Orthogonal Re-Basin: Beyond Permutations in Model Merging

August 29, 2025

Rosamaria Graziosi

Abstract

The Git Re-Basin framework demonstrates that independently trained neural networks can be aligned by permuting hidden units, enabling model merging. However, this method is restricted to permutations, a small subgroup of the orthogonal group. This project has the aim of relaxing this restriction, aligning models via general orthogonal transformations. Before doing so the weight structures are compared through Procrustes analysis and CKA similarity, and the effects of non-linearities are observed on cycle-consistency.

GitHub repo with code and further data: https://github.com/WrongMedal/ML_proj_Orthogonal_Re-Basin/tree/main

1. Introduction

Ensembling and merging independently trained models is an effective strategy in deep learning. A central challenge is aligning neurons across models due to weight permutation symmetries. In this work, we go beyond permutations and explore the use of orthogonal matrices to align models, introducing *Orthogonal Re-Basin*. This allows rotations and reflections of hidden representations, potentially capturing richer symmetries. We study this phenomenon on a MLP trained on MNIST and on a CNN trained on CIFAR10.

2. Related Work

Permutation symmetries. Neural networks are invariant to hidden unit permutations (Frankle et al., 2020). Git Re-Basin (Ainsworth et al., 2022) leveraged this property for model alignment, without loss in accuracy.

Email: Rosamaria Graziosi <graziosi.2083375@studenti.uniroma1.it>.

Machine Learning 2025, Sapienza University of Rome, 2nd semester a.y. 2024/2025.

Mode connectivity. Linear and non-linear interpolation between models has been studied in (Garipov et al., 2018; Frankle et al., 2020), showing that many minima are connected via low-loss paths.

Representation similarity. Metrics such as Centered Kernel Alignment (CKA) (Davari et al., 2023; Kornblith et al., 2019) are widely used to quantify alignment between representations.

3. Method

We represent model parameters as W_1, W_2 and seek an orthogonal matrix $Q \in O(n)$ such that

$$\min_{Q \in O(n)} \|W_1 - QW_2\|_F^2. \tag{1}$$

This is the classical orthogonal Procrustes problem, solved by singular value decomposition (SVD). Let $W_1^\top W_2 = U\Sigma V^\top$, then the optimal transformation is

$$Q^\star = UV^\top. \tag{2}$$

This idea is used first neuron-wise but yields bad results. Therefore we also explore layer-wise alignment and cycle consistency constraints to enforce global coherence, especially since the CKA results are encouraging.

CKA - Centered Kernel Alignment - is a widely used approach to quantify similarities of internal representations, that takes into consideration the activations functions of a model, introduced by Kornblith.

Let $X \in \mathbb{R}^{n \times d_x}$ and $Y \in \mathbb{R}^{n \times d_y}$ denote the activation matrices of two layers (or two models), where n is the number of examples and d_x, d_y are the respective representation dimensions. Each row corresponds to the activations for one input sample.

CKA is based on the Hilbert–Schmidt Independence Criterion (HSIC) between the Gram (kernel) matrices of X and Y. For the commonly used linear kernel, the similarity is defined as

$$\mathrm{CKA}(X, Y) = \frac{\|Y^\top X\|_F^2}{\|X^\top X\|_F \, \|Y^\top Y\|_F}, \tag{3}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

This measure has several desirable properties:

- **Invariant to isotropic scaling:** multiplying activations by a scalar does not affect the similarity.
- **Invariant to orthogonal transformations:** rotating the feature space leaves the measure unchanged, which is very useful in our case.
- **Boundedness:** $\text{CKA}(X, Y) \in [0, 1]$, where values closer to 1 indicate more similar representations.

4. Results

4.1. MLP on MNIST

- **Procrustes on neurons**. Disparities were high across all the neurons: a one by one match with and orthogonal transformation would not have worked.
- **CKA (Centered Kernel Alignment)**. The CKA values were high, suggesting that, despite these disparities in weights, the models are functionally very similar in their internal representations.
- **Orthogonal weight alignment / Procrustes on layers**. Layer-wise disparities and cosine similarity were evaluated after aligning the layers. Results showed large disparities (Table 1) and low cosine similarity (~ 0.18), indicating that these operations are not sufficient to align activations. This misalignment is likely due to nonlinearities (non incorporated in the orthogonal transformations) and possibly subsequent layers like BatchNorm.

Table 1. MLP layer disparities after orthogonal alignment.

Layer	Disparity
Linear 0	23.59
Linear 1	9.15
Linear 2	7.90
Linear 3	0.48
Mean	10.28

This may also explain these (Fig1) cycle consistency results.

4.2. CNN on CIFAR10

Similar results with the CNNs showed reduced disparity and high CKA alignment (0.76–0.99). A big difference with the MLP models is seen in the orthogonal weight alignment: the deeper the layer, the higher is the disparity, except for the last layer that classifies. This last layer behavior checks out with the fact that both models reach about the same accuracy.

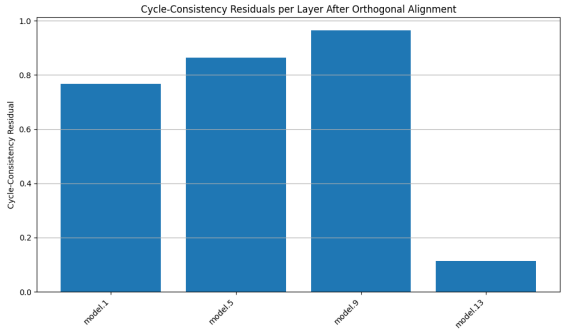

Figure 1. Cycle consistency across layers for MLP models.

4.3. Interpolation

We also tried to interpolate between aligned models using both linear interpolation and spherical linear interpolation (SLERP) but with no success. As seen in (Fig 2), the accuracy as we tend towards the other model doesn't rise again.

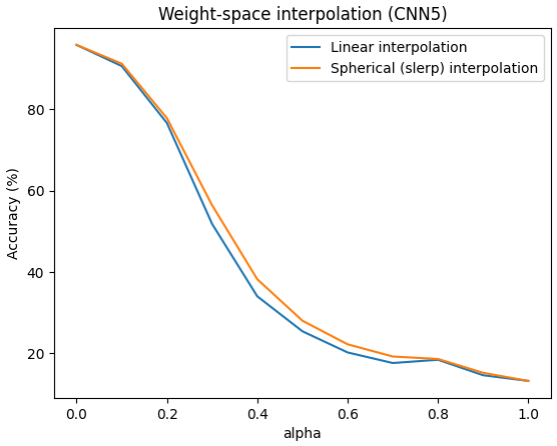

Figure 2. Cycle consistency across layers for CNN models.

References

Ainsworth, S., Hayase, J., and Srinivasa, S. Git re-basin: Merging models modulo permutation symmetries. *arXiv preprint arXiv:2209.04836*, 2022.

Davari, M., Horoi, S., Natik, A., Lajoie, G., Wolf, G., and Belilovsky, E. Reliability of cka as a similarity measure in deep learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. OpenReview preprint.

Frankle, J., Dziugaite, G. K., Roy, D. M., and Carbin, M. Linear mode connectivity and the lottery ticket hypothesis. *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D., and Wilson, A. G. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Advances in Neural Information Processing Systems*, 2018.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, 2019.