Null-Space Filtering for Data-Free Continual Model Merging: Preserving Stability, Promoting Plasticity

Zihuan Qiu[1,3], Lei Wang[1], Yang Cao[3], Runtong Zhang[1], Bing Su[3], Yi Xu[2], Fanman Meng[1], Linfeng Xu[1], Qingbo Wu[1], Hongliang Li[1]

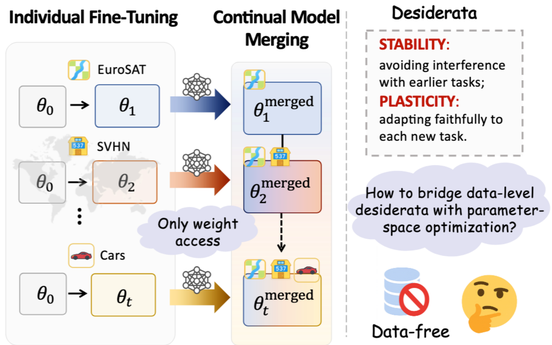
[1] University of Electronic Science and Technology of China, [2] Dalian University of Technology, [3] Jiigan Technology

Paper Code ICLR

Background and Motivation

Data-free continual model merging (DFCMM)

- A sequence of task-specific models are merged into a single backbone
- Tasks arrive sequentially and must be integrated without retraining
- Only the current model and merged model are accessible at each step

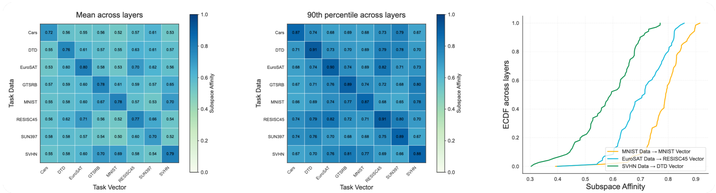


Key challenges

- Stability–plasticity trade-off : Preserve previous knowledge while adapting to new tasks
- Parameter interference: Task updates are entangled and may conflict in parameter space
- Fundamental gap: Data-level objectives vs parameter-space operations

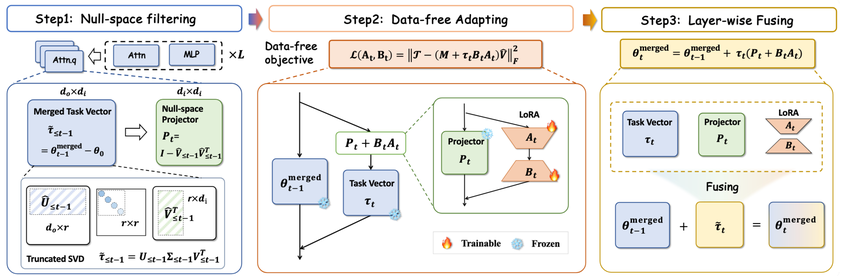
Key Insight: Task vectors align with representation subspaces

Implication: Parameter updates encode data-level structure



NUFILT: Null-Space Filtering

Data-free framework for continual model merging



Step 1: Filtering (Null-space projector preserves prior responses)

Project new task vector onto null-space of previous directions:

$$P_t^{(l)} = I - \hat{V}_{\le t-1}^{(l)} \hat{V}_{\le t-1}^{(l)\top} \quad \text{where } \hat{V}_{\le t-1}^{(l)} \text{ is the top-}r_p \text{ right singular vectors of merged vector}$$

Property: for $x^{(l)} \in span\left(\hat{V}_{\le t-1}^{(l)}\right)$, $P_t^{(l)} x^{(l)} = 0$

Step 2: Adapting (LoRA adapter enables plasticity)

$$\mathcal{L}_{\text{stab}} = \|(\tilde{\tau}_{\le t} - \tilde{\tau}_{\le t-1}) X_o^\top\|_F^2 \le 2\,\sigma_1(X_o)^2 \left(\|(\tilde{\tau}_{\le t} - \tilde{\tau}_{\le t-1}) \hat{V}_{\le t-1}\|_F^2 + r_o \zeta_o^2 \|(\tilde{\tau}_{\le t} - \tilde{\tau}_{\le t-1})\|_2^2 \right)$$

$$\mathcal{L}_{\text{plas}} = \|(\tilde{\tau}_{\le t} - \tau_t) X_n^\top\|_F^2 \le 2\,\sigma_1(X_n)^2 \left(\|(\tilde{\tau}_{\le t} - \tau_t) \hat{V}_t\|_F^2 + r_n \zeta_n^2 \|(\tilde{\tau}_{\le t} - \tau_t)\|_2^2 \right)$$

