Topological Data Analysis for Trustworthy AI

Victor Toscano Durán[1]

[1]Department of Applied Mathematics I, University of Sevilla, Sevilla, Spain

Abstract

Artificial Intelligence (AI) is transforming industries by analyzing large amounts of data to find patterns and make decisions more efficiently than ever before. Neural networks, which are inspired by the human brain, are a key part of AI but often work in ways that are hard to understand, leading to concerns about their reliability. This doctoral research proposal, titled "Topological Data Analysis for Trustworthy AI," aims to tackle these issues by using Topological Data Analysis (TDA) and Computational Topology. The research will develop a new method to compare piecewise neural networks with ReLU activation functions using topological entropy, which could help make these networks more transparent. It will also apply TDA techniques to improve the analysis of time series data in neural networks, aiming to enhance prediction accuracy and understanding of how these networks work over time. Additionally, the study will look at applying TDA to recurrent neural networks like LSTM and potentially to Transformer models. This research aims to make AI systems more reliable and understandable, with benefits for areas like healthcare and autonomous systems. The proposal also includes plans for attending conferences and publishing research findings.

Keywords

Artificial Intelligence, Neural Networks, Topological Data Analysis, Time series, reliability

1. Context and Motivation

Artificial Intelligence (AI) [1] stands at the forefront of technological advancement, reshaping industries, economies, and our interaction with the world. From virtual assistants like Siri and Alexa to self-driving cars and advanced medical diagnostics, AI is pervasive in modern life. Its growth trajectory has been nothing short of remarkable, with exponential leaps in capability and application.

The importance of AI lies in its ability to process vast amounts of data, identify patterns, and make decisions with a level of efficiency and accuracy unmatched by human counterparts. This capacity has propelled AI into domains once deemed exclusive to human intelligence, from natural language processing and image recognition to strategic decision-making and problem-solving.

Central to the functionality of AI systems are neural networks [2], inspired by the biological neural networks of the human brain. These interconnected layers of nodes, or artificial neurons, simulate the processes of learning and adaptation through iterative training on labeled datasets.

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

vtoscano@us.es (V. T. Durán)

https://victosdur77.github.io/ (V. T. Durán)

0009-0006-1316-9026 (V. T. Durán)

As more data is fed into the network, it adjusts its internal parameters to optimize performance, enabling it to recognize complex patterns and make predictions with increasing accuracy.

However, alongside its tremendous potential, AI also presents significant challenges, chief among them being the issue of reliability, particularly concerning so-called "black box" algorithms [3]. While neural networks excel at solving complex problems, their inner workings often remain opaque to human understanding. This lack of transparency raises concerns regarding the reliability and trustworthiness of AI systems, especially in high-stakes applications such as autonomous vehicles, medical diagnosis, and financial forecasting.

The problem of reliability in AI underscores the need for transparency and accountability in algorithmic decision-making. Efforts to address this issue include research into explainable AI, which aims to develop models that not only produce accurate results but also provide insights into the reasoning behind those decisions. By making AI systems more interpretable and understandable to human users, researchers hope to build trust and mitigate the risks associated with opaque algorithms.

Despite these challenges, the potential benefits of AI are undeniable, with far-reaching implications for virtually every sector of society. From improving healthcare outcomes and optimizing resource allocation to enhancing cybersecurity and mitigating climate change, AI offers solutions to some of humanity's most pressing problems. As we continue to harness the power of artificial intelligence, it is essential to remain vigilant, balancing innovation with ethical considerations and ensuring that AI serves the collective good.

In the ever-evolving landscape of artificial intelligence (AI), where innovation is the norm and breakthroughs are constant, emerging fields like Topological Data Analysis (TDA) [4, 5] and Computational Topology [6] are gaining recognition for their potential to augment the efficiency and capabilities of neural networks and AI systems as a whole.

At its core, TDA is a branch of mathematics that leverages tools from algebraic topology to analyze the shape, structure, and connectivity of complex data sets. By applying topological principles to high-dimensional data, TDA seeks to extract meaningful insights that may be obscured by traditional statistical or geometric methods. This approach allows researchers and practitioners to uncover hidden patterns, identify critical features, and gain a deeper understanding of the underlying structure inherent in the data.

Computational Topology, on the other hand, focuses on the development and implementation of algorithms and computational techniques for solving topological problems. It bridges the gap between theoretical concepts in topology and practical applications in fields such as computer science, engineering, and data analysis. Through the use of advanced computational tools, Computational Topology enables researchers to tackle complex problems in data analysis, visualization, and machine learning [2, 7].

