VIPA
Visual Intelligence and Pattern Analysis

Training Free Pretrained Model Merging

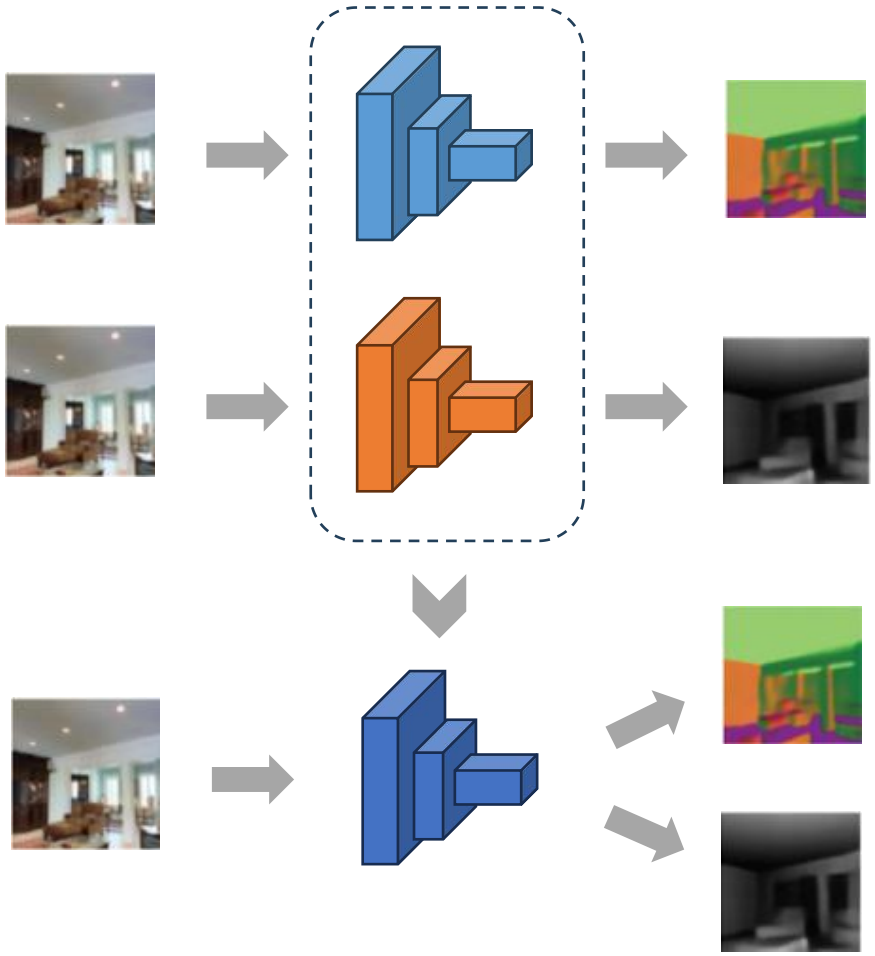
Zhengqi Xu[1], Ke Yuan[1], Huiqiong Wang[2], Yong Wang[3], Mingli Song[1], Jie Song[1*]

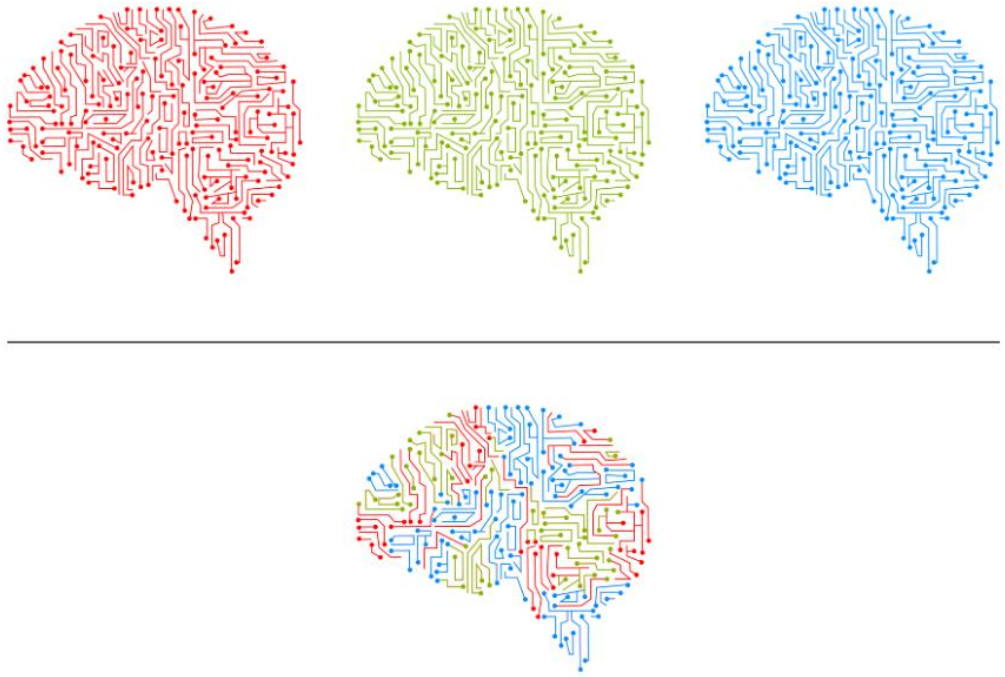
Zhejiang University[1],
Ninbo Innovation Center, Zhejiang University[2],
State Grid Shangdong Electric Power Company[3]

