COSE474-2024F: Final Project Proposal "Analyzing and Improving Geometric Math Problem-Solving Ability of LLaMa3.2-3B"

Won Park

1. Introduction

There has been remarkable improvement of large language models (LLMs) on recent years. Accordingly, performance of small language models (SLMs) such as Meta's LLaMa3.2-3B and Google's Gemma-2B has also improved. However, SLMs still have weaknesses in solving mathematical word problems, especially graphic and geometric problems. The mathematical abilities of SLMs are significantly lower than those of LLMs, as reflected in GSM8K (Grade School Math 8K) benchmark scores. While large models such as Abel(83.85) and GPT-4 (93.25) have achieved high scores on the GSM8K test, smaller models like Abel-7B (59.51) and MetaMath-Mistral (77.79) still lack sufficient mathematical reasoning abilities. (Li et al., 2024)

One of the main issues of the reasoning ability was geometrical position recognition. These models cannot answer properly when they have to use positional information. Therefore, in this study, we are going to find reason for geometric misconception on the LLaMa3.2-3B's mathematical problem solving, and we focus on improving its answer generation.

2. Problem definition & challenges

Main problem of our research are:

1) Analyzing why the language model generates wrong answer in geometric problems.

2) Modifying the model to improve its ability solving graphical problems. Appending layers and prompt engineering would be our methods to enhance the model.

3. Related Works

(Yu et al., 2024) introduced LLaMa-based mathematical reasoning model, MetaMath. They enhanced performance by bootstrapping questions and rewriting the question from multiple perspectives.

(Tong et al., 2024) is not focused on math word problem but math vision language model. However, this study suggests that performance improves when training focuses on problem that are frequently answered incorrectly.

4. Datasets

We are going to use **OpenMathInstruct-2** dataset for training and validation. NVIDIA developed the dataset and already enhanced LLM's performance on math word problems through full fine-tuning. (Toshniwal et al., 2024) The dataset will be filtered to include only geometrical problems.

5. State-of-the-art methods and baselines

When focusing on models with fewer than 7 billion parameters, the MetaMath-Mistral model achieves a score of 77.79 on the GSM8k benchmark, which is performance of State-of-the art. (Li et al., 2024)

However, we will use LLaMa3.2-3B to analyze and improve the model. First reason we use LLaMa as a baselines instead of MetaMath-Mistral is that most of the small mathematics models are based on LLaMa. Moreover, LLaMA has the advantage of allowing prompt engineering, unlike MetaMath. Finally, a method that improves the performance of LLaMA may also lead to significant improvements in MetaMath-Mistral.

References

Li, Q., Cui, L., Zhao, X., Kong, L., and Bi, W. Gsm-plus: A comprehensive benchmark for evaluating the robustness of llms as mathematical problem solvers, 2024. URL https://arxiv.org/abs/2402.19255.

Tong, Y., Zhang, X., Wang, R., Wu, R., and He, J. Dart-math: Difficulty-aware rejection tuning for mathematical problem-solving, 2024. URL https://arxiv.org/abs/2407.13690.

Toshniwal, S., Du, W., Moshkov, I., Kisacanin, B., Ayrapetyan, A., and Gitman, I. Openmathinstruct-2: Accelerating ai for math with massive open-source instruction data, 2024. URL https://arxiv.org/abs/2410.01560.

Yu, L., Jiang, W., Shi, H., Yu, J., Liu, Z., Zhang, Y., Kwok, J. T., Li, Z., Weller, A., and Liu, W. Metamath: Bootstrap your own mathematical questions for large language models, 2024. URL https://arxiv.org/abs/2309.12284.