NEURAL INFORMATION PROCESSING SYSTEMS

Samsung Research

LittleBit: Ultra Low-Bit Quantization via Latent Factorization

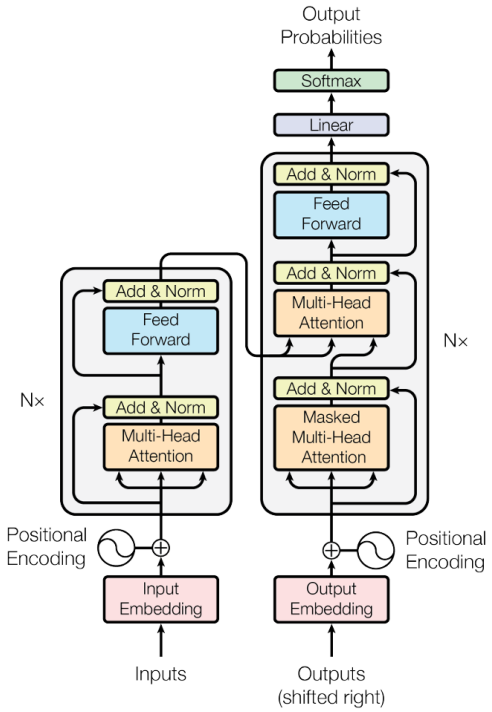
Banseok Lee*, Dongkyu Kim*, Youngcheon You, Youngmin Kim†

Samsung Research

*: Equal Contribution, †: Corresponding author

{bs93.lee, dongkyu.k, y01000.you, ym1012.kim}@Samsung.com

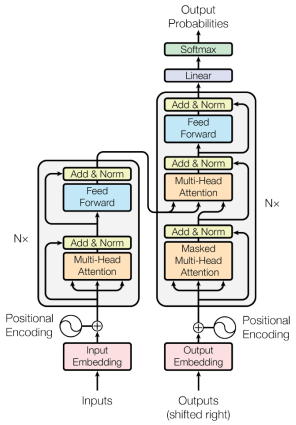
Motivation: LLMs on Mobile Devices?



70B LLMs [138GB] VRAM

Mobile [12GB] VRAM

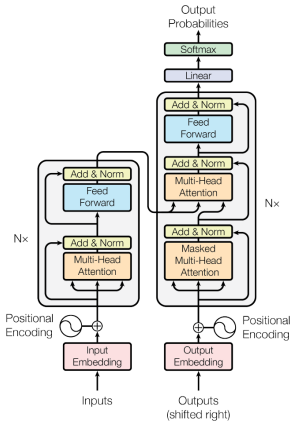
Motivation: LLMs on Mobile Devices?



1bit?

Mobile [12GB] VRAM

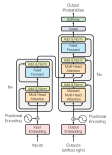
Motivation: LLMs on Mobile Devices?



70B 1Bit [14GB] VRAM

Mobile [12GB] VRAM

Motivation: LLMs on Mobile Devices?



70B 0.1Bit [2GB] VRAM

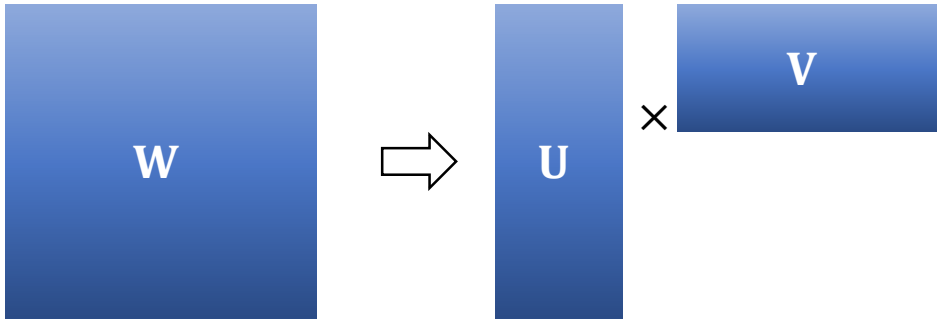
Mobile [12GB] VRAM

Our Approach

1. **Compress**
   (Latent Factorization)

2. Recover
(Quantization-Aware Training (QAT))

Factorization & Initialization (Dual-SVID)