Paper: https://arxiv.org/abs/2403.01753
Code: https://github.com/zju-vipa/training_free_model_merging

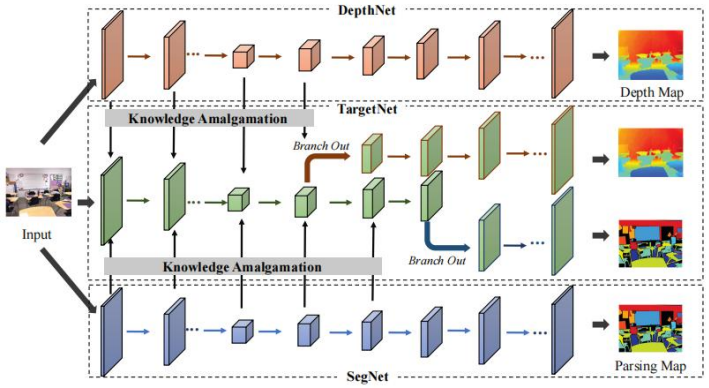
Introduction

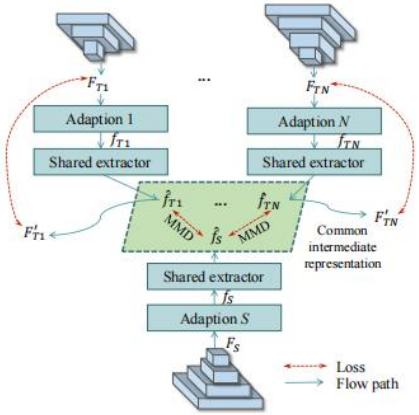
What is Model Merging?





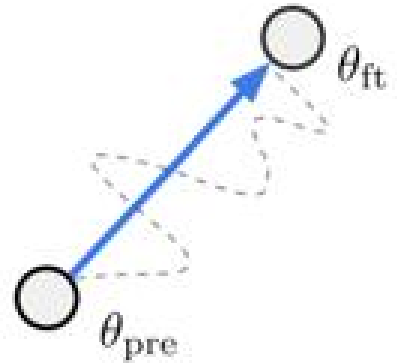
Related Work



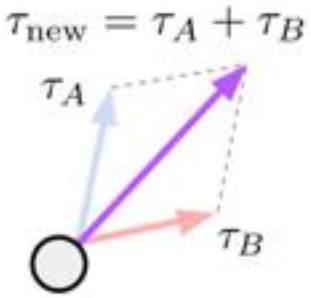


Training-free Model Merging

- Merging model by distilling knowledge from multiple teacher models into student models.
- Need additional training.



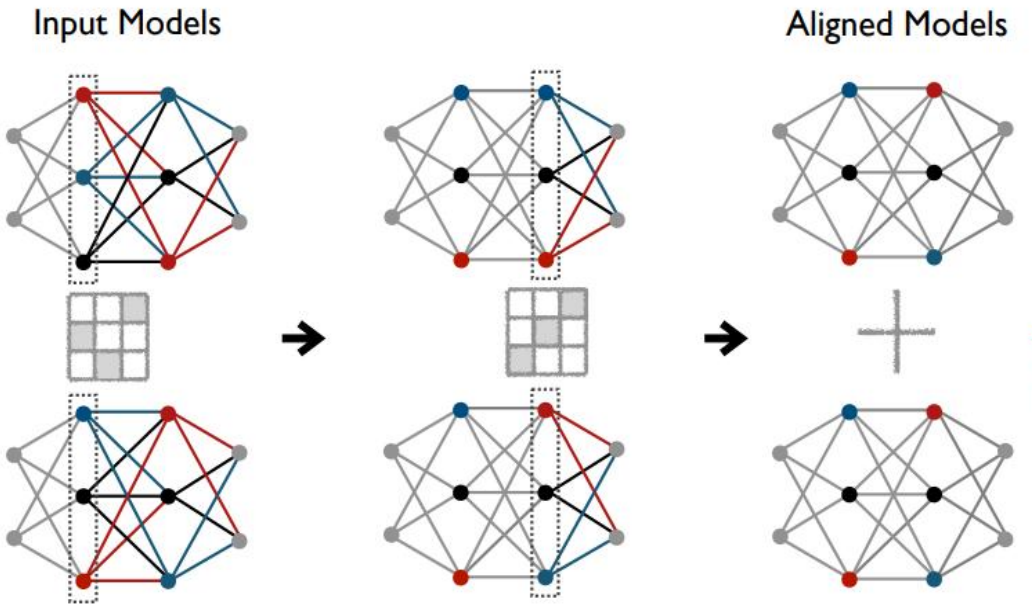
$$\tau = \theta_{\text{ft}} - \theta_{\text{pre}}$$

