ICLR 2025

[Spotlight]

The Interplay between Quantization and Sparsity

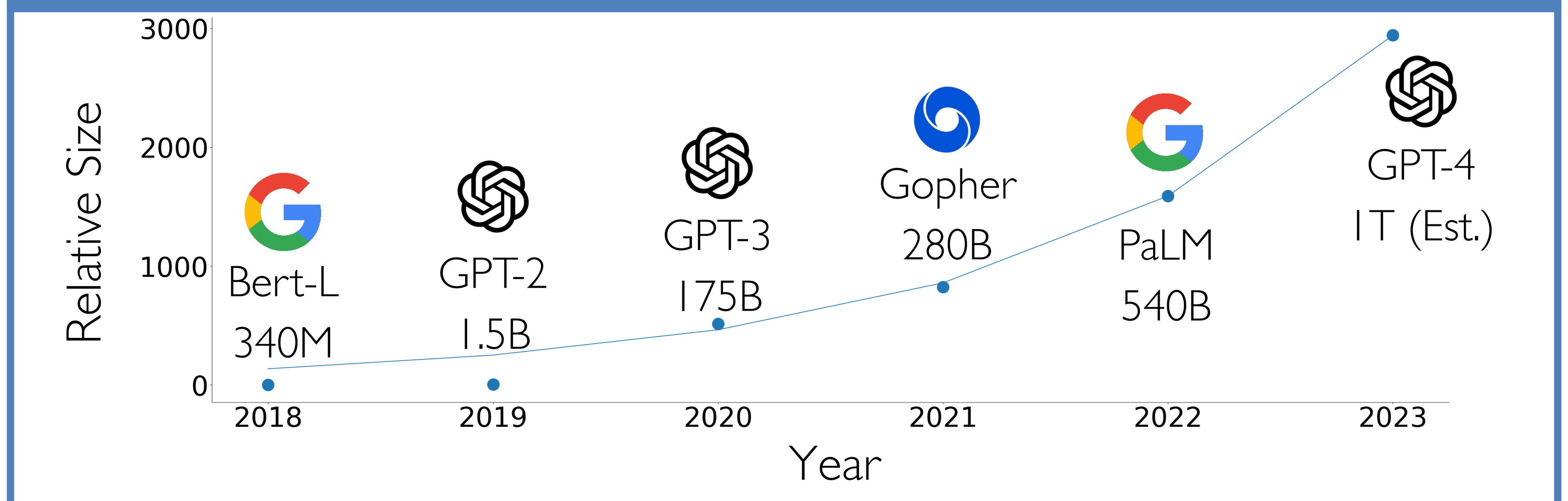
PARALLEL SYSTEMS ARCHITECTURE LAB

Simla Burcu Harma †, Ayan Chakraborty †, Elizaveta Kostenok †, Danila Mishin †, Dongho Ha ⁺,

Babak Falsafi †, Martin Jaggi †, Ming Liu ★, Yunho Oh ⁺, Suvinay Subramanian ★, Amir Yazdanbakhsh ‡

