SparseLoRA: Accelerating LLM Fine-Tuning with Contextual Sparsity

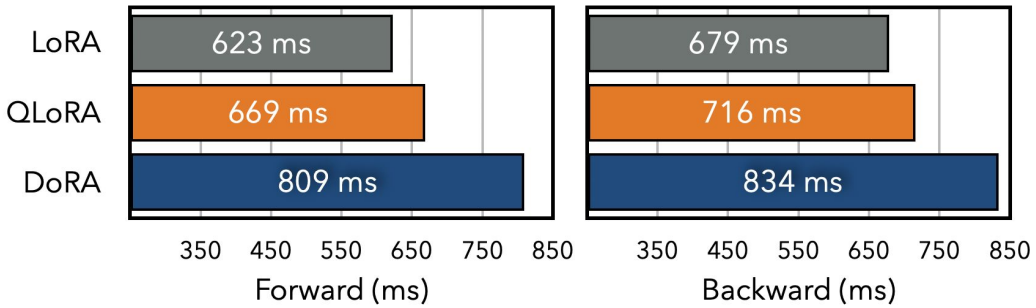
ICML 2025

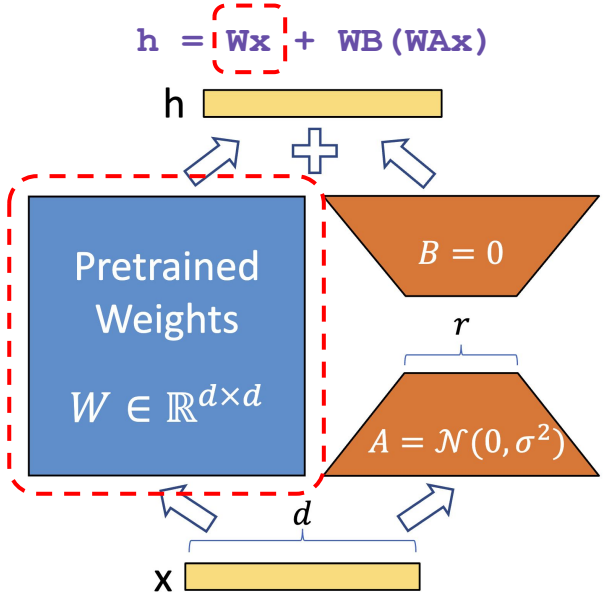
Samir Khaki*[1], Xiuyu Li*[2], Junxian Guo*[3], Ligeng Zhu[3], Konstantinos N. Plataniotis[1], Amir Yazdanbakhsh[4], Kurt Keutzer[2], Song Han[3], Zhijian Liu[3]

[1]University of Toronto [2]UC Berkeley [3]MIT [4]Google DeepMind

LoRA Fine-tuning is not Fast

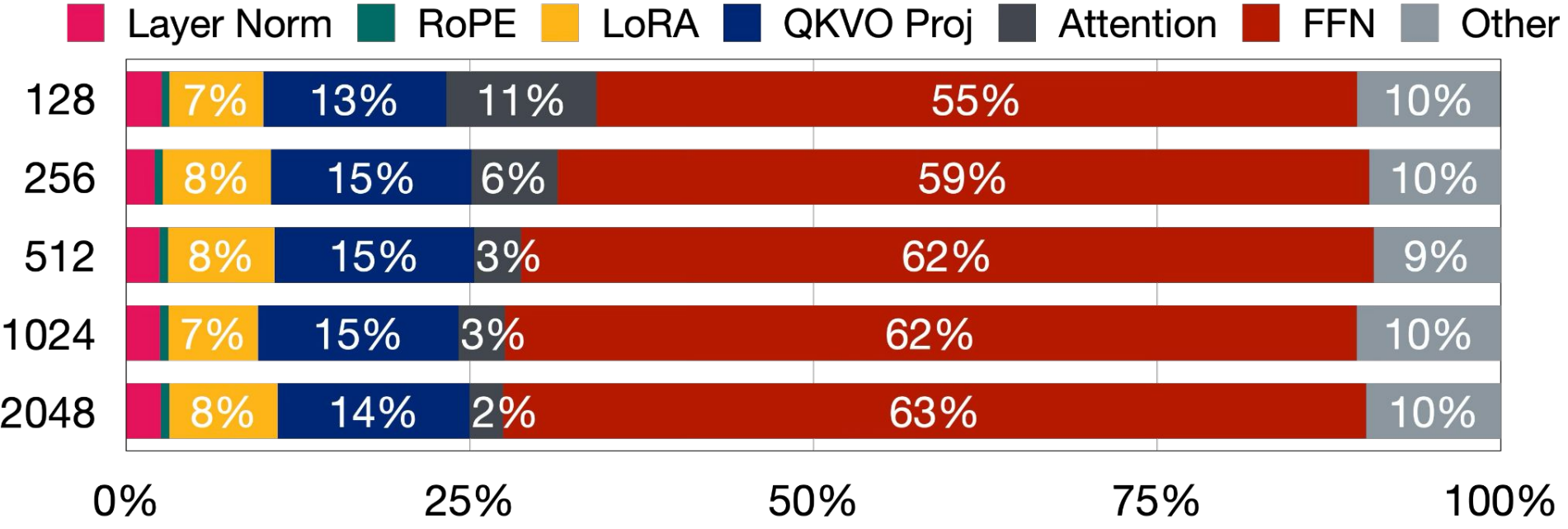
- **Existing methods mainly aim to save memory**
  - But have worse compute efficiency…





