DisTaC: Conditioning Task Vectors via Distillation for Robust Model Merging

Kotaro Yoshida[1] Yuji Naraki[2] Takafumi Horie[3]
Ryotaro Shimizu[4] Ioannis Mitliagkas[5,6] Hiroki Naganuma[5,6]

[1]Institute of Science Tokyo [2]Independent Researcher [3]Kyoto University
[4]ZOZO Research [5]Mila [6]Université de Montréal

ICLR 2026 Poster

Outline

1. Background: Model Merging for Multi-Task Learning
2. Failure Modes in Model Merging
3. Proposed Method: DisTaC
4. Experiments & Results
5. Discussion & Guidelines
6. Conclusion

Background: Model Merging for Multi-Task Learning

Goal: Combine multiple fine-tuned models into a single multi-task model *without retraining*.

Task Vector (Ilharco et al., 2023):

$$\tau_t = \theta_t - \theta_{\text{pre}}$$

Merged model:

$$\theta_{\text{mtl}} = \theta_{\text{pre}} + \mathcal{M}(\tau_1, \tau_2, \ldots \tau_T)$$

- **Advantages over conventional MTL:**
  - No need to aggregate all task-specific labeled data
  - Easy to add/remove skills after deployment
- **Existing methods:** Task arithmetic, TIES-Merging, Consensus Merging, TSVM, ...

Problem: Existing Benchmarks Are Too Idealized

- State-of-the-art merging methods are evaluated on highly favorable benchmarks
  - the same training configuration
- In practice, source models are trained with **diverse training recipes**
  - Different learning rates, steps, weight decay, label smoothing, Mixup, ...
- **Robustness in realistic settings is largely unexplored**

Research Question:

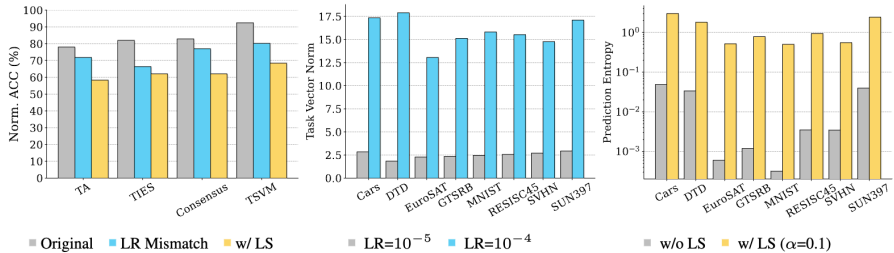
Where do model merging methods break down, and how can we fix it?

Failure Modes in Model Merging

We identify **two critical failure modes**:

(i) Task Vector Norm Disparity	(ii) Low-Confidence Source Models
e.g. different lr / steps / weight decay	e.g. label smoothing / Mixup / focal loss

- For example, in a vision MTL setting, even SOTA merging methods exhibit up to a 41.3% drop due to norm mismatch and a 53.3% drop due to low-confidence source models in Norm. ACC.



Failure Mode 1: Norm Disparity — Theoretical Analysis

Proposition 1

Let $\tau_1, \tau_2 \in \mathbb{R}^d$ with $\tau_1 \perp \tau_2$ and $\delta := \|\tau_1\|/\|\tau_2\|$. For $\tau_{\text{merge}} = \tau_1 + \tau_2$:

$$\cos(\tau_{\text{merge}},\, \tau_2) = \frac{1}{\sqrt{1+\delta^2}} \;\geq\; 1 - \tfrac{1}{2}\delta^2, \qquad \cos(\tau_{\text{merge}},\, \tau_1) = \frac{\delta}{\sqrt{1+\delta^2}} \;\leq\; \delta$$

Interpretation:

- When $\delta \ll 1$: merged model is almost entirely aligned with the **high-norm** task
- The low-norm task's contribution vanishes at $O(\delta)$
- Under the NTK approximation, functional shift is determined by task vector **direction**
- $\Rightarrow$ High-norm vector **dominates**, low-norm task's knowledge is lost

Failure Mode 2: Low Confidence — Theoretical Analysis

Consider a calibrated objective that adds a penalty $C_i(\theta)$ to the CE loss:

$$J_i^{\text{CAL}}(\theta) \;=\; J_i^{\text{CE}}(\theta) + \lambda_i\, C_i(\theta) \qquad \text{(e.g., label smoothing, focal loss)}$$

