One Shot Spatial Learning through Replay in a Hippocampus-Inspired Reinforcement Learning Model

Adedapo Alabi
Dept. of Electrical Engineering
and Computer Science
University of Cincinnati
Cincinnati, USA
alabiaa@mail.uc.edu

Ali A. Minai
Dept. of Electrical Engineering
and Computer Science
University of Cincinnati
Cincinnati, USA
minaiaa@ucmail.uc.edu

Dieter Vanderelst
Dept. of Electrical Engineering
and Computer Science
University of Cincinnati
Cincinnati, USA
vanderdt@ucmail.uc.edu

***Abstract*—The neural basis of spatial cognition and learning in mammals has been studied extensively for several decades. Research has focused in particular on the place cells of the hippocampus and the grid cells found in the entorhinal cortex. In turn, these studies have inspired several models for robotic navigation. One interesting, experimentally observed, feature of spatial learning in rodents is the importance of replay, where animals replay sequences of spatial representations they have experienced in order to learn and make decisions. This feature too has been incorporated into some computational models. In this paper, we describe a new approach to learning navigation in mazes using replay of intrinsically generated sequences rather than relying only on experienced sequences. We show that this improves generalization, and leads to effective one-shot learning that is closer to what is observed in animals.**

I. Introduction

Spatial mapping and navigation are of central importance in animals such as mammals and birds, and understanding their biological basis is critical to the understanding of cognition and behavior. This topic has been studied extensively for several decades through experiments, modeling, and computational simulation. More recently, the results from these studies have also inspired methods for mapping and navigation in robots, typically in combination with neural learning algorithms, as is the case in this paper.

It is well-known that the hippocampus and its surrounding regions play a central role in spatial cognition in mammals [1]–[3]. This brain area contains (among others) place cells [2], [4], [5], grid cells [6], [7], and head-direction cells [8]. Major research efforts in this area over the past 30 years have led to a detailed understanding of these cells and the way they may support the construction of cognitive maps.

In brief, cognitive maps are supported primarily by place cells [2], [9] in the hippocampus and grid cells [1], [10]–[12] in the entorhinal cortex. During the formation of a cognitive map, each place cell becomes associated with a particular location in the environment, i.e., the place field of the cell [2], [4], [13], [14].

Grid cells are organized in a way that maximizes their spatial resolution for the fewest number of cells [15], [16], forming discrete modules differing in scale, ranging from a few centimeters to several meters in rodents [17]. A leading hypothesis about grid cell function is that they provide a path integration-based input to place cells [10], [18], [19]. Together, the place and grid systems allow animals to localize and navigate by integrating sensory and ideothetic information.

These insights from neuroscience, in turn, have inspired several research groups to develop methods for mapping and navigation in robots. These efforts have yielded impressive results. For example, RatSLAM [20]–[23] has been shown to be able navigate office spaces [24] and map large outdoor environments [25]. The algorithm has also been used successfully with multiple sensors (including cameras, sonar, and electromagnetic sensors) and to support sensor fusion [26]–[28].

Reinforcement learning (RL) is a biologically-inspired learning method for training agents by optimizing rewards. It has proven to be a powerful technique, in particular when complete knowledge of the environment is unavailable, rendering supervised training impossible. The primary inspiration for reinforcement learning is the naturally occurring learning by trial and error in animals, which can explore new environments and quickly learn how to maximize rewards in those environments, e.g., learning where to find food. However, despite the progress in the underlying neuroscience, biological RL is far from fully understood. In particular, computational models of RL are unable to reproduce the ability of animals to learn from a small number of trials. One potential explanation for this ability is that animals can augment real-time learning with off-line learning by replaying experiences. Such replay has been observed in the hippocampus during sleep [29]–[31]. Replay has also incorporated into several RL models of hippocampal function [32]–[34] and generated interesting results.

In this paper, we follow these previous models in assuming that hippocampal replay plays a key part in natural RL for navigation tasks. However, our model differs from previous

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

ones because we focus on *awake* replay in sharp wave ripples during the receipt of rewards, and propose that the learning process is better modelled as *asynchronous* dynamic programming estimating the value function rather that the traditional view of temporal difference (TD) learning.