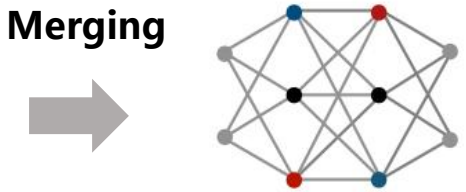


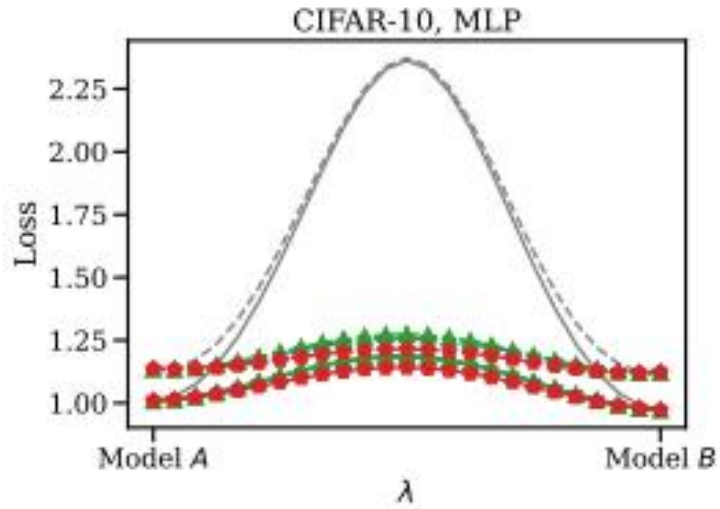
Training-free Model Merging

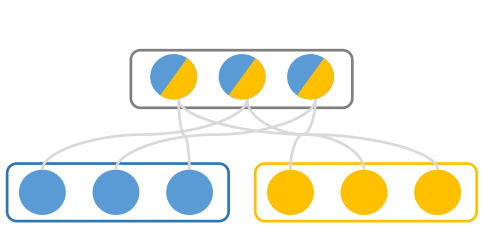
- Merging model by linearly adding Task Vector τ onto the pretrained model.
- Training free, but models must be initialized with the same pre-trained model.

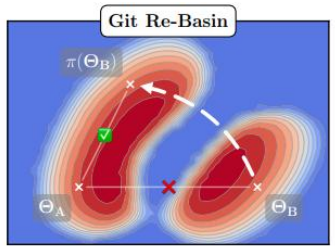
Related Work Training-free Model Merging -Unit Alignment

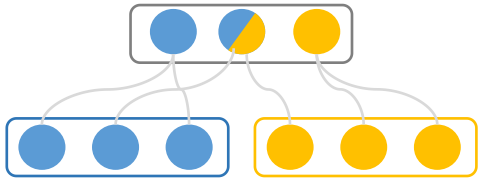


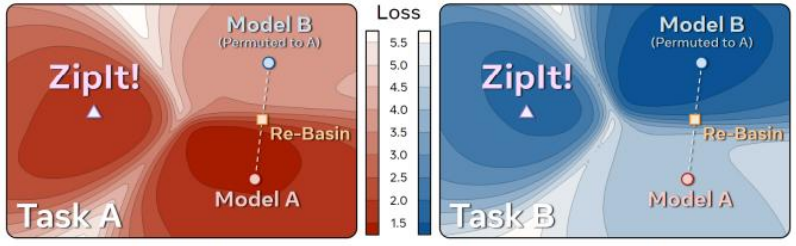










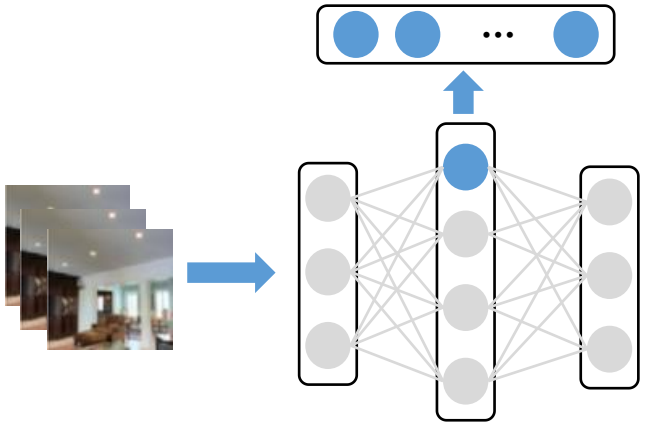


- **Git Rebasin**: Align units across models
- **Zipit**: Align units across&within models

