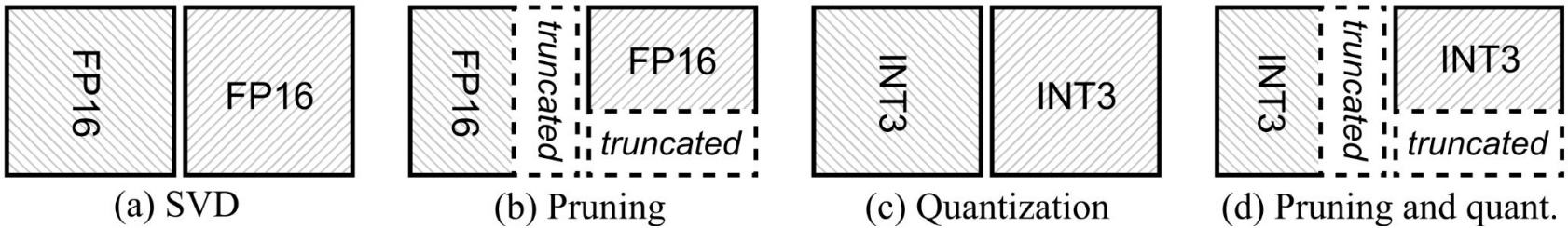
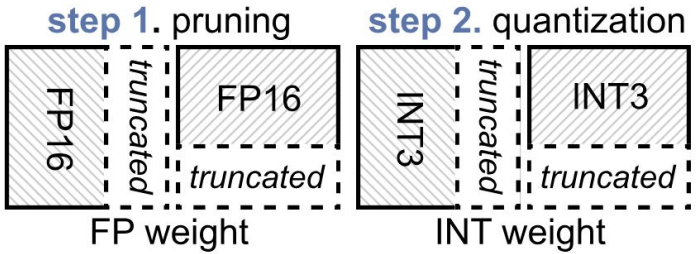


Figure 1: Motivation illustration.

Concept

Problem: existing approaches apply pruning and quantization in (a) **Sequential** manner, which make it difficult to find a balance between them.



(a) **Sequential** optimization

Figure 2: Our compression concept illustration.

Concept

Problem: existing approaches apply pruning and quantization in (a) **Sequential** manner, which make it difficult to find a balance between them.

Solution: (b) **Co**-optimization method that learn component-wise bit-widths, allowing less important ones to be assigned 0-bit precision and pruned away.

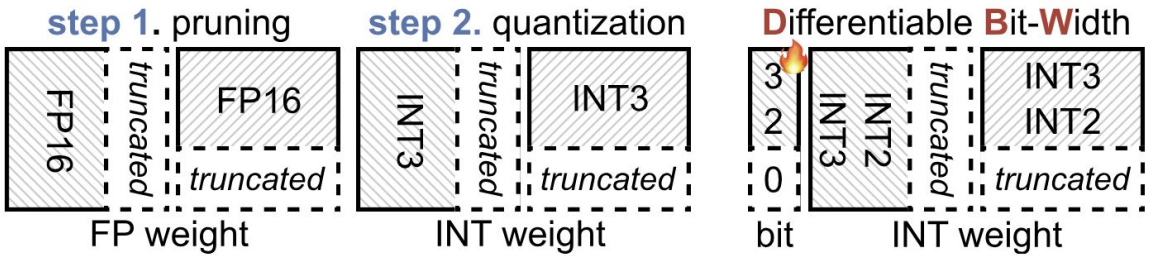
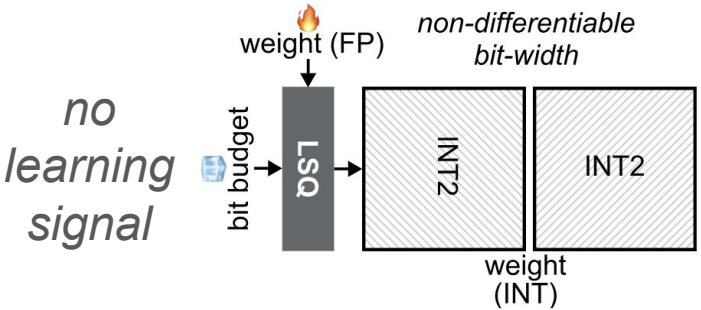


Figure 2: Our compression concept illustration.

Method

- The challenge is that bit-widths are non-differentiable during backpropagation (Esser et al., 2020).