One of the most promising aspects of TDA and Computational Topology is their potential to enhance the efficiency and effectiveness of neural networks and AI algorithms. By incorporating topological insights into the design and training of neural networks, researchers can develop more robust and adaptive models capable of handling diverse and complex data sets. TDA techniques such as persistent homology have been successfully applied to tasks such as image recognition, natural language processing, and time-series analysis, demonstrating their efficacy in extracting meaningful features and improving classification accuracy.

In recent years, there has been a growing interest in interdisciplinary research at the intersec-

tion of TDA, Computational Topology, and artificial intelligence. Collaborative efforts between mathematicians, computer scientists, and domain experts have yielded novel approaches and techniques for solving complex problems in data analysis and machine learning. This convergence of disciplines holds great promise for advancing the capabilities of AI systems and unlocking new opportunities for innovation across a wide range of applications.

As we continue to explore the synergies between TDA, Computational Topology, and artificial intelligence, it is clear that these fields will play an increasingly important role in shaping the future of data-driven decision-making, enabling more efficient, reliable, and interpretable AI systems.

In summary, in this doctoral proposal named "Topological Data Analysis for Trustworthy AI", I will focus on the application of Topological Data Analysis and Computational Topology as a fundamental tool for improving the reliability of artificial intelligence in challenging contexts.

2. Related Work

In recent years, there has been a surge of interest in integrating artificial intelligence (AI) with topological data analysis (TDA) to enhance the efficiency, robustness, and interpretability of AI systems. This section explores some of the key contributions and advancements in this interdisciplinary research area.

Enhanced Data Analysis with Topological Summaries

An emerging focus is the application of topological summaries to improve data analysis itself. These summaries are mathematical tools used in topological data analysis (TDA) to capture and characterize the underlying structure of complex data sets by focusing on the intrinsic structure of complex datasets, rather than relying on traditional geometric methods. These summaries, rooted in concepts like homology and persistent homology, capture the fundamental shapes and features of data, providing stable representations that resist noise and variations. By integrating these topological descriptors, researchers can gain deeper insights into the data, leading to more informed decisions and enhanced analysis outcomes.

TDA for Feature Extraction in AI

Topological Data Analysis (TDA) has proven to be a powerful method for extracting meaningful features from high-dimensional data, which traditional techniques often overlook. Persistent homology, a key TDA tool, captures stable topological features like connected components and loops across multiple scales, making it effective for AI tasks such as image recognition. By identifying essential features that enhance classification accuracy, TDA has been successfully applied in areas like object detection and texture classification, where the data's inherent shape is critical for distinguishing categories.

Topological Representations for Neural Networks

Incorporating topological representations into neural network design and training offers promising improvements in generalization, overfitting reduction, and interpretability. Topological regularization, which imposes topological constraints during learning, helps neural networks capture essential data structures, stabilizes training, and increases resilience against adversarial attacks. Additionally, using topological insights to refine decision boundaries enhances the robustness and reliability of AI models, contributing to the development of more effective and interpretable neural networks.

Interpretable AI with TDA

The integration of Topological Data Analysis (TDA) into AI models has advanced the field of interpretable AI by making complex systems more transparent and accountable. TDA-based methods provide topological explanations for AI decisions, offering insights into how data features influence predictions, particularly in critical areas like healthcare, finance, and autonomous systems. This aligns with the goals of explainable AI (XAI), where the focus extends beyond performance to include the interpretability and trustworthiness of AI outputs, addressing the increasing demand for transparency in AI technologies.

Through these advancements, TDA not only contributes to the development of more efficient and robust AI systems but also addresses the growing demand for interpretability and transparency in AI technologies.

3. Research Questions, Hypothesis, and Objetives

In this project, we embark on the application of topology as a fundamental tool to enhance the reliability of neural networks in challenging contexts. I aim to initially focus on achieving satisfactory results in the previous research done by my advisors, and then applying Topological Data Analysis (TDA) techniques to analyze time series data in neural networks, aiming to improve prediction accuracy and understand temporal dynamics.

Questions

1. How can topology be leveraged to improve the reliability of neural networks in challenging contexts?
2. What role does topological entropy play in measuring similarities between piecewise neural networks using activation functions like ReLU?
3. How can TDA be extended to analyze time series data in neural networks, and what insights can be gained from this analysis?

Hypothesis

1. Piecewise neural networks employing ReLU activation functions can be evaluated for similarity using topological entropy, leading to greater transparency in their operation.

2. The application of TDA to time series analysis in neural networks will yield valuable insights into the temporal dynamics of network behavior and improve predictive performance.