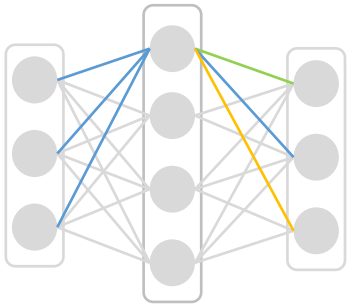
Motivation

How to represent neuronal units?

Collect activations for the unit



Select weight vector for the unit



Weights of units

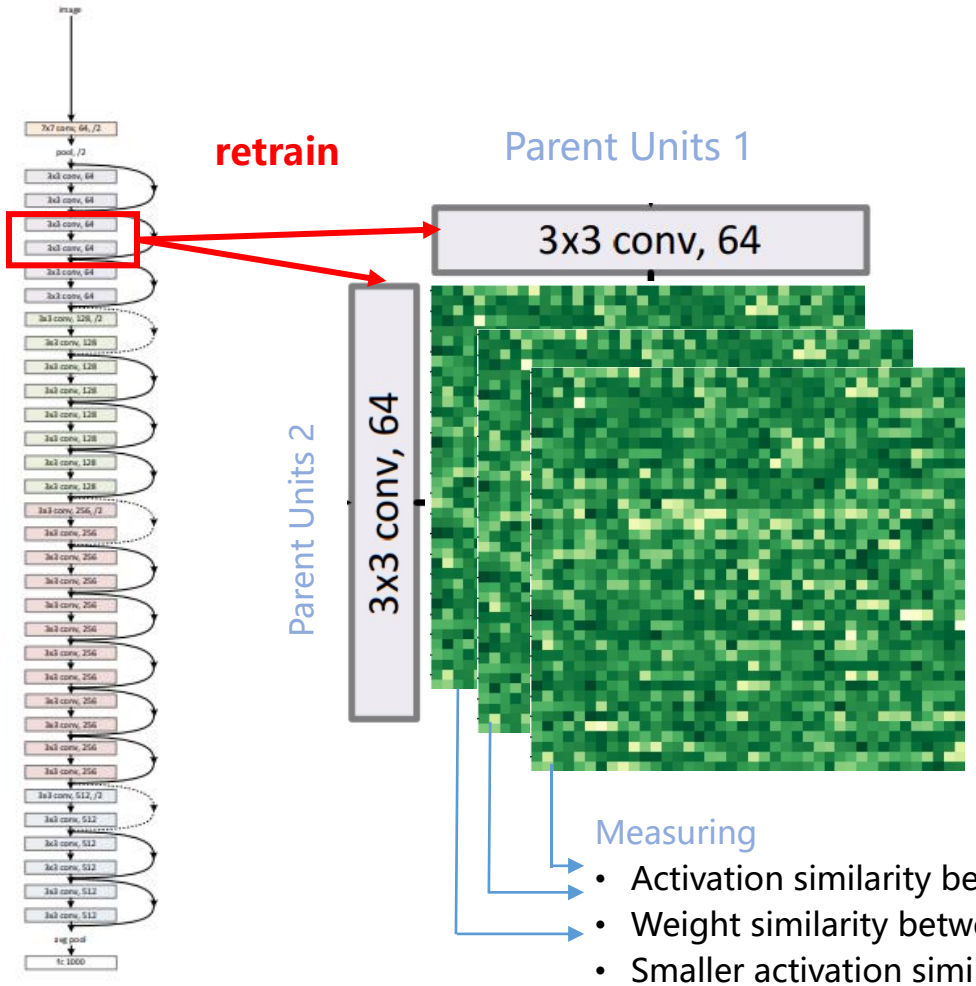
- Reflects its connection strength with the units in the previous layer.
- Cannot capture the information of specific features (i.e. scale, offset) due to the absence of input data.

