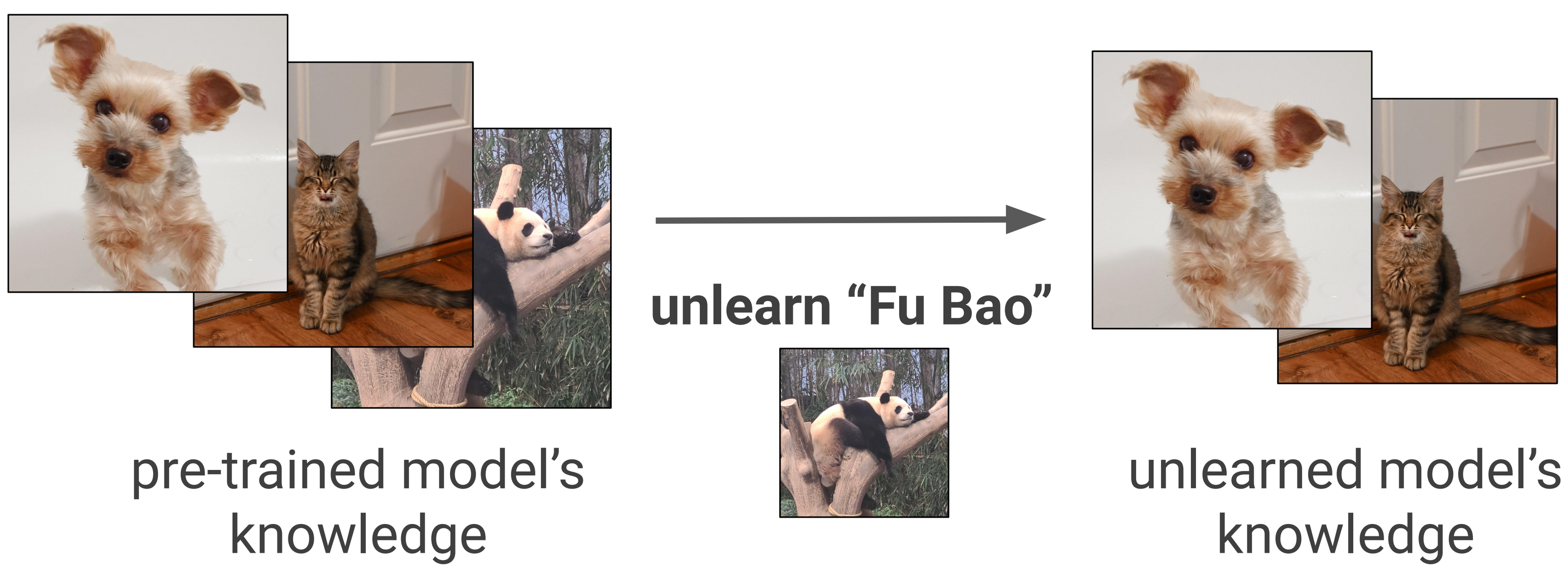
NegMerge: Consensual Weight Negation for Strong Machine Unlearning

Hyoseo Kim[1,2*], Dongyoon Han[1†], Junsuk Choe[2†]
[1]NAVER AI LAB, [2]Sogang University
*Work done during an internship at NAVER AI Lab, †Corresponding author

Poster Code Paper

What is Machine Unlearning?

