Exploring Topological Properties in Artificial Neural Networks

Sophia Pi
Northwestern University Department of Computer Science
633 Clark Ave
Evanston, IL 60208 USA
sophiapi2026.1@u.northwestern.edu

Abstract

Biological neural systems often represent information on low-dimensional manifolds that reflect the topology of their encoded variables. This suggests that neural activity can be naturally organized in geometrically meaningful ways, as seen in rodent head direction cells forming circular manifolds. This proposal examines whether artificial neural networks (ANNs) trained on tasks with well-defined topologies—such as planar or spherical coordinates from autonomous driving datasets like Apolloscape, cyclic temporal variables, or graph-structured road networks—develop similar low-dimensional representations aligned with the variables' inherent topology. We consider convolutional and vision transformer models for image data, graph neural networks for road network graphs, and 3D or point-based models for LIDAR point clouds, analyzing their internal activations with dimensionality reduction and topological data analysis. If successful, this approach not only elucidates the nature of internal representations in ANNs but also offers insights into the computational principles that bridge artificial systems and biological cognition.

Introduction

Neural activity in biological systems often resides on low-dimensional manifolds that reflect the topology of the encoded variables. For instance, head direction (HD) cells in rodents produce population activity tracing a one-dimensional loop, capturing the circular nature of head orientation (Peyrache et al. 2015; Chaudhuri et al. 2019). Such alignments suggest that ensembles of neurons—biological or artificial—might represent information in ways that preserve intrinsic topological properties.

We ask if artificial neural networks (ANNs) can develop similar topological alignments. Consider a vision model trained on geographical tasks: if the model's internal representations map onto a manifold reflecting planar or spherical coordinates, it parallels biological strategies for representing space. Datasets like Apolloscape (Huang et al. 2018), with dashcam images, LIDAR data, and road network graphs, offer natural testbeds to investigate such questions. Beyond computer vision, tasks involving temporal cycles (e.g., day of the week) or three-dimensional orientations prompt investigations into how networks encode loops, spheres, and other topological structures.

If ANNs reveal manifold structures aligned with topology, it suggests a deeper connection to biological cognition, where neural codes also reflect such intrinsic geometric properties. Understanding these correspondences may enhance our ability to interpret and improve ANN architectures and bridge insights between neuroscience and AI research.

Background

In neuroscience, manifold-like representations have been found in multiple systems: rodent HD cells encode angles as loops (Peyrache et al. 2015; Chaudhuri et al. 2019), motor cortex dynamics fall into low-dimensional embeddings tied to action sequences (Elsayed et al. 2016), and hippocampal representations capture geometric abstractions (Bernardi et al. 2020). Theoretical work has explored how intrinsic dimensionality and embedding shape relate to neural computation (Jazayeri and Ostojic 2021).

In deep learning, representations often become low-dimensional in deeper layers (Ansuini et al. 2019), and manifold geometry evolves across layers to support complex classification tasks (Cohen et al. 2020). It remains unclear, however, if these low-dimensional embeddings can reflect topologies inherent in the data, as seen in biological systems. Prior work on language models or molecular embeddings found structured low-dimensional representations but did not explicitly test for topological correspondences.

This project focuses on tasks with known topologies to test the hypothesis more directly. Autonomous driving data from Apolloscape (Huang et al. 2018) provides camera images suitable for convolutional neural networks (CNNs) or vision transformers (ViTs), and three-dimensional LIDAR point clouds for 3D CNNs or point-based architectures. Road network graphs can be processed by graph neural networks (GNNs), potentially revealing embeddings that reflect graph loops or connectivity. Time-series models like RNNs, LSTMs, or transformers can capture periodic patterns and may yield loop-like internal state trajectories. Identifying such structures in ANNs would strengthen analogies to biological cognition, suggesting common computational principles underlying representation formation.

Approach

We will select tasks with clear topologies and train models suited to the data. For planar or spherical structures (e.g., mapping coordinates in an urban environment), we will use CNNs or ViTs on images from Apolloscape, which includes diverse urban scenes with known GPS locations. This setup tests if the model's internal activations form manifolds aligned with underlying spatial topologies.

When processing road network data, GNNs allow direct encoding of graph structure, providing a natural substrate for discovering loop-like embeddings if the data include closed routes. For LIDAR data, which represent 3D geometry, 3D CNNs or point-based models (e.g., PointNet) can be used to test whether representations align with continuous surfaces or spherical topologies. For time-based cyclic tasks, like predicting the day of the week from sequences, recurrent or transformer-based sequence models can be used, examining whether their hidden states form closed manifold loops.

Following training, we will extract intermediate representations and analyze their geometric organization. Dimensionality reduction methods such as PCA, t-SNE, UMAP, Isomap, and LLE will help visualize and identify whether data cluster on low-dimensional manifolds reflecting the task's topology. Persistent homology and Betti numbers from topological data analysis (TDA) will quantify topological features, providing rigorous criteria to evaluate whether observed embeddings align with the expected topology.

Evaluation

The evaluation will focus on both qualitative visualization and quantitative topological analysis. Dimensionality reduction techniques (PCA, t-SNE, UMAP, Isomap, LLE) will help visualize extracted representations. For instance, if a network trained on a circular variable exhibits a ring-like structure in a reduced embedding space, this is a strong qualitative indicator of topological alignment.

