NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2024 Spotlight:

EMR-Merging:

Tuning-Free High-Performance Model Merging

Chenyu Huang[1†], Peng Ye[1,3†], Tao Chen[1*], Tong He[2], Xiangyu Yue[3], Wanli Ouyang[3]

[1] Fudan University; [2] Shanghai AI Laboratory;

[3] The Chinese University of Hong Kong

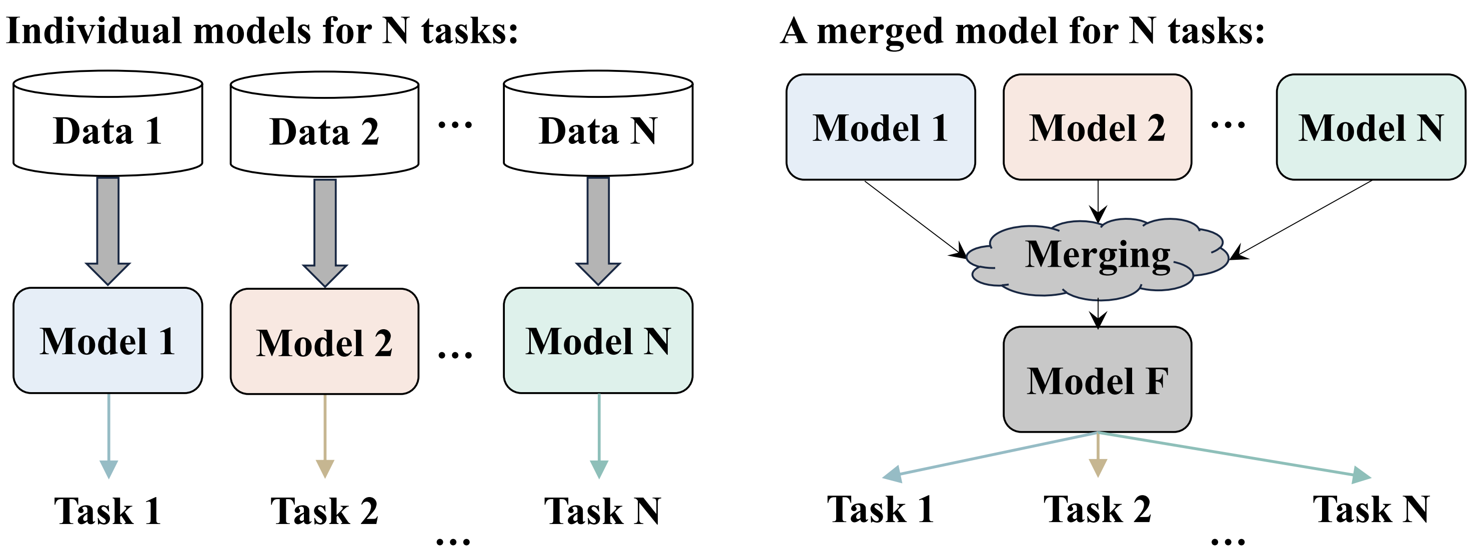
[†] Equal Contribution; [*] Corresponding Author: eetchen@fudan.edu.cn

arXiv GitHub

1. Background

Model Merging:

Combining weights instead of additional training to render a model multi-task capabilities.

