

Investigating the Interplay Between Memory and Map-Like Image Input in Reinforcement Learning Agents for Exploring Task

Yuki Ishihara

Asahigaoka High School

February 22, 2024,

Abstract

There was a new finding that cognitive maps are structured in the long-short-term memory (LSTM) during navigation tasks for artificial intelligence (AI). This is an example of the connection between machine learning and psychology. In this paper, it is shown that cognitive maps do not affect performance without memory again and the external visit information is a natural solution for exploring tasks. The agents try the task of touching two objects in the right order in a complex environment. In this condition, I investigate the effect of the input like a map image on their performance and behavior. As a result, agents with memory are likely to be less influenced by the new input than those without memory. Both types of agents also showed a slight difference in their behavior.

1. Introduction

Edward C. Tolman previously found that rats in a maze memorized the environment they were in without receiving any reward. It was claimed that rats ran around the environment to build a spatial representation in their brain after navigating the maze. It was also argued that humans usually construct maps or other representations in their brains. These representatives are called “cognitive maps”.¹ Animals have the cognition to mediate between stimuli and action to reach the destination.

Recently, it was shown that cognitive maps are represented in the memory of an AI model. The reinforcement learning model decides the agents’ actions based on their location in an environment, for example a house. The movement of agents is decided randomly at first. As training proceeds, they are optimized to receive a higher reward upon reaching the destination. A spatial representative emerged in the memory and helped the agents achieve a navigation task. (Wijmans et al, 2023) trained a separate network to predict the map given the LSTM state and examined if the agents could navigate well from the goal to the start with only the final LSTM state. As a result, the two tasks were successful, and the presence of the internal spatial representative was confirmed based on the evidence in that work. This also inferred building cognitive maps is a natural solution for navigation tasks. This is an example of a connection between organisms and AI models.²

¹ Edward C. Tolman. Cognitive maps in rats and men. *Psychological Review*, 55(4):189–208, 1948. doi: 10.1037/h0061626.

² Wijmans, Erik. “Emergence of Maps in the Memories of Blind Navigation Agents.” ICLR 2023 submitted, 2023.

Using not only recurrent neural networks (RNN) but also frame stacking, which summarizes features of sequential input data into one data, is also effective for speech recognition tasks. In a traditional RNN with LSTM, the speed of learning gradually decreases as the data to decode increases in number. However, frame staking of the audio input reduces the time for training while the accuracy is still competitive with the original RNN model.³

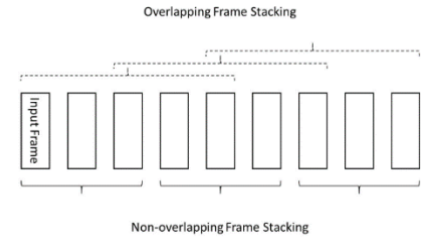


Figure 1: an image of frame stacking³

This attempt is also reported to be effective for Proximal Policy Optimization (PPO)⁴, an algorithm for reinforcement learning. PPO with a recurrent network is an effective way to learn sequential observations in reinforcement learning. However, it is reported that PPO with frame-stacking (giving a history of observation as input) is usually quite competitive if not better, and faster than recurrent PPO.⁵

There might be another connection between research results from psychology and the behavior of AI. The focus of this research is “map-like external input.” This paper shows what happens in its training speed, performance, and behavior if adding map-like input. People often use external information to organize their memory and keep their memory longer. There might be a specific phenomenon by this. The hypotheses of this research are to investigate:

1. Providing map-like external input changes agents’ performance and behavior.
2. The performance does not differ even if leveraging the effectiveness of both internal spatial representation and external abstracted map compared to the agent with only memory because it has been reported that LSTM-based agents naturally acquire cognitive maps in their memory.
3. Leveraging the external input reduces the time for training.

2. Methodology

In this experiment, agents attempt to reach the goal object in the environment after interacting with a switch-like object in an environment separated into 36 areas (Figure 2a). This is an exploring task for agents. The environment has the switch object, the goal object, and 7 boxes scattered across it as a guide for agents (Figure 2b). They can recognize this environment with sensors like rays from themselves, which can inform them of the kind of objects (like a wall, switch, guide box, etc.). The model of reinforcement learning is PPO. The rewards are given only when the agents reach the goal after interacting with the switch. Agents decide their actions based on the current observation and optimize their neural network to get higher rewards. While agents first move randomly, the actions are optimized as the training proceeds.

³ Tian, X., Zhang, J., Ma, Z., He, Y., & Wei, J. (2017). Frame Stacking and Retaining for Recurrent Neural Network Acoustic Model. *ArXiv*. /abs/1705.05992

⁴ Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal Policy Optimization Algorithms. *ArXiv*. /abs/1707.06347

⁵ RAFFIN, Antonin. “PPO VS Recurrentppo (Aka PPO LSTM) on Environments with Masked Velocity (SB3 Contrib).” *W&B*, 1 May 2022, wandb.ai/sb3/no-vel-envs/reports/PPO-vs-RecurrentPPO-aka-PPO-LSTM-on-environments-with-masked-velocity-SB3-Contrib---VmlldzoxOTI4NjE4#ppo-lstm-vs-ppo-(no-framestack).

Four types of agents were created and tested, depending on the use of map-like images (Figure 2c) and LSTM. Two of the agents use the map-like image as visitation information, with colors of tiles indicating the areas visited by the agents. Black tiles indicate unvisited areas and white tiles indicate visited areas. Two types of agents are trained using LSTM and a recurrent network. The memory is maintained during each episode. Both use the same basic reinforcement learning algorithm, employing PPO.

To compare the performance of two types of agents, these agents trained for 3 million episodes and an analysis was done on how fast the progress of the training was with the curve of their reward or the Policy Loss values (correlates to how appropriate the best action agents predicted was)⁶⁷. The way of examining the agents' behavior is by counting the areas agents had visited and observing the agents who completed training for an episode.