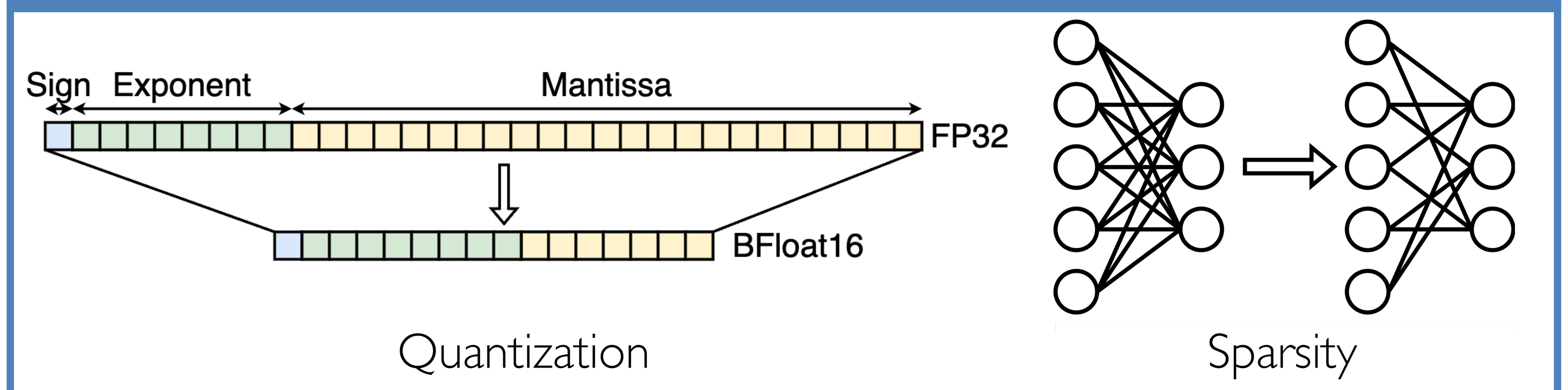
† EcoCloud, EPFL ⁺ Yonsei University ★ Google ‡ Google DeepMind

DNN model sizes are exploding!



- Memory footprint becomes a severe bottleneck during inference

Model compression



- Their combination can provide a huge reduction in memory footprint
- However, what is their combined impact on model accuracy?

Research question and contributions

- When sparsity and quantization are combined, are there additional errors introduced beyond those of each method individually?
- To answer this question, we conduct mathematical analysis of their combination

- We mathematically define two tensor transformations f and g to be *orthogonal* if no additional error is introduced upon their combination:

$\|\varepsilon_{f \circ g}(x)\| \le \|\varepsilon_f(x)\| + \|\varepsilon_g(x)\|$ and $\|\varepsilon_{g \circ f}(x)\| \le \|\varepsilon_f(x)\| + \|\varepsilon_g(x)\|$ for any input tensor x, where $\varepsilon_f(x) \coloneqq x - f(x)$

- We mathematically demonstrate the non-orthogonality of sparsity and quantization at the (a) tensor level, and (b) dot-product level
- We empirically validate our mathematical findings and demonstrate end-to-end non-orthogonality across a diverse range of SOTA models

Tensor-level analysis

- Our mathematical analysis centers on:
  - Block-wise quantization
  - Magnitude-based sparsity

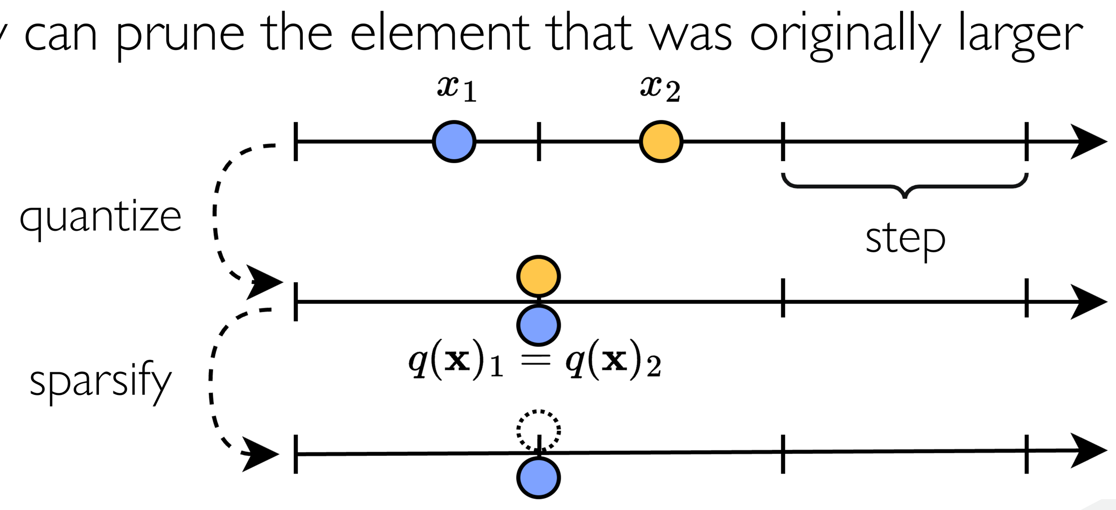
- If sparsity is applied before quantization, no additional error occurs

