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.

- **Artificial Text Detection via Examining the Topology of Attention Maps.** It this article authors (Kushnareva L., 2022) introduced three new types of interpretable topological features for natural language processing (NLP) tasks, utilizing Topological Data Analysis (TDA). Authors' empirical findings demonstrate that features derived from the BERT model surpass traditional count- and neural-based approaches by up to 10% across three standard datasets. Moreover, through probing analysis, authors observed that TDA features are sensitive to both surface-level and syntactic properties of language. Overall, results of this research highlight the potential of TDA in enhancing NLP tasks, particularly those involving surface and structural information.

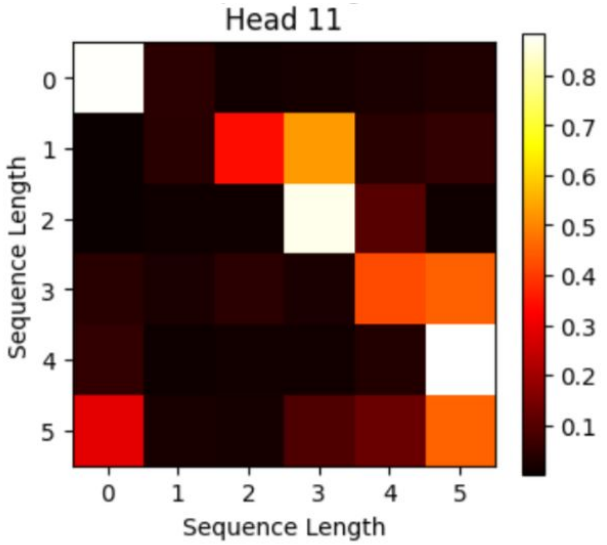

Figure 1. An attention map obtained from the 1st layer and 11th head

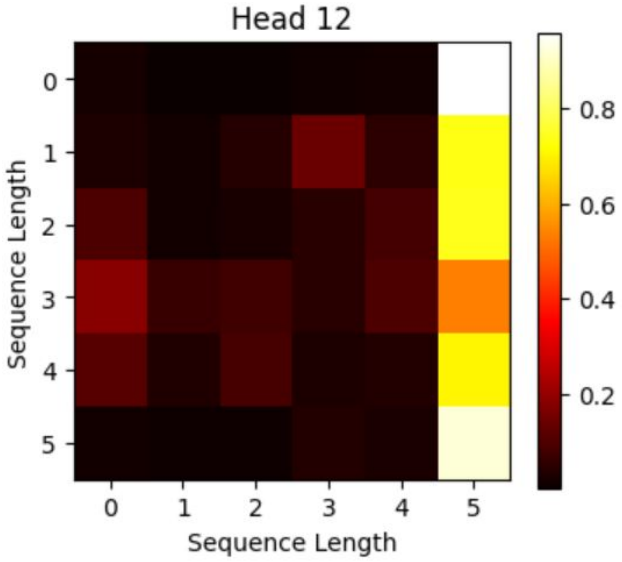

Figure 2. An attention map obtained from the 9th layer and 12th head

3. Methodology. Barcode features

3.1. Attention maps

Attention maps represent a fundamental concept in the realm of neural network architectures, particularly within models like Transformers, where they play a pivotal role in capturing dependencies and relationships between different elements of input sequences. In tasks like machine translation, sentiment analysis, and text generation, attention maps provide valuable context for understanding model predictions

	Method 1	Method 2
En-BERT + linear layer(baseline)	73.9	68.2
Baseline + undirected TDA	73.4	70.0
Baseline + directed TDA	73.0	-

Table 1. Results of classification for various models: En-BERT with linear layer(baseline), En-BERT with undirected TDA and En-BERT with directed TDA. The table shows the accuracies of the two feature acquisition methods. Method 1: TDA-features obtained only from the last layer and from the last head of attention, dataset size for this method: 25000 texts; method 2: TDA-features obtained from all layers and from all heads of attention, dataset size for this method: 12500 texts.

and decision-making processes. By visualizing attention patterns, researchers can identify linguistic structures, semantic relationships, and syntactic dependencies captured by the model.

Attention maps are generated dynamically as the model processes input sequences. These maps depict the distribution of attentional weights or probabilities assigned to each token in the input sequence relative to every other token. The process involves computing pairwise attention scores between tokens, often followed by a softmax normalization to obtain attention probabilities. These probabilities form the basis of the attention map, which is typically visualized as a heatmap, with intensity indicating the strength of attention.

For example, figure 1 and figure 2 show attention maps obtained for various layers and heads for the sentence: “It was raining yesterday”.

