WILLIAM FORGET, Bishop's University
Understanding Neural Networks Through the Knowledge Matrix

This project investigates formal methods for analysing neural network performance using a compact knowledge matrix that captures relationships among learned features across all layers of the model. For each input, the knowledge matrix is treated as a point in a high-dimensional space, forming a point cloud that reflects the network's internal representation of the data. We then apply dimensionality reduction techniques such as PCA and LDA to study the structure of these representations in lower dimensions. The resulting embeddings are compared with representations from the penultimate and output layers in order to evaluate how well the knowledge matrix preserves information about the network's behaviour.