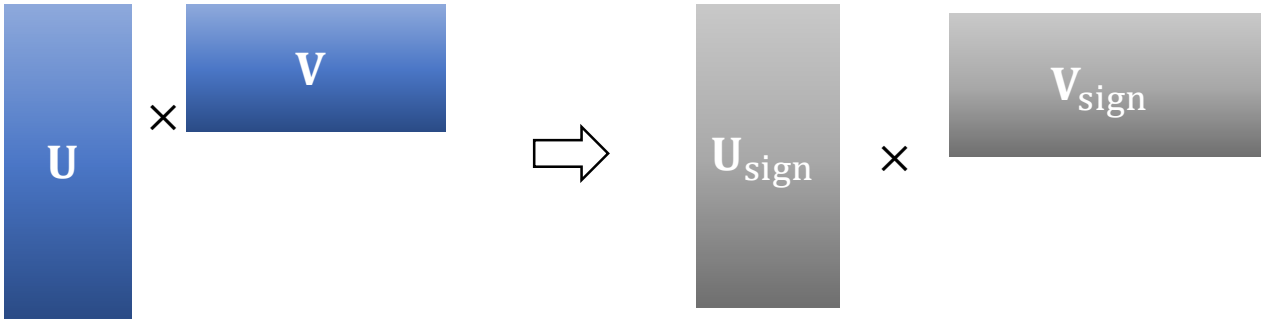


1. Low-rank Factorization (SVD)

FP16

Binary(1bit)

Factorization & Initialization (Dual-SVID)

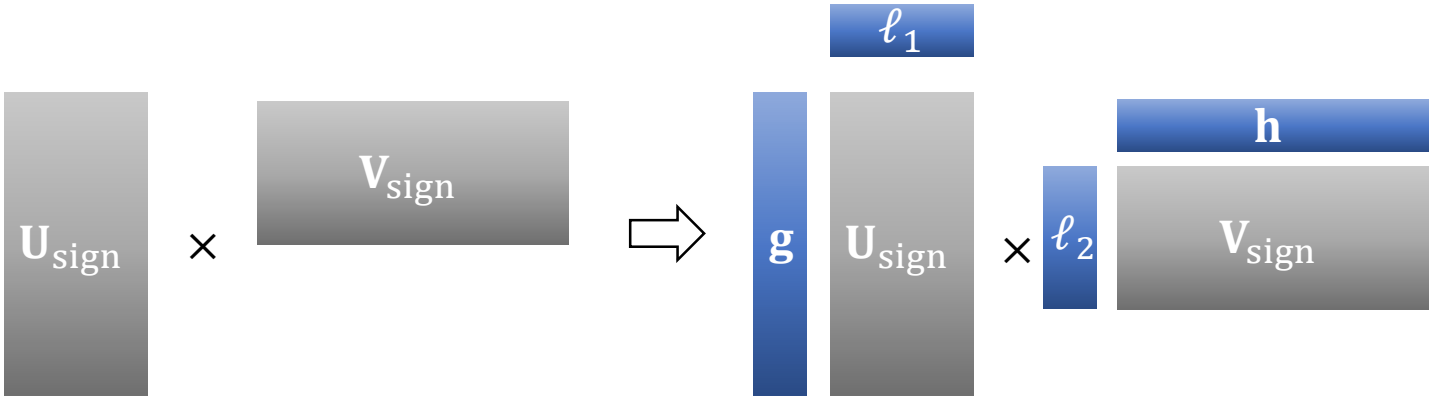


2. Binarization

FP16

Binary(1bit)

Factorization & Initialization (Dual-SVID)