Objectives

Firstly, I will extend previous research carried out by my advisors in [8] on piecewise neural networks [9], particularly those using ReLU activation functions, by developing a new approach based on topological entropy to measure similarities between these networks. In addition, evaluate the effectiveness of the proposed approach in improving the transparency and reliability of piecewise neural networks. In summary, the research conducted by my advisors in [8] will be the starting point of my thesis.

Secondly, my research will extend into the application of Topological Data Analysis (TDA) techniques to the analysis of time series data within neural networks. This will involve leveraging existing knowledge in time series analysis and integrating recent advancements in the field. A significant aspect of this part of the research will be to explore how incorporating TDA can improve prediction accuracy and provide deeper insights into temporal dynamics. For that, topological descriptors, also known as topological summaries, will be used, which are mathematical tools used in topological data analysis to capture and characterise the underlying structure of complex data sets. Unlike traditional data analysis methods that rely on specific geometry and Euclidean metrics, topological descriptors focus on the intrinsic properties of the data space, providing a robust and stable representation in the face of noise and variations in the data. Additionally, I will evaluate the applicability of TDA techniques to recurrent neural networks, such as Long Short-Term Memory (LSTM) networks [10, 11, 12, 13, 14, 15, 16, 17, 18, 19].

Finally, contingent on achieving positive results and having sufficient time, there may be an opportunity to expand the research to include "Transformers" models [20, 21].

Moreover, as part of my professional development, I intend to attend numerous conferences to stay abreast of the latest advancements and network with peers in the field. I also plan to write articles related to my thesis and present them at conferences to contribute to the academic community.

Stay

I have arranged two research stays as part of my thesis. The first will take place in October 2024 at the Institute of Electronic, Informatics and Telecommunications Engineering, National Research Council in Genoa, under the guidance of Prof. Maurizio Mongelli, who specializes in machine learning applied to bioinformatics and cyber-physical systems. During this stay, I will focus on advancing my understanding of machine learning techniques, particularly in the context of explainable AI. The second stay is planned for Summer 2025 at Bastian Grossenbacher Rieck's laboratory in Helmholtz Munich, a leading center for machine learning research, especially in computational healthcare. There, I will work with the AIDOS Lab, under the guidance of Prof. Bastian Rieck [1], who is focus on geometry and topology in machine learning with a keen

[1] https://bastian.rieck.me/

interest in biomedical applications, to deepen my knowledge of topological machine learning techniques and their application in healthcare.

4. Research Approach and Methods

Approach

The research approach for this study combines theoretical exploration, algorithm development, and empirical validation to investigate the application of topological data analysis (TDA) in enhancing the reliability and transparency of neural networks, particularly in challenging contexts.

Methods

The methodological approach of this research involves three main phases:

- First, a theoretical exploration will be conducted through a comprehensive review of existing literature on TDA, neural networks, and topological concepts like persistent homology and topological entropy, aiming to develop new methodologies for assessing neural network reliability and enhancing interpretability.
- The second phase focuses on the development of novel algorithms to apply TDA to neural network architectures, designing techniques for measuring similarities using topological summaries like persistent entropy, with particular emphasis on ReLU networks. Moreover, this phase will focus in extend TDA techniques to analyze time series data and integrating them into neural networks for time series, like recurrent neural networks.
- Finally, empirical validation will be carried out by implementing these algorithms on real-world datasets to evaluate their effectiveness in improving the reliability, interpretability, and performance of neural networks, and comparing the outcomes with traditional approaches.

The underlying hypothesis is that integrating TDA techniques, which offer a unique perspective on data structure and relationships, can overcome the limitations of traditional neural networks, especially in complex and high-dimensional data domains, thereby enhancing their performance and transparency, and enhance the analysis of time series.

5. Preliminary Results and Contributions

The research is currently in its early stages, focusing on a comprehensive literature review to identify relevant methodologies and theoretical frameworks for integrating Topological Data Analysis (TDA) with neural networks. While concrete results are not yet available, initial findings suggest promising avenues for enhancing AI interpretability and reliability through topological methods, establishing a strong foundation for future empirical research and experimentation.

6. Expected Next Research Steps

We plan to refine our approach for measuring similarities between piecewise neural networks using topological entropy. This involves enhancing mathematical models that quantify the relationships between different neural network components based on their topological properties, aiming to develop a robust metric that better captures the complexity and behavior of these networks. Additionally, we will explore advanced techniques for applying topological data analysis (TDA) to time series within neural networks, with the goal of improving the predictive power and interpretability of AI models.

