Human Instruction Following: Graph Neural Network Guided Object Navigation

Hongyi Chen[1] Letian Wang[2] Yuhang Yao[2] Ye Zhao[1] Patricio A. Vela[1]
[1]Georgia Institute of Technology [2]Carnegie Mellon University
[hchen657, ye.zhao, pvela]@gatech.edu, [letianw, yuhang]@andrew.cmu.edu

Abstract

Home-assistant robots (e.g., mobile manipulator) following human instructions is a long-standing topic of research whose main challenge comes from the interpretation of diverse instructions and dynamically-changing environments. This paper proposes a hybrid planner for parsing human instruction and task planning, and a graph-based object navigation method to search unknown objects by exploiting a partially known semantic map. We present preliminary evaluations of the proposed methods on human instruction parsing and object-to-object link prediction, and demonstrate their effectiveness in human instruction following tasks.

1. Introduction

Home-assistant robots share living and working spaces with humans, and assist them by interpreting human instructions and performing corresponding tasks. Early symbolic works exploited the syntactic structure of language to understand human instructions and statically generated a sequence of actions [12] [4] [6]. However, this type of approach fails to interpret the diverse human instructions nor captures semantic meaning in incomplete sentences. To avoid processing natural language based on engineered symbolic structure, the recent deep learning methods can automatically learn the linguistic features via deep neural networks [3] [1]. However, it is difficult to plan a sequence of actions through end-to-end training on neural network. To leverage the strengths of symbolic and learning based approaches, we adopt a hybrid approach which combines the deep learning methods for goal learning and the symbolic approaches for task planning.

With the planned action sequence from Planning Domain Definition Language (PDDL), a robot will reason where the needed objects locate and then find them out. In previous object navigation tasks, the robot searched for an instance of an object category in an unseen environment without prior knowledge [2] [14] [10]. But real home-assistant robots are often equipped with certain level of semantic knowledge about the environment, regions, and objects [5] [7]. In our experiments, we assume that the robot is equipped with a partially known semantic map, which contains the information of some objects' positions but is unaware of others due to the environment changes. To solve the problem, we build a graph to represent the relationship among objects, and use a graph neural network to reason the possible positions of the unknown target object and guide the search process. Once the target object is found, the robot will execute the planned action sequence on the object.

2. Proposed Approach

Goal Learning and Symbolic Task Planning: Given a natural language sentence L composed of K words, we first pass it into a linguistic encoder to generate an embedding vector q. The classifier then parses the embedding vector q to predict the action a, subject s, and object o. For symbolic task planning, we employ the Planning Domain Definition Language (PDDL), a widely used symbolic planning language. With a list of pre-defined objects and their corresponding predicates (such as dirty, graspable), a domain consists of primitive actions and corresponding effects. Besides, the planning problem is to transfer from the initial state to the desired goal state, where the initial state is formed with a list of objects with corresponding predicates and the goal state is estimated from the classifier. From the domain and problem specification, a PDDL planner produces a sequence of primitive actions to reach the goal state when executed, a simple example is shown in green part of Fig. 1.

Semantic Graph Neural Network: To improve the efficiency of object navigation, we exploit the fact that the target unknown objects are usually located closely with some known objects. For example, the remote is usually placed close to the TV. To this end, we model the object-to-object relationship in the form of graph representation and use Graph Attention Networks (GAT) [13] to compute

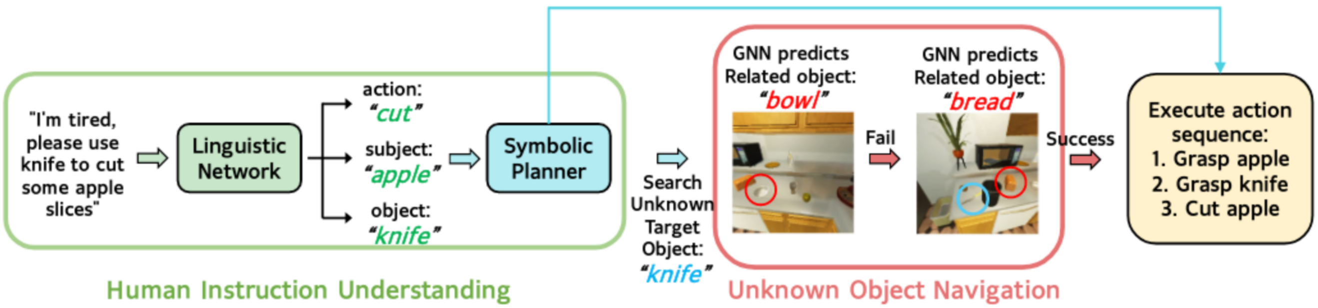

Figure 1. Illustration of symbolic goal learning and searching for unknown object "knife" with graph neural network (GNN).

relational features on the graph. We denote our graph by $G = (V, E)$, where V and E denote the nodes and the edges between nodes, respectively. Specifically, each node $v \in V$ denotes an object category, and each edge $e \in E$ denotes a relationship between a pair of object categories. The input to each node v is a feature vector x_v which includes object category and attribute information. Compared to other traditional machine learning algorithms that find related objects like clustering, graph neural network has greater generalization and extensibility: it can not only find out related objects using edge prediction but also encode spatial relationships between different object categories.

