First project peer review

Multimodal chain of thought

Sergey Kastryulin

The project is dedicated to solve the following problems in large language models (LLMs): improve efficiency of the chain of thought mechanism and generalize the framework to new modality (3D data). The first problem is proposed to be solved by introduction of an additional autoencoder model to compress input representations and second problem is to be addressed by fine-tuning on a newly annotated dataset of 3D data.

On project report structure and completeness of descriptions.
The project report contains the required sections i.e., problem statement, description of the main idea. However, I could not find any notes about the protocol of comparison with relevant methods. In the Datasets section it is said that the approach will be evaluated on the ScienceQA benchmark. However, it is still unclear whether the authors aim to assess the end quality of the resulting model or the performance gains due to introduction of the input compression model. My recommendation here would be to make the description of the test protocol more apparent and explicit.

On styling, quality and structure.
The proposed method is well motivated, the text is easy to follow and understand. I would positively assess the styling and quality parts, while the structure part is more nuanced. In general, there is a best-practice to dedicate a scientific artefact (white paper, technical report) to a single question or problem. Otherwise, the message might get diluted and the work loses its focus. Structure-wise it is typically seen as the paper having two or more weakly connected contributions. This is what I can see in this work. Both proposed problems are valid and valuable to explore. However, I doubt that it is optimal to have them interconnected in a single project. It means that it is going to be practically difficult to explain to a reader why experiments on compression of chain of thought representations are constantly altered with multimodal fine-tuning experiments. I would propose authors to either give a strong motivation why these two contributions belong to the same paper or to split the work into two relatively independent assets.

On code and completed experiments.
To the moment of writing the review (the 18th or May, 19:50), the repository has no training code. I could find dataloaders and tools for data visualization, which are also described in the readme. My genuine hope is that the authors will have time to upload their code and results in the future because the project seems reasonable and interests me.