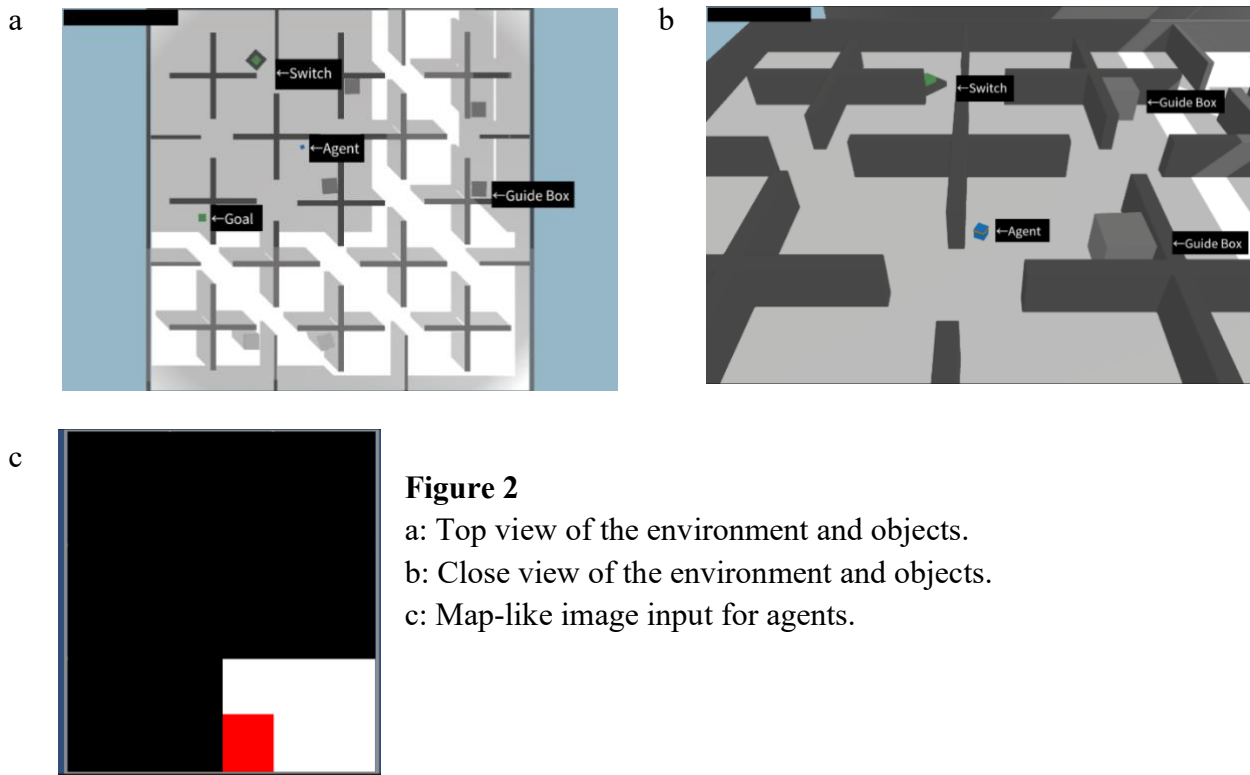


Figure 2

a: Top view of the environment and objects.

b: Close view of the environment and objects.

c: Map-like image input for agents.

⁶ “Learning Rate, Epsilon and Policy Loss.” *Unity Forum*, 8 August 2020, <https://forum.unity.com/threads/learning-rate-epsilon-and-policy-loss.948134/>. Accessed 5 January 2024.

⁷ van Heeswijk, Wouter. “The Four Policy Classes of Reinforcement Learning | by Wouter van Heeswijk, PhD.” *Towards Data Science*, 11 May 2021, <https://towardsdatascience.com/the-four-policy-classes-of-reinforcement-learning-38185daa6c8a>. Accessed 5 January 2024.

3. Results

3.1 The effect on performance by map-like input and LSTM

To investigate the effect on performance in the navigation task by map-like input and LSTM, the transition of agents was compared, and the difference in the value was tested by statistical hypothesis testing.

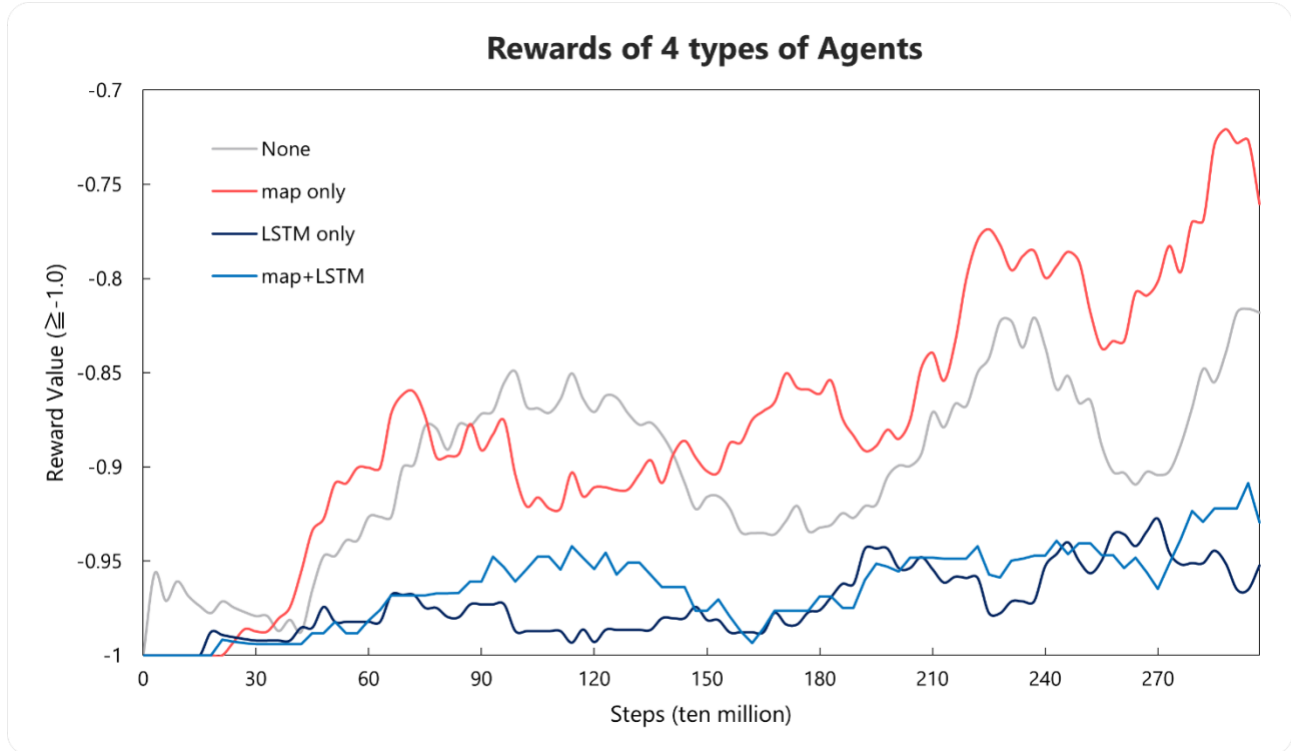


Figure 3: This shows that the agent with the map-like input successfully got more rewards than one without a map if they do not have memory. On the other hand, agents with memory and map-like images did not much improve their performance.

Result of statistical hypothesis testing of agents without LSTM

	None	map only
Reward Average	-0.8924	-0.8659
Standard Deviation	0.0918	0.1173
Degree of Freedom	187	
P value (one side)	0.0383	

Table 1: $t(187) = t$ statistic, $P < 0.05$

Result of statistical hypothesis testing of agents with LSTM

	LSTM only	LSTM + maps
Reward Average	-0.9604	-0.9703
Standard Deviation	0.0513	0.0576
Degree of Freedom	195	
P value (one side)	0.1002	

Table 2: $t(195) = t$ statistic, $P < 0.05$

3.2 The learning progress of each agent.

To check if the agents successfully learned, the transition of policy loss was compared. Policy loss correlates to how appropriate the best action agents predicted are and indicates progress in training.

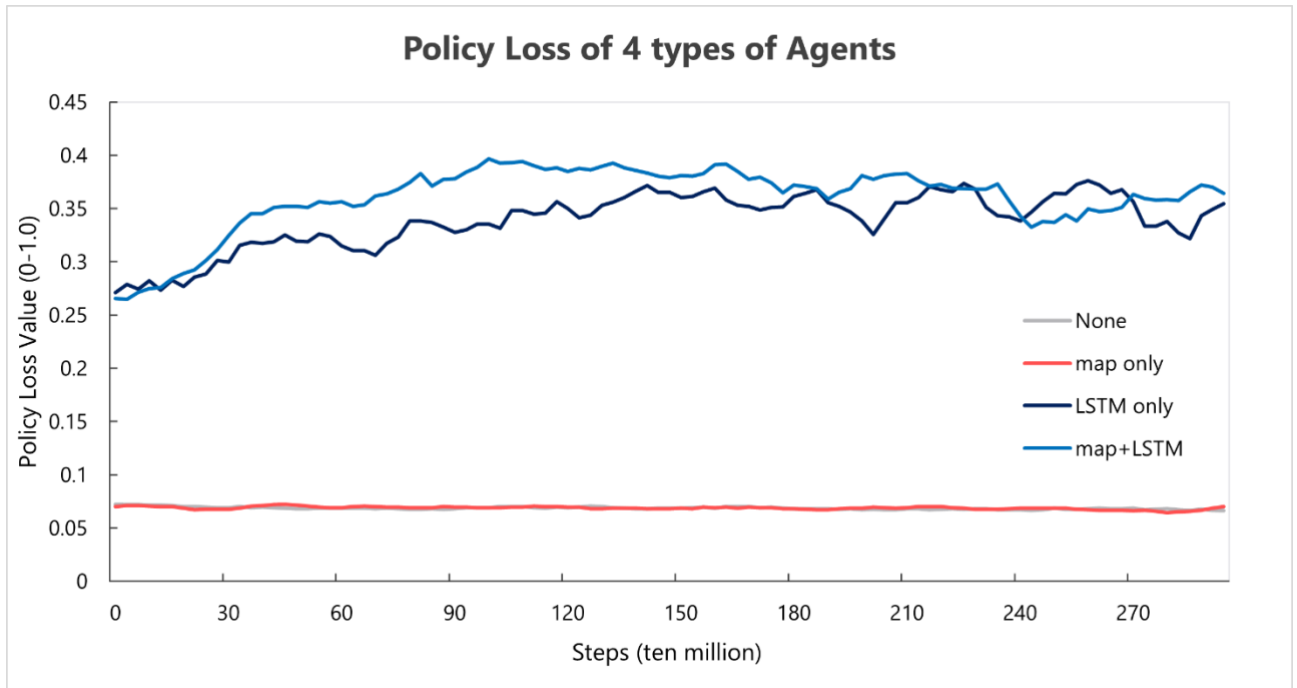


Figure 4: The policy loss of agents with memory remained high throughout the experiment.

