CAT Merging: A Training-Free Approach for Resolving Conflicts in Model Merging

Wenju Sun[1], Qingyong Li[1], Yangli-ao Geng[1], Boyang Li[2]

[1]Beijing Jiaotong University

[2]Nanyang Technological University

Presented by: Wenju Sun

The 42nd International Conference on Machine Learning (ICML 2025)

Content

ICML
International Conference
On Machine Learning

Introduction

- Model Merging
- Baseline: Task Arithmetic
- Knowledge Conflict

Method

- Isolating Knowledge Conflict to Each Layer
- Objective
- Example: for Linear Weight

Experiment

Model Merging

ICML
International Conference
On Machine Learning

Definition: Considering a pretrained model W_0 and a set of finetuned models $\{W_i\}_{i=1}^K$ with corresponding downstream tasks $\{D_i\}_{i=1}^K$.

Our goal is to merge all K models into a unified model W_{mtl} **without redundant retraining**. The unified model W_{mtl} should **perform well on all downstream tasks**.

Baseline: Task Arithmetic

Task Arithmetic[1]: Considering a pretrained model W_0 and a set of finetuned models $\{W_i\}_{i=1}^k$ with corresponding downstream tasks $\{D_i\}_{i=1}^k$, the task vectors $\{\tau_i\}_{i=1}^k$ are defined as $\tau_i = W_i - W_0$.

Task vectors can be applied to W_0 with a scaling term λ, i.e.,$W_{mtl} = W_0 + \alpha \sum_i \tau_i$, which allows to control the behavior of the edited model via simple arithmetic operations on task vectors.

Knowledge Conflict

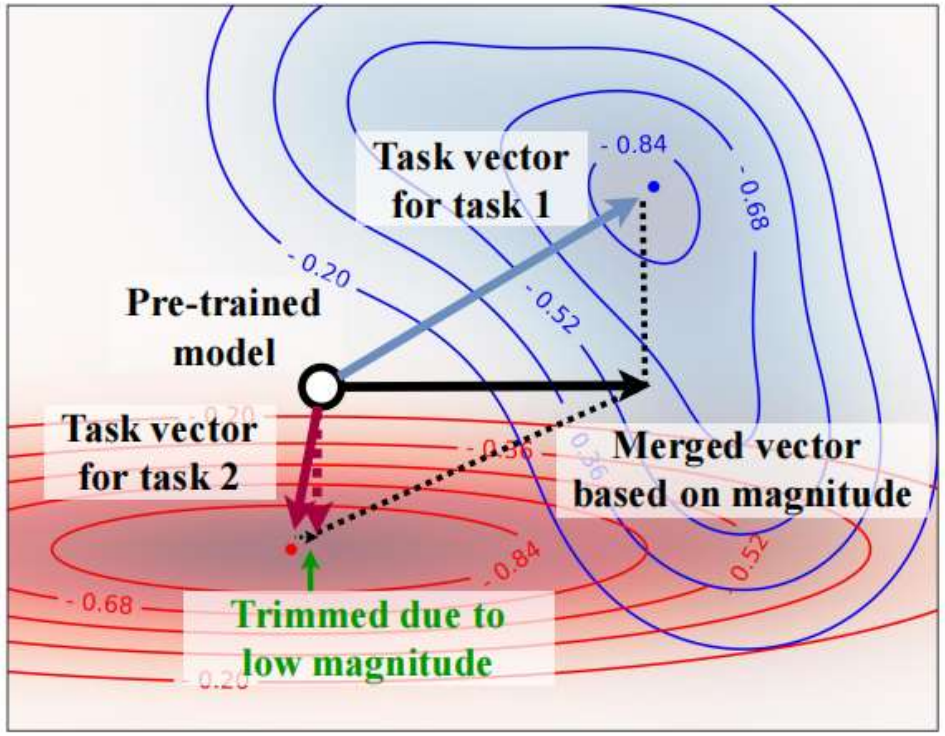
Definition (for task arithmetic): the increase in task-specific loss incurred by incorporating **other** task vectors.

