Structure Is Not Enough: Leveraging Behavior for Neural Network Weight Reconstruction

Léo Meynent[1], Ivan Melev[2], Konstantin Schürholt[1], Göran Kauermann[2], Damian Borth[1]

[1] School of Computer Science, University of St.Gallen, Switzerland

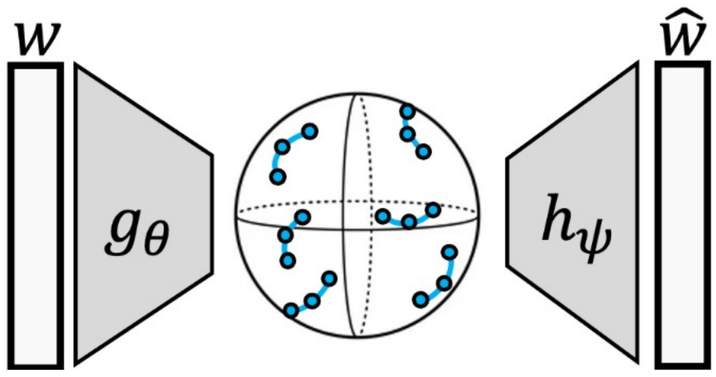
[2] Department of Statistics, Ludwig-Maximilians-University Munich, Germany

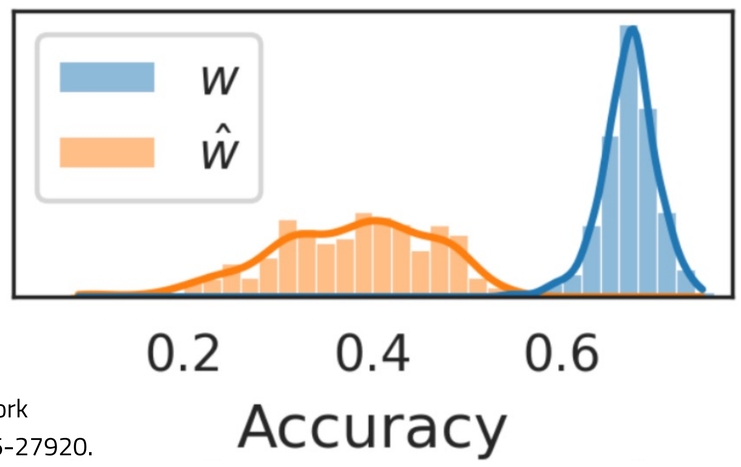
ICLR International Conference On Learning Representations

Workshop on Neural Network Weights as a New Data Modality

Motivation

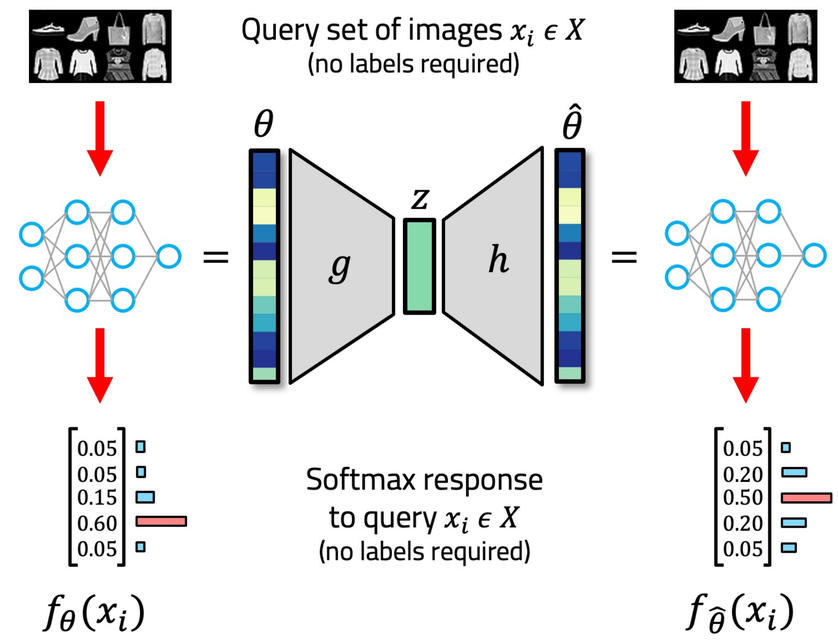
Weight-space autoencoders fail to reconstruct model weights that match the performance of the original ones.





Schürholt, K., Knyazev, B., Giró-i-Nieto, X., & Borth, D. (2022). Hyper-representations as generative models: Sampling unseen neural network weights. *Advances in Neural Information Processing Systems*, 35, 27906-27920.