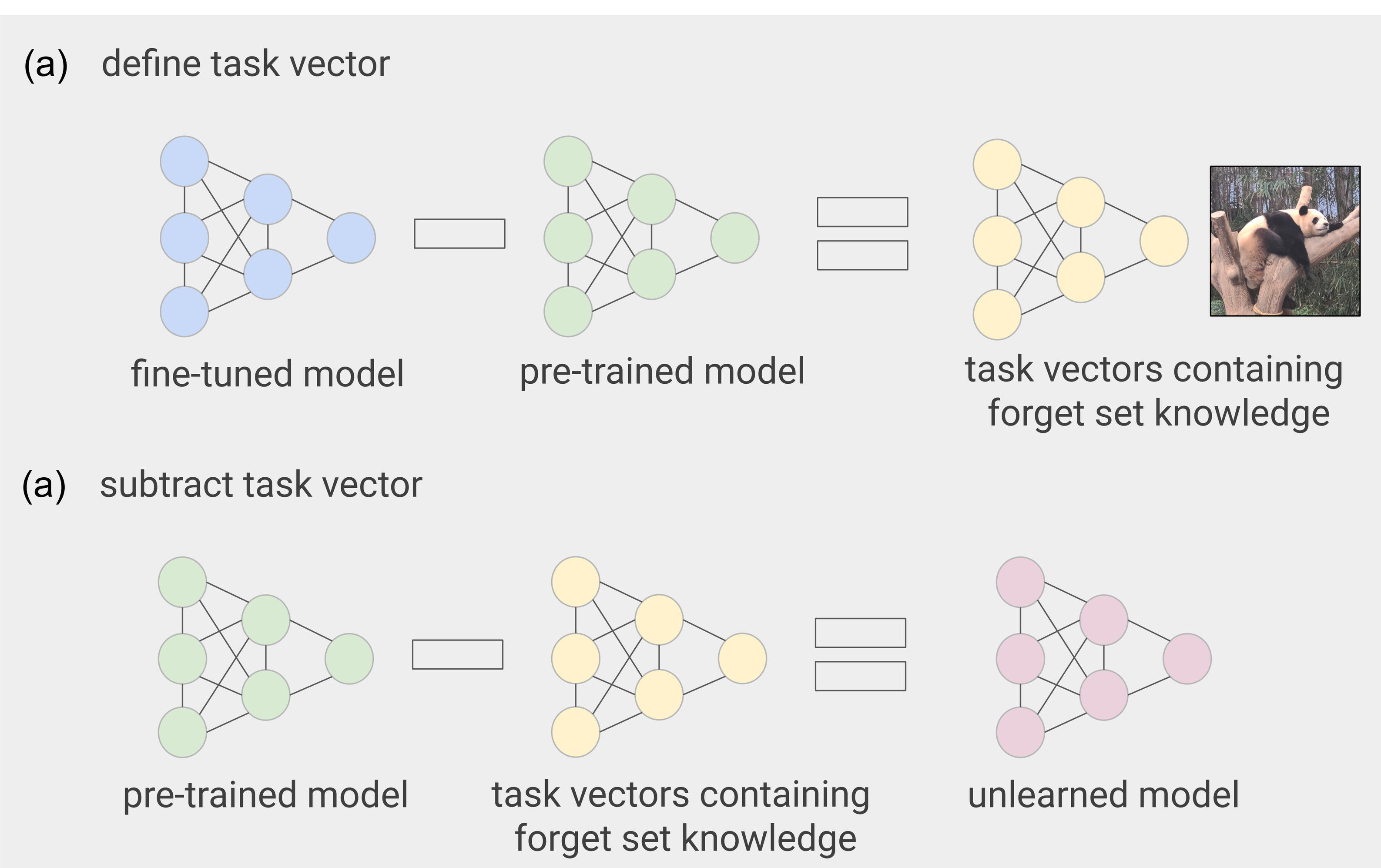
- Removal of specific knowledge from a pre-trained model without impacting its remaining knowledge.
- Applications: protecting privacy and rights under regulations like GDPR, correcting inaccuracies, and removing harmful data.



How to Unlearn?

- Forgetting via negation[1]
  - Adjust the model by subtracting the sum of the task vectors.