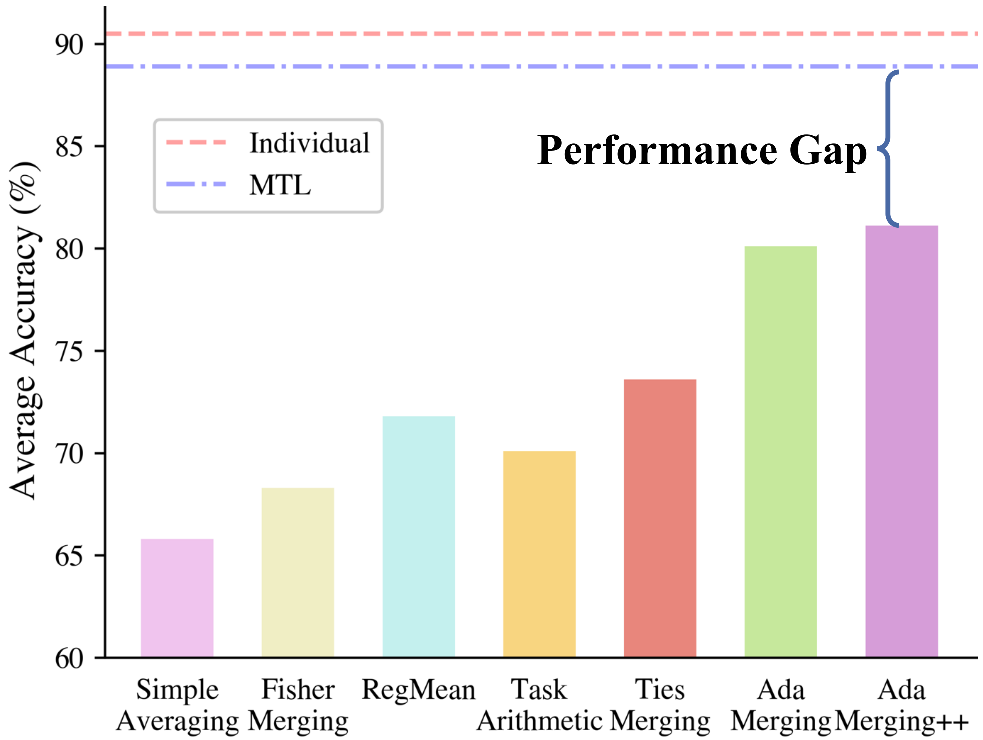


- Reducing **storage** and **deployment** cost.
- No additional training or training data.

1. Background

Current Problem 1:

