An Analysis of Visually Grounded Instructions in Embodied AI Tasks

An Analysis of Visually Grounded Instructions in Embodied AI Tasks

Marco Grazioso[1,2,*,†], Alessandro Suglia[3,4,†]

[1]Interdepartmental Research Center Urban/Eco, University of Naples Federico II, Italy
[2]Logogramma s.r.l, Italy
[3]Heriot-Watt University, Scotland, UK
[4]AlanaAI, Edinburgh, United Kingdom

Abstract

Thanks to Deep Learning models able to learn from Internet-scale corpora, we observed tremendous advances in both text-only and multi-modal tasks such as question answering and image captioning. However, real-world tasks require agents that are embodied in the environment and can collaborate with humans by following language instructions. In this work, we focus on ALFRED, a large-scale instruction-following dataset proposed to develop artificial agents that can execute both navigation and manipulation actions in 3D simulated environments. We present a new Natural Language Understanding component for Embodied Agents as well as an in-depth error analysis of the model failures for this challenge, going beyond the success-rate performance that has been driving progress on this benchmark. Furthermore, we provide the research community with important directions for future work in this field which are essential to develop collaborative embodied agents.

Keywords

embodied AI, situated interaction, visual grounding, deep learning

1. Introduction

In recent years, we experienced tremendous improvements in Natural Language Understanding (NLU) tasks thanks to powerful Large Language Models (e.g., [1, 2, 3]). These models are trained by leveraging internet-scale textual data. However, by having access to text only, they leverage only a part of the rich multi-modal training data that can be derived from interaction with the world and with other agents [4]. Embodied Artificial Intelligence (EAI) is the field of AI that aims at developing agents that can perceive the environment via multi-modal inputs, and that can execute actions in the world.

Many benchmarks have been proposed so far in EAI. For instance, Vision+Language Navigation [5] aims at studying the capabilities of EAI agents to follow natural language instruction in 3D simulated environments. However, the agent can only output navigation actions limiting the richness of concepts that the agent can learn. To simulate a scenario that is closer to the real-world usage of these systems, Shridhar et al. [6] proposes ALFRED, a new instruction-following benchmark that facilitates the study of both situated language understanding as well as visual memory, commonsense reasoning, as well as long-term action planning.

So far, progress on ALFRED has been driven by accuracy-based metrics on the official leaderboard (e.g., [7, 8, 9, 10]). However, considering that the success rate on this benchmark is still below production-level performance (∼40%), this calls for a more in-depth analysis of model failures. In this paper, we provide two main contributions: 1) we train a novel Natural Language Understanding component for an EAI agent trained using multi-task learning that has a 0.117 error rate on the validation unseen of ALFRED, an improvement over the one proposed by Min et al.[7]; 2) we provide an in-depth analysis of our model's failures highlighting lack of important situated language understanding capabilities that are key for an EAI agent such as referential expression resolution, and conversational grounding [11].

2. The ALFRED dataset

In this study, we use ALFRED [6], a benchmark aimed at assessing the ability of embodied agents to learn from natural language instructions and egocentric vision to generate sequences of actions for household tasks. The ALFRED dataset comprises 25,743 human-annotated language directives corresponding to 8,055 expert demonstration episodes. Each directive includes a high-level goal and a set of step-by-step instructions. Directives fall under one of the following seven tasks parameterised

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
*Corresponding author.
†These authors contributed equally.
marco.grazioso@unina.it (M. Grazioso); a.suglia@hw.ac.uk (A. Suglia)
0000-0002-4056-544X (M. Grazioso); 0000-0002-3177-5197 (A. Suglia)