II. Background

A. Reinforcement Learning

Reinforcement learning problems are typically formulated as Markov decision processes (MDPs) described by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R})$ where:

- $\mathcal{S}$ is the set of agent-environment states.
- $\mathcal{A}$ is the set of available actions to the agent.
- The *state transition function*, $\mathcal{T}(s, a) \rightarrow s'$, maps a state-action pair to the next state – deterministically or stochastically.
- The *reward function*, $\mathcal{R}(s'|s, a) \rightarrow r$, indicates the reward in transitioning to state s' from state s by taking action a.

The task is then to find a deterministic *policy*, $\pi(s) \rightarrow a$, that maps states to actions to maximize the expected sum of future rewards i.e.

$$\pi^*(s) := argmax_a \sum_{s'} (R(s'|s, a) + \gamma V(s')) \tag{1}$$

where $V(s)$ is the *value function* which provides the sum of discounted rewards that can be earned from state s by following the policy $\pi(s)$ discounted by a factor $\gamma \in [0, 1]$ at every step, i.e.

$$V(s) := \sum_{s'} (R_{\pi(s)}(s'|s) + \gamma V(s')) \tag{2}$$

Solving this problem usually involves updating both the value and policy functions recursively using approaches such as temporal difference (TD) learning, Monte Carlo methods, or dynamic programming when complete knowledge of the environment and rewards is available. TD and Monte Carlo generally work by using the error δ_t between the estimated value of the current state $\hat{V}(s)$ and the better estimate $R_{t+1} + \gamma \hat{V}(s') - \hat{V}(s)$ obtained either after each state transition (TD) or after entire episodes of experience (Monte Carlo). Dynamic programming, which is more computationally intensive, can be thought of as proceeding backward in time to back-propagate the maximum value reachable from each state, discounted by the number of steps taken and combining it with the reward at the state to update its value. The policy is then computed from this value function using Equation 1. Explicitly:

$$V_{i+1}(s) := \max_a \sum_{s'} (R(s'|s, a) + \gamma V_i(s')) \tag{3}$$

where i is the iteration number.

B. Spatial Cognition in the Hippocampus

As discussed above, spatial cognition in mammals is mediated by spatially-tuned cells in the hippocampus and surrounding brain regions. The details of these representations have been elucidated by many studies, but three features are of special relevance to the model presented in this paper.

1) Markovian Aspect of Place Cells: Neuroscientists have discovered a great variety of spatially modulated cells that are believed to be crucial for spatial cognition. For this work, we focus on hippocampal *place cells*, which are neurons that increase their firing in localized regions of the environment termed their place fields. It has been observed that place cells appear to learn a Markovian (rather than Euclidean) state-space representation [33], i.e., it reflects the sequential nature of the animal's experience during the traversal of the environment rather than purely metric relationships between locations. Given the strongly recurrent structure of the CA3 region of the hippocampus where place cells are often found, spike time-dependent plasticity (STDP) can cause their synapses to represent sequential information obtained during navigation. This is also reflected in the fact that place fields typically do not straddle impenetrable boundaries [35] in the environment – presumably because locations across the boundary are never experienced sequentially, and the correct action at each one is unlikely to be the same despite their proximity in Euclidean space.

2) Place Cell Backward Replay: A particularly interesting phenomenon observed in place cell networks is the backward replay of the place cell sequences that led to rewards during the consumption of the rewards [36]. Dynamic programming has traditionally been seen as a purely algorithmic formulation with no equivalent in biological neural networks as it proceeds backward in time. However, the backward replay of sequences observed at reward sites can potentially enable a similar process to occur.

3) Reward Cells: Rewards, or the expectation thereof, appear to be represented by neurons in the ventral striatum that show ramping activity as the animal moves toward rewarded sites [37]–[39]. This is a downstream structure from the hippocampus that shows high synchronization with the hippocampus during active movement.

III. The Model

We define and implement a hippocampally-inspired model of mapping and navigation for a simple animat (simulated animal). The model involves four types of cells: Head-direction cells, boundary vector cells, place cells, and reward cells that produce a reward of +1 at specific locations. The current version of the model does not include grid cells, which will be added in future versions.