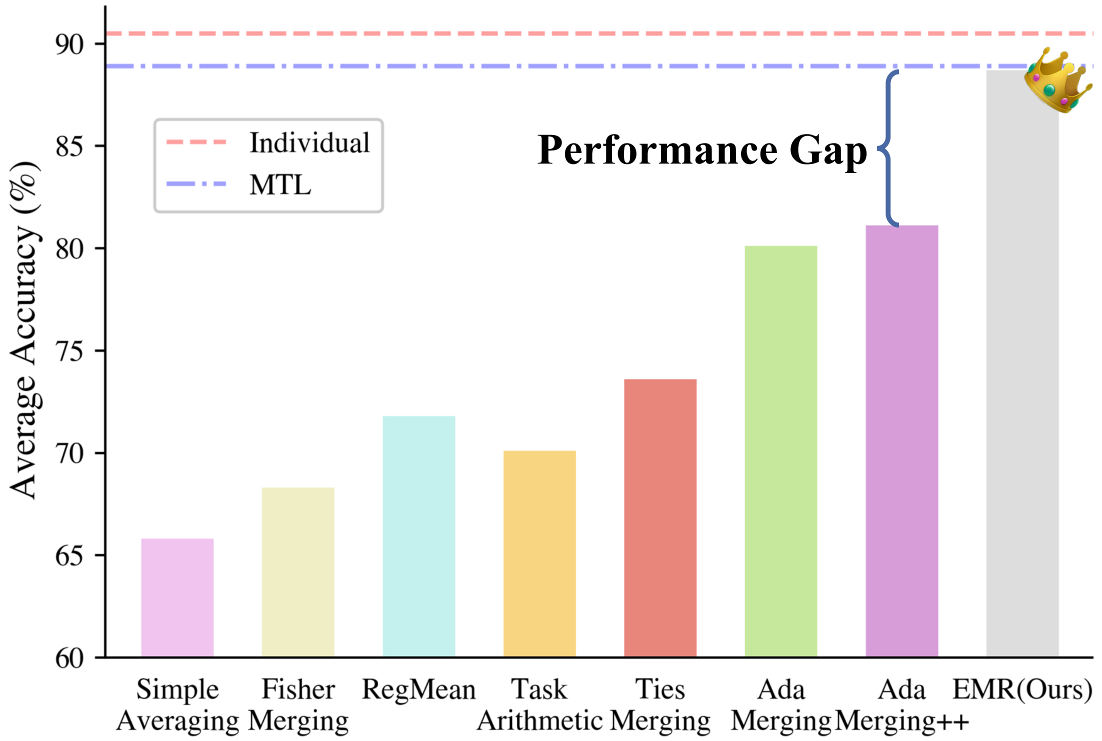
- Significant performance degradation



1. Background

Current Problem 1:

- Significant performance degradation



1. Background

Current Problem 2:

- Additional Tuning

Others:

Tuning by training data
- Multi-Task Learning

Tuning by validation data
- Task Arithmetic
- Ties-Merging
- RegMean
- Fisher-Merging

Hyper-parameter tuning
- Task Arithmetic
- Ties-Merging
- DARE

Tuning by additional training
- Multi-Task Learning
- AdaMerging

1. Background

Current Problem 2:

- Additional Tuning

Others:

Tuning by training data
- Multi-Task Learning

Tuning by validation data
- Task Arithmetic
- Ties-Merging
- RegMean
- Fisher-Merging

Hyper-parameter tuning
- Task Arithmetic
- Ties-Merging
- DARE

Tuning by additional training
- Multi-Task Learning
- AdaMerging

Ours:

Tuning-Free
- ~~Training Data~~
- ~~Test Data~~
- ~~Training or finetuning~~
- ~~Hyper-Parameter~~

2. Motivation

- **Finding #1**: A single model weight is hard to simulate all the models' performance due to **weight interference.**
- **Finding #2**: Focusing on model weights instead of data may **avoid additional tuning.**

Modifying the merging paradigm:

- Decoupling model merging into 1) **a unified model** and several 2) **task-specific modules.**

Light-weight
task-specific modules

- Original:

$$W_M = \mathcal{M}\left([W_1..W_N]\right)$$

× Severe weight interference
× Additional tuning needed

- New Paradigm:

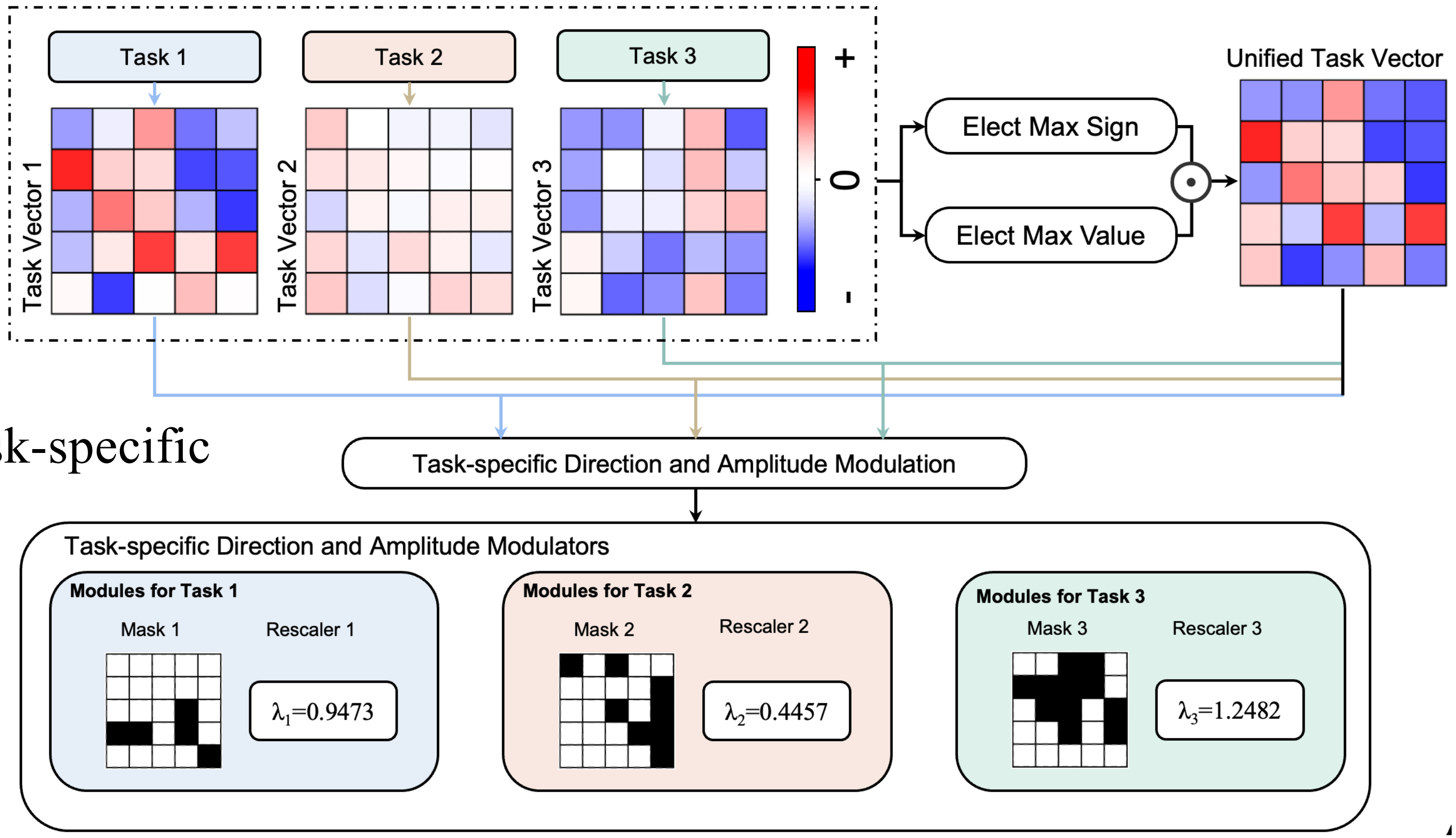
$$W_{uni}, [E_1..E_N] = \mathcal{M}'\left([W_1..W_N]\right)$$