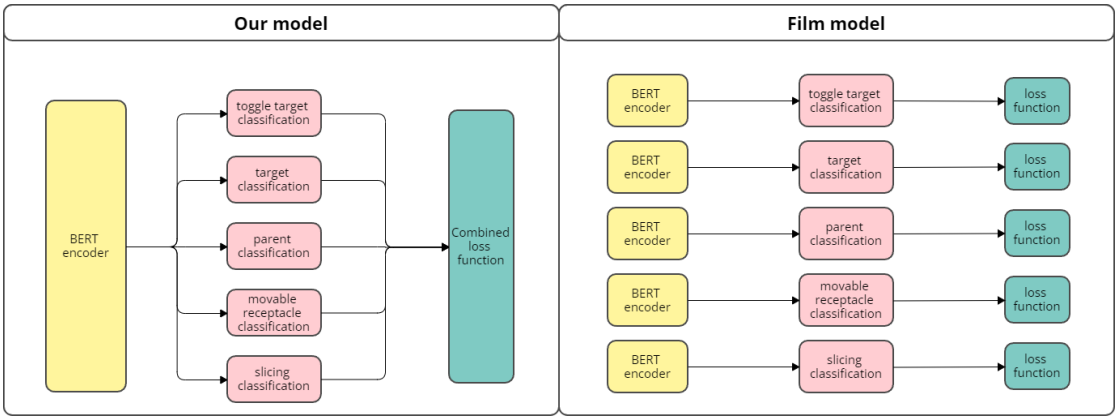

Figure 1: Comparison between our model and the one proposed in the FILM paper.

by 84 object classes in 120 scenes: pick and place, stack and place, pick two and place, clean and place, heat and place, cool and place, and examine in light. In addition to the task category, for each instruction, the dataset also provides a few relevant annotations: 1) **target object**: the principal object involved in the interaction; 2) **parent object**: the final destination of the target object (sink, counter, and similar); 3) **movable receptacle**: a movable object containing the target one, e.g. a spoon in a mug; 4) **slicing action**: true or false respectively if the target object must be cut or not; 5) **toggle target**, indicating an object to toggle on/off (oven, microwave, etc.).

To better estimate the models' capability to generalise to new environments, the validation set is composed of two subsets called seen and unseen, respectively. In the former, the agent has to complete tasks in rooms/scenarios that have already been seen during training, while the latter provides examples in unseen scenarios to assess the ability of the agent to generalise.

3. Baseline model and error analysis

We implemented the solution proposed in the FILM paper [7] which is the base approach for many of the state-of-the-art models for ALFRED (e.g., [8]). FILM is a modular architecture that is composed of a trained natural language component that derives a semantic representation represented in terms of intents and slot values akin to conventional NLU systems (e.g., [12]). This representation is in turn converted into an action plan by a rule-based component.

In this paper, we focus on improving the language understanding component of FILM, which is essential for instruction interpretation. Concretely, the FILM implementation casts instruction understanding as a classification task and fine-tunes one BERT-based model [13] for each of the following tasks: 1) **task classification**: the instruction sequence is classified into one of the seven task categories; 2) **target classification**: the instruction sequence is classified into one of the allowed target objects; 3) **movable receptacle classification**: the instruction sequence is classified into one of the allowed movable receptacle objects; 4) **parent classification**: the instruction sequence is classified into one of the allowed parent objects; 5) **slicing classification**: the instruction sequence is binary classified to be a slicing/non-slicing action.

However, training five different models each one having its own BERT encoder can be suboptimal because 1) it has a high computational cost, and 2) it does not consider the semantic relationships between a task and the objects required to solve that task. To take advantage of these relationships, we implemented a multi-task model by fine-tuning a single BERT encoder on all the tasks [14] (see Figure 1). As shown in Table 2, thanks to this multi-task setup, our model obtains an improvement in the overall language understanding performance measured using the error rate which we define as the proportion of examples for which no mistakes are made (i.e., neither on the high-level task nor on the single slots). Additionally, we report in Table 1, our model performance on specific high-level tasks measured using F1-score.

