Multimodal Chain-of-Thought Reasoning in Language Models

Zhuosheng Zhang, Aston Zhang, Mu Li, hai zhao, George Karypis, Alex Smola

Published: 03 Jun 2024, Last Modified: 17 Sep 2024 Accepted by TMLR Everyone Revisions BibTeX CC BY 4.0

Abstract: Large language models (LLMs) have shown impressive performance on complex reasoning by leveraging chain-of-thought (CoT) prompting to generate intermediate reasoning chains as the rationale to infer the answer. However, existing CoT studies have primarily focused on the language modality. We propose Multimodal-CoT that incorporates language (text) and vision (images) modalities into a two-stage framework that separates rationale generation and answer inference. In this way, answer inference can leverage better generated rationales that are based on multimodal information. Experimental results on ScienceQA and A-OKVQA benchmark datasets show the effectiveness of our proposed approach. With Multimodal-CoT, our model under 1 billion parameters achieves state-of-the-art performance on the ScienceQA benchmark. Our analysis indicates that Multimodal-CoT offers the advantages of mitigating hallucination and enhancing convergence speed. Code is publicly available at https://github.com/amazon-science/mm-cot.

Submission Length: Regular submission (no more than 12 pages of main content)
Code: https://github.com/amazon-science/mm-cot
Assigned Action Editor: Alessandro Sordoni
Submission Number: 2124.

Filter by reply type... Filter by author... Search keywords... Sort: Newest First

Everyone

7 / 7 replies shown

Add: **Public Comment**

Decision by Action Editor UvPB

Decision by Action Editor UvPB (Alessandro Sordoni) 05 May 2024, 01:06 (modified: 09 May 2024, 21:39) Everyone Revisions

Claims And Evidence:
Claims about the effectiveness of the proposed architecture are convincing and accurate.

Audience:
TMLR audience interested in multi-modality and chain-of-thought reasoning in LLMs, especially in some of the failures incurred in the multi-modal setting, would be interested in this paper.

Recommendation: Accept as is

Comment:
Reviewers concur that this paper presents a valuable analysis for the application of CoT reasoning in multimodal models. One of the striking weaknesses of this paper is basically novelty of the architecture, given that it shares a lot of similarity with the LLAVA approach. Even if novelty is not a determining factor for acceptance at TMLR, the novelty of the proposal might affect the interestingness of the work for TMLR, which is a determining factor. Nevertheless, it caught to my attention that the authors state "It is worth noting that Chameleon, LLaMA-Adapter, LLaVA, and InstructBLIP are concurrent works released several months after our work", this prompted me to verify that the first version of this paper appeared on arXiv a while ago, in 2023. In this light, I will consider this work indeed concurrent to LLaVA and the preceeding works. I appreciated the effort that the authors made to rerun their experiments (Table 2) and include more recent baselines (i.e. LLaVA in Table 5) along with the experiments on MMMU, which were in response to a reviewer concern.

Add: **Public Comment**

Review of Paper2124 by Reviewer Cqj7

Review by Reviewer Cqj7 08 Mar 2024, 10:16 (modified: 08 Mar 2024, 10:16) Everyone Revisions

Summary Of Contributions:
This paper presents Multimodal CoT, which includes training a small-size vision-language model, and a two-stage prompting pipeline that separates rationale and answer generation. They first motivate their two-stage prompting design with a text-only model, and show that without vision features, CoT performs worse than direct prompting. Then they show that two-stage prompting improves the performance with vision features. Next, they demonstrate their fine-tuned vision-language models based on FLAN-Alpaca, and the evaluation on 2 benchmarks show that their models with <1B parameters perform on par with LLaVA, which is a much larger model.

Strengths And Weaknesses:
Strengths:

The evaluation results are decent considering the small model size.

