Learning by Asking Questions

Chaonan Song

Jun 18, 2018

Abstract

Today reads the article Embrided Question Answering by Dr. Abhishek. They propose a new AI task - a specific question answer (EmbodiedQA) - An intermediary generates and asks a question at a random position in the 3D environment ('What colour is the car?'). To answer, the agent must first navigate intelligently to explore the environment, collect the necessary visual information through the first-person visual, and then answer the question ('orange'). EmbassyQA requires a series of AI skills - language understanding, visual recognition, active awareness, goal-driven navigation, common sense reasoning, long-term memory, and language integration. In this work, they developed questions and answers data sets in the House3D environment, evaluation indicators, and hierarchical models trained by imitative and reinforcement learning.

1. Introduction

Abhishek team's long-term goal is to build smart agents that can sense their environment, communicate, and act. In order to move toward goal-driven agents who can perceive, communicate, and act, Dr. Abhishek present a new AI task-¨C Embodied Question Answering(EmbodiedQA), and problem data sets in the virtual environment, evaluation indicators, and deep reinforcement learning (RL)model. Specifically, Figure **??** shows the task of EmbassyQA - an agent who generates a random position in an environment and asks a question (for example, 'What color is the car?'). Agents perceive their environment through a first-person perspective and can perform some subtle movements (forward, turn, turn, *etc*.). The agent's goal is to intelligently browse the environment and collect the visual information needed to answer questions. EmbodiedQA is a challenging task that contains several basic issues as sub-tasks. Obviously, the agent must understand the language (what is the question?) and the vision (what does the "car" look like?), but it must also learn:

Active Perception: Agents may be generated anywhere in the environment and may not immediately "see" pixels containing answers to visual questions. Therefore, the agent must successfully control the pixels it perceives. Agents must learn to map their visual input to the correct behavior based on their perception of the world, potential physical limitations, and understanding of the problem.

Commonsense Reasoning: The agency does not provide a floor plan or an environmental map and must navigate independently from the self-center view. Therefore, it must learn common feelings, similar to how humans drive in unfamiliar houses.

Language Grounding: A common drawback of modern visual and linguistic models is that they lack a foundation - these models often fail to associate the entities in the text with the corresponding image pixels, but rely on data set deviations, even when they notice unrelated areas. Respond intelligently [4, 5]. In EmbitureQA,Dr. Abhishek took a fundamental, goal-oriented view - our agents based their problem not on a pixel-by-pixel basis but on a series of operations.

Contributions. The contribution of Dr. Abhishek's article is shown in Table 1.

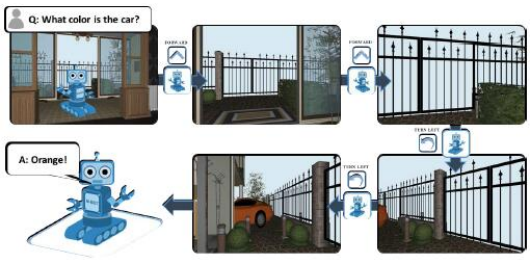

Figure 1. Embodied Question Answering ¨C EmbodiedQA¨C tasks agents with navigating rich 3D environments in order to answer questions. These agents must jointly learn language understanding, visual reasoning, and goal-driven navigation to succeed.

2. Related Work

VQA: Vision + Language. Like Embassy QA, image and video question answering tasks Antol *et al.* [1] need to reason about natural language issues raised by visual content. The key difference is the lack of control - these tasks

Contribution 1	Presented a new AI task: EmbodiedQA
Contribution 2	Introduced a hierarchical navigation module
Contribution 3	Use an imitation learning initialization agent
Contribution 4	Evaluating agents in House3D
Contribution 5	Introduced the EQA data set

Table 1. Contribution of this paper.

give the respondent a fixed view of the circumstances in which the agent must answer the question, never allowing the agent to perceive it. Instead, EmbodiedQA agents control their trajectories.

Visual Navigation: Vision + Action. The long-term research of navigation problems in visual perception-based environments has been conducted in vision and robotics. Classical technology divides navigation into two distinct phases - drawing and planning. The recent deep RL development proposes a fusion architecture that visually observes navigation actions directly from an ego-center Brahmbhatt and Hays [2]. Abhishek team model the agent as a similar pixel to action navigator. The key difference in EmbitureQA is how to set goals. Embodying QA's goal of assigning agents through language is inherently combinative and provides different strategies for each task (problem).

Situated Language Learning: Language + Action. Inspired by Winograd's classic work, some of the recent works assign tasks to agents by placing them in a simple, globally aware environment and assigning them to natural language-specific goals. Of course, one of the key differences in EmbitureQA is that the visual perception - the environment can only be partially observed, that the agent can not access the floor plan, object tags, attributes, etc., and must be purely from the first person visual sense. **Embodiment: Vision + Language + Action. EmbassyQA** The closest work is to extend the paradigm of language learning in place to the perception that agents' perceptions are local, purely visual, and change based on their behavior - Abhishek team call it embodied language learning. In contrast, our Embodied QA environment consists of multi-room dwellings (¡°8 per household¡±), which are densely populated with various objects (54 unique objects per house).

Interactive Environments. There are many common interactive environments in the community, ranging from simple 2D grid worlds to 3D game-like environments with limited realism (*eg*, Doom [3]). In this work, Abhishek team use the House3D environment because it achieves a useful middle ground between being realistic enough and providing a large number of different room layouts and object class sets.

Hierarchical Agents. Abhishek team modeled our EmbodiedQA agent as a deep-seated network, decomposing the overall control problem so that higher-level planners call lower-level controls to issue primitive operations. Their model is also inspired by Graves's work on adaptive computing time.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence, and D. Parikh. VQA: Visual question answering. In *CVPR*, 2015.

[2] S. Brahmbhatt and J. Hays. Deepnav: Learning to navigate large cities. In *CVPR*, 2017.

[3] D. S. Chaplot, K. M. Sathyendra, R. K. Pasumarthi, D. Rajagopal, and R. Salakhutdinov. Gated-attention architectures for task-oriented language grounding. In *AAAI*, 2018.

[4] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90-100, 2017.

[5] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.