CS539 - Embodied AI
Paper Review - Multi-Target Embodied Question Answering

The paper proposes Multi-Target Embodied Question Answering (MT-EQA) task, associated dataset and a new model with a modular architecture. MT-EQA generalizes and extends EQA to a more challenging multi-target setting, requiring an agent to navigate to multiple locations and perform comparative reasoning before answering a question. The MT-EQA dataset consists of questions relating to objects color comparisons within/across rooms, objects size comparisons within/across rooms, objects distance comparisons within a room, room size comparisons, overall 19,287 questions across 588 environments. The modular architecture consists of 4 components, a program generator, a navigator, a controller and a VQA module. Overall the experiments demonstrate that the model significantly outperforms baselines on both question answering and navigation tasks, detailed ablative analysis for each component also support effectiveness of these components.

Strengths of the paper are that it introduces a new task of MT-EQA, extending an existing task to a more challenging one. Paper also contributes the dataset to support training/evaluation of the task. Novelty also lies in using a modular architecture for the model, there is novelty in their navigator, controller and VQA answering module. The experiments do show a significant increase in performance over baselines on the 4 question types, given an oracle setting the Controller and cVQA modules show significant bump over baselines supporting its learned understanding. As for the navigator, when fine tuned with RL the metrics show fair improvement of not overshooting the target, decreased episode length and overall an improvement to the EQA question answering.

Weakness of the paper were that it could not show improvements on across the room question answering tasks, showing deficiency in long term planning for the navigator. Even under an oracle setting, the model couldn't perform well on distance comparison tasks, showing some deficiency in the image feature extractor. The metrics on medium length and hard length tasks show that EQA metrics are not that much high as compared to random, showing that EQA module might need some improvement.

Reflections are that the paper might benefit from maintaining and updating a Global Semantic Memory, rather than having all that information in the LSTM with limited memory, a dedicated Memory block as used in IQA: Visual Question Answering in Interactive Environments, would help improve the metrics on distance comparisons and cross room comparison tasks. I guess next in this line of work would be more refinements in the model, addition of some sort of dedicated memory blocks.

Most interesting thought about the paper is that all the elements for the task are setup, ie the dataset, modular architecture, balanced question generation and comparision to baselines, so now most of the work can be focused on improving the model. Interesting bits from the model were the use of Controller module to select and store image frames and then feed it to the VQA, I'm not too sure if the VQA conditioned on the attribute(color), op(equal) with all the linear layers followed by ReLU is a good choice or not, this wasn't actually fully understood.