Language-Conditioned Policy Learning for Long-Horizon Robot Manipulation Tasks

Oier Mees and Wolfram Burgard
Department of Computer Science, University of Freiburg, Germany

I. Introduction

"... spend the summer linking a camera to a computer and getting the computer to describe what it saw."

Marvin Minsky on the goal of a 1966 undergraduate summer research project [2].

A long-standing goal in robotics is to build robot systems that can perform a wide range of everyday tasks given onboard sensor data and instructions from the user. Doing so requires the robot to acquire a diverse repertoire of general-purpose skills and non-expert users to be able to effectively specify tasks for the robot to solve. This stands in contrast to most current end-to-end models, which typically learn individual tasks one at a time from manually-specified rewards and assume tasks being specified via goal images [16] or one-hot skill selectors [13], which are not practical for untrained users to instruct robots. Natural language presents a promising alternative form of specification, providing an intuitive and flexible way for humans to communicate tasks and refer to abstract concepts. Despite the tremendous progress made in visual and language understanding since this now famously ambitious summer project from one of the AI pioneers, we are far away from achieving robots that can learn to relate human language to their world model. As robots become ubiquitous across human-centered environments the need for intuitive task specification grows: how can we scale robot learning systems to autonomously acquire general-purpose knowledge that allows them to compose long-horizon tasks by following unconstrained language instructions?

Understanding and following unconstrained language instructions is a notoriously challenging problem, subsuming many long term problems in AI [9, 28]. For example, a robot presented with the command "fetch the banana and place it left of the bottom object" must be able to relate language to its low-level perception (what does a banana look like?). It must perform visual and spatial reasoning about where to place the "banana" relative to the "bottom object" in order to reproduce the spatial relation "to the left of", which is inherently ambiguous as natural language placement instructions do not uniquely identify a location in a scene. Additionally, it must solve a complex sequential decision problem (what commands do I send to fetch an object, or to do a relative placement?). In my work, I have focused on addressing the challenging problem of relating human language to a robot perceptions and action by introducing techniques that leverage self-supervision and structural priors to enable sample-efficient learning of language-conditioned manipulation tasks.

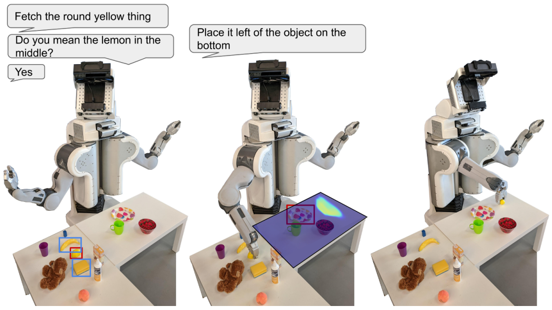
Fetch the round yellow thing
Do you mean the lemon in the middle?
Yes
Place it left of the object on the bottom

Fig. 1. The goal of my work is to control a robot to perform tabletop manipulation tasks via natural language instructions.

II. Grounding Objects and Spatial Relations

Visual grounding of referring expressions has been addressed in robotics by training classifiers to recognize predefined object categories [7, 23]. These approaches restrict themselves to tasks covered by predefined visual concepts and simple language expression templates. I developed solutions that leverage advances in modular networks [30, 1] for joint referential expression comprehension and generation [17]. This allows decomposing free-form language into modular components related to subject appearance, location, and relationship to other objects, flexibly adapting to expressions containing different types of information in an end-to-end fashion. I leveraged our approach to enable a PR2 robot to sequentially compose unconstrained pick-and-place language instructions, as shown in Figure 1. If the referred object cannot be uniquely identified by the grounding model, the system needs to ask for clarification from the human operator. Inspired by recent advances in image caption generation and understanding [11, 29], I proposed incorporating a captioning module to our grounding network that allows the robot to describe each detected object with a natural language description. To generate discriminative sentences, we leveraged a maximal mutual information constraint that encourages the generated expression to describe the target object better than the other objects within the image. We demonstrated that by jointly training the comprehension and language generation modules, they regularize each other and improve their respective performances, showcasing the effectiveness of multitask learning [5]. Exploiting multitask learning and modular networks, our architecture achieved a state-of-the-art performance on the challenging RefCOCO benchmark [14] for referential expression comprehension and generation.

