Visualization of Activation Space in ReLU Neural Networks through Topological Data Analysis

Syed Fahim Ahmed
UID: u1419916

May 1, 2024

This project delves into the application of Topological Data Analysis (TDA) for visualizing and understanding the activation spaces within ReLU neural networks, focusing on how these spaces manifest as polyhedral complexes. By employing the Mapper algorithm, we explore the intricate, high-dimensional activation landscapes that underpin the decision-making processes of neural networks. To make our analysis interpretable, we utilized a simplified model consisting of a single hidden layer with three neurons equipped with ReLU activation functions. This configuration allows us to demonstrate that the activation space, in this case, forms a polyhedral complex resembling tetrahedra, as shown in Figures 1, 2, and 3. Through extended training sessions and the comparative analysis of different datasets—specifically the single and double annulus datasets—this study highlights distinct clustering and geometrical partitioning within these activation spaces. The findings not only advance our understanding of neural network behaviors but also enhance the toolkit for developing more interpretable and reliable AI systems. This report integrates comprehensive methodologies, discusses insights from both datasets and proposes future research directions for further enhancing the transparency of neural networks through TDA.

Keywords: Topological Data Analysis, Neural Networks, ReLU Activation, Mapper Algorithm, Visualization, AI Interpretability, Polyhedral Complex.

1 Introduction

In the field of artificial intelligence, neural networks equipped with ReLU (Rectified Linear Unit) activation functions are ubiquitous due to their efficacy in various applications ranging from image recognition to natural language processing. Despite their widespread use, a significant challenge persists in understanding and interpreting the high-dimensional spaces that these networks operate within. The activation space of a neural network, which represents the output of its layers, can be complex and unintuitive, especially when viewed through the lens of traditional analysis techniques.

The activation spaces of neural networks, particularly those with ReLU activations, possess structures that can be interpreted as polyhedral complexes. This complex geometric representation holds the key to understanding how neural networks process and transform input data into outputs. However, the inherent complexity and high dimensionality of these spaces make them difficult to study and understand using conventional methods.

The need for transparency and interpretability in AI systems is critical, especially in sectors where decisions have significant ethical and safety implications. By demystifying the internal mechanisms of neural networks, we can enhance trust, ensure reliability, and facilitate deeper insights into their functioning, which is crucial for advancing AI technology responsibly.

This project employs Topological Data Analysis (TDA), specifically the Mapper algorithm, to visualize and analyze the activation spaces of neural networks. We focus on making the complex mathematical concepts of TDA accessible and interpretable to a broad audience by employing a simplified model—a single hidden layer neural network with three neurons using ReLU activation. This setup helps illustrate how the activation space forms a polyhedral complex, specifically tetrahedral structures, as detailed in Figures 1 and 2. Additionally, in Figure 3, we extend our analysis to demonstrate similar topological data patterns, reinforcing the polyhedral nature of these activation spaces.

By systematically applying TDA to both the single and double annulus datasets, this study not only clarifies the geometrical and topological underpinnings of activation spaces in neural networks but also showcases how these insights can be visually and conceptually communicated. The simplified model choice, with its clear tetrahedral activation patterns, serves as a compelling example, making the abstract concept of high-dimensional data analysis more tangible and comprehensible.

2 Background

Deep Neural Networks (DNNs) play a crucial role in advancing artificial intelligence, capable of processing complex datasets in a way that mimics human cognitive functions. This project utilizes a simplified neural network model with one hidden layer containing three neurons activated by ReLU (Rectified Linear Unit) functions. This setup is crucial for mapping and visualizing the transformations that inputs undergo within the network.

Activation spaces in such networks, mainly when ReLU functions are used, can be conceptualized as polyhedral complexes. This geometric representation helps in dissecting the multi-dimensional transformations occurring within the network, making it possible to study how inputs are partitioned and processed across different layers.

Polyhedral complexes describe how activation spaces in ReLU networks are segmented into geometrically distinct units, simplifying the high-dimensional outputs into more manageable forms. Topological Data Analysis (TDA) offers powerful methods for visualizing these spaces, enabling more precise insights into the data's underlying structures.