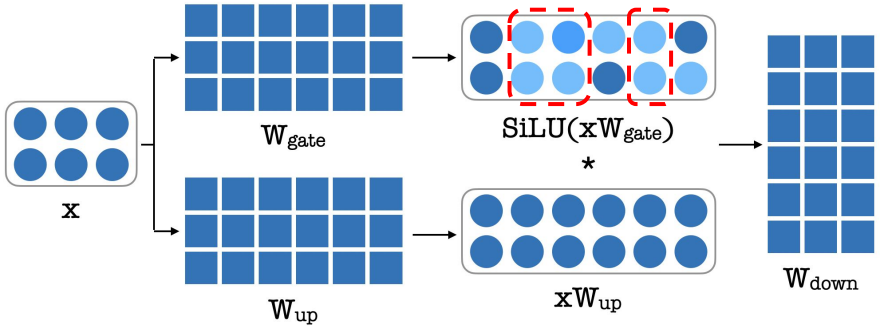
What is the bottleneck?

- **Interestingly the Frozen Linear Layers bottleneck fine-tuning latency**



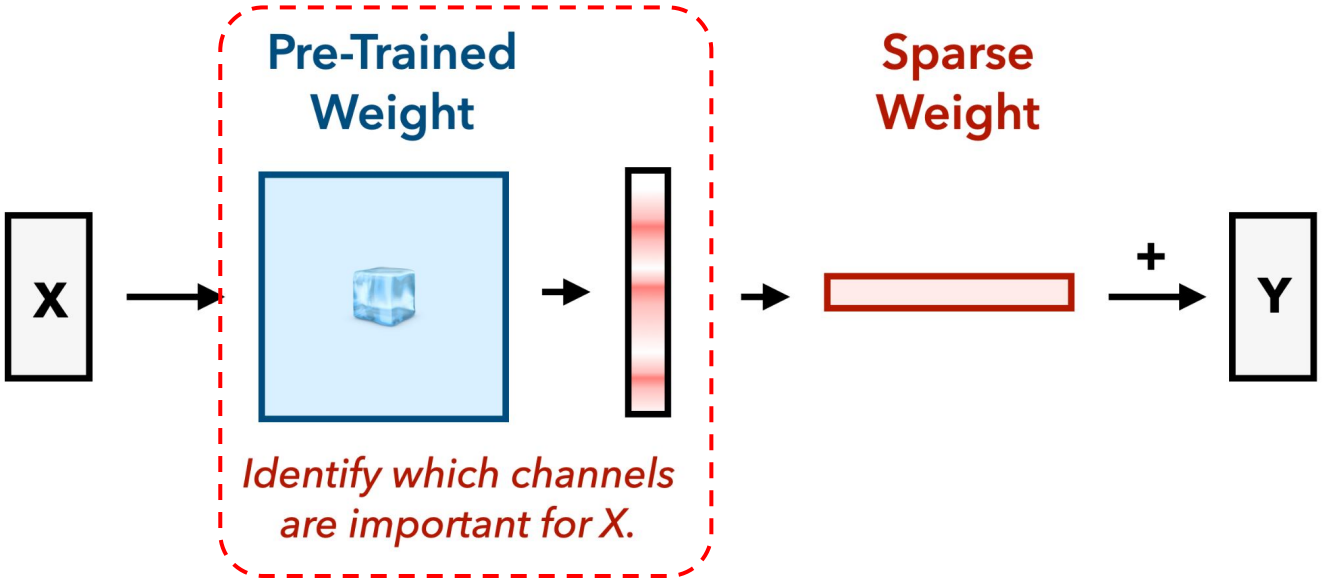
Leveraging Contextual Sparsity in Fine-tuning