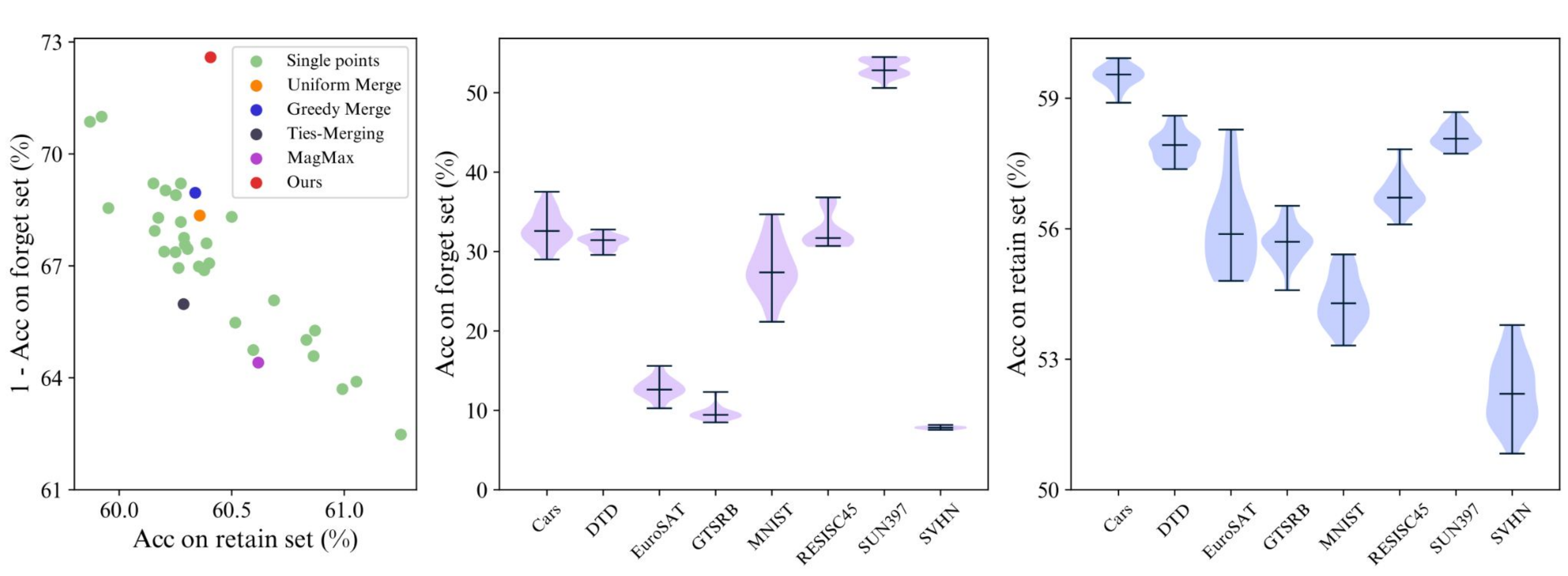
$$\theta_{unlearn} = \theta_{pre} - \lambda\underbrace{(\theta_{ft}^{forget} - \theta_{pre})}_{\text{sum of the task vectors}}$$



