Cognitive Mapping and Planning for Visual Navigation

Saurabh Gupta[1,2] James Davidson[2] Sergey Levine[1,2] Rahul Sukthankar[2] Jitendra Malik[1,2]
[1]UC Berkeley [2]Google
[1]{sgupta, svlevine, malik}@eecs.berkeley.edu, [2]{jcdavidson, sukthankar}@google.com

***Abstract*—We introduce a neural architecture for navigation in novel environments. Our proposed architecture learns to map from first-person views and plans a sequence of actions towards goals in the environment. The Cognitive Mapper and Planner (CMP) is based on two key ideas: a) a unified joint architecture for mapping and planning, such that the mapping is driven by the needs of the planner, and b) a spatial memory with the ability to plan given an incomplete set of observations about the world. CMP constructs a top-down belief map of the world and applies a differentiable neural net planner to produce the next action at each time step. The accumulated belief of the world enables the agent to track visited regions of the environment. Our experiments demonstrate that CMP outperforms both reactive strategies and standard memory-based architectures and performs well in novel environments. Furthermore, we show that CMP can naturally achieve semantically specified goals, such as "go to a chair".**

I. Introduction

As humans, when we navigate through novel environments we draw on our previous experience in similar conditions. We reason about free-space, obstacles and the topology of the environment, guided by common sense rules and heuristics for navigation. The goal of this paper is to design a learning framework for acquiring such expertise, and demonstrate it for the problem of robot navigation in novel environments.

In contrast, classic approaches to navigation rarely make use of such common sense patterns. Classical SLAM based approaches [2, 8] first build a 3D map using LIDAR, depth or structure from motion, and then plan paths in this map. These maps are built purely geometrically, and nothing is known until it has been explicitly observed, even when there are obvious patterns. This becomes a problem for goal directed navigation. Humans can often guess, *e.g.* where they will find a chair or that a hallway will probably lead to another hallway, but a classical robot agent can at best only do uninformed exploration. The separation between mapping and planning also makes the overall system unnecessarily fragile.

Inspired by this reasoning, recently there has been an increasing interest in end-to-end learning-based approaches that go directly from pixels to actions [4, 5] without going through explicit model or state estimation steps. These methods thus enjoy the power of being able to learn behaviors from experience. However, it is necessary to carefully design architectures that can capture the structure of the task at hand. For instance Zhu *et al.* [10] use reactive memory-less vanilla feed forward architectures for solving visual navigation problems. In contrast, experiments by Tolman [9] have shown that even rats build sophisticated representations for space in the form of 'cognitive maps' as they navigate, giving them the ability to reason about shortcuts, something that a reactive agent is unable to.

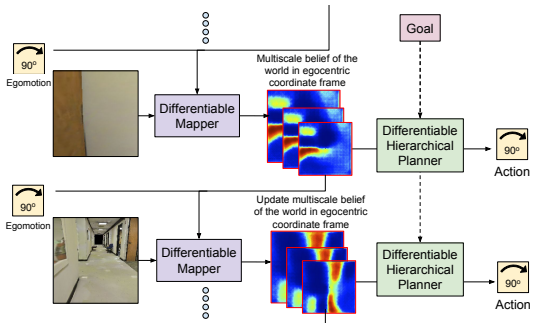

Fig. 1: Overall network architecture: Our learned navigation network consists of mapping and planning modules. The mapper writes into a latent spatial memory that corresponds to an egocentric map of the world, while the planner uses this memory with the goal to output navigational actions. The map is not supervised explicitly but emerges from the learning process.

II. Approach

This motivates our Cognitive Mapping and Planning (CMP) approach for visual navigation (Figure 1). CMP consists of a) a spatial memory to capture the layout of the world, and b) a planner that can plan paths given partial information. The mapper and the planner are put together into a unified architecture that can be trained to leverage regularities of the world. The mapper fuses information from input views as observed by the agent over time to produce a metric egocentric multi-scale belief about the world in a top-down view. The planner uses this multi-scale egocentric belief of the world to plan paths to the specified goal and outputs the optimal action to take. This process is repeated at each time step to convey the agent to the goal.

At each time step, the agent updates its belief of the world from the previous time step by a) using the ego-motion to transform the belief from the previous time step into the current coordinate frame and b) incorporating information from

This work presents a two page version of the full paper [3] that will appear at CVPR 2017. Please refer to [3] for details. Project website with code, models, simulation environment and videos: https://sites.google.com/view/cognitive-mapping-and-planning/.

Work done when S. Gupta was an intern at Google.

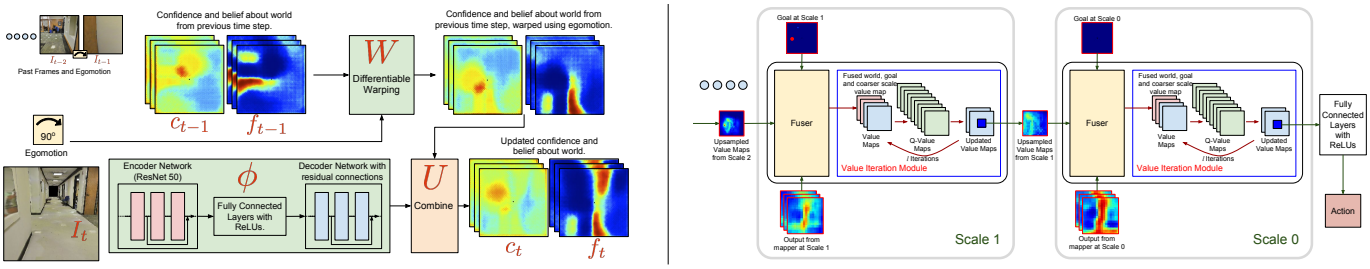

Fig. 2: Architecture of the mapper (left): The mapper module processes first person images from the robot and integrates the observations into a latent memory, which corresponds to an egocentric map of the top-view of the environment. The mapping operation is not supervised explicitly – the mapper is free to write into memory whatever information is most useful for the planner. In addition to filling in obstacles, the mapper also stores confidence values in the map, which allows it to make probabilistic predictions about unobserved parts of the map by exploiting learned patterns. **Architecture of the hierarchical planner (right)**: The hierarchical planner takes the egocentric multi-scale belief of the world output by the mapper and uses value iteration expressed as convolutions and channel-wise max-pooling to output a policy. The planner is trainable and differentiable and back-propagates gradients to the mapper. The planner operates at multiple scales (scale 0 is the finest scale) of the problem which leads to efficiency in planning.