Weaknesses:

1. The fundamental limitation of this work is about its novelty and significance. Despite that the approach is called multimodal CoT, the approach is not a new prompting technique. Instead, the major part of this work is on how to have vision features and small-scale pretrained language models and then fine-tune the resulted vision-language model to perform multimodal reasoning. However, this submission misses the recent rapid progress of large multimodal models. For example, Gemini and GPT-4V have demonstrated impressive performance across various multimodal reasoning benchmarks, while there are neither discussion nor empirical comparison to such SOTA models. Also, the authors need to evaluate those multimodal benchmarks that are used in the Gemini technical report, instead of only evaluating on 2 benchmarks.
2. This work designs the two-stage prompting and demonstrates that it improves the performance. However, note that the two-stage prompting trains 2 separate models for rationale and answer generation, which increases the total model size. From this perspective, the effectiveness of two-stage prompting mainly comes from the insufficient model size, and it is unclear whether it is necessary with more powerful and larger multimodal models.
3. Table 7 is confusing. Did the authors use the text-only chatGPT? If so, how did they feed images into the model? Also, I don't think the results with generated rationales and human annotations are "comparable" as described in the paper, the performance with generated rationales is clearly worse.

Requested Changes:
1. Justify the novelty and significance of the approach, especially two-stage prompting.
2. Add missing related work discussion and empirical comparison to recent large multimodal models.
3. Add evaluation on more recent multimodal reasoning benchmarks, such as those evaluated in the Gemini technical report.
4. Explain how to compute the results with image input for chatGPT in Table 7.

Broader Impact Concerns:
No concern.

Claims And Evidence: Yes
Audience: Yes

Add: **Public Comment**

Responses to Reviewer Cqj7

Official Comment by Authors 20 Mar 2024, 12:48 (modified: 21 Mar 2024, 00:46) Everyone Revisions

Comment:
Thank you Reviewer Cqj7 for valuable feedback.

> RC1. Justify the novelty and significance of the approach, especially two-stage prompting.

Our work is in sharp contrast with previous studies in the following three aspects.

Firstly, this work is the first to study CoT reasoning in different modalities among existing publications.

Secondly, as a pioneer and effective approach to deal with the multimodal CoT problem, our technique is different from either previous text-only reasoning or VQA approaches by providing a means of explicit intermediate reasoning with multimodal input powered by a frozen image encoder and two-stage framework. We show that our framework offers the advantages of mitigating hallucination and enhancing convergence speed.

Thirdly, our two-stage approach elicits the analysis of why the naive way of employing CoT fails in the context and how incorporating vision features alleviates the problem. The approach has been shown generally effective across tasks and backbone models. Furthermore, we conduct in-depth studies to investigate the benefits and limitations of Multimodal-CoT to facilitate future studies.

> RC2. Add missing related work discussion and empirical comparison to recent large multimodal models.

We have added more recent studies and empirical comparisons to recent large multimodal models in Section 1 and Section 6.6, respectively.

> RC3. Add evaluation on more recent multimodal reasoning benchmarks, such as those evaluated in the Gemini technical report.

We have added experiment results on another widely-used multimodal reasoning benchmark, MMMU. It is a comprehensive benchmark designed to evaluate multimodal models on massive multi-discipline tasks demanding college-level subject knowledge and deliberate reasoning. MMMU meticulously collected multimodal questions from college exams, quizzes, and textbooks, covering six core disciplines: Art & Design, Business, Science, Health & Medicine, Humanities & Social Science, and Tech & Engineering.

We test Multimodal-CoT on MMMU without additional training. It is evident that Multimodal-CoT demonstrates effective generalization to MMMU, achieving better performance than various larger models around 8B. We have added the results and discussions in Section 6.6.

Method	Size	Accuracy
Kosmos-2	1.6B	24.4
Fuyu	8B	27.9
OpenFlamingo-2	9B	28.7
MiniGPT4-Vicuna	13B	26.8
Multimodal-CoT	738M	28.7
GPT-4V(ision)	-	56.8
Gemini Ultra	-	59.4

In addition, we also report the performance of GPT-4V and Gemini Ultra only for reference because it is not a fair comparison due to the large model scale and the lack of training details of those models.