3.2. Barcode features

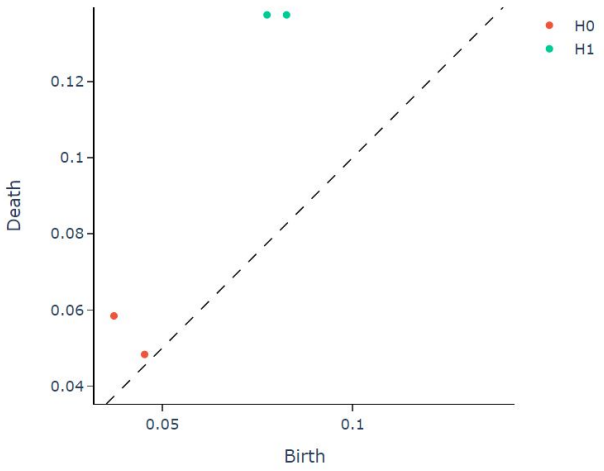

Figure 3. TDA features obtained for the sentence: “It was raining yesterday”

In this research we study barcode features (features that were extracted from barcodes) inferred from directed graphs and undirected graphs.

To calculate this features we used a high-performance topological machine learning toolbox in Python built on top of scikit-learn giotto-ai. Using this tool one can compute topological summaries, called persistence diagrams, from collections of point clouds or weighted graphs.

In the analysis of each text sample, we perform calculations to obtain the barcodes associated with the first two persistent homology groups, designated as H_0 and H_1, across every attention head of the BERT model. Our research covers several key characteristics of these barcodes, that form features:

- Total aggregate length of bars
- Mean length of bars
- Bars lengths variance
- Birth and death times of the longest bar
- Overall count of bars present
- Barcode’s entropy assessment

For example, the TDA features obtained for the sentence: “It was raining yesterday”, are shown in figure 3.

3.3. Representing Attention Maps as Directed and Undirected Graphs

Attention maps can be represented as both directed and undirected graphs to visualize dependencies among sequence elements. In a directed graph, each node represents a token, and each directed edge corresponds to the attention weight from one token to another, illustrating the flow and strength of attention from one token to another. This representation is particularly useful for understanding the directional influence and hierarchical relationships within the sequence.

On the other hand, an undirected graph simplifies the attention map by treating the attention weights symmetrically. Here, each edge represents a mutual relationship between tokens, irrespective of direction, which can be useful for highlighting the overall connectivity and bidirectional dependencies within the sequence.

3.4. Topological Feature Extraction from Undirected Graphs

Feature extraction from undirected graphs involves the following steps:

1. Increasing Parameter ϵ:

   As ϵ increases from 0 to infinity, we consider subgraphs that include:all vertices from the original graph and edges with weights less than or equal to the current ϵ.

2. Forming Simplicial Complexes:

   Next these subgraphs are extended to simplicial complexes, which include: vertices and edges or $k-$simplices, defined as $(k+1)-$cliques in the subgraph, for instance, a 2-simplex (triangle), 3-simplex (tetrahedron), etc. By definition, vertices are $0-$simplices and edges are $1-$simplices.

3. Tracking Topological Changes:

   As ϵ increases, we track:

   - The creation and merging of connected components due to the appearance of vertices (all vertices appear at $\epsilon = 0$) and new edges.
   - The formation of $d-$dimensional voids, such as 1-dimensional holes or 2-dimensional cavities, that are not bounded by $(d+1)-$simplices.
   - The filling of $d-$dimensional voids by newly appearing $(d+1)-$simplices.

This process, known as Vietoris-Rips persistent homology, records the topological evolution of the graph across all edge weights.

3.5. Topological Feature Extraction from Directed Graphs

The ideas and constructions underlying the algorithm in this case are very similar to the ones described above for the undirected case. Again, we threshold the graph and its directed edges according to an ever-increasing parameter and the edge weights. And again we look at "cliques" of vertices to define simplices and hence a "complex" for each value of the parameter. The main difference is that here simplices are ordered sets (tuples) of vertices, and that in each instantaneous complex the "clique" $(v_0, v_1, ..v_k)$ is a $k-$simplex if and only if, for each $i < j$, (v_i, v_j) is a currently present directed edge.

The figure 4 shows an example in which the left complex, in which the edges of the triangle cycle in one direction, contains a one-dimensional hole, while in the picture on the right there is no one-dimensional hole.

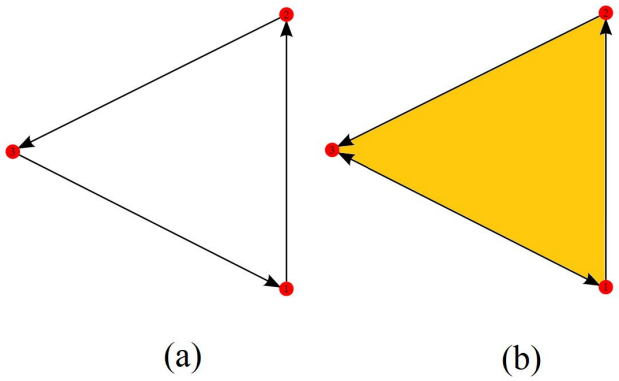

Figure 4. (a): (1, 2), (2, 3) and (3, 1) form a 1D hole. (b): (1, 2), (2, 3) and (1, 3) form the boundary of (1, 2, 3) – not a 1D hole

4. Dataset and feature extraction methods

4.1. Dataset

In our analysis, we selected the IMDb Movie Reviews dataset as the basis for our investigation. This dataset is widely known for its extensive collection of film reviews. IMDb contains movie reviews labeled as positive or negative sentiment. It's commonly used for sentiment analysis tasks. For our task, we intend to employ this dataset for binary classification purposes, specifically categorizing instances into two classes: positive and negative.