The use of tools such as the Mapper algorithm in TDA helps translate these complex geometric data into comprehensible visual summaries. This project leverages such tools to render the intricate activation spaces of neural networks into formats that are easier to interpret and analyze.

The foundational theories for this study stem from Chakir, Schütte, and Sunkara's (2023)[1] analysis of ReLU networks, which articulates how these networks' activation spaces can form polyhedral complexes. Inspired by these findings, this project extends those theoretical models into applied visualizations aimed at demystifying the operational mechanisms of neural networks.

Further enhancing our toolkit, Zhou et al. (2021)[2] developed ”Mapper Interactive,” an innovative platform for dynamically exploring high-dimensional data. This tool has been instrumental in this project, facilitating practical and interactive analysis of complex data patterns and enhancing the interpretability of neural network behaviors.

The exploration of activation spaces is enriched by further studies like those by Raghu et al. (2017)[3] and Jia et al. (2019)[4], which delve into neural networks' geometric and topological dynamics. These studies underscore the critical need for tools to dissect and articulate networks' complex behaviors, highlighting the intersection of theoretical research and practical applications.

3 Methodology

3.1 Neural Network Architecture and Training Process

The neural network used in this study is a straightforward feedforward model consisting of a single hidden layer equipped with three neurons. This architecture facilitates a focused examination of activation space transformations.

Model Structure and Equations:

1. Input to Hidden Layer:

- $L_1 = W_1x + B_1$: Linear transformation where W_1 and B_1 are the weights and biases of the hidden layer, and x represents the input features.
- $N_1 = \phi(L_1)$: Application of the ReLU activation function ϕ, which introduces non-linearity by outputting the input directly if it is positive, else zero.

2. Hidden Layer to Output:

- $L_2 = W_2N_1 + B_2$: The activated output N_1 from the hidden layer undergoes another linear transformation to the output layer.
- $y = \text{Sigmoid}(L_2)$: The sigmoid activation function converts the final layer's output to a probability between 0 and 1, suitable for binary classification.

Training Configuration:

- Epochs: The model is trained over 1000 epochs to allow adequate learning by minimizing the error between the predicted outputs and the actual data labels.
- Loss Function: Binary cross-entropy loss is employed, which is ideal for binary classification tasks.
- Optimization: Gradient descent optimizes the weights and biases, updating them to reduce the loss function effectively.

3.2 Data Preparation and Visualization

Data is sampled from defined geometric distributions to form specific shapes, like single and double annuli, characterized by distinct angular intervals. The labels are generated based on these distributions, facilitating the model's training on complex patterns that challenge its classification capabilities.
Data Sampling:

- Single Annulus Dataset: Generated using specific angular intervals to create a circular data distribution with varying densities.
- Double Annulus Dataset: Combines data from two distinct circular distributions, each with its angular and radial parameters, creating a more complex pattern.

Topological Data Analysis (TDA) and Mapper Algorithm
After training, the activation outputs of the hidden layer are analyzed using the Mapper algorithm, a TDA tool that simplifies high-dimensional data into understandable visualizations.
Mapper Configuration:

- Data Input: Activation outputs from the hidden layer are extracted and saved into a CSV file for processing.
- Parameters: The Mapper algorithm is configured with parameters such as $min_samples = 2$ and $epsilon = 0.1$, using DBSCAN for clustering, with an interval count of 30 and an overlap of 50%.

Visualization:

- Hyperplane and Activation Mapping: The decision boundaries (hyperplanes) determined by the network's weights are visualized across the data space alongside the areas activated by each neuron.
- Mapper Graphs: Produced to display the topological structure of the activation space, highlighting how the network perceives and segments the input space.

Training and Visualization Scripts
Python and **PyTorch** are utilized to implement the neural network and conduct all analyses. Libraries such as **matplotlib** are used for plotting and **torch.utils.data.DataLoader** for efficient data handling supports the robust testing and visualization of the network's performance and its interpretative capabilities through TDA. Moreover, we employed **kmapper** and **mapper interactive tool** to build the mapper graphs of the activation space.

