Embodied Question Answering

Wenjie Niu

June 6,2018

Abstract

*[5]We present a new AI task- **Embodied Questi-n Answering** (EmbodiedQA) where an agent is spawned at a random location in a 3D environment and asked a question ('What color is the car?'). In order to answer, the agent must first intelligently navigate to explore the environment, gather necessary visual information through first-person (egocentric) vision, and then answer the question ('orange').*

EmbodiedQA requires a range of AI skills language understanding, visual recognition, active perception, goaldriven navigation, commonsense reasoning, long-term memory, and grounding language into actions. In this work, we develop a dataset of questions and answers in House3D environments [6], evaluation metrics, and a hierarchical model trained with imitation and reinforcement learning.

1. Introduction

Our long-term goal is to build intelligent agents that can perceive their environment (through vision, audition, or other sensors), communicate (i.e., hold a natural language dialog grounded in the environment), and act (e.g. aid humans by executing API calls or commands in a virtual or embodied environment). In addition to being a fundamental scientific goal in artificial intelligence (AI), even a small advance towards such intelligent systems can fundamentally change our lives from assistive dialog agents for the visually impaired, to natural-language interaction with selfdriving cars, in-home robots, and personal assistants.

As a step towards goal-driven agents that can perceive, communicate, and execute actions, we present a new AI task-**Embodied Question Answering**(EmbodiedQA) along with a dataset of questions in virtual environments, evaluation metrics, and a deep reinforcement learning (RL) model. Concretely, the EmbodiedQA task is illustrated in Fig. 1 an agent is spawned at a random location in an environment (a house or building) and asked a question (e.g. 'What color is the car?'). The agent perceives its environment through first-person egocentric vision and can perform a few atomic actions (move-forward, turn, strafe, etc.). The goal of the agent is to intelligently navigate the environment and gather visual information necessary for answering the question.

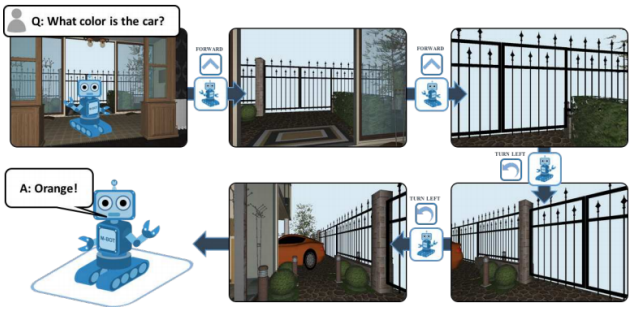

Figure 1. Embodied Question Answering EmbodiedQA tasks agents with navigating rich 3D environments in order to answer questions. These agents must jointly learn language understanding, visual reasoning, and goal-driven navigation to succeed.

2. Interactive Environments

There are several interactive environments commonly used in the community, ranging from simple 2D grid-worlds (e.g. XWORLD [7]), to 3D game-like environments with limited realism (e.g. DeepMind Lab [3] or Doom [4]), to more complex, realistic environments (e.g. AI2-THOR [8], Matterport3D [1], Stanford 2D-3D-S [2]). While realistic environments provide rich representations of the world, most consist of only a handful of environments due to the high difficulty of their creation. On the other hand, large sets of synthetic environments can be programmatically generated; however, they typically lack realism (either in appearance or arrangement). In this work, we use the House3D [6] environment as it strikes a useful middle-ground between being sufficiently realistic and providing access to a large and diverse set of room layouts and object categories.

References

[1] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Snderhauf, I. Reid, S. Gould, and A. V. D. Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, 2017. 1

[2] I. Armeni, S. Sax, A. R. Zamir, and S. Savarese. Joint 2D-3D-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. 1

[3] C. Beattie, J. Z. Leibo, D. Teplyashin, T. Ward, M. Wainwright, H. Küttler, A. Lefrancq, S. Green, V. Valdés, A. Sadik, J. Schrittwieser, K. Anderson, S. York, M. Cant, A. Cain, A. Bolton, S. Gaffney, H. King, D. Hassabis, S. Legg, and S. Petersen. Deepmind lab. *arXiv preprint arXiv:1612.03801*, 2016. 1

[4] D. S. Chaplot, K. M. Sathyendra, R. K. Pasumarthi, D. Rajagopal, and R. Salakhutdinov. Gated-attention architectures for task-oriented language grounding. In *AAAI*, 2017. 1

[5] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra. Embodied question answering. In *CVPR*, 2018. 1

[6] Y. Wu, Y. Wu, G. Gkioxari, and Y. Tian. Building generalizable agents with a realistic and rich 3D environment. *arXiv preprint arXiv:1801.02209*, 2018. 1

[7] H. Yu, H. Zhang, and W. Xu. A deep compositional framework for human-like language acquisition in virtual environment. *arXiv preprint arXiv:1703.09831*, 2017. 1

[8] Y. Zhu, D. Gordon, E. Kolve, D. Fox, F. F. Li, A. Gupta, R. Mottaghi, and A. Farhadi. Visual semantic planning using deep successor representations. In *ICCV*, 2017. 1