Identify existing sparsity in activations using L2 Norm for the given input during fine-tuning



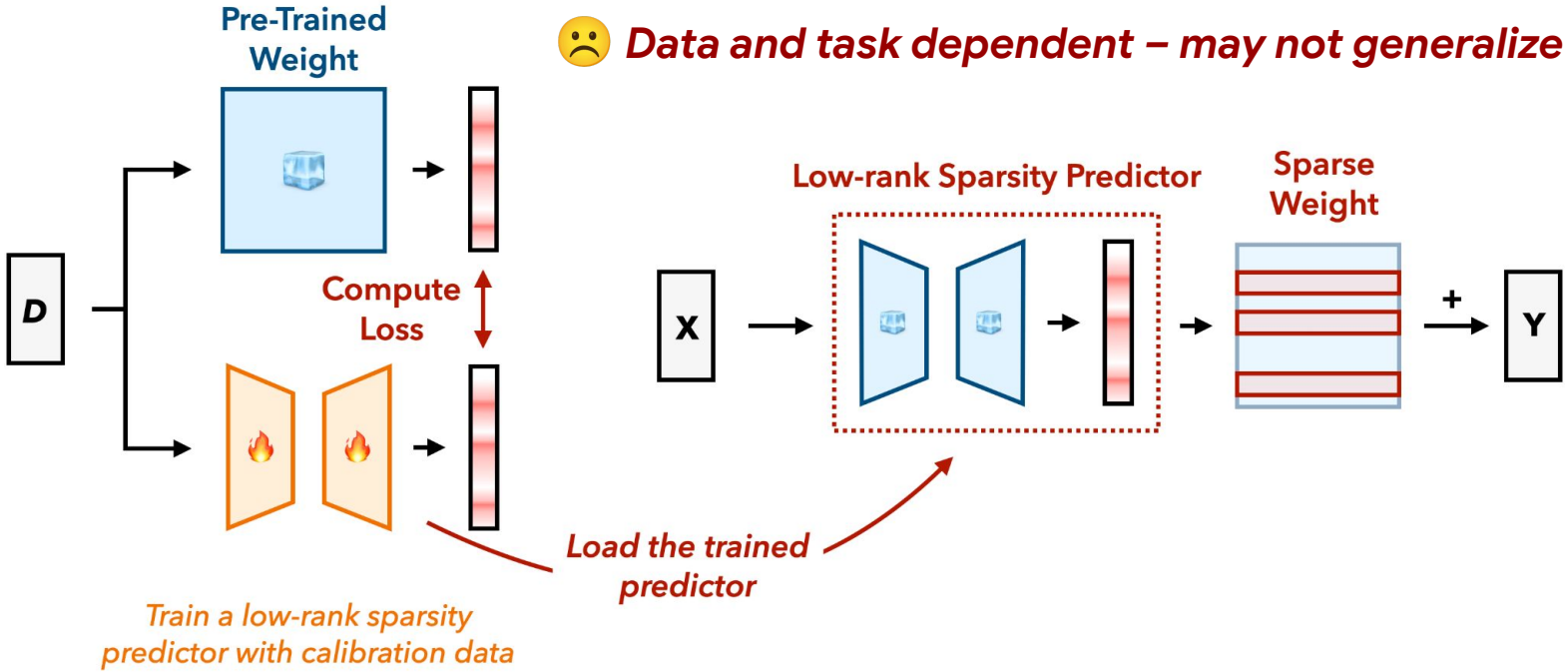
Leveraging Contextual Sparsity in Fine-tuning

Need to know the sparsity mask before going through the weight!



How to Identify Sparsity On the Fly?