4 Results and Discussion

This project explored the complex activation spaces within a simple neural network configured with ReLU activations, aiming to deepen our understanding of how such networks process and interpret data. The results presented herein not only confirm theoretical insights from current literature but also reveal novel observations regarding the network's behavior across varied datasets.
Visualization of Neural Activation
Figure 1 showcases the activation patterns within our neural network when trained on a two-dimensional input space. This visualization provides a dual perspective: the partitioning in the input domain and the transformed codomain.

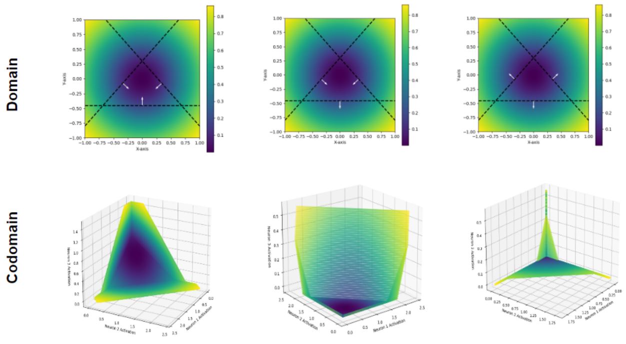
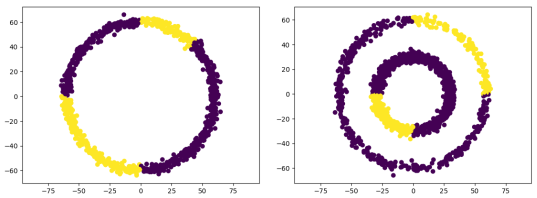


Figure 1: Top row: Partitioning of the two-dimensional input space by three neurons, demonstrating distinct activation regions. Bottom row: Corresponding 3D plots of the activation landscape, depicting the peaks (high activation) and plains (low activation) formed by each neuron.

Domain Analysis:
The images in the top row of Figure 1 display a clear segmentation of the input space. Each neuron carves out distinct regions where it predominates, similar to territories defined on a geographical map.
Codomain Analysis:
The bottom row of Figure 1 transitions these divisions into a three-dimensional codomain, where the tetrahedral structures represent the complex polyhedral shapes formed by neuron activations. These shapes depict how the network processes and categorizes different inputs, providing a geometric perspective on neural activation.
Interpretation of Results
The visualization of tetrahedral structures in the activation space is significant as it conveys the non-linear transformations that inputs undergo in a neural network. Each tetrahedron represents a region in the input space where the combination of neuron activations results in a specific output pattern:
Peaks and Valleys: The vertices and edges of the tetrahedra indicate points of high neural activity, whereas the faces and interiors might represent areas

of lesser activity. This topography illustrates the varying influence of neurons across the input space.

Implications for Classification: The sharp boundaries of the tetrahedra suggest that the network has learned to clearly distinguish between different types of inputs. However, the complexity of these shapes also hints at the potential challenges in generalizing beyond the training data, particularly in regions near the edges of the tetrahedra.



(a) Single Annulus (b) Double Annulus

Figure 2: (a) Single annulus dataset showing the clear division between two distinct classes. (b) The double annulus dataset illustrates a more complex scenario with two concentric classes, challenging the network's classification abilities.

Figure 2 illustrates the two different datasets: the single annulus dataset (a) and the double annulus dataset (b). These datasets were chosen to test the network's ability to classify complex geometric patterns, which involve discerning data points based on their positioning within concentric rings.

