FY2025/2025年度

Tensor Learning Team

Qibin Zhao

テンソル学習チーム　趙 啓斌

RIKEN Center for Advanced Intelligence Project

Research Overview

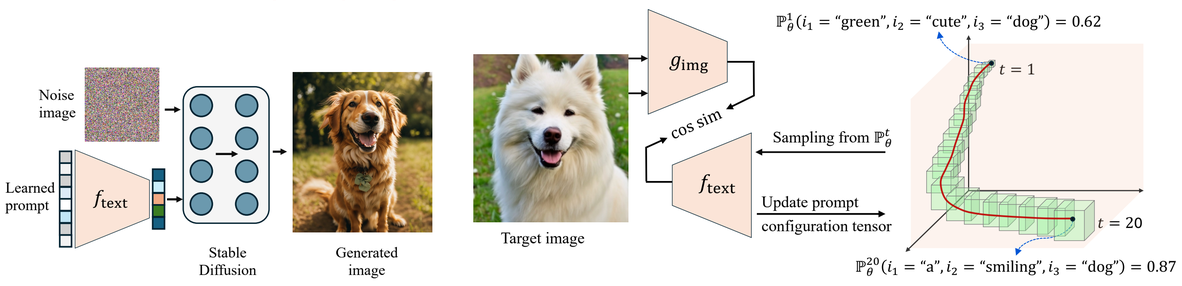
- **Research Directions**: Our team conducts research from several perspectives including tensor representation for machine learning, trustworthy machine learning and quantum machine learning.
- **Research Achievements (FY2025)**: We developed low-rank and tensor representation-based technology to improve the efficiency and scalability of modern machine learning models. Our key achievements include 1) efficient hard prompt optimization; 2) memory-efficient long-context inference for large language models (LLMs); 3) storage-efficient class incremental learning; 4) parameter-efficient model adaptation and transfer learning; and 5) tensor manifold frameworks and their associated learning theory.

Inference-Time and Continual Learning

Motivation: Our research focuses on developing efficient and adaptive learning mechanisms for large models across different stages of their lifecycle. These works address the challenges lying in prompt learning, inference-time approach for long-context reasoning and continual learning.

STEPS: Sequential Probability Tensor Estimation for Text-to-Image Hard Prompt Search (Qiu et al. CVPR 2025)

- **Hard Prompt Optimization**: Projected gradient descent suffers from slow convergence, while gradient-based token search is easy to get stuck in local minima and cannot make use of full candidate prompt space.
- **Methodology**: To search or optimize a prompt in a high-dimensional and discrete space, we propose to formulate loss landscape function as an approximate low-rank tensor and develop a sequential probability tensor estimation algorithm for efficient and effective hard prompt optimization.

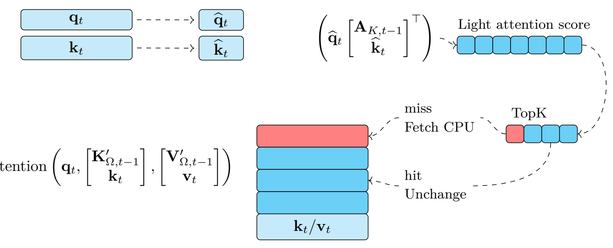


Efficient Low Rank Attention for Long-Context Inference in Large Language Models (Li et al. NeurIPS 2025)

- **Motivation**: Long context inference (i.e., >32K) induces linearly increasing KV cache, causing out-of-memory (OOM) failures in LLM inference.
- **Key Contribution**: To maintain accuracy and avoid heavy data transfers, we propose a low-rank attention and a dynamic CPU-GPU KV management by offloading KV cache to CPU and maintaining only a few pairs for GPU inference.

Prefill (prompt processing)

$$\arg\min_{\mathbf{A}_Q,\mathbf{B}_Q,\mathbf{A}_K,\mathbf{B}_K} \frac{1}{2}\left\|\mathbf{Q}\mathbf{K}^\top - \mathbf{A}_Q\mathbf{A}_K^\top\right\|_{\mathrm{F}}^2, \text{s.t. } \mathbf{Q} = \mathbf{A}_Q\mathbf{B}_Q, \mathbf{K} = \mathbf{A}_K\mathbf{B}_K.$$



