Topological Data Analysis for Neural Network Analysis: A Comprehensive Survey

Rubén Ballester RUBEN.BALLESTER@UB.EDU
Carles Casacuberta CARLES.CASACUBERTA@UB.EDU
Departament de Matemàtiques i Informàtica
Universitat de Barcelona
Gran Via de les Corts Catalanes, 585, 08007 Barcelona, Catalonia, Spain

Sergio Escalera SESCALERA@UB.EDU
Departament de Matemàtiques i Informàtica
Universitat de Barcelona
Gran Via de les Corts Catalanes, 585, 08007 Barcelona, Catalonia, Spain
Computer Vision Center
Edifici O, Campus UAB, 08193 Bellaterra, Catalonia, Spain

Abstract

This survey provides a comprehensive exploration of applications of Topological Data Analysis (TDA) within neural network analysis. Using TDA tools such as persistent homology and Mapper, we delve into the intricate structures and behaviors of neural networks and their datasets. We discuss different strategies to obtain topological information from data and neural networks by means of TDA. Additionally, we review how topological information can be leveraged to analyze properties of neural networks, such as their generalization capacity or expressivity. We explore practical implications of deep learning, specifically focusing on areas like adversarial detection and model selection. Our survey organizes the examined works into four broad domains: 1. Characterization of neural network architectures; 2. Analysis of decision regions and boundaries; 3. Study of internal representations, activations, and parameters; 4. Exploration of training dynamics and loss functions. Within each category, we discuss several articles, offering background information to aid in understanding the various methodologies. We conclude with a synthesis of key insights gained from our study, accompanied by a discussion of challenges and potential advancements in the field.

Keywords: Topological data analysis, persistent homology, Mapper, deep learning, neural networks, topological machine learning

1 Introduction

Over the past few years, deep learning has consolidated its position as the most successful branch of artificial intelligence. With the continuous growth in computational capacity, neural networks have expanded in size and complexity, enabling them to effectively tackle progressively difficult problems. However, their increased capacity has made it more challenging to comprehend essential properties of the networks such as their interpretability, generalization ability, or suitability for specific problems. From both theoretical and practical standpoints, this is undesirable, especially in critical contexts where AI decisions could

lead to catastrophic consequences, such as medical diagnosis (Yang et al., 2021) or autonomous driving (Wäschle et al., 2022), among others.

Topological Data Analysis (TDA) has emerged as a subfield of algebraic topology that offers a framework for gaining insights into the *shape* of data in a broad sense. Topological data analysis has found application across a wide array of experimental science disciplines, spanning from biomedicine (Skaf and Laubenbacher, 2022) to finance (Gidea and Katz, 2018), among numerous others. One of its most prolific areas of application is machine learning, particularly in the domain of deep learning. A basic introduction to topological machine learning can be found in Hensel et al. (2021). Topological data analysis, specifically homology, persistent homology and Mapper, has been used to analyze various aspects of neural networks. Broadly, these aspects can be categorized in the following four groups: 1. Structure of neural networks; 2. Input and output spaces; 3. Internal representations and activations; 4. Training dynamics and loss functions.

Figure 1 visually delineates these four categories. The first category involves examining unweighted graphs associated with neural networks and their properties, such as depth, layer widths, and graph topology, among others. The second category encompasses the analysis of neural network input and output spaces, including decision regions and boundaries for classification problems, as well as the study of latent spaces for generative models. The third category, which currently holds the largest number of contributions in the literature, focuses on the analysis of hidden and output neurons in a broad sense. Lastly, the fourth category involves the analysis of neural network training procedures, including the study of loss functions.

Many components studied within the preceding categories play a pivotal role in understanding some of the fundamental traits of deep learning, such as interpretability or the generalization capacity of neural networks. Moreover, these elements inherently exhibit geometrical and topological characteristics, rendering them exceptionally suitable for the application of topological data analysis methods.