We will then apply topological data analysis (TDA) methods, such as persistent homology, to assess the presence and prominence of topological features like loops or holes in the learned representations (Carlsson 2009; Edelsbrunner and Harer 2010). By computing Betti numbers at various scales, we can quantitatively confirm the dimensionality and connectivity of the manifold. For a circular topology (e.g., a cyclic time variable), we expect to find a single prominent one-dimensional hole; for a spherical topology, we may find the characteristic topological invariants of a sphere. When processing Apolloscape images with a CNN or ViT, if the network's intermediate activations cluster on a manifold that resembles a 2D surface, Betti numbers may characterize its connectivity and dimensionality.

Statistical validation, including nonparametric tests or bootstrapping (Efron and Tibshirani 1994), will ensure observed patterns are not due to chance. Comparisons with control experiments—where the network is trained on tasks lacking intrinsic topological structure—will help confirm that topological features arise due to the inherent nature of the data. If we observe consistent and statistically significant topological invariants that match theoretical expectations (e.g., circular manifolds for cyclic tasks, loops corresponding to road network structure, etc.), the evaluation criteria will be met.

Discussion

If we find that ANNs form embeddings aligned with a task's topology, it could suggest that artificial systems and biological networks share underlying principles for representing complex variables. Such a result would indicate that, like biological neurons that naturally cluster activity on low-dimensional manifolds matching the structure of encoded features, ANNs can discover and leverage the topology inherent in their training data. This similarity may help explain why certain types of representations are so effective for tasks involving spatial, spherical, or cyclic variables and could guide the development of new architectures that intentionally incorporate topological constraints.

Alternatively, if ANNs don't exhibit such alignment, it may indicate that they do not naturally develop topologically aligned structures under current training regimes. New approaches, such as regularizers or specialized loss terms might be introduced to promote topological coherence in the latent space, guiding the network to represent cyclic or spherical variables as loops or spheres in its intermediate layers (Rifai et al. 2011). These interventions could improve the interpretability and stability of the learned representations, potentially leading to better performance on tasks requiring a deep understanding of the underlying structure.

Positive results could bridge part of the gap between neuroscience and AI (Kriegeskorte and Diedrichsen 2019), highlighting that both fields might converge on common strategies for information encoding. Such parallels can support the application of neuroscientific principles to AI, leading to models that mimic brain-like efficiency and adaptability. They also have significant implications for interpretability. If we know that certain layers represent features in a topologically meaningful manner, we can better understand the network's internal reasoning, improving transparency.

Conclusion

This work aims to determine whether internal representations in ANNs trained on tasks with known topologies reflect those underlying structures. By applying CNNs, ViTs, GNNs, 3D models, and sequence models to datasets like Apolloscape and tasks involving cyclic or spherical variables, and rigorously evaluating with dimensionality reduction and topological data analysis, we seek evidence of topologically grounded representations. Confirming such structures would align artificial neural computation with biological principles and illuminate fundamental aspects of how brains and machines organize information.

Acknowledgements

I would like to express my sincere gratitude to my mentor, Jeshwanth Challagundla, for his invaluable guidance and support throughout this project, and to Dr. Jason Grant, Chair of the AAAI-UC, for his leadership and dedication to fostering undergraduate research.

References

Ansuini, A.; Laio, A.; Macke, J. H.; Zoccolan, D.; and Stella, L. 2019. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32: 6109–6119.

Bernardi, S.; Benna, M. K.; Rigotti, M.; Munuera, J.; Fusi, S.; and Salzman, C. D. 2020. The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell*, 183(4): 954–967.e21.

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

Chaudhuri, R.; Gerçek, B.; Pandey, B.; Peyrache, A.; and Fiete, I. 2019. The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep. *Nature Neuroscience*, 22(9): 1512–1520.

Cohen, U.; Chung, S.; Lee, D. D.; and Sompolinsky, H. 2020. Separability and geometry of object manifolds in deep neural networks. *Nature Communications*, 11(1): 746.

Edelsbrunner, H.; and Harer, J. L. 2010. *Computational topology: an introduction*. American Mathematical Society.

Efron, B.; and Tibshirani, R. J. 1994. *An introduction to the bootstrap*. CRC press.

Elsayed, G. F.; Lara, A. H.; Kaufman, M. T.; Churchland, M. M.; and Cunningham, J. P. 2016. Reorganization between preparatory and movement population responses in motor cortex. *Nature Communications*, 7: 13239.

Huang, Q.; Wang, W.; Zhang, Y.; Wu, L.; Thalmann, N.; Chen, D.; and Yu, L. 2018. The Apolloscape Dataset for Autonomous Driving and its Application to Localization and 3D Scene Understanding. *CVPR Workshop on Autonomous Driving*.

Jazayeri, M.; and Ostojic, S. 2021. Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. *Current Opinion in Neurobiology*, 70: 113–120.

Kriegeskorte, N.; and Diedrichsen, J. 2019. Peeling the onion of brain representations. *Annual Review of Neuroscience*, 42: 407–432.

Peyrache, A.; Lacroix, M. M.; Petersen, P. C.; and Buzsáki, G. 2015. Internally organized mechanisms of the head direction sense. *Nature Neuroscience*, 18(4): 569–575.

Rifai, S.; Vincent, P.; Muller, X.; Glorot, X.; and Bengio, Y. 2011. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th International Conference on Machine Learning*, 833–840.