SD-LoRA: Scalable Decoupled Low-Rank Adaptation for Class Incremental Learning

Yichen Wu[1,2*], **Hongming Piao**[1*], **Long-Kai Huang**[4†], **Renzhen Wang**[3], **Wanhua Li**[2], **Hanspeter Pfister**[2], **Deyu Meng**[3,6], **Kede Ma**[1†], **Ying Wei**[5†]
[1]City University of Hong Kong, [2]Harvard University, [3]Xi'an Jiaotong University, [4]Tencent AI Lab, [5]Zhejiang University, [6]Pengcheng Laboratory

- Problem / objective
  - 기존 Continual Learning 방법들의 scalability 문제
- Contribution / Key idea
  - Scalable Decoupled LoRA (SD-LoRA)
    - Incrementally adds LoRA components by separating the magnitude and direction learning.

WU, Yichen, et al. S-LoRA: Scalable Low-Rank Adaptation for Class Incremental Learning. *arXiv preprint arXiv:2501.13198*, 2025.

ICLR 2025 Oral

- **SD-LoRA: CL method with foundation models**

1. Rehearsal-free
2. Inference-efficient
3. End-to-end optimized

Table 1: Comparisons of existing CL methods with foundation models in terms of three desirable properties: 1) *Rehearsal-free* (*i.e*, without memory for sample storage), 2) *inference efficiency* (*i.e*, without additional computational overhead during inference), and 3) *end-to-end optimization* (of all model parameters for CL objectives).

Method	Rehearsal-free	Inference Efficiency	End-to-end Optimization
L2P (Wang et al., 2022b)	✓	✗	✗
DualPrompt (Wang et al., 2022a)	✓	✗	✗
CODA-Prompt (Smith et al., 2023)	✓	✗	✓
HiDe-Prompt (Wang et al., 2024a)	✗	✗	✓
InfLoRA (Liang & Li, 2024)	✗	✓	✓
SD-LoRA(Ours)	✓	✓	✓

전유진

- **two variants of SD-LoRA**

  1. Rank reduction
  2. Knowledge distillation

  ,based on the observations: "The importance of the incrementally learned LoRA directions diminishes as CL progresses"

- **Preliminaries: Class-Incremental Learning**

  - 목표: Perform well on both the current task and all previous tasks.
  - 학습: CE loss of task t. When training on $\mathcal{D}_t$, no data from previous tasks $\{\mathcal{T}_k\}_{k=1}^{t-1}$ is accessible.

$$\ell\left(\mathcal{D}_t;\boldsymbol{\theta}\right)=\frac{1}{|\mathcal{D}_t|}\sum_{i=1}^{|\mathcal{D}_t|}\ell\left(f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_t^{(i)}\right),y_t^{(i)}\right), \quad (1)$$

  - 평가: Average loss across all tasks encountered so far.

$$\frac{1}{t}\sum_{k=1}^{t}\ell(\underbrace{\mathcal{V}_k}_{\text{test split of } \mathcal{T}_k};\boldsymbol{\theta})$$

  - Notation:
    a. N sequential classification tasks: $\{\mathcal{T}_1,\mathcal{T}_2,\ldots,\mathcal{T}_N\}$
    b. Training split of $\mathcal{T}_t$: $\mathcal{D}_t=\{\boldsymbol{x}_t^{(i)},y_t^{(i)}\}_{i=1}^{|\mathcal{D}_t|}$ (: image, label)

- **Preliminaries: LoRA (Low-Rank Adaptation)**

  - 정의: Constrain the parameter updates during fine-tuning to lie in a low-rank subspace.
  - For a given layer of f_{θ}, the LoRA-updated output :

$$h' = \mathbf{W}_0 x + \Delta\mathbf{W} x = (\mathbf{W}_0 + \mathbf{AB})x. \quad (2)$$

parameter update expressed as the product of two learnable matrices.