Insert low-rank LoRA adapter:

$$P_t^{(l)} \to P_t^{(l)} + B_t^{(l)} A_t^{(l)}$$

Data-Free Objective:

$$\mathcal{L} = \left\| \mathcal{T} - (M + \tau_t^{(l)} B_t^{(l)} A_t^{(l)}) \hat{V} \right\|_F^2$$

$\mathcal{T}$: target behavior (preserve old tasks + match new task)

$\tau_t B_t A_t$: adaptive update (plasticity) → learns new task

M: filtered update (stability) → keeps previous knowledge

$$\mathcal{T} = \begin{bmatrix} \tilde{\tau}_{\le t-1}^{(l)} \hat{V}_{\le t-1}^{(l)} \\ \tau_t^{(l)} \hat{V}_t^{(l)} \end{bmatrix}, \quad \hat{V} = \begin{bmatrix} \hat{V}_{\le t-1}^{(l)} \\ \hat{V}_t^{(l)} \end{bmatrix}, \quad M = \tilde{\tau}_{\le t-1}^{(l)} + \tau_t^{(l)} P_t^{(l)}.$$

Step 3: Fusing (Absorb extra params)

$$\theta_t^{\text{merged},(l)} = \theta_{t-1}^{\text{merged},(l)} + \tau_t^{(l)} \left(P_t^{(l)} + B_t^{(l)} A_t^{(l)} \right).$$

Absorbed into weights, with no extra parameters or inference cost

Results

Effectiveness of each components

Method	Component	ACC(%) ↑			BWT(%) ↑		
	Null-space / LoRA	8 tasks	14 tasks	20 tasks	8 tasks	14 tasks	20 tasks
(1) Naive Merging ($\theta_t^{\text{merged}} = \theta_{t-1}^{\text{merged}} + \tau_t$)	✗ / ✗	62.1±0.0	46.5±0.0	34.3±0.0	-18.5±6.2	-25.8±2.2	-24.7±5.1
(2) Only Null-space ($\theta_t^{\text{merged}} = \theta_{t-1}^{\text{merged}} + \tau_t P_t$)	✓ / ✗	80.0±1.1	76.7±1.0	67.0±0.7	-1.7±0.9	-4.2±0.9	-6.2±1.8
(3) Data-free Objective							
w/o null-space filter ($\theta_t^{\text{merged}} = \theta_{t-1}^{\text{merged}} + \tau_t B_t A_t$)	✗ / ✓	75.8±1.9	63.7±1.6	51.7±1.0	-10.2±4.3	-17.1±2.7	-20.6±4.9
full method ($\theta_t^{\text{merged}} = \theta_{t-1}^{\text{merged}} + \tau_t (P_t + B_t A_t)$)	✓ / ✓	83.6±0.2	78.0±0.2	71.0±0.9	-2.7±0.7	-5.7±0.9	-8.9±2.3

State-of-the-art Results

Table 1: Comparative results of continual merging methods, reporting average accuracy and backward transfer over ten task orders (mean±std). EP and DA denote method assumptions: the need for extra parameters or data access. Best results are in **bold**, and the second best are underlined.