Decode (autoregressive generation)

$$\arg\min_{\hat{\mathbf{q}}_t,\hat{\mathbf{k}}_t} \frac{1}{2}\left\|\hat{\mathbf{q}}_t\mathbf{B}_{Q,t-1} - \mathbf{q}_t\right\|_{\mathrm{F}}^2 + \frac{1}{2}\left\|\hat{\mathbf{k}}_t\mathbf{B}_{K,t-1} - \mathbf{k}_t\right\|_{\mathrm{F}}^2,$$

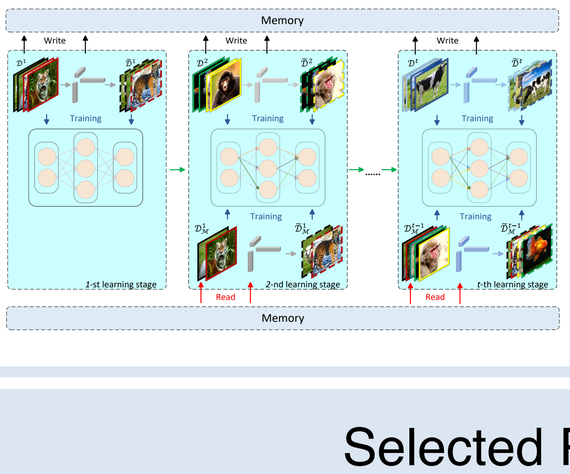
$$\text{s.t. } \hat{\mathbf{q}}_t\hat{\mathbf{k}}_t^\top = \mathbf{q}_t\mathbf{k}_t^\top, \hat{\mathbf{q}}_t\mathbf{A}_{K,\Omega,t-1}^\top = \mathbf{q}_t\mathbf{K}_{\Omega,t-1}^\top.$$

$$\text{Attention}\left(\mathbf{q}_t, \begin{bmatrix}\mathbf{K}'_{\Omega,t-1}\\ \mathbf{k}_t\end{bmatrix}, \begin{bmatrix}\mathbf{V}'_{\Omega,t-1}\\ \mathbf{v}_t\end{bmatrix}\right)$$

Select top-k relevant tokens; Fetch corresponding KV cache; Update low-rank basis.

Tensor Decomposition Based Memory-Efficient Incremental Learning (Li et al. ICML 2025)

- **Motivation**: Continual/Lifelong learning has unavoidable forgetting due to class and data distribution shifts. Replay based methods aim to address this issue but demand large storage for re-training.
- **Methodology**: We propose an adaptive training framework leveraging tensor decomposition, alongside a hybrid exemplar selection strategy to improve the storage efficiency and exemplar diversity.



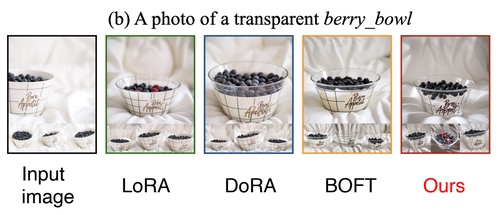
Low-rank Tensor for Model Adaptation

Transformed Low-rank Adaptation via Tensor Decomposition (Tao et al. ICCV 2025)

- **Motivation**: LoRA assumes fine-tuning adapter to be low-rank for parameter efficiency, while full fine-tuning adapter tends to be high-rank.
- **Methodology**: We propose the Transformed Low-rank Adaptation (TLoRA) method that is a mixture of transform adaptation and residual adaptation. By leveraging TN representation, TLoRA achieves better approximation to full fine-tuning as well as the improved parameter efficiency. Its application to Text-to-Image models validates the high quality of generated samples.

Transform adaptation preserving the pre-trained information — Residual adaptation learning compact task-specific knowledge

$$\boldsymbol{y}' = (\boldsymbol{W}_0\boldsymbol{T} + \boldsymbol{\Delta})\boldsymbol{x},$$



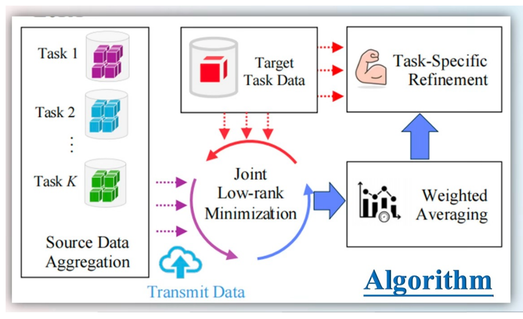
$$\boldsymbol{T}[\overline{i_1\cdots i_D}, \overline{j_1\cdots j_D}] = \mathrm{tr}(\mathbf{A}^1[i_1, j_1, :, :]\cdots\mathbf{A}^D[i_D, j_D, :, :]).$$

$$\boldsymbol{\Delta}[\overline{i_1\cdots i_D}, \overline{j_1\cdots j_D}] = \mathrm{tr}(\mathbf{B}^1[i_1, :, :]\cdots\mathbf{B}^D[i_D, :, :]\mathbf{C}^1[j_1, :, :]\cdots\mathbf{C}^D[j_D, :, :]).$$

