Human Instruction Following: Graph Neural Network Guided Object Navigation

Anonymous CVPR submission

Paper ID *****

Abstract

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

1. Introduction

Home-assistant robots share the living and working spaces with human, and assist them by interpreting human instructions and performing their corresponding tasks. Early works employ semantic parsing and task planning to first map natural language into certain representations and then generate a sequence of actions [7] [14] [5]. However, semantic parsing relies on syntactic structure of natural language and fails to capture the abstract or vague semantic meaning [15] [1] [12]. Recently, the rise of deep learning methods provided a means to avoid processing natural language based on engineered symbolic representations, by automatically learning linguistic features via deep neural networks [4] [2]. But end-to-end network design makes the training harder and slower, and suffers from performance drops in testing stage. To leverage strengths of symbolic and learning based approaches to compensate limitation of each other, we adopt a hybrid approach which combines the deep learning methods for goal learning and the symbolic approaches for task planning.

Besides, the objects needed to accomplish the planned action sequence may not exist in robot's view. For instance, given the instruction "cut the apple slices", the robot needs to reason and find out where are the apple and knife. In previous object navigation tasks, the robot search for an instance of an object category in an unseen environment without prior knowledge [3] [17] [11]. But real home-assistant robots are usually equipped with some level of semantic knowledge about the environment, regions and objects [6] [8]. In our experiments, we assume the robot is equipped with a partially known semantic map, which contains some objects' positions information but misses others due to environment changing. Then, to solve the problem, we build a graph to represent the relationship between objects, and use graph neural network to reason the possible positions of the unknown object and search until it finds.

2. Proposed Approach

Goal Learning and Symbolic Task Planning: Given a natural language sentence L composed of K words, we first pass it into the linguistic encoder that generate a single embedding vector q which represents the semantic meaning of the entire sentence. Later, using the embedding vector, the classifier predicts the action a, subject s and object o, which form the structure of the goal state. 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 learned above. 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: With the planned action sequence from PDDL, the next step is to find out the objects and then execute the actions. 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

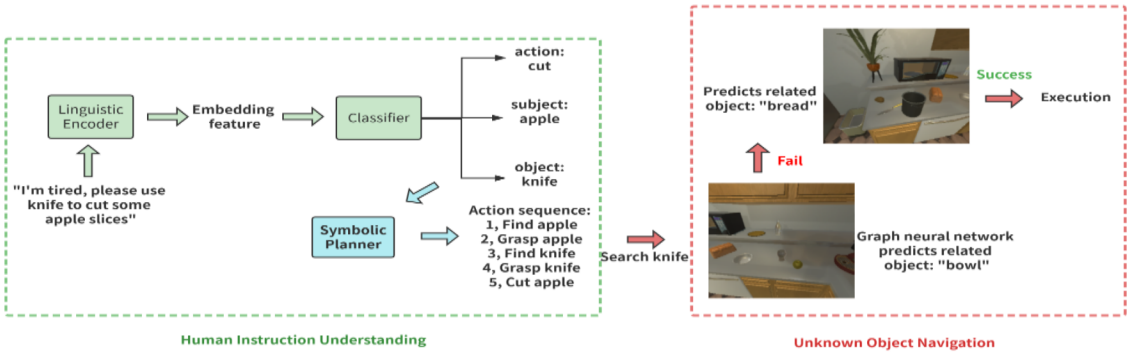

Figure 1. Illustration of symbolic goal learning and searching for unknown object "knife" with graph neural network.

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) [16] to compute 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 attributes 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.

In detail, we use the Visual Genome dataset [10], where each image is annotated with objects, attributes and the relationships between objects, to build the graph. We count the occurrence of object-to-object relationships in the Visual Genome 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 keeps detecting the object-to-object relationships during object navigation, incrementally provide these information for link prediction and search the places of known objects with high $\hat{y}_{uv}$ values until we find.

3. Experimental Results

Goal Learning: For learning symbolic goal representation from language, we adopt the Symbolic Goal Learning Dataset[1], and pick out 8163 explicit human instructions[2] which cover 33 objects and 4 daily activities which are cutting, cooking, cleaning and pick-and-place. We adopt the MMF [13] framework 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 get 115 graphs from the Visual Genome dataset including 108 different object categories in AI2THOR [9]. GAT model is trained for 500 epochs and the experiments are repeated for 5 times. The averaged link prediction accuracy is 89.66%, 88.28%, 87.58% in train, validation and test set 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 our object detector and test the whole pipeline in 40 different scenes including kitchens, bedrooms and living rooms in AI2THOR. If the robot correctly predict the goal state and find out the unknown object, the human instruction is regarded as completed. During the navigation, the robot will always search the area of the most related object and then predict the next most related one if it doesn't find. The overall success rate is 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. Success rate of the human instruction following.

	cook	clean	cut	pick-and-place
Success rate	80%	100%	90%	80%

[1] https://smartech.gatech.edu/handle/1853/66305
[2] explicit human instruction contains the subject and object inside.

4. Conclusion and Future Work

In this paper, a solution, consisting of a hybrid planner to understand human instruction and a semantic graph network to guide the object navigation, is developed for home assistant robots. The results demonstrate that the hybrid planner module performs perfectly in explicit instruction dataset, and the graph network guides the unknown object navigation and leads to high success rate. 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 detail robot execution and manipulation.

References

[1] Alexandre Antunes, Lorenzo Jamone, Giovanni Saponaro, Alexandre Bernardino, and Rodrigo Ventura. From human instructions to robot actions: Formulation of goals, affordances and probabilistic planning. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5449–5454. IEEE, 2016. 1

[2] 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

[3] 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

[4] 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

[5] 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

[6] 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

[7] Sergio Guadarrama, Lorenzo Riano, Dave Golland, Daniel Go¨hring, Yangqing Jia, Dan Klein, Pieter Abbeel, and Trevor Darrell. Grounding spatial relations for human-robot interaction. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1640–1647, 2013. 1

[8] 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

[9] 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

[10] 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

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

[12] Pradip Pramanick, Chayan Sarkar, and Indrajit Bhattacharya. Your instruction may be crisp, but not clear to me! In *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1–8. IEEE, 2019. 1

[13] 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

[14] 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

[15] Moritz Tenorth and Michael Beetz. Knowrob: A knowledge processing infrastructure for cognition-enabled robots. *The International Journal of Robotics Research*, 32(5):566–590, 2013. 1

[16] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *stat*, 1050:20, 2017. 2

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