For a given task *k*, associated with the loss function $\mathcal{L}_k(.)$, the conflict introduced by another task vector $T_i (i \neq k)$ is quantified as:

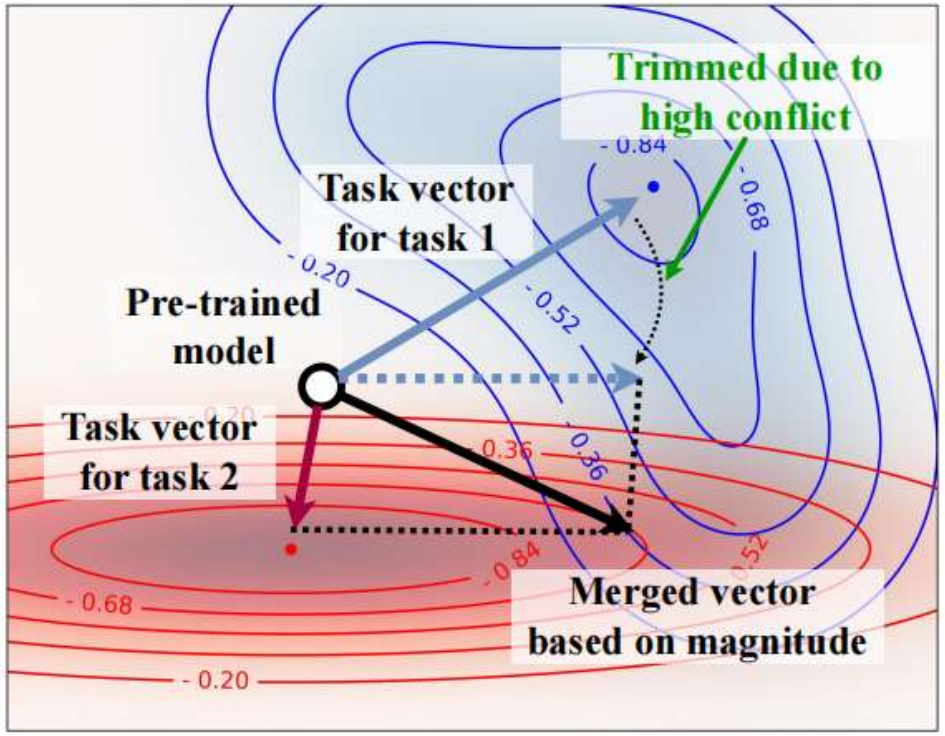
$$\Delta\mathcal{L}_{k|i} = \mathcal{L}_k\left(W_k + T_i\right) - \mathcal{L}_k\left(W_k\right)$$

Knowledge Conflict

ICML
International Conference
On Machine Learning



Magnitude based Methods



CAT Merging (ours)

Isolating Knowledge Conflict

Theorem 4.4 (An Upper Bound on Knowledge Conflict):
Suppose that within the range of model merging, the function of layer l is γ_l-Lipschitz continuous with respect to its input, and the loss function L is β-Lipschitz continuous with respect to the final output of the network. Then, the knowledge conflict follows:

$$|\Delta\mathcal{L}_{k|i}| \leq \beta \sum_{l=1}^{L} \Big(\prod_{m=l+1}^{L} \gamma_m \Big) \|\Delta \hat{f}^l_{k|i}\|$$

Objective

ICML
International Conference
On Machine Learning

$$\min_{\Phi_k} \sum_{i \neq k} \underbrace{\left\| \hat{f}_k\left(W_k + \Phi_k(T_i)\right) - f_k(W_k) \right\|^2}_{\text{Inter-task Knowledge Conflict } \Delta \hat{f}^l_{k|i}} + \underbrace{\lambda \left\| \hat{f}_i(W_0 + \Phi_k(T_i)) - f_i(W_i) \right\|^2}_{\text{Intra-task Knowledge Deviation}}$$

We propose to balances two objectives:

(1) Minimizing interference between tasks;

(2) Preserving the knowledge encoded in task vectors.

For Linear Weights

We introduce a **removal** basis B_k, such that the trimming projector $\Phi_k(T_i) = T_i - T_i B_k B_K^\top$ maximize the following objective:

$$\max_{B_k} \sum_{i \neq k} \left(\left\| X_k T_i B_k B_k^\top \right\|_F^2 - \lambda \left\| X_i T_i B_k B_k^\top \right\|_F^2 \right)$$

The optimal basis of B_k is constructed from the top-c eigenvectors of the matrix:

$$\sum_{i \neq k} T_i^\top (X_k^\top X_k - \lambda X_i^\top X_i) T_i$$

Experiments

ICML
International Conference
On Machine Learning

Table 1. Multi-task performance when merging ViT-B/32 models on eight vision tasks. The best and second-best performances are writter in **bold** and underlined text. The "#best" column represents the number of datasets where the method performs the best.

