Language Guided Meta-Control for Embodied Instruction Following

Divyam Goel
Indian Institute of Technology Roorkee
dgoel@bt.iitr.ac.in

Kunal Pratap Singh
Allen Institute for AI
kunals@allenai.org

Jonghyun Choi
Yonsei University
jc@yonsei.ac.kr

Abstract

Embodied Instruction Following (EIF) is a challenging problem requiring an agent to infer a sequence of actions to achieve a goal environment state from complex language and visual inputs. We propose a generalised Language Guided Meta-Controller (LMC) for better language grounding in the large action space of the embodied agent. We additionally propose an auxiliary reasoning loss to improve the 'conceptual grounding' of the agent. Our empirical validation shows that our approach outperforms strong baselines on the Execution from Dialogue History (EDH) benchmark from the TEACh benchmark.

1. Introduction

Robotic assistants need to understand natural language and perceive their surroundings to interact with the environment. Many tasks and benchmarks have been proposed [1, 7, 9] to encourage the development of language-driven embodied agents. Building such embodied agents is a challenging problem on the cusp of important research directions in robotics, computer vision, and natural language processing. The agent must learn to infer a sequence of actions, including navigation and object manipulation, to attain a target-environment state from complex natural language directives and egocentric visual inputs. However, even with natural language instructions, this task is often challenging for the agent without any additional supervision. Additional supervision solves the following problems: 1) resolving ambiguities in natural language directives, 2) grounding instructions to environments with a rich action space, and 3) planning for long-horizon action sequences while recovering from possible failure modes [4].

Recent works suggest obtaining clarification to the ambiguities in the natural language instructions via simulated interactions [3, 6] or learning from human-human dialogue [10, 11] as possible directions to improving language-driven embodied navigation. We take a step forward in the task of embodied instruction following (EIF) with learning from human-human dialogues. Specifically, we present a Language Guided Meta-Controller (LMC) designed to improve language grounding in the agent's action space. The proposed method is based on an explicit high-level multisensory integration between learned natural language representation and the predicted actions.

2. Method

Here, we describe the details of the proposed model architecture of our Language Guided Meta-Controller (LMC) and the design of the auxiliary reasoning loss. We empirically discuss the progress monitor used in the ablation study. We train our agents using imitation learning, specifically behaviour cloning.

2.1. Language Guided Meta-Controller (LMC)

We illustrate the proposed LMC in Figure 1. It predicts the type of action to be taken next, either navigation or interaction, based solely on the natural language directive. The representations learned on account of the meta-controller are further updated based on the features extracted from the egocentric visual inputs before being used to predict the following action in the sequence. This two-step approach for predicting the action sequence of the agent based on the natural language directives helps improve language grounding in the large action space of the agent.

2.2. Auxiliary Reasoning Loss

Human acquisition of semantic representations is grounded in our visual and proprioceptive interactions with the environment. This phenomenon discusses the natural exploitation of complementarity between multiple senses and is commonly referred to as 'conceptual grounding' [2]. The proposed auxiliary reasoning loss asks the agent to predict the type of action to be taken next based on the entire observation space available to the agent, *i.e.*, the natural language directives, the egocentric visual observations, and the action history of the agent.

2.3. Progress Monitor

The auxiliary task of progress estimation [5] has been proposed to improve the generalization performance of em-

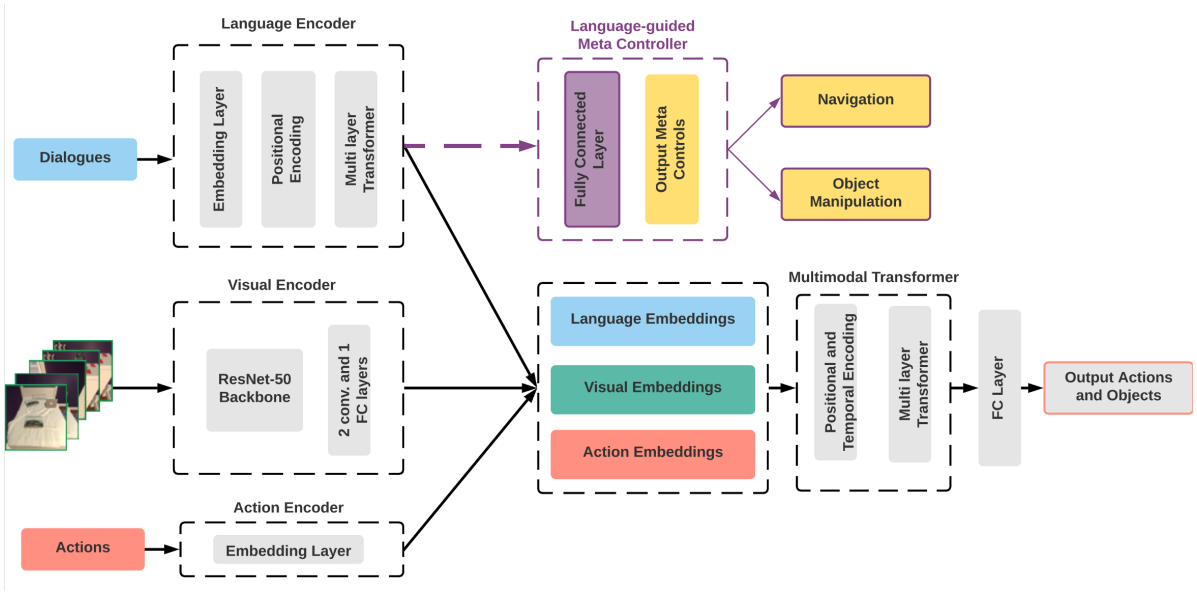

Figure 1. **Proposed Language Guided Meta-Controller for Embodied Instruction Following given a dialogue history between the commander and driver agents.** During training all predicted actions are used for gradient descent. At the time of inference, the last action is chosen and applied to the environment.

bodied agents in both seen and unseen environments. The Episodic Transformer baseline adapted for the EDH benchmark by [7] did not use any progress estimation signals. Here, we adapt the progress monitor [8] to the EDH benchmark.

