Monarch: Expressive Structured Matrices for Efficient and Accurate Training

Tri Dao, Beidi Chen, Nimit Sohoni, Arjun Desai, Michael Poli, Jessica Grogan, Alexander Liu, Aniruddh Rao, Atri Rudra, Christopher Ré

Stanford University   University at Buffalo

Sparse Training for Large Models

Challenges with structured linear maps (low-rank, sparse, Fourier):

Sparse end-to-end training   Dense-to-sparse finetuning

Efficiency-quality tradeoffs:

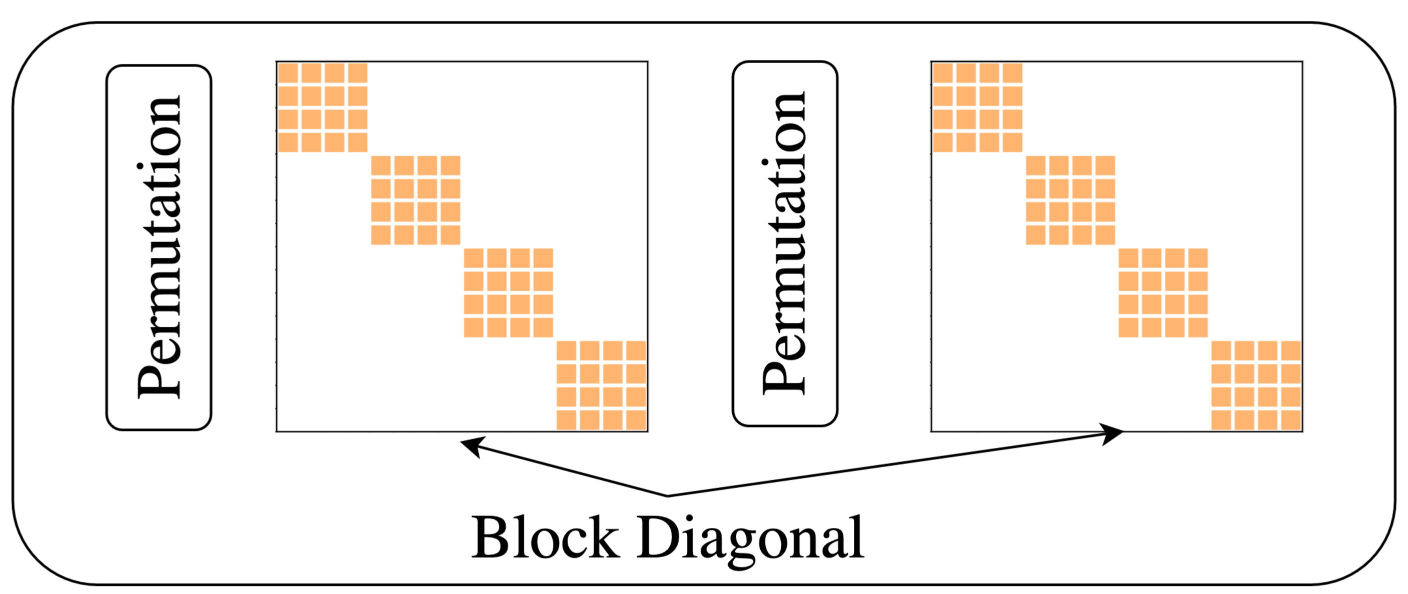
1. Efficiency: on modern hardware (GPU)

2. Quality: how expressive are the weight matrices (can they represent commonly used transforms)

3. Projection: How to find a sparse/structured matrices closest to a pretrained dense weight matrix

Are there structured matrices that are efficient, expressive, and with tractable projection algorithm? Yes