Activations of units

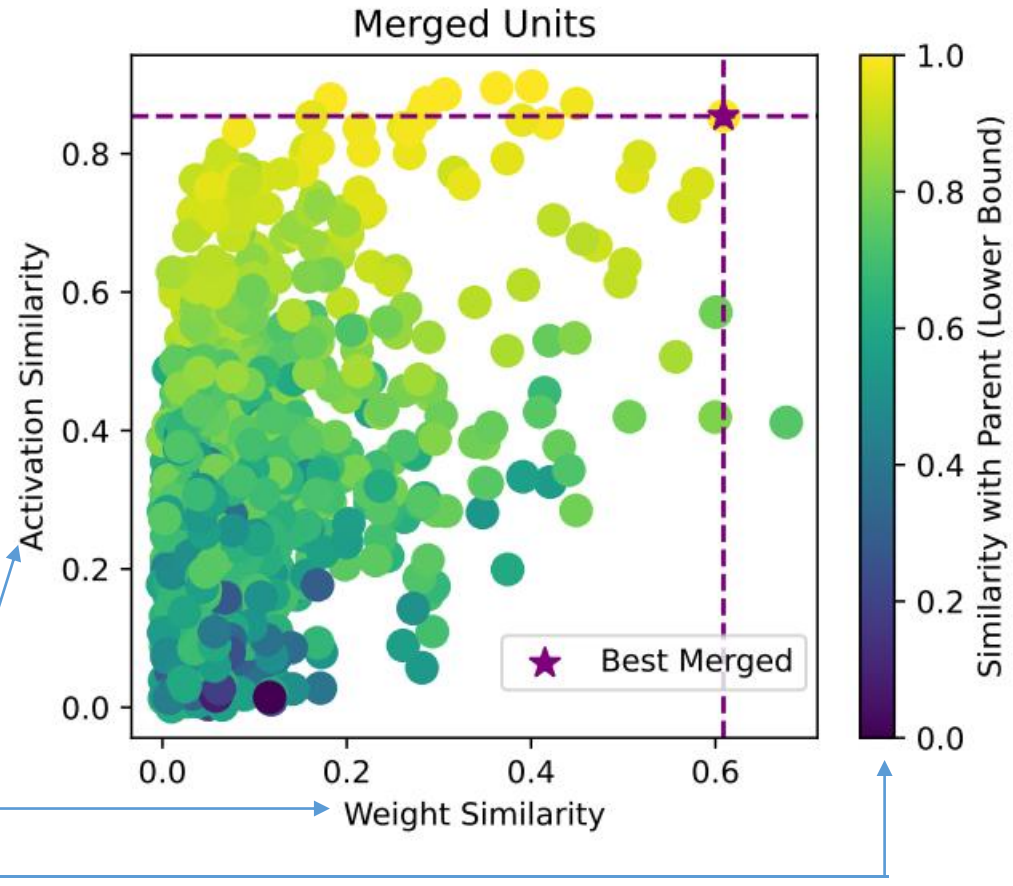
- The same task can be achieved by distinct parameters.

Motivation

X-axis: Weight similarity of parents
Y-axis: Activation similarity of parents
Color: Merged units' smaller act. similarity with its parents



Measuring
- Activation similarity between parents
- Weight similarity between parents
- Smaller activation similarity between merged units and their parents



Resnet50 trained on CIFAR10

Method

Merging under Dual-Space Constraints (MuDSC)

A General Procedure of Model Merging

$$\boldsymbol{W}_l' = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{P}_l^{(n)})^\top \boldsymbol{W}_l^{(n)} \boldsymbol{P}_{l-1}^{(n)},\ l \in \{1,2,\ldots,L\}$$

Solve P_l

Activation-based Matching

$$P_l = \Psi(\mathbb{C}(A_l)), l \in \{1,2,\ldots,L\}$$

Weight-based Matching

$$\begin{cases} \boldsymbol{P}_l^1 = \mathbb{I} \\ \boldsymbol{Z}_l^{t+1} = \upsilon(\boldsymbol{P}_{l-1}^t, \boldsymbol{P}_{l+1}^t, \boldsymbol{W}_l^t, \boldsymbol{W}_{l+1}^t) \\ \boldsymbol{P}_l^{t+1} = \Psi(\mathbb{C}(\boldsymbol{Z}_l^{t+1})) \end{cases}$$

MuDSC Matching

$$\begin{cases} \boldsymbol{P}_l^1 = \Psi(\mathbb{C}(\boldsymbol{A}_l)) \\ \boldsymbol{Z}_l^{t+1} = \upsilon(\boldsymbol{P}_{l-1}^t, \boldsymbol{P}_{l+1}^t, \boldsymbol{W}_l^t, \boldsymbol{W}_{l+1}^t) \\ \boldsymbol{P}_l^{t+1} = \Psi(\alpha\mathbb{C}(\boldsymbol{Z}_l^{t+1}) + (1-\alpha)\mathbb{C}(\boldsymbol{A}_l)) \end{cases}$$

The MuDSC balances the inconsistency of unit similarity in weight space and activation space when merging models by linearly combining the similarity matrices both of the weights and the activations of the units to seek a better permutation matrix.