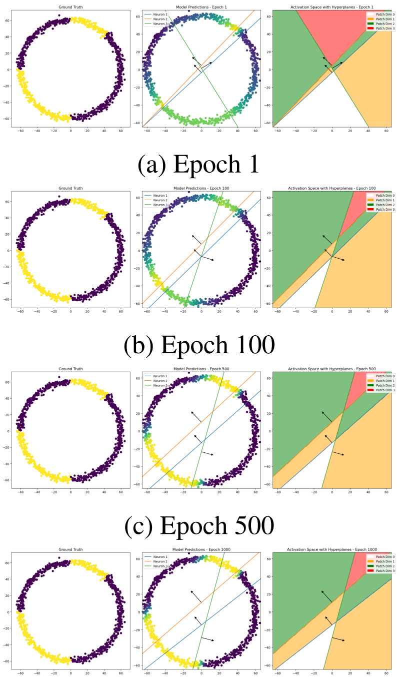
Single Annulus Analysis:

Figure 3 illustrates the progression of the neural network's learning over time, depicted through snapshots at various epochs (1, 100, 500, and 1000 epochs). These visualizations help elucidate how the neural model's predictions evolve as it trains on the single annulus dataset, particularly focusing on the decision boundaries established by its neurons and how they adapt to better classify the two distinct regions within the dataset.

The top right of Figure 3 showcases the corresponding activation space, now partitioned into polyhedra (2D space) by the neurons' activations. Each segment is shaded differently, which signifies a distinct area where specific neuron activations dominate, demonstrating how the network differentiates one class from another.

Figure 4 presents a detailed view of the neural network's understanding and classification capabilities through two different visualization techniques: the 3D Activation Space and a Topological Data Analysis (TDA) Mapper graph.

The 3D Activation Space (a) provides a vivid representation of how each of the three neurons responds to the dataset inputs, laid out across three dimensions. This visualization highlights distinct clusters that correlate with the ground truth classes. The separation of these clusters signifies the network's proficiency in learning and distinguishing between different classes:



(a) Epoch 1

(b) Epoch 100

(c) Epoch 500

(d) Epoch 1000

Figure 3: Progression of the neural network's learning and prediction accuracy over time, demonstrating the evolving precision in demarcating the inner and outer regions of the annulus. Each panel shows the decision boundaries at different epochs, indicating the refinement in the network's ability to distinguish between the two classes.

Distinct Clusters: The clearer and more separated the clusters, the more effectively the network has learned to categorize the inputs based on the learned features.

Neuron Activation Patterns: Each axis in the space represents the activation level of a neuron, showing how different neurons activate to varying degrees depending on the input, thus creating a multi-dimensional perspective of the data's characteristics.

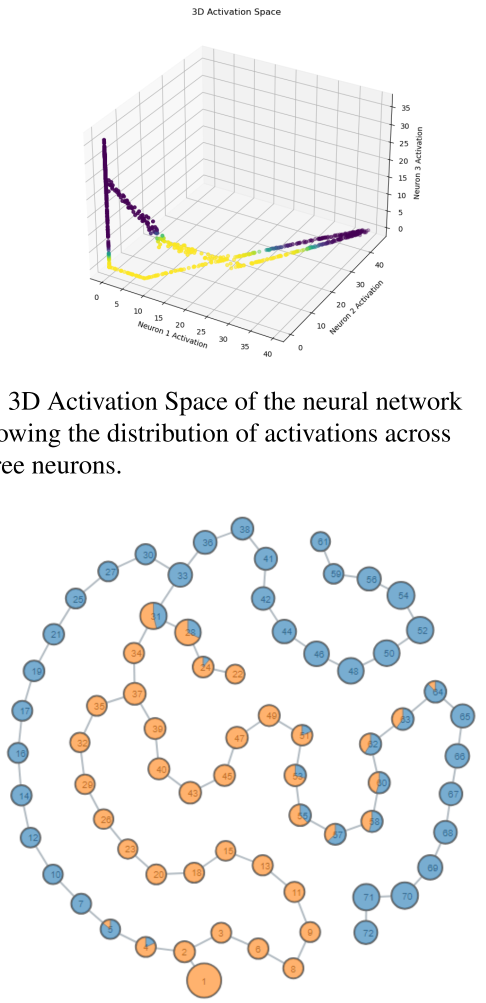
Mapper Graph Interpretation

The Mapper graph (Figure 4b) utilizes TDA to further elucidate the neural network's internal representations, translating the high-dimensional activation data into a simplified yet informative topological summary. This graph offers insights into how the neural network perceives and processes the data:

Node Representation: Each node in the graph represents a cluster of data points that have similar activation patterns, where the size of the node indicates the number of points within the cluster.

