Topology in Neural Machine Translation: A Topological Study of Transformers through Attention

Yuwei (Johnny) Meng

23 Dec 2025

Abstract

1. Introduction

2. Background

2.1. Algebraic Topology

Topology is a branch of mathematics that characterizes shapes, spaces, and sets by their connectivity (Guss & Salakhutdinov, 2018). Algebraic topology, more sophisticatedly, is a subfield of topology that attributes algebraic properties such as groups and chains to topological spaces in order to make explanations and interpretations more expressive. Formally, let X be a compact metric space. We can define a p-simplex to be a collection of points $\{x_0, \cdots, x_n\} \subseteq X$ in p-dimension. Depending on the value of p, these simplices bear different names:

- $p = 0$: point
- $p = 1$: line
- $p = 2$: triangle
- $p = 3$: tetrahedron
- $\cdots$

Now consider a collection of such p-simplices, called $\mathcal{K}$. Then $\mathcal{K}$ is called a *simplicial* complex if it satisfies these conditions:

1. If $\sigma \in \mathcal{K}$ and τ is a face of σ, then $\tau \in \mathcal{K}$;
2. If $\sigma_1, \sigma_2 \in \mathcal{K}$, then $\sigma_1 \cap \sigma_2 = \emptyset$ or $\sigma_1 \cap \sigma_2 \in \mathcal{K}$.

Given a simplicial complex $\mathcal{K}$ in the compact metric space X, one method that is frequently used to study $\mathcal{K}$ is homology. The core idea of homology is to construct chains, cycles, and boundaries from the simplices in $\mathcal{K}$ and analyze their relationships. Given a dimension n, the nth homology group of the compact metric space X is defined as $H_n(X) = \mathbb{Z}^{\beta_n}$, where β_n is called the *nth Betti number*. For $n \geq 1$, the nth Betti number β_n measures the number of n-dimensional holes in the space X, while β_0 measures the number of connected components in X. For example, a torus is a 3-dimensional object with 1 connected component, 2 1-dimensional holes, and 1 2-dimensional void. Therefore, the homology groups of the torus are $H_0(X) = \mathbb{Z}^1$, $H_1(X) = \mathbb{Z}^2$, and $H_2(X) = \mathbb{Z}^1$.

2.2. Persistent Homology

Realistically, given a collection of n-dimensional points in $\mathbb{R}^n$, we would like to extract meaningful topological information that characterizes these points. Persistent homology is thus one method that computes topological characteristics of this collection of points. Given a collection of points P, we first construct a simplicial complex $\mathcal{K}$ known as the Vietoris-Rips (VR) complex. The vertices in $\mathcal{K}$ are just the points in P. To build the edges that connect the points, we consider an increasing sequence of radii. For each radius r in the sequence, we superimpose a circle of radius r on each point in P. If for some radius r_1 the circle at point x_1 starts to cover another point x_2, then we connect x_1 and x_2 by an edge at r_1.

Notice that as the radius r increases, more and more edges would be connected, leading to emergence and disappearance of topological features. We thus can characterize these topological features by their emergence time and disappearance time, or birth time and death time using standard topology terminology. For instance, recall our previous example concerning points x_1 and x_2. If a 1-dimensional hole α emerges because of the addition of the edge between them at r_1, then we would say that α has a birth time of r_1. Similarly, if at some later radius $r_2 > r_1$ that α disappears because of adding another edge in the simplicial complex, then we would denote r_2 as the death time of α.

Figure 1 below shows a visualization of computing persistent homology using the method described above on a set of points in $\mathbb{R}^2$. The left figure shows the sequence of radii increasing from 0 to 6.15, along which edges are added to the simplicial complex. Note that at $r = 5.6$, the addition of the edge between p_1 and p_3 make the simplicial complex into a quadrilateral, which results in a 1-dimensional hole. Subsequently, at $r = 6.15$, the quadrilateral is destroyed by the edge between p_1 and p_4, causing the 1-dimensional hole to disappear. The simplicial complex constructed by increasing the radius r is called a *filtration*. On the other hand, the right figure shows the persistence diagram of this filtration. Note that the 1-dimensional feature described previously corresponds to the orange point at $(5.6, 6.15)$ in the persistence diagram.