4.2. Feature extraction methods

In this work, the following feature extraction methods were carried out:

1. First, we extracted features only for the attention map of the 12th head of the last layer of BERT for each text from the dataset. We decided to conduct such an experiment because article (Kushnareva L., 2022) reported that the features obtained from this attention map are representative. Since this method of calculating features is relatively fast, TDA-features were calculated using this method for the entire dataset (25,000 texts).

2. Next we decided to calculate topological features for each attention map from each head and from each layer of BERT for each text from the dataset. Since the model we used (BERT-base-uncased) had 12 heads and 12 layers, for each text from the dataset we received 12x12x14 = 2016 features in total. Where 14 corresponds to the number of features we consider, which were listed above. Due to the fact that this feature extraction method turned out to be computationally expensive, we used half of the texts from the dataset to train and validate the model (so we used 12,500 texts).

5. Experiments and Results

For each feature extraction method, we trained 3 models: En-BERT with linear layer(baseline), En-BERT with undirected TDA and En-BERT with directed TDA. The results of all experiments are presented in Table 1.

5.1. Baseline

For the baseline in this work, a combination of BERT-base-uncased and one linear layer was used. We leveraged the BERT model to generate representations for our textual data. Specifically, we utilize the pooled output, often associated with the special [CLS] token, which serves as a condensed representation of the entire input sequence. This pooled output encapsulates the semantic information learned by the BERT model from the input text. And after obtaining the pooled outputs from the pre-trained BERT model, the next step in our methodology involved training a classifier using these representations as features. As a result, the accuracy of our baseline solution on the entire dataset was 73.9%, while on half of the dataset the accuracy of baseline solution was 68.2%.

5.2. TDA features from Undirected attention maps

Next, we trained linear neural network on vectors obtained by concatenating pooled outputs from BERT and TDA-features obtained by considering attention maps as undirected graphs. In our research we obtained TDA features in 2 different ways as was described above: firstly we obtained TDA-features only from the last layer and from the last head of attention, and secondly we obtained TDA-features from all layers and from all heads of attention. Moreover, due to the high computational cost of calculating the TDA-features using the second method, we trained and validated models on only half of the texts from the dataset, when using the second method of calculating features.

As a result, by calculating TDA-features only on the last attention layer and head we were able to achieve an accuracy of 73.4% on the full dataset, while by calculating features on the all attention layers and all attention heads we were able to achieve an accuracy of 70.0% on the half of the dataset.

5.3. TDA features from Directed attention maps

Next, we trained an ensemble of models consisting of baseline and RFC (random forest classifier) trained on TDA-features obtained by considering attention maps as directed graphs. Similar to the methods described above for the undirected attention maps, we were able to achieve an accuracy of 73.0% on the full dataset by calculating features only on the last attention layer and head, while by calculating features on the all attention layers and all attention heads we could not calculate TDA features, because even on half of the dataset it would take a huge amount of time and resources that we did not have. This is due to the fact that FlagserPersistence, which we used to calculate bar codes, is poorly parallelized and requires a lot of calculation time.

6. Discussion

The results of our experiments show that the features obtained for the last attention layer and the last head (Method 1) turned out to be unrepresentative for both the case of directed graphs and the case of undirected graphs. It can be seen that the accuracy of the baseline solution turned out to be 0.5-1% higher compared to models trained on vectors that were formed by concatenating pooled outputs from BERT and TDA-features.

At the same time, the features obtained from all layers and heads of attention, if we consider attention maps as undirected graphs, turned out to be representative and the model, which is an ensemble of RFC (random forest classifier) trained on these TDA-features and the baseline model, turned out to be 2% higher than the baseline.

7. Conclusion

To summarize, we can note that topological analysis can be highly beneficial for NLP tasks. For our task, TDA features obtained by considering attention maps as undirected graphs increased the prediction accuracy by 2%. However, using attention maps as directed graphs and employing the giotto-tda toolbox did not lead to any success due to the computational complexity and inefficiency of the giotto-tda library.

It is important to note that a 2% improvement is relatively modest. For instance, in article (Cherniavskii D., 2022), topological analysis enhanced the accuracy of predicting grammatically correct sentences by 10%. This discrepancy can be explained by the nature of the tasks: the attention mechanism effectively captures grammatical relationships within tokens, making TDA valuable for predicting grammatical correctness. In contrast, our task involves sentiment analysis, where the attention mechanism might find it more challenging to understand the semantic relationships between words.

References

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

Kushnareva L., e. a. Artificial text detection via examining the topology of attention maps, 2022.

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