Cognitive Planning and Navigation for Visual Navigation: Paper Review

The authors introduce a neural network architecture- Cognitive Mapper and Planner (CMP) to learn policies for visual navigation tasks in novel indoor environments. The mapper architecture produces a metric-egocentric multiscale belief about the world, and planner uses this belief to plan paths to the desired specified goal, and outputs the optimal action to reach the goal. In this unified architecture framework, the mapper updates the belief of the world using an ego-motion to transform the belief from the last time step to the current coordinate frame, and then updating the belief using the current view of the world, thereby improving the model as it moves around. The belief update is learned using a CNN with first person view as the input. Planer uses a trainable, differentiable and hierarchical version of value iteration, to deal with partial observability and to reduce the time complexity to reach distant goals.
The proposed architecture has been tested for Geometric tasks and Semantic tasks on Stanford large scale 3D indoor spaces (S3DIS). The performance has been compared with the reactive agent which uses the first person view of the world with sequence of previous frames, memory(LSTM) based agent, as well as classical methods of purely geometric mapping with access to depth, images. The results demonstrate that the proposed method outperforms all learning based methods across all metrics as well as classical purely geometric method with only RBG images, but classical approach performs better with depth images.

One of the strengths of the paper is in their approach where the network learns the belief of the world, instead of analytically predicting it, which enables it to learn the statistical regularity of the environment, and could be one of the core reasons for the out-performance of the model on novel environments. Overall, the paper is written in very detailed way, and the obtained results have been analysed and explained in very details. Further, the baselines used for comparing the performance are quite intuitive and makes full sense, for instance the classical approach with RGB and depth image, and the one with memory based model. Also, the real-time deployment on Turtle-bot to validate the approach is also one of the strengths of the paper.

There are few weaknesses as following. Firstly, in the mapping architecture, it is not clear how the confidence value along with the cumulative estimate of the free space helps in improving the performance. There should be some experiment without using confidence value to update the function U. Further, by asserting that using additional dataset (for instance Matterport 3D) improves the performance, it is not clear if the improvement in performance is due to the additional dataset, or is there any special features in MP3D dataset, which helps in generalizing the learning which is improving the performance. Further, I would be interesting to know how the map has been retrieved from the latent space, what metric has been used to obtain similarity of maps with the ground truth free-space. Also, there is no analysis to support how hierarchical value iteration reduced the time complexity, and show computation time and memory requirement.

Overall, the idea of an end-to-end learning for mapping and planning using metric beliefs is novel. It has set a good benchmark for similar future works in the area of visual navigation, and could be combined with various other datasets with more complex environment to test the performance. I would be curious to apply to extend this work on a robot with odometer errors, as perfect odometry is one of the assumptions (they mentioned in the paper). One possible way could that in addition to learn the belief update, network will also be learning to model the odometer errors and update the beliefs of the odometer errors.