	Method	Requirement	ViT-B/32			ViT-B/16			ViT-L/14		
		EP / DA	8 tasks	14 tasks	20 tasks	8 tasks	14 tasks	20 tasks	8 tasks	14 tasks	20 tasks
	Pre-Trained	– / –	48.1	56.9	55.6	55.4	62.0	59.8	64.9	69.1	65.6
	Individual	– / –	90.4	89.3	89.8	92.4	91.3	91.6	94.3	93.4	93.5
	C. Fine-Tuned	– / –	79.8	67.4	62.6	82.9	72.2	68.2	90.0	70.9	77.7
ACC (%) ↑	Weight Averaging	✗ / ✗	66.3±0.0	65.4±0.0	61.1±0.0	72.3±0.0	69.7±0.0	64.8±0.0	80.0±0.0	77.5±0.0	71.1±0.0
	Task Arithmetic	✗ / ✗	67.5±0.0	66.5±0.0	60.0±0.0	77.1±0.0	70.9±0.6	64.2±0.0	82.1±0.0	77.9±0.0	70.3±0.0
	Ties-Merging	✗ / ✗	49.0±0.2	66.2±0.6	59.9±0.7	66.8±3.7	70.5±0.8	63.0±1.6	64.3±7.0	78.0±0.6	68.3±0.9
	MagMax-Ind	✗ / ✗	70.7±0.0	67.0±0.0	61.2±0.0	76.7±1.8	67.0±0.0	62.5±0.0	83.4±0.0	71.2±0.0	71.2±0.0
	LW AdaMerging	✗ / ✓	53.4±3.2	59.8±1.6	59.7±7.4	59.9±2.3	64.3±1.2	61.5±1.1	68.8±2.9	73.1±5.7	66.9±1.1
	LoRA-WEMOE	✓ / ✓	68.8±7.8	63.8±3.4	49.6±15.4	72.6±3.7	67.9±2.9	55.0±7.0	75.6±7.8	74.0±5.0	56.9±19.8
	WUDI-Merging	✗ / ✗	74.7±6.6	67.0±6.9	63.7±3.8	81.0±4.7	75.0±4.1	69.6±4.7	87.5±3.3	84.2±3.7	78.1±2.8
	Iso-C	✗ / ✗	71.7±1.2	73.2±1.8	67.6±0.8	78.5±1.2	79.7±1.3	73.0±1.1	86.9±0.5	86.9±1.8	80.9±0.8
	KnOTS-TIES	✗ / ✗	54.4±6.9	67.8±0.4	60.5±1.4	57.9±8.4	71.6±0.3	62.7±1.1	68.3±5.7	78.8±0.3	69.7±0.8
	TSV-M	✗ / ✗	68.2±4.8	63.3±4.8	58.8±3.3	75.4±4.0	69.2±3.1	63.1±1.3	82.2±3.6	78.1±3.6	70.5±1.2
	OPCM	✗ / ✗	75.5±0.5	71.9±0.3	65.7±0.2	81.8±0.3	77.1±0.5	70.3±0.2	87.0±0.4	83.5±0.2	76.0±0.2
	NUFILT (Ours)	✗ / ✗	**83.6**±0.2	**78.0**±0.2	**71.0**±0.9	**87.3**±0.1	**83.1**±0.3	**78.1**±0.9	**91.6**±0.1	**89.2**±0.1	**84.7**±0.8
BWT (%) ↑	Weight Averaging	✗ / ✗	-11.5±2.2	-8.0±1.3	-7.1±2.1	-9.7±1.5	-7.1±1.4	-7.3±1.7	-7.3±1.4	-5.8±1.0	-6.4±1.5
	Task Arithmetic	✗ / ✗	-9.6±1.5	-1.3±1.6	-3.4±1.0	-4.2±1.0	-1.3±0.4	-3.6±0.4	-7.1±0.8	-1.8±0.3	-3.3±0.3
	Ties-Merging	✗ / ✗	-15.3±8.0	**1.9**±0.6	-1.5±0.7	-5.5±0.4	**1.4**±0.7	**-1.5**±1.2	-13.0±5.7	-1.1±0.4	-2.9±1.0
	MagMax-Ind	✗ / ✗	-8.3±1.3	-7.4±1.4	-7.2±1.6	-6.1±1.3	-7.4±2.0	-8.0±2.2	-5.0±0.8	-6.0±2.1	-6.5±2.1
	LW AdaMerging	✗ / ✓	-32.5±3.6	-24.1±1.7	-22.7±4.3	-27.8±2.7	-22.1±1.4	-21.4±1.2	-24.3±3.3	-19.6±1.7	-21.7±1.1
	LoRA-WEMOE	✓ / ✓	-20.4±9.0	-20.2±3.9	-24.5±10.0	-18.0±6.2	-18.8±3.4	-25.8±7.9	-17.8±5.9	-16.8±3.3	-27.9±17.2
	WUDI-Merging	✗ / ✗	-17.0±7.5	-22.8±7.3	-26.0±4.1	-12.6±5.4	-16.9±4.4	-18.5±14.2	-7.3±3.7	-9.4±4.0	-15.8±2.9
	Iso-C	✗ / ✗	-10.2±1.2	-10.4±1.9	-10.3±1.4	-6.7±0.5	-7.1±1.1	-9.9±1.7	-3.7±0.6	-3.7±1.7	-5.5±1.3
	KnOTS-TIES	✗ / ✗	-12.6±3.9	**1.9**±0.5	**-1.3**±0.7	-13.5±5.5	1.0±0.1	-2.3±0.6	-11.6±3.6	**0.3**±0.3	**-2.3**±0.7
	TSV-M	✗ / ✗	-24.0±5.6	-26.7±4.9	-31.7±3.4	-18.5±4.6	-22.7±3.1	-28.2±1.8	-13.0±4.0	-15.6±3.9	-23.3±1.3
	OPCM	✗ / ✗	-6.3±1.1	-6.0±1.0	-7.8±1.5	-4.8±0.7	-5.1±1.4	-6.3±2.2	-2.6±1.0	-4.3±0.7	-6.5±1.8
	NUFILT (Ours)	✗ / ✗	**-2.7**±0.7	-5.7±0.9	-8.9±2.3	**-1.6**±0.5	-3.5±0.6	-7.1±1.9	**-1.1**±0.3	-2.0±0.3	-4.6±0.7

Conclusion:

1. Established subspace alignment property of task vectors
2. NUFILT = null-space filtering + projection-aware adaptation
3. SOTA performance, minimal forgetting, no extra data/params