Spatial relations play a crucial role in understanding natural language instructions [8, 24] as objects are often described in relation to others. Modeling spatial relations is a challenging problem [18], as natural language placement instructions do not uniquely identify a location in a scene. I advocate to model such spatial relations using distributions to capture the inherent ambiguity. To this end, I proposed a novel approach that combines referring expression comprehension with the grounding of spatial relations to enable complex object placement commands such as "place the ball inside the left box" without the need of pixelwise ground-truth data [19]. Concretely, in this work I proposed a convolutional network for estimating pixelwise object placement probabilities for a set of spatial relations from a single input image. We addressed the problem of the unavailability of ground-truth pixelwise annotations of spatial relations from the perspective of auxiliary learning. Though classifying two objects into a spatial relation does not carry any information on the best placement location to reproduce a relation, inserting objects at different locations in the image would allow to infer a distribution over relations. Most commonly, "pasting" objects in an image requires access to 3D models and silhouettes and creates subtle pixel artifacts that lead to noticeably different features and to the training erroneously focusing on these discrepancies [6]. To this end, our approach receives the learning signal by classifying hallucinated scene representations as an auxiliary task. Concretely, deep features of objects are implanted into a pretrained auxiliary classifier to compute a posterior class probability over spatial relations. By rearranging deep features, we can reason over what relation would most likely be formed if we placed an object at the given location without modifying the input image. Unlike previous approaches that considered only grounding fetching instructions [8, 26], combining my two aforementioned methods enabled us to tackle temporally more extended tasks, leading to the first comprehensive system for controlling a PR2 robot to sequentially compose unconstrained pick-and-place language instructions [17].

III. Language-conditioned Policy Learning

Thus far, I have introduced a method for picking-and-placing objects based on language instructions that can solve ambiguities through dialog. However, if we want to command the robot to solve more complex tasks, such as opening a drawer, extending our approach is not trivial. Towards developing generalist robots, it is not only imperative to ground object semantics and spatial relations, but also to be able to ground a diverse repertoire of robot skills. To this end, I advocate defining skills as being continuous instead of discrete [20], endowing the agent of task-agnostic control: the ability to reach any reachable goal state from any current state [12]. In recent work, I have proposed a new open-source simulated benchmark, coined CALVIN, that links human language to robot motor skills, behaviors, and objects in interactive visual environments [21]. In this setting, a single agent must solve complex manipulation tasks by understanding a series of language expressions in a row, e.g., "open the drawer ... pick up the blue block ... push the block into the drawer ... open the sliding door". Furthermore, to evaluate the agents' ability for long-horizon planning, agents in this scenario are expected to be able to perform any combination of subtasks in any order. Our framework has been developed from the ground up to support training, prototyping, and validation of language conditioned policies over a range of four indoor environments. To establish baseline performance levels, we evaluate an approach that uses relabeled imitation learning to distill reusable behaviors into a language-based goal-directed policy [15]. This is the first public benchmark of instruction following, to our knowledge, that combines: natural language conditioning, multimodal high-dimensional inputs, 7-DOF continuous control, and long-horizon robotic object manipulation. While recent advances have been made in language-driven robotics by leveraging end-to-end learning from pixels [10, 27, 15], there is no clear and well-understood process for making various design choices due to underlying variation in setups. In an effort to standardize research, I conducted an extensive study of the most critical challenges in learning language conditioned policies to identify which components matter most [22]. I further identified architectural and algorithmic techniques that improve performance, such as a hierarchical decomposition of the robot control learning, a multimodal transformer encoder, discrete latent plans and a self-supervised contrastive loss that aligns video and language representations. This open-sourced work is currently state-of-the-art on the challenging CALVIN benchmark.

IV. Future Work

The overall goal of my work is to develop generalist robots that can solve a wide range of everyday tasks from onboard sensors and following natural language instructions. While my current approach achieves state-of-the-art performance in the CALVIN benchmark, training the same approach on a real robot might require a large-scale data collection effort. To this end, I am working on learning language-conditioned visual affordances to improve policy sample-efficiency, by extending a self-supervised approach to obtain affordance labels I recently proposed [3]. Furthermore, to open the door for the future development of agents that can generalize abstract concepts to unseen entities the same way humans do, I plan to take inspiration from foundation models, such as GPT-3 [4] or CLIP [25]. The direction I wish to pursue is to explore how natural language can act as a common grounding across otherwise incompatible embodiments and foundation models. Learning multimodal, multitask foundation models with complementary forms of commonsense will unlock combinatorial generalization of robots to novel behaviors and improve human-robot interaction by generating free-form answers to contextual reasoning questions. Overall, I believe that since Marvin Minsky's summer project, these are the most exciting times to work towards general-purpose robots that can relate human language to their perception and actions.

