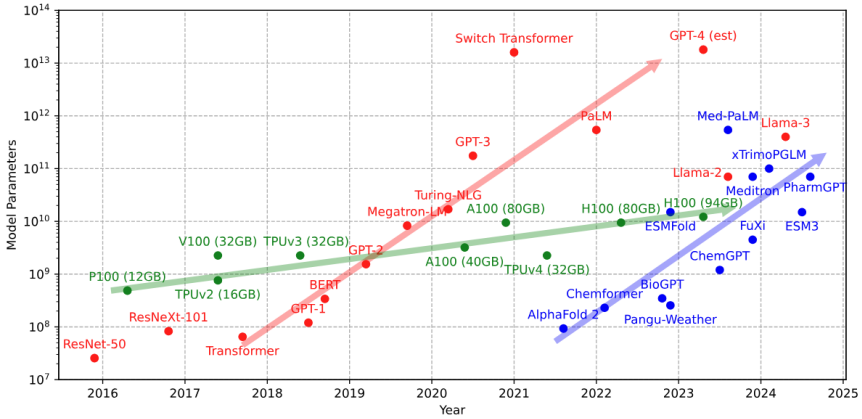
A Survey on Memory-Efficient Transformer-Based Model Training in AI for Science

Kaiyuan TIAN, Linbo QIAO, Baihui LIU, Gongqingjian JIANG, Shanshan LI, Dongsheng LI

Frontiers of Computer Science, DOI: 10.1007/s11704-025-50302-6

Problems & Ideas

- Problems of existing reviews on memory-efficient LLM training:
  - Existing reviews lack a dedicated and comprehensive focus on memory-efficient training specifically for large-scale transformers in scientific AI.
  - There is a notable absence of a comprehensive review that summarizes the application of memory-efficient training techniques within the domain of AI for Science.
- Ideas: We present a systematic review to bridge the gap between memory-efficient training methodologies and the needs of scaling scientific models.

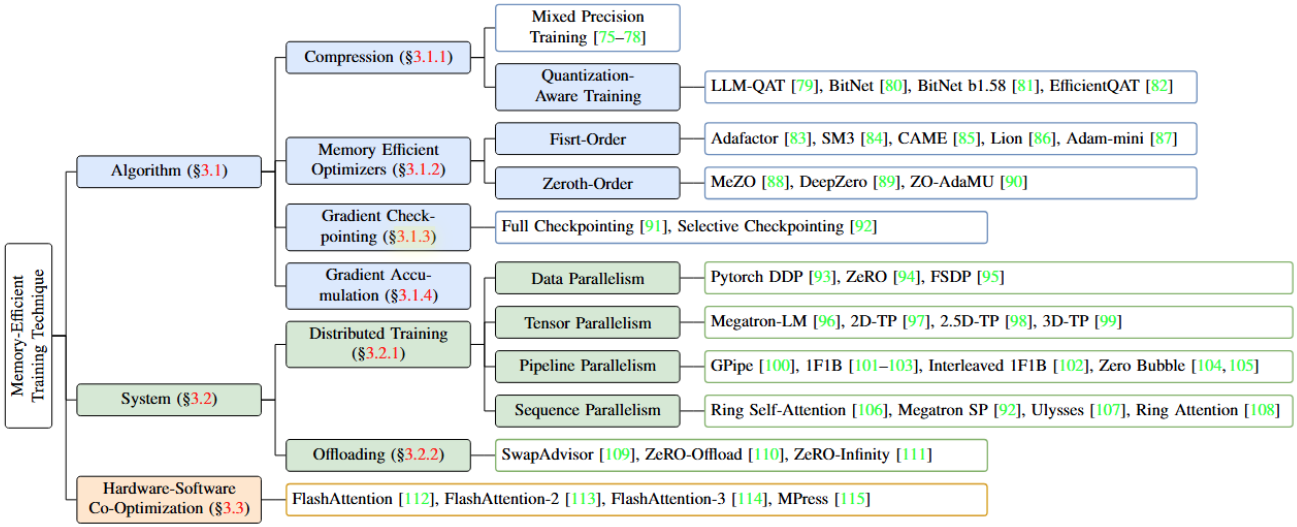


Field	Work	Backbone	Main Building Block	#Parameters	Memory Cost (est.)	Optimizations	Tasks
Biology	AlphaFold 2 [2]	-	Evoformer [2]	93M	1.45 GB	DP mixed-precision GC	protein structure prediction
	RosettaFold [40]	-	SE(3)-Transformer [41]	130M	2.03 GB	DP GA	protein structure prediction
	AlphaFold 3 [42]	-	Pairformer [42]	Unreported	-	Unreported	biomolecular complex structure prediction
	OpenFold [43]	-	Evoformer	93M	1.45 GB	DP (ZeRO-2) mixed-precision offloading GC, GA	protein structure prediction
	FastFold [44]	-	Evoformer	93M	1.45 GB	DP DAP	protein structure prediction
	ScaleFold [45]	-	Evoformer	97M	1.52 GB	DP DAP	protein structure prediction
	ESMFold [46]	ESM-2 15B	Transformer [17]	15B	240 GB	DP (FSDP)	protein structure prediction
	xTrimoPGLM [47]	xTrimoPGLM-100B	Transformer	100B	1.56 TB	DP (ZeRO-1) PP (1F1B) TP (Megatron-LM) mixed-precision GC	protein understanding protein generation
	ESM3 [48]	-	Transformer	1.4B / 7B / 98B	1.53 TB	DP (FSDP) mixed-precision	protein reasoning protein generation

Left: AI memory wall. The growth of model parameters has surpassed the increase in memory capacity of accelerators; Right: Applications of LLMs in biology field and corresponding training optimization strategies.

Main Contributions

- Contributions:
  - We comprehensively summarized the applications of LLMs across various scientific domains, including biology, medicine, chemistry, meteorology, and geoscience;
  - We offered a systematic taxonomy on memory-efficient pre-training techniques for transformers, covering algorithm-level, system-level, and hardware-software co-optimization;
  - We emphasized the specific memory-related challenges and opportunities for transformer-based models in AI for science.



The overview of memory-efficient training techniques for transformers.