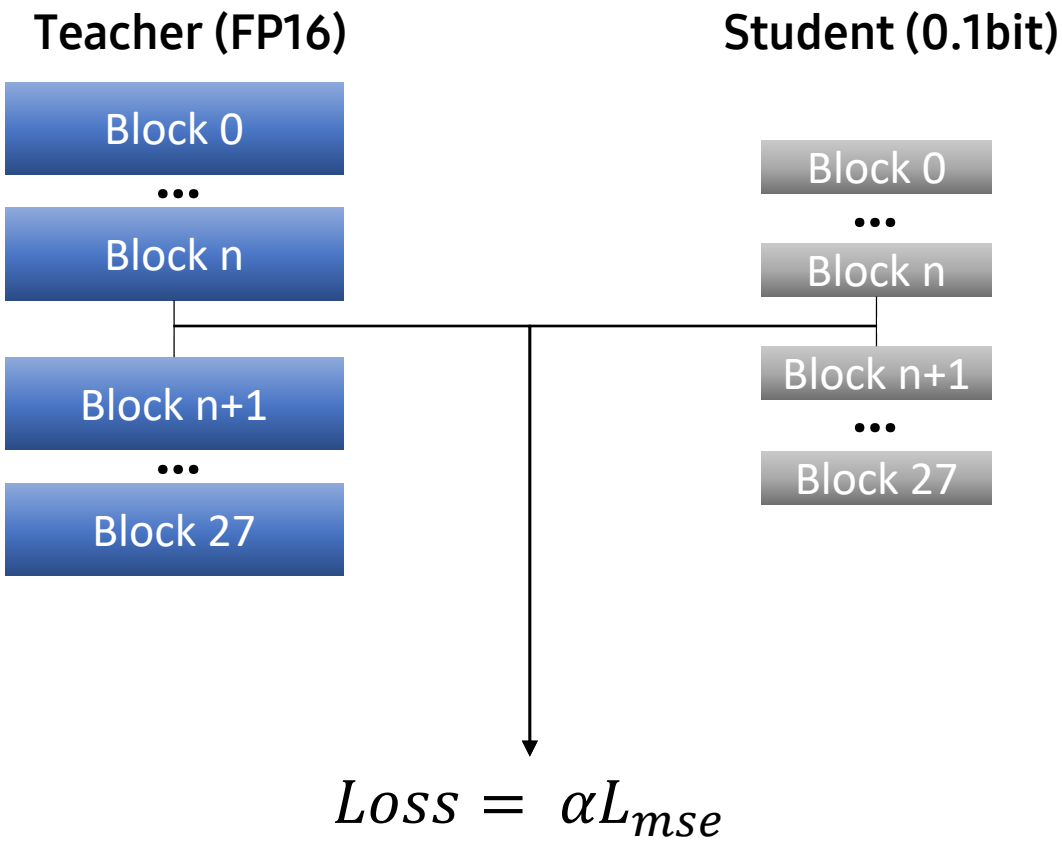


3. Multi-scale Compensation

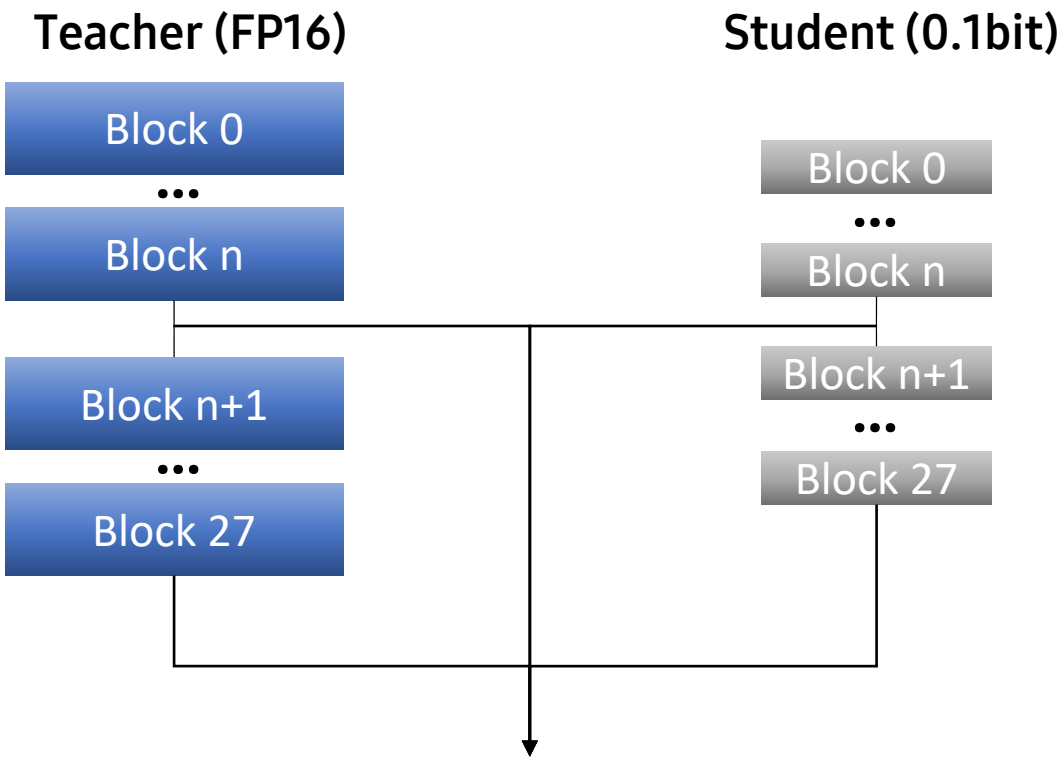
FP16

Binary(1bit)

