Topological Data Analysis of Neural Network Layer Representations

Archie Shahidullah
Computing + Mathematical Sciences
California Institute of Technology
Pasadena, CA 91125
archie@caltech.edu

Abstract

This paper is a cursory study on how topological features are preserved within the internal representations of neural network layers. Using techniques from topological data analysis, namely persistent homology, the topological features of a simple feedforward neural network's layer representations of a modified torus with a Klein bottle-like twist were computed. The network appeared to approximate homeomorphisms in early layers, before significantly changing the topology of the data in deeper layers. The resulting noise hampered the ability of persistent homology to compute these features, however similar topological features seemed to persist longer in a network with a bijective activation function.

1 Introduction

In recent years, deep neural networks have revolutionized many computing problems once thought to be difficult. However, they are often referred to as black boxes and the mechanisms through which they learn have remained elusive. Insight into how a neural network internally represents the dataset it is trained on will provide understanding into what makes for effective training data, and how a neural network extracts relevant features from a dataset.

1.1 Layer Representations

A feedforward neural network, N, that acts on input data x can be viewed as a composition of n functions, where each function represents a layer, L_i,

$$N(x) = L_n(L_{n-1}(\dots L_1(x)))$$

Each layer, acting on input z, consists of a weights matrix W, bias vector b, and (usually) nonlinear activation function σ,

$$L_i(z) = \sigma(Wz + b)$$

The output of L_i is what we will refer to as a layer representation. It is important to note that more complicated architectures are not restricted to a linear chain structure and thus can have skip and recurrent connections that complicate their graph structure. However, skip connections simply take in multiple inputs and recurrent (and recursive) connections can be unrolled and layer representations can be recovered.

Preprint. Under review.

1.2 Overview of Paper

Section 2 gives relevant background about topology and its relation to neural networks. Section 3 gives an overview of persistent homology, a technique to compute the topological features of a point-cloud dataset. Section 4 outlines the experiment performed and its results. Finally, Section 5 has the discussion of results and directions for future work.

2 Background and Previous Work

We will define topology, give motivation for the homology groups of a topological space, its applicability to neural networks, and review previous work on the subject.

2.1 Definition of Topology

Topology is the study of geometric objects with a particular structure that allows for a rigorous treatment of the concepts of "bending" and "twisting" a space and how properties of the space are preserved under these transformations. Formally, a topological space is the tuple (X, τ), where X is some set and τ is a multiset consisting of subsets of X. Armstrong (1983) gives the following definition of a topological space,

Definition 2.1 (Topological space)**.** *Given a set X and a multiset τ that consists of subsets of X, a topological space is the tuple (X, τ) that fulfills the following axioms,*

1. $\emptyset \in \tau$ *and* $X \in \tau$
2. $\forall S_i \in \tau, \bigcup_i S_i \in \tau$ *for finite or infinite unions*
3. $\forall S_i \in \tau, \bigcap_i S_i \in \tau$ *for only finite intersections*

Any member of τ is termed an open set (and its complement is a closed set), and τ is the topology on X.

Another central concept in topology is continuous deformation between topological spaces, and this is formalized in the idea of a homeomorphism, defined as,

Definition 2.2 (Homeomorphism)**.** *A homeomorphism $f : X \to Y$ is an isomorphism between two topological spaces X and Y and therefore fulfills the following critera,*

1. *f is bijective*
2. *f is continuous*
3. *f^{-1} is continuous*

2.2 Homology

The homology of a topological space informally characterizes the number of "holes" in the space. If we take a cycle to be a generalization of a closed loop on some space, such as on the surface of the sphere S^2, we can classify cycles by whether or not they can be continuously deformed into each other. If two cycles cannot be deformed into each other, it is said that there exists a hole on the topological space.

A 0-dimensional hole is a connected component, a 1-dimensional hole is a loop, a 2-dimensional hole is a shell, and so on. The study of these holes requires a formal description of the boundaries on topological spaces. Boundaries in general are linear combinations of more basic geometric objects, which motivates us to introduce some sort of structure to allow for this. The homology of a topological space X is formally represented by its homology groups,

$$H_0(X), H_1(X), H_2(X), \ldots$$

Each homology group $H_k(X)$ has an abelian group structure, and as such we naturally use $\mathbb{Z}$. We refer to the rank of a homology group as the Betti number b_k. In general,

$$H_k(X) = \mathbb{Z} \times \cdots \times \mathbb{Z} = Z_{b_k}$$

A Betti number counts the number of holes in a topological space and is topologically invariant, which makes it an ideal candidate to judge the topological features of a space. Under homeomorphism, the Betti numbers of our domain and codomain remain unchanged. It is important to note that the Betti numbers do not account for all topological invariants, such as torsion. Torsion refers to features such as the twist of a Möbius strip.

The Betti numbers of S^1, the circle, are $b_0 = 1, b_1 = 1$, and 0 otherwise. This is because there is one connected component, and one loop (1-dimensional boundary). The Betti numbers of S^2 are $b_0 = 1, b_2 = 1$, and 0 otherwise. This is because there is 1 connected component and one 2-dimensional boundary (around the interior of the sphere). Notably, there are no loops that can be drawn on the surface that cannot be deformed to a single point. The torus T^2, a doughnut-shaped object, has Betti numbers $b_0 = 1, b_1 = 2, b_2 = 1$, and 0 otherwise. There is one connected component, and two loops (one around the ring and another around the "main hole"). Lastly, there is a 2-dimensional boundary around the interior. It is interesting to note $T^2 = S^1 \times S^1$.

2.3 Application to Neural Networks

We are interested in whether the layer representation L_i acts similar to a homeomorphism in that it allows a neural network to represent the topological features of a dataset. Unfortunately, neural network layers are not in general homeomorphisms,.

Theorem 2.1. *A neural network layer, $L(x) = \sigma(Wx + b)$, need not be a homeomorphism.*

Proof. If W is not a member of the general linear group $\mathrm{GL}_n(\mathbb{R})$, it is not invertible and therefore no homeomorphism exists by 2.2. Additionally, if σ is not bijective, no homeomorphism can exist also by 2.2. □

However, the topological approach is still useful. If layer representations are examined and are found to resemble the topological features of the dataset, this suggests a reason neural networks are effective is that they learn a robust representation of the topological features of a space.

2.4 Previous Work

Studying neural network learning with topology has been a popular approach in theoretical machine learning. This is best exemplified with the concept of the manifold hypothesis. Fefferman, et al. (2013) gives the following definition of the manifold hypothesis,

Definition 2.3 (Manifold Hypothesis)**.** *High-dimensional data tends to lie near low-dimensional manifolds.*

This concept is best explained with an example. Imagine there exists a model to classify $m \times n$ images between cats and dogs. The data space is $\mathbb{R}^{mn}$. However, this space clearly contains data not relevant to the task at hand (such as images of flowers or random noise), and the manifold hypothesis conjectures that there exists a much lower-dimensional submanifold of the data space that approximates the relevant data.

Given that manifolds are topological spaces, it is reasonable to assume methods from topology can be used to gain insight into the learning process, should the manifold hypothesis be correct. There exists a field called topological data analysis (TDA) precisely focused on extracting topological features from data. The most common technique used in TDA is persistent homology. Persistent homology will be described in the next section, but it essentially computes whether the features corresponding to each homology group exist in a point-cloud, e.g. if one can draw loops via linear combinations of datapoints. Montúfar, et al. (2020) showed that a neural network can be trained to approximate topological features similar to persistent homology. While this paper will use persistent homology to compute topological features, it is encouraging that a neural network can be trained to compute the topological features of a dataset.