Transporting task vectors across different architectures

Filippo Rinaldi, Aniello Panariello, Giacomo Salici, Angelo Porrello, Simone Calderara

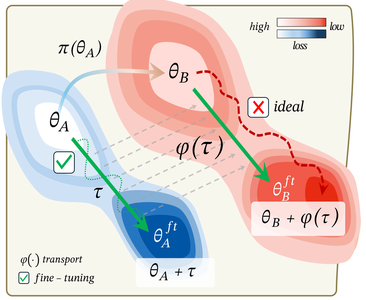
AImageLab UNIMORE Università degli Studi di Modena e Reggio Emilia

ICML International Conference On Machine Learning

TL;DR: THESEUS transfers fine-tunings between **different architectures** by aligning their representation spaces.

Problem

- Fine-tuning is costly.
- Existing task-vector transfer works only for identical architectures.
- Different model widths make parameter spaces incompatible.



Background

Rebasin: Transporting the previous model's task vector ($\tau_A = \theta_A^{ft} - \theta_A$) recycles past adaptation without training from scratch.

The Transfer Failure: Naive transfer across different pre-trained models fails. Misaligned parameter spaces **introduce harmful directions.**

Previous Work: *TransFusion* **aligns models** via weight permutations; *Gradfix*, computes the gradient of the target model.

The Limitation: Transported vectors **are not guaranteed** to align with the new model's local loss geometry.

Method

training-free · closed-form · no backprop · reduced memory*

How to compute the transport update τ_B ?

Compute activations cross-covariances:

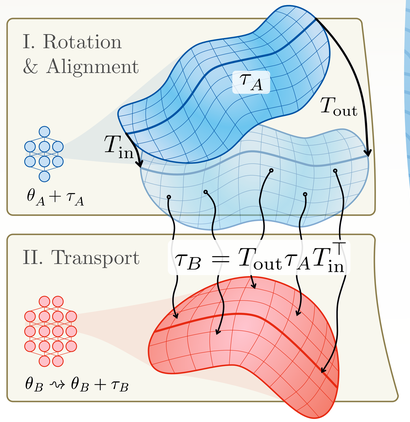
$$C_{in} \leftarrow H_{in,A}^\top H_{in,B}$$
$$C_{out} \leftarrow H_{out,A}^\top H_{out,B} \quad [1]$$

Compute SVD:

$$C_{in} = U_{in}\Sigma_{in}V_{in}^\top$$
$$C_{out} = U_{out}\Sigma_{out}V_{out}^\top \quad [2]$$

Solve orthogonal Procrustes:

$$T_{in} \leftarrow U_{in}V_{in}^\top$$
$$T_{out} \leftarrow U_{out}V_{out}^\top \quad [3]$$



*thanks to the incremental covariance accumulation: $O(d_A \times d_B)$ memory

Experiments

Width scaling (ViT-B/16 → ViT-B/16+)

Vision Classification (Accuracy)

Model	$\mathcal{B}$	EUROSAT	SVHN	GTSRB	RESISC45	DTD	CARS	MNIST	SUN397	AVG (ΔAcc)
θ_B *zero-shot*	–	50.92	39.23	49.63	64.53	55.48	84.53	57.06	68.67	58.76 (+0.00)
θ_B *fine-tune*	–	98.96	91.08	98.63	92.59	77.81	87.65	99.63	76.76	90.39 (+31.63)
θ_A *fine-tune*	–	98.69	97.45	98.64	95.65	82.24	91.53	99.61	79.86	92.96 (+34.20)
$\theta_B + \tau_A^{\text{random}\to B}$	2	50.96	39.41	49.75	65.60	55.48	84.41	56.62	69.51	58.97 (+0.21)
THESEUS	2	64.03	59.30	59.55	68.12	57.86	84.89	68.97	70.12	**66.60** (+7.84)
THESEUS	10	65.91	61.59	61.23	68.73	59.52	84.93	74.46	70.42	**68.35** (+9.59)
THESEUS	20	67.01	62.52	61.01	69.27	60.48	85.04	76.66	70.65	**69.08** (+10.32)
THESEUS	50	67.13	64.29	59.46	70.01	61.28	85.19	77.23	71.28	**69.48** (+10.72)

Depth and width scaling (ViT-B/16 → ViT-L/14)

Vision Classification (Accuracy)

Model	$\mathcal{B}$	EUROSAT	SVHN	GTSRB	RESISC45	DTD	CARS	MNIST	SUN397	AVG (ΔAcc)
θ_A *interpolated*	–	16.04	18.37	6.44	4.78	3.99	0.72	16.58	1.07	8.50 (−65.13)
θ_B *zero-shot*	–	68.40	67.44	58.96	71.27	66.81	93.01	86.84	76.34	73.63 (+0.00)
θ_A *fine-tune*	–	98.66	97.70	98.77	95.40	83.29	92.43	99.64	79.91	93.23 (+19.60)
θ_B *fine-tune*	–	99.00	98.22	98.42	96.73	86.88	94.80	99.76	82.94	94.59 (+20.96)
THESEUS	10	76.78	78.34	73.59	75.02	72.71	93.10	95.18	77.16	**80.24** (+6.61)

Width scaling (T5-Flan → T5-v1.1)

Text Summarization (ROUGE Score)

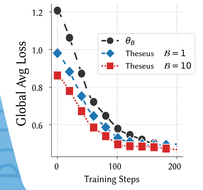
Configuration	α	$\mathcal{D}=200$	$\mathcal{D}=500$	$\mathcal{D}=800$	Average
θ_B	–		0.1053		
θ_B *fine-tune*	–		0.2821		
THESEUS	1.0	0.1278	0.1337	0.1437	0.1351
THESEUS	2.0	0.1237	0.1189	0.1423	0.1283
THESEUS	4.0	0.0876	0.0512	0.0879	0.0756

Width scaling (T5-3B → T5-Lg)

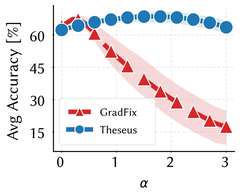
Text Classification, linear probing (Accuracy)

Encoder	$\mathcal{B}$	MNLI	QNLI	RTE	SCITAIL	SNLI	AVG (ΔAcc)
θ_A *fine-tune*	All	89.88	95.78	87.73	92.45	90.45	91.26 (-)
θ_B *fine-tune*	All	86.34	92.14	81.23	91.12	88.75	87.92 (-)
θ_A	20	66.84	80.85	71.84	69.40	45.06	66.80 (+8.39)
θ_B	20	64.87	66.61	64.62	49.85	43.42	57.87 (+0.00)
THESEUS	20	78.67	86.09	74.00	84.74	76.04	**79.91** (+22.04)

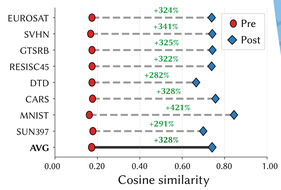
Earlier convergence during fine-tuning



Robust to update scaling (α)



Higher activation cosine similarity



The Theseus Paradox: If you replace every plank of a ship, is it still the same ship? **We replaced every parameter;** the task was preserved.

Paper: Code: