Closed-Form Merging of Parameter-Efficient Modules for Federated Continual Learning

Riccardo Salami, Pietro Buzzega, Matteo Mosconi,
Jacopo Bonato, Luigi Sabetta, Simone Calderara

AImageLab
University of Modena and Reggio Emilia, Modena, Italy

ICLR 2025

Code: `https://github.com/aimagelab/mammoth`

Federated Learning

- Distributed training
- Local data privacy
- Global model aggregation

$$\min_w F(w) = \sum_{k=1}^{K} p_k F_k(w), \quad F_k(w) = \mathbb{E}_{x_k \sim \mathcal{D}_k}\left[f_k(w; x_k)\right]$$

FedAvg

- Privacy-preserving
- Communication-efficient
- Scalable

$$w^{t+1} = \sum_{k=1}^{K} p_k w_k^{t+1}, \quad \text{where} \quad w_k^{t+1} = w^t - \eta \nabla F_k(w^t)$$

Continual Learning

- Incrementally learns new tasks while retaining previous knowledge.
- Addresses the issue of **catastrophic forgetting**.
- Enables ongoing adaptation to new and changing tasks.

Role of Pretrained Models

- Serve as a **robust starting point**, reducing training time for new tasks.
- Enhance generalization by using previously learned representations.
- Facilitate rapid and effective adaptation through **parameter-efficient** techniques.

Parameter-efficient Fine-tuning

- Minimizes the number of parameters modified during fine-tuning.
- Reduces memory requirements.
- Enables **fast adaptation** of large pre-trained models to new tasks.

LoRA (Low-Rank Adaptation)

- Optimizes **residual low-rank** matrices on top of pretrained weights to improve fine-tuning efficiency:

$$W' = W + BA \quad \text{where} \quad B \in \mathbb{R}^{d \times r}, A \in \mathbb{R}^{r \times k}$$

- Significantly reduces the number of trainable parameters and communication costs
- Maintains performance comparable to full fine-tuning methods.

- Distills knowledge between models **without data**:

$$\Omega = \sum_{i=1}^{N} \|(W_0 + B_M A_M)X_i - (W_0 + B_i A_i)X_i\|_2^2.$$

- **Alignment** of linear layers through **closed-form** solution [1]
- Useful with transformers, mainly composed by linear layers
- As LoRA optimizes A and B simultaneously, setting both derivatives to zero results in an **inteterminate system** when solving for the two variables.

- Closed-form computation for optimal layer-wise distillation

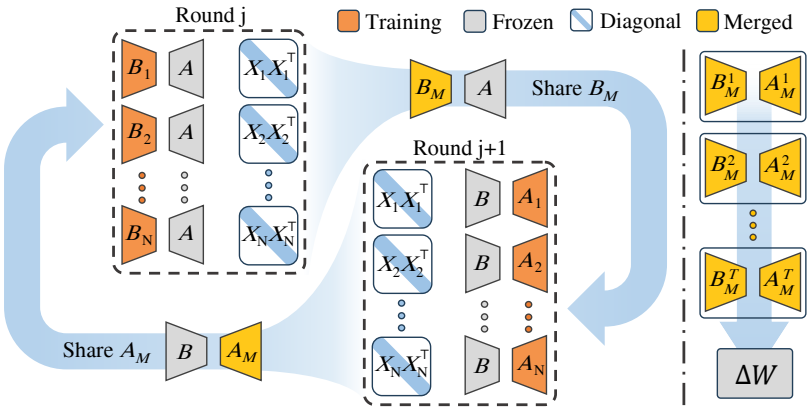
$$A_M = \left(\sum_{i=1}^{N} A_i X_i X_i^\top \right) \left(\sum_{i=1}^{N} X_i X_i^\top \right)^{-1}$$

$$B_M = \left(\sum_{i=1}^{N} B_i A X_i X_i^\top \right) A^\top \left(A \sum_{i=1}^{N} X_i X_i^\top A^\top \right)^{-1}$$

- Alternating optimization of A and B LoRA matrices
- Efficient training and low communication costs
- Additional theoretical results for other PEFT techniques in the paper.

Algorithm Overview

1. Train LoRA modules locally
2. Compute closed-form optimal merge
3. Aggregate modules globally
4. Update global model and redistribute it



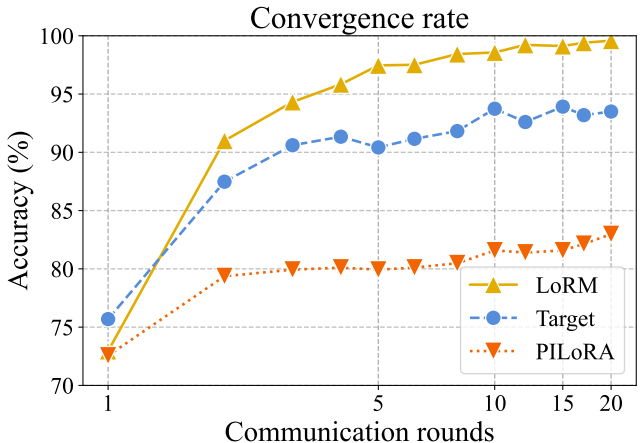
Main Results: In-distribution (Table 1)

- Significant accuracy improvement
- Great performance both in and out-of-domain
- See the paper for more results (also text)

Method	CIFAR-100	ImageNet-R	ImageNet-A	EuroSAT	Cars-196	CUB-200
Joint	**92.75**	**84.02**	**54.64**	**98.42**	**85.62**	**86.04**
EWC	78.46	58.93	10.86	64.12	19.55	31.46
LwF	62.87	54.03	8.89	31.91	20.84	25.25
FisherAVG	76.10	58.68	11.59	58.84	26.03	30.45
RegMean	59.80	61.18	8.56	48.74	21.83	35.57
CCVR	79.95	70.00	**39.50**	64.44	38.99	62.67
L2P	83.88	42.08	20.14	40.63	35.49	56.23
CODA-P	82.25	61.18	18.30	73.38	28.04	42.53
FedProto	75.79	58.52	9.87	58.79	26.08	30.22
TARGET	74.72	54.65	10.27	52.74	28.65	39.30
PILoRA	76.48	53.67	19.62	48.35	37.57	61.11
LoRM (ours)	**86.95**	**72.48**	37.26	**84.23**	**54.41**	**64.60**

Convergence Speed Ablation

- Faster convergence with closed-form
- Stability across learning phases
- Consistent performance gap



Thank you for your attention!

[1] Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng.
Dataless knowledge fusion by merging weights of language models.
In *International Conference on Learning Representations*, 2023.