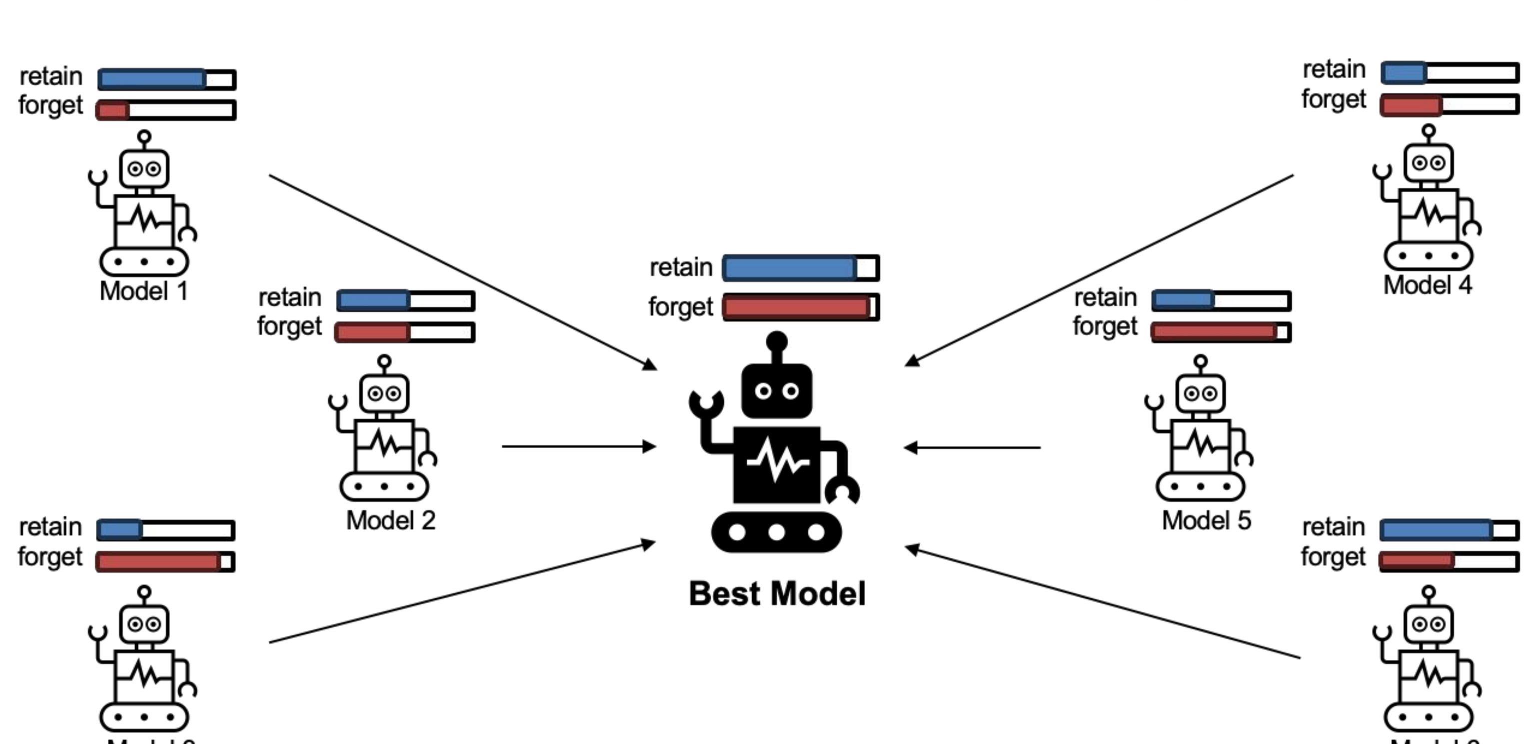
[1] Ilharco, Gabriel, et al. "Editing models with task arithmetic." The Eleventh International Conference on Learning Representations.

Challenges of Machine Unlearning

- Highly sensitive to the hyperparameters used for fine-tuning.
- Trade-off in unlearning performance and retaining performance.



