Review of the report: Exploring Directed vs. Undirected Graph-Based Topological Data Analysis of Transformer Attention Maps

May 31, 2024

1 Report Review

1.1 Clarity

The paper delves into the sophisticated application of topological data analysis (TDA) to transformer attention maps, presenting a clear and well-defined problem statement. The review of related works and the methodology are thorough, making the context accessible even to those less familiar with the subject. The methods are detailed, and the results are presented with clarity. Nonetheless, several sections of the paper could benefit from enhanced clarity to improve the overall comprehensibility:

- The abstract offers a high-level overview but lacks specific details about the key findings.
- The introduction effectively establishes the context but would benefit from more information about NLP tasks in general, their significance, and a brief discussion of existing approaches. Additionally, the project plan section appears less useful and could be replaced with bullet points highlighting the paper's contributions, such as a summary of the proposed method and its advantages.
- The related works section comprehensively captures existing solutions and applications for topological data analysis. However, a broader discussion on topological analysis in general would enhance the reader's understanding of the context. For instance, while the authors introduce the directed TDA approach, it would be beneficial to discuss any similar works that have successfully utilized directed TDA.
- The methodology is detailed and accessible even to readers without prior knowledge of the topic. However, the flow could be improved. It would be more logical to begin with a discussion of Topological Feature Extraction as a general concept, and then move on to its application to attention maps. Additionally, the introduction of H_0 and H_1 is lacking; the authors mention these terms but do not provide definitions.
- The experiments are well-defined, but it is puzzling why the authors use a linear model for the undirected attention map features while employing a random forest classifier for the directed ones. This discrepancy makes the comparison of the methods unfair. Additionally, the discussion of the results is not fully comprehensive.

For instance, there is no explanation for why the directed attention map results are lower than those for the undirected maps, nor is there any discussion on how the method could be improved in the future.

1.2 Format, styling and grammar

The paper is well-written and easy to follow, though there are some minor issues that need addressing:

- **Figure 2 caption**: The caption states that the maps are for the 9th and 12th heads, but the images show the 11th and 12th heads.
- **Proofreading**: The paper requires proofreading to correct minor misprints. For example, in "En-BERT with linear layer(baseline)," a space is missing. Despite these minor issues, the overall text is very good.

2 Repo quality

The repository quality is good and easy to understand. However, there are a few areas that need improvement:

- **Results**: The repository lacks the results of the experiments conducted. Including these would provide a clearer understanding of the findings.
- **Usability**: It is not very clear how one can apply this code to their own tasks. Adding detailed instructions or examples would enhance usability.
- **Consistency** The code uses a linear model instead of the random forest mentioned in the paper, leading to inconsistencies between the paper and the code.

3 Reproducibility

The code is reproducible; however, the provided Google Colab links are missing some crucial code parts, such as defining the model and optimizer. I was able to reproduce the results because the full notebooks are available in the repository, but the notebooks in Colab should be fixed to include all necessary code parts.