A. Head Direction Network

The head direction (HD) network consists of head direction cells with neurons that fire maximally when the agent's allocentric head direction is aligned with the cell's preferred direction. Experimental findings indicate the presence of head direction cells in the entorhinal cortex [40] and the postsubiculum [41]. The model proposed in this work uses only eight HD cells tuned at 45° intervals for simplicity. Mathematically,

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

The activity vector **H** of the neurons in the HD network is modeled as in the work of Erdem and Hasselmo [42]:

$$\mathbf{D} = \left[\begin{array}{c} cos(0), cos(45°)...cos(315°) \\ sin(0), sin(45°)...sin(315°) \end{array} \right] \tag{4a}$$

$$\mathbf{H} = \mathbf{v} \cdot \mathbf{D} \tag{4b}$$

where **D** is the tuning kernel, **v** is the velocity (row) vector, and some fixed distal cue is taken as the 0° heading.

B. Place Cell Network

Place cells are implemented using the previously proposed boundary vector cell model [43]. Boundary vector (BV) cells are neurons found in the subiculum of rodents that encode insurmountable obstacles or boundaries at specific distances and allocentric directions from the animal. This translates to BV cells having band-like receptive fields as illustrated in Fig 1 (left) for a circular environment.

In the boundary vector cell model of place cells, place cells multiplicatively combine the firing of multiple boundary vector cells with different preferred directions. We implement this model using boundary vector cells tuned to the eight directions represented in the head direction network with randomly distributed preferred distances. Place cells receive connections from a set of eight BV cells, one for each encoded head direction, with random preferred distances. Mathematically, for place cell i, its gross firing rate $\hat{f}_i(x)$ at a position x is given as:

$$\hat{f}_i(x) = \prod_{j=1}^{8} \exp\left[-\frac{(r_j - d_j)^2}{2\sigma_{rad}^2} \right] \tag{5}$$

where r_j is the distance of the nearest boundary in the preferred direction of boundary vector cell j and d_j is the preferred distance. The gross firing rates are normalized at each time-step as a fraction of the maximum firing rate at the current position across all place cells to get the final firing rate:

$$f_i(x) = \frac{\hat{f}_i(x)}{\hat{f}_j(x)} \tag{6}$$

where j is the index of the place cell with the maximum firing rate at x.

C. Markov Decision Process Problem Formulation

1) $\mathcal{S}$: The set of agent-environment states, $\mathcal{S}$, can be taken as the vector of place cell firing rates, i.e given n place cells;

$$\mathcal{S} = \{f_1(x), f_2(x), ..., f_n(x)\} \tag{7}$$

2) $\mathcal{A}$: For simplicity, we assume the animat moves at a constant speed $|\mathbf{v}|$ in any direction, which is represented by head direction cells [41] coding for the eight allocentric cardinal directions. Therefore, the set of available actions, $\mathcal{A}$, is,

$$\mathcal{A} = [|\mathbf{v}|\underline{\angle 0°}, |\mathbf{v}|\underline{\angle 315°}) \tag{8}$$

3) $\mathcal{V}$: The value function, $\mathcal{V}$, is simply the firing rate of the ventral striatum *reward cell* associated with the currently sought reward which also receives projections from place cells. We assume that a higher level brain structure which we do not model assigns reward cells to the rewards found, and selects which reward cells to evaluate when seeking known goals.

$$\mathcal{V} = r + \mathbf{W}_{SR}\mathcal{S} \tag{9}$$

where r is a binary value representing the receipt of the reward and W_{SR} is the weight matrix from the place cells to the reward cell.

4) $\mathcal{T}$: As the reward locations and place field adjacencies are initially unknown, the policy is randomly initialized to allow exploration. We represent the state transition function $\mathcal{T}(s, a)$ as a 3 dimensional place cell to place cell synapse matrix $\mathbf{W}_{SS}$ where $\mathbf{W}_{SS}[k][i][j]$ represents the strength of connection from the i^{th} place cell to the j^{th} in the k^{th} head direction. During exploration, $\mathbf{W}_{SS}$ is updated after every time-step using Hebbian learning modulated by the head direction network activity:

$$\mathbf{W}'_{SS} = \mathbf{H}(\mathbf{W}_{SS} + \eta_1((\mathcal{S}'\mathcal{S}^T)|1 - \mathbf{W}_{SS}|) \tag{10}$$

where **H** and $\mathcal{S}$ are the head direction and state vectors respectively as previously described, η_1 is the learning rate, and self connections are not allowed i.e the diagonals are zero.$\mathbf{W}'_{SS}$ and $\mathcal{S}'$ represent $\mathbf{W}_{SS}$ and $\mathcal{S}$ on the next timestep.

