Exploring the Logical and Mathematical Capabilities of the BERT Embedding Space using Contrastive Learning

Stanford CS224N Custom | Default Project

Mona Anvari
Department of Computer Science
Stanford University
monaavr@stanford.edu

Abstract

Natural language understanding is a critical challenge in the field of artificial intelligence. BERT is a widely used model for natural language processing tasks, but its ability to perform logical and mathematical reasoning is not well understood. In this project, we investigate whether BERT's embedding space can learn logical and mathematical relationships and perform well on benchmarks designed for such tasks, using contrastive learning methods to measure similarity and improve embeddings. Our main goal is to look at the potential and limitations of BERT-like models for these types of tasks. We also aim to compare the performance of BERT models trained with contrastive learning to those trained with other techniques, and potentially extend our investigation to the domain of mathematical reasoning. Our main findings will dependant on the performance of BERT models trained with contrastive learning on logical and mathematical reasoning tasks. We hope to demonstrate the potential of contrastive learning to improve the performance of BERT on these tasks and provide insights into the capabilities of BERT-like models for logical and mathematical aspects of natural language understanding.

1 Key Information to include

N/A

2 Approach

After performing the base task of the default project, I am planning on using benchmark datasets such as the Logical Inference (LI) dataset and the MathQA dataset to evaluate. To prepare the datasets, I will tokenize the sentences and questions and convert them into BERT-compatible format. For the LI dataset, I will convert the labels into a similarity score between the two sentences, with a higher score indicating a stronger relationship. The datasets will be split into train, validation, and test sets, with proportions of 70%, 15%, and 15%.

I will use BERT-based models I have completed as my baseline for logical and mathematical reasoning tasks. I might also use pre-trained BERT models available in the Hugging Face Transformers library depending on time. I will compare the performance of my models with previously published scores for the LI and MathQA datasets.

I will use contrastive loss and contrastive learning techniques to investigate BERT models' logical and mathematical reasoning abilities. We will use cosine similarity as the contrastive loss function, and we will project sentence embeddings onto a hypersphere to help improve the quality of the embeddings.

If time allows, I might change the loss function or investigate methods to better capture the kind of logical relationship between sentence embeddings.

I will implement a basic version of miniBERT as a starting point, following the default project descriptions. I will fine-tune the miniBERT model on the LI and MathQA datasets using contrastive learning. If resources and time allow, I may explore the performance of other models designed for logical and mathematical reasoning, such as NALUs. I will compare their performance with that of our BERT-based models.

For quantitative evaluation, I will use specific evaluation metrics for each task. For the sentence-pair similarity task on the LI dataset, I will use the commonly used F1 score. For MathQA , I will use accuracy as evaluation metric. I will compare our models' performance with previously published scores for the LI and MathQA datasets.

While my primary focus is quantitative evaluation, I may also conduct qualitative evaluation, potentially involving human evaluation on a small subset of the results, depending on howmuch time I have left.

3 Experiments

I have received permission from course staff, Elaine Sui, to skip this part of the milestone. I have implemented a version of miniBERT that passes the sanitycheck. However, even though I reached out to the teaching team about not having a GPU quota on AWS two weeks ago, I didn't have a GPU available to me on my AWS account until 6:59pm tonight. I am emailing my successful miniBERT code to course staff at the same time as I am uploading this report. However I was unable to run experiments. Since I finally have access to a GPU please review my plan for the following week in the future work section.

4 Future work

- use the miniBERT implementation I have and fine-tune it on the Logical Inference (LI)[1] and MathQA[2] datasets using contrastive learning.
- Preprocess the datasets by tokenizing the sentences and converting them into BERT-compatible format.
- Split the datasets into train, validation, and test sets, with proportions of 70%, 15%, and 15%, respectively.
- Use cosine similarity as the contrastive loss function and project sentence embeddings onto a hypersphere[3] to help improve the quality of the embeddings.
- Evaluate the performance of the BERT-based models using specific evaluation metrics for each task, such as F1 score for the sentence-pair similarity task on the LI dataset and accuracy for the MathQA dataset.
- Compare the performance of the BERT-based models with previously published scores for the LI and MathQA datasets and state-of-the-art performance on the MathQA dataset. Potentially conduct qualitative evaluation, involving human evaluation on a small subset of the results, if time permits.
- If time allows, investigate methods to better capture the kind of logical relationship between sentence embeddings and explore other models designed for logical and mathematical reasoning, such as NALUs.[4]
- Write up the findings of the investigation in a research paper, including the methods used, results obtained, and analysis of the strengths and limitations of the approach.

References

[1] Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas, November 2016. Association for Computational Linguistics.

[2] Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[3] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *CoRR*, abs/2104.08821, 2021.

[4] Andreas Madsen and Alexander Rosenberg Johansen. Neural arithmetic units. In *International Conference on Learning Representations*, 2020.