WILLIAM FORGET, Bishop's University

Evaluating Neural Networks Through the Lens of Topology: A Persistent Homology Approach

The goal of this project is to study formal methods to analyse a neural network's performance metrics. To do so we use a compact "knowledge matrix", which captures the relationships among learned features across all layers. The idea is to treat this matrix as a point cloud in a high-dimensional feature space, and apply topological data analysis (more precisely persistent homology) to extract its multi-scale topological features such as connected components, loops and voids that persist across filtration levels. Our investigation follows two complementary paradigms. In Experiment 1, we fix a trained (or untrained) network and compute knowledge matrices for every input in our dataset; we then either vectorize these matrices or analyse them directly via persistent homology, comparing topological invariants against accuracy, robustness and generalisability. In Experiment 2, we fix a single input and trace its evolving knowledge matrix at successive training epochs, revealing how topological structure emerges and stabilizes during learning. To assess resilience, we introduce adversarial and corruption-based attacks into both paradigms and compare the resulting homological features to those of the unperturbed network.