D. Replay

As previously stated, rodents are known to replay the place cell activation sequences in reverse order upon reaching rewards at the end of runs [36]. This occurs during a very synchronous neural state termed sharp wave ripples. This is believed to emerge due to the absence of the theta-modulated inhibition of the place cell to place cell connections during exploration [44], [45]. Since the synapses between place cells with adjacent place fields would have been potentiated during exploration, a temporary lack of theta mediated inhibition causes the currently active place cells, i.e., those active at the goal location, to depolarize place cells adjacent to them, which then do the same and so on, resulting in a time-compressed backwards replay of all (or many) of the previously experienced paths to the goal. If the active ventral striatum reward cell receives projections from the place cells, spike-time-dependent-plasticity during this backward spread of activity will modify the synapses from each activated place cell to the active reward cell in inverse proportion to the distance from that place cell's field to the reward location. Crucially, if a location's place cell has been part of several previous paths to the goal, its first activation in the backward replay corresponds to the shortest known path from that location to the goal. Thus, the learning process automatically adjusts the synaptic modification to account for the most accurate value estimate. Over the course of experience, which is reflected in the increase of its synaptic strength to the reward cell.

We approximate this in our rate-coded model in the following manner. When receiving a reward, the selected reward

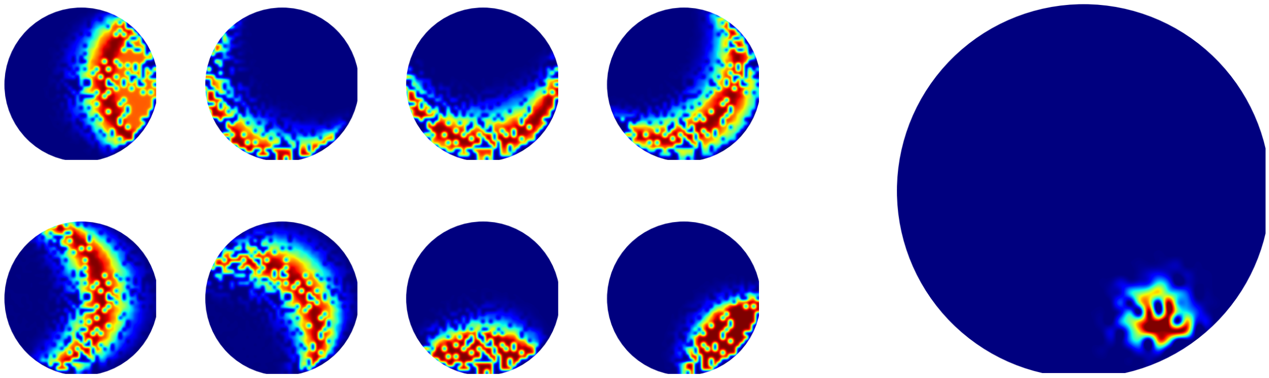

Fig. 1. Left: The smoothed spatial activity patterns in a circular arena during a random run are shown for eight boundary vector cells that are tuned to boundaries at different distances in the eight cardinal directions. Right: The resulting place field of a place cell receiving input from the eight boundary vector cells and competing with other place cells for excitation.

cell receives a signal $r = 1$ and becomes active. Hebbian learning then occurs so that the synapses from the currently active place cells get strengthened. Then backward replay ensues from the currently active place cells spreading out to all connected neurons along the strongest synapse i.e., the head direction most strongly connecting them. The strongest synapse is used as it represents the most efficient known action or shortest path. As place cells might have shown some activation during the previous time-step due to place field overlap and/or future time-steps due to the undirected nature of the spreading activation replay, the place cell to reward cell synapse is set to be the maximum of its previous value and currently implied value. This indicates when the place cell and its field are at the peak of the spreading activation. Activation then spreads from these place cells after which learning occurs again and so on for a specified number of steps, n_{rs}, forming what can be thought of as a ripple in physical space starting at the reward location and spreading out along the known paths. To prevent the network becoming unstable, the net activation is normalized at every time-step.

This entire process is summarized in Algorithm 1, where η_2 is the time-decay learning rate, $A \succcurlyeq_{i,j} B$ is used to mean $A[i,j] \geq B[i,j]$ and $\max_k \mathbf{W}_{SS}$ is the maximum of $\mathbf{W}_{SS}$ along the direction axis i.e. the maximum synapse strength between each place cell pair.