Recovery via QAT (Knowledge Distillation)

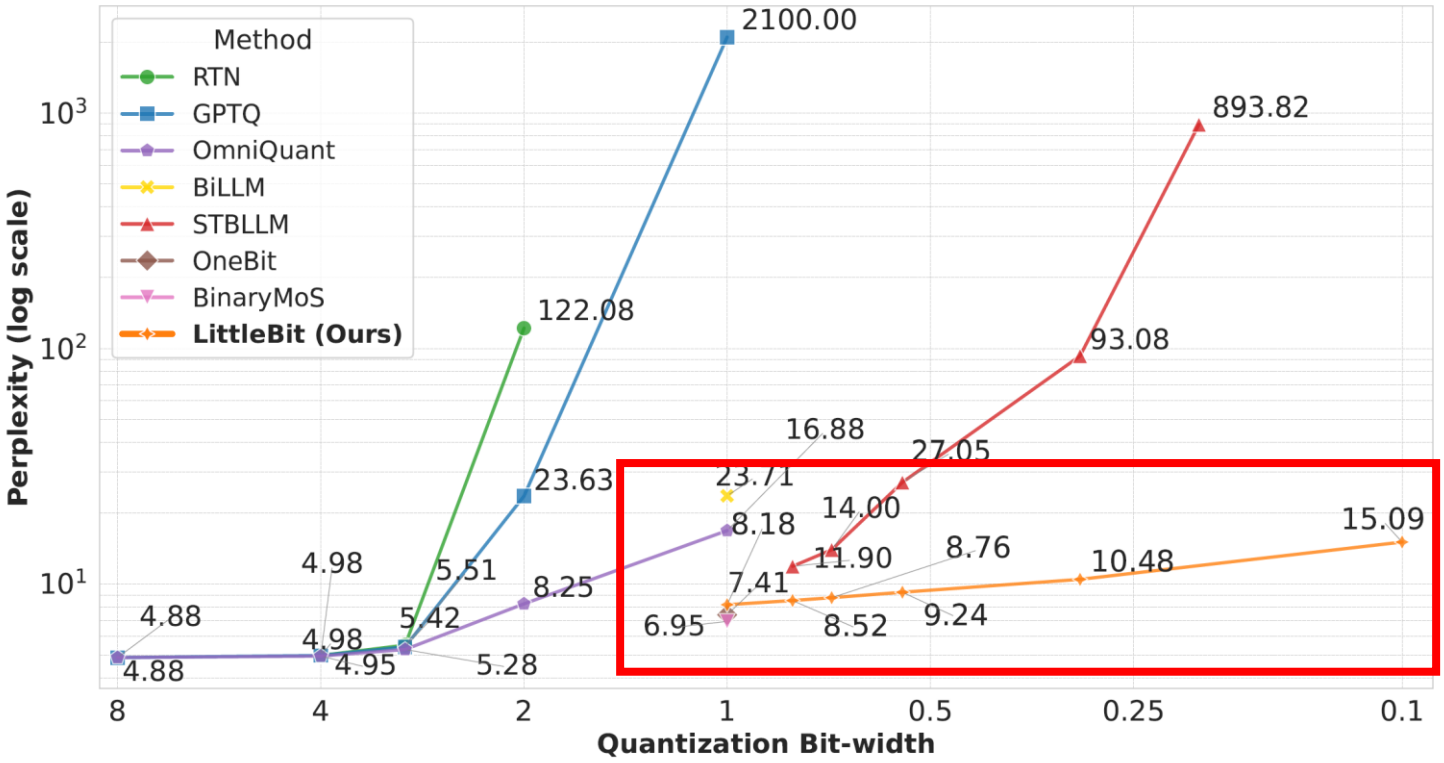


Recovery via QAT (Knowledge Distillation)