original weight matrix of a layer in the classifier f_{θ}

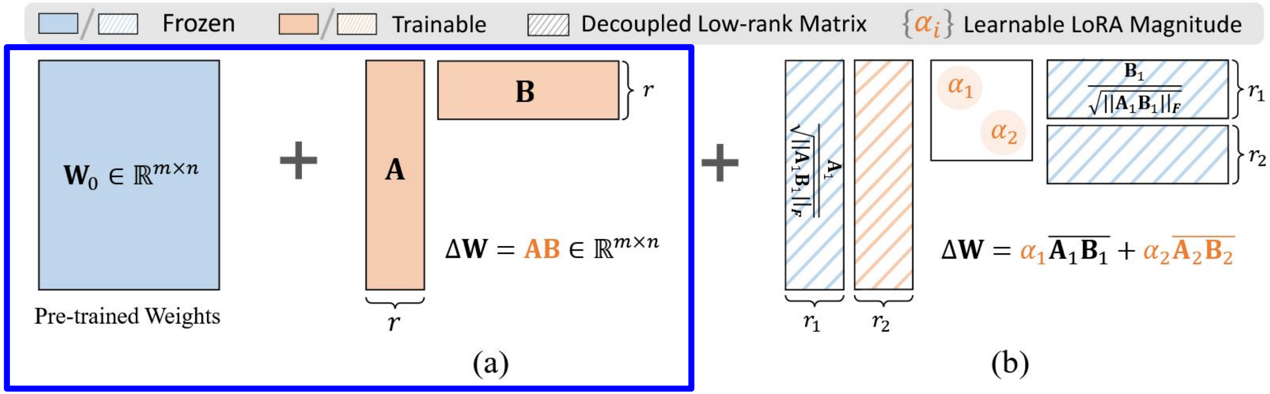

Figure 1: Illustration of the parameter update in **(a)** Vanilla LoRA and **(b)** the proposed SD-LoRA, where the current task index is $t = 2$ and $r, r_1, r_2 \ll \min\{m, n\}$.

WU, Yichen, et al. S-LoRA: Scalable Low-Rank Adaptation for Class Incremental Learning. *arXiv preprint arXiv:2501.13198*, 2025.

ICLR 2025 Oral

- **SD-LoRA**

aa

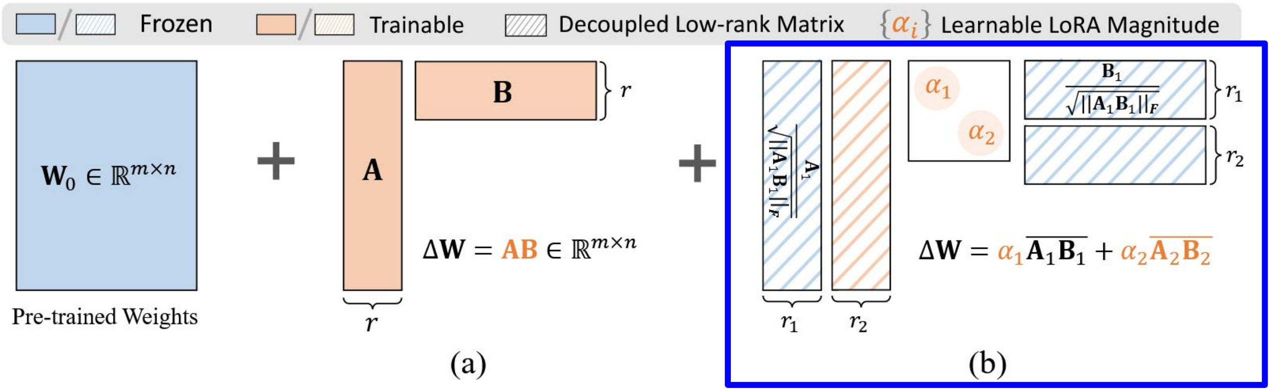

Figure 1: Illustration of the parameter update in **(a)** Vanilla LoRA and **(b)** the proposed SD-LoRA, where the current task index is $t = 2$ and $r, r_1, r_2 \ll \min\{m, n\}$.

- **How SD-LoRA mitigates catastrophic forgetting?**

aa