$$\|\varepsilon_{q \circ s}(x)\| \le \|\varepsilon_q(x)\| + \|\varepsilon_s(x)\|$$

- However, applying quantization before sparsity yields additional error

$$\|\varepsilon_{s \circ q}(x)\|_1 \le \|\varepsilon_q(x)\|_1 + \|\varepsilon_s(x)\|_1 + \underbrace{2 \cdot \mathrm{step} \cdot \frac{\mathrm{M}-\mathrm{N}}{\mathrm{M}} \cdot \mathrm{n}}_{\text{additional error}}$$

- Reason of the additional error:
  - Quantization can equalize elements
  - Sparsity can prune the element that was originally larger



- Therefore, applying sparsity before quantization is optimal

Dot-product-level analysis

- Our dot-product analysis focuses on the following set-up:
  - Weights are both sparsified and quantized
  - Activations are only quantized

- Sparsity and quantization combined yield additional error in both orders
- Therefore, quantization and sparsity are non-orthogonal

- Moreover, the additional error has an upper bound:

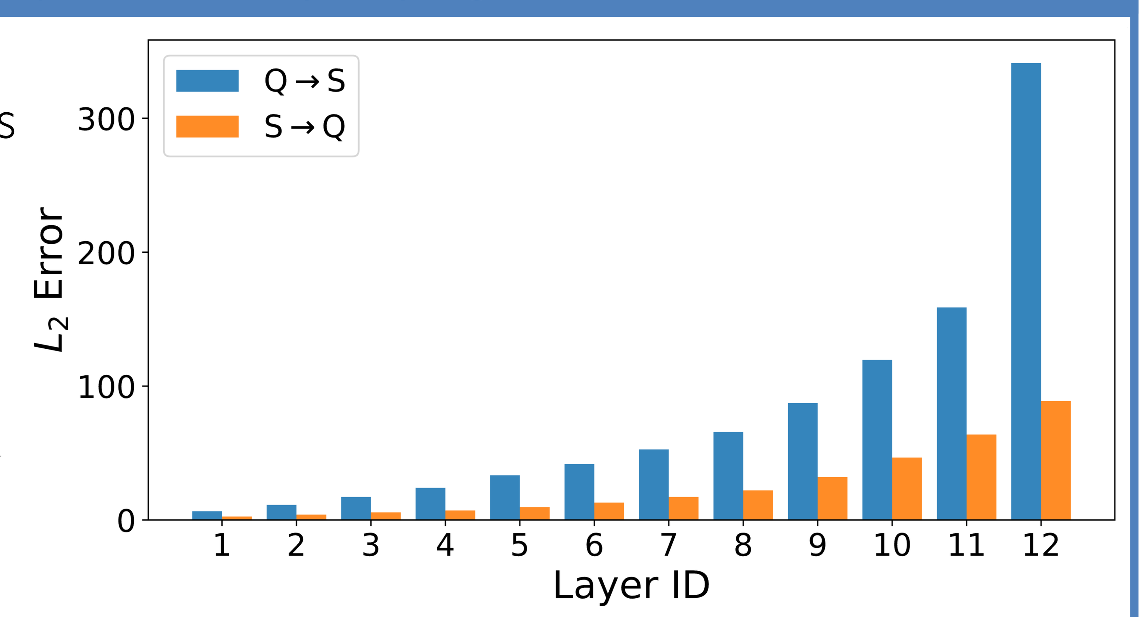
$$\|\varepsilon^{\mathrm{D}}_{\mathrm{q,c}}(\mathrm{x},\mathrm{w})\| \le \|\varepsilon^{\mathrm{D}}_{\mathrm{I,s}}(\mathrm{x},\mathrm{w})\| + \|\varepsilon^{\mathrm{D}}_{\mathrm{q}}(\mathrm{x},\mathrm{w})\| + \underbrace{\|\langle \mathrm{q}(\mathrm{x}), \tilde{\varepsilon}_{\mathrm{c}}(\mathrm{w})\rangle\| + \|\langle \varepsilon_{\mathrm{q}}(\mathrm{x}), \varepsilon_{\mathrm{s}}(\mathrm{w})\rangle\|}_{\text{additional error}}$$