3. Results and Discussion

We replicate the results of the Episodic Transformer (E.T.) model presented in [7] to establish the baseline performance for the Execution from Dialogue History (EDH) benchmark. Note that the results obtained for the E.T. baseline are higher than originally reported. This is because of a change in the evaluation pipeline, which now uses the action and image frames from the history of the EDH instance.

The proposed LMC helps the E.T. model draw some purchase from the input text. E.T. equipped with our meta-controller, significantly outperforms all the baseline models (See Table 1), including the E.T. trained with the auxiliary reasoning loss (+Aux) and the progress monitor (+PM). The self-attention layers of the multimodal transformer perform joint multimodal modelling in the E.T. baseline. The proposed auxiliary loss tries to improve this joint multimodal modelling of the agent. In contrast, the LMC induces a direct information flow from the learned language representations to the large action space of the model. This difference in modelling characteristics may be the reason for the performance gap noted here. Results imply that agents trained with the auxiliary loss and the progress monitor are better at transferring to unseen environments.

	Seen		Unseen	
Model	**SR [TLW]**	**GC [TLW]**	**SR [TLW]**	**GC [TLW]**
Random	0.82 [0.62]	0.75 [0.43]	1.34 [0.43]	0.41 [0.07]
Lang	0.99 [0.28]	1.04 [0.29]	2.36 [0.23]	0.78 [0.29]
Vision	5.1 [1.15]	6.96 [1.76]	3.89 [0.61]	3.56 [0.73]
E.T.	9.5 [2.8]	10.0 [7.5]	7.6 [2.2]	9.1 [7.3]
+LMC (Ours)	18.55 [5.6]	19.0 [12.1]	12.5 [3.8]	12.0 [11.5]
+Aux	10.9 [2.6]	11.6 [8.0]	10.7 [2.8]	11.0 [10.4]
+PM	8.2 [3.6]	9.5 [8.1]	10.5 [3.6]	10.4 [11.1]

Table 1. **EDH validation.** The Language Guided Meta-Controller outperforms all the baseline models. We empirically validate an auxiliary reasoning loss (Aux) and a progress monitor (PM). Metrics are success rate (SR), goal-conditioned success rate (GC) and trajectory length weighted metrics [in brackets]. All values are percentages. The higher the better.

4. Conclusion

We present a language guided meta-controller that enables a more robust grounding of the language directives into the agents' action space. We investigate baselines trained with a progress monitor and a novel auxiliary reasoning loss. We empirically validate by comparing the proposed methodologies with the Execution from Dialogue History (EDH) benchmark. The results establish the effectiveness of our approach as the proposed meta-controller outperforms strong baselines for the task of embodied instruction following, given a dialogue history.

References

[1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683, 2018. 1

[2] Lisa Beinborn, Teresa Botschen, and Iryna Gurevych. Multimodal grounding for language processing. *arXiv preprint arXiv:1806.06371*, 2018. 1

[3] Ta-Chung Chi, Minmin Shen, Mihail Eric, Seokhwan Kim, and Dilek Hakkani-tur. Just ask: An interactive learning framework for vision and language navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2459–2466, 2020. 1

[4] Xiaofeng Gao, Qiaozi Gao, Ran Gong, Kaixiang Lin, Govind Thattai, and Gaurav S Sukhatme. Dialfred: Dialogue-enabled agents for embodied instruction following. *arXiv preprint arXiv:2202.13330*, 2022. 1

[5] Chih-Yao Ma, Zuxuan Wu, Ghassan AlRegib, Caiming Xiong, and Zsolt Kira. The regretful agent: Heuristic-aided navigation through progress estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6732–6740, 2019. 1

[6] Khanh Nguyen and Hal Daumé III. Help, anna! visual navigation with natural multimodal assistance via retrospective curiosity-encouraging imitation learning. *arXiv preprint arXiv:1909.01871*, 2019. 1

[7] Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. Teach: Task-driven embodied agents that chat. *arXiv preprint arXiv:2110.00534*, 2021. 1, 2

[8] Alexander Pashevich, Cordelia Schmid, and Chen Sun. Episodic transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15942–15952, 2021. 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749, 2020. 1

[10] Alane Suhr, Claudia Yan, Jacob Schluger, Stanley Yu, Hadi Khader, Marwa Mouallem, Iris Zhang, and Yoav Artzi. Executing instructions in situated collaborative interactions. *arXiv preprint arXiv:1910.03655*, 2019. 1

[11] Jesse Thomason, Daniel Gordon, and Yonatan Bisk. Shifting the baseline: Single modality performance on visual navigation & qa. *arXiv preprint arXiv:1811.00613*, 2018. 1