The resulting task vector acquires a **perturbation** δ_i:

$$\tau_i^{\text{CAL}} \;=\; \tau_i^{\text{CE}} + \delta_i + O(\lambda_i^2), \qquad \delta_i := -\lambda_i H_i^{-1} \nabla C_i(\theta_0)$$

Proposition 2 (informal)

Unless δ_i is aligned with the CE descent direction for *all* tasks, there exist merge coefficients such that:

$$J_i^{\text{CE}}\big(\theta_{\text{merge}}^{\text{CAL}}\big) \;>\; J_i^{\text{CE}}\big(\theta_{\text{merge}}^{\text{CE}}\big) \quad \text{by} \quad O(\lambda\|\tau\|)$$

Interpretation:

- Calibration penalties push task vectors into directions **misaligned** with the CE landscape
- This introduces a first-order degradation in CE loss upon merging
- Pure CE vectors do *not* suffer this first-order penalty

Proposed Method: DisTaC

Distillation for **Ta**sk vector **C**onditioning — a KD-based **pre-conditioning** method.

Two conditioning strategies:

1. **Norm Conditioning**
   Rescale $\tau_t \rightarrow \kappa_t \tau_t$, then recover performance via KD from the original source model $\theta_{\text{pre}} + \tau_t$
2. **Confidence Conditioning**
   Use asymmetric temperatures ($T_{\text{stu}} > T_{\text{tcr}}$) to sharpen the student's predictions

Key properties:

- Uses only unlabeled data
- Only a small number of steps needed.

Algorithm 1 DisTaC

Require: Pre-trained parameters $\boldsymbol{\theta}_{\text{pre}}$, task vector $\boldsymbol{\tau}_t$, scaling factor κ_t, temperature pair $(T_{\text{tcr}}, T_{\text{stu}})$, regularization weight β, unlabeled dataset $\tilde{\mathcal{D}}_t^u$ drawn from the distribution of task t, learning rate η, number of steps K
Ensure: Fine-tuned student parameters $\boldsymbol{\theta}$

$\boldsymbol{\theta}_0 \leftarrow \boldsymbol{\theta}_{\text{pre}} + \kappa_t \boldsymbol{\tau}_t$ ▷ Anchor point
$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_0$ ▷ Student initialization
for $k = 1, 2, \ldots, K$ **do**
  Sample mini-batch $\mathcal{B}_t \subset \tilde{\mathcal{D}}_t^u$
  $L \leftarrow 0$
  for all $\boldsymbol{x}_t \in \mathcal{B}_t$ **do**
    $\boldsymbol{z}_{\text{tcr}} \leftarrow f(\boldsymbol{x}_t; \boldsymbol{\theta}_{\text{pre}} + \boldsymbol{\tau}_t)$
    $\boldsymbol{z}_{\text{stu}} \leftarrow f(\boldsymbol{x}_t; \boldsymbol{\theta})$
    $\boldsymbol{s}_{\text{tcr}} \leftarrow \sigma(\boldsymbol{z}_{\text{tcr}} / T_{\text{tcr}})$
    $\boldsymbol{s}_{\text{stu}} \leftarrow \sigma(\boldsymbol{z}_{\text{stu}} / T_{\text{stu}})$
    $L \leftarrow L + T_{\text{tcr}} T_{\text{stu}} \mathrm{KL}(\boldsymbol{s}_{\text{tcr}} \| \boldsymbol{s}_{\text{stu}})$
  end for
  $L \leftarrow \frac{L}{|\mathcal{B}_t|} + \beta \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2^2$
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} L$ ▷ Gradient step
end for

Main Results — CLIP on Vision Tasks

Tasks: 8 vision classifications (Cars, DTD, EuroSAT, GTSRB, MNIST, RESISC45, SUN397, SVHN)