Despite the superior performance of our multitask model, the capabilities of this model were still limited. Therefore, we performed a manual error analysis based on 821 instructions from the validation unseen split of the ALFRED dataset. Particularly, as shown in Table 3, the most common errors are about a wrong object classification and, even if the object was correctly classified, a

Class	Ours F1-score	FILM F1-score
look_at_obj_in_light	0.994	1
pick_and_place_simple	0.985	0.91
pick_and_place_with_movable_recep	0.991	0.94
pick_clean_then_place_in_recep	0.991	0.98
pick_cool_then_place_in_recep	1	0.93
pick_heat_then_place_in_recep	0.996	0.97
pick_two_obj_and_place	0.988	0.9

Table 1
Comparison between our model scores and FILM's model scores in task classification on unseen validation set.

Model	Language processing error rate
FILM	0.196
Ours	**0.117**

Table 2
Comparison between our model error rate and FILM's model error rate on all language processing tasks on unseen validation set.

Error type	Subtype	Rate
Referential ambiguity	Mismatching	40/821
	Underspecification	24/821
	Others	32/821
Target object search	Object not visible	171/821
	Spatial understanding	106/821
Others interaction errors		218/821

Table 3
Error rate for each error type derived from our error analysis.

failure to find it in the environment. Specifically, we can categorise errors in two main classes namely REFERENTIAL AMBIGUITY, and TARGET OBJECT SEARCH, which we further divide into the following classes:

MISMATCHING OBJECT REFERENCE: the user refers to an object with a non-conventional name or a particular linguistic form due to visual ambiguity (*brown ball* or *potato* instead of *egg*).

UNDERSPECIFIED OBJECT REFERENCE: the user refers to an object using a name which could be ambiguous because not precise enough (typically "soap" is used to refer to a *soap bar* or a *soap bottle*).

OBJECT NOT FOUND BECAUSE NOT VISIBLE: this can happen when the target object is contained in other objects (e.g., spoons are contained in drawers).

SPATIAL UNDERSTANDING: the user gives nuanced spatial references for the object but the system does not understand them (e.g., *pick up the salt which is inside the cabinet under the coffee machine*).

Finally, we use a third class (OTHERS) which includes other interaction errors that do not depend on the language understanding component.

4. Challenges for embodied instruction following

Thanks to our error analysis, we derive that an embodied agent faces several challenges when fusing multiple modalities. Moreover, it must take care of the basic concepts of human-to-human communication [11].

In this context, the agent's reasoning can be seen as a sequential process in which it implements a set of strategies to follow the current language instruction. An embodied agent must rely on visual context, commonsense knowledge, and interactive skills. For instance, when the user asks for an object, e.g., soap, the agent must be able to understand that "soap", "soap bar" and "soap bottle" share enough features to define them as similar objects. Additionally, it should take advantage of the visual context to resolve ambiguities (if the only soap in the agent's field of view is a soap bar, this should be the target). Therefore, multi-modal information becomes crucial to understanding visually grounded instructions, going from spatial language instructions to multi-modal input ones [15]. Finally, if no other strategy resulted in a solution, it should ask for human intervention, e.g. using clarification strategies [16]. Furthermore, integrating commonsense knowledge can result in better interpretation (e.g., by leveraging knowledge graphs [17]) as well as better action plans by reasoning over pre-conditions and post-conditions of the actions.

In collaborative tasks [18, 19, 20], agents have to build common ground to successfully complete their tasks and adapt to new situations [11]. Therefore, negotiating meanings becomes a fundamental skill that allows the agent to learn how the user refers to the environment, and understand user preferences which will lead to a more effective interaction.

5. Conclusion

In this work, we used the ALFRED dataset as a benchmark to investigate the language understanding abilities of state-of-the-art EAI models. We started by improving the model originally proposed by Min et al. [7] by training using multi-task learning and we showed that even by using the new model several issues remain unsolved. We categorised these problems into different classes to facilitate our analysis. This classification led us to the conclusion that an EAI agent must leverage multi-modal signals, commonsense knowledge, and interaction with the user to solve embodied problems in an effective way.

According to Schlangen [21], *situated interaction is a direct, purposeful encounter of free and independent but similar agents.* Following this definition, in the ALFRED tasks there are two different agents: a follower and a leader. The leader is intended as an oracle that provides

instructions in one go without conversing with the follower. Moreover, the leader assumes that the follower has perfect capabilities to follow the provided instructions without considering the notion of uncertainty or potential mistakes. Finally, there is no concept of conversational grounding intended as a joint activity in which the two agents have to negotiate meanings that are required to solve the task effectively and efficiently. In this sense, even if the ALFRED dataset still represents a challenging task, it is far from providing a benchmark that can be used to develop artificial agents able to collaboratively solve tasks using natural language.

References

[1] 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, Advances in neural information processing systems 33 (2020) 1877–1901.

[2] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).

[3] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al., Bloom: A 176b-parameter open-access multilingual language model, arXiv preprint arXiv:2211.05100 (2022).

[4] Y. Bisk, A. Holtzman, J. Thomason, J. Andreas, Y. Bengio, J. Chai, M. Lapata, A. Lazaridou, J. May, A. Nisnevich, et al., Experience grounds language, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 8718–8735.

[5] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, A. 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, 2018, pp. 3674–3683.

[6] M. Shridhar, J. Thomason, D. Gordon, Y. Bisk, W. Han, R. Mottaghi, L. Zettlemoyer, D. Fox, ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020. URL: https://arxiv.org/abs/1912.01734.

[7] S. Y. Min, D. S. Chaplot, P. Ravikumar, Y. Bisk, R. Salakhutdinov, Film: Following instructions in language with modular methods, 2021. arXiv:2110.07342.

[8] Y. Inoue, H. Ohashi, Prompter: Utilizing large language model prompting for a data efficient embodied instruction following, arXiv preprint arXiv:2211.03267 (2022).

[9] A. Suglia, Q. Gao, J. Thomason, G. Thattai, G. Sukhatme, Embodied bert: A transformer model for embodied, language-guided visual task completion, arXiv preprint arXiv:2108.04927 (2021).

[10] A. Pashevich, C. Schmid, C. Sun, Episodic transformer for vision-and-language navigation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15942–15952.

[11] H. H. Clark, S. E. Brennan, Grounding in communication., in: Perspectives on socially shared cognition., American Psychological Association, 1991, pp. 127–149.

[12] Q. Zhu, Z. Zhang, Y. Fang, X. Li, R. Takanobu, J. Li, B. Peng, J. Gao, X. Zhu, M. Huang, Convlab-2: An open-source toolkit for building, evaluating, and diagnosing dialogue systems, arXiv preprint arXiv:2002.04793 (2020).

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[14] X. Liu, P. He, W. Chen, J. Gao, Multi-task deep neural networks for natural language understanding, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4487–4496. URL: https://aclanthology.org/P19-1441. doi:10.18653/v1/P19-1441.

[15] M. Grazioso, A. S. Podda, S. Barra, F. Cutugno, Natural interaction with traffic control cameras through multimodal interfaces, in: International Conference on Human-Computer Interaction, Springer, 2021, pp. 501–515.

[16] V. Russo, A. Mancini, M. Grazioso, M. Di Bratto, Graph-based representations of clarification strategies supporting automatic dialogue management, IJCoL. Italian Journal of Computational Linguistics 8 (2022).

[17] A. Origlia, M. Di Bratto, M. Di Maro, S. Mennella, A multi-source graph representation of the movie domain for recommendation dialogues analysis, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 1297–1306.

[18] H. De Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, A. Courville, Guesswhat?! visual

object discovery through multi-modal dialogue, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5503–5512.

[19] N. Ilinykh, S. Zarrieß, D. Schlangen, Meet up! a corpus of joint activity dialogues in a visual environment, in: Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue-Full Papers, 2019.

[20] A. Suhr, C. Yan, J. Schluger, S. Yu, H. Khader, M. Mouallem, I. Zhang, Y. Artzi, Executing instructions in situated collaborative interactions, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 2119–2130.

[21] D. Schlangen, What a situated language-using agent must be able to do: A top-down analysis, arXiv preprint arXiv:2302.08590 (2023).