Chapter 1

Background

1.1. Weight Space Learning

Weight space learning is a field within machine learning that focuses on understanding and leveraging the structure of the neural network weight space. The central aim is to model how network parameters are shaped by data, architecture, and training dynamics, and to capture these relationships within a learnable representation.

At its core, weight space learning seeks to construct *meta-models*—models that learn from other models. Unlike standard machine learning models that capture patterns in data, meta-models capture patterns in the *weights* of networks trained on that data. In this way, the goal shifts from learning a direct input–output mapping to learning the structure that governs how such mappings are formed.

Due to the enormous scale and dimensionality of modern neural networks [], it is typically infeasible to operate directly on raw model weights. Furthermore, redundancy is well known to exist in neural networks; smaller architectures can often achieve comparable performance to larger ones []. These challenges motivate one of the central subproblems of weight space learning: the discovery of low-dimensional representations of weight space.

A *latent representation* of weight space provides a compact and structured encoding of a model's parameters. The transformations—linear or non-linear—that map weights into this latent space are learned to preserve the essential information required to reconstruct or analyse the original weights. Among the many dimensionality reduction techniques available, a useful distinction can be made between *reversible* and *non-reversible* methods.

Reversibility is of particular importance in weight space learning. While encoding weights into a latent representation (real $\rightarrow$ latent) is informative, the ability to reconstruct the original weights (real $\rightarrow$ latent $\rightarrow$ real) is far more valuable. This reversibility enables the synthesis of entirely new weight configurations, supporting generative applications such as zero-shot model creation and performance-guided model generation. Consequently, reversible latent representation methods are the most prevalent within weight space learning, especially for

encoding and decoding neural network weights.

This chapter proceeds as follows. Section 1.2 introduces Principal Component Analysis (PCA), a linear and probabilistic approach that provides a simple yet effective reversible dimensionality reduction method. Section 1.3 discusses reversible, non-linear approaches, focusing on autoregressive encoder architectures and presenting the Sequential Autoencoder for Neural Embeddings (SANE) []. Finally, Section 1.4 explores a non-reversible, modality-heterogeneous technique, contrastive learning, including a description of the NT-Xent loss and its implementation in CLIP [].

1.2. Principal Component Analysis

1.3. Autoregressive Encoders

1.3.1. Sequential Autoencoder for Neural Embeddings

1.4. Contrastive Learning

1.4.1. NT-Xent Loss

1.4.2. CLIP

Bibliography