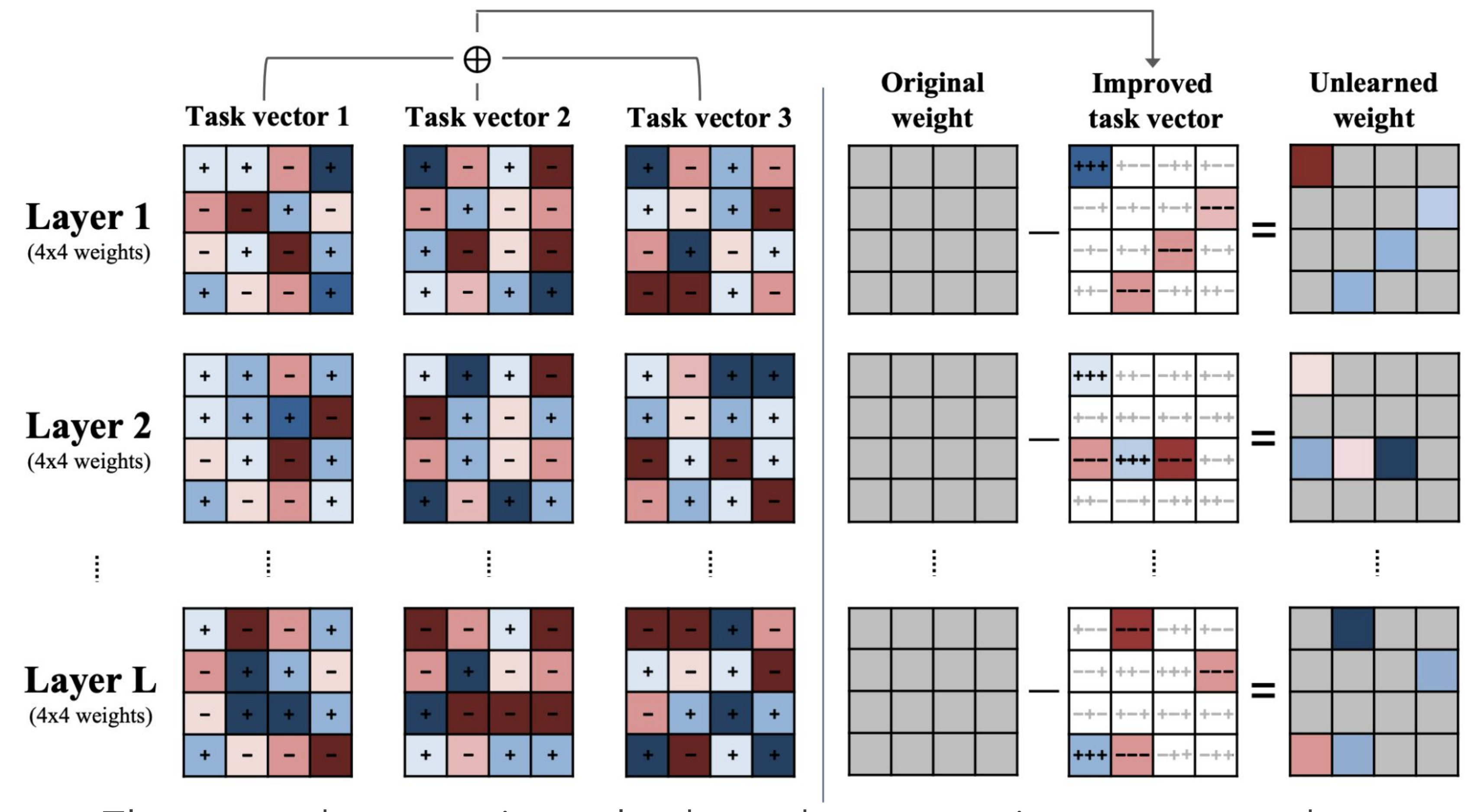
- Effective unlearning requires the fine-tuned model to maintain high performance on the forget set without degrading performance on the retain set.
- To achieve this, merge them all instead of selecting just one.



Overview of NegMerge

- Hyperparameter tuning generates multiple fine-tuned models.
- Merge all models based on the sign consensus of task vectors.

$$\tau_{\text{merged}} = \left(\frac{1}{n}\sum_{k=1}^{n}\tau_k\right) \odot \mathbf{1}_{\text{signs are equal}}$$



- Elements that consistently show the same sign across task vectors are attributed to the forget set.
- Components that exhibit differing signs are considered less related to the forget set.

Experiment Results

- Unlearning on CLIP ViT Models.
  - achieves the best reduction in accuracy on the forget set.

Method	ViT-B/32		ViT-B/16		ViT-L/14		Time (sec)
	Acc D_f(↓)	Acc D_r(↑)	Acc D_f(↓)	Acc D_r(↑)	Acc D_f(↓)	Acc D_r(↑)	
Pre-trained	48.13	63.33	55.49	68.32	65.19	75.54	-
Task Arithmetic							
*Paper number**	24.00	60.90	21.30	65.40	19.00	72.90	-
Single Best Model†	23.63	60.60	20.64	64.04	19.17	72.09	-
Uniform Merge	22.50	60.55	21.51	64.60	18.10	71.91	$12_{\pm0.1}$
Greedy Merge‡	23.31	60.75	21.34	64.54	17.71	71.99	$607_{\pm2.6}$
TIES-Merging	26.21	61.08	23.78	64.72	22.70	72.41	$128_{\pm10.1}$
MagMax	25.24	60.95	24.45	64.78	21.71	72.55	$24_{\pm1.8}$
NegMerge (Ours)	**20.76**	60.36	**19.24**	64.54	**17.32**	72.08	$37_{\pm1.2}$
Linear Task Arithmetic							
*Paper number**	10.90	60.80	11.30	64.80	-	-	-
Single Best Model†	8.88	60.16	6.92	64.62	-	-	-
Uniform Merge	9.12	60.47	6.84	65.26	-	-	$19_{\pm2.3}$
Greedy Merge‡	8.73	60.27	6.80	64.72	-	-	$1696_{\pm35.3}$
TIES-Merging	10.66	60.38	8.44	65.12	-	-	$378_{\pm8.0}$
MagMax	11.33	60.67	8.65	65.17	-	-	$164_{\pm2.4}$
NegMerge (Ours)	**8.03**	60.58	**6.60**	65.40	-	-	$194_{\pm1.6}$

- 10% Random Data Forgetting on CIFAR-10 using ResNet-18.
  - achieves the smallest average gap of 1.07.