Algorithm 1 MuDSC Merging

Require: The weights $\{\boldsymbol{W}_l\}_{l=1}^L$ and the activations $\{\boldsymbol{A}_l\}_{l=1}^L$ of the models, the function for computing the similarity matrix $\mathbb{C}$, selected matching algorithm Ψ, the balanced factor α. RP returns a random permutation of the sequence.

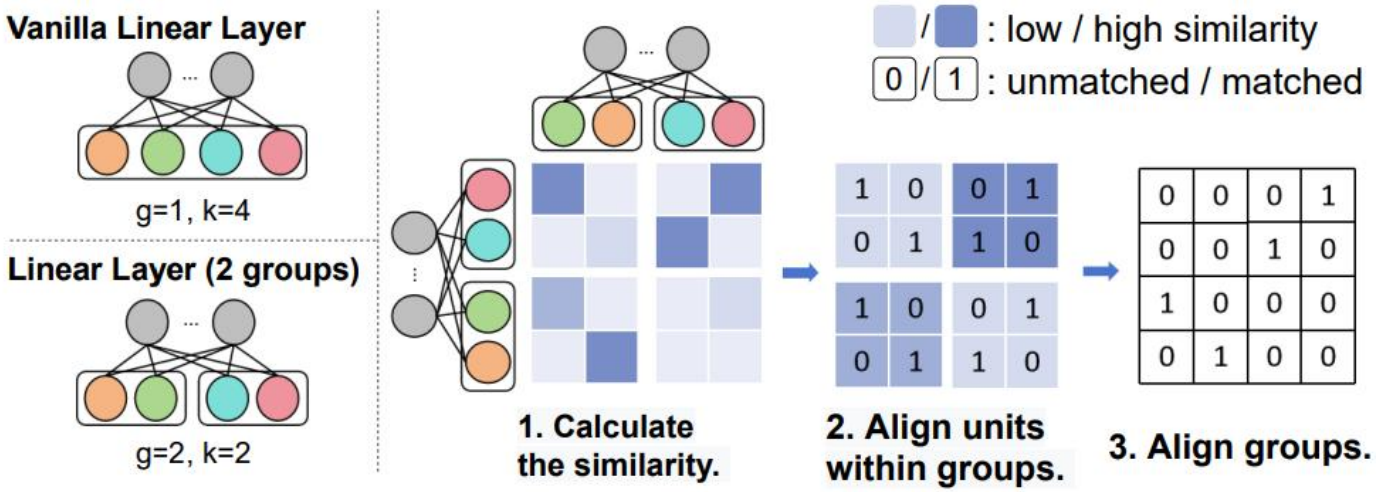
for $l = 1, 2, \ldots, L$ **do**
  $\boldsymbol{C}_l' \leftarrow \mathbb{C}(\boldsymbol{A}_l)$
  $\boldsymbol{P}_l \leftarrow \Psi(\boldsymbol{C}_l')$
end for
repeat
  for $l = RP(1, 2, \ldots, L)$ **do**
    $\boldsymbol{Z}_l \leftarrow \upsilon(\boldsymbol{P}_{l-1}, \boldsymbol{P}_{l+1}, \boldsymbol{W}_l, \boldsymbol{W}_{l+1})$
    $\boldsymbol{C}_l' \leftarrow \alpha\mathbb{C}(\boldsymbol{Z}_l) + (1-\alpha)\boldsymbol{C}_l'$
    $\boldsymbol{P}_l \leftarrow \Psi(\boldsymbol{C}_l')$
  end for
until convergence
Get the merged weights $\{\boldsymbol{W}_l'\}_{l=1}^L$ by Eq. (2)
return $\{\boldsymbol{W}_l'\}_{l=1}^L$

Method Group Alignment for Group Normalization and Multi-Head Attention

1. Calculate the similarity between units.

2. compute permutation and then calculate the average of matched similarity within each pairs of groups.

3. compute permutation for each pairs of groups and then set the permutation of unmatched pairs to zeros.