Method	Mean		75th %ile		Success %age	
	RGB	Depth	RGB	Depth	RGB	Depth
Geometric Task						
Initial	25.3	25.3	30	30	0.7	0.7
Reactive (1 frame)	20.9	17.0	28	26	8.2	21.9
Reactive (4 frames)	14.4	8.8	25	18	31.4	56.9
LSTM	10.3	5.9	21	5	53.0	71.8
Our (CMP)	**7.7**	**4.8**	**14**	**1**	**62.5**	**78.3**
Semantic Task (Aggregate)						
Initial	16.2	16.2	25	25	11.3	11.3
Reactive (4 frames)	14.2	14.2	22	23	23.4	22.3
LSTM	13.5	13.4	20	23	23.5	27.2
Our (CMP)	**11.3**	**11.0**	**18**	**19**	**34.2**	**40.0**

TABLE I: Navigation Results: We report the mean distance to goal, 75th percentile distance to goal and success rate after executing the policy for 39 time steps. The top part presents results for the case where the goal is specified geometrically in terms of position of the goal in the coordinate frame of the robot. The bottom part presents aggregate results for the case where the goal is specified semantically in the form of 'go to a chair' (or door or table).

the current view of the world to update the belief (shown in Figure 2(left)). This allows the agent to progressively improve its model of the world as it moves around. The most significant contrast with prior work is that our approach is trained end-to-end to take good actions in the world. To that end, instead of analytically computing the update to the belief (via classical structure from motion) we frame this as a learning problem and train a convolutional neural network to predict the update based on the observed first person view. We make the belief transformation and update operations differentiable thereby allowing for end-to-end training. This allows our method to adapt to the statistical patterns in real indoor scenes without the need for any explicit supervision of the mapping stage.

Our planner (shown in Figure 2(right)) uses the metric belief of the world obtained through the mapping operation described above to plan paths to the goal. We use value iteration as our planning algorithm but crucially use a trainable, differentiable and hierarchical version of value iteration [7]. This has three advantages, a) being trainable it naturally deals with partially observed environments by explicitly learning when and where to explore, b) being differentiable it enables us to train the mapper for navigation, and c) being hierarchical it allows us to plan paths to distant goal locations in time complexity that is logarithmic in the number of steps to the goal.

The mapper and the planner are encapsulated in a unified neural architecture (Figure 1) that is trained jointly with imitation learning using DAGGER [6]. This allows learning a mapper that is driven by the needs of the planner.

Our approach is a reminiscent of classical work in navigation that also involves building maps and then planning paths in these maps to reach desired target locations. However, our approach differs from classical work in the following significant way: except for the architectural choice of maintaining a metric belief, everything else is learned from data. This leads to some very desirable properties: a) our model can learn statistical regularities of indoor environments in a task-driven manner, b) jointly training the mapper and the planner makes our planner more robust to errors of the mapper, and c) our model can be used in an online manner in novel environments without requiring a pre-constructed map.

III. Experiments

We conduct experiments on static simulated real world environments obtained from large-scale scans of indoor buildings [1]. We assume that the robot lives in a grid world and has primitives that allow it to move forward, or turn left or right by 90°. The robot observes the world through a regular RGB camera or a depth camera. We study two tasks, a geometric task where the goal is specified in the coordinate frame of the robot and a semantic task where the goal is to reach an object of a desired target category (chair, table or door). Testing is done in novel environments using scans from a different building with different floor-plan and appearance from the ones used for training.

Our experiments test the utility of spatial memory and a unified mapping and planning architecture for the task of visual navigation. To understand this we compare against a number of neural network based baselines, in particular a reactive policy based on a feed forward network that uses last few frames, a vanila neural network memory in the form of a LSTM. Results are presented in Table I. Our approach outperforms both these baselines consistently across both the geoemetric and the semantic task, thereby supporting our neural network design. The full version of the paper [3] shows comparisons to classical mapping, more extensive comparisons to LSTMs, analysis of individual components of the model and visualizations of successful and failed navigations.

REFERENCES

[1] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3D semantic parsing of large-scale indoor spaces. In *CVPR*, 2016. 2

[2] Andrew J Davison and David W Murray. Mobile robot localisation using active vision. In *ECCV*, 1998. 1

[3] Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive mapping and planning for visual navigation. In *CVPR*, 2017. Project website: https://sites.google.com/view/cognitive-mapping-and-planning/. 1, 2

[4] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *JMLR*, 2016. 1

[5] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 2015. 1

[6] Stéphane Ross, Geoffrey J Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *AISTATS*, 2011. 2

[7] Aviv Tamar, Sergey Levine, and Pieter Abbeel. Value iteration networks. In *NIPS*, 2016. 2

[8] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic robotics*. MIT press, 2005. 1

[9] Edward C Tolman. Cognitive maps in rats and men. *Psychological review*, 1948. 1

[10] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *ICRA*, 2017. 1