- The upper bound is significantly lower for $S \to Q$ order than $Q \to S$

Per-layer error propagation

- Error accumulates across layers regardless of the order

- However, $S \to Q$ order consistently yields lower error than $Q \to S$



Optimal order of sparsity and quantization

- Sparsity followed by quantization is the optimal order
- The sub-optimal order can cause up to 7.96 point increase in perplexity

Sparsity type	Order	LLaMA-2-7B					
		FP32	INT8	MXFP8	MXFP6	HBFP8	HBFP6
dense	-	5.12	5.15	5.17	5.16	5.12	5.24
50%	$S \to Q$	6.31	6.94	6.4	6.38	6.32	6.51
	$Q \to S$	-	8.13	8.47	9.32	9.86	10.2
2:4	$S \to Q$	9.3	9.37	9.35	9.32	9.39	10.68
	$Q \to S$	-	14.65	14.35	14.5	14.98	18.64

Non-orthogonality of sparsity and quantization

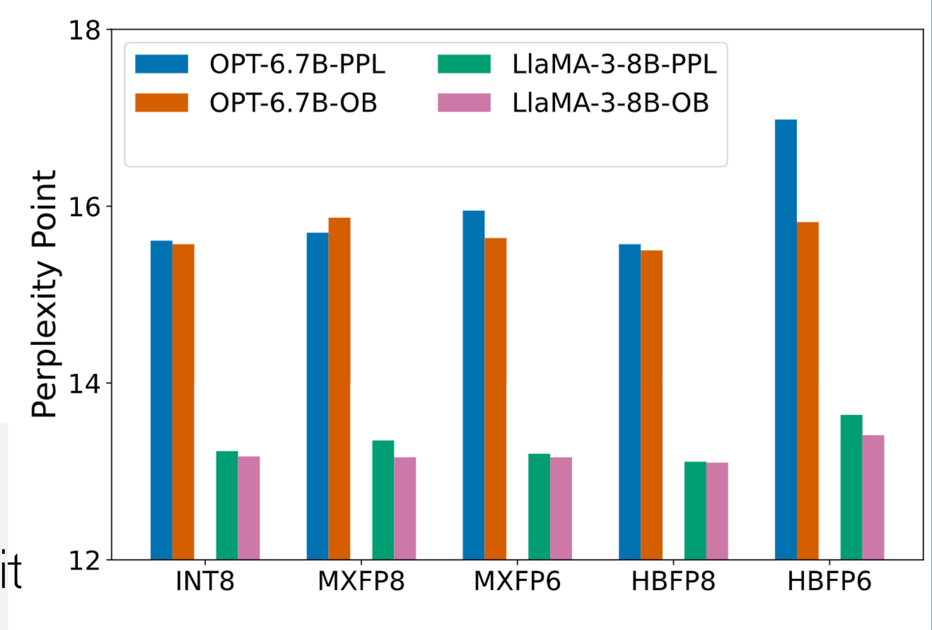
- Orthogonality bound (OB) is

$$OB = PPL + Err_q + Err_s$$

- We use the optimal order

- PPL exceeds OB in most cases

- Difference is prominent for <8 bit



For more details and experimental results, please check out our paper!