Review10: IQA: Visual Question Answering in Interactive Environments

Vivswan Shitole

CS539 Embodied AI

1 Summary

The paper has the following two contributions: (a) it introduce novel tasks based on Interactive Question-Answering (IQA), which are similar to question answering tasks but the agent needs to interact with the environment in order to be able to answer the question. The Interactive Question-Answering dataset (IQUAD v1) is built upon AI2-THOR, a photo-realistic, customizable simulation environment for indoor scenes integrated with the Unity physics engine. IQUAD v1 consists of over 75000multiple choice questions for 3 different question types (existence, counting and spatial relationship), each question accompanied by scene identifier and a unique arrangement of movable objects in the scene. The wide variety of scene configurations prevent the model from learning simple rules for solving the IQA tasks. (b) Hierarchical Interactive Memory Network (HIMN) which is a novel modular controller architecture corresponding to the behaviour of an agent for solving the IQA task. The HIMN comprises of a "Planner", which is a high level controller that chooses the task to be performed (navigation/ manipulation/ answering) and generates a command for the chosen task. Tasks specified by the planner are executed by a set of low level controllers (Navigator, Manipulator, Detector, Scanner and Answerer) which return the control to the Planner when a task termination state is reached. In addition to the Planner and the low level controllers, the HIMN consists of a spatial and semantic memory used by the agent to maintain a spatial and semantic representation of the environment to solve the IQA task. This memory is implemented using a novel Egocentric Spatial GRU (esGRU). Owing to the factorized set of controllers, HIMN agent can operate at multiple levels of temporal abstraction.

More concretely, each location in the esGRU implemented memory consists of a feature vector encoding object detection probabilities, free space probability (2D occupancy grid), coverage and navigation intent. The esGRU maintains an external global spatial memory represented as a 3D tensor. At each time step, the esGRU swaps in local egocentric copies of this memory into the hidden state of GRU, preforms computation using current inputs and swaps out the resulting hidden state into the global memory at the predetermined location. The low level controllers have read-write access to the memory, while the Planner has read-only access. The Planner is a high level controller that yields commands to invoke one of the low level controllers (Navigator, Manipulator, Detector, Scanner and Answerer), for a total of 32 discrete actions. The Planner consists of a GRU which accepts at each time step the current viewpoint (encoded by a CNN) and the previous action. The output of the GRU is combined with the question embedding (encoded by a LSTM) and an embedding of the nearby semantic spatial memory to predict the policy π and a value v for the current state. The agent receives fixed rewards/penalties based on the correctness of the answer, the time taken to answer, the number of invalid actions attempted and the environment coverage. The Navigator runs A^* search to find the shortest path to the goal specified by the Planner. As the navigator moves through the environment, it uses the esGRU to produce a local (5x5) occupancy grid to yield updated global occupancy estimate, a new shortest path computation and a termination signal when suitable. The Scanner rotates the agent's camera up, down, left or right and invokes the Detector on each image while maintaining the agent's location. The Detector uses fine-tuned YOLO-v3 for object detection and FRCN depth estimation network for estimating depth of an object. The detection probabilities are incorporated into the spatial memory. The Manipulator manipulates the current state of an object using open/close actions. The Answerer uses the current image, the full spatial memory and the question embedding vector to predict answer probabilities a_i for each possible answer to the question. Since the individual tasks of the controllers are mostly independent, each controller is pretrained separately, followed by joint training of the HIMN model.

The experiments compare the mean accuracies of the HIMN model + YOLO detections with a random baseline (most likely answer per question type), A3C agent with ground truth detections and human performance. The HIMN model yields highest accuracies for all 3 question types (existence, counting and spatial relationship). Ablations are conducted over the HIMN model, providing it with Yolo detections, ground truth detections and ground truth navigator

(HIMN-GT), HIMN-GT without question-access and HIMN-GT with no loss for invalid actions. HIMN-GT yields highest accuracies. Supervised loss for invalid actions is found to be more effective than penalties. When testing generalization to unseen environment, HIMN model loses only a few percentage points.

2 Strengths

The paper has many contributions (IQUAD-v1 dataset, HIMN model and the esGRU implemented spatial semantic memory). The authors have done a good job of analyzing the performance of HIMN model over multiple settingd and ablations, considering tht large number of moving parts in the HIMN model. The IQUAD-v1 dataset introduces active interaction, compared to the passive setting in other datasets for question answering. The spatial semantic memory is very rich as it can be used for navigation, planning, question-answering and enables long term recall from very old observations. Learning of trivial solutions and simple rules is prevented by associating each question with multiple scene configurations that result in different answers to the question. The percentage of invalid actions for HIMN-GT model is lower than that of human. The HIMN model generalizes well to unseen environments.

3 Weaknesses

The multiple choice answers to the questions in the IQUAD-v1 dataset form a limited answer set that may not be generalizable. Top-1 accuracy is used as the evaluation metric, which may be cherry picking the best runs. Its not clear how the local egocentric copies of the global spatial memory are obtained and updated. Even though the training methodology is specified, no final training objective or loss function is provided (no equations in the paper). The threshold for passing the terminal signal predicted by the Navigator to the Planner is not mentioned. The choice of A3C agent as a baseline is questionable when its performing worse than a random agent. The Scanner and Manipulator are not trained, their behavior is instead predetermined (trainable Scanner and Manipulator can affect the results obtained). In Table 3, its not clear why the lengths for spatial relationship questions shorter than those for existence questions. The inference from section 5.2 is that majority of failed actions are due to navigation failures. But we see a higher improvement in accuracies in Table 3 when going from YOLO to GT detections (row-1 to row-2), rather than going from GT detections to oracle navigator (row-2 to row-3). In section 5.5, its mentioned that HIMN is unable to differentiate between an object being inside a container and being on top of the container. This limitation will be critical for the Manipulator.

4 Reflections

The paper has many novel contributions and avenues for future research. The contributions are a big step in the Interactive Question-Answering literature, but the generalizability of the contributions to neighbouring domains is questionable (particularly the HIMN model).

5 Most Interesting Thought

I would like to know the implementation details for the spatial semantic memory using esGRU and the training objective for a model with so many moving parts.