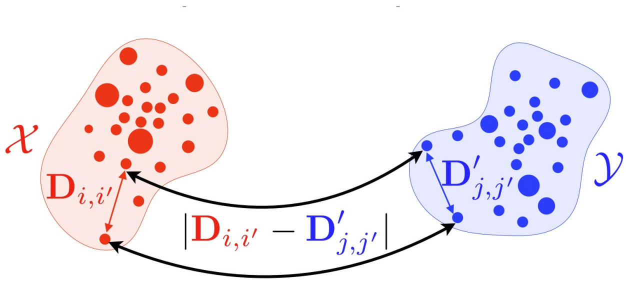
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Prof. R. Wattenhofer

Optimal Transport for Federated Learning: Bridging Semantics and Alignment through Mechanistic Interpretability



Federated learning is a machine-learning approach where multiple devices or organizations collaboratively train a shared model without sharing their raw data. Each participant trains the model locally and sends only information about the model to a central server for aggregation, which enhances privacy and reduces data transfer. A key challenge in this area is effectively combining models. In simpler terms, imagine we have two models and want to optimally merge them into a single, unified model.

Several strategies are currently used to combine models in federated or distributed settings. A common one is parameter averaging, where model weights or gradients are aggregated into a single global model. Another approach uses ensemble methods, where participating models remain separate, and their outputs are combined without merging their parameters into a shared space. Some other approaches rely on knowledge distillation, where each model's predictions on a shared public or proxy dataset are used to train a new unified model. However, these methods each have limitations: parameter averaging can fail in the presence of heterogeneous data, ensembles are more computationally intensive at inference time and do not produce a single compact model, and distillation requires suitable public data and the training of a new model.

In this project, we will focus on parameter averaging, based on recently proposed optimal transport–based alignment methods [3]. Such methods aim to address the issue of the missing one-to-one correspondence between the weights of two different neural networks, even when they have the same architecture and initialization, due to heterogeneous training data. They fuse models by first matching neurons in each layer using optimal transport (OT) to find the lowest-cost pairing based on activation or weight similarity. After aligning neurons, corresponding weights are averaged, effectively forming a Wasserstein barycenter of the models' layer-wise weight distributions. This represents a significant improvement over vanilla averaging, where weights are simply averaged based on their position rather than the semantics of the connected neurons. The goal of this project is to build on inspiration from mechanistic interpretability to find weaknesses from current OT methods and propose improvements.

One key insight is that optimal transport aligns neurons with similar semantics. However, in mechanistic interpretability, it has been shown that features, which represent semantically meaningful concepts, can be encoded across multiple neurons (see [4] as discussed in section 3.1). Therefore, neurons can be polysemantic [2], meaning they are part of the encoding of multiple concepts rather than having a single, concrete semantic meaning. Given this, it should be researched both theoretically and empirically how features encoded in different bases than the neuron basis in the hidden space can harm the proposed optimal transport methods. These insights may contribute to methods for aligning features rather than neurons for more effective weight averaging. Sparse autoencoders[1][1] could be used to find features for alignments.

Interested? Contact us for more information

- Denis Sutter: densutter@ethz.ch, ETZ G97
- Dr. Antonio Di Maio: adimaio@ethz.ch, ETZ G97

Last update: December 2025

References

[1] Hoagy Cunningham et al. *Sparse Autoencoders Find Highly Interpretable Features in Language Models*. 2023. arXiv: 2309.08600 [cs.LG]. URL: https://arxiv.org/abs/2309.08600.

[2] Adam Scherlis et al. *Polysemanticity and Capacity in Neural Networks*. 2025. arXiv: 2210.01892 [cs.NE]. URL: https://arxiv.org/abs/2210.01892.

[3] Sidak Pal Singh and Martin Jaggi. *Model Fusion via Optimal Transport*. 2023. arXiv: 1910.05653 [cs.LG]. URL: https://arxiv.org/abs/1910.05653.

[4] Denis Sutter et al. *The Non-Linear Representation Dilemma: Is Causal Abstraction Enough for Mechanistic Interpretability?* 2025. arXiv: 2507.08802 [cs.LG]. URL: https://arxiv.org/abs/2507.08802.

[1]https://transformer-circuits.pub/2023/monosemantic-features/index.html