Rapid Learning Without Catastrophic Forgetting in Multiple Morris Water Mazes

Raymond L Wang **Jaedong Hwang** **Akhilan Boopathy** **Ila Fiete**

MIT

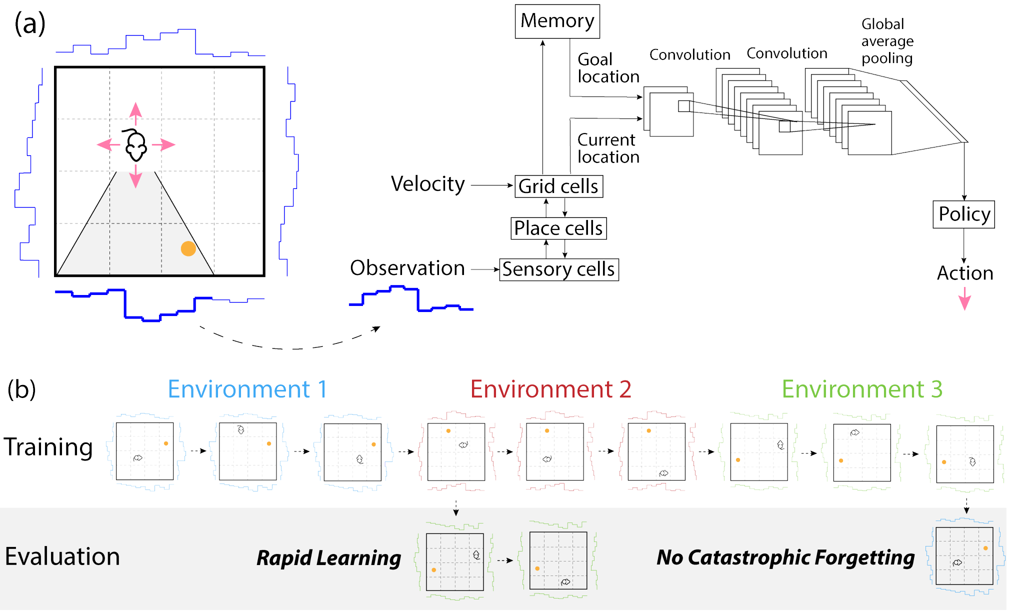

Figure 1: Schematic of our water maze environment and training setup. The rodent icon indicates the agent, arrows indicate the rodent's allowed actions, gold circles indicate the hidden platforms, and curves parallel to the walls of each environment indicate patterns along the walls. **(a)** shows the water maze environment and our model architecture. The agent observes a portion of the wall pattern. Observations, along with velocity inputs, are fed into a MESH network that produces grid cell activations representing the agent's location. An external memory module stores the grid code of the goal location. Grid codes of the current location and goal location are fed to a spatially-invariant convolutional neural network to produce a representation of the relative goal position. This is fed to a policy that produces actions. **(b)** shows our training setup. Agents are trained sequentially over multiple environments, with multiple trajectories in each environment. The agent is evaluated on rapidly learning to navigate in new environments and remembering navigation strategies in previously seen environments.

Abstract

Machine learning models can struggle to rapidly transfer their knowledge to new tasks while retaining their abilities on trained tasks; by contrast, animals perform this easily. We propose a novel neural model of this rapid learning process for the Morris Water Maze task. We combine 1) the Memory Scaffold with Heteroassociation (MESH) architecture, a model of the entorhinal-hippocampal circuit, to retain knowledge of many environments over time, and 2) a spatially invariant convolutional network architecture to rapidly transfer to new unseen environments. Experimentally, unlike baseline neural networks, our model both retains knowledge of previously seen environments and rapidly learns to navigate in new environments.

Keywords: Morris Water Maze, Catastrophic Forgetting, Neural Modeling

Introduction

Animals can quickly learn new tasks that are conceptually similar to previously encountered tasks, but have different inputs and surface-level details; importantly, *they retain the ability to solve the previous tasks*. Neural modeling of this process of rapid conceptual knowledge transfer has been limited. We develop a neural model of this process in the classic Morris Water Maze (Morris, 1981; Vorhees & Williams, 2006), in which a rodent must navigate to an unseen platform in a pool of water across multiple trials in multiple environments.

Conventional, unstructured neural networks suffer from catastrophic forgetting: a phenomenon where networks trained on one task fail to perform well on previously trained tasks (McCloskey & Cohen, 1989). Unstructured neural networks generally also lack an intrinsic ability to generalize to unseen tasks; rapid learning of unseen tasks typically requires extensive training on many previous tasks (e.g. using multi-task learning techniques). To avoid these shortcomings, we use a structured neocortical-entorhinal-hippocampal circuit, the Memory Scaffold with Heteroassociation (MESH) architecture (Sharma et al., 2022), to enable such generalization in the Water Maze *after training only on a single environment*.

Avoiding Catastrophic Forgetting With MESH

Morris Water Maze Task

We design a version of the Morris Water Maze task that assesses an artificial rodent's ability to retain knowledge of previously navigated environments and learn new ones quickly. In this task, the rodent is placed in a pool of water with distal cues, and its goal is to find an unseen platform to avoid staying afloat. Our version involves multiple water maze environments with different goal locations and distal cues, and the rodent must move in one of four directions (north, south, east, or west) based on its current position. The rodent is sequentially trained in each environment from randomly selected starting locations and tested on its ability to navigate from *unseen* starting locations in both *seen* and *unseen* environments. Our task evaluates both the rodent's ability to retain knowledge and rapidly learn. Figure 1 illustrates our setup.

Catastrophic Forgetting

We first tested a baseline neural network trained via supervised learning to map observations to movements. As Figure 2(b) shows, the model effectively learned the task when trained in a single environment, but when subsequently trained in a new environment, although it performed well in the new environment, it exhibited catastrophic forgetting: performance in the original environment degraded rapidly. This is not consistent with the behavior of biological rodents; next, we propose a model that avoids catastrophic forgetting.

Associating Grid Code Displacements with Movements

To address the problem of catastrophic forgetting, we use the Memory Scaffold with Heteroassociation (MESH) architecture, a structured neocortical-entorhinal-hippocampal circuit. MESH uses online learning rules to associate sensory cues in sensory cortical areas to grid cell codes of spatial location in the entorhinal cortex. Importantly, MESH can retain these associations over *many environments* without catastrophic forgetting: as the number of environments grows large, it exhibits only a gradual degradation in its recall performance.

Next, we develop a model of how rodents rapidly learn to navigate in new environments. Using a fixed convolutional neural network (CNN), our model maps the rodent's current and goal locations (encoded in a grid code) to a spatially-invariant representation of the *displacement* of the goal relative to its current position. A learned policy network then maps this displacement to the appropriate movement. Our architecture's spatial invariance allows for *forward transfer to new environments* as we show next. Figure 1(a) illustrates our model architecture.

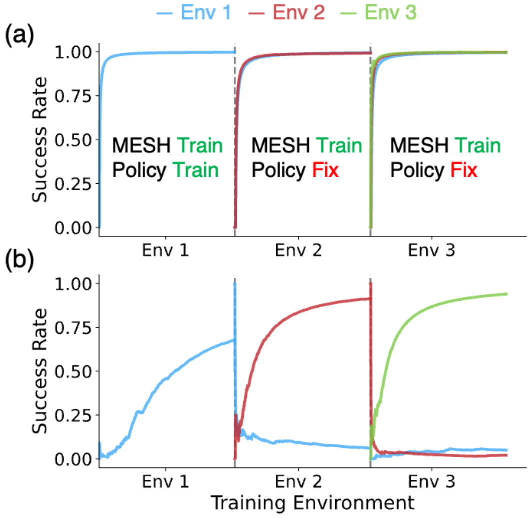

Figure 2: MESH and baseline comparison. **(a)** MESH training and evaluation curve. The x-axis represents the environment MESH is trained on, with the action-producing policy only trained in Environment 1. After training on each subsequent environment, all previous environments are tested. The rise from 0 for previous environments' success rate indicates the agent taking a few trajectories of reorienting itself and recognizing the prior environment. **(b)** Baseline training and evaluation curve. Each new environment is trained and evaluated on previous environments. Catastrophic forgetting occurs shortly after training on a few new trajectories.

Results & Discussion

In Figure 2, we observe that after being trained on the first environment, our approach rapidly learns to navigate subsequent environments *without any policy training*. This is attributed to the acquisition of a general, transferable navigation policy from the initial environment. Furthermore, our method prevents catastrophic forgetting by recalling past environments after recognizing the current environment through a few trajectories. In contrast, the baseline experiences catastrophic forgetting of previous environments almost immediately after learning a new environment.

In conclusion, our results demonstrate that our novel neural model, using the MESH architecture, is capable of rapidly learning and retaining knowledge across multiple environments while effectively transferring to unseen environments. This research underscores the potential of specialized neural models in addressing challenges that conventional deep learning methods struggle with, such as rapid learning, generalization, and avoiding catastrophic forgetting. Our findings highlight how structured neural models can tackle complex, real-world tasks, and their promising role in the context of artificial intelligence and cognitive science.

Acknowledgments

IF is supported by the Simons Foundation, the Office of Naval Research, and the Howard Hughes Medical Institute (HHMI).

References

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation* (Vol. 24, pp. 109–165). Elsevier.

Morris, R. G. (1981). Spatial localization does not require the presence of local cues. *Learning and motivation*, *12*(2), 239–260.

Sharma, S., Chandra, S., & Fiete, I. (2022). Content addressable memory without catastrophic forgetting by heteroassociation with a fixed scaffold. In *Icml.*

Vorhees, C. V., & Williams, M. T. (2006). Morris water maze: procedures for assessing spatial and related forms of learning and memory. *Nature protocols*, *1*(2), 848–858.