Unified Symbolic Modeling and Adaptive Tensor Partitioning for Efficient Deep Learning Parallelization

Hongxing Wang[1], Zhengdao Yu[1,2], Chong Li[1], Serge Petiton[3]

[1]DPSL-Paris, [2]STL Sorbonne Université, [3]CRIStAL & CNRS

Background

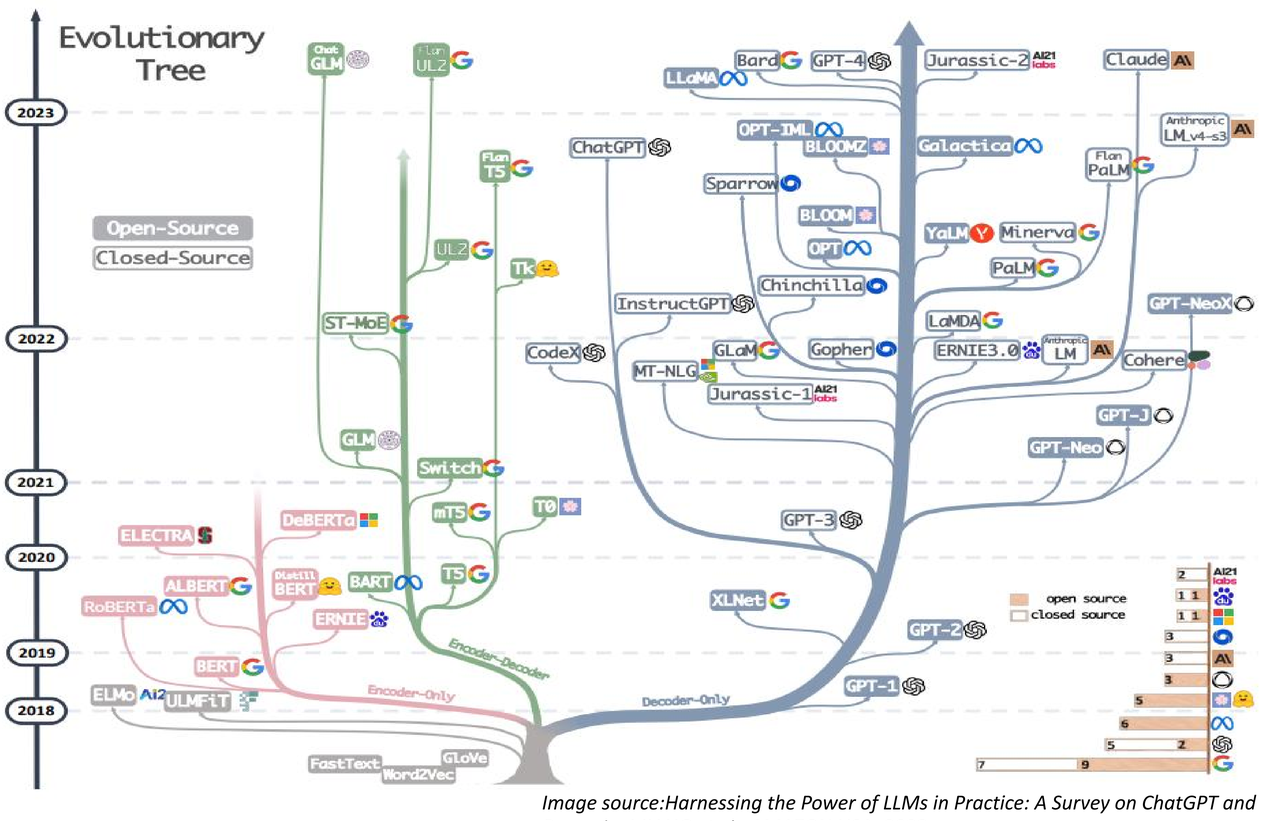


Image source:Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond，J JIANG et al., AMAZON USA, 2023

Rapid Evolution of LLMs

The evolution of LLMs remains extraordinarily rapid:

- **Structural diversity:** from CNNs to ResNets, recently to Transformer and MoE
- **Operator innovations:** ex: Attention -> FlashAttention
- **Expansion of parameters:** from millions to hundreds of billions

which made distributed training essential.

However, this rapid evolution of LLMs further **exacerbates the difficulty of adapting parallelization strategies to changing models**.

