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 established them as a foundational technology in modern artificial intelligence. However, as their use has expanded, so too has attention to their inherent mechanistic limitations. Two major drawbacks of NNs are their lack of explainability—the “black-box” effect—and the substantial computational cost of training. With platforms like Hugging Face and GitHub hosting over one million pre-trained models [2], interest in addressing these limitations has intensified. Motivated by this vast availability of models, a novel research direction—weight space learning—has emerged as a promising approach to better understand these limitations.

Formally, the weight space of a neural network refers to the set of all possible configurations of its parameters $W \in \mathbb{R}^n$, where n is the total number of trainable weights. Each point in this space represents a unique model with a distinct mapping from inputs to outputs. Weight space learning thus concerns learning representations or distributions over this space, capturing how variations in W relate to model behaviour and how variations in model training influence W. The field generally considers two types of tasks: discriminative and generative.

In discriminative applications, models use the weights of pre-trained networks—often collected into a model zoo [3]—as input to predict meta-information about the original models [4]. The quality of a weight space representation is typically evaluated by the performance of a simple multi-layer perceptron (MLP) in predicting such meta-information, conditioned only on the model’s weight embedding.

Common meta-information metrics include the model’s final performance and its generalisation gap (the difference between training and validation loss). [5] demonstrated that weight embeddings can encode key training characteristics, such as recovering the size of the dataset used for training. These discriminative tasks serve both to validate the quality of the derived weight representations and to provide practical predictive value.

In generative applications, researchers aim to model the underlying distribution of neural network weights W, conditioned on additional information or reference models, $P(W \mid \dots)$. Sampling from this distribution enables the generation of entirely new model weights.

[6] used an autoencoder with a bottleneck layer to generate hyper-representations of multiple model zoos. Building on this idea, [7] introduced the Sequential Autoencoder for Neural Embeddings (SANE), which improved scalability and enabled work on much larger models.

To model $P(W \mid \mathcal{D})$, [8] employed a Vector Quantised Variational Autoencoder (VQ-VAE), which incorporates dataset information when learning latent weight representations. Similarly, [9] modelled $P(W \mid R)$ by incorporating behavioural differences between reconstructed and original models into the embedding learning process.

While existing weight space learning methods have made significant progress, they typically address isolated aspects of the learning process. Current approaches model either $P(W \mid \mathcal{D})$ or $P(W \mid R)$, but rarely both simultaneously. This is a key limitation: in practice, a model's weights are influenced by both the data it was trained on and the results it achieved. Understanding the joint relationship $P(W \mid \mathcal{D}, R)$ is essential for explaining model behaviour and for generating models with desired characteristics.

Contrastive learning has emerged as a powerful paradigm for learning unified representations across modalities, as demonstrated by models such as CLIP [10], which bridge vision and language. A contrastive objective pulls related samples closer in representation space while pushing unrelated samples apart, forming a meaningful joint embedding space for heterogeneous data types. This property makes contrastive learning particularly suitable for modelling the complex distribution $P(W \mid \mathcal{D}, R)$ — encompassing visual datasets, high-dimensional weight tensors, and performance metrics — without requiring a shared native representation.

In this 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 another for weight embeddings using an autoencoder architecture—alongside a binned result embedding table. These encoders are trained using the contrastive objective NT-Xent [11], which encourages related triplets $(\mathcal{D}, W, R)$ to be close in the shared latent space. Figure 1.1 depicts a high-level view of the full embedding pipeline.

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

Interpretability and Analysis: Examining geometric relationships within the shared embedding space can reveal how dataset characteristics shape learned weights and model behaviour. By establishing a meaningful relationship between the triplet $(\mathcal{D}, W, R)$, one can systematically

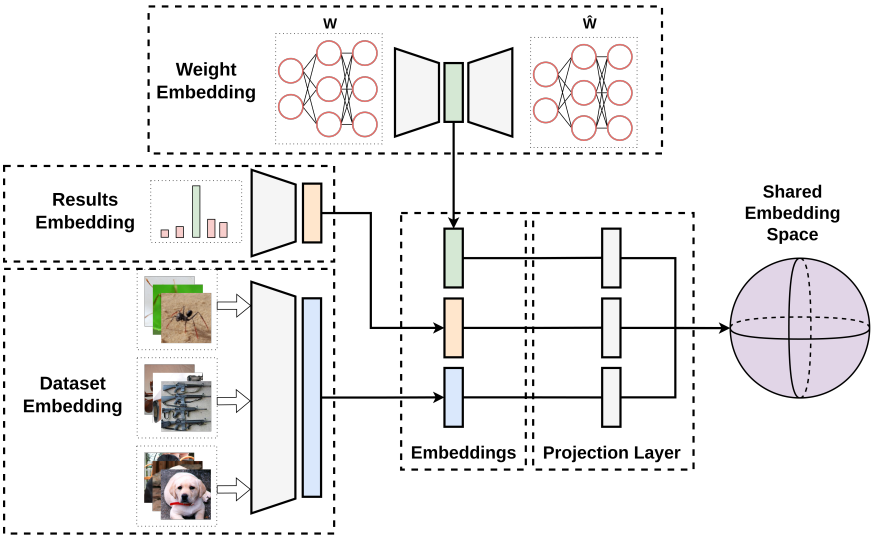

Figure 1.1: AA figure illustrating the process of embedding a dataset, model weights and results in a shared embedding space.

perturb any of the three components while keeping the others fixed, and observe how such changes affect the reconstructed—and thus the real—weight space. This capability enables a structured and extensive exploration of how dataset properties and training outcomes influence learned representations.

Conditional Model Sampling: The learned distribution enables sampling of model weights conditioned on both dataset properties and target performance metrics–approximating $P(W \mid \mathcal{D}, R)$. Such a capability supports zero-shot model generation, where high-performing networks can be synthesised directly from their latent representations, bypassing the need for gradient-based optimisation. By leveraging the structure of the learned weight manifold, this approach has the potential to significantly reduce training costs while enabling the targeted creation of models optimised for specific datasets or performance objectives.

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," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8748–8763. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

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