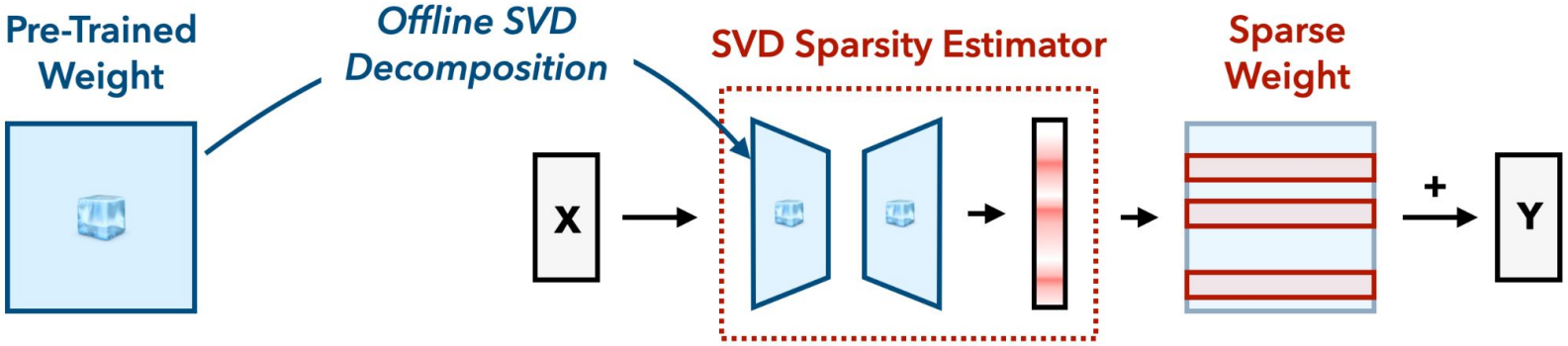
- **Prior work: train a low-rank predictor for each layer to identify channels to prune**



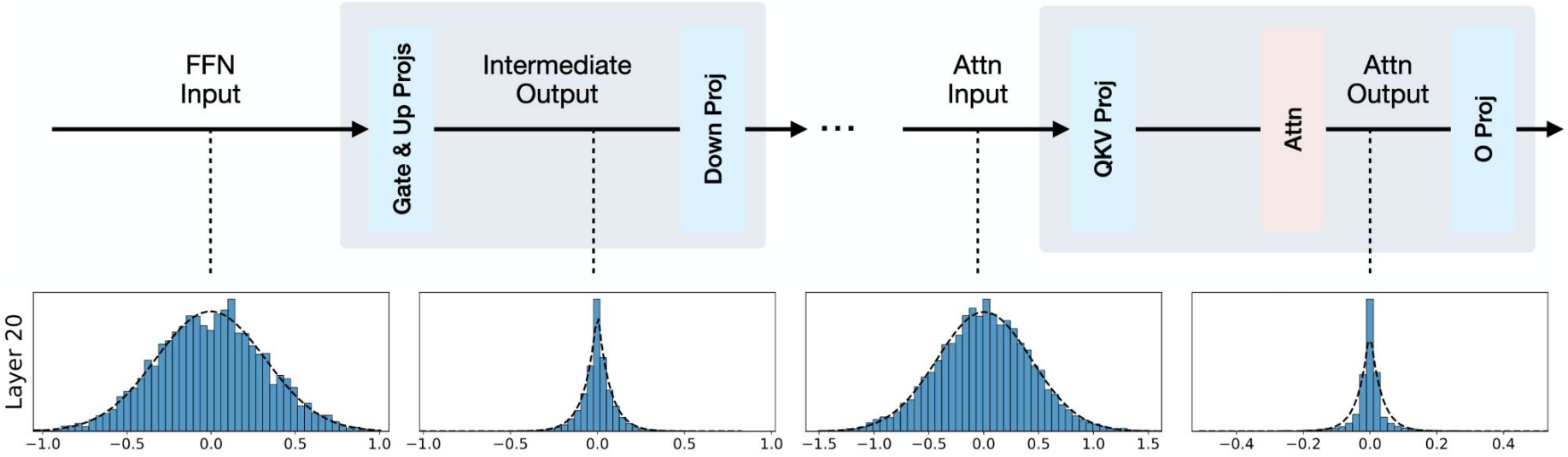
How to identify sparsity on-the-fly?

- **Our solution: use the top-k SVD of the model weights as the sparsity estimator**

😊 Data and task independent – simple & generalize!