Behavioral Loss Function



Experimental setup

Loss functions

- $\mathcal{L}_C$: Contrastive loss
- $\mathcal{L}_S$: Structural loss (L^2 reconstruction)
- $\mathcal{L}_B$: Behavioral loss

Model zoos: CNNs, 1'200/zoo

SVHN | CIFAR-10 | EuroSAT

Gradient Analysis

$$\mathcal{L}_B = \frac{1}{2kn} \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| f_{\hat{\theta}_j}(x_i) - f_{\theta_j}(x_i) \right\|^2$$

Queries; Reconstructed weights; Original weights

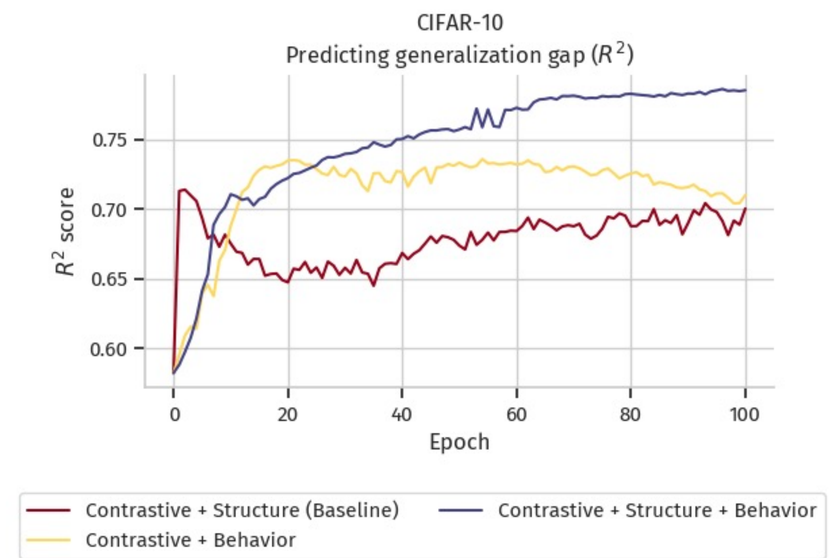
$$\frac{\partial \mathcal{L}_S}{\partial w} = \frac{1}{k} \sum_{j=1}^{k} \Delta\theta_j^\top \frac{\partial \hat{\theta}_j}{\partial w} \qquad \frac{\partial \mathcal{L}_B}{\partial w} \approx \frac{1}{k} \sum_{j=1}^{k} \Delta\theta_j^\top F_j \frac{\partial \hat{\theta}_j}{\partial w}$$

Structural loss gradient | Behavioral loss gradient

$$F_j = \frac{1}{n} \sum_{i=1}^{n} J_{\theta_j}(x_i)^\top J_{\hat{\theta}_j}(x_i)$$

modulates the loss with the original and reconstructed models' sensitivity to changes in the weights

Discriminative Downstream Tasks

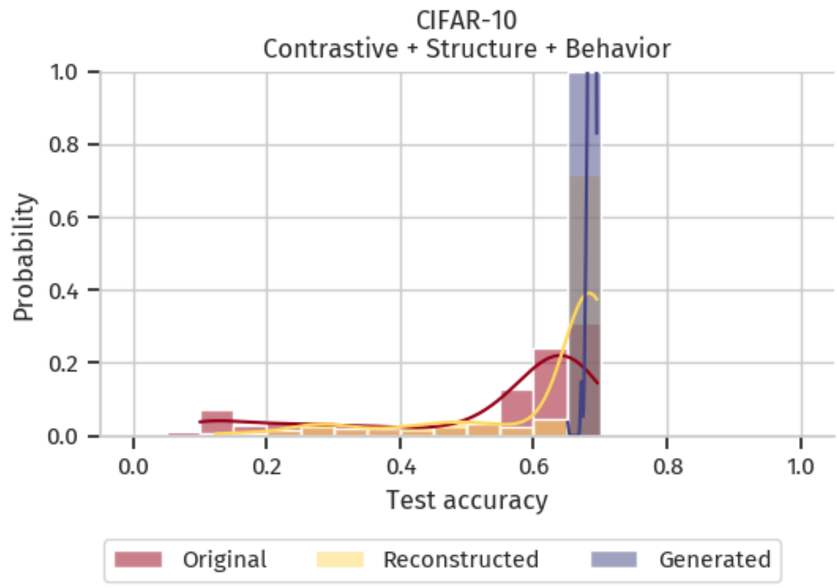


Reconstructive Downstream Tasks

We compare the pairwise distance between the original model and its reconstruction, both in terms of **structure** (L^2 distance between weights) and **behavior** (model agreement)

Generative Downstream Tasks

We look into the distribution of model test accuracies for the **original** model zoos, their respective **reconstructions**, and model weights **generated** based on a KDE of high-performing models' embeddings.



Contribution

- Analysis of the importance of the behavioral loss
- Uncovering of a synergy between structural and behavioral signals
- Improved performance on discriminative downstream tasks
- Strongly improved performance on reconstructive and generative downstream tasks

Further Material

Paper:
arxiv.org/abs/2503.17138
Code:
github.com/HSG-AIML/ICLR_WSL_2025-Structure_is_not_enough