Methods	Used Splits	Acc D_r(≃)	Acc D_f(≃)	Acc D_{test}(≃)	MIA(≃)	Avg. Gap(↓)
Retrain *	retain	$100.00_{\pm0.00}$	$94.76_{\pm0.69}$	$94.26_{\pm0.02}$	$12.88_{\pm0.09}$	0.00
Random Labeling *	all	$99.67_{\pm0.14}$	$92.39_{\pm0.31}$	$92.83_{\pm0.38}$	$37.36_{\pm0.06}$	7.15
Influence *	all	$99.20_{\pm0.22}$	$98.93_{\pm0.28}$	$93.20_{\pm1.03}$	$2.67_{\pm0.01}$	4.06
SalUn *	all	$99.62_{\pm0.12}$	$97.15_{\pm0.43}$	$93.93_{\pm0.29}$	$14.39_{\pm0.82}$	1.15
Finetune *	retain	$99.88_{\pm0.08}$	$99.37_{\pm0.55}$	$94.06_{\pm0.27}$	$2.70_{\pm0.01}$	3.78
$\ell1$-sparse *	retain	$97.74_{\pm0.33}$	$95.81_{\pm0.62}$	$91.59_{\pm0.57}$	$9.84_{\pm0.00}$	2.26
Gradient Ascent *	forget	$99.50_{\pm0.38}$	$99.31_{\pm0.54}$	$94.01_{\pm0.47}$	$1.70_{\pm0.01}$	4.12
Boundary Shrink *	forget	$98.29_{\pm2.50}$	$98.22_{\pm2.52}$	$92.69_{\pm2.99}$	$8.96_{\pm0.13}$	2.67
Boundary Expanding *	forget	$99.42_{\pm0.33}$	$99.41_{\pm0.30}$	$93.85_{\pm1.02}$	$7.47_{\pm1.15}$	2.76
Random Labeling	forget	$99.99_{\pm0.00}$	$99.98_{\pm0.02}$	$95.04_{\pm0.11}$	$2.15_{\pm1.94}$	4.19
SalUn	forget	$99.88_{\pm0.04}$	$99.89_{\pm0.04}$	$94.42_{\pm0.05}$	$9.51_{\pm2.07}$	2.20
Task Arithmetic						
Single Best Model†	forget	$98.36_{\pm0.51}$	$94.85_{\pm0.16}$	$91.49_{\pm0.80}$	$10.91_{\pm0.72}$	1.62
Uniform Merge	forget	$98.70_{\pm0.91}$	$95.83_{\pm2.17}$	$92.36_{\pm1.16}$	$10.14_{\pm2.93}$	1.75
TIES-Merging	forget	$98.38_{\pm0.17}$	$95.45_{\pm0.32}$	$92.23_{\pm0.14}$	$9.36_{\pm0.31}$	1.96
MagMax	forget	$98.38_{\pm0.12}$	$97.97_{\pm0.77}$	$91.53_{\pm0.00}$	$8.45_{\pm2.60}$	3.00
NegMerge (Ours)	forget	$99.15_{\pm0.24}$	$96.63_{\pm0.59}$	$92.71_{\pm0.39}$	$12.87_{\pm1.29}$	**1.07**

- Impact of Sign Conflicts on Unlearning: Grad-CAM

(a) Original image (b) Original model (c) Conflict (d) Non-conflict (ours)

Conclusion

- Propose a novel machine unlearning technique, NegMerge, based on task arithmetic and model merging.
- Tested on the CLIP ViT models and the standard ResNet18 classifier, achieving SOTA across nine datasets.