REFERENCES

[1] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *CVPR*, 2016.

[2] Margaret A. Boden. *Mind as machine: A history of cognitive science*. Oxford University Press, 2008.

[3] Jessica Borja-Diaz, Oier Mees, Gabriel Kalweit, Lukas Hermann, Joschka Boedecker, and Wolfram Burgard. Affordance learning from play for sample-efficient policy learning. In *ICRA*, 2022.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020.

[5] Rich Caruana. Multitask learning. *Machine learning*, 1997.

[6] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *ICCV*, 2017.

[7] Sergio Guadarrama, Lorenzo Riano, Dave Golland, Daniel Go, Yangqing Jia, Dan Klein, Pieter Abbeel, and Trevor Darrell. Grounding spatial relations for human-robot interaction. In *IROS*, 2013.

[8] Jun Hatori, Yuta Kikuchi, Sosuke Kobayashi, Kuniyuki Takahashi, Yuta Tsuboi, Yuya Unno, Wilson Ko, and Jethro Tan. Interactively picking real-world objects with unconstrained spoken language instructions. In *ICRA*, 2018.

[9] Karl Moritz Hermann, Felix Hill, Simon Green, Fumin Wang, Ryan Faulkner, Hubert Soyer, David Szepesvari, Wojciech Marian Czarnecki, Max Jaderberg, Denis Teplyashin, et al. Grounded language learning in a simulated 3d world. *arXiv preprint arXiv:1706.06551*, 2017.

[10] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. BC-0: Zero-Shot Task Generalization with Robotic Imitation Learning. In *CoRL*, 2021.

[11] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, 2016.

[12] Leslie Pack Kaelbling. Learning to achieve goals. In *IJCAI*, pages 1094–1099, 1993.

[13] Dmitry Kalashnikov, Jacob Varley, Yevgen Chebotar, Benjamin Swanson, Rico Jonschkowski, Chelsea Finn, Sergey Levine, and Karol Hausman. MT-Opt: Continuous Multi-Task Robotic Reinforcement Learning at Scale. *arXiv preprint arXiv:2104.08212*, 2021.

[14] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014.

[15] Corey Lynch and Pierre Sermanet. Language Conditioned Imitation Learning Over Unstructured Data. In *RSS*, 2021.

[16] Corey Lynch, Mohi Khansari, Ted Xiao, Vikash Kumar, Jonathan Tompson, Sergey Levine, and Pierre Sermanet. Learning latent plans from play. In *CoRL*, 2019.

[17] Oier Mees and Wolfram Burgard. Composing pick-and-place tasks by grounding language. In *ISER*, 2021.

[18] Oier Mees, Nichola Abdo, Mladen Mazuran, and Wolfram Burgard. Metric learning for generalizing spatial relations to new objects. In *IROS*, 2017.

[19] Oier Mees, Alp Emek, Johan Vertens, and Wolfram Burgard. Learning object placements for relational instructions by hallucinating scene representations. In *ICRA*, 2020.

[20] Oier Mees, Markus Merklinger, Gabriel Kalweit, and Wolfram Burgard. Adversarial skill networks: Unsupervised robot skill learning from videos. In *ICRA*, 2020.

[21] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *arXiv preprint arXiv:2112.03227*, 2021.

[22] Oier Mees, Lukas Hermann, and Wolfram Burgard. What matters in language conditioned robotic imitation learning. *arXiv preprint arXiv:2204.06252*, 2022.

[23] Dejan Pangercic, Benjamin Pitzer, Moritz Tenorth, and Michael Beetz. Semantic object maps for robotic housework-representation, acquisition and use. In *IROS*, 2012.

[24] Rohan Paul, Jacob Arkin, Nicholas Roy, and Thomas M Howard. Efficient grounding of abstract spatial concepts for natural language interaction with robot manipulators. In *RSS*, 2016.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[26] Mohit Shridhar and David Hsu. Interactive Visual Grounding of Referring Expressions for Human-Robot Interaction. In *RSS*, 2018.

[27] DeepMind Interactive Agents Team, Josh Abramson, Arun Ahuja, Arthur Brussee, Federico Carnevale, Mary Cassin, Felix Fischer, Petko Georgiev, Alex Goldin, Tim Harley, et al. Creating Multimodal Interactive Agents with Imitation and Self-Supervised Learning. *arXiv preprint arXiv:2112.03763*, 2021.

[28] Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:25–55, 2020.

[29] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. A joint speaker-listener-reinforcer model for referring expressions. In *CVPR*, 2017.

[30] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, 2018.