Can Intermediate Reasoning Chains Rationalize Better in MultiModals?
Team Pilot (Mentor: Shirley Anugrah Hayati)

Mani Deep Cherukuri University of Minnesota cheru050@umn.edu

Shunichi Sawamura University of Minnesota sawam002@umn.edu

Shashank Sharma University of Minnesota sharm964@umn.edu

Nicole Vu University of Minnesota vu000166@umn.edu

1 Member's Roles

- **Mani Deep Cherukuri**: Set up MSI, Deploying the Multi-model CoT, Finetune the model, extracting vision features, Extract Captions, Inference/Result Analysis, Project Proposal, Final Report

- **Shashank Sharma**: Deploy the LLaVa on MSI, Project Proposal, Cleaning and running tests, Result Analysis, Final Report

- **Shunichi Sawamura**: Literature Research, Project Proposal, Deploying the GPT-3, Programming/Code Modifying, Result Analysis, Final Report

- **Nicole Vu**: Programming/Code Modifying, Project Proposal, Deploy the VisualBERT, Cleaning and running tests, Result Analysis, Final Report

2 Motivation

The advancement of large multi-model pre-trained language models has opened up new possibilities for natural language understanding in conjunction with visual content. These models exhibit remarkable capabilities in tasks like image captioning, question answering, and text generation. However, while they demonstrate the potential for reasoning over multimodal information, challenges remain in harnessing their abilities to generate informative explanations, particularly in visual question-answering (VQA) tasks.

In many VQA tasks, generating informative explanations or rationales is crucial for enhancing user understanding and trust in model responses. Yet, current multimodal models often struggle to produce informative rationales, leading to a gap between model performance and human-level reasoning. This challenge is even more pronounced in tasks like Science Question Answering (ScienceQA), where questions often require multi-hop reasoning and explanations that involve both textual and visual modalities.

This project seeks to address this gap by investigating whether an intermediate reasoning chain that incorporates language (text) and vision (image) modalities can lead to better information rationales and improved understanding in multimodal tasks. We aim to analyze various multimodal language models' performance in generating informative explanations for ScienceQA questions. By comparing these models to a baseline model, we aim to understand their potential to enhance information extraction and reasoning across diverse contexts.

Our project's findings have broader implications, extending to various fields such as education, healthcare, business, and media. Enhanced multimodal reasoning and explanation capabilities can revolutionize content generation and improve user engagement and understanding. However, we also acknowledge that ethical considerations, bias mitigation, and responsible deployment of multimodal models are essential to ensure the technology's responsible and equitable use.

3 Literature Survey

3.1 ScienceQA

Science Question Answering (ScienceQA) is a benchmark with approximately 21,000 multimodal multiple-choice questions covering diverse science topics like natural science, language science, and social science. The dataset includes annotations for answers along with corresponding lectures and explanations. This dataset is suitable for training or testing the performance of language models that are designed to mimic the human multi-hop reasoning process and generate correct answers with reasoning. ScienceQA significantly differs from Visual Question Answering (VQA) (Agrawal et al., 2015) datasets in terms of difficulty, multi-modal contexts, and diversity of topics. ScienceQA mainly consists of questions with 1) image context, 2) text context, and 3) both. Image context means the questions are supposed to be answered based on a given image. Text context indicates that additional information is given with the questions that could lead to the direction of the solutions or hints.(Lu et al., 2022)

3.2 GPT-3

GPT-3 (Brown et al., 2020) is the state-of-the-art model that can be applied to various tasks and can be the baseline model in our analysis. We plan to implement few-shot learning so that the GPT-3 model can answer questions without parameter updates. GPT-3 will be instructed to answer the ScienceQA by providing in-context examples with components of the question text, options, and the correct answer text. It encourages the GPT-3 to lead the answer via chain-of-thought prompting. In the leaderboard, the GPT-3 performs 73.97% accuracy with 2-shot learning. (Lu et al., 2022)

3.3 VisualBERT

VisualBERT (Li et al., 2020) uses the self-attention mechanism within the Transformer to implicitly align text elements with image regions. The model maintains a consistent configuration with BERT-Base, featuring 12 layers, a hidden size of 768, and 12 self-attention heads. In addition, VisualBERT introduces visual embeddings to represent image features. These visual embeddings are processed through a multi-layer Transformer alongside text embeddings, enabling the model to discover implicit alignments between text and images and construct joint representations.

The analysis of VisualBERT applied to the ScienceQA dataset yields an average accuracy rate of 61.87%, showcasing notable performance variations with a peak accuracy of 69.18% observed for social science questions and a lower accuracy of 58.54% noted for questions lacking contextual information.