✓ Resolving weight interference
✓ Needs no tuning

3. Method: EMR-Merging

Merging Procedure

- Elect the unified task vector

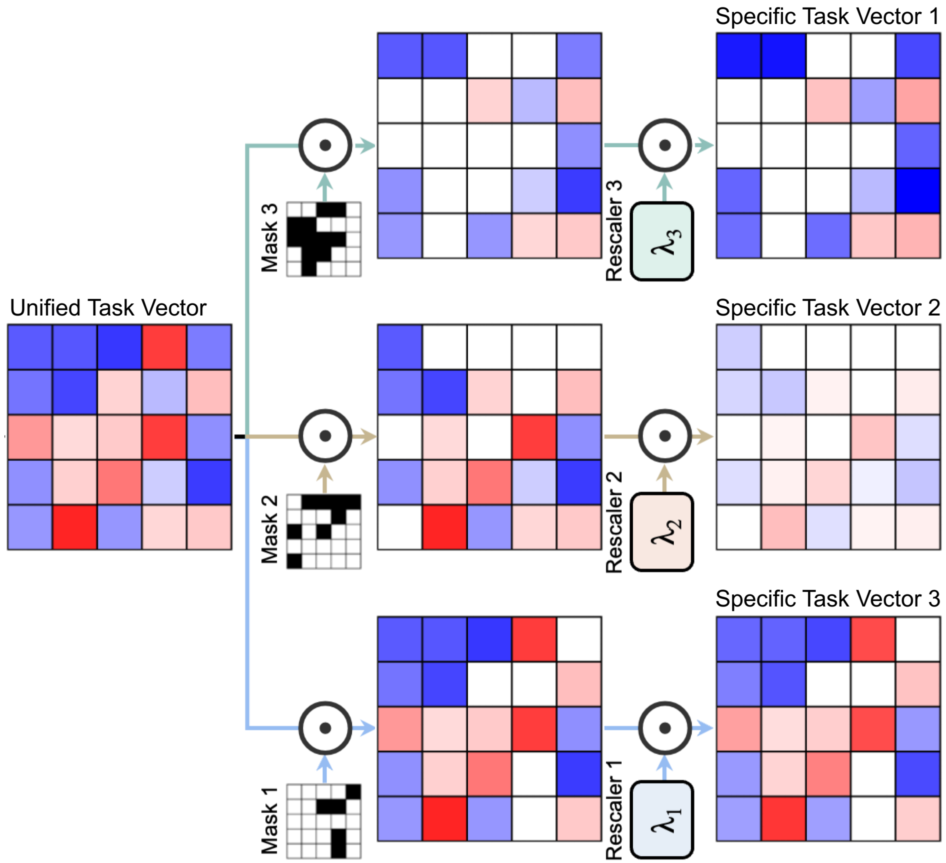


- Generate task-specific modulators

3. Method: EMR-Merging

Inference Procedure

- Apply the specific modulators before inference on a task



Algorithm Flow

Algorithm 1 EMR-MERGING Procedure

Input: Finetuned models $W_{1..N}$, pretrained model W_{pre}
Output: Unified task vector τ_{uni}, task-specific masks $M_{1..N}$, task-specific rescalers $\lambda_{1..N}$

for t ***in*** $1, ..., N$ **do**
  ▷ Create task vectors.
  $\tau_t = W_t - W_{pre}$
end
▷ Step 1: Elect the unified task vector.
$\gamma_{uni} = sgn(\sum_{t=1}^{n} \tau_t)$
$\epsilon_{uni} = zeros(d)$
for t ***in*** $1, ..., N$ **do**
  for p ***in*** $1, ..., d$ **do**
    if $\gamma_{uni}^p \cdot \tau_t^p > 0$ **then**
      $\epsilon_{uni}^p = max\,(\epsilon_{uni}^p, abs\,(\gamma_{uni}^p))$
    end
  end
end
$\tau_{uni} = \gamma_{uni} \odot \epsilon_{uni}$.
for t ***in*** $1, ..., N$ **do**
  ▷ Step 2: Generate task-specific masks.
  for p ***in*** $1, ..., d$ **do**
    $M_t^p = bool(\tau_t^p \cdot \tau_{uni}^p > 0)$
  end
  ▷ Step 3: Generate task-specific rescalers.
  $\lambda_t = \frac{sum(abs(\tau_t))}{sum(abs(M_t \odot \tau_{uni}))}$
end

3. Method: EMR-Merging

Charastics:

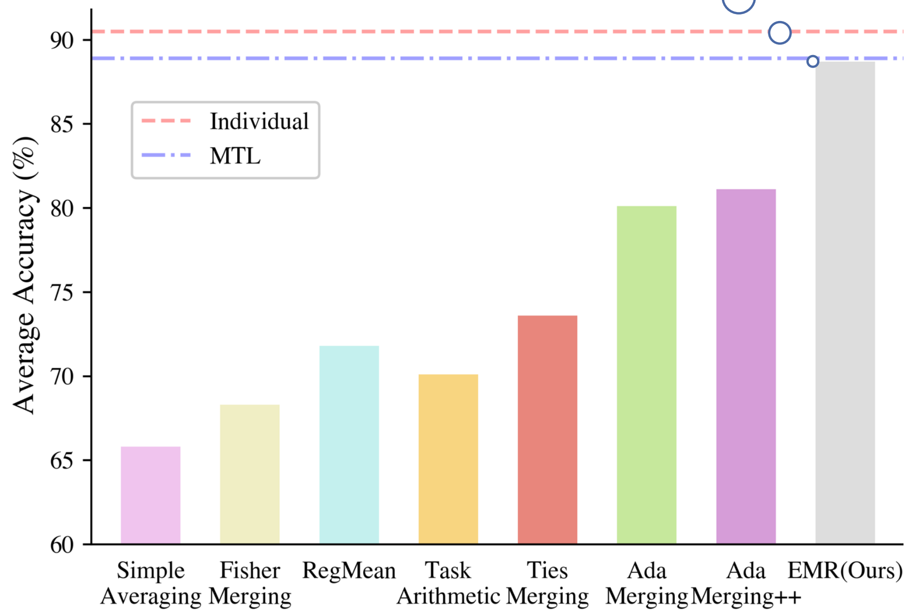
- ✓ **Tuning-Free**
- ✓ High **applicability**
- ✓ Good **performance**.

