Building Controllable Multimodal Language Model Agents

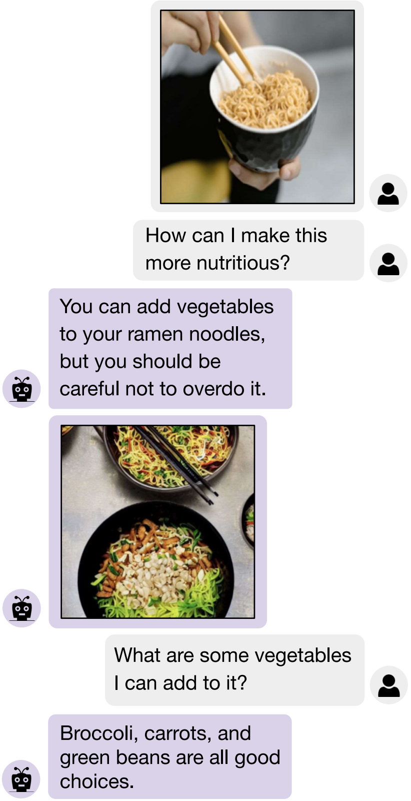
Jing Yu Koh (jykoh.com, jingyuk@cs.cmu.edu) *Carnegie Mellon University*

My research aims to build controllable machine learning models that integrate language, vision, and more, to achieve strong performance on reasoning and decision making tasks. While existing large language models (LLMs) exhibit impressive performance on processing and generating text, the text-only domain imposes inherent limitations. Text-only models are limited in their capabilities for understanding visual settings (such as web pages), producing visual outputs, or taking actions. For example, consider the simple task of finding a dinner spot. Instead of reading text about a restaurant, a user would likely derive much greater value from a multimodal agent which can autonomously navigate the restaurant's website to make reservations, identify and display relevant pictures of food, and answer specific questions from the user. Multimodal language model agents capable of doing such tasks autonomously would unlock many new applications beyond the capabilities of text-only models.

In the next few years, I intend to focus on several complementary research threads. I'm interested in finding methods to (1) improve the grounded language understanding and generation abilities of existing pretrained LLMs, and (2) benchmark and develop stronger multimodal language model agents towards building models that can perform *any* computer task. Successfully achieving these goals would pave the way towards automating many routine tasks, augmenting human productivity and user experience.

Grounding LLMs. State-of-the-art LLMs today boast impressive capabilities [2, 3]. However, they are usually trained on text-only data, and not explicitly exposed to the rich visual cues present in the real world. Despite recent efforts on building multimodal models, many LLMs today are still unable to effectively solve visually grounded problems [4, 5], and display far below human performance. My research aims to make progress towards efficiently enabling existing text-only LLMs to generalize to new applications, such as multimodal tasks. Towards this goal, our recent work [6, 1] propose the first models capable of processing image-text inputs and generating text-and-image outputs (Fig. 1). We efficiently ground text-only LLMs to images, enabling LLMs to generate both images and text. We demonstrated the effectiveness of our approach on few-shot multimodal tasks, significantly improving over existing models on multimodal dialogue and conversational text-to-image generation. By leveraging abilities learnt from large-scale text pretraining, our models exhibit improved understanding of user intents, and generate more relevant outputs. More generally, this work introduced parameter-efficient methods for training LLMs to use external tools to achieve new capabilities (i.e., image generation).

Building Multimodal Language Agents. Autonomous agents capable of planning, reasoning, and taking actions offer a promising avenue for automating computer tasks. However, the majority of existing benchmarks primarily focus on text-based language agents [7, 8], neglecting the direct utilization of visual inputs. Given that most computer interfaces cater to human perception, visual information often augments textual data in ways that text-only models struggle to harness effectively. My current project aims to bridge this gap by introducing *VisualWebArena*, a visually grounded web agent benchmark that builds upon the WebArena [8] work. *VisualWebArena* is a set of over 900 realistic, diverse, and complex web-based tasks that evaluate the capabilities of autonomous agents in handling visually grounded tasks which involve processing image inputs. To do well on this benchmark, autonomous agents have to accurately process image-text inputs, interpret instructions in natural language, and execute actions to accomplish user-defined objectives on realistic web pages. This benchmark provides a framework for evaluating and improving multimodal autonomous language agents, and our analysis provides insights towards building stronger autonomous agents. We hope to release this benchmark in early 2024, and believe that it will further spur research into strong autonomous web agents.



How can I make this more nutritious?

You can add vegetables to your ramen noodles, but you should be careful not to overdo it.

What are some vegetables I can add to it?

Broccoli, carrots, and green beans are all good choices.

Figure 1: GILL [1] grounds a language model to the visual domain, enabling it to process arbitrarily interleaved image-text inputs and generate coherent image-text outputs. Speech bubbles in purple are model generated, grey bubbles are input prompts.

Further Applications. Building upon the foundation of these advances, I plan to work on building strong multimodal language model agents equipped with the skills to understand, navigate, and interact with web pages. I envision future work on developing improved model architectures for autonomous agent tasks, finetuning multimodal language models for web understanding, and building strong benchmarks for concretely measuring progress in the field. I also plan to continue working on grounded language understanding problems, bridging text-only LLMs to visual and embodied settings, in order to apply their strong reasoning, generation, and dialogue capabilities on a wider range of real world applications. I am excited to continue working on these problems, and pave the way towards more communicative, robust, and capable artificial intelligence.

References

[1] J. Y. Koh, D. Fried, and R. Salakhutdinov, “Generating images with multimodal language models,” *NeurIPS*, 2023.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, “Language models are few-shot learners,” *NeurIPS*, 2020.

[3] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, *et al.*, “Palm: Scaling language modeling with pathways,” *arXiv preprint arXiv:2204.02311*, 2022.

[4] T. Thrush, R. Jiang, M. Bartolo, A. Singh, A. Williams, D. Kiela, and C. Ross, “Winoground: Probing vision and language models for visio-linguistic compositionality,” in *CVPR*, 2022.

[5] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, and W. Chen, “Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi,” 2023.

[6] J. Y. Koh, R. Salakhutdinov, and D. Fried, “Grounding language models to images for multimodal generation,” *ICML*, 2023.

[7] E. Z. Liu, K. Guu, P. Pasupat, T. Shi, and P. Liang, “Reinforcement learning on web interfaces using workflow-guided exploration,” *arXiv preprint arXiv:1802.08802*, 2018.

[8] S. Zhou, F. F. Xu, H. Zhu, X. Zhou, R. Lo, A. Sridhar, X. Cheng, Y. Bisk, D. Fried, U. Alon, *et al.*, “Webarena: A realistic web environment for building autonomous agents,” *arXiv preprint arXiv:2307.13854*, 2023.