Color Coding: Nodes are color-coded to reflect their corresponding class in the ground truth, with orange representing the 'yellow' class and blue representing the 'purple' class from the dataset.

Mixed Nodes: The presence of mixed nodes where both colors appear indicates regions in the activation



(a) 3D Activation Space of the neural network showing the distribution of activations across three neurons.

(b) Mapper graph of the neural network's activation space, illustrating the clustering of data points and the topology of the dataset.

Figure 4: Visualization of the neural network's interpretation of data through activation patterns and topological summarization.

space where the prediction quality is ambiguous or where the network's decision boundaries are less clear. This mixed coloring mirrors the areas of potential confusion and highlights the need for further model tuning to enhance classification accuracy.

Comparative Analysis

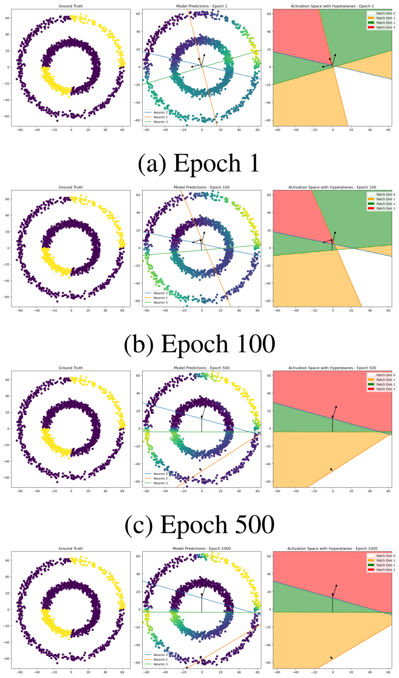
The side-by-side visualization of the 3D Activation Space and the Mapper graph allows for a comprehensive understanding of the network's processing dynamics. While the 3D space shows the direct response of each neuron, the Mapper graph abstracts these responses into a topological form, providing a macroscopic view of the data's topology and the network's interpretative patterns:

Pattern Similarity: Both visualizations, though different in form, depict similar patterns of clustering and separation, underscoring the consistency in how the network processes and understands the dataset.

Insight into Learning Dynamics: Together, they illustrate not only the end result of the learning process but also the evolution of the network's internal representations over time.

Double Annulus Analysis:

Figures 5 and 6 extend the analysis to a more complex double annulus dataset, displaying similar patterns in learning dynamics and topological data analysis visualizations as observed with the single annulus dataset. These figures further demonstrate the neural network's ability to adapt and refine its understanding of increasingly complex spatial distributions.



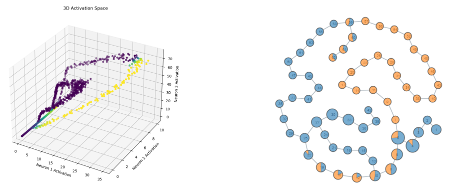
(a) Epoch 1

(b) Epoch 100

(c) Epoch 500

(d) Epoch 1000

Figure 5: Model predictions over various epochs demonstrating the learning progression on a double annulus dataset. Each panel shows how the model's predictions evolve, becoming more refined at each stage.



(a) 3D Activation Space (b) Mapper Graph

Figure 6: Visualization of the neural network's understanding and classification of the double annulus dataset through 3D activation space and Mapper graph analysis.

Analysis of Double Annulus Dataset Learning Progression (Figure 5): Across different epochs, the neural network shows marked improvement in identifying and classifying the two concentric rings of the double annulus dataset. Early epochs demonstrate broad and imprecise boundaries which progressively sharpen and become more accurate with further training.