vision, language, PEFT, and multi-modal models

Methods	Training-Data Tuning	Valid-Data Tuning inputs	Valid-Data Tuning labels	Tuning by Training
Weight Averaging	×	×	×	×
Traditional MTL	✓	×	×	✓
Fisher-Merging [46]	×	✓	×	×
RegMean [33]	×	✓	×	×
Task Arithmetic [30]	×	✓	✓	×
Ties-Merging [84]	×	✓	✓	×
AdaMerging [85]	×	✓	×	✓
EMR-Merging(Ours)	×	×	×	×

Lowest Cost

Close to individual models' performance



4. Experimental Results

Merging Vision Models:

- **Results on merging eight ViT-B/32 models**

Methods	Averaging	Fisher	RegMean	Task Arithmetic	Ties-Merging	AdaMerging	EMR-Merging(Ours)
Performance	65.8	68.3	71.8	70.1	73.6	81.1	**88.7 (↑7.6)**

- **Results on merging eight ViT-L/14 models**

Methods	Averaging	Fisher	RegMean	Task Arithmetic	Ties-Merging	AdaMerging	EMR-Merging(Ours)
Performance	79.6	82.2	83.7	84.5	86.0	91.0	**93.7 (↑2.7)**

- **Results on merging 30 ViT-B/16 models**

Methods	Averaging	RegMean	Task Arithmetic	Ties-Merging	AdaMerging	EMR-Merging(Ours)
Performance	42.5	68.1	48.9	37.5	60.3	**89.5 (↑21.4)**

Reproduce our experiments: https://github.com/harveyhuang18/EMR_Merging

4. Experimental Results

Merging Language Models:

➢ Results on merging eight RoBERTa models

Methods	Averaging	RegMean	Task Arithmetic	Ties-Merging	EMR-Merging(Ours)
Performance	51.3	70.0	66.7	64.0	**80.2 (↑10.2)**

➢ Results on merging seven GPT-2 models

Methods	Averaging	Fisher	RegMean	Task Arithmetic	Ties-Merging	EMR-Merging(Ours)
Performance	56.1	58.7	68.8	70.0	70.0	**80.4 (↑10.4)**

➢ Results on merging eleven PEFT $(IA)^3$ models

Methods	Averaging	Fisher	RegMean	Task Arithmetic	Ties-Merging	EMR-Merging(Ours)
Performance	58.0	62.2	58	63.9	66.4	**67.1 (↑0.7)**

Reproduce our experiments: https://github.com/harveyhuang18/EMR_Merging

4. Experimental Results

Merging Vision-Language Models:

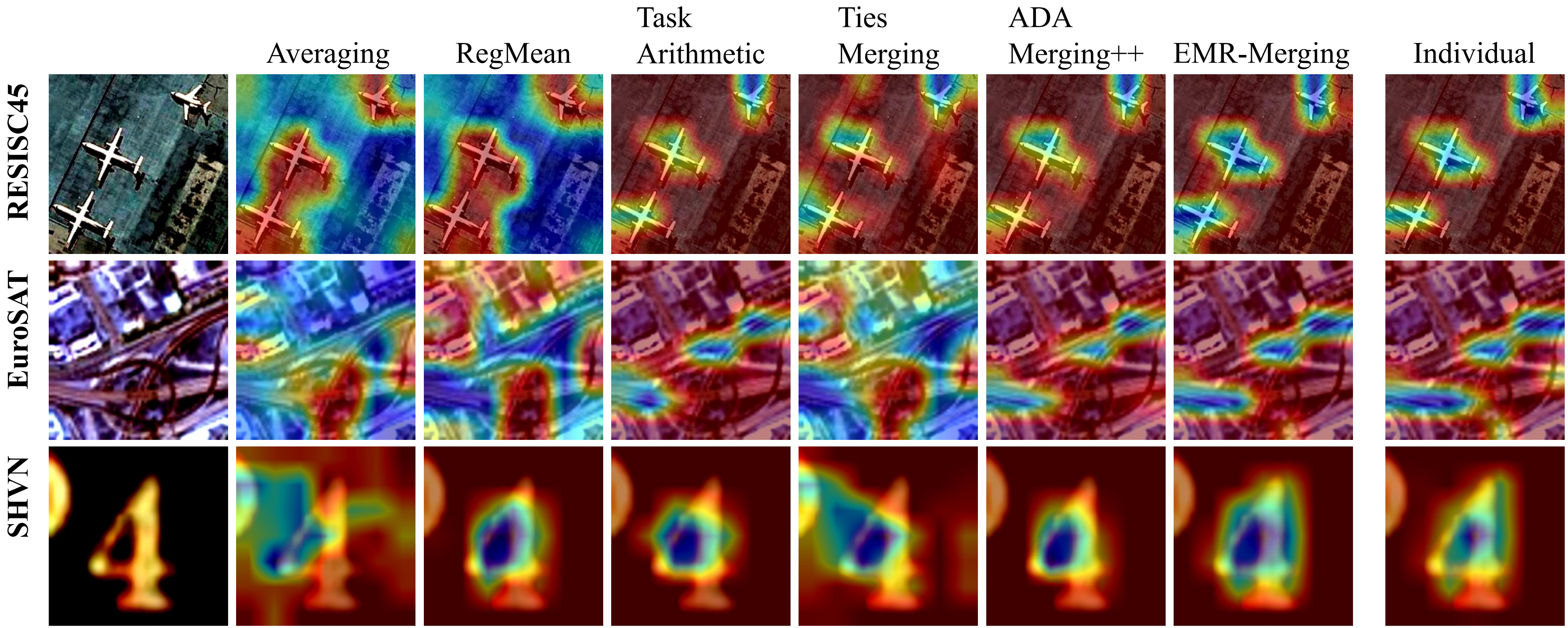
- **Results on merging five BEiT-3 models**

Methods	Task Metric	**COCO-Retrieval** Accuracy(↑)	**COCO-Captioning** BLEU4(↑)	CIDEr(↑)	METEOR(↑)	ROUGE-L(↑)	**ImageNet-1k Classification** Accuracy(↑)	**NLVR2** Accuracy(↑)	**VQAv2** Accuracy(↑)
Individual		0.8456	0.394	1.337	0.311	0.601	0.8537	0.7765	0.8439
Weight Averaging		0.1893	0.031	0.001	0.115	0.159	0.6771	0.2800	0.6285
Task Arithmetic [30]		0.3177	0.033	0.000	0.118	0.176	0.7081	0.3809	0.6933
Ties-Merging [84]		0.3929	0.029	0.001	0.108	0.167	0.6978	0.3206	0.6717
EMR-MERGING(Ours)		**0.7946**	**0.289**	**1.060**	**0.272**	**0.534**	**0.7742**	**0.7475**	**0.7211**