[1] Yue, Xiang, et al. "Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi." arXiv preprint arXiv:2311.16502 (2023).

> RC4. Explain how to compute the results with image input for ChatGPT in Table 7.

We use the text-only chatGPT. For computing the results with image input for ChatGPT in Table 7, we only pass the question to ChatGPT, without the image input.

Because ScienceQA contains questions with images and without images, we leverage InstructBLIP and ChatGPT to generate the rationales for questions with paired images and questions without paired images, respectively. Then, we combine both of the generated pseudo-rationales as the target rationales for training (Multimodal-CoT w/ Generation) instead of relying on the human annotation of reasoning chains (Multimodal-CoT w/ Annotations). We have added the detailed explanation in Section 6.2.

The key takeaway message in Table 7 is that Multimodal-CoT is free from human-annotated rationales by making better use of automatically generated rationales. With those generated rationales, Multimodal-CoT achieves much better performance than using InstructBLIP or ChatGPT.

Add: **Public Comment**

Review of Paper2124 by Reviewer omWv

Review by Reviewer omWv 05 Mar 2024, 23:59 (modified: 08 Mar 2024, 10:16) Everyone Revisions

Summary Of Contributions:
This paper proposes a two-stage pipeline for multi-modal question answering. The pipeline contains two stages. In the first stage it generates rationale and then it generates the final answer using a separate network, conditioning on the original questions as well as the generated rationales. In both stages it trains a multi-modal network that takes both text and images as input. The authors analyze how hallucinated rationales mislead the performance of the network. In experiments, they demonstrate that their methods can outperform several large vision language models with relative small networks with limited sizes.

Strengths And Weaknesses:
Strength:

- The paper provides a nice analysis about the motivation of introducing multimodal inputs for question answering and demonstrated the bad effects of the hallucinated rationales. The explanation is clear and reasonable .
- The proposed two stage pipeline sounds reasonable for me. It achieves good performance in various dataset

Weakness: the proposed method seems straightforward and not novel. However, as I am not the expert in the field. I can only make an educational guess.

Requested Changes:
No requested changes

Broader Impact Concerns:
no broader concerns

Claims And Evidence: Yes
Audience: Yes

Add: **Public Comment**

Responses to Reviewer omWv

Official Comment by Authors 20 Mar 2024, 12:49 Everyone

Comment:
Thank you Reviewer omWv for valuable feedback.

> Weakness: the proposed method seems straightforward and not novel. However, as I am not the expert in the field. I can only make an educational guess.

Our work is in sharp contrast with previous studies in the following three aspects.

Firstly, this work is the first to study CoT reasoning in different modalities among existing publications.

Secondly, as a pioneer and effective approach to deal with the multimodal CoT problem, our technique is different from either previous text-only reasoning or VQA approaches by providing a means of explicit intermediate reasoning with multimodal input powered by a frozen image encoder and two-stage framework. We show that our framework offers the advantages of mitigating hallucination and enhancing convergence speed.

Thirdly, our two-stage approach elicits the analysis of why the naive way of employing CoT fails in the context and how incorporating vision features alleviates the problem. The approach has been shown generally effective across tasks and backbone models. Furthermore, we conduct in-depth studies to investigate the benefits and limitations of Multimodal-CoT to facilitate future studies.

Add: **Public Comment**

Review of Paper2124 by Reviewer MP8C

Review by Reviewer MP8C 25 Feb 2024, 20:06 (modified: 08 Mar 2024, 10:16) Everyone Revisions

Summary Of Contributions:
The paper proposes a method of leveraging Chain-of-Thought (CoT) prompting in a multimodal setting, to improve performance on vision-language tasks. They analyze both prompt based methods and fine-tuned models (of around 1B parameters). The paper also investigates why models of 1B-parameter scale fail at CoT reasoning, which they find is partly due to hallucinated rationales. The proposed method has two stages: (1) rationale generation and (2) answer inference. They show that the proposed method outperforms much larger models on the ScienceQA and A-OKVQA datasets.