More specifically, we use the Visual Genome dataset [9], where each image is annotated with objects and the relationships between objects, to build the graph. We count the occurrence of object-to-object relationships in the dataset and connect two nodes when the occurrence frequency of any relationship is more than three. We build multiple graphs from the dataset by constructing a new graph every 20,000 relationships and each graph is represented as a binary adjacency matrix A. The training task is the link prediction by using node embeddings $h_v = GAT(x_v, A)$, which is the hidden layer output after GAT information propagation and aggregation. After we get the node embeddings, we use another neural network to predict the link probability, $\hat{y}_{uv} = Predictor(h_u, h_v)$. During testing, the robot predicts the relationships between the unknown object and other known objects, and search the place of known object with highest $\hat{y}_{uv}$. If not found, we remove that known object from the graph and repeat the process.

3. Experimental Results

Goal Learning: For learning symbolic goal representation from language, we adopt the Symbolic Goal Learning Dataset[1], and select 8163 explicit human instructions[2] which cover 33 objects and 4 daily activities, i.e., cutting, cooking, cleaning, and pick-and-place. We adopt the Multi Modal Framework (MMF) [11] and only train the language encoder with human instructions and corresponding ground-truth goal states. Our goal learning network achieves 100% prediction accuracy in 1024 unseen explicit human instructions, and the PDDL planner works perfectly once the goal state is correctly learned.

Graph Neural Network Link Prediction: We obtain 115 graphs from the Visual Genome dataset including 108 different object categories in AI2THOR [8]. The GAT model is trained for 500 epochs and the experiments are repeated 5 times. The averaged link prediction accuracy is 89.66%, 88.28%, 87.58% in training, validation, and test sets, respectively. This result demonstrates that our GAT can predict the related objects with high accuracy and help to guide the unknown object navigation.

Human Instruction Following: We adopt MaskRCNN as object detector and test the pipeline in 20 different scenes, including kitchens, bedrooms, apartments and living rooms in AI2THOR. If the robot correctly predicts the goal state and finds out the unknown object, the human instruction is treated as completed. The instruction success rate and object navigation search efficiency (success weighted by path length, SPL) are summarized in Tab. 1. There are two failures cases: Firstly, the detector fails to detect small objects like butter knife and saltshaker. Secondly, the robot sometimes needs to crouch to find the book in the lower shelf or open the fridge to find the food.

Table 1. Experimental results of the human instruction following.

	cook	clean	cut	pick-and-place
Success rate	97.76%	88.23%	97.14%	86.98%
SPL	0.68	0.59	0.64	0.59

4. Conclusion and Future Work

In this letter, we developed and demonstrated that the hybrid planner performs perfectly with explicit instructions and GAT could guide the unknown object navigation which leads to high success rate for human instruction following tasks. Our future work will focus on two aspects: (1) encode the spatial relationship into the graph and estimate the specific spatial region of unknown objects. (2) implementation of detailed robot execution and manipulation.

[1] https://smartech.gatech.edu/handle/1853/66305

[2] Explicit human instruction contains the subject and object inside which would make the training easier.

References

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

[2] Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33:4247–4258, 2020. 1

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[4] Cameron Finucane, Gangyuan Jing, and Hadas Kress-Gazit. Ltlmop: Experimenting with language, temporal logic and robot control. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1988–1993, 2010. 1

[5] Cipriano Galindo, Juan-Antonio Fernández-Madrigal, Javier González, and Alessandro Saffiotti. Robot task planning using semantic maps. *Robotics and autonomous systems*, 56(11):955–966, 2008. 1

[6] Kai-yuh Hsiao, Stefanie Tellex, Soroush Vosoughi, Rony Kubat, and Deb Roy. Object schemas for grounding language in a responsive robot. *Connect. Sci*, 20(4):253–276, 2008. 1

[7] Zhe Hu, Jia Pan, Tingxiang Fan, Ruigang Yang, and Dinesh Manocha. Safe navigation with human instructions in complex scenes. *IEEE Robotics and Automation Letters*, 4(2):753–760, 2019. 1

[8] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017. 2

[9] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. 2

[10] Xiaotian Liu and Christian Muise. A neural-symbolic approach for object navigation. In *CVPR Embodied-AI Workshop*, 2021. 1

[11] Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Mmf: A multimodal framework for vision and language research, 2020. 2

[12] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R. Walter, Ashis Gopal Banerjee, Seth Teller, and Nicholas Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, AAAI'11, page 1507–1514. AAAI Press, 2011. 1

[13] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *stat*, 1050:20, 2017. 1

[14] Wei Yang, Xiaolong Wang, Ali Farhadi, Abhinav Gupta, and Roozbeh Mottaghi. Visual semantic navigation using scene priors. *arXiv preprint arXiv:1810.06543*, 2018. 1