Challenges

Challenge1: Parallelism strategies are handled independently via fixed rules. Strategy combinations yield uncontrollable performance impacts, as positive effects from individual strategies may be compromised when combined.

Parallelism	Benefits	Costs	Publication Time
Data Parallelism(DP)	Compute scalability	Gradient synchronization	2012
Tensor Parallelism(TP)	Memory efficiency	Frequent collective communications	2019
Expert Parallelism(EP)	Parameter scaling	Expensive AllToAll communication	2021
Sequence Parallelism(SP)	Long-sequence support	Expensive synchronization across time steps	2021

Challenge2: Existing DL frameworks rely on predefined rules and paradigms. Such design lacks adaptability to evolving model architectures and parallelism strategies.

- Megatron-LM:
  Megatron hardcodes DP and TP within Transformer blocks,while assuming uniform structure across layers, therefore lacking adaptability to diverse architectures or emerging parallelism strategies
- Alpa:
  Despite automatic optimization, Alpa remains constrained by predefined DP/TP/PP paradigms, unable to adapt to emerging strategies like SP or EP

Our Solution: SymTensor

Dimension-wise Representation of parallelization strategies

DL model is described as a directed acyclic graph G = (V, E), where

- each node v ∈ V corresponds to a operator
- each edge **e ∈ E** corresponds to a **tensor flowing** between operators.

↓ For each **tensor T_e ∈ E**

we define tensor's shape as:
shape(T_e) = (d_1, d_2, . . . , d_k), where d_i ∈ ℕ* denotes the size of a tensor dimension

↓ Map parallelism to dimension-wise partitioning

we define a strategy over **T_e** as:
strategy(T_e) = (s_1, s_2, . . . , s_k)
where s_i ∈ ℕ* specifies the number of partitions along dimension d_i.

Unified Symbolic Cost Model

Total Cost of operator v

$$Cost_{total}(v) = \sum_{tensors\ T\ of\ v} \left(\frac{Cost_{mem}(T)}{\gamma} + Cost_{comm}(T)\right)$$

