Interpreting Deep Learning Models using Persistent homology

Sudhanva Manjunath Athreya, Dr. Paul Rosen
University of Utah

Deep Learning Interpretability

In Deep Learning, understanding how models make decisions is crucial. Interpretability reveals the reasoning behind predictions, while explainability provides insights into model behavior. While deep learning models often achieve high accuracy, they can be harder to interpret. Balancing accuracy with interpretability helps ensure responsible and effective AI deployment.

Persistent Homology

Persistent homology is a method in computational topology that analyzes the shape of data by tracking the creation and disappearance of topological features (like connected components, loops, and voids) across different spatial scales.

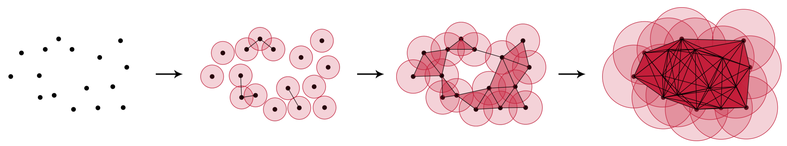

Figure 1. Filtration

How Persistent Homology Works

1. **Filtration:** Constructing a sequence of simplicial complexes by gradually increasing a proximity threshold, linking points into edges, triangles, and higher-dimensional simplicial complexes. Topological features like **Connected Components** (0D), **Loops** (1D), **Voids** (2D+) are tracked during the filtration process.
2. **Tracking Features:** Topological features **appear** (birth) and **disappear** (death) as filtration progresses, and when points are close enough, they form a **connected component**.
3. **Persistence Diagram & Barcode:** The persistence of topological features is visualized using two main tools: persistence diagrams and barcodes.

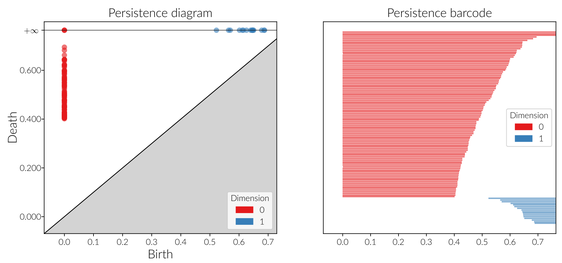

Figure 2. Persistence Homology on a Toy Point Cloud

Where is the PH applied?

The activations from a deep learning model for a given input are retrived using a hook function. Then the pairwise distances for the given inputs are calculated using various distance metrics. The distance metrics are then forwarded to the filtration function.

The choice of distance metric directly affects the filtration process and influences which topological features are captured and their persistence.

By analyzing the same activation data with multiple distance metrics, we can triangulate a more complete understanding of how information flows through the network and which structural patterns contribute to the model's decisions.

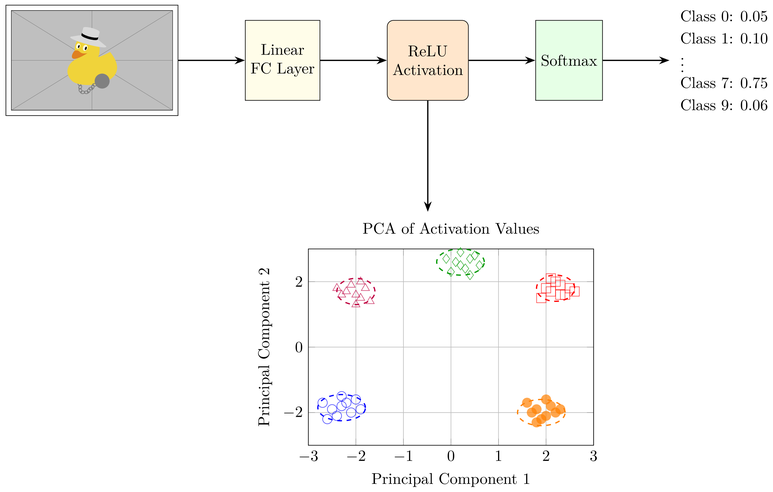

Figure 3. Model Activations

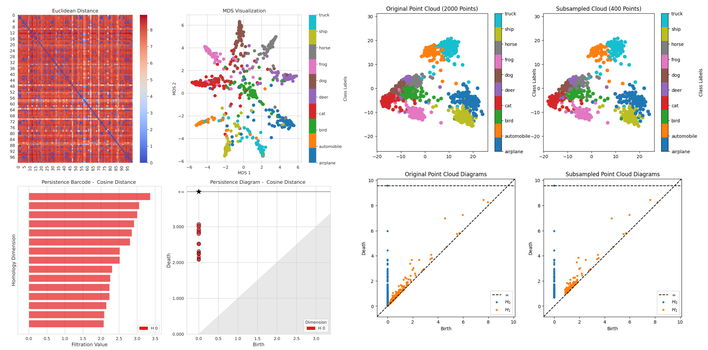

Figure 4. Resnet50 (Trained on CIFAR10)

Figure 5. Persistent Homology on ViT model's Layer 12 activations.

Distance Metrics in Filtration

The choice of distance metric plays a crucial role in persistent homology by defining when simplices form connections in the filtration process. During our experiments we tried Euclidean distance, Mahalanobis distance, and a density-based normalization applied to the the former metrics.

1. **Euclidean Distance (L_2 norm)**

$$d_2(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_i (x_i - y_i)^2}$$

2. **Mahalanobis Distance**

$$d_M(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mathbf{y})}$$

3. **Cosine Distance**

$$d_{\cos}(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

4. **Relative Neighborhood-based distance normalization**:
Here k is the k-th nearest neighbour, a hyperparameter chosen by the user.

$$d_{\text{VR-RNG}}(x, y) = \frac{d(x, y)}{\max\left(d(x, x_k), d(y, y_k)\right)}$$

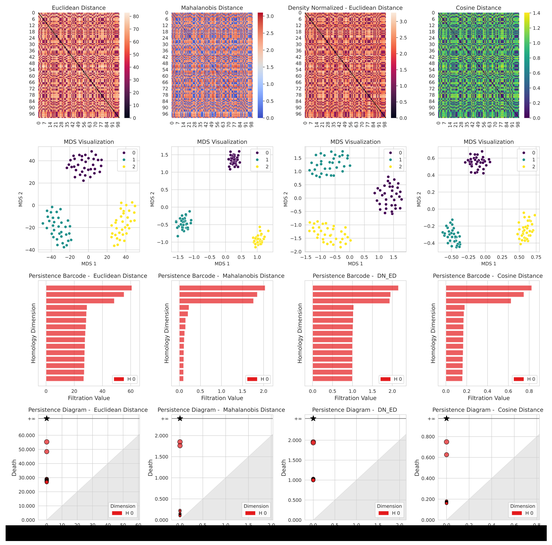

Figure 6. Persistent Homology Sample