$$Loss = L_{kl} + \alpha L_{mse}$$

Result1: SOTA Perplexity

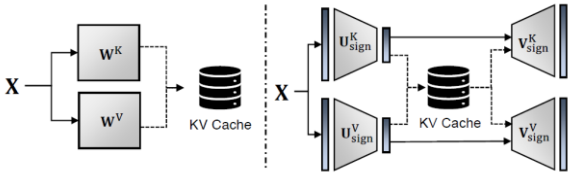


Result2: Memory Reduction

GPU Memory ↓

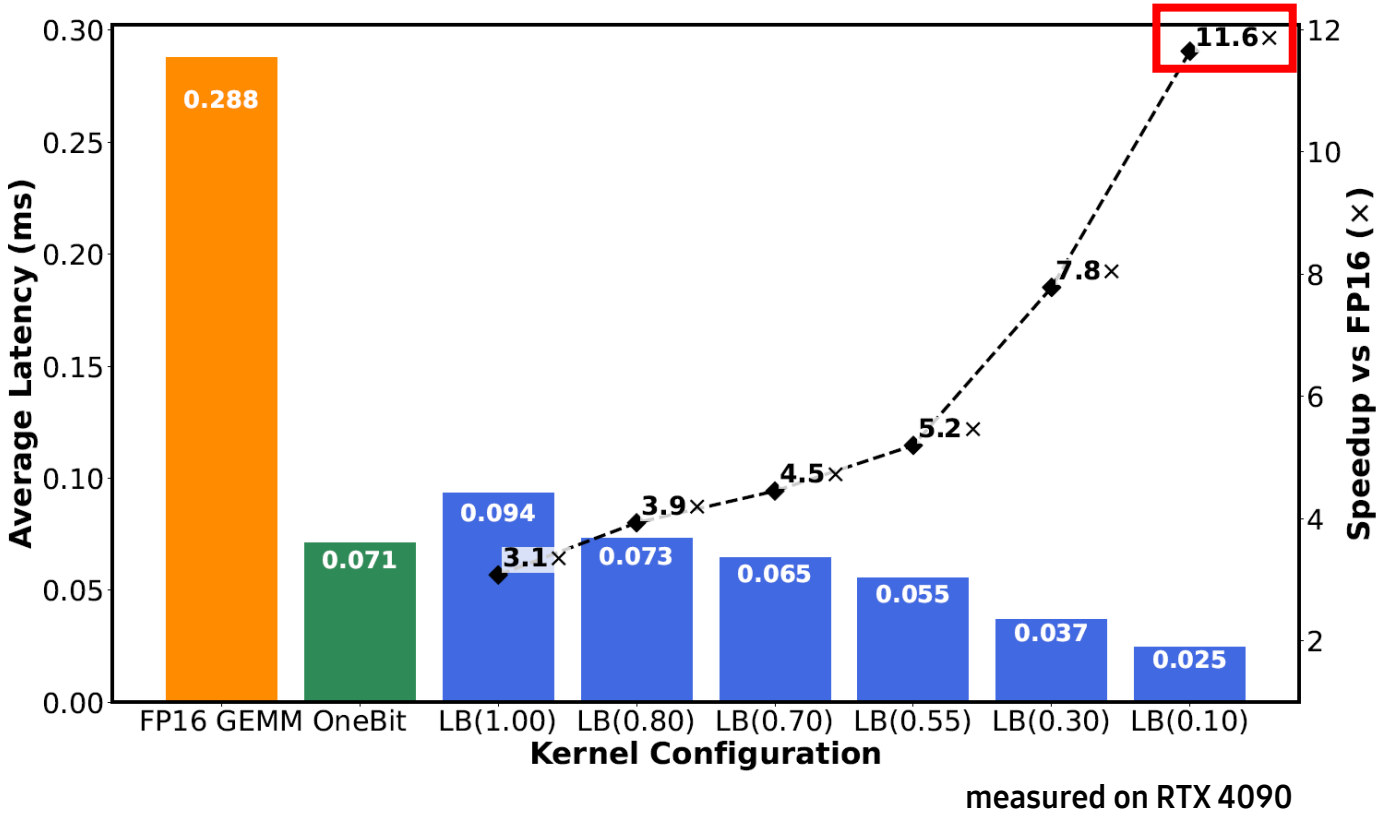
Model	FP16	OneBit [13]	**LittleBit (Ours)**	
BPW	16	1	0.3	0.1
Llama2-7B	13.49 GB	1.36 GB (9.92×)	0.79 GB (17.08×)	0.63 GB (21.41×)
Llama2-13B	26.06 GB	2.28 GB (11.43×)	1.15 GB (22.66×)	0.84 GB (31.02×)
Llama2-70B	138.04 GB	9.72 GB (14.20×)	3.70 GB (37.31×)	1.98 GB (69.72×)

KV Cache ↓



BPW	KV Latent Rank (r)	KV Cache Reduction Factor
0.80	1,624	∼2.5×
0.55	1,112	∼3.7×
0.30	600	∼6.8×
0.10	192	∼21.3×

Result3: Up to 11.6x Kernel Speed-Up



Conclusion

LittleBit: SOTA Sub-1-bit Quantization