

Scrutinizing the Training Dynamics of Neural Network Representations: An Approach from Graph-based Representation Dissimilarity

Isaac Zúñiga Arias*

September 2026

Abstract

Deep neural networks achieve remarkable predictive performance, yet understanding how their internal representations evolve throughout training remains a fundamental challenge. This thesis investigates theoretically and empirically the *Shape-Aware Graph Distance* (SAGD), a graph-based framework for comparing neural network representations through their intrinsic geometric structure. Representations are modeled as weighted graphs, whose pairwise Commute Time Distances are summarized by empirical cumulative distribution functions and compared using the 1-Wasserstein distance. This construction provides a global measure of representational dissimilarity that is naturally invariant to permutations of node labels and captures structural relationships beyond standard Euclidean comparisons.

The theoretical framework combines spectral graph theory, random walks, and 1-Wasserstein distance to establish the mathematical foundations of SAGD and to analyze its principal properties. Robustness is evaluated through controlled experiments involving synthetic Gaussian data, high-dimensional Gaussian mixtures, and Watts–Strogatz small-world networks, demonstrating that the proposed methodology is stable under moderate variations in graph construction and stochastic perturbations.

Finally, SAGD is applied to study the evolution of internal representations in a ResNet-18 model trained on the CIFAR-10 dataset. By combining SAGD with the graph-based visualization method; Shape-Aware Stochastic Neighbor Embedding (SASNE), the analysis reveals distinct geometric trajectories across network layers and training epochs. The results show that representations undergo rapid structural reorganization during the early stages of optimization, followed by gradual refinement, and continue to evolve even after predictive performance has largely converged. Overall, this work presents a graph-theoretic perspective on representation learning and provides a mathematically grounded framework for analyzing the geometry and dynamics of neural network representations.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden.
E-mail: isaac.actuario@gmail.com. Supervisor: Chun-Biu Li.