A Planning based Neural-Symbolic Approach for Embodied Instruction Following

Xiaotian Liu
ServiceNow Research
Montreal, QC, Canada
jack.liu.to@gmail.com

Hector Palacios
ServiceNow Research
Montreal, QC, Canada
hector.palacios@servicenow.com

Christian Muise
Queen's University Canada
Kingston, ON, Canada
christian.muise@queensu.ca

Abstract

The ALFRED environment features an embodied agent following instructions and accomplishing tasks in simulated home environments. However, end-to-end deep learning methods struggle at these tasks due to long-horizon and sparse rewards. In this work, we propose a principled neural-symbolic approach combining symbolic planning and deep-learning methods for visual perception and NL processing. The symbolic model is enriched as exploration progress until a full plan can be obtained. New perceptions are added to a discrete graph representation that is used for producing new planning problems. Empirical results demonstrate that our approach can achieve high scalability with SOTA performance of 36.04% unseen success rate in the ALFRED benchmark. Our work builds a foundation for a neural-symbolic approach that can act in unstructured environments when the set of skills and possible relationships is known.

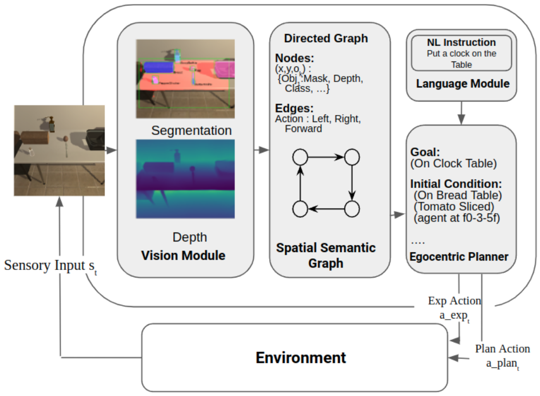

Figure 1. Model Overview

1. Introduction

Embodied instruction following require an agent to process multimodal information and plan over long task horizons. Recent advancements in deep learning (DL) models have made grounding visual and natural language information faster and more reliable [7] As a result, embodied task-oriented agents have been the subject of growing interest [9, 11, 12]. Benchmarks such as *The Action Learning From Realistic Environments and Directives* (ALFRED) was proposed to test embodied agents' ability to act in an unknown environment and follow language instructions [9]. The success of DL has led researchers to attempt end-to-end neural methods [10, 13]. In an environment like ALFRED, these methods are mostly framed as imitation learning, where neural networks are trained via a set of expert trajectories. However, end-to-end optimization leads to entangled latent state representation where compositional and long-horizon tasks are difficult to solve. Other approaches use neural networks to ground visual information into persistent memory structures to store information [1, 6]. These approaches rely on templates of existing tasks, making them difficult to generalize to new problems or unexpected action outcomes. Agents for ALFRED need to deal with the composition of fixed skills and long horizons despite being deterministic, and the environment remains unchanged except for the effect of the agent's actions. This motivates using methods specialized in the composition of skills. Classical planning is the most natural candidate given their high scalability [3]. However, classical planners assumes full observability. Our main innovation relies on combining DL models for perception and NLP with a new egocentric planner based on successive planning problems formulated using the PDDL syntax [2], both for exploration and task accomplishment.

We evaluated our approach on the ALFRED dataset and achieved the SOTA success rate of 36.04% on the unseen tasks. Compared with previous methods, our planning framework can naturally recover from action failures at any stage of the planned trajectory. In addition, by specifying a set of objects and skills, our agent can be easily generalized to other tasks with different goal compositions.

	Unseen				Seen			
	SR	GC	PLWSR	PLWGC	SR	GC	PLWSR	PLWGC
Our Approach	0.36	0.40	0.03	0.03	0.40	0.44	0.03	0.04
LGS-FR	0.34	0.40	0.15	0.20	0.43	0.46	0.20	0.26
FILM	0.28	0.39	0.11	0.15	0.29	0.40	0.11	0.16

Table 1. Performance Comparison

2. Approach

Our proposed method consist of a visual module for semantic segmentation and depth estimation, a language module for goal extraction, a semantic spatial graph for scene memorization, and an egocentric planner to conduct planning and inference. At time $t = 0$, we extract goal information from the high-level language instruction. The agent is then given a random exploration budget of 500 steps to explore the environment. Then, at $t = 500$, we convert information gathered via semantic spatial graph as a PDDL problem for the agent. We then use a novel open-loop replanning approach powered by an off-the-shelf planner to facilitate exploration and goal planning. An overview of our method can be seen in Figure 1

Vision and Language Module: We pre-trained MaskR-CNN and U-Net models on the AI2-THOR environment to conduct semantic segmentation and depth estimation [4, 5]. These models are used for identifying objects and obstacles in the environment. For task goals, we extracted features directly from labels produced by FILM authors that are trained on multiple transformers [6].

Semantic Spatial Graph: The spatial graph serves as the persistent memory of the agent during exploration. We use a directed graph with location and orientation as nodes and actions as edges. Object classes, segmentation masks, and depth information are stored in the nodes. For every task, we initialize the agent at position $(0, 0, 0)$, and the graph will be continually expanded via agent movements.

