Visual Mathematical Reasoning using Vision-Language Models

Mounika Mukkamalla, Neeharika Gupta, Taejas Gupta, Vishal Singh, Vishesh Agrawal

Abstract

Mathematical reasoning has been a recent area of focus for Large Language Models (LLMs), leveraging their comprehension abilities for problem solving. When the problems are presented visually in the form of visual puzzles or problems, the modality is no longer supported by LLMs alone. Over the course of the semester, we propose to perform various analytical experiments on different types of problems present in the MathVista dataset and evaluate the performance of Vision-Language Models (VLMs). Additionally, our analysis will present the performance of VLMs on different types of visual problems and provide possible justifications behind the performances. For a stretch goal, we also propose using external mathematical tools in conjunction with a VLM in an attempt to get improved results over the base VLM.

1 Introduction

Mathematical reasoning has been a common presence in testing human intelligence, be it IQ tests or mental ability evaluations. Since a young age, humans are taught mathematical concepts and reasoning to develop intellect and thinking. Drawing inspiration from humans, to actually develop 'intelligent' machine learning models, it is important to ensure that these models are capable of some level of mathematical reasoning. Current research in mathematical reasoning for LLMs primarily handles problems presented in a textual format as input. However, there exist mathematical problems that require visual inputs along with the textual query. Such problems include mathematical puzzles, geometric questions, and visual question-and-answer problems. In these cases, LLMs alone do not suffice.

VLMs possess the ability to process both visual and textual input data. The visual input is usually in the form of images, and the textual input is in the form of queries. The architecture comprises of an image encoder, a text encoder, and a fusion strategy that is able to combine the two encodings together. Modern VLMs primarily use transformer-based encoders for both text and image inputs. As a result of their design, VLMs possess the ability to take visual mathematical problems as input along with the query prompts, and perform reasoning on those.

1.1 Literature Review

Nowadays, LLMs have become the go-to solution for any given problem due to their sophisticated model architectures and the extensive data that they have been trained on. There has been a lot of work on text and vision modalities. Some of the recent approaches range from unimodal models like LLMs for text (GPT-3 [Brown et al., 2020], T5 [Raffel et al., 2020]) and vision models (Vision Transformer [Dosovitskiy et al., 2020], ResNet [He et al., 2016]) to LMMs catering to multiple modalities (GPT-4 [Achiam et al., 2023], LLaMA [Touvron et al., 2023]).

Mathematical problem solving is an intricate task of language modeling and understanding. Various types of mathematical problems (Ahn et al., 2024) have been dealt using different strategies. Models are required to be trained specifically to tailor the problems in the domain of mathematical problem solving. MAmmoTH (Yue et al., 2023) – a series of open-source LLMs – is trained via hybrid instruction tuning, LLemma (Azerbayev et al., 2023) is capable of formal theorem proving without any further finetuning, Minerva (Lewkowycz et al., 2022) generates solutions without relying on external tools, and WizrdMath (Luo et al., 2023) empowers reasoning via reinforced evol-instruct.

Reasoning is one of the most fundamental aspects of mathematical solving. Recent models are focusing not only to answer correctly but also to provide the right reasoning and rationale behind the answer generated. (Lu et al., 2023) evaluates mathematical reasoning in visual contexts, (Kazemi et al., 2023) provides reasoning behind geometric problems, (Wang et al., 2023) seamlessly integrates code to enhance the models reasoning capabilities and (Imani et al., 2023) boosts model confidence by providing more samples via zero shot chain of thought prompting.

2 Dataset

For our project, we plan to work with the MathVista dataset (Lu et al., 2023) which consists of a wide variety of mathematical problems. It consists of five main categories of problems – math word problems (MWP), figure question answering (FQA), geometry problem solving (GPS), textbook question answering (TQA), and visual question answering (VQA). It combines 31 datasets on these five types of categories. Over the course of our project, we will assess the models/techniques on the basis of these five main categories. This would likely give us as an insight into where the models work better, and where do they lack.

The dataset designates a subset of 1000 samples as the testmini set, which also have their associated correct answer. For the scope of this project, we plan to evaluate our approaches on the testmini set only. For each instance, the dataset provides an image, a question, an associated query to prompt the LLM, and other associated tags. For further analysis of mathematical reasoning, we plan on using a selected pool of datapoints, manually write out their reasoning, and use that as a gold standard for analysing the reasoning abilities of different models – not just comparing the final answer.

3 Project Plan

In this section, we go over our plan for the project over the remainder of the semester. First, we plan on picking a small subset ($\sim$100) of mathematical visual problems picked from the testmini set, and manually annotating their solutions' reasoning. We plan on having all five members of our group perform this manual annotation task so that we are able to neutralize the various styles of reasoning used by different individuals. Our goal with this sample is to not just analyse the answers from the models as a metric, but to also analyse the mathematical reasoning provided by the models. Based on these findings, we can potentially suggest future scopes of improvements. To compare the reasoning, we plan on using a metric like BLEURT with all our five manually annotated reasonings, and then use a threshold to determine whether they are correct or not. This task would require about two weeks.

Next, we plan on developing a basic VLM pipeline to determine its findings on the entire testmini set. To evaluate these models on the problems, we can directly use accuracy as a metric. Additionally, we aim to use models other than VLMs if possible to evaluate and analyze their performance on different types of problems. We plan on using open-source VLMs, and if possible, GPT-4, which also supports multimodal input. This task would require at least four weeks to code and run experiments.

Finally, our stretch goal would be to use external tools to further improve the performance of the VLMs based on the findings of the reasoning dataset. This is purely experimental and we may not have the required resources to achieve this. This work would possibly require two or more weeks for implementation and experimentation.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2023. Llemma: An open language model for mathematics. *arXiv preprint arXiv:2310.10631*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. MathPrompter: Mathematical reasoning using large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 37–42, Toronto, Canada. Association for Computational Linguistics.

Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. 2023. Geomverse: A systematic evaluation of large models for geometric reasoning. *arXiv preprint arXiv:2312.12241*.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ke Wang, Houxing Ren, Aojun Zhou, Zimu Lu, Sichun Luo, Weikang Shi, Renrui Zhang, Linqi Song, Mingjie Zhan, and Hongsheng Li. 2023. Mathcoder: Seamless code integration in llms for enhanced mathematical reasoning. *arXiv preprint arXiv:2310.03731*.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*.