Strengths And Weaknesses:
Strengths:

- The proposed method is quite general, and strong for its size (< 1B). It outperforms most other larger models (in the 6-11B scale) on ScienceQA. The method is also complementary with stronger VLMs, and the authors show (Sec. 6.2) that these models can be used to further improve results by training on their generated outputs rather than on human annotated reasoning chains.
- The analysis on failure modes of CoT and introducing multimodal inputs in Sec. 3 are informative.

Weaknesses:

- As all the models are benchmarked on fine-tuned performance, it's unclear how well the proposed method generalizes outside of the fine-tuning domain. It would be helpful to see such experiments, e.g., testing the ScienceQA fine-tuned model on MMMU [1] (both on the science and non-science domains).
- Similar to the above, does the fine-tuned model still do well on other vision-language tasks such as image captioning? Or is it highly specialized for VQA-type problems?
- Ablations on the importance of the visual backbone and language backbone would be appreciated (e.g., model size, image resolution, etc.). It would be useful to have information on how using larger visual encoders and language models affect performance.

References

[1] Yue, Xiang, et al. "Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi." arXiv preprint arXiv:2311.16502 (2023).

Requested Changes:
Would strengthen the work:

- Results on generalization or transfer to datasets that the model was not trained on.
- Analysis on different language model and visual backbones.

Broader Impact Concerns:
There are no major issues with the ethical implications or broader impact of the work.

Claims And Evidence: Yes
Audience: Yes

Add: **Public Comment**

Responses to Reviewer MP8C

Official Comment by Authors 20 Mar 2024, 12:50 (modified: 20 Mar 2024, 13:14) Everyone Revisions

Comment:
Thank you Reviewer MP8C for valuable feedback.

> RC1. Results on generalization or transfer to datasets that the model was not trained on.

Great point. Following your comment, we have added experiment results on another widely-used multimodal reasoning benchmark, MMMU. We directly test Multimodal-CoT on MMMU without additional training.

It is evident that Multimodal-CoT demonstrates effective generalization to MMMU, achieving better performance than various larger models around 8B. We have added the results and discussions in Section 6.6.

Method	Size	Accuracy
Kosmos-2	1.6B	24.4
Fuyu	8B	27.9
OpenFlamingo-2	9B	28.7
MiniGPT4-Vicuna	13B	26.8
Multimodal-CoT	738M	28.7
GPT-4V(ision)	-	56.8
Gemini Ultra	-	59.4

In addition, we also report the performance of GPT-4V and Gemini Ultra only for reference because it is not a fair comparison due to the large model scale and the lack of training details of those models.

> RC2. Analysis on different language model and visual backbones.

Yes. We conduct analysis on different language model and visual backbones to gain insights in the following ways.

Firstly, we alter the underlying LMs to other variants in different types. We see that our approach is generally effective for the different language models.

Language Backbone	Accuracy
UnifiedQA	82.55
FLAN-T5	83.19
FLAN-Alpaca	85.31

Secondly, we compare three widely-used types of vision backbones. We observe that VIT achieves relatively better performance.

Vision Backbone	Accuracy
VIT	85.31
CLIP	84.27
DETR	83.16

Thirdly, we have tried intricate alignment strategies inspired by BLIP, i.e., Image-grounded text encoder according to Section 3.1 in BLIP. This alignment approach injects visual information by inserting one additional cross-attention (CA) layer between the self-attention (SA) layer and the feed-forward network (FFN) for each transformer block of the text encoder. Our current strategy in the paper is similar to the Unimodal encoder as in BLIP, which is used for comparison. We see that using other alignment strategies also contributes to better performance than direct answering.

Method	Accuracy
Direct Answering	82.62
Unimodal encoder	85.31
Image-grounded text encoder	84.60

Add: **Public Comment**

OpenReview is a long-term project to advance science through improved peer review, with legal nonprofit status through Code for Science & Society. We gratefully acknowledge the support of the OpenReview Sponsors. © 2024 OpenReview