Sparse Neuron Selection across Layers



- O projections inputs follow a Laplace distribution as well – can use L2 Norm
- QKV projections inputs follow a Normal distribution – L2 norm not applicable

Sparse Neuron Selection on QK Projections

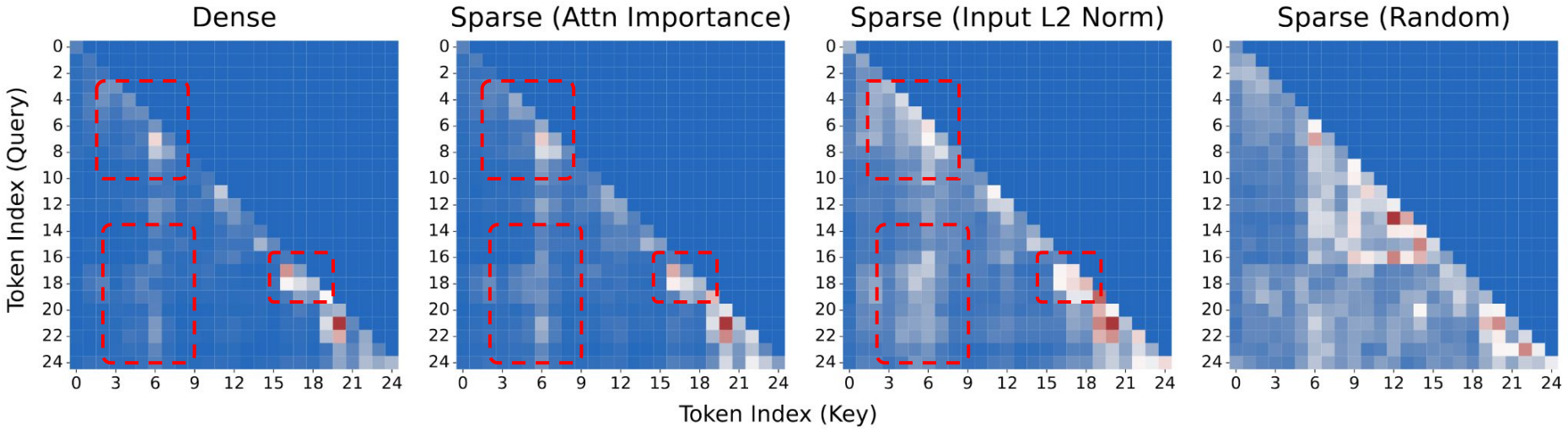
- **Attention scores guided criteria – sparsifying channels with minimal contributions**
  - Define a proxy metric that quantifies each channel's importance

$$\mathbf{q} = \|\mathbf{Q}\|_2, \quad \mathbf{k} = \|\mathbf{K}\|_2. \quad \mathbf{s} = \mathbf{q} \odot \mathbf{k}.$$

Sparse Neuron Selection on QK Projections

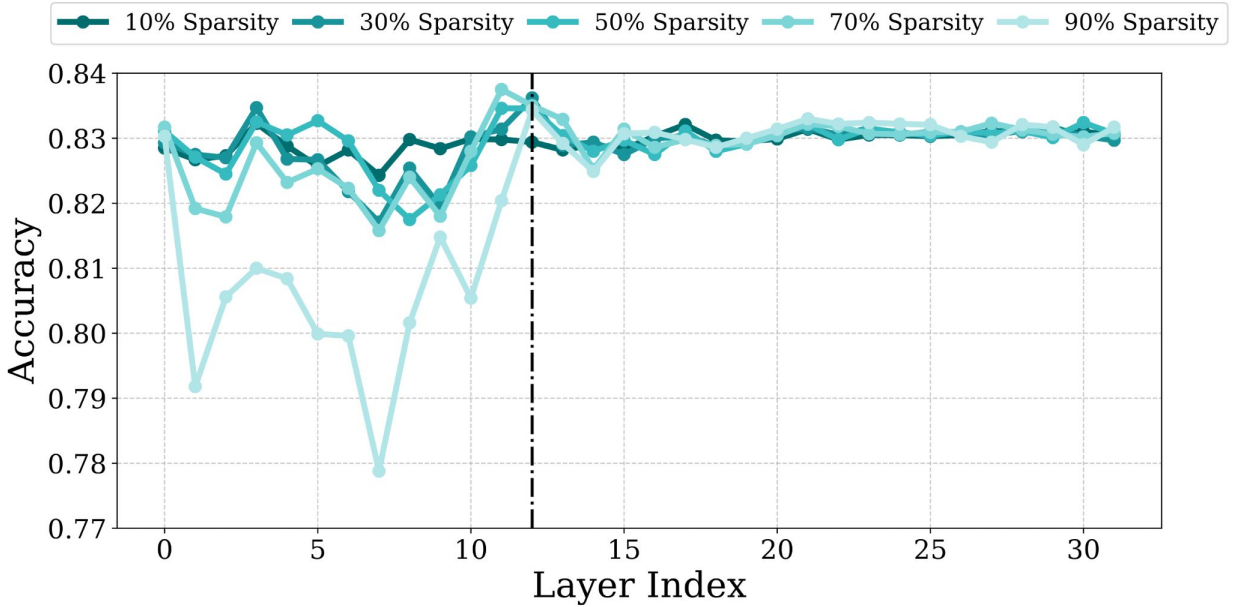
- **Attention scores guided criteria – sparsifying channels with minimal contributions**
  - Define a proxy metric that quantifies each channel's importance