Monarch Matrices: Efficient and Expressive

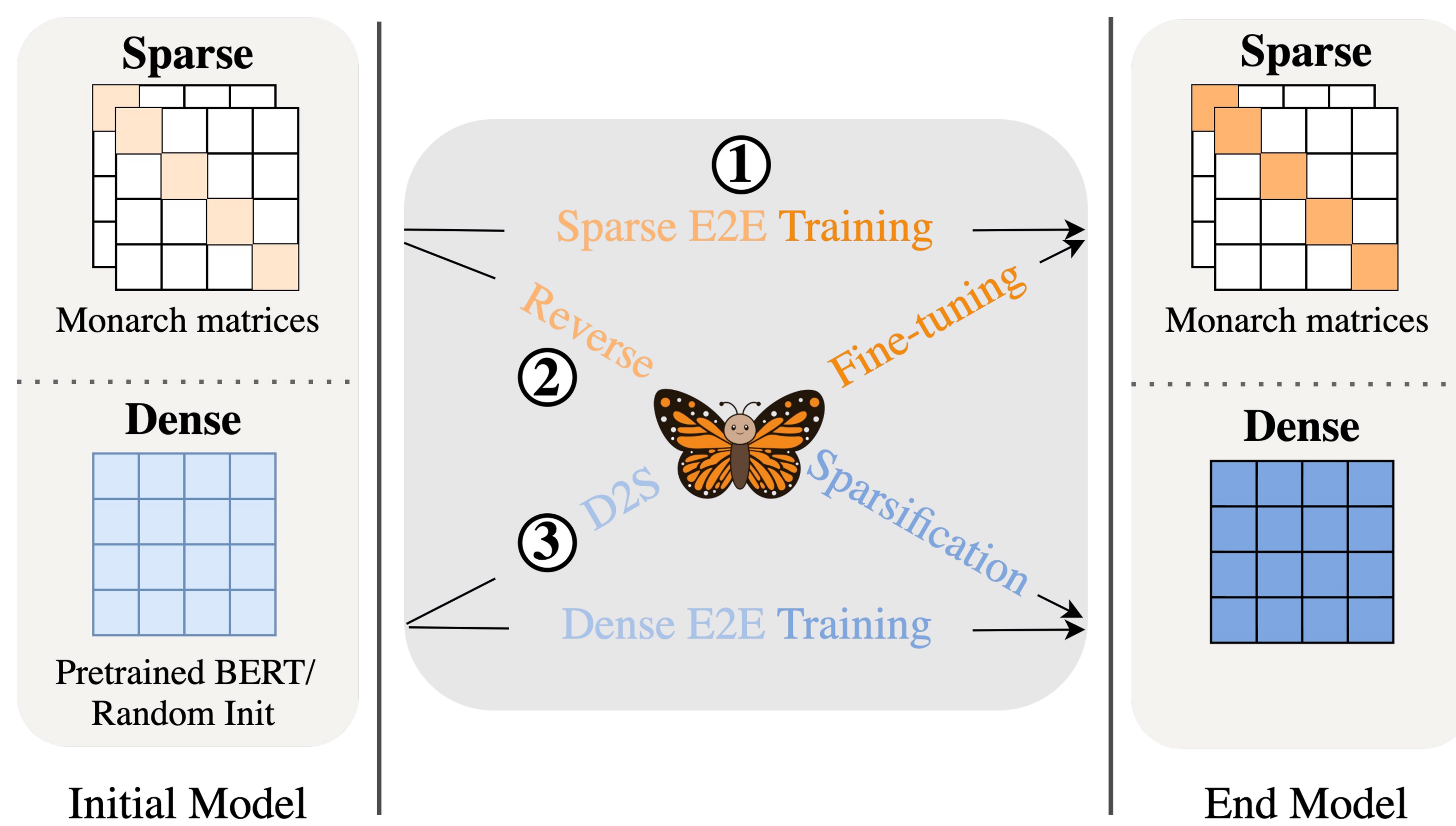


1. Hardware-efficient: Block-diagonal leverages efficient batch-matrix-multiply on GPU.

2. Expressive: contains butterfly matrices (and Fourier, DST, DCT, convolution, Hadamard, etc.)

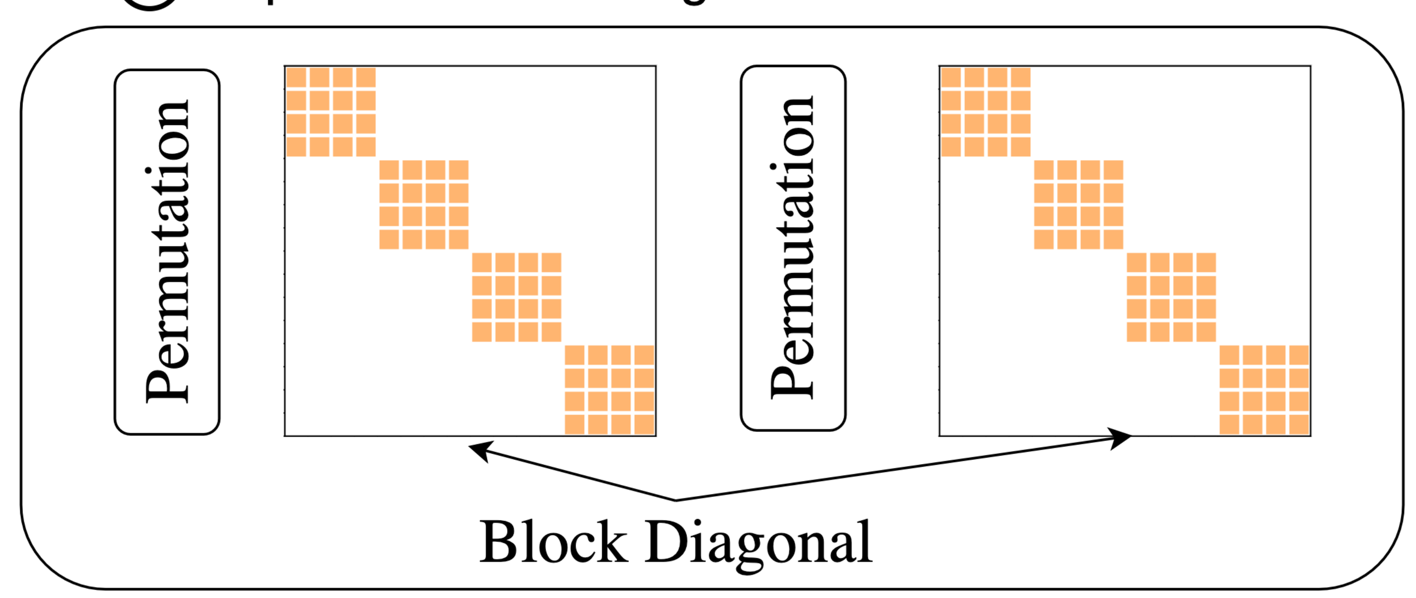
