Modular Framework for Visuomotor Language Grounding

Kolby Nottingham, Litian Liang, Daeyun Shin, Charless C. Fowlkes, Roy Fox, Sameer Singh
University of California Irvine
knotting,litianl1,daeyuns,fowlkes,royf,sameer@uci.edu

Introduction

- Instruction following is useful for robotics and language grounding
- High quality data collection is expensive
- Current methods rely too much on labeled data
- We propose separating tasks into independent language, action, and vision (LAV) modules

ALFRED Benchmark

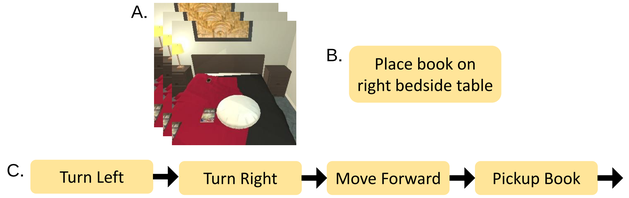

Figure 1: ALFRED Data Instance

ALFRED [2] instances are composed of:

- Fig 1.A: RGB observations
- Fig 1.B: Natural language instructions
- Fig 1.C: Discrete action trajectories

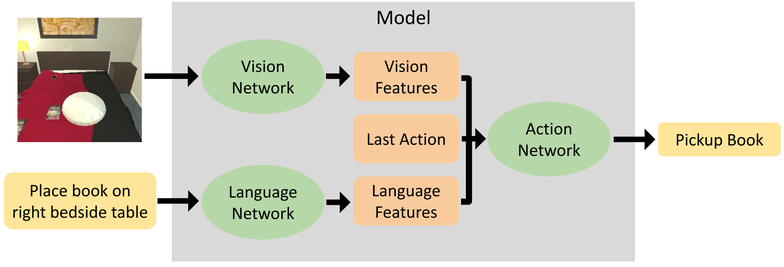

Figure 2: Baseline Approach

LAV Framework

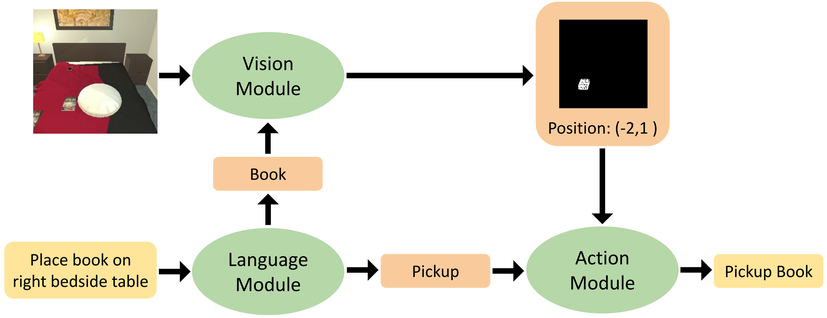

Figure 3: Language Action Vision Framework (LAV)

Language Module: Predicts subtask from language instructions

- Finetuned T5 language model
- Non-navigation actions used as subtasks
- Only module dependent on labeled data

Action Module: Completes subtask with target object

- Hardcoded depth first search to target object
- Attempts to execute non-navigation action
- Independent of labeled data

Vision Module: Estimates mask and position of target object

- Mask obtained from class segmentation model
- Position obtained from depth model
- Trained on simulator data, independent of labeled data

Results

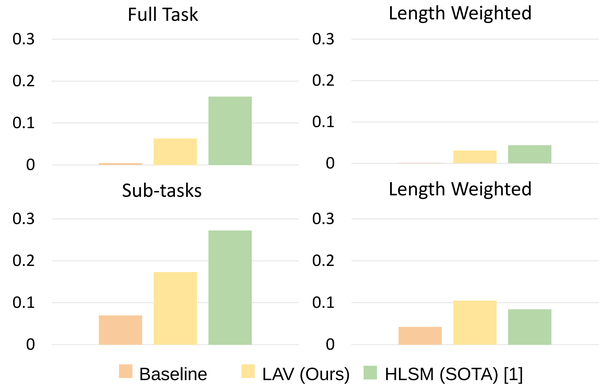

Figure 4: Success Rates

- Significant improvement over baseline
- Failures indicate need for improved action module

Conclusion

- Supervised task reduces to predicting subtasks and target objects
- Results indicate potential with more robust implementation
- Future work can ground language in other modalities
  - Details about visible objects from vision module
  - Probability of success over subtasks from action module

References

[1] Blukis, Valts, Chris Paxton, Dieter Fox, Animesh Garg, and Yoav Artzi “HLSM.” ALFRED Leaderboard. Accessed June 17, 2021. https://leaderboard.allenai.org/alfred/submissions/public.

[2] Shridhar, Mohit, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. "Alfred: A benchmark for interpreting grounded instructions for everyday tasks." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10740-10749. 2020.