Review9: Vision-and-Dialog Navigation

Vivswan Shitole

CS539 Embodied AI

1 Summary

In the spirit of smart home robots that navigate the environment and interact with humans or other agents via question-answer dialogs, this paper introduces ”Cooperative Vision-and-Dialog Navigation” (CVDN), a dataset of embodied human-human dialogs situated in the Matterport R2R simulation environment. The dataset comprises of dialogues between a navigator that intends to reach a goal region from a starting node and asks questions when it needs guidance, and answers from an oracle that has privileged access to the next best steps that the navigator should take according to a shortest path planner. Based on the CVDN dataset, the paper introduces a ”Navigation from Dialog History” (NDH) task that can be used to train navigation agents for search of a goal region or agents that answer navigation-related questions given expert knowledge of the environment to enable automated language guidance for humans in unfamiliar places. The agent trained for the NDH task must learn to infer navigation actions towards the goal in unexplored environments given the dialog history from the CVDN dataset. The authors train a multi-modal sequence-to-sequence model using student-forcing for the NDH task and demonstrate that looking farther back in the dialog history improves performance.
The CVDN dataset comprises of 2050 human-human navigation dialogs over 7K navigation trajectories punctuated by question-answer exchanges across 83 Matterport houses. The dataset was collected through Amazon Mechanical Turk. Every dialog was instantiated with a randomly chosen prompt (S,t_0,p_0,G_j), where S is the house scan, p_0 is the starting panorama where the navigator starts, t_0 is an ambiguous hint for the goal location (in text) and G_j is the goal region. The dialogs exhibit complex phenomena that require both dialog and navigation history to resolve. For each dialog with a prompt (S,t_0,p_0,G_j), an NDH instance is created. The training input is t_0 and a (possibly empty) history of questions and answers $(Q_{1:i},A_{1:i})$. The task is to predict navigation actions that bring the agent closer to the goal location G_j starting from the previous node N_{i-1} (or p_0 for N_0). Two forms of supervision is provided for the task: N_i, the navigation steps taken by the navigator after question-answering exchange i, and O_i, the shortest path steps shown to the oracle and used as context to provide answer A_i. 7415 NDH instances are extracted from the 2050 navigation dialogs in CVDN that are divided into 4742 training, 382 seen validation, 907 unseen validation and 1384 unseen test instances. Performance on the task is evaluated by measuring the reduction in the distance from the goal region G_j at the beginning versus at the end node of the path.
The experiments involve training a multi-modal sequence-to-sequence model for the NDH task. The dialog history is encoded using an LSTM and used to intialize the hidden state of an LSTM decoder. whose observations are visual frames from the environment and outputs are actions. The agent is trained using student forcing for 20000 iteration of batch size 100 and validated every 100 iterations. At each time step the agent executes its inferred action $\hat{a}$ and is trained using cross entropy loss against the action a^* that is next along the shortest path to the end node. Ablations are performed at training time over the distance of dialog history encoded and the supervisory signal (navigator, oracle or mixed supervision). The sequence-to-sequence agent is compared with other baselines involving a full-state information shortest path agent, a non-learning random agent and unimodal baselines (agents that consider only visual input or only language input). The sequence-to-sequence agent with access to sufficient dialog hostory outperforms unimodal baselines but underperforms in comparison to the shortest path agent, particularly in unseen environments (as the shortest path agent with navigator supervision approximates human performance). The sequence-to-sequence agent performs better under mixed aupervision, with the performance improving with the length of dialog history provided to the agent, verifying the authors' hypothesis that looking farther back in the dialog history improves performance.

2 Strengths

The CVDN dataset and the NDH task form a great contribution that is closer to the real world setting and can lead to novel theoretical insights as the agent trained on such a task need to grapple with complex multimodal input and needs to be interactive via question-answering rather than passively conducting the navigation task. The live interface demo provided with the paper greatly helps in understanding the CVDN data creation. The authors do a good job of comparing the CVDN dataset to other datsets in the related work and in Table 1. The experiments involve many ablations (Table 2) that point out the important aspects of the sequence-to-sequence agent, such as the role of mixed supervision and the length of dialog history.

3 Weaknesses

No details are provided on the shortest path planner that is used to inform the oracle about the next best steps. The authors' claim that other datasets for VLN tasks use language instructions that are unambiguous and fully specified does not seem to be correct. When comparing with the ”Talk the Walk” task, authors claim that language grounding centers around semantic elements such as ”bank” and ”restaurant”. Same can be true for the NDH task since the CVDN dialogs include semantic elements. On the other hand, the CVDN datset is not sufficient to learn a language grounding as over 90% of all dialogs contain egocentric references requiring the agent's orientation and position to interpret. Table 3 shows a drop in performance for the sequence-to-sequence agent when provided with dialog history in the validation (seen) fold under mixed supervision. This contradicts the suthors' hypothesis that looking farther back in the dialog history improves performance.

4 Reflections

The paper is well positioned in the embodied navigation literature as it contributes a novel dataset and a training task which requires agents to actively dialog with humans or other agents in the environment rather than passively conducting the navigation task. It has multiple directions for future research such as using RL based formulation for training the agent's policy or inverting the NDH task to yield agents trained to provide navigation guidance in unseen environments.

5 Most Interesting Thought

The most interesting idea in the paper for me is that the navigation learning task is interspersed with the language grounding task. It would be great to see if we can verify that the trained agent asks a question only when it really needs the answer, or it just picks up a temporal pattern for asking the questions.