Algorithm 1: Replay Generation Method

```
for $r_i \leftarrow 1$ $\mathit{to}$ $n_{rs}$ do
    $\hat{\mathbf{W}}_{SR} \leftarrow e^{\frac{-\eta_2}{n_{rs}}} \mathcal{S}$;
    if $\hat{\mathbf{W}}_{SR} \succcurlyeq_{i,j} \mathbf{W}_{SR}$ then
        $\mathbf{W}_{SR}[i][j] \leftarrow \hat{\mathbf{W}}_{SR}[i][j]$;
    end
    $\hat{\mathcal{S}} \leftarrow \max_k \mathbf{W}_{SS} \mathcal{S}$;
    $\mathcal{S} \leftarrow \frac{\hat{\mathcal{S}}}{||\hat{\mathcal{S}}||}$;
end
```

The net effect of the replay and associated learning is to establish in the place-to-reward cell synapses an implicit value map for the environment with respect to that goal. If there can be multiple reward locations, each with its own reward cell, different maps can be learned without mutual interference for each such location using the same place fields. When a specific goal/reward is sought, the animat can estimate the value of each possible move via the activity each proposed location induces in the appropriate reward cell.

E. Exploration vs Exploitation

The animat operates in two modes:

During *exploration*, it moves around randomly and so that it is equally likely to maintain its current heading or to make a 90° turn that is equally likely to be to the right or the left – all the while avoiding obstacles. This enables it to learn a partial value map over the locations visited during exploration, which is generalized by off-line reverse replay. During *exploitation*, the animat moves based on the value map it has inferred, taking the highest value action at each step.

In simulation, once a value map has been learned, the animat employs a stochastic greedy exploration/exploitation strategy where the probability of exploitation is inversely exponentially proportional to the expected reward value.

In a familiar environment, rodents are known to stop and vicariously sample the different options at decision points [46]–[48]. This behaviour which manifests physically as small head movements alternating between the potential choices has more recently been discovered to be accompanied by an activation of the place cell sequences associated with the sampled paths [49] and the reward cells [50]. If the strength of the synapses between place cells and the reward cells reflect the proximity of the place cell's field to the reward location as we propose, this provides a way for the animal to evaluate the value of the available actions at each state i.e. $\mathcal{V}(s'|s,a)$.

We approximate this in the following way for the actions being considered next from the current location. Given the current state $\mathcal{S}$:

1) For each proposed next action $a \in \mathcal{A}$, generate $\mathcal{S}'$ – the place cell activity for the new location – by letting

activation spread between place cells along the synapse representing the action:

$$\mathcal{S}' = \mathbf{W}_{SS}[i]\mathcal{S} \tag{11}$$

where i is the index of the evaluated action a in $\mathcal{A}$.

2) Evaluate the value $\mathcal{V}$ for each proposed location $\mathcal{S}'$:

$$\mathcal{V}(\mathcal{S}') = \mathbf{W}_{SR}\mathcal{S}' \tag{12}$$

3) Choose the action a^* that gives the maximum value and move to the corresponding location.

The complete process is explained in Algorithm 2 with $\mathcal{V}$ as defined in (12) and π^* as defined in (1).

Algorithm 2: Exploration/Exploitation Strategy

```
while $r < 1$ do
    if reward previously found then
        Select most recently activated reward cell;
        Evaluate $\mathcal{R}(s'|s,a)$ for all $a \in \mathcal{A}$ i.e
          $\mathcal{V}(s'|s,a) - \mathcal{V}(s)$;
        if $max(\mathcal{R}(s'|s,a)) > 0$ then
            $v^* = \mathcal{V}(s) + max(\mathcal{R}(s'|s,a))$;
            $p \leftarrow \exp(-\beta/v^*)$;
            Follow $\pi^*(s)$ with a probability $p$ or
              explore with a probability $1-p$;
        else
            Explore
        end
    else
        Explore
    end
end
```

IV. Experimental Environment

We implemented our model on the Khepera IV robot in Webots [51], a commercial mobile robot simulation software developed by Cyberbotics Ltd. In addition to the standard package, the robot was equipped with a 360° LIDAR sensor to detect obstacle distances, and a compass to get its allocentric heading.

The model was tested on the classic reference memory water maze task. The Morris water maze task [52], [53] involves placing a rat into a circular tank of colored water where there is a hidden platform. Its natural aversion to swimming (though rats are perfectly capable of it) motivates the rat to find the platform which is considered the end of the episode.

