TensLoRA: Tensor Alternatives for Low-Rank Adaptation

Axel Marmoret[1], Reda Bensaid[1,2], Jonathan Lys[1], Vincent Gripon[1], and François Leduc-Primeau[2]

[1]IMT Atlantique, Lab-STICC, UMR CNRS 6285, F-29238, Brest, France
[2]Polytechnique Montréal, Canada

TL;DR

- **A Unified Framework:** TensLoRA systematically models tensor-based low-rank adaptations by aggregating LoRA updates into higher-order tensors.
- **Structure Impacts Performance:** Tensor architecture critically impacts model performance $\implies$ redundancy across attention modes is not uniform.
- **Outperforming LoRA:** With similar parameter budgets, some TensLoRA configurations consistently surpass standard LoRA.

What is LoRA?

Low-Rank Adaptation (LoRA) [1] adapts large pre-trained models via learnable low-rank matrices. Given an input $\mathbf{x} \in \mathbb{R}^d$, the LoRA update for an attention projection writes as:

$$\mathbf{h} = \mathbf{W_0 x} + \mathbf{ABx}, \quad \text{with} \quad \mathbf{W_0} \in \mathbb{R}^{d\times d}, \mathbf{A} \in \mathbb{R}^{d\times r}, \mathbf{B} \in \mathbb{R}^{r\times d}$$

where $\mathbf{AB}$ are the trainable matrices, and $\mathbf{W_0}$ the original pretrained weights.

- **Pros:** Highly parameter-efficient, minimal training costs, and zero inference latency.
- **Cons:** Updates layers and projections (Query, Key, Value) independently.
  $\implies$ This ignores correlations, creates redundancy, and scales parameters linearly.

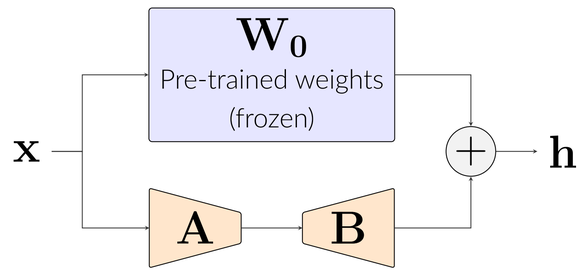


Figure 1. Schematic representation of LoRA.

Tensor-based LoRA

Motivation: Aggregate LoRA updates into higher-order tensors to reduce redundancy and maximize parameter efficiency.

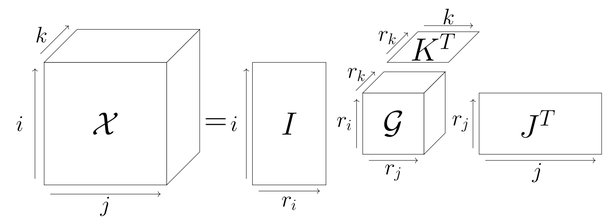
Gap in Existing Works: Prior works (FacT [2], LoTR [3], LoRTA [4], CaRA [5]) **lack a systematic framework** comparing tensor constructions.

Why TensLoRA?

→ **Generalizes** existing methods into a unified paradigm.
→ **Systematically compares** different tensor architectures.
→ Introduces **mode-specific compression rates** (exploiting non-uniform redundancy).

Tucker Factorization [6]

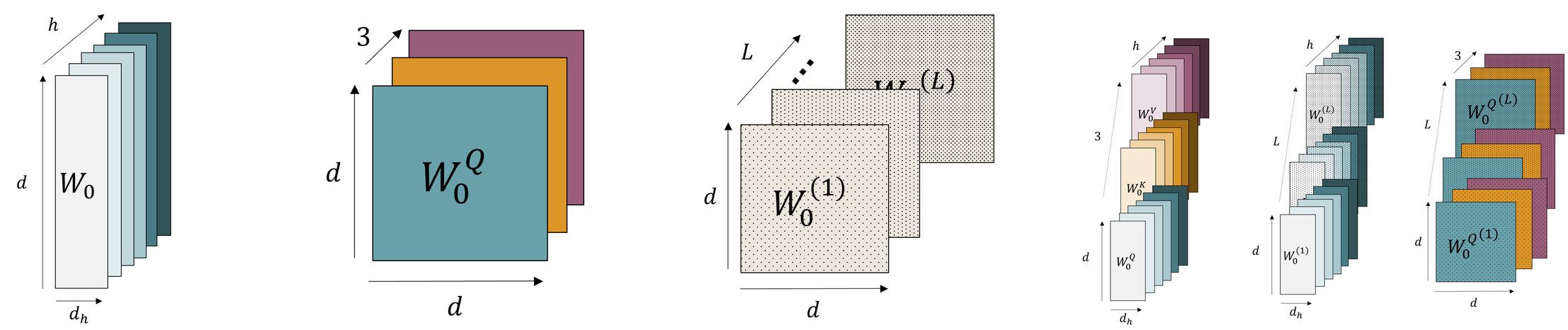
For a 3rd-order tensor $\mathcal{X} \in \mathbb{R}^{i\times j\times k}$: $\mathcal{X} \approx \mathcal{G} \times_1 \mathbf{I} \times_2 \mathbf{J} \times_3 \mathbf{K}$



Motivation:

→ **Reduces parameter count** while preserving expressivity.
→ One specific rank r_c per mode $\implies$ **compression can be optimized per dimension**.