Experiments

Merging Models of Homogeneous Tasks

Merging models from random initialization

Model	Resnet20								Resnet20GN			
Dataset	CIFAR100(50+50)				CIFAR10(5+5)				CIFAR100(50+50)			
Method	Joint	Avg	T. A	T. B	Joint	Avg	T. A	T. B	Joint	Avg	T. A	T. B
Average	16.52	24.22	23.22	25.21	54.42	75.24	79.58	70.90	5.63	11.06	9.67	12.44
Rebasin	41.33	56.94	57.31	56.58	60.61	88.57	88.46	88.68	13.85	22.18	22.99	21.37
A. Align	44.33	61.13	61.61	60.66	**61.71**	89.21	88.63	**89.78**	29.37	42.05	41.05	43.05
MuDSC$_{Align}$	**45.50**	**62.81**	**63.06**	**62.56**	60.84	**89.34**	**89.04**	89.63	**31.84**	**45.31**	**45.34**	**45.29**
Zipit	54.69	66.78	67.11	66.44	82.44	94.61	94.22	95.00	29.93	41.20	39.99	42.41
W.Zip	55.16	67.65	68.58	66.71	82.85	94.71	94.42	94.99	14.28	20.95	19.17	22.72
MuDSC$_{Zip}$	**56.01**	**68.13**	**68.80**	**67.47**	**83.09**	**94.88**	**94.56**	**95.21**	**30.05**	**41.52**	**40.39**	**42.65**

Merging models from same pretrained model

Model	Resnet26				Resnet50GN				ViT			
Method	Joint	Avg	T. A	T. B	Joint	Avg	T. A	T. B	Joint	Avg	T. A	T. B
Average	61.44	74.75	74.46	75.05	74.52	84.78	85.06	84.50	70.16	84.32	84.32	84.32
Rebasin	61.39	74.79	74.48	75.10	74.52	84.78	85.06	84.50	**70.16**	84.32	84.32	84.32
A. Align	61.91	75.41	75.03	75.79	74.44	84.77	84.99	84.56	69.99	84.22	84.20	84.24
MuDSC$_{Align}$	**62.84**	**76.14**	**75.87**	**76.40**	**74.66**	**84.91**	**85.25**	**84.58**	70.09	**84.39**	**84.38**	**84.40**
Zipit	60.23	73.68	73.20	74.17	72.05	82.99	83.06	82.92	68.57	83.05	82.79	83.30
W.Zip	61.28	74.69	74.42	74.96	74.52	84.78	85.06	84.50	**70.16**	84.32	84.32	84.32
MuDSC$_{Zip}$	**61.58**	**75.01**	**74.61**	**75.41**	**74.71**	**84.88**	**85.14**	**84.62**	70.10	**84.38**	**84.41**	**84.36**

Merging models on Imagenet (Resnet50)

Method	Joint Acc	Avg Acc	T. A	T. B
Average	44.85	62.31	61.93	62.68
Rebasin	44.85	62.31	61.93	62.68
A. Align	44.86	62.62	62.56	62.67
MuDSC$_{Align}$	**44.87**	**62.66**	**62.57**	**62.74**
Zipit	44.22	62.25	62.23	62.26
W.Zip	44.85	62.31	61.93	62.68
MuDSC$_{Zip}$	**44.86**	**62.59**	**62.45**	**62.72**

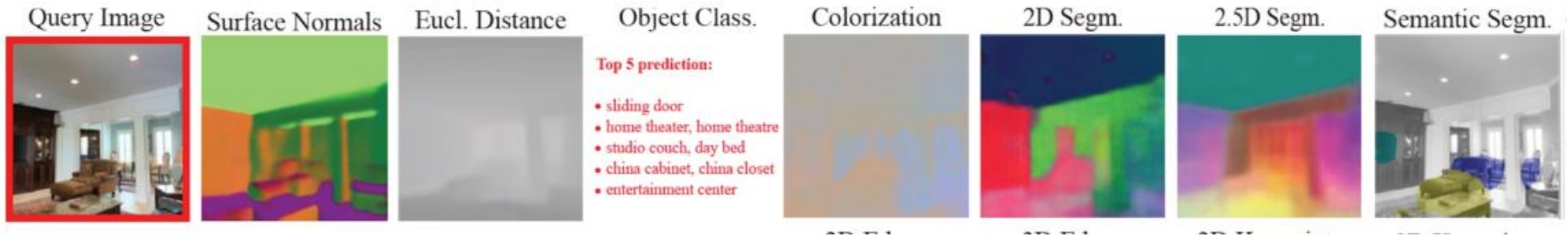
- Ours MuDSC methods bring significant improvement in manifold multi-task scenarios

Experiments

Merging Models of Heterogenous Tasks

Taskonomy Dataset

- 12 Task&Pretrained model
- Merging pairwise



Scaled Performence

$$\mathcal{L}_{SP} = \frac{\mathcal{L}_\theta - \mathcal{L}_0}{\mathcal{L}_1 - \mathcal{L}_0},$$