3D Activation Space and Mapper Graph (Figure 6)

3D Activation Space: This visualization (Figure 6a) exhibits the activation levels of the three neurons in response to the dataset inputs. The 3D representation helps in understanding the complex interaction between neuron activations and their impact on the classification boundaries. Mapper Graph: Figure 6b provides a topological summary of the dataset, where nodes represent clusters of similar activation patterns. The graph depicts the separation between clusters associated with different classes and mixed nodes where classifications are ambiguous.

Implications of Findings

The analysis of the double annulus dataset emphasizes the network's capacity to handle complex, nested spatial structures. The progression in learning, observed through the epochs and visualized in both the activation space and the mapper graph, highlights:

Network Adaptability: Demonstrates the network's ability to adjust its internal representations to effectively manage more challenging classification tasks.

Topological Insights: The use of TDA techniques, such as the Mapper algorithm, provides deeper insights into the network's classification logic, revealing not just how well the network performs but why certain decisions are made.

Discussion

These patterns' consistency across the 3D Activation Space and the Mapper graph underscores the network's learning behaviors. The correct classifications and areas of confusion are faithfully translated from the activation space to the topological representation, providing a comprehensive view of the network's performance and areas ripe for refinement.

5 Conclusions and Future Directions

The exploration of neural network capabilities through both traditional model predictions and advanced topological data analysis has provided significant insights into the intricate processing patterns and decision-making processes within neural networks. The utilization of both 3D activation spaces and Mapper graphs has not only demonstrated the effectiveness of these models in classifying complex datasets but also highlighted areas for improvement and further research. Our studies have shown that with increased training and refinement, neural networks are capable of distinguishing and classifying complex spatial patterns with high accuracy. The progression in the network's ability to interpret the single and double annulus datasets illustrates its adaptability and the potential for deep learning models to handle similarly complex real-world tasks. The application of topological data analysis has also enriched our understanding by providing a new perspective on the internal dynamics of neural networks, revealing not just outcomes but the structural reasons behind those outcomes.

5.1 Future Directions

To build upon the findings and address the limitations identified during this research, we propose the following future directions:

Expand Dataset Diversity: Future research will benefit from incorporating a wider variety of datasets, including those with higher complexity and real-world applicability. Expanding the diversity of datasets will help in testing the robustness and adaptability of neural networks across different scenarios, which is critical for practical applications.

Enhance TDA Techniques: Improving the methods used in topological data analysis will allow for more nuanced insights into the high-dimensional spaces that neural networks operate within. This could involve integrating more sophisticated TDA tools or developing new algorithms that can more accurately capture the essence of neural activations.

Deepen Neural Network Analysis: Further studies should aim to deepen the analysis of neural networks by exploring different architectures, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), and their specific impacts on the classification tasks. Understanding how different architectures influence learning and decision-making will provide valuable guidelines for designing more effective neural models.

Real-time Visualization Tools: Developing real-time visualization tools for monitoring the training process and the evolution of neural network predictions can significantly enhance the understanding and debugging of these models. Such tools would provide immediate feedback on the effects of adjustments to the network's parameters, facilitating more dynamic and informed development cycles.

References

[1] M. Chaukair, C. Schütte, and V. Sunkara, "On the activation space of relu equipped deep neural networks," *Procedia Computer Science*, vol. 222, pp. 624–635, 2023.

[2] Y. Zhou, N. Chalapathi, A. Rathore, Y. Zhao, and B. Wang, "Mapper interactive: A scalable, extendable, and interactive toolbox for the visual exploration of high-dimensional data," in *2021 IEEE 14th Pacific Visualization Symposium (PacificVis)*. IEEE, 2021, pp. 101–110.

[3] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein, "Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability," *Advances in neural information processing systems*, vol. 30, 2017.

[4] Y. Jia, H. Wang, S. Shao, H. Long, Y. Zhou, and X. Wang, "On geometric structure of activation spaces in neural networks," *arXiv preprint arXiv:1904.01399*, 2019.