Method	SUN397	Cars	RESISC45	EuroSAT	SVHN	GTSRB	MNIST	DTD	Avg Acc	#best
Pretrained	62.3	59.7	60.7	45.5	31.4	32.6	48.5	43.8	48.0	-
Individual	75.3	77.7	96.1	99.7	97.5	98.7	99.7	79.4	90.5	-
Traditional MTL	73.9	74.4	93.9	98.2	95.8	98.9	99.5	77.9	88.9	-
Weight Averaging	65.3	63.4	71.4	71.7	64.2	52.8	87.5	50.1	65.8	0
Fisher Merging	**68.6**	**69.2**	70.7	66.4	72.9	51.1	87.9	59.9	68.3	2
RegMean	65.3	63.5	75.6	78.6	78.1	67.4	93.7	52.0	71.8	0
Task Arithmetic	55.2	54.9	66.7	78.9	80.2	69.7	97.3	50.4	69.1	0
Ties-Merging	59.8	58.6	70.7	79.7	86.2	72.1	98.3	54.2	72.4	0
TATR	62.7	59.3	72.3	82.3	80.5	72.6	97.0	55.4	72.8	0
Ties-Merging & TATR	66.3	65.9	75.9	79.4	79.9	68.1	96.2	54.8	73.3	0
Consensus Merging	65.7	63.6	76.5	77.2	81.7	70.3	97.0	57.1	73.6	0
PCB Merging	63.8	62.0	77.1	80.6	**87.5**	**78.5**	**98.7**	58.4	75.8	3
CAT Merging (ours)	68.1	65.4	**80.5**	**89.5**	85.5	**78.5**	98.6	**60.7**	**78.3**	4

Experiments

ICML
International Conference
On Machine Learning

Table 2. Multi-task performance when merging ViT-L/14 models on eight vision tasks.

Method	SUN397	Cars	RESISC45	EuroSAT	SVHN	GTSRB	MNIST	DTD	Avg Acc	#best
Pretrained	66.8	77.7	71.0	59.9	58.4	50.5	76.3	55.3	64.5	-
Individual	82.3	92.4	97.4	100.0	98.1	99.2	99.7	84.1	94.2	-
Traditional MTL	80.8	90.6	96.3	96.3	97.6	99.1	99.6	84.4	93.5	-
Weight Averaging	72.1	81.6	82.6	91.9	78.2	70.7	97.1	62.8	79.6	0
Fisher Merging	69.2	**88.6**	87.5	93.5	80.6	74.8	93.3	70.0	82.2	1
RegMean	73.3	81.8	86.1	**97.0**	88.0	84.2	98.5	60.8	83.7	1
Task Arithmetic	73.9	82.1	86.6	94.1	87.9	86.7	98.9	65.6	84.5	0
Ties-Merging	<u>76.5</u>	85.0	89.3	95.7	90.3	83.3	99.0	68.8	86.0	0
TATR	74.6	83.7	87.6	93.7	88.6	88.1	99.0	66.8	85.3	0
Ties-Merging & TATR	76.3	85.3	88.8	94.4	<u>90.8</u>	88.7	<u>99.2</u>	68.8	86.5	0
Consensus Merging	75.0	84.3	89.4	95.6	88.3	82.4	98.9	68.0	85.2	0
PCB Merging	76.2	86.0	<u>89.6</u>	95.9	89.9	<u>92.3</u>	<u>99.2</u>	<u>71.4</u>	<u>87.6</u>	0
CAT Merging (ours)	**78.7**	<u>88.5</u>	**91.1**	<u>96.3</u>	**91.3**	**95.7**	**99.4**	**75.7**	**89.6**	6

Experiments

ICML International Conference On Machine Learning

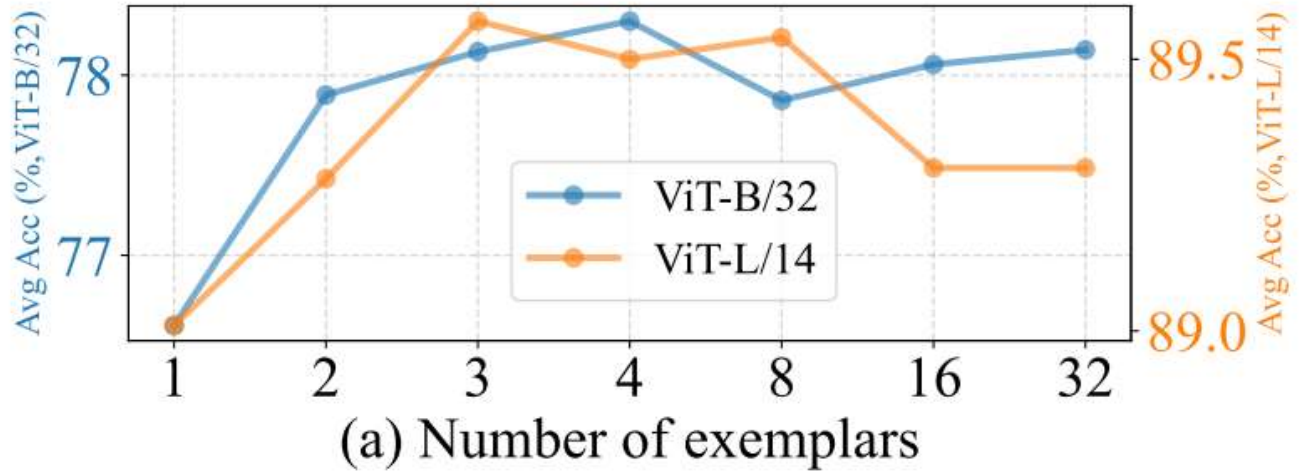
Table 3. Multi-task performance when merging RoBERTa models on eight NLP tasks.