Reproduce our experiments: https://github.com/harveyhuang18/EMR_Merging

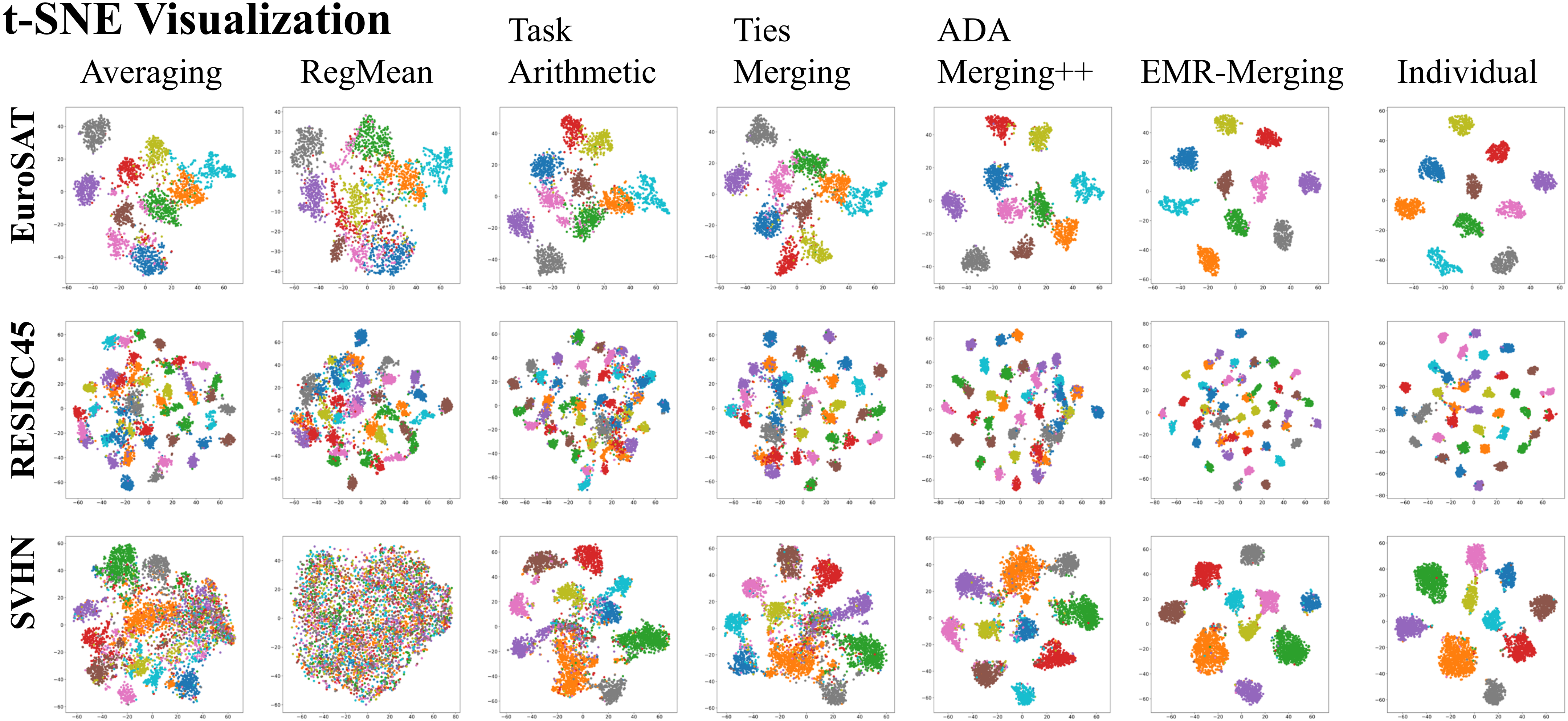
5. Visualization Results

Grad-CAM Visualization



5. Visualization Results

t-SNE Visualization



NEURAL INFORMATION
PROCESSING SYSTEMS

FUDAN UNIVERSITY 1905

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

Thanks!

E-mail: **cyhuang24@m.fudan.edu.cn**

arXiv GitHub EDLab