11-777 Spring 2021 Class Project
ALFRED Background and Related Works

Christian Deverall* **Jingyuan Li** * **Artidoro Pagnoni** *
{cdeveral, jingyua4, apagnoni}@andrew.cmu.edu

1 Related Work

In this part, we investigate existing work related to our project. We focus on related techniques and tasks. The supporting techniques discussed here are (1) program synthesis/induction, (2) language-image grounding. The related task discussed here is multimodal embodied interaction.

1.1 Program Synthesis

One challenge of dealing with instructions expressed in natural language is that natural language is noisy. The same instructions can be expressed through different paraphrases and depending on the context the same instructions might describe different actions. The task of parsing natural language instructions into a representation that has a deterministic execution is generally referred to as program synthesis or inductive programming (Summers, 1977; Muggleton and De Raedt, 1994).

Neural Code Generation Previous work in program synthesis has explored the task of parsing natural language descriptions into source code written in a general-purpose programming language. Yin and Neubig (2017) proposes a neural architecture which uses a grammar to explicitly capture the syntax of the target language in an Abstract Syntax Tree. Dong and Lapata (2018) use a different approach which is based on coarse-to-fine decoding and decomposes the parsing process into two stages: a rough sketch of the meaning of input utterance is first generated, and it is further refined by filling in the missing information at a second stage.

Compositionality Instructions and questions are generally compositional. Being able to capture the compositional linguistic structure of natural language questions can help create and combine specialized modules. Andreas et al. (2016) uses a natural language parser to dynamically lay out a network composed of reusable modules that are jointly trained. The CLEVR dataset (Johnson et al., 2017) was proposed to test model's ability to perform compositional reasoning such as recognizing novel attribute combinations. In this sense it is similar to the ALFRED Shridhar et al. (2020a) instructions which test the performance of an agent in new action-object-scene combinations. Santoro et al. (2017) propose Relation Networks to augment CNNs and perform relational reasoning on the CLEVR dataset.

Semantic Parsing The problem of code program synthesis is related to semantic parsing where a natural language sentence is mapped into a cpmplete, formal meaning representation (Mooney, 2007). Several approaches have been proposed to leverage unlabeled data for the task of semantic parsing using question-answering and paraphrase models (Berant et al., 2013; Berant and Liang, 2014) .

1.2 Image Understanding

Natural Language Object Retrieval Traditional image recognition tasks such as object detection and semantic segmentation have performed extremely well when the textual input is highly defined. Often bounding boxes and segmentation masks are directly associated with a tag or label that defines the object. One limitation of this method is that it cannot not leverage the complex structural information inherent to natural language. In (Xiao et al., 2017), natural language object retrieval is performed on a relatively small dataset using weak supervision. More specifically, the authors represent sentences as a parse tree, which enables learning at several levels of the tree for the same image caption. This method outperformed baselines that did not consider linguistic structure on the MS COCO and

*Everyone Contributed Equally – Alphabetical order

Visual Genome datasets. For the same task, (Wu et al., 2017) uses deep reinforcement learning to iteratively reshape a bounding box to localize the object.

Visual QA In contrast to the previous papers which only localize objects within images, the visual question answering task involves answering a more complex question about the image. In (Antol et al., 2015), which initially proposed the VQA task, the best performing model uses an LSTM to separately encode the text while a CNN encodes the image. (Yang et al., 2015) improves on this by introducing stacked attention networks, which iteratively queries the image. By repeatedly refining queries that combine both image and textual data, the latter attention layers can focus on the most relevant parts of the image. Recently, papers such as (Wu et al., 2019) have provided innovations in the way that the image and textual encodings are fused together. Typical models take the product between both encodings, however this paper proposes the idea that the product of the difference between feature elements is a superior fusion.