7. Expected final contribution to knowledge

The expected outcome of this research is a significant contribution to the field of AI, particularly in the areas of reliability, transparency, and interpretability. By integrating Topological Data Analysis with neural networks, the research aims to produce AI systems that are not only more accurate but also more understandable to human users. This integration has the potential to revolutionize how neural networks are designed and applied, particularly in high-stakes areas where trust and transparency are paramount. Ultimately, the research aspires to bridge the gap between complex AI models and human interpretability, contributing to the development of AI systems that are both powerful and ethically sound.

Acknowledgments

Thanks to my thesis tutor, Rocío González Díaz, for her invaluable help, advice, and for this incredible opportunity, and to my thesis supervisors, Miguel Ángel Gutiérrez Naranjo and Matteo Rucco, for their guidance and support. This work was supported in part by the European Union HORIZON-CL4-2021-HUMAN-01-01 under grant agreement 101070028 (REXASI-PRO) and by TED2021-129438B-I00 / AEI/10.13039/501100011033 / Unión Europea NextGenerationEU/PRTR.

References

[1] P. Winston, Artificial Intelligence, A-W Series in Computerscience, Addison-Wesley Publishing Company, 1992. URL: https://books.google.es/books?id=b4owngEACAAJ.

[2] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–44. doi:10.1038/nature14539.

[3] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature (2019). doi:10.1038/s42256-019-0048-x.

[4] G. Carlsson, Topology and data, Bulletin of The American Mathematical Society - BULL AMER MATH SOC 46 (2009) 255–308. doi:10.1090/S0273-0979-09-01249-X.

[5] F. Chazal, B. Michel, An introduction to topological data analysis: Fundamental and practical aspects for data scientists, Frontiers in Artificial Intelligence 4 (2017). doi:10.3389/frai.2021.667963.

[6] H. Edelsbrunner, J. Harer, Computational Topology: An Introduction, Springer, 2010. doi:10.1007/978-3-540-33259-6_7.

[7] G. F. Marcus, Deep learning: A critical appraisal, ArXiv abs/1801.00631 (2018). doi:10.48550/arXiv.1801.00631.

[8] M. Rucco, R. Gonzalez-Diaz, M.-J. Jimenez, N. Atienza, C. Cristalli, E. Concettoni, A. Ferrante, E. Merelli, A new topological entropy-based approach for measuring similarities among piecewise linear functions, Signal Processing 134 (2017) 130–138. doi:10.1016/j.sigpro.2016.12.006.

[9] Q. Tao, L. Li, X. Huang, X. Xi, S. Wang, J. Suykens, Piecewise linear neural networks and deep learning, Nature Reviews Methods Primers 2 (2022) 42. doi:10.1038/s43586-022-00125-7.

[10] Y. Zhou, Persistent homology on time series, 2016. doi:https://doi.org/10.7939/R3K931F13.

[11] N. Ravishanker, R. Chen, An introduction to persistent homology for time series, WIREs Computational Statistics 13 (2021). doi:10.1002/wics.1548.

[12] S. C. di Montesano, H. Edelsbrunner, M. Henzinger, L. Ost, Dynamically maintaining the persistent homology of time series, ArXiv abs/2311.01115 (2023). doi:10.48550/arXiv.2311.01115.

[13] C. M. Pereira, R. F. de Mello, Persistent homology for time series and spatial data clustering, Expert Systems with Applications 42 (2015) 6026–6038. doi:10.1016/j.eswa.2015.04.010.

[14] T. Ichinomiya, Time series analysis using persistent homology of distance matrix, Nonlinear Theory and Its Applications, IEICE 14 (2023) 79–91. doi:10.1587/nolta.14.79.

[15] Y.-M. Chung, W. Cruse, A. Lawson, A persistent homology approach to time series classification (2020). doi:10.48550/arXiv.2003.06462. arXiv:2003.06462.

[16] L. Leaverton, Analysis of financial time series using tda: theoretical and empirical results, 2020. URL: http://hdl.handle.net/2445/163638.

[17] G. Ma, Using topological data analysis to process time-series data: A persistent homology way, Journal of Physics: Conference Series 1550 (2020) 032082. doi:10.1088/1742-6596/1550/3/032082.

[18] Y. Umeda, J. Kaneko, H. Kikuchi, Topological data analysis and its application to time-series data analysis, Fujitsu Scientific and Technical Journal 55 (2019) 65–71.

[19] M. L. Sánchez, Persistent homology: Functional summaries of persistence diagrams for time series analysis, 2021. URL: http://hdl.handle.net/2445/181324.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., 2017, p. 6000–6010. doi:10.48550/arXiv.1706.03762.

[21] R. Reinauer, M. Caorsi, N. Berkouk, Persformer: A transformer architecture for topological machine learning (2022). doi:10.48550/arXiv.2112.15210. arXiv:2112.15210.