Exploring Directed vs. Undirected Graph-Based Topological Data Analysis of Transformer Attention Maps
(Selected Topics in DS 2024 Course)

Kamil Garifullin [1] **Ilya Trofimov** [1]

Abstract

This project investigates the application of directed graph analysis to transformer attention maps, comparing it with traditional undirected graph methods. Using persistent homology, we examine the topological features of attention maps to evaluate their impact on downstream classification tasks. The results of this research show the efficacy of utilizing directed graph representations, providing valuable insights into the dynamic flow of information within transformer models.

Github repo: github.com/TDA-for-Transformers

1. Introduction

In the realm of natural language processing (NLP), transformer models have revolutionized the field by effectively capturing contextual relationships between tokens in textual data. Central to these models are attention mechanisms, which generate attention maps highlighting the interactions between tokens across different layers and heads.

Being rich in information, attention maps often require sophisticated analysis techniques to fully exploit their potential. One such promising direction is the application of persistent homology to extract meaningful signals from these attention maps (Tulchinskii E., 2022), (Cherniavskii D., 2022). By characterizing the topological structure of the attention graphs, persistent homology allows us to understand the underlying relationships between lexemes and their importance in the context of various NLP tasks.

However, existing research works has focused on analyzing attention maps as undirected graphs, ignoring the inherent directional nature of the relationships encoded in them. This oversight prevents us from obtaining valuable information about the flow and dynamics of information propagation in transformer models.

Fortunately, tools and methodologies exist for computing homologies of directed graphs. This gives us the opportunity to explore the advantages of using attention map representations based on directed graphs.

Therefore, the main goal of this project is to systematically compare the performance and efficiency of persistent homology-based feature extraction from attention maps represented as directed graphs with their non-directed counterparts.

The project plan consists of the following steps:

1. **Dataset and Model Selection**: took the dataset and a pre-trained transformer model (BERT) suitable for the downstream classification task from the article (Cherniavskii D., 2022).
2. **Pipeline Development**: developed a pipeline for extracting topological features from attention maps, implementing two variants: directed and undirected - to cater to different graph representations.
3. **Performance Comparison**: compared the classification performance on the features obtained by the two different methods, drawing conclusions regarding the effectiveness of using features to improve model performance.
4. **Feature Importance Analysis**: analyzed the importance of features obtained in two different ways (directed and undirected).

2. Related works

The study of attention mechanisms in natural language processing (NLP) and their role in encoding linguistic knowledge has attracted considerable attention in recent years. Various studies have focused on understanding how attention heads contribute to different linguistic tasks and phe-

[1] Skolkovo Institute of Science and Technology, Moscow, Russia.. Correspondence to: Kamil Garifullin <kamil.garifullin@skoltech.ru>.

Final Projects of the Machine Learning 2024 Course, Skoltech, Moscow, Russian Federation, 2024.

nomena. Below we review related works, on the study of the topology of attention maps.

- **Acceptability Judgements via Examining the Topology of Attention Maps.** The work presented by (Cherniavskii D., 2022) introduces one of the first attempt to analyze attention heads in the context of linguistic acceptability (LA) using topological data analysis (TDA), that enabled to examine graph representations of transformers' attention maps. This work demonstrated that features obtained by TDA can be used for further acceptability classification task with results that outperform the established baselines in three Indo-European languages (English, Italian, and Swedish) and confirmed the hypothesis that grammatical phenomena can be encoded through topological properties of the attention map.

- **Topological Data Analysis for Speech Processing.** In the second paper (Tulchinskii E., 2022) the authors explore the application of topological data analysis (TDA) to tackle speech classification problems and understand how a pre-trained speech model, HuBERT, works internally. To do this, authors introduced new features derived from Transformer attention maps. Surprisingly, with the usage of these new features a simple linear classifier performs better than a classifier specifically fine-tuned for the task. Moreover, authors found that on one dataset called CREMA-D, new TDA features set a new state-of-the-art performance with an accuracy of 80.155. Overall, this research shows that TDA could be a really useful approach for analyzing speech, especially in tasks where predicting structure is important.

3. Algorithms and models

4. Experiments and Results

5. Conclusion

References

Cherniavskii D., e. a. Acceptability judgements via examining the topology of attention maps, 2022.

Tulchinskii E., e. a. Topological data analysis for speech processing., 2022.