Embodied Question Answering

Xuewen Yang

August 11 2018

Abstract

The author present a new AI task C Embodied Question Answering (EmbodiedQA) C 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 C language understanding, visual recognition, active perception, goaldriven navigation, commonsense reasoning, long-term memory, and grounding language into actions. In this work, the author develop a dataset of questions and answers in House3D environments, evaluation metrics, and a hierarchical model trained with imitation and reinforcement learning.

1. EQA Dataset: Questions In Environments

1.1. House3D: Simulated 3D Environments

The author instantiate EmbodiedQA in House3D, a recently introduced rich, simulated environment based on 3D indoor scenes from the SUNCG dataset [2]. Concretely, SUNCG consists of synthetic 3D scenes with realistic room and furniture layouts, manually designed and crowdsourced using an online interior design interface. Each scene was further verified as realistic by majority vote of 3 human annotators. In total, SUNCG contains over 45k environments with 49k valid floors, 404k rooms containing 5 million object instances of 2644 unique objects from 80 different categories. House3D converts SUNCG from a static 3D dataset to a set of simulated environments, where an agent (approximated as a 1 meter high cylinder) may navigate under simple physical constraints (not being able to pass through walls/objects).

1.2. QuestionAnswer Generation

The author draw inspiration from the CLEVR [1] dataset, and programmatically generate a dataset (EQA) of grounded questions and answers. This gives the ability to carefully control the distribution of question-types and answers in the dataset, and deter algorithms from exploiting dataset bias.

2. Imitation Learning and Reward Shaping

The navigation module is trained to mimic the shortest path in a teacher-forcing setting - i.e., given the navigation history encoded in the hidden state of the planner, question encoding, and the current frame, the model is trained to predict the action that would keep it on the shortest path. The author use a cross-entropy loss and train the model for 15 epochs. They find that even in this imitation learning case, it is essential to train the model under a distance-based curriculum. In the first epoch, the author backtrack 10 steps from the target along the shortest path and initialize the agent at this point with full history of the trajectory from spawned location. They step back an additional 10 steps at each successive epoch. They train for 15 epochs total with batch size ranging from 5 to 20 (depending on path length due to memory limitations). The answering module is trained to predict the correct answer from the question and shortest path navigation frames, with cross-entropy loss for 50 epochs and batch size of 20.

References

[1] J. Johnson, B. Hariharan, V. D. M. Laurens, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2016. 1

[2] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser. Semantic scene completion from a single depth image. In *CVPR*, 2016. 1