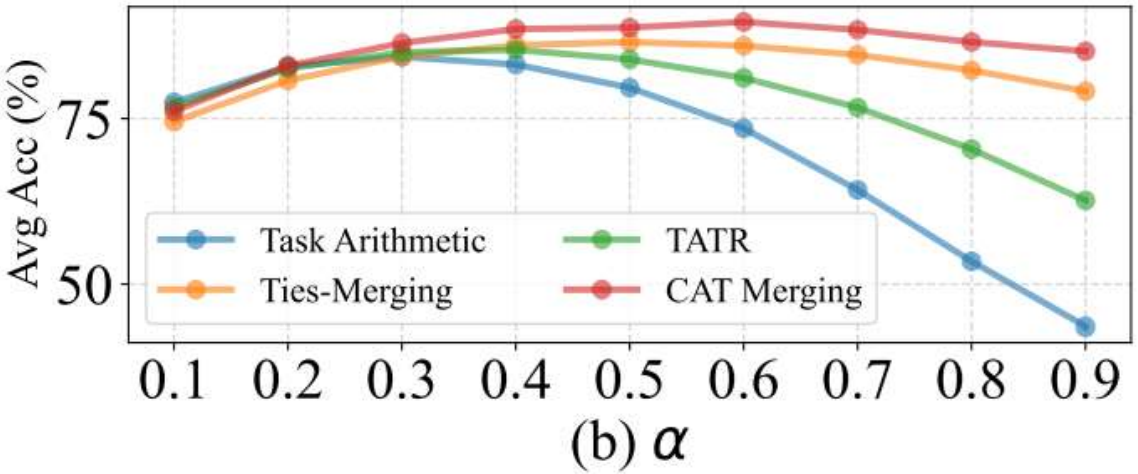
Method	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST2	STS-B	Average	#best
Task Arithmetic	6.68	66.23	**78.46**	78.62	72.69	53.43	83.49	27.10	58.34	1
Ties-Merging	9.46	59.34	74.71	65.93	41.29	47.29	72.13	9.21	47.42	0
TATR	10.20	65.44	72.56	75.73	74.58	55.18	78.87	37.46	58.39	0
PCB Merging	11.40	50.85	77.63	78.22	55.78	60.29	75.57	**67.01**	59.59	1
CAT Merging (ours)	**33.20**	**72.33**	68.22	**82.92**	**76.05**	**62.82**	**89.33**	15.57	**62.56**	**6**

Table 4. Multi-task performance when merging BLIP models on six vision-language tasks.

Method	COCO Caption	Flickr30k Caption	Textcaps	OKVQA	TextVQA	ScienceQA	#best
	CIDEr	CIDEr	CIDEr	Accuracy	Accuracy	Accuracy	
Pretrained	0.07	0.03	0.05	42.80	21.08	40.50	-
Task Arithmetic	0.86	0.50	**0.39**	17.71	0.49	40.10	1
Ties-Merging	0.53	0.27	0.22	27.95	0.57	40.35	0
TATR	0.46	0.31	0.21	28.30	14.74	42.98	0
PCB Merging	0.71	0.52	0.30	36.04	1.88	43.01	0
CAT Merging (ours)	**0.91**	**0.53**	0.36	**44.07**	**19.69**	**46.36**	5

Experiments





- Only need 3 or 4 samples for each task.
- Robust to the value of α.

THANKS FOR WATCHING

CAT Merging: A Training-Free Approach for Resolving Conflicts in Model Merging

Presented by: Wenju Sun

北京交通大學
BEIJING JIAOTONG UNIVERSITY

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

ICML
International Conference On Machine Learning