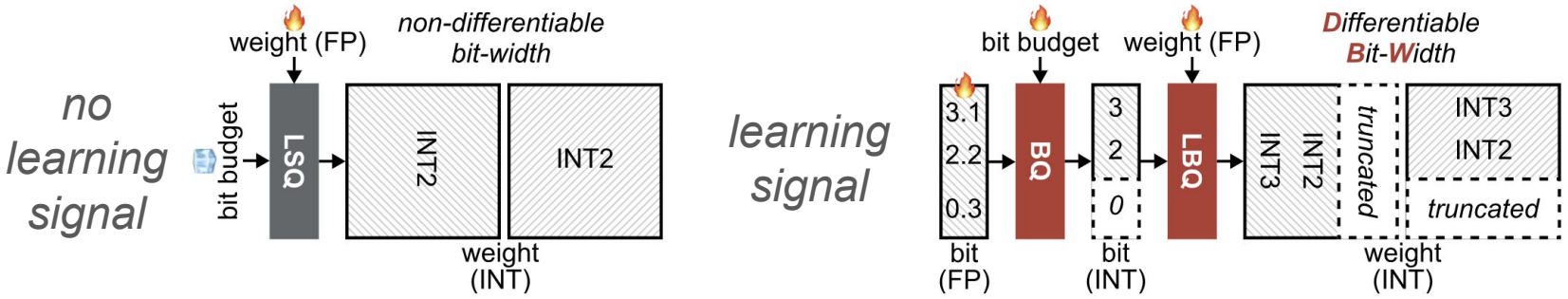


(a) Non-differentiable bit

Figure 3: Our differentiable method illustration.

Method

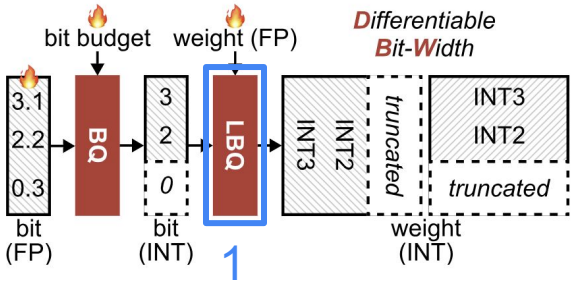
- The challenge is that bit-widths are non-differentiable during backpropagation.
- To overcome the challenge, we present a ***D***ifferentiable ***B***it ***W***idth method.



(a) Non-differentiable bit (b) Our differentiable bit

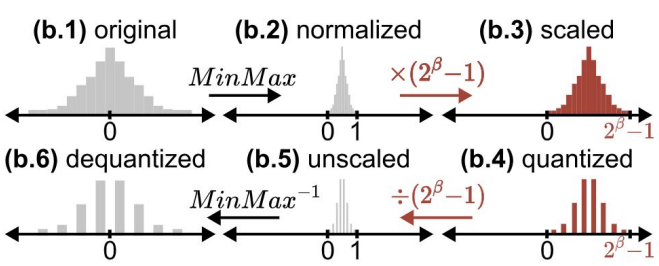
Figure 3: Our differentiable method illustration.

Method Details

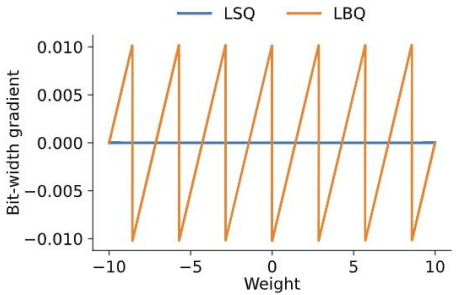


DBW is composed of two components:

- Learned bit-width quantizer (LBQ) enables bit-width parameter differentiable.



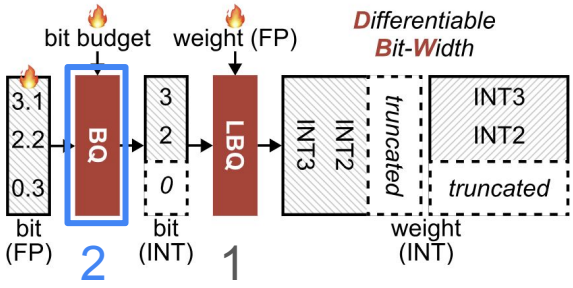
(a) LBQ illustration



(b) Bit-width gradient analysis

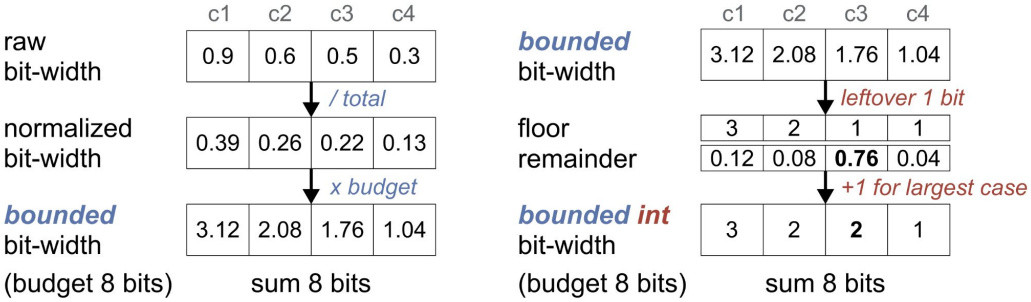
Figure 4: Our LBQ illustration.

