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

1. Introduction 1

Bibliography 4

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. With platforms like Hugging Face and GitHub providing access to over one million pre-trained models [2] interest in these limitations has grown. Fueled by this enormous availability of models, a novel field, weight space learning, emerges as a promising area to further our understanding of these mechanistic limitations.

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 referred to as a model zoo [3], as input to accurately predict meta-information about the original model [4]. 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). [5] 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 . These discriminative tasks serve the dual purpose of both validating the quality of the derived weight representations and possessing significant practical value.

In the generative application, researchers make use of varous methods to model the underlying distribution of NN weights W conditioned on additional information or reference models $P(W \mid \ldots)$. The process of sampling from this distruibution is what enables eintrely new weight generation

[6] made use of an autoencoder with a bottleneck layer to generate a hyper-representations of various model zoo's. Expanding on this concept [7] makes use of a Sequential Autoencoder for Neural Embeddings (SANE) to improve the scaliablity of the method, allowing work on much larger models.

In order to model the distribution $P(W \mid \mathcal{D})$, [8] use a Vector Quantized Variational Autoencoder (VQ-VAE), which takes in information about the dataset as part of the process of find latent weight representations. [9] models $P(W \mid R)$ through including the difference in behaviour between the reconstructed and original models in the process of learning model embeddings.

While existing weight space learning methods have made significant progress, they typically focus on isolated aspects of the model learning process. Current approaches model either $P(W \mid \mathcal{D})$ or $P(W \mid R)$, but not both simultaneously. This represents a significant limitation: in practice, a model's weights are shaped by both the data it was trained and the results it achieved when training on a certain objective. Understanding the more complex joint relationship — $P(W \mid D, R)$ — is essential for both explaining model behavior and for generating new models with desired characteristics.

Contrastive learning has emerged as a powerful paradigm for learning unified representations across different modalities, as demonstrated by the success of models like CLIP [10] in bridging vision and language. A learned embedding space is learned through the use of a loss objective function which pulls related samples together in representation space, and pushes unrelated samples away from each other. This allows for the formation of a meaningful represnetation space, of heretogenous data types. It is this feaute of contrasitve learning which makes it particularly usefull to model the complex distribution $P(W \mid D, R)$ — visual datasets, high-dimensional weight tensors, and performance metrics — without requiring them to share the same native representation.

In this we report we develop a contrastive learning framework to create a unified embedding space that jointly represents neural network weights (W), the datasets they were trained on $(\mathcal{D})$, and their resulting performance characteristics (R). Specifically, we construct two separate encoders—one for dataset embeddings using pre-trained CLIP features and one for weight embeddings using an autoencoder [cite] architecture, while creating a binned result embedding table, and train these encoders using a contrastive objective, NXTEnt [11], that encourages related triplets $(\mathcal{D}, W, R)$ to be proximal in the shared latent space. Figure 1.1 depicts a high-level view of the full embedding pipeline.

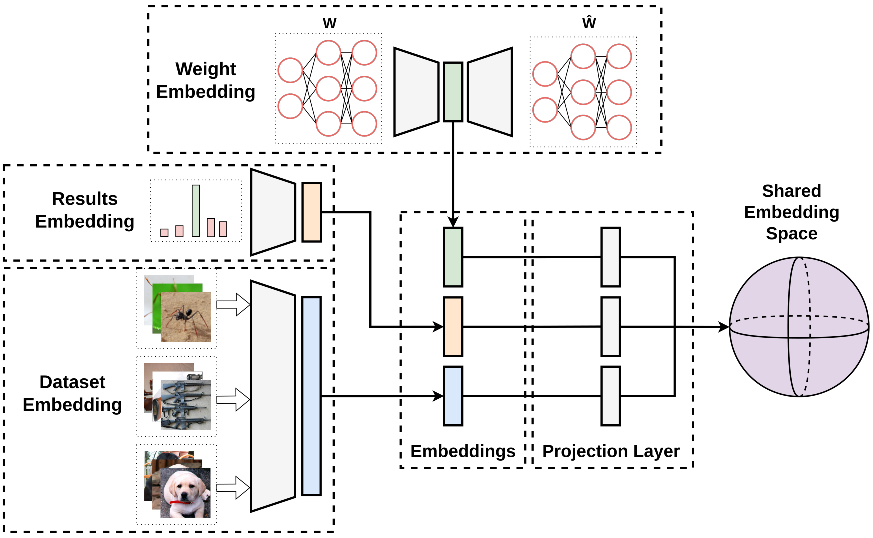

Figure 1.1: AA figure illustrating the process of embedding a dataset, model weights and results in a shared embedding space.

Our central hypothesis is that this unified representation space will enable two key capabilities:

Interpretability and Analysis: By examining the geometric relationships in the shared embedding space, we can better understand how dataset characteristics influence the learned weights and resulting model behaviors. This includes investigating questions such as: Do models trained on similar datasets cluster together in weight space? Can we identify which data characteristics most strongly determine weight patterns?

Conditional Model Sampling: The learned distribution can enable sampling of model weights conditioned on both desired dataset properties and target performance metrics—that is, approximating P(W—D,R)—which could allow for more efficient model generation than training from scratch.

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] Hugging Face, "Open-source AI: Year in Review 2024," https://huggingface.co/spaces/huggingface/open-source-ai-year-in-review-2024, 2024, accessed: 14 October 2025.

[3] K. Schürholt, D. Taskiran, B. Knyazev, X. G. i Nieto, and D. Borth, "Model zoos: A dataset of diverse populations of neural network models," in *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. [Online]. Available: https://openreview.net/forum?id=MOCZI3h8Ye

[4] T. Unterthiner, D. Keysers, S. Gelly, O. Bousquet, and I. Tolstikhin, "Predicting neural network accuracy from weights," 2021. [Online]. Available: https://arxiv.org/abs/2002.11448

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

[6] K. Schürholt, B. Knyazev, X. G. i Nieto, and D. Borth, "Hyper-representations as generative models: Sampling unseen neural network weights," 2022. [Online]. Available: https://arxiv.org/abs/2209.14733

[7] K. Schürholt, M. W. Mahoney, and D. Borth, "Towards scalable and versatile weight space learning," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*. PMLR, 2024.

[8] B. Soro, B. Andreis, S. Chong, and S. J. Hwang, "Instruction-guided autoregressive neural network parameter generation," in *Workshop on Neural Network Weights as a New Data Modality*, 2025. [Online]. Available: https://openreview.net/forum?id=QutFK34ea1

[9] L. Meynent, I. Melev, K. Schürholt, G. Kauermann, and D. Borth, "Structure is not enough: Leveraging behavior for neural network weight reconstruction," in *Workshop on Neural Network Weights as a New Data Modality*, 2025. [Online]. Available: https://openreview.net/forum?id=APsHrpqO3W

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[11] W. Ågren, "The nt-xent loss upper bound," 2022. [Online]. Available: https://arxiv.org/abs/2205.03169