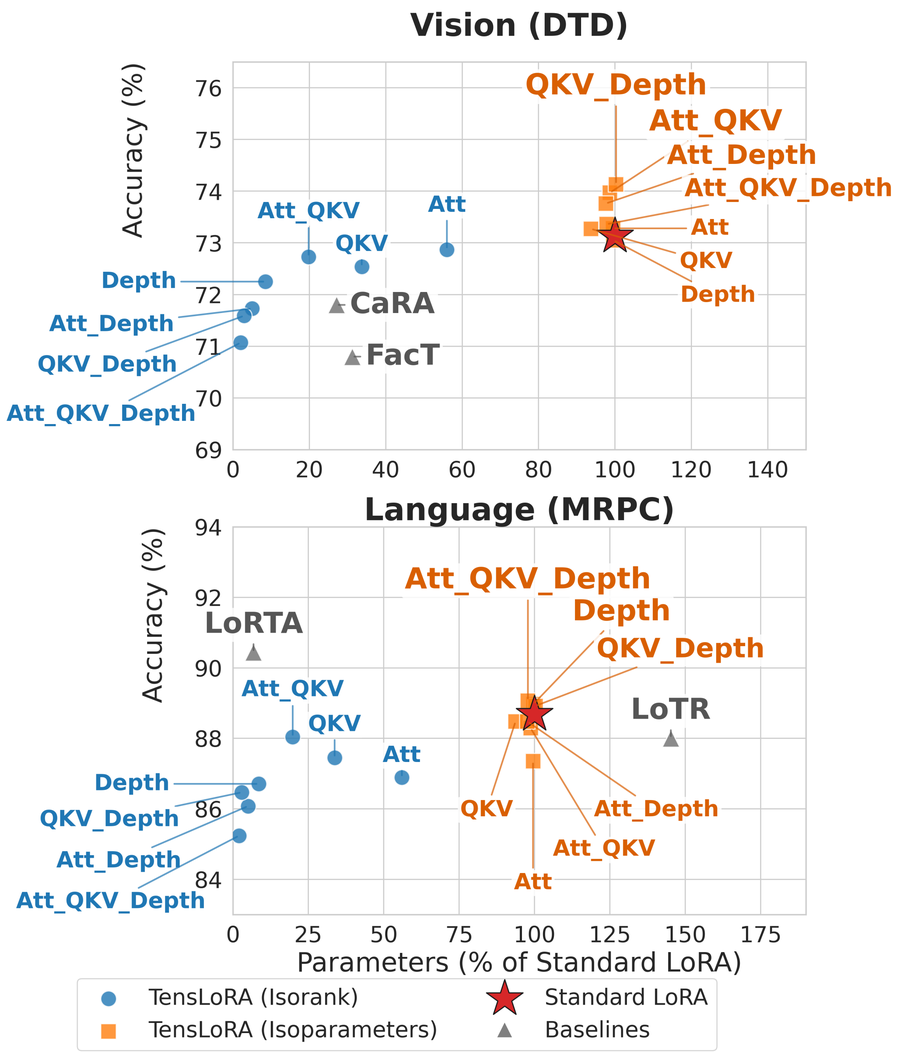
TensLoRA Representations: Aggregating LoRA Updates



(a) Grouping Attention Layers (**Att**) (b) Grouping Key/Query/Value (**QKV**) (c) Grouping Layers (**Depth**) (d) Combinations (4D tensors)

Figure 2. TensLoRA representations for **Att**, **QKV**, **Depth**, and combinations of two of them (4D tensors).

Experimental Results



Discussion

- **Optimal Aggregations:** Grouping QKV and depth (**QKV_Depth**) yields the best results overall.
  $\implies$ it probably exploits the most redundant adaptation subspaces.
- **The Head Bottleneck:** Aggregating attention heads (**Att**) is consistently the least effective.
  $\implies$ forcing a shared representation seems to degrade their specialized roles.
- **Compression Trade-offs:** Extreme compression (*isorank*) underperforms the LoRA baseline, whereas similar parameter budgets (*isoparameters*) may surpass it.
  $\implies$ Strategic parameter allocation is essential for performance.

Conclusion & Future Work

Conclusion:

- TensLoRA provides a unified and systematic framework for tensor-based low-rank adaptations.
- Some tensor alternatives outperform standard LoRA within similar parameter budgets.

Future Work:

- Investigate advanced rank allocation strategies and ablations.
- Explore alternative tensor factorizations (CP, Tensor-Train).

References

[1] E. J. Hu *et al.*, "Lora: Low-rank adaptation of large language models.," *ICLR*, vol. 1, no. 2, p. 3, 2022.
[2] S. Jie and Z.-H. Deng, "FacT: Factor-tuning for lightweight adaptation on vision transformer," in *Proc. AAAI Conf. Artificial Intelligence*, 2023.
[3] D. Bershatsky *et al.*, "LoTR: Low tensor rank weight adaptation," *arXiv preprint arXiv:2402.01376*, 2024.
[4] I. Hounie *et al.*, "LoRTA: Low rank tensor adaptation of large language models," *arXiv preprint arXiv:2410.04060*, 2024.
[5] L. Veeramacheneni *et al.*, "Canonical rank adaptation: An efficient fine-tuning strategy for vision transformers," in *ICML*, 2025.
[6] L. R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, 1966.

2026 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). 4-8 May, Barcelona, Spain.