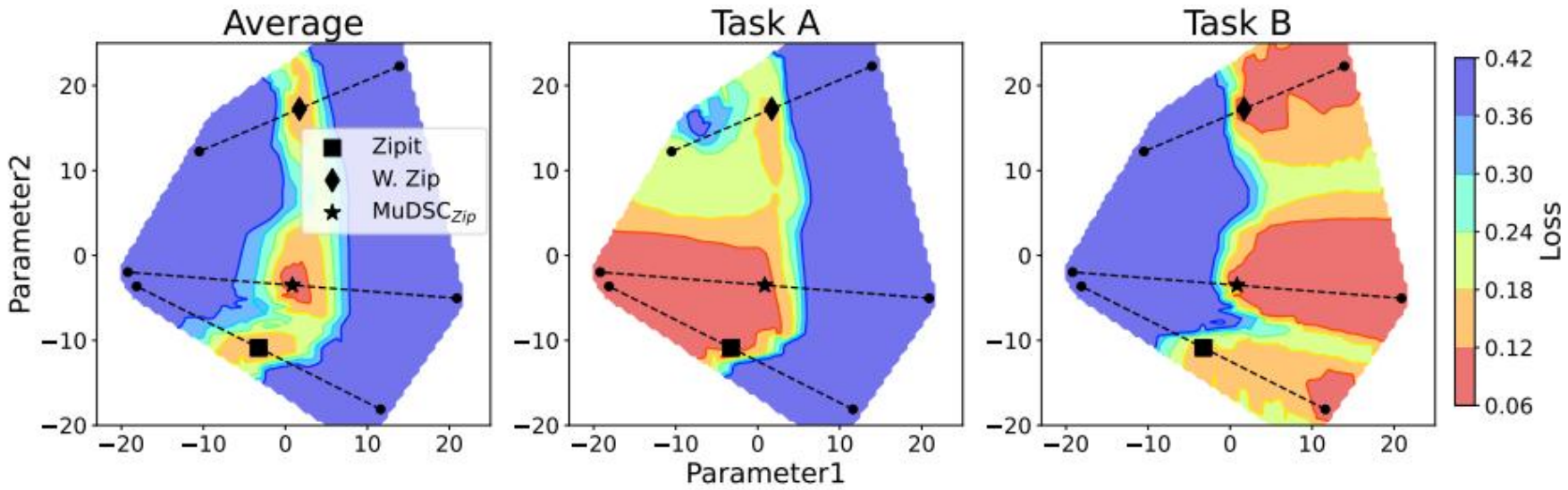
$\mathcal{L}_1$: Original loss

$\mathcal{L}_0$: loss of average estimator

Visual Task	Method						
	Weight Average	Rebasin	Act. Align	MuDSC$_{Align}$	Zipit	Weight Zip	MuDSC$_{Zip}$
Class Object	76.15	80.76	**89.75**	**89.75** +0.00	84.84	80.93	**86.04** +1.20
Segment Semantic	23.01	30.59	52.54	**55.24** +2.69	32.51	33.77	**36.22** +2.45
Rgb2depth	95.42	95.90	98.69	**98.70** +0.01	99.30	99.07	**99.35** +0.05
Rgb2mist	94.56	95.00	**98.28**	**98.28** +0.00	99.08	98.59	**99.15** +0.00
Edge3D	79.79	80.44	**91.04**	90.77 −0.27	**93.22**	86.42	92.81 −0.40
Edge2D	68.54	75.48	**81.37**	81.24 −0.12	90.39	88.85	**93.09** +2.69
Keypoints2D	71.33	75.89	81.99	**82.30** +0.31	93.38	93.62	**94.73** +1.11
Keypoints3D	91.54	91.99	96.16	**96.17** +0.02	**96.99**	96.00	96.95 −0.04
Reshading	-4.72	8.52	60.23	**61.85** +1.61	**69.18**	41.66	68.71 −0.47
Rgb2sfnorm	26.62	29.43	65.46	**65.56** +0.09	76.62	64.43	**77.50** +0.88
Autoencoding	21.96	35.65	51.84	**54.21** +2.37	86.55	77.58	**87.99** +1.45
Denoising	33.92	33.57	52.37	**54.40** +2.03	84.28	76.14	**85.49** +1.22
Total Average	56.51	61.10	76.64	**77.37** +0.73	83.86	78.09	**84.84** +0.98

Experiments

Loss Landscape Visualization



VIPA
Visual Intelligence and Pattern Analysis

Training Free Pretrained Model Merging

Zhengqi Xu, Ke Yuan, Huiqiong Wang, Yong Wang, Mingli Song, Jie Song

Paper: https://arxiv.org/abs/2403.01753
Code: https://github.com/zju-vipa/training_free_model_merging