Method Details



DBW is composed of two components:

- ❏ Learned bit-width quantizer (LBQ) enables bit-width parameter differentiable.
- ❏ Bit-width quantizer (BQ) allows a sum of bit-widths remains within a bit budget.



(a) Normalization (b) Hamilton method

Figure 5: Our BQ illustration.

Empirical Findings

- DBW performs well against heuristic approach in PPL (↓) and Acc (↑) .
- DBW adapts the bit allocation to a singular-value spectrum of each matrix.

LLaMA	Method	PPL↓	Acc.↑
1-7B	Uniform	690.8	25.06
	Heuristic	11.71 (-679.09)	45.54 (+20.48)
	Learnable	10.65 (-680.15)	48.98 (+23.92)
2-7B	Uniform	650.81	25.05
	Heuristic	11.69 (-639.12)	45.10 (+20.05)
	Learnable	10.54 (-640.27)	49.06 (+24.01)

Table 1: Controlled factor analysis of SVD-space bit allocation.

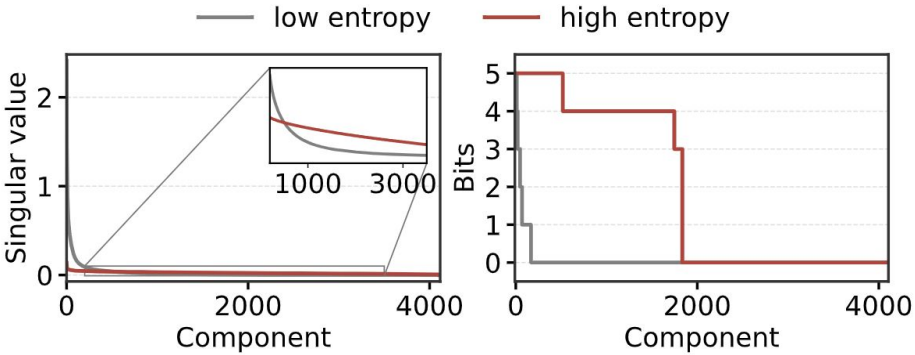


Figure 6: Learned bit-width analysis.

Comprehensive Benchmark

Co-optimization using DBW works well against the sequential baseline.

Type	Method (Bit)	LLaMA-7B				LLaMA-2-7B				LLaMA-3-8B			
		Ratio	*Mem*	*PPL*↓	*Acc.*↑	*Ratio*	*Mem*	*PPL*↓	*Acc.*↑	*Ratio*	*Mem*	*PPL*↓	*Acc.*↑
-	Baseline (FP16)	0.00	12.31	6.38	63.06	0.00	12.31	6.22	63.19	0.00	13.98	7.51	65.98
sequential optimization	ASVD (FP16)	46.33	6.72	3.0e4	25.02	46.33	6.72	4.5e4	23.94	46.33	7.96	2.6e4	24.43
	ASVD+GPTQ (1.61)	46.33	1.56	3.0e4	25.00	46.33	1.56	4.5e4	24.63	46.33	2.40	2.6e4	24.26
	ASVD+OmniQ (1.61)	46.33	1.56	3.1e4	25.02	46.33	1.56	8.9e4	24.89	46.33	2.40	5.3e4	24.97
	SVD-LLM (FP16)	46.33	6.72	14.08	47.21	46.33	6.72	59.28	31.59	46.33	7.96	53.76	36.55
	SVD-LLM+GPTQ (1.61)	46.33	1.56	22.45	37.17	46.33	1.56	116.38	27.80	46.33	2.40	102.31	30.26
	SVD-LLM+OmniQ (1.61)	46.33	1.56	17.22	42.84	46.33	1.56	77.10	29.35	46.33	2.40	89.41	30.52
co-optimization (our DBW) (1.61)		47.21	1.56	10.65	48.98	46.95	1.56	10.54	49.06	44.71	2.40	15.96	44.38

Table 2: Sequential vs. co-optimization of pruning and quantization.

Conclusion

- We present a differentiable method for learning component-wise bit-widths, allowing less important ones to be assigned 0-bit precision and pruned away.
- We find out that the differentiable method adapts to a singular-value spectrum of each matrix.
- We demonstrate that co-optimization using the differentiable method works well against the two-stage baselines.

Reference

Steven K. Esser et al., “Learned Step Size Quantization”, ICLR 2020.

Xin Wang et al., “SVD-LLM: Truncation-aware Singular Value Decomposition for Large Language Model Compression,” ICLR 2025.

Zhihang Yuan et al., “ASVD: Activation-aware Singular Value Decomposition for Compressing Large Language Models,” 2023.