Vision and Language Navigation in Continuous Environments:An Improvement from Generative Pre-trained Models

Yijie XU, He ZHANG, Jinhui YE
The Hong Kong University of Science and Technology (Guangzhou)
No. 1, Duxue Rd., Nansha, Guangzhou, China
{yxu409, hzhang757, jye624}@connect.hkust-gz.edu.cn

In this proposal, we aim to explore the complex **V**ision and **L**anguage **N**avigation in **C**ontinuous **E**nvironments (**VLN-CE**) problem, which involves interpreting first-person RGB-D visuals and textual instructions to navigate a sequence of goals in continuous environments using discrete actions. The problem's significance arises from its complexity, multi-modal nature, and its integration of Computer Vision and Natural Language Processing techniques, making it a relevant and challenging research area. It also encompasses perception and decision-making aspects within reinforcement learning and simulation-to-reality transfer. To address **VLN-CE**, we propose to utilize the **R**oom-Across-**R**oom (**RxR**) dataset [2], based on the Matterport3D collection of indoor scenarios, featuring 16.5K navigation paths, multi-lingual annotations, 126K instructions, and time-aligned pose traces for seamless operation.

Owing to space limitations, an exhaustive elucidation of the baseline method has been relegated to the original publication, which may be perused in [2] for a more comprehensive understanding. In the present section, we provide a concise overview of the input and output formats employed by the chosen baseline method. As demonstrated in Figure 1, the agent is furnished with visual information from both RGB and depth cameras, in conjunction with a natural language directive delineating the pathway to the desired room, for each action executed. Utilizing this information, the agent selects the appropriate action for the current context (e.g., turn left, turn right, move forward, or stop), consequently reaching a new location and commencing a fresh decision-making process, which persists until the output action is "stop."

As depicted in Figure 1, the instructions may exhibit a high degree of intricacy, potentially posing a challenge for the machine's comprehension and rendering them impractical in certain scenarios. Within the scope of this study, our foremost objective is to replicate the model delineated in the cited paper, given the considerable complexity of the task at hand. As an ambitious goal, we aim to leverage pre-trained large-scale models to reconstruct the linguistic guidance, thereby rendering the instructions more pragmatic and comprehensible.

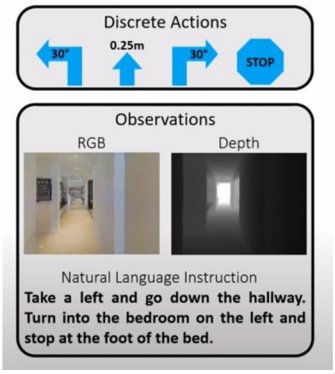

Figure 1. Input format of dataset

Our primary objective is to replicate the existing methodology on our designated apparatus. Subsequently, we propose integrating a "Planner Model," designed to serve as an intermediary between human directives and the navigation robot. As previously discussed, instructions in the dataset tend to be excessively intricate, often proving challenging for humans to comprehend promptly. We hypothesize that our Planner Model will have the capacity to interpret general directives and transform them into more explicit instructions, thereby simplifying decision-making for the robot.

Drawing inspiration from [1], we posit that the implementation of a Large Language Model, in conjunction with the integration of pertinent information, can facilitate the development of our planner model. We anticipate that incorporating additional information into the planner model will yield improvements in success rate (SR) and success weighted by inverse path length (SPL), while simultaneously reducing training cost (TC). Consequently, our evaluation metrics consist of SR, SPL, and TC, which collectively embody our goals for enhancing the performance of the existing model in addressing this problem.

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can and not as i say: Grounding language in robotic affordances. In *arXiv preprint arXiv:2204.01691*, 2022. 1

[2] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-Across-Room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Conference on Empirical Methods for Natural Language Processing (EMNLP)*, 2020. 1