Method		Original		Norm Mismatch		Low Confidence	
		ViT-B-32	ViT-L-14	ViT-B-32	ViT-L-14	ViT-B-32	ViT-L-14
Pre-trained		47.3	65.1	47.3	65.1	47.3	65.1
Individual		89.9	93.7	89.3	93.3	89.8	94.0
MTL		87.8	92.6	-	-	-	-
Task arithmetic		70.4 (78.0)	84.0 (89.3)	63.6 (71.8)	78.6 (84.2)	51.0 (58.3)	66.9 (71.5)
Task arithmetic	+ **DisTaC**	-	-	**70.0** (78.2)	**83.9** (89.6)	**63.6** (72.2)	**77.6** (83.3)
TIES		74.0 (82.0)	85.0 (91.9)	59.1 (66.4)	74.0 (79.5)	54.5 (62.0)	68.3 (73.0)
TIES	+ **DisTaC**	-	-	**73.1** (81.0)	**84.4** (90.2)	**68.7** (77.9)	**79.4** (85.4)
Consensus TA		74.8 (82.8)	85.3 (90.7)	68.8 (77.0)	82.0 (87.6)	54.6 (62.0)	68.6 (73.2)
Consensus TA	+ **DisTaC**	-	-	**73.7** (82.2)	**84.9** (90.7)	**67.7** (76.5)	**80.0** (85.8)
EMR-Merging		88.5 (98.4)	93.0 (99.6)	80.0 (88.7)	87.6 (93.6)	39.2 (45.1)	27.4 (30.1)
EMR-Merging	+ **DisTaC**	-	-	**88.1** (97.3)	**92.7** (99.0)	**70.3** (79.2)	**92.3** (98.1)
TSVM		83.3 (92.4)	90.5 (96.3)	72.2 (80.2)	84.8 (90.7)	60.7 (68.4)	71.6 (76.4)
TSVM	+ **DisTaC**	-	-	**82.9** (91.8)	**90.3** (96.6)	**81.5** (91.8)	**89.7** (96.2)
Iso-CTS		81.0 (89.7)	90.4 (96.4)	78.1 (86.2)	90.8 (96.9)	72.5 (81.1)	80.8 (86.0)
Iso-CTS	+ **DisTaC**	-	-	**80.3** (88.9)	90.1 (96.1)	69.0 (78.1)	**86.1** (91.5)
WUDI-Merging		85.5 (93.9)	91.7 (97.7)	49.2 (52.6)	57.9 (60.8)	38.0 (40.8)	28.0 (29.2)
WUDI-Merging	+ **DisTaC**	-	-	**84.4** (93.2)	**91.4** (97.5)	**73.8** (83.3)	**91.6** (97.3)

- DisTaC consistently improves all merging methods under both failure modes
- Gains up to +42.5% (ViT-B-32) and +68.1% (ViT-L-14) in Norm. ACC
- Performance matches the original (ideal) benchmark in most cases

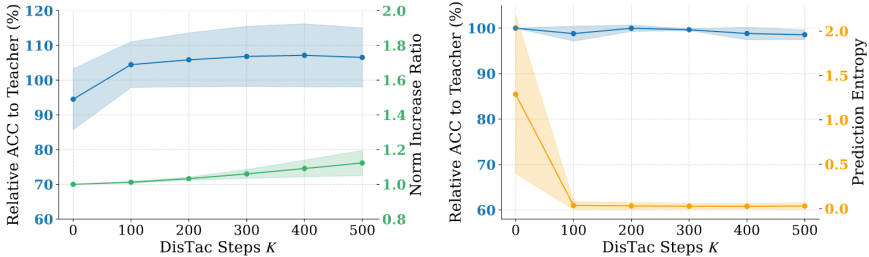
Main Results — NLP Tasks

Tasks: GLUE (CoLA, MRPC, RTE, SST-2)