3. Tractable projection: find a Monarch matrix closest to a given dense matrix.

Three Ways to Use Sparse Models



Sparse End-to-End Training

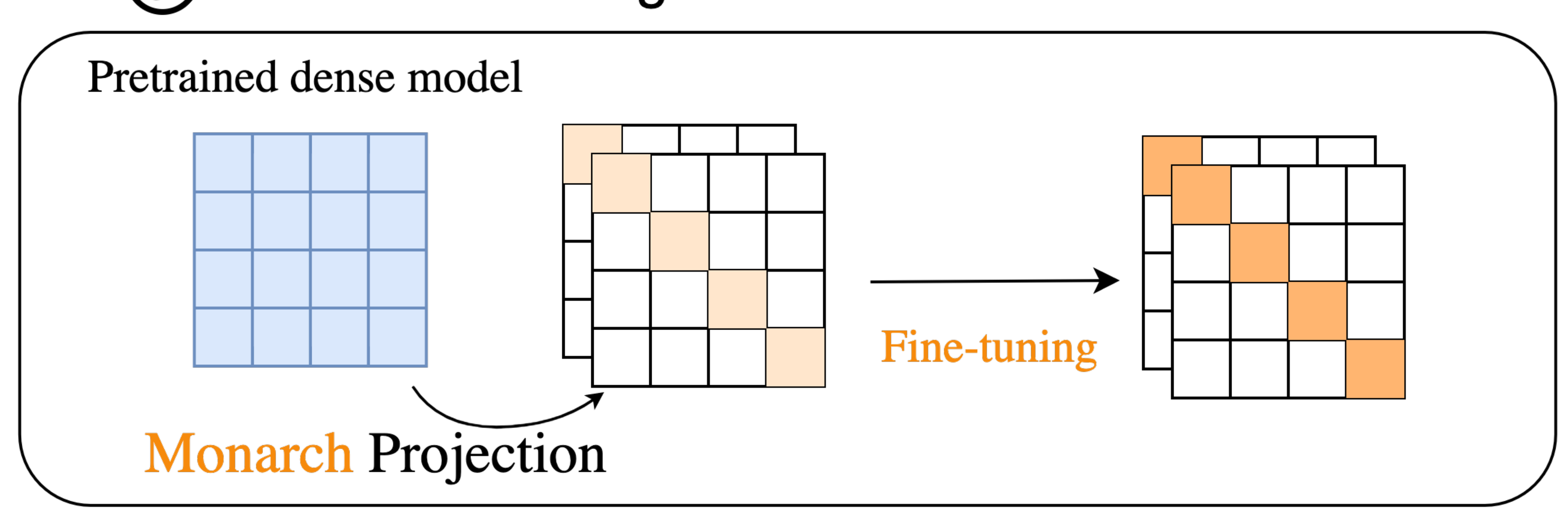
① Sparse E2E Training



Replace dense weight matrices (e.g., attention & FFN) with Monarch matrices for efficiency.

