Cover Letter

Amita Kamath

Microsoft ROAR Internship Applicant · https://amitakamath.github.io · (she/her)

My name is Amita Kamath. I am a PhD student with Prof. Kai-Wei Chang at UCLA, a visiting PhD student with Prof. Ranjay Krishna at UW, and a frequent collaborator with the Allen Institute for Artificial Intelligence (AI2). Previously, I worked with Percy Liang's group during my MS at Stanford.

Here are some projects I've been thinking about that I believe would be of interest to you:

Improving Calibration of Multimodal LLMs

My thesis work at Stanford revolved around selective prediction in NLP: improving the ability of models to know what they don't know, so they abstain from prediction rather than returning to the user an incorrect output (Kamath et al., 2020; Kamath, 2020). The importance of this problem has become more pressing with the rise of LLMs and multimodal LLMs: on one hand, we have an increase in use of these models by people not well acquainted with machine learning; on the other, we have increasingly complex models that remain poorly calibrated, more so after the RLHF finetuning that is critical to their ability (Ouyang et al., 2022).

For the past three years, I have been researching vision-language models: identifying flaws in their architecture (Kamath et al., 2023a) and their reasoning abilities (Kamath et al., 2023b), tracing their behavior back to their pre-training (in ongoing work). I believe that my experience has set me up to tackle this challenge: how can we improve uncertainty estimation of multimodal LLMs, enabling us to deploy them responsibly and without the fear of hallucination?

Pre-training Techniques to Improve Reasoning Capabilities of Multimodal LLMs

As mentioned above, my current work in collaboration with AI2 studies what we call a *reporting bias in vision-language*: specifically, identifying what people do and don't mention when captioning images in web-scale corpora, and how that impacts what models do and don't learn when trained on these corpora. This work has raised interesting questions about what makes a good caption to pre-train multimodal models: for example, visually descriptive captions work well to train image-to-text generation models, as in DALLE-3 (Betker et al., 2023), but what kind of captions are needed for *reasoning* tasks?

I believe there is potential for an interesting study here to improve the reasoning capabilities of vision-language models, without causing adverse effects such as hallucination (going hand-in-hand with my first proposal).

I'm excited to discuss these topics with you, along with ideas I have about drawbacks to current pre-training techniques, decoding methods to improve model performance, and others. I look forward to hearing from you!

References

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. 2023. Improving image generation with better captions.

Amita Kamath. 2020. Selective prediction under domain shift for question answering. *MS Thesis*.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023a. Text encoders bottleneck compositionality in contrastive vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore. Association for Computational Linguistics.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023b. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore. Association for Computational Linguistics.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *ACL*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.