# Topological Analysis of Neural Network Transfer function

BMVC 2021 Submission # ??

**Abstract**

In Spite of the tremendous success of deep neural networks for solving complex tasks in domains like computer vision, speech recognition and natural language processing, there has been little work on understanding the dynamics of the training process and also the working of a trained model. In this study, we analyze desired characteristics of the transfer function of each layer, and hence the whole network, based on the topological transformations of the space of training samples. We introduce a new family of activation functions that provides faster convergence for classification tasks. Based on our empirical results, we also propose a novel method for pruning and reducing the size of a trained model. We performed experiments on previously studied, synthetic binary classification datasets with complex topology using Multi Layer Perceptron.We use betti numbers to quantify topological complexity and our experimental results show that the proposed activation function results in much faster convergence and the betti number reduces at a faster rate across layers with the proposed activation function. We verified our results on popular image classification datasets like fashion MNIST, CIFAR-10 and cat-vs-dog classification datasets by running experiments using Convolutional Neural Networks.

## 1 Introduction

Deep neural networks have become the default choice for solving many tasks, which was otherwise either partially solved or not solved at all, in domains like computer vision, speech and natural language processing where a lot of data is available. However, choosing the right architecture (like the selection of hyper parameters activation function, number of layers and number of units per layer) for a specific task is mostly based on trial and error or based on the previous empirical results. The transfer function of each layer and the entire neural network is just treated as a complex, unknown and nonlinear function parameterized by weights and biases. Not much work has been done in identifying the impact of various favourable and unfavourable characteristics of this transfer function.

In this study we investigate some of the desired characteristics a neural network architecture should have for solving a classification task. We derive our results based on the topology of the space of training data and how this topology changes as data is transformed by each layer.

Topology is a field of mathematics that studies the shape of objects and associated invariances like connectedness, number of holes etc. It is observed that many real datasets when viewed as point cloud in a high dimensional space follow certain topology. For example the

study in [**?** ] shows that the image patches obtained from natural images follow the topology of a Klein bottle. Topological Data Analysis[**?** ][**?** ] uses topological tools ( like persistent topology ) for analysing point cloud dataset in order to identify and characterize underlying structures in a dataset.

Most of the existing work on the analysis of a trained model is limited to either looking at the learned weights and feature maps of each layer, with the goal of reusing these features across tasks, or techniques like Class activation map ( or saliency map[**?** ] where the intention is to identify the relevant feature of input that is responsible for producing the output. In their recent study, Gregory et. al. [**?** ] quantifies topological complexity using Betti numbers. They observe that topological changes to data across layers of a network remain robust under different instances of training. They further observe that, compared to smooth activation functions like sigmoid and tanh, non-homeomorphic activation functions like ReLU helps in changing the topology of data faster.

Our work is motivated by [**?** ] where the transfer function of each layer is looked at based on how the layer changes the topology of the data. Most real world datasets have non-trivial complex topology, and in order to perform classification each layer of the neural network transforms the entire space of data to a simpler topology. This leads us to the conclusion that in order to achieve classification, each layer of the neural networks should be able to change the topology of data and hence we need a non-homeomorphic transformation at each layer. This is achieved by activation functions with discontinuity like Relu. We followed the approach in [**?** ] and used betti numbers to quantify topological complexity of the point cloud dataset.

In addition to the above insight that the activation function should be non-homeomorphic, we also hypothesize that a many-to-one transfer function can help to bring samples from the same class closer in the transformed space( See appendix A for a skeleton of proof). Based on this hypothesis, we introduce a new activation function with multiple many-to-one regions and multiple discontinuities. Results of our experiments show that, with the proposed activation function the network converges faster as compared to commonly used activation functions like ReLU and sigmoid. It is also seen that the betti number, computed using persistent homology [**?** ] reduces faster when with the proposed activation function.

The major contributions of our paper are:

1. We provide new guidelines for designing activation functions for supervised classification tasks. We illustrate the guideline by proposing a new family of activation functions.

2. We propose an easy technique for neural network pruning (i.e reducing the parameters in a trained model without significant decrease in accuracy) using betti numbers computed on the output feature space of each layer.

# 2 Related Work

# 3 Proposed Method

## 3.1 Problem Formulation

We restrict our analysis to the task of supervised classification of 3-dimensional synthetic datasets using Multi Layer Perceptron(MLP ) and classification of images using Convo-

lutional Neural Network. The classification task can be viewed as a many-to-one mapping, $f : R^d \mapsto \{c_1, c_2, \ldots c_k\}$ ( $f : R^{H \times W} \mapsto \{c_1, c_2, \ldots c_k\}$ in case of images). We denote the domain of the mapping with a discrete set $X$ with finite number of samples i.e $X = \{x_1, x_2, x_3, \ldots x_n\}$. The set of all samples $X$ form a point cloud dataset on $d$-dimensional space ($d = H \times W$ in case of image of height $H$ and width $W$). The function $f$ is a composition of multiple functions, $g_l$, where $1 \leq l \leq L$ and each $g_l$ refers to the transfer function of layer $l$ and $L$ is the total number of layers in the neural network. As detailed in [**?** ], for classification task a non-homeomorphic transfer function is required for each layer as it can change the topology of point cloud dataset. In order to achieve classification one needs to change the topology from an initial complex topology to a simple and contractible topology for each class. Another important characteristic of layer transfer function is that each layer reduces the dimensionality of the input data.