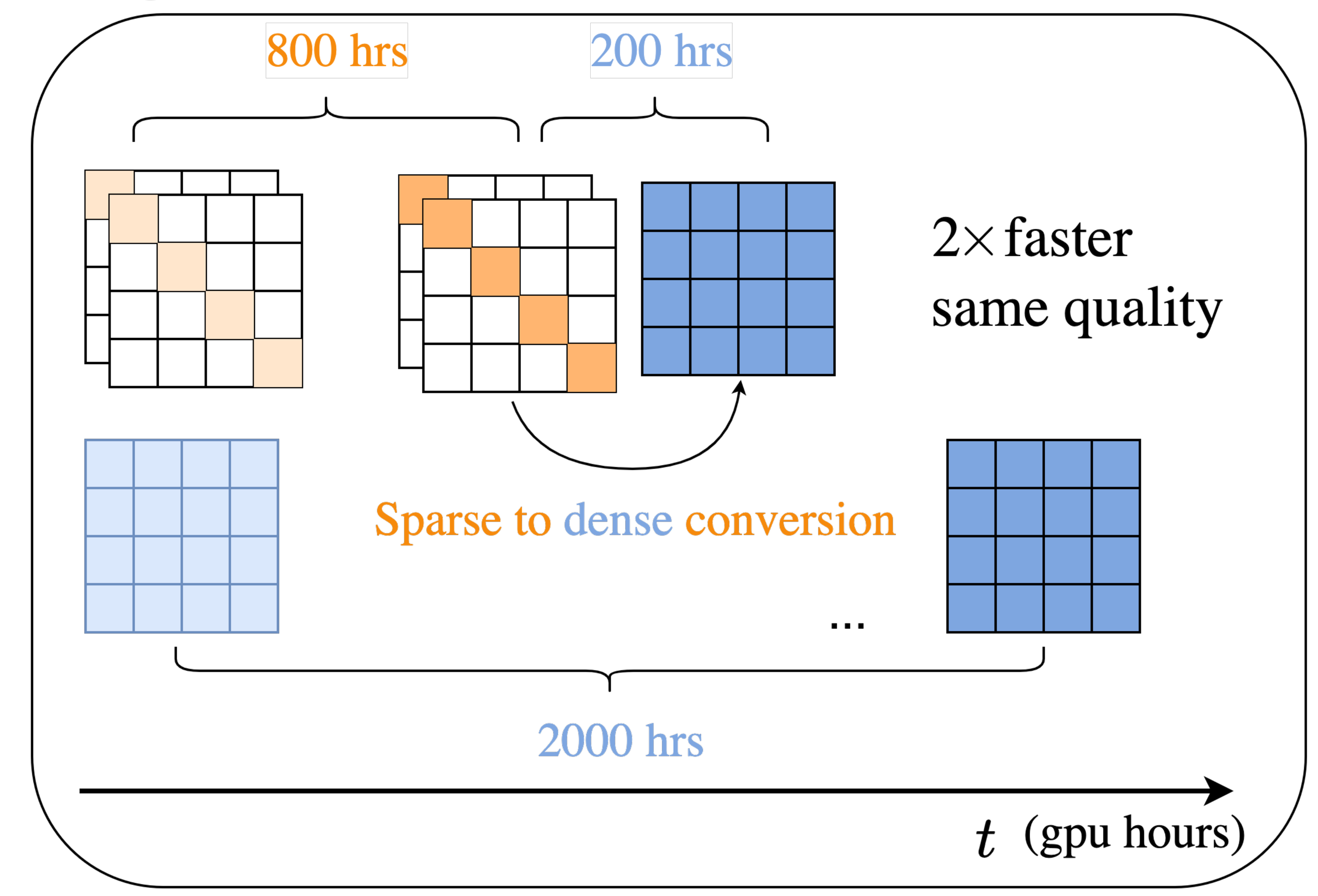
Dense-to-sparse Finetuning

③ D2S Fine-tuning



Sparse-to-Dense Training

② Reverse Sparsification



Sparse-to-dense speeds up training without losing performance 🚀!

Results: Monarch speeds up training from scratch and Finetuning

Sparse end-to-end training

Model	WikiText 103(ppl)	Speedup
GPT-2 Small	20.6	-
Monarch GPT-2-small	20.7	**1.8 x**

Sparse-to-dense training

Implementation	BERT-large training time on 8xA100s (h)
HuggingFace	84.5
MegaTron	52.5
Nvidia MLPerf 1.1	30.2
Monarch	**23.8**

Dense-to-sparse finetuning

Model	GLUE (avg)	Speedup
BERT-large	80.4	-
Monarch BERT-large	79.6	**1.7 x**

Up to 3.5x training speedup without performance loss.