

MoARa: Module-Aware Rank Allocation and Structure-Preserving Decomposition for Low-Rank LLM Pre-training

Keunyoung Kim, Nojun Kwak

Seoul National University, South Korea



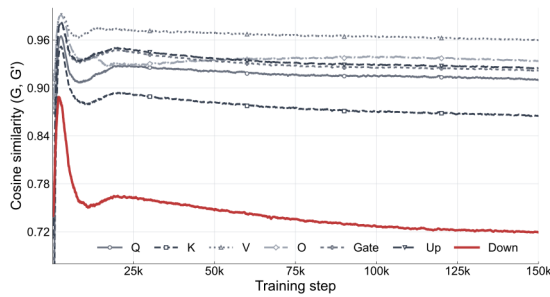
TL;DR

1. Uniform projection ranks overlook differences in module sensitivity.
2. Projecting raw gradients can distort both local magnitudes and directions.
3. MoARa combines projection rank reallocation and gradient decomposition, cutting time-to-target by $\approx 34\%$ on 7B scale vs. GaLore.

MOTIVATION

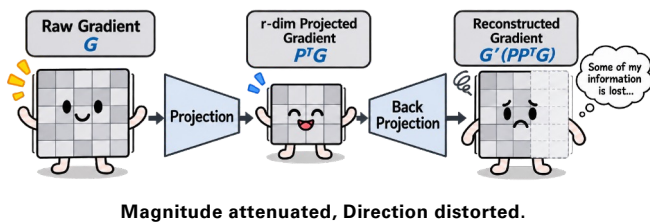
Low-rank projection saves optimizer memory.
Can we reach the same training quality faster?

Uniform Projection Rank Allocation



- MLP.Down shows the lowest gradient reconstruction alignment.
- Rank-reduction tests identify Q/K as robust donors.

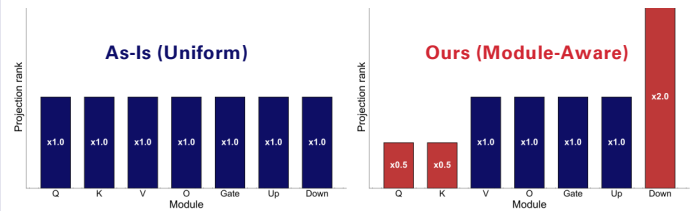
Information Compression via Projection



Magnitude attenuated, Direction distorted.

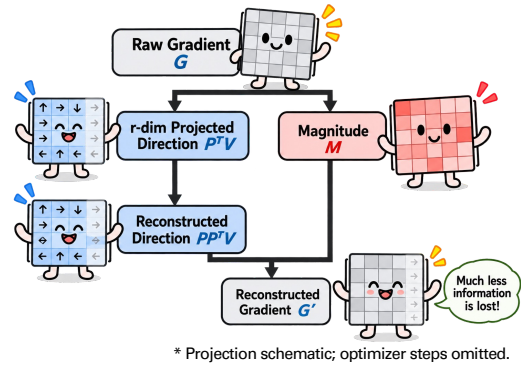
METHOD

Module-Aware Projection Rank Allocation



- Same total projection rank budget, allocation fixed during training.
- Profile once, reuse at multiple scale in the same architecture.

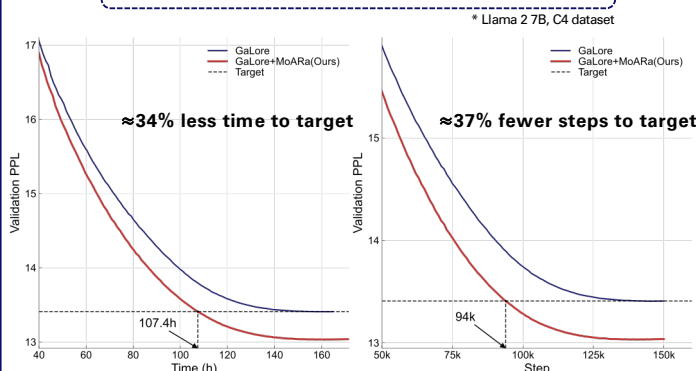
Block-wise Gradient Decomposition



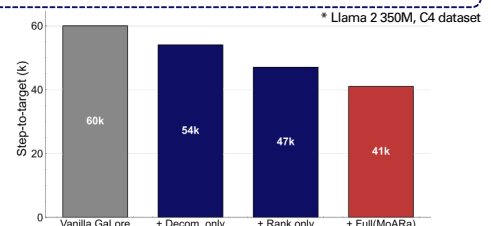
* Projection schematic; optimizer steps omitted.

RESULTS

Same Target Perplexity, Less Training Time.



Contribution Breakdown



Tested Architectures, Scales, and Methods

- MHA (350M, 1B, 7B) / GQA (300-350M) / MLA (350M)
- GaLore / SubTrack++ / OSD / LDAdam / Fira / Q-GaLore
- Projection rank allocation improves all tested methods.
- Decomposition gains depend on the host projection method.