3.3 The change in agents' behavior by map-like input and LSTM

The purpose of this experiment is to investigate the difference in the behavior of agents. The focused behavior is exploring since this exploration is important for this task.

The average number of areas where an agent visited for 5 minutes

	With map	Without map
With Memory	28.5	26.0
Without Memory	31.5	30.5

Table 3: The average number of areas where both agents with memory and those without memory visited was a little increased with visit information from the map image. The amount of change in exploring behavior was so close between agents with memory and those without memory.

3.4 Findings from Observation

From observation of the agents' behavior, the agents with memory seem to move and decide their way more slowly while the agents without memory move fast and often collide with the walls and explore across many areas.

3.5 The difference in training time

This is data to investigate the difference in the training time of each agent for 3 million steps.

The times of training for 3 million steps

	With map	Without map
With Memory	2h47m56s	1h41m35s
Without Memory	2h44m59s	1h25m25s

Table 4: The agents with map-like images spent significantly longer time, and the ones with LSTM did a little additional time for their training.

4. Discussion

The maps would help the agents without LSTM explore many areas (Table 3), which were connected to higher rewards (Figure 3). In contrast, Because the reward value of the agent with both a map and memory were less than that with only a map (Figure 3), taking advantage of internal spatial cognitive maps and external map-like input was not effective for this task in terms of performance and training time (Table 2). There is a possibility that cognitive maps in their memory might prevent another map-like input from improving the agents' performance. In other words, they already have information about where they visit; they no longer need other input about the visit information. This supports the idea that this representation works as cognitive maps for AI again. (Wijmans et al, 2023)

On the other hand, the agents seemed not to be able to learn during only 3 million steps because policy loss of agents with memory remained high throughout the experiment (Figure 4). According to a researcher, Erik Wijmans, "While an LSTM can learn to build a map, the map may not be as accurate as an external one and it takes more steps to learn to build one than to simply use the provided one. So, the LSTM agent may still benefit from being given the external map." Whereas longer training time may solve this deficient performance of the agents with both a map and memory based on this comment, further investigation was impossible due to lack of machine power. Another cause is hyperparameters. Wijmans said, "Using an RNN also tends to be more sensitive to initialization and hyperparameters... it's quite hard to find those"

The external input improved the performance (Figure 3). The external input was more effective in improving the performance in just 3 million steps. Now, training some AI models is time-consuming and costs a lot. Combining both external guides for the training and internal representation flexibly according to its task might be one of the effective ways to train an AI model.

In this experiment, the way to observe the environment around the agents was different from (Wijmans et al, 2023); the rays around the agents informed the model about the environment. That might affect the structure of cognitive maps in their memory. Due to the lack of methodology to search for internal representation, this remains unclear.

The external map-like input was not effective in improving training time. (Table 4) The cause is expected to be that this map-like image increases the amount of information the AI model must deal with. This input did not work as a flame-stacking like (Tian et al, 2017), which combined the input into simple information and improved the effectiveness of training of data in chronological order. Especially in this experiment, the map-like input was provided to the model through the camera sensor, and it gave a big load to the learning process.

5. Conclusion

The agents used map-like information and changed their behavior toward venturing into many areas, but the effect of the memory is unclear. However, the performance improvements by agents with LSTM were limited, possibly because of the existence of cognitive maps in their memory, the lack of training steps or hyperparameters.

To investigate more about the connection between psychology and machine learning, it is necessary to examine the longer memory across the episode because it may get closer to human brains, which have both implicit memory and explicit memory. However, it might be overflowing with data in their memory. I assumed that adjusting forgetting gates of LSTM enables the model to use their limited memory more efficiently and with high performance from the point of view of psychology and biology.

6. Acknowledgement

I would like to express my sincere thanks for Erik Wijmans's assistance. His valuable suggestions have provided me with new insights and have motivated me to further improve my research.

7. Environment

- Unity 2020.3.32f1
- ML-Agents 0.30.0
- PyTorch: 1.11.0