# CS539 - Embodied AI

## Paper Review - IQA: Visual Question Answering in Interactive Environments

The paper introduces the task of Interactive Question Answering, where an agent given a scene and a question must navigate within the scene, acquire visual understanding, interact with objects to deduce an answer for the question. The paper proposes a Hierarchical Interactive Memory Network (HIMN) to address the above mentioned task, which consists of a hierarchy of controllers and a rich semantic memory that aids in navigation, interaction and question answering. The paper also introduces IQUAD V1, a new dataset of 75,000 questions and scene configurations in a simulated photorealistic environment. Experiments show that the proposed model outperforms popular single controller based methods, their semantic memory unit has been shown to generalize the model well across seen and unseen environments.

Strengths of the paper are that it is a formulation of a new task IQA, proposing a new method to address it and introduction of a dataset for training models for the task. The Suggestions about Hierarchical Planner, Semantic Memory and Egocentric Spatial GRU show merit and are all new ideas. The experiments were fairly executed showing merit to their proposal of Semantic Memory Unit and that model could have benefited from better object detection than YOLOV3.

Reflections are that the paper relates to Cognitive Mapping and Planning for Visual Navigation paper sharing the idea of learning an egocentric map of its environment using a hierarchical planner, some ideas about free space prediction, its model architecture could be borrowed here. The next research direction I guess would be to improve their Global Semantic Memory block, store more spatial information, which would improve its results on all question types. I guess a good choice of object detector and adding cross modal matching constraints would help improve results further.

Most interesting thought about the paper is its way of storing information about the global map via its Global Semantic Memory block, it keeps track of object detection probabilities, occupancy grid, coverage, navigation intent etc. It would be interesting to store more spatial information, object information here, 3d scene information and that would surely translate to improved performance on Counting and Spatial Relationships question types.