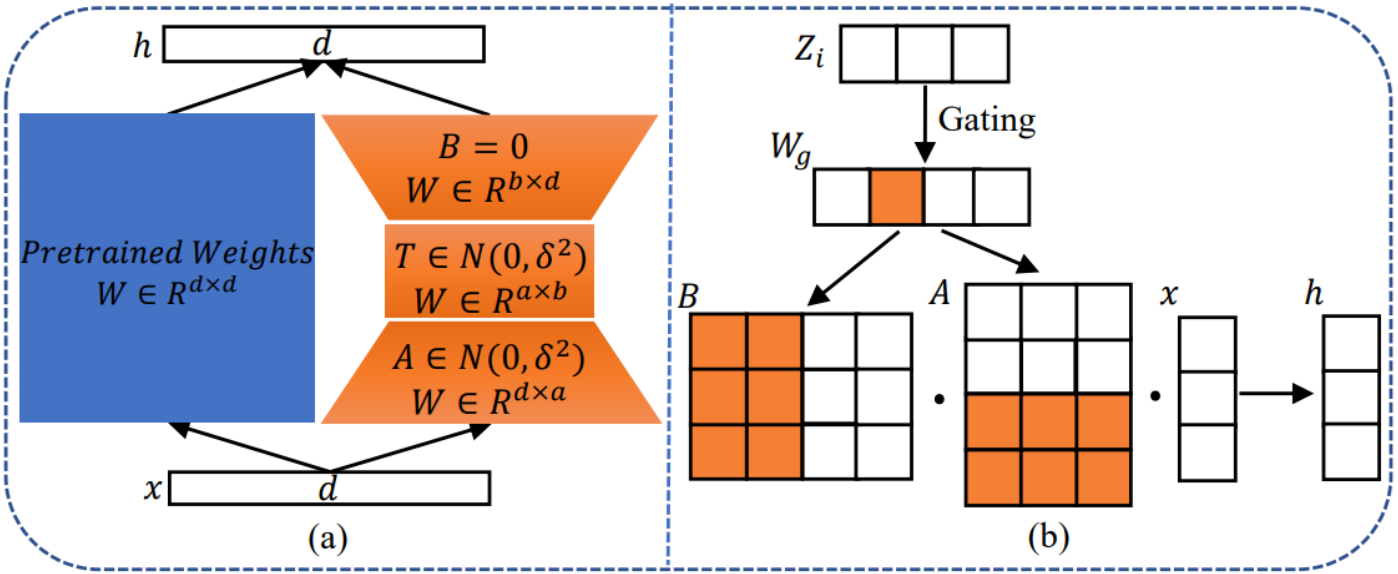
Optimizing Low-Rank Adaptation with Decomposed Matrices and Adaptive Rank Allocation

Dacao ZHANG, Fan YANG, Kun ZHANG, Xin LI, Si WEI, Richang HONG, Meng WANG

Frontiers of Computer Science, DOI: 10.1007/s11704-024-40317-w

Problems & Ideas

- Problems of Low-Rank Adaption (LoRA) for Large Language Models (LLMs):
  - It lacks a granular consideration of the relative importance and optimal rank allocation in the decomposed matrices.
  - It fails to account for the inherent varying rank requirements in multi-task scenarios.
- Ideas: exploring distinct rank settings in the decomposed matrices in single-task scenarios and designing an adaptive rank selector for multi-task learning.



Main Contributions

- Contributions:
  - A newly designed enhanced matrix decomposition strategy in single-task scenarios, which assigns different ranks to the decomposed matrices to enhance the flexibility of the fine-tuned module;
  - A novel task-specific rank allocation module for multi-task learning, which treats each rank in LoRA as an expert and uses task embeddings as the router to select suitable rank;
  - Extensive experiments over different tasks and backbone models have been done to demonstrate the effectiveness of the proposed strategies.

Table 1 The results of our methods on glue tasks. The enhanced matrix decomposition for single-task training and the task-specific rank allocation for multi-task training.

Model & Method	params/per task	MNLI	SST-2	MRPC	CoLA	QNLI	QQP	RTE	STS-B	AVG
Single-Task Training										
$RoB_{large}(LoRA_{r=8})$	3M	89.71	95.87	90.67	66.52	94.67	91.28	84.48	91.62	88.10
$RoB_{large}(Ours)$	3.01M	90.36	95.99	90.93	70.02	94.53	91.30	85.56	92.32	**88.87**
Multi-Task Training										
$RoB_{base}(LoRA_{r=8})$	0.24M	84.39	93.46	88.48	63.16	90.68	87.12	75.81	-	83.30
$RoB_{base}(LoRA_{r=16})$	0.48M	84.79	94.50	88.97	60.32	91.12	88.12	75.81	-	83.38
$RoB_{base}(Ours)$	0.48M	85.40	93.92	88.73	64.18	91.27	88.33	76.17	-	**84.00**

Table 2 The results of our decomposition method compared with Dylora [3] and Adalora [2].

Method	SST-2	MRPC	CoLA	QNLI	STS-B	AVG
DyLoRA	94.26	89.46	59.51	92.22	91.06	85.30
AdaLoRA	94.49	**90.19**	61.64	93.08	**91.16**	86.11
Ours	**94.84**	89.73	**63.31**	**93.88**	90.98	**86.55**

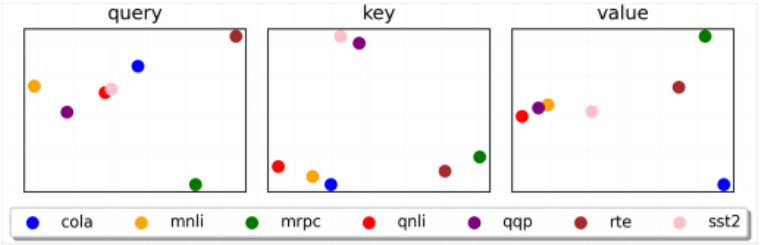


Fig. 2 Visualization of the last layer task embeddings.