Method		Original			Norm Mismatch			Low Confidence		
		RoBERTa-b	RoBERTa-l	Llama2-7b	RoBERTa-b	RoBERTa-l	Llama2-7b	RoBERTa-b	RoBERTa-l	Llama2-7b
Task arithmetic		60.9 (73.5)	68.3 (82.4)	75.9 (91.7)	56.8 (68.5)	46.0 (58.1)	55.3 (64.7)	61.3 (72.6)	64.5 (73.9)	75.7 (95.1)
Task arithmetic	+ **DisTaC**	-	-	-	**59.9 (71.7)**	**64.4 (80.5)**	**75.0 (91.1)**	**62.5 (74.6)**	**70.0 (82.3)**	73.0 (95.9)
Ties-merging		60.9 (74.8)	65.7 (80.7)	58.3 (80.7)	39.9 (46.1)	40.8 (51.3)	40.6 (47.7)	65.4 (79.1)	71.8 (84.0)	38.3 (47.5)
Ties-merging	+ **DisTaC**	-	-	-	**62.4 (76.4)**	**59.4 (75.9)**	**44.0 (51.6)**	64.4 (78.0)	**72.5 (86.4)**	**58.9 (78.4)**
TSVM		65.8 (80.8)	72.0 (87.8)	66.1 (78.5)	58.8 (71.1)	48.0 (60.7)	55.5 (65.5)	69.6 (84.3)	73.3 (85.6)	68.1 (84.6)
TSVM	+ **DisTaC**	-	-	-	**65.1 (79.6)**	**66.3 (84.1)**	**64.6 (77.4)**	67.5 (82.4)	**75.8 (90.8)**	**72.4 (97.1)**
Consensus-merging		61.3 (73.7)	67.9 (81.4)	74.5 (89.7)	58.1 (70.0)	38.1 (47.3)	58.3 (68.6)	61.2 (72.3)	65.2 (75.5)	65.0 (79.1)
Consensus-merging	+ **DisTaC**	-	-	-	**60.5 (72.5)**	**63.4 (79.0)**	**68.4 (82.3)**	**62.2 (74.3)**	**69.8 (82.3)**	**72.0 (94.9)**

- Observations similar to the vision task
- Gains up to +31.7% (Norm Mismatch) and +30.9% (Low Confidence) in Norm. ACC

Efficiency of DisTaC

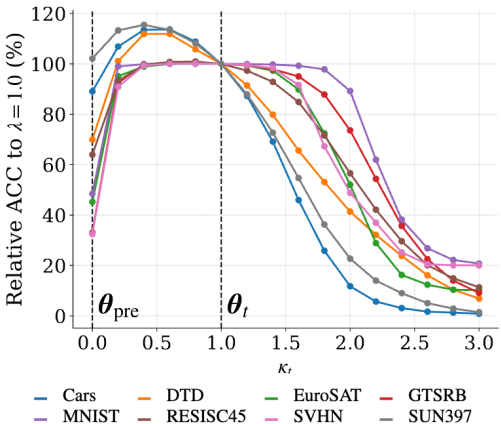


(a) Norm Mismatch (b) Low Confidence

- **Norm Mismatch (left):** accuracy recovers to teacher level within ∼100 steps; norm stays near κ_t-adjusted target
- **Low Confidence (right):** entropy drops substantially within ∼100 steps; accuracy fully preserved
- **Minimal computational overhead** — only 500 steps with unlabeled data

Guideline 1: Shrink, Don't Stretch

When norms differ, shrink the longer vectors rather than stretching shorter ones.



- Shrinking ($\kappa_t < 1$) retains or even *improves* accuracy
- Stretching ($\kappa_t > 1$) degrades accuracy; by $\kappa_t = 3.0$, worse than zero-shot
- Smaller displacements stay in the local linear regime around θ_{pre} (NTK perspective)
- Consistent with theoretical bounds (Wei et al., 2025)

Guideline 2: Overconfident → Merge → Calibrate

Better to merge overconfident models, then calibrate post-hoc.

- Well-calibrated / underconfident models are **fragile** for merging
- Overconfident models are **robust** to the merge process
- Post-hoc calibration (e.g., temperature scaling) can fix overconfidence
- But merging underconfident models causes **irreversible** performance loss

Conclusion & Future Work

Summary:

- **Identified two failure modes** in model merging: norm disparity & low confidence
- **Proposed DisTaC:** lightweight KD-based pre-conditioning
- **Results:** restores performance to ideal benchmark levels across all merging methods
- **Guidelines:** shrink long task vectors; merge overconfident, then calibrate post-hoc

Future Work:

- Explore failure modes in **generative tasks** (e.g., LLMs, diffusion models) and apply DisTaC
- Investigate DisTaC with **surrogate-task unlabeled data** — can we condition task vectors using out-of-distribution data from non-target tasks?

Code: `https://github.com/katoro8989/DisTaC`