Egocentric Planning: We first specify the ontology of the planning problems which includes action schemas, object types, and possible facts schemas. The natural language task description is then converted into a planning goal. After the initial exploration phase, we use and update the semantic spatial graph —initially empty—, indicating the new position of the agent and what it perceives. The algorithm iterates over these steps. a) it attempt to find a plan for achieving the goal, and return it in case of success; b) if there is no plan, we replace the goal with another fact called `(explore)` associated with not visited states. We reduce the risk of executing irreversible actions by only attempting them when we have obtained a plan that should achieve the goal. In each iteration, we updated a semantic spatial graph to be used to build the new initial state of the agent, allowing an incremental egocentric view of the environment.

3. Experiments and Results

We use the same metric provided by the ALFRED leaderboard as evaluation criteria and benchmarked our result against FILM [6], which is the highest non-anonymous model on the leaderboard. Our model achieved an unseen success rate (SR) of 0.36 and 0.40 seen SR. This is an increase of 6%(21% relative) improvement on unseen tasks and 11%(38% relative) improvement on seen tasks. Our qualitative examination of the generated trajectories indicates that the planner's ability to handle failure recovery contributes the most to the performance of our method. There is also an anonymous entry, LGS-FR, that has relatively close performance to our model with an unseen SR of 34%. However no detail is shown regarding the approach which making comparison difficult. Overall, our method achieved SOTA on the ALFRED dataset in terms of both seen and unseen SR. A detailed comparison can be seen in Table 1. We also use our planner feedback to conduct error analysis on the test set. The analysis shows that task failures are mostly due to depth and agents being unable to locate the objects of interest. This failure is a combination of objects spawning inside of a receptacle, inefficient exploration, and failures in object recognition. The failure mode is shown in Table 2.

	seen	unseen
Objects Not Found	45%	24%
Collision	36%	64%
Interactions	9%	8%
Others	10%	4%

Table 2. Failure Modes

4. Conclusion and Future Work

In this work, we proposed a novel iterative replanning approach to solve embodied instruction following problems and achieved SOTA performance on the ALFRED benchmark. We demonstrated that automated planning powered by off-the-shelf planners could significantly improve reasoning through task decomposition and fault recovery. For future research, we would like to improve location mapping to account for geometric relationships among objects. We might explore extracting more information from step-by-step instructions to facilitate better exploration. We want to improve our egocentric planner to take into account non-deterministic effects, extending the scope of our method beyond ALFRED. That is a promising direction as there are efficient planning methods supporting actions with non-deterministic effects [8].

References

[1] Valts Blukis, Chris Paxton, Dieter Fox, Animesh Garg, and Yoav Artzi. A persistent spatial semantic representation for high-level natural language instruction execution. In *CoRL*, volume 164 of *Proceedings of Machine Learning Research*, pages 706–717. PMLR, 2021. 1

[2] Maria Fox and Derek Long. PDDL2.1: an extension to PDDL for expressing temporal planning domains. *J. Artif. Intell. Res.*, 20:61–124, 2003. 1

[3] Hector Geffner and Blai Bonet. A concise introduction to models and methods for automated planning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 8(1):1–141, 2013. 1

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *ICCV*, pages 2980–2988. IEEE Computer Society, 2017. 2

[5] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from CT volumes. *IEEE Trans. Medical Imaging*, 37(12):2663–2674, 2018. 2

[6] So Yeon Min, Devendra Singh Chaplot, Pradeep Ravikumar, Yonatan Bisk, and Ruslan Salakhutdinov. FILM: following instructions in language with modular methods. *CoRR*, abs/2110.07342, 2021. 1, 2

[7] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[8] Christian J. Muise, Vaishak Belle, and Sheila A. McIlraith. Computing contingent plans via fully observable non-deterministic planning. In *AAAI*, pages 2322–2329. AAAI Press, 2014. 2

[9] Mohit Shridhar, 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 *CVPR*, pages 10737–10746. Computer Vision Foundation / IEEE, 2020. 1

[10] Alessandro Suglia, Qiaozi Gao, Jesse Thomason, Govind Thattai, and Gaurav S. Sukhatme. Embodied BERT: A transformer model for embodied, language-guided visual task completion. *CoRR*, abs/2108.04927, 2021. 1

[11] Luca Weihs, Matt Deitke, Aniruddha Kembhavi, and Roozbeh Mottaghi. Visual room rearrangement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 1

[12] Karmesh Yadav, Santhosh Kumar Ramakrishnan, John Turner, Aaron Gokaslan, Oleksandr Maksymets, Rishabh Jain, Ram Ramrakhya, Angel X Chang, Alexander Clegg, Manolis Savva, Eric Undersander, Devendra Singh Chaplot, and Dhruv Batra. Habitat challenge 2022. https://aihabitat.org/challenge/2022/, 2022. 1

[13] Yichi Zhang and Joyce Chai. Hierarchical task learning from language instructions with unified transformers and self-monitoring. In *ACL/IJCNLP (Findings)*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4202–4213. Association for Computational Linguistics, 2021. 1