In our experiment, the animat moved in a circular environment, with a platform placed at a random location (Figure 2). Arriving at the platform produced a reward. The experiment was run ten times with random initial starting positions and platform (reward) locations. We considered a single run to be four episodes from random initial starting locations, no prior knowledge of the environment and a fixed goal location.

Variable parameters were set to the values described in the table below:

Parameter	η_1	η_2	n_{rs}	β
Value	1	0	3	.135

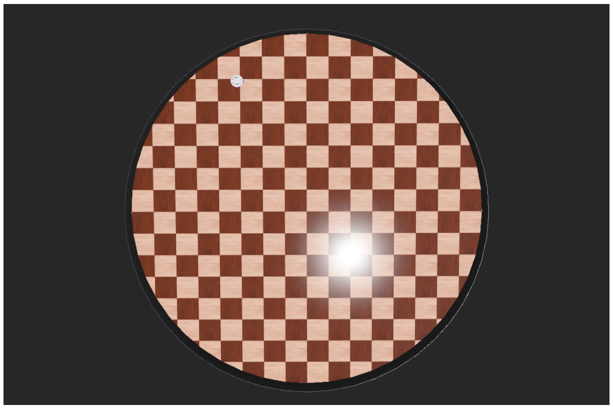

Fig. 2. The simulation environment is shown with the robot in the upper left corner and the goal represented in white towards the bottom right corner.

V. Results

Figure 3 shows the mean normalized escape latency (MNEL) – the time from start to reaching the platform scaled by the distance between the start point and the platform – for the four episodes in each trial. The rapid decrease in the MNEL from the first to the second trial indicates that the system exhibits strong one-shot learning, i.e., one exploratory episode is sufficient for the animat to build a reasonable value map of the environment. As can be expected, however, when the initial run only covers a limited part of the environment, the animat needs to explore further on a subsequent run that does not start in a part of the environment it had previously seen. This is the cause of the relatively high escape latency variance on the second run. By the third and fourth run, most of the environment has been explored, leading to almost direct paths. A sample trial is shown in Figure 4, where the animat explored the environment extensively before finding the goal

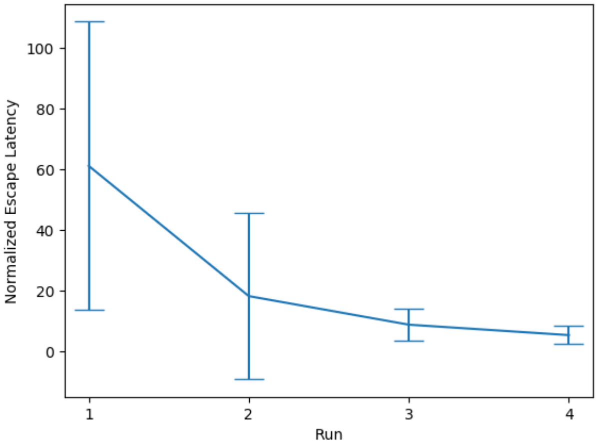

Fig. 3. The normalized escape latency is shown over four runs. It drops consistently over the four runs, indicating that the animat is able to build and successfully use a value map from experience.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

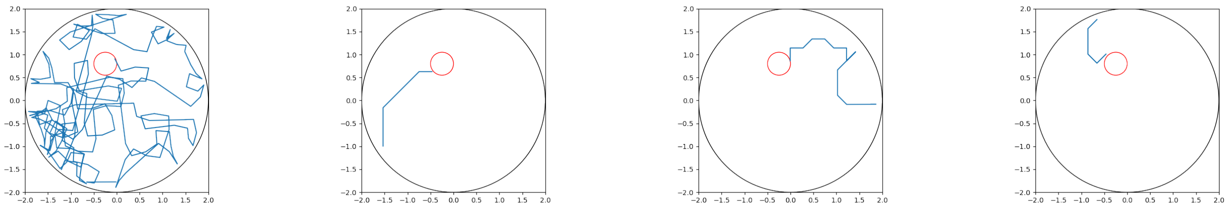

Fig. 4. A sample trial where the animat had explored most of the environment before finding the escape platform. This enabled the animat to build an extensive value map resulting in it finding fairly direct paths from the multiple different starting locations on subsequent runs.