Low-Rank Tensor Transitions (LoRT) for Transferable Tensor Regression (Wang et al. ICML 2025)

Tensor regression from scarce data is difficult $y_i^{(0)} = \langle\mathcal{X}_i^{(0)}, \mathcal{W}_\star^{(0)}\rangle + \epsilon_i^{(0)}$ ⬅ **borrowing from data rich tasks faces distribution shifts** $y_i^{(k)} = \langle\mathcal{X}_i^{(k)}, \mathcal{W}_\star^{(k)}\rangle + \epsilon_i^{(k)}$

- **Motivation:** Tensor regression suffers from data insufficiency and faces distribution shifts when using transfer learning.
- **Methodology:** Low-rank tensor transition (LoRT) for transferable tensor regression with theoretical guarantees.



Error Bounds *under certain conditions*

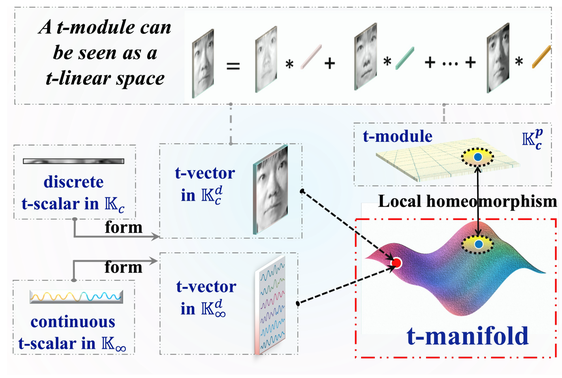
$$\|\hat{\mathcal{W}}_{\text{lort}}^{(0)} - \mathcal{W}_\star^{(0)}\|_{\mathrm{F}}^2 \lesssim \frac{rd_1d_3}{N} + \bar{h}\sqrt{\frac{d_1}{N_T}}$$

$\frac{rd_1d_3}{N}$ *statistical efficiency from multi-task learning* — *improves over target-only data* $\frac{rd_1d_3}{N_T}$

$\bar{h}\sqrt{\frac{d_1}{N_T}}$ *captures residual error due to imperfect source-target parameter alignment (model shifts)*

Towards a Geometric Understanding of Tensor Learning via the t-Product (Wang et al. NeurIPS 2025)

- **Motivation**: t-scalar-based low-rank representations fail to represent nonlinear geometry of complex tensor data, while classical differential geometry remains incompatible with t-scalar algebra.
- **Key Contribution**: Establish t-product geometry framework that defines t-manifolds and associated sheaf-theoretic structures, and fundamental learning theory on t-manifolds.



Selected Publications

- T. Li, G. Zhou, X. Zhao, Y. Qiu and Q. Zhao, “Efficient Low Rank Attention for Long-Context Inference in Large Language Models,” *NeurIPS 2025*.
- A. Wang, Y. Qiu, H. Huang, Z. Jin, G. Zhou and Q. Zhao, “Towards a Geometric Understanding of Tensor Learning via the t-Product,” *NeurIPS 2025*.
- Z. Tao, Y. Takida, N. Murata, Q. Zhao and Y. Mitsufuji, “Transformed Low-rank Adaptation via Tensor Decomposition and Its Applications to Text-to-image Models,” *ICCV 2025*.
- Y. Li, G. Zhou, Z. Huang, X. Chen, Y. Qiu, and Q. Zhao, “Tensor Decomposition Based Memory-Efficient Incremental Learning,” *ICML 2025*.
- A. Wang, Y. Qiu, Z. Jin, G. Zhou, and Q. Zhao, “Low-Rank Tensor Transitions (LoRT) for Transferable Tensor Regression,” *ICML 2025*.
- Y. Qiu, A. Wang, C. Li, H. Huang, G. Zhou and Q. Zhao, “STEPS: Sequential Probability Tensor Estimation for Text-to-Image Hard Prompt Search,” *CVPR 2025*.