$$\mathbf{q} = \|\mathbf{Q}\|_2, \quad \mathbf{k} = \|\mathbf{K}\|_2. \quad \mathbf{s} = \mathbf{q} \odot \mathbf{k}.$$



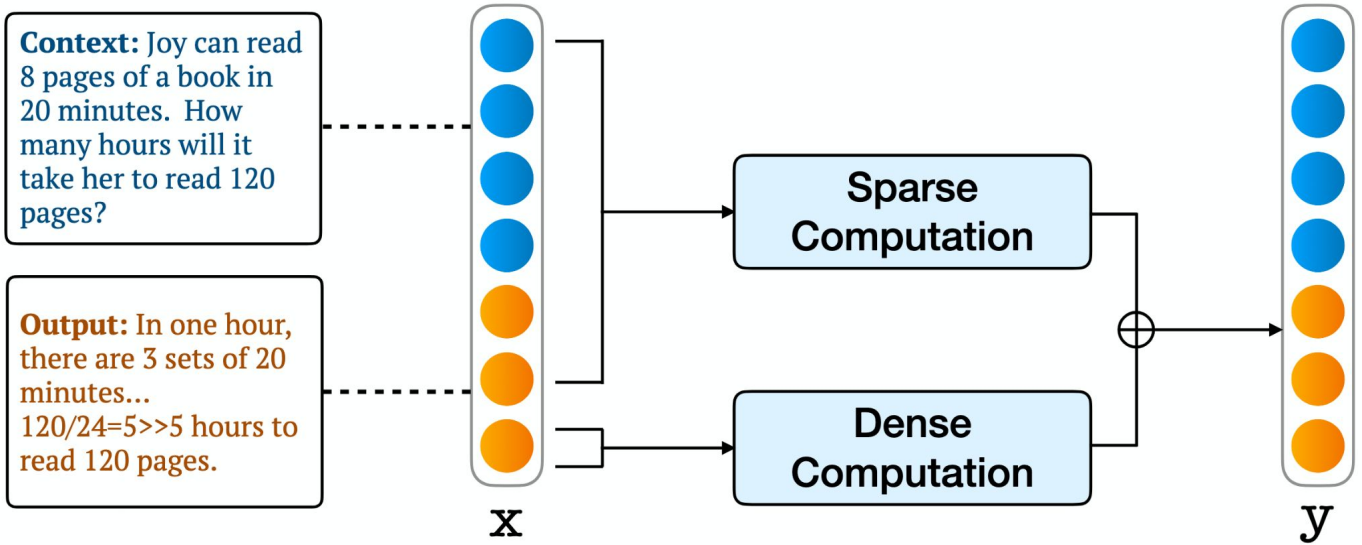
Layer Sensitivity: Adaptive Sparsity Configuration

- **The importance of individual layers and their contributions can differ substantially**
  - Conduct systematic layer sensitivity analysis with one task subset
  - Derive a layer-specific sparsity configurations for optimal performance