Our study focuses on two important aspects of neural network design.

1. Design of an optimum non-linear activation function for supervised classification task. See subsection **??** below.

2. Propose an easy technique for neural network pruning (i.e reducing the parameters in a trained model without significant decrease in accuracy) using betti numbers computed on the output feature space of each layer. See subsection **??** for details.

## 3.2 Dataset

We use two 3-dimensional simulated datasets, nine ring dataset and nine sphere dataset used in [**?** ]. The nine ring dataset, as shown in Figure 1, consists of two classes of data, colored Green and Red, interlocked together. The nine sphere dataset consists of nine Green spheres and 18 Red Spheres enclosing each other as shown in Figure 2. Both the datasets contain 16000 samples for training and 2000 samples for testing.

We also performed experiments and provide results on the following real datasets:

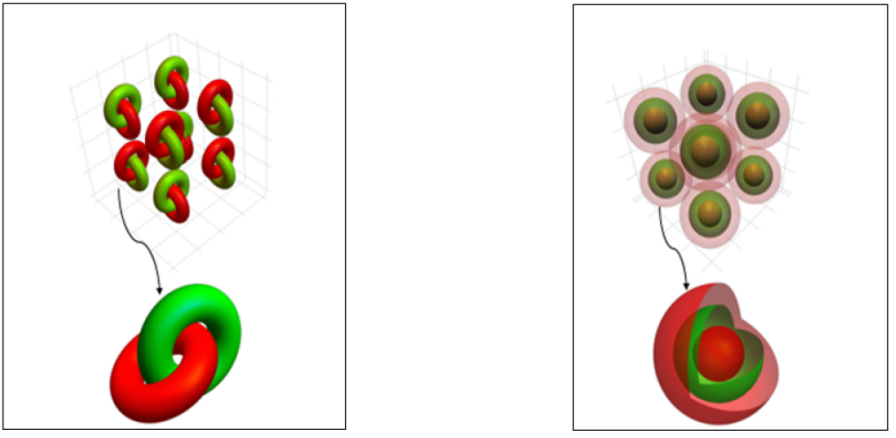
1. Cat-Dog dataset in Kaggle - The dataset consists of images of size $32 \times 32$. The dataset is divided into training and testing sets, with 8000 training images and 2000 testing images, each set with an equal number of images of cats and dogs.

2. Fashion Minst - The dataset consists of 70,000 grayscale images of size $28 \times 28$ with 10 different categories. The dataset is divided into training and testing sets, with 60,000 training images and 10,000 test images.

3. CIFAR-10 - The dataset consists of 60,000 colour images of size $32 \times 32$ with 10 different categories. Each category consists of 6000 images. The dataset is divided into training and testing sets, with 50,000 training images and 10,000 test images.

Some sample images from each of the above datasets are shown in Figure 2, Figure 3 and Figure 4.

The datasets were converted to gray scale and normalized. No other preprocessing was performed.



(a) Nine ring dataset (b) Nine sphere dataset

Figure 1: 3 dimensional synthetic datasets with two different classes(Red and Green) used for training MLP

## 3.3 Design of Layer Transfer Function

The layer transfer function is composed of an affine transformation and a non-linear activation function. In the subsections below, we provide the details of desired characteristics for activation function from a topological point of view.

### 3.3.1 Betti numbers and their significance on layer transfer function

Betti numbers, denoted as $\beta_k(X)$ are used to quantify the topological complexity of a $d$-dimensional topological space $X$ where $0 \leq k \leq d$. The $0^{th}$ betty number $\beta_0(X)$ is the number of connected components, the first betti number $\beta_1(X)$ is the number of one dimensional holes, the second betti number is the number of two dimensional holes and so on. Figure 6 shows some topological spaces and their corresponding betti numbers. For efficient classification, one needs to transform the original point data cloud, $X$, from a high dimensional space with large betti numbers to a low dimensional latent representation with $0^{th}$ betti number (number of connected components) equal to the number of classes $K$ and all other betti numbers to zero. As evident from Figure 6, this ensures that each connected component corresponds to a single class ( either Red or Green) and there are no holes within the connected components. Hence each connected component is contractible to a single point. It is easy to find a decision boundary if there are no holes on the manifold formed by samples from a single class. In the rest of this document wherever we mention the term topological complexity, we mean the betti numbers of the topological space.

### 3.4 Pruning of Convolutional Neural Network

## 4 Results

## 5 Conclusion

## References

[] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[] Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *arXiv preprint arXiv:1710.04019*, 2017.

[] Fanman Meng, Kaixu Huang, Hongliang Li, and Qingbo Wu. Class activation map generation by representative class selection and multi-layer feature fusion. *arXiv preprint arXiv:1901.07683*, 2019.

[] Gregory Naitzat, Andrey Zhitnikov, and Lek-Heng Lim. Topology of deep neural networks. *Journal of Machine Learning Research*, 21(184):1–40, 2020.