Topological Data Analysis and Interpretability of 3D-Convolutional Neural Networks

Sadie Lee[1]

[1]University of British Columbia

1. Introduction

With the increasing ubiquity of deep learning models in many areas, the field of interpretability has emerged to understand what exactly a model is doing and how it is learning, given its complexity and opaqueness. The demand for tools to help interpret and gain confidence in these models' performance has thus grown. A common task of deep learning models is classification, which has been used for two and three dimensions of images, videos, and objects, and adopted by various applications such as computer vision in autonomous vehicles, robotics, and mixed reality. Convolutional neural networks, in particular, are often used for such classification tasks. This proposal aims to address the following questions: What is the shape and structure of an activation space for three-dimensional objects? How are activations organized within and across layers of a three-dimensional convolutional neural network (3D-CNN)? And how can these patterns be visualized to explore a 3D-CNN and its training process? I propose to apply tools from topological data analysis, specifically the Mapper algorithm, to capture the overall shape and pattern of activation vectors within a 3D-CNN, and provide an interface for visualization.

2. Related Work

Topology in Neural Networks. Laying the groundwork for the use of topology to better interpret neural networks, it has been found that neural networks preserve topological features of the data when it is projected into space with low dimensionality (Polianskii 2018). Furthermore, the process of training has been associated with simplifying topological descriptors of the data manifolds in the internal representation of a CNN (Magai & Ayzenberg 2022)—in the first few epochs, topology changes insignificantly whereas in later epochs in which the network is more well-trained, the topology of data is quickly changed throughout the layer hierarchy.

This proposal is largely based on the work of TopoAct (Rathore et al. 2021), a visual exploration system on topological summaries of activation vectors using the Mapper graph, so that insights into learned representations by a neural network can be gained. However, TopoAct focuses primarily on two-dimensional images and objects, whereas this proposal focuses on three-dimensions of images and objects, specifically in three-dimensional convolutional neural networks.

Prior Work by the Applicant. This proposal is derived from my curiosity of neural network processes during my work using convolutional neural networks to classify brain tumors from MRI data with UBC Multifaceted Innovations in Neurotechnology (MINT), as well as during my studies in Cognitive Systems that further propelled my interest in both the mathematics and implications of neural networks. My prior work is therefore adjacent to this proposal, and contributes to my understanding of deep learning.

3. Theoretical background

Mapper Algorithm. The Mapper algorithm defined by Singh et al. (2007), is a method in topological data analysis based on the notion of partial clustering in the data and can be used to reduce high-dimensional datasets into simplicial complexes that capture topologic and geometric information of interest at a specific resolution. A simplicial complex is a combinatorial object in which its evolution is observed as the resolution scale varies: Let P be a discrete set. Then, an abstract simplicial complex is a finite collection K of finite, nonempty sets of P (simplices). A k-simplex is defined where σ is the convex hull of $k+1$ affinely independent points. For example, a 0-simplex is a point, 1-simplex an edge, 2-simplex a triangle, 3-simplex a tetrahedron, and so on. The Mapper algorithm and its graph is aptly suited for visualizing functional structure, such as that

of neural networks, because it can capture an understanding of structure by characterizing the relationship between a feature space and prediction space, even with a highly sparse representation of the dataset.

4. Approach

First, activation vectors as high-dimensional point clouds are collected from a chosen layer of a 3D-CNN by feeding binary 3D tensors (representing a shape) into the network. The dimension of each activation vector depends on the neuron count within a layer, and a collection of activations from overlapping spatial patches in the tensor. These activation vectors are used to compute Mapper graphs, which summarize the topological information of the dataset. Each node in the graph represents a cluster of activation vectors, and an edge connects two nodes if their corresponding clusters have a non-empty intersection. A k-nearest neighbors (K-NN) graph is used to select a more adaptive ε value for the activation space of a layer. Feature visualization is then applied to each node in the Mapper graph, transforming high-dimensional vectors into more semantically meaningful representations by synthesizing an idealized object that would have produced vector $h_{x,y}$ through an iterative optimization process. This process is similar to back-propagation and applies gradient descent in a Fourier basis, resulting in corresponding objects that are likely to have produced such an activation.
With this topological information, we can then create a user interface that visualizes topological patterns such as branches and loops, as well as a global view. Branches are seen when bifurcations occur among object classes. Loops are seen as clusters, when different features of objects are captured. A global view allows us to observe the overall distribution of topological structures.

5. Evaluation and Discussion

To evaluate this framework, I propose using datasets such as the Princeton Shape Benchmark (Shilane et al. 2004) and ShapeNet which consists of 3D CAD models (Chang et al. 2014) to observe the shape of learning representations within 3D-CNNs. Moreover, we may be able to evaluate the generalizability of the framework to other types of data and network architectures such as two-dimensional or three-dimensional image data on the MNIST or CIFAR datasets in standard CNN (not 3D-CNN) architectures.

The most prominent implication of this framework is anticipated to be improvement in the interpretability of 3D-CNN models, and dependent upon results, other neural network architectures. A user interface allows people coming from various backgrounds to visualize the structure of a 3D-CNN and its learning process.
Furthermore, this framework may provide for corrective actions during training, including humans in the loop to improve accuracy of the network. If, for example, we observe that two classes bifurcate at a certain layer and continue to be misclassified, the network width can be increased or training data can be selectively augmented. Additionally, this framework may be useful in analyzing the effect of adversarial attacks at different layers of the network by visualizing how an attack alters the activations. This framework proposes understanding topological structure of activation spaces within a 3D-CNN to explore the interpretability of deep learning models through visualization.

References

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., ... & Yu, F. (2015). Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Magai, G., & Ayzenberg, A. (2022). Topology and geometry of data manifold in deep learning. *arXiv preprint arXiv:2204.08624*.

Polianskii, V. (2018). An Investigation of Neural Network Structure with Topological Data Analysis.

Rathore, A., Chalapathi, N., Palande, S., & Wang, B. (2021, February). TopoAct: Visually exploring the shape of activations in deep learning. In *Computer Graphics Forum* (Vol. 40, No. 1, pp. 382-397).

Shilane, P., Min, P., Kazhdan, M., & Funkhouser, T. (2004, June). The princeton shape benchmark. In *Proceedings Shape Modeling Applications, 2004.* (pp. 167-178). IEEE.

Singh, G., Mémoli, F., & Carlsson, G. E. (2007). Topological methods for the analysis of high dimensional data sets and 3d object recognition. *PBG@ Eurographics*, *2*, 091-100.