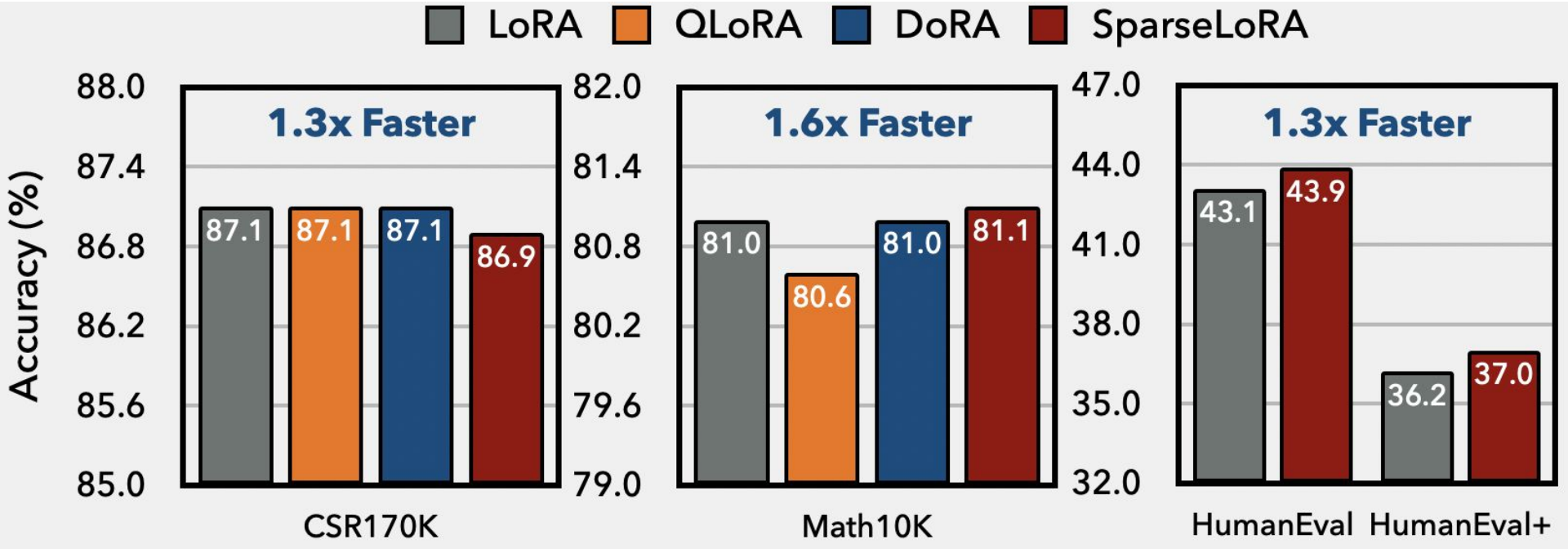
Token Sensitivity: Context-Output Aware Sparsity

- **Effectiveness of sparsity varies across tokens within a sequence**
  - All ***context*** tokens (the prefix tokens provided as input) are insensitive to sparsity
  - Some ***output*** tokens (the target tokens for loss computation) should be kept dense



Results

SparseLoRA maintains **lossless performance** on Math Reasoning & Code Generation and achieves up to **1.6x speedup** (on LLaMA3-8B)



SparseLoRA unlocks faster training

🔥 **1.5x Faster** 🔥

🔥 **1.8x Faster** 🔥

SparseLoRA accelerates PEFT LoRA 🤗

Results

SparseLoRA can be seamlessly combined with QLoRA,
achieving both **lower memory consumption** and **improved runtime efficiency**

	CSR170K			Math10K		
	#FL.	Spd.	Acc.	#FL.	Spd.	Acc.
LLaMA3-8B	–	–	62.5	–	–	33.5
+ QLoRA	100%	1.0×	87.1	100%	1.0×	80.6
+ **SparseQLoRA**	65%	1.2×	86.9	60%	1.3×	80.8

Samir Khaki*, Xiuyu Li*, Junxian Guo*, Ligeng Zhu, Konstantinos N. Plataniotis, Amir Yazdanbakhsh, Kurt Keutzer, Song Han, Zhijian Liu, Accelerating LLM Fine-Tuning with Contextual Sparsity, ICML 2025.

Thank you!

https://z-lab.ai/projects/sparselora

Samir Khaki*[1], Xiuyu Li*[2], Junxian Guo*[3], Ligeng Zhu[3], Konstantinos N. Plataniotis[1], Amir Yazdanbakhsh[4], Kurt Keutzer[2], Song Han[3], Zhijian Liu[3]