Review11: Embodied Question Answering in Photorealistic Environments with Point Cloud Perception

Vivswan Shitole

CS539 Embodied AI

1 Summary

This paper studies the Visual Question Answering (VQA) task, but with three variations to the traditional VQA setting: (1) Instead of the ”internet vision” style problem, we have an ”embodied” setting where an agent must navigate in the environment in order to be able to answer the given question. The authors generalize the Embodied Question Answering task (originally proposed in synthetic SUNCG scenes) to photorealistic 3D reconstructions from Matterport 3D, as an effort to be closer to reality. For the question-answer sets, authors introduce MP3D-EQA dataset, consisting of 1136 questions and answers grounded in 83 environments. (2) Instead of using 2D RGB images, the authors explore 3D point cloud representations of the environment as input to the agents. The agents are developed as end-to-end trainable models with point cloud perception, going from raw 3D point clouds to goal-driven navigation policies. (3) Authors introduce a novel loss weighing scheme called ”Inflection Weighting” - balancing the contributions to the cross-entropy loss between inflections (states where the ground truth action differs from the previous one) and non-inflections. The authors claim it to be an effective technique for training recurrent navigation models using behaviour cloning with a shortest path expert.

More concretely, the authors examines both RGB and RGB-D perception in Matterport3D environments. For RGB, they take renders from the mesh reconstructions and for point clouds, they operate directly on the aligned point clouds. TO render the 2.5 RGB-D frames, they first construct a global point cloud from all of the panoramas provided. Then the agent’s current position, camera parameters and the mesh reconstruction are used to determine which points are within its view. The questions are programatically generated based on Matterport 3D annotations using 3 templates: location questions, color questions and color-room questions. Instead of unique rooms and unique objects, the dataset has only unique room-object pairs, providing the navigator significantly less information. The agents are tested on entirely new homes, testing both semantic and perceptual generalization. The point cloud representations are learnt using PointNet++, which is a 3D network architecture that alternates between spatial clustering and feature summarization - resulting in a hierarchy of increasingly course point clusters with associated feature representations summarizing their members. As an agent navigates an environment, the number of points it perceives can vary. This problem is addressed by discretizing the possible number of points in any given point cloud into 5 equal sized bins, resulting in a 32-d sparsity embedding that is concatenated with the PointNet++ output vector. The agent is separately pretrained on 3 tasks: semantic segmentation, color autoencoding and structural autoencoding. The question encodings are learnt using two-layer LSTMs with 128-d hidden states. The authors experiment over 3 types of question-answering models (Question-only, Attention-based, Spatial Attention based), two navigation models (Forward-only and Random), two navigation architectures (Reactive and Memory-based), four perception variations (Blind, PC, RGB and PC+RGB) and two language variations (w/o question). A dataset of expert trajectories is created to train for navigation. The stopping criteria is defined by the best view of the target, determined by the IOU of segmentation mask of target object with a pre-determined bounding box. The training is conducted using behaviour cloning with inflection weighting. Inflection weighting may be viewed as a generalization of class-balanced loss methods that are commonly used in supervised learning under heavily imbalanced class distributions.

Obtained results show that forward-only and random baselines are strong baselines for the proposed task, Inflection weighting improves navigation, particularly for memory-based models, memory-based navigators perform better than their reactive counterparts, vision helps gaze detection metrics such as IOU criteria for stopping and question-answering accuracy, but hurts for distance-based navigation metrics, questions are of little help for navigation agents trained using behaviour cloning, point clouds provide a richer signal for obstacle avoidance, while RGB provides richer semantic information. PC+RGB provides best of both worlds.

2 Strengths

The paper studies a novel setting of Embodied VQA which is different from the traditional VQA setting in three important ways (embodied setting with photorealism, 3D point cloud representation input and inflection weighting in behaviour cloning training). The paper presents a large-scale exhaustive evaluation of design decisions, training a total of 16 navigation models (2 architectures, 2 language variations, and 4 perception variations), 3 visual question answering models, and 2 perception models - ablating the effects of perception, memory, and goal specification. The paper presents a good explanation of PointNet++ (instead of just citing it) for processing 3D point cloud input, which is a crucial section of the paper. Experiments and results are provided in quite detail along with the appendix.

3 Weaknesses

There are many under-explained nuances in the results of the experiments that do not align with the inferred conclusion headlines. Strong performance of naive baselines such as Forward-only and Random question the complexity of the navigation task. Its not clear why improved navigation by virtue of inflection weighting does not translate to improved question-answering accuracy. The effect of ablation in vision inputs are not clear - they mostly seem to harm performance which is counter-intuitive. Moreover the performance depreciation trends are different for reactive and memory-based agents (Figure-6 column-1). The models don't seem to properly leverage information from questions. Its not clear how the bounding box used for stopping criteria was pre-determined. Its not clear why the question encoding for navigation and question-answering are learned separately. All tasks have the start and goal positions on the same floor. Its not clear why inflection weighting is claimed to be effective specifically for recurrent models to be trained on long trajectories.

4 Reflections

The paper is an important contribution in embodied question answering since it introduces three novelties over the traditional VQA setting. Many findings are non-intuitive and open avenues for future research. Some practices such as inflection weighting and processing 3D point clouds over 2D RGB inputs can be adopted in other areas of AI research in addition to Embodied VQA. It will be interesting to perform the ablation of given navigation metrics over more complex navigation tasks, where forward-only is not a strong baseline.

5 Most Interesting Thought

I found PointNet++ to be very interesting architecture for its use in processing 3D point clouds and its analogy with 2D convolution and pooling operations. As for inflection weighting, I have an experience working with class-balanced loss methods in supervised learning with imbalanced classes, and the method of adding balancing weights did not seem to generalize well across different datasets.