Review of Paper2124 by Reviewer Czj7

Review by Reviewer Czj7 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 finetuned 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 fuse vision features and small-scale pretrained language models and then finetune 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**