Stellenbosch
UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT

A Contrastive Approach to Weight Space Learning

by

D.J. Swanevelder

MSc Machine Learning/ Artificial Intelligence

Department of Applied Mathematics

Faculty of Science

Stellenbosch University

Supervisor: Ruan van der Merwe

November 2025

Contents

Chapter 1

Introduction

> ”We don’t tell [computers] what to do, we give them examples... The problem is, sometimes we don’t understand how it figured it out.”
>
> – Jeff Dean, Head of Google AI [1]

The prevalence of neural networks (NNs) has positioned them as a foundational technology in modern artificial intelligence. However, as their use has grown, so too has the focus on their inherent mechanistic limitations. Two of the most significant drawbacks of NNs are their lack of explainability—the ”black-box” effect—and the immense computational cost required for training. A novel field, weight space learning, emerges as an field showing promise to further our understanding of these very challenges.

Weight space learning primarily focuses on developing methods to represent the high-dimensional, weights of NN models in a lower-dimensional latent space, in order to facilitate further exploration of the weight space. The field generally considers two types of tasks: discriminative and generative.

In the discriminative application, models use the weights of already-trained models, with a specific collection of models reffered to as a model zoo, as input to accurately predict meta-information about the original model. The quality of weight space representation is often quantitatively measured by a simple Multi-layer perceptron (MLP)’s ability to predict this meta-information, conditioned only on the model’s weight space representation.

Common meta-information metrics investigated include predicting the model’s final performance and the generalization gap (the difference between training and validation loss). For instance, the Sequential Autoencoder for Neural Embeddings (SANE) [2] demonstrated strong performance in meta-prediction tasks. Furthermore, research has shown that weight embeddings can encode fundamental training characteristics, such as accurately recovering the size of the dataset on which the model was originally trained [3]. These discriminative tasks serve the dual purpose of both validating the quality of the derived weight representations and possessing significant practical value.

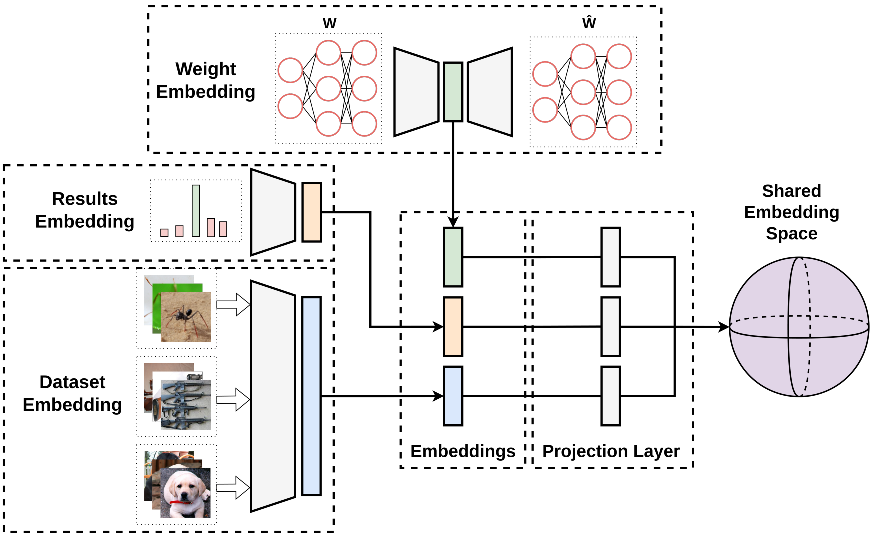

Figure 1.1: A representation of the process of embedding a dataset, model weights and results in a shared embedding space. A dataset embedded using a pretrained CLIP encoder, averaging over all images. Model weights embedded using a PCA enhanced autoregressive encoder. Model results embeddings and the shared embedding space created through contrastive learning

In the generative application, researchers leverage the structure of the latent weight space to create new, functional models. For example, [2] (SANE) makes use of an autoencoder for sequentially encoding and decoding models, with a bottleneck layer generating the latent vector, and then a decoder to generate a full model. Another paper uses a Vector Quantized Variational Autoencoder (VQ-VAE), which takes in information about the dataset as part of the process of find latent weight representations, with the ultimate going of instruction-guided paramter generation [4]. [5] makes use of information on the model's performance and behavior in generating the model embeddings. With $P(W \mid \dots)$ being the distribution sampled from to generate new model weights, currently no existing paper conditions on both dataset information $(\mathcal{D})$ and model performance (R) simultaneously, modeling the conditional distribution:

$$P(W \mid \mathcal{D}, R)$$

Problem Statement: To what extent can we develop contrastive methods to project model input and model predictive performance to a latent weight space, thereby accurately modeling the complex relationship between a model's weights, its inputs, and its resulting behavior?

The focus of this report is to develop a extend latent weight representations into a shared, multimodal representation space. Figure 1.1 outlines this proposed pipeline.

The results found from the report, and how it's limited

The remainder of this report is structured as follows.:

In Section 1 we will discuss x, next y, concluded by z in attempt to showcase that this report is the best

Bibliography

[1] J. Dean, "Keynote speech on the future of ai," https://www.google.com/search?q=https://events.google.com/io/, 2017, statement widely attributed to the keynote speech, highlighting the challenge of interpretability in deep learning systems.

[2] K. Schürholt, M. W. Mahoney, and D. Borth, "Towards scalable and versatile weight space learning," 2024. [Online]. Available: https://arxiv.org/abs/2406.09997

[3] M. Salama, J. Kahana, E. Horwitz, and Y. Hoshen, "Dataset size recovery from lora weights," 2024. [Online]. Available: https://arxiv.org/abs/2406.19395

[4] S. Bedionita, B. Andreis, S. Chong, and S. J. Hwang, "Instruction-guided autoregressive neural network parameter generation," 2025. [Online]. Available: https://arxiv.org/abs/2504.02012

[5] L. Meynent, I. Melev, K. Schürholt, G. Kauermann, and D. Borth, "Structure is not enough: Leveraging behavior for neural network weight reconstruction," 2025. [Online]. Available: https://arxiv.org/abs/2503.17138