1.1 Contribution

In this work, we offer a comprehensive overview of applications of topological data analysis in analyzing neural networks across the aforementioned four categories. However, we have omitted many relevant works related to the general use of topological data analysis in deep learning. For example, we omit work on the branch of Topological Deep Learning, which involves developing neural networks tailored for specific topological data. A recent survey on Topological Deep Learning is available in Papillon et al. (2023). We have also omitted the majority of applications using topological data analysis to construct loss functions, since our focus remains on the analysis of neural networks rather than their enhancement through the imposition of specific topological structures on data, unrelated to the particular machine learning algorithm used for the task. Given the substantial volume of papers analyzed, our discussion is limited to peer-reviewed papers, with occasional exceptions made for relevant and credible sources.

This survey aims to be self-contained and approachable for readers unfamiliar with topological data analysis or deep learning. However, it is advisable that readers have a background in at least one of these areas. For mathematicians interested in how topology

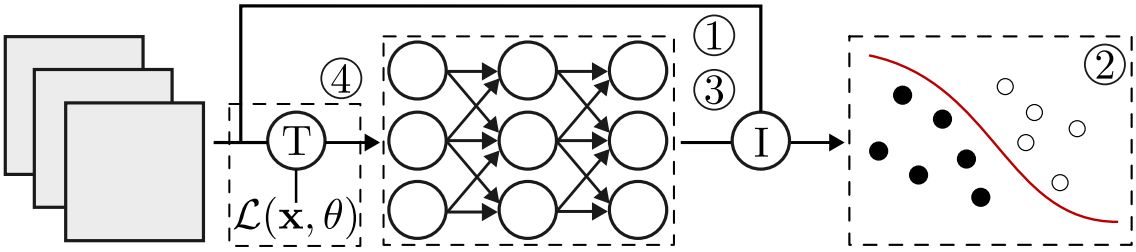

Figure 1: Diagram showing the usual lifecycle of a neural network $\mathcal{N}$. First, an architecture $a(\mathcal{N})$ is selected based on the task to be solved. This architecture is independent of the learned parameters $\theta(\mathcal{N})$ or the specific input data used to train or test the network, denoted $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$, respectively. Second, the architecture is trained (T) using a specific training algorithm $\mathcal{A}$, which generally minimizes the empirical risk of a loss function $\mathcal{L}$ evaluated on the training dataset $\mathcal{D}_{\text{train}}$. Once the network is trained, inference (I) is performed with data coming from the same distribution $\mathbb{P}$ from which the training data were sampled. For trained neural networks, input and output spaces gather several interesting structures, such as decision regions and boundaries for classification problems or latent spaces for generative models, among others. Each dashed box is related to one of the categories, labeled 1 to 4, in which topological data analysis has been used to analyze neural networks. The categories are the following: (1) Structure of the neural network; (2) Input and output spaces; (3) Internal representations and activations; (4) Training dynamics and loss functions. The leftmost box contains the training part of the lifecycle of a neural network and is related to category 4. The central box contains the neural network and is concerned with categories 1 and 3. The rightmost box contains the decision regions and boundaries of the output space of a neural network after training, which are related to category 2.

can be applied to study modern AI and deep learning systems, we recommend reading the first chapter of the book by Grohs and Kutyniok (2022), that provides a concise introduction to deep learning from a mathematical viewpoint. For machine learning scientists curious about how advanced topological methods can enhance the understanding of deep learning systems, we recommend the survey on Topological Machine Learning by Hensel et al. (2021), where topological data analysis is introduced in a friendly manner within the scope of machine learning.

1.2 Outline

In Section 2 we introduce notation, definitions, and basic results of topological data analysis and deep learning needed to follow the survey. Section 3 constitutes the core content of our survey, structured into the four aforementioned categories. Finally, Section 4 addresses limitations, challenges, and potential future directions concerning the application of TDA in deep learning up to the time of publication of this survey.