- Memory Cost of tensor T

$$Cost_{mem}(T) = \sum_{i=1}^{k} \frac{d_i}{s_i}$$

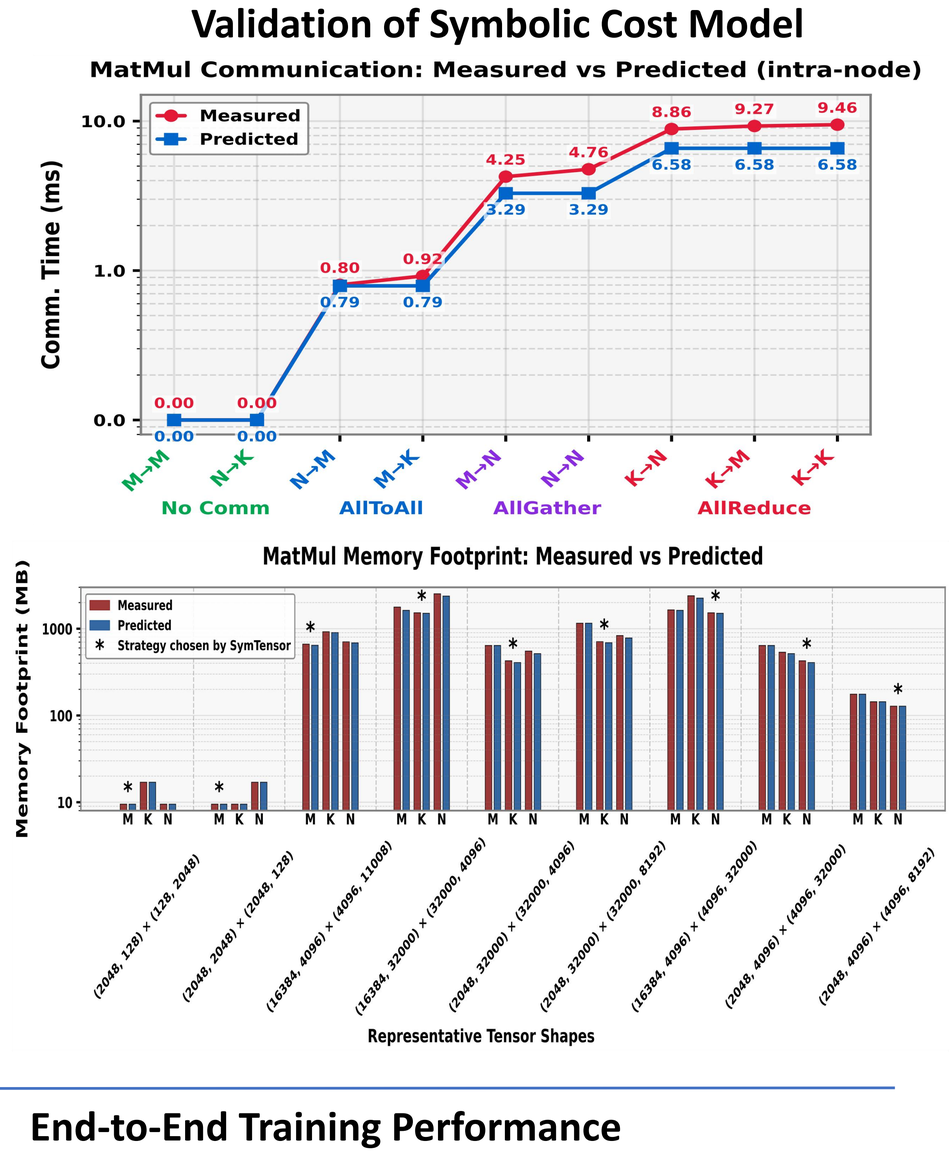
- Communication Cost of tensor T

$$Cost_{comm}(T) = CollectiveCost(T)$$

- Computation Cost

We exclude computation cost to simplify the cost model, since an operator's total FLOPs remain unchanged across different partitioning schemes.

Validation of Symbolic Cost Model



Experiments

Adaptability Evaluation

This experiment evaluates SymTensor's ability to adapt to common structural changes encountered in practice:

- **LLama2-13B**: replace self-attention modules with FlashAttention
- **Mixtral 8x7B**: which features a novel MoE design
- **Qwen-7B-LoRA:** increase the batch size from 8 to 32

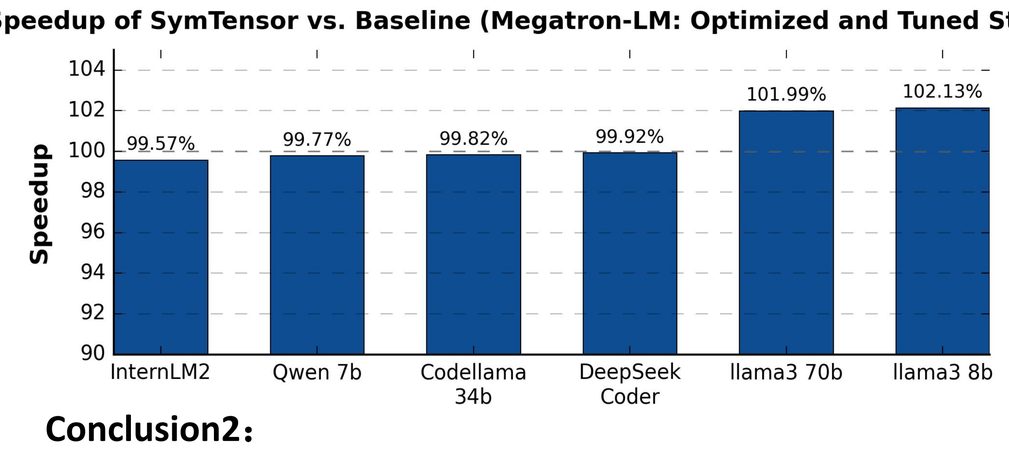
Table 1: Training Throughput Comparison (in tokens/sec)

Model	Megatron-LM	SymTensor	Speedup
LLaMA2 13B	10961	15845	144.56%
Mixtral 8×7B	2936	6506	221.59%
Qwen 7B LoRA	OOM	17626	∞

Conclusion1：
SymTensor resolves adaptability & scalability limitations across changing model architectures through adaptive strategy selection.

End-to-End Training Performance

This experiment evaluates the overall effectiveness and generalization capability of our strategy generation method.



Conclusion2：
SymTensor preserves generalizability across diverse model architectures while reducing manual optimization from days to seconds-level search.