VLanGOGh: Vision + Language-guided Generalized Object Grasping

Team: **StarkInc**

Nikhilanj Pelluri (pellu003@umn.edu)

Introduction

Large Language Models (LLMs) like GPT-3, GPT-4 show SOTA performance in Natural Language Understanding. They also encode a vast amount of general context about the world as they're trained on internet-scale corpora. We aim to use these capabilities to build **a robot that can understand natural language instructions.** However, to be usable in the real-world, LLMs need to understand their current context. We ground LLMs using visual input, use SOTA prompting methods, and achieve excellent results.

Discussion

Conclusions

-- Grounding for LLMs is an active research area. We showcase **visual grounding for GPT-3/3-5**.

-- Chain-of-Thought reasoning and few-shot prompts help **fine-tune output** and reduce LLM "hallucination".

-- Relying only on prompting tricks to fine-tune LLMs is not a viable strategy for production/critical apps.

Future Work

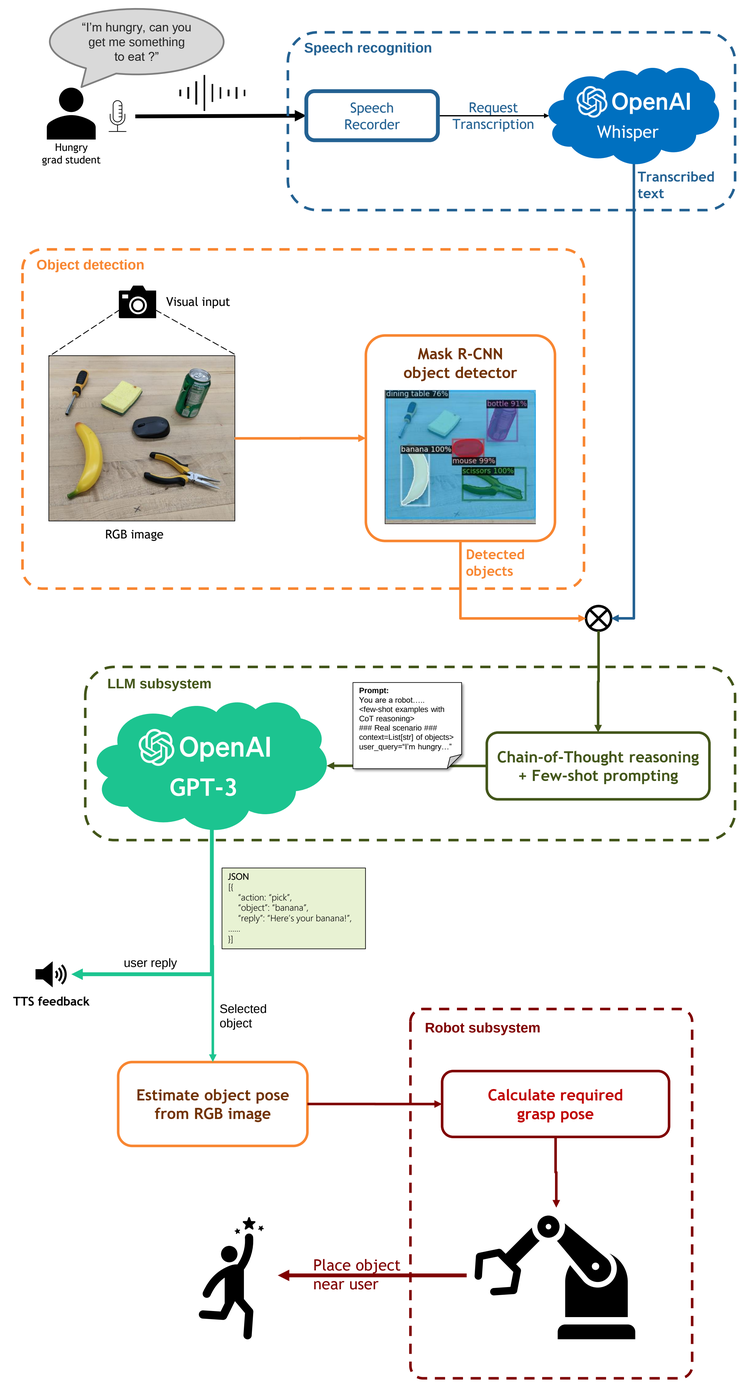
-- Make the system interactive – offer users options, etc.

-- Use multimodal models like CLIP to match objects in a single stage, instead of detection + matching.

-- Extend the system to unseen objects using models like OWL-ViT.

-- Extend the system to complex multi-step tasks.

Methods



Results

We demonstrate that:

-- GPT-3 shows remarkable ability in understanding the task at hand with just a few manual annotations.

-- LLMs can be tuned to perform a wide range of tasks using other sensory inputs to build a "local world" model for the LLM.

A short demo video of the system is available here: https://z.umn.edu/vlangogh

References

[1] M. Ahn *et al.*, "**Do as I can, not as I say: Grounding language in robotic affordances**" arXiv, Aug. 16, 2022. *doi: 10.48550/arXiv.2204.01691.*

[2] W. Huang et al., "**Inner Monologue: Embodied Reasoning through Planning with Language Models**" arXiv, Jul. 12, 2022. *doi: 10.48550/arXiv.2207.05608.*

[3] T. Brown *et al.*, "**Language models are few-shot learners**" in *Advances in Neural Information Processing Systems*, 2020, pp. 1877–1901.

[4] J. Wei *et al.*, "**Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,**" in *Advances in Neural Information Processing Systems*, 2022, pp. 24824–24837.

[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "**Mask R-CNN,**" in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 2980–2988. *doi: 10.1109/ICCV.2017.322.*

[6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "**Robust Speech Recognition via Large-Scale Weak Supervision**" arXiv, Dec. 06, 2022. *doi: 10.48550/arXiv.2212.04356.*