3.4 LLaVa

LLaVA, or "Visual Instruction-Tuning,"(Liu et al., 2023) is a multimodal model designed for instruction-following and reasoning. It combines LLaMA(Touvron et al., 2023) with the CLIP visual encoder (ViT-L/14) (Radford et al., 2021). LLaVA connects image features to language embeddings using a simple linear layer.

For training, LLaVA utilizes multi-turn conversation data, organizing it as a sequence of turns with the assistant's responses and instructions aligned with the image. Instruction-tuning on the LLM aims to optimize the probability of generating target answers through an auto-regressive training objective.

In the ScienceQA benchmark, LLaVA achieves close to state-of-the-art accuracy (90.92%) when utilizing visual features before the last layer of CLIP. This model outperforms GPT-3.5 and achieves a new state-of-the-art accuracy of 92.53% when combining the outcomes of LLaVA and GPT-4.

3.5 Multimodal CoT

Multimodal Chain-of-Thought Reasoning (Zhang et al., 2023) incorporates text and vision into a two-stage framework. Creating a justification based on multi-modal data is the first stage. The first stage involves the rationale generation based on the given multi-modal information. This means that given text and images, the model is supposed to come up with a justification for how the two are related. The second stage of the framework is the answer inference phase. This is where the model uses the information rationale that it generated in the first stage to infer the correct answer to the question. In general, there are two ways to elicit the multimodal-CoT reasoning: (i) transform the inputs of different modalities into one modality and prompt LLMs to perform CoT and (ii) fine tuning smaller models by fusing multimodal features.

4 Problem Definition

To evaluate the effectiveness of an intermediate reasoning chain that integrates language and vision modalities in generating information rationales for visual question answering, particularly focusing on the ScienceQA dataset, with the goal of enhancing information extraction and reasoning in multimodal models.

5 Proposed Idea and Hypothesis

5.1 Proposed Idea

The primary objective of this project is to examine the effectiveness of an intermediate reasoning chain that combines language (text) and vision (image) modalities to produce more informative information rationales.

To accomplish this objective, our proposed approach involves generating answers from the ScienceQA dataset using three distinct multimodal pre-trained language models, namely VisualBERT, Llava, and Multimodal CoT. Following the generation phase, we will evaluate and compare the accuracy scores of these models with those of the GPT3 baseline model. Our analysis will focus on the right combination of textual and visual data in these models, with the aim of assessing whether this integrated approach has the potential to enhance the quality and depth of information explanations in visual question answering. Simultaneously, we will investigate which language model exhibits better performance in both general and specific contexts.

Science QA:
The train dataset has: 12.7k rows
The validation dataset has: 4.24k rows
The test dataset has: 4.24k rows
Features of the dataset:
image, question, hint, task, lecture, solution

LLMs used : GPT-3, LlaVA, VisualBert, Multimodal-COT

Error Analysis : We randomly pick up 50 samples that are correct and 50 samples that are incorrect. We try to figure out whether the correct samples contain a low amount of incorrect chain-of-thought to identify if the modal can predict the correct answer by ignoring the incorrect rationale. Similarly, for the incorrect samples whose answers are incorrect, we analyze the impact of common sense mistakes where answering the questions requires common sense knowledge, e.g., understanding maps, counting numbers in the images, and utilizing the alphabet.

5.2 Hypothesis

The key challenge observed in most of the language models under 100 billion parameters is that they tend to generate hallucinated rationales that mislead the answer inference. To mitigate this and facilitate the interaction between the modules, our project hypothesizes that fine-tuning a smaller language model (LM) by fusing multi-modal features with intermediate reasoning can enhance the inference of the given information.

6 Broader Impact

The result of the project would allow us to analyze the impact of an intermediate reasoning chain that incorporates language (text) and vision (image) modalities to generate information rationales to explore the potential to improve information extraction and increase other cross-modal integration in various fields. For example, a combination of visuals and text can help students ask questions about what they see in a picture and receive answers. It can also be a valuable tool in business, marketing, entertainment, and media, offering more engaging and informative content to users.

7 Feedback

In our pitch presentation, we received valuable feedback from the audience, which has a substantial impact on our current decision for the approach of this project. The key points of feedback are detailed below.

One of the pieces of feedback we received during the pitch highlighted the potential time and power constraints associated with training a multimodal language model, which could potentially extend beyond the project timeline. To address this issue, we have changed the direction of the project to use the pre-trained or fine-tuned models for running the dataset.

Another recommendation was to consider replacing the language models initially selected for the presentation with alternatives such as VisualBERT, Llava, and BLIP-2. After researching those models, we decided to use VisualBERT and Llava, but not BLIP-2 because we didn't think that it would be suitable for our project and dataset.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2015. Vqa: Visual question answering. *International Journal of Computer Vision*, 123:4 – 31.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020.

Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020. What does BERT with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275, Online. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models.