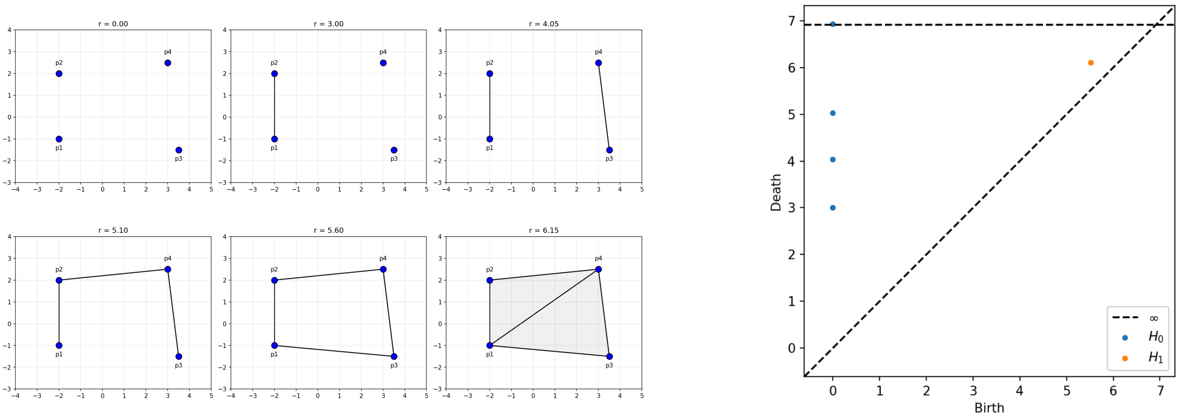

Figure 1: Visualization of Vietoris-Rips filtration on a set of points in $\mathbb{R}^2$.

Now given two sets of points, we can compute separately their persistence diagrams using persistent homology, but we need a metric to compare the two persistence diagrams. *Wasserstein distance* is the most common metric for this task. Given persistence diagrams D_1 and D_2, Wasserstein distance finds the best bijection between points in D_1 and D_2 and computes the sum of distances between matched points. Formally, let p be a fixed dimension. The p-Wasserstein distance between D_1 and D_2 is defined as:

$$W_p(D_1, D_2) = \inf_{\phi : D_1 \to D_2} \left(\sum_{x \in D_1} \| x - \phi(x) \|^p \right)^{1/p}.$$

The Wasserstein distance between D_1 and D_2 is thus:

$$W(D_1, D_2) = \sum_{n=0}^{\infty} W_n(D_1, D_2).$$

3. Related Work

4. Methodology

5. Results & Discussion

6. Conclusion

Bibliography

[1] W. H. Guss and R. Salakhutdinov, “On Characterizing the Capacity of Neural Networks using Algebraic Topology,” *CoRR*, 2018.

[2] A. Vaswani *et al.*, “Attention is All you Need,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017.

[3] M. Bianchini and F. Scarselli, “On the Complexity of Neural Network Classifiers: A Comparison Between Shallow and Deep Architectures,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 8, pp. 1553–1565, 2014.

[4] S. Fitz, “The Shape of Words - topological structure in natural language data ,” in *Proceedings of Topological, Algebraic, and Geometric Learning Workshops 2022*, in Proceedings of Machine Learning Research, vol. 196. PMLR, 2022, pp. 116–123.

[5] O. Draganov and S. Skiena, “The Shape of Word Embeddings: Quantifying Non-Isometry with Topological Data Analysis,” in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 12080–12099.

[6] S. H. Meirom and O. Bobrowski, “Unsupervised Geometric and Topological Approaches for Cross-Lingual Sentence Representation and Comparison,” in *Proceedings of the 7th Workshop on Representation Learning for NLP*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 173–183.

[7] V. Ravishankar, A. Kulmizev, M. Abdou, A. Søgaard, and J. Nivre, “Attention Can Reflect Syntactic Structure (If You Let It),” in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online: Association for Computational Linguistics, Apr. 2021, pp. 3031–3045.

[8] L. Kushnareva *et al.*, “Artificial Text Detection via Examining the Topology of Attention Maps,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 635–649.

[9] A. Uchendu and T. Le, “Unveiling Topological Structures from Language: A Comprehensive Survey of Topological Data Analysis Applications in NLP.” 2025.

[10] N. Team *et al.*, "No Language Left Behind: Scaling Human-Centered Machine Translation ." 2022.