1.3 Multimodal Embodied Interaction

The ability to understand and follow instructions is of great importance for robotic systems that aims to assist human in the real world. For its practicability, methods allowing the robots to follow instructions and complete specific goals have attracted great attention. One of the related tasks is called visual language navigation, which aims to generate sequences of actions from human instructions that guide the agent to complete tasks in specific scenes. The Room-to-Room dataset was proposed to simulate the real-world circumstance where instructions, scene images, and corresponding actions are provided, to accelerate the development of visual language navigation (Anderson et al., 2018). To deal with the visual language navigation task, the multi-modal mapping between action sequence and instructions is built, which not only allows the agent to plan before taking actions but also enables data augmentation which improves the robustness of the model (Fried et al., 2018). To allow the learning from environmental exploration and failure experiences, reinforcement learning techniques are also used in the navigation task (Ma et al., 2019; Wang et al., 2018). Besides performance improvements, many related works are also working on addressing specific assumptions made in the Room-to-Room dataset, including the discrete space assumption (Krantz et al., 2020), and the known environment assumption (tan). While the Room-to-Room dataset has modeled VLN tasks comprehensively, one limitation that cannot be addressed is the lack of interaction which is one of the eventual goals of real-world robotic systems. To take a step forward, the ALFRED dataset is proposed (Shridhar et al., 2020b), with both navigation and interaction taken into consideration. Due to the requirement of interaction, the agent needs not only to decide the actions to take but also to figure out which object in the environment should be acted on, introducing extra complexity to the already hard problem. To address the issue, MOCA (Singh et al., 2020) with modules dealing with visual perception and action policy was proposed, to make the task feasible by dividing different components of the targets. The above describes works and tasks related to our project. In this project, we are intended to work on the exploration of and interaction with environment by the agent.

References

P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3674–3683.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. *CoRR*, abs/1505.00468.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544.

Jonathan Berant and Percy Liang. 2014. Semantic parsing via paraphrasing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425.

Li Dong and Mirella Lapata. 2018. Coarse-to-fine decoding for neural semantic parsing. *arXiv preprint arXiv:1805.04793*.

Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. 2018. Speaker-follower models for vision-and-language navigation. In *Neural Information Processing Systems (NeurIPS)*.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910.

Jacob Krantz, Erik Wijmans, Arjun Majundar, Dhruv Batra, and Stefan Lee. 2020. Beyond the nav-graph: Vision and language navigation in continuous environments. In *European Conference on Computer Vision (ECCV)*.

Chih-Yao Ma, Zuxuan Wu, Ghassan AlRegib, Caiming Xiong, and Zsolt Kira. 2019. The regretful agent: Heuristic-aided navigation through progress estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Raymond J Mooney. 2007. Learning for semantic parsing. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 311–324. Springer.

Stephen Muggleton and Luc De Raedt. 1994. Inductive logic programming: Theory and methods. *The Journal of Logic Programming*, 19:629–679.

Adam Santoro, David Raposo, David GT Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. 2017. A simple neural network module for relational reasoning. *arXiv preprint arXiv:1706.01427*.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020a. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020b. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kunal Pratap Singh, Suvaansh Bhambri, Byeonghwi Kim, Roozbeh Mottaghi, and Jonghyun Choi. 2020. Moca: A modular object-centric approach for interactive instruction following. *arXiv preprint arXiv:2012.03208*.

Phillip D Summers. 1977. A methodology for lisp program construction from examples. *Journal of the ACM (JACM)*, 24(1):161–175.

Xin Wang, Wenhan Xiong, Hongmin Wang, and William Yang Wang. 2018. Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. In *ECCV (16)*, pages 38–55.

Chenfei Wu, Jinlai Liu, Xiaojie Wang, and Ruifan Li. 2019. Differential networks for visual question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8997–9004.

Fan Wu, Zhongwen Xu, and Yi Yang. 2017. An end-to-end approach to natural language object retrieval via context-aware deep reinforcement learning. *CoRR*, abs/1703.07579.

Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. 2017. Weakly-supervised visual grounding of phrases with linguistic structures. *CoRR*, abs/1705.01371.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. 2015. Stacked attention networks for image question answering. *CoRR*, abs/1511.02274.

Pengcheng Yin and Graham Neubig. 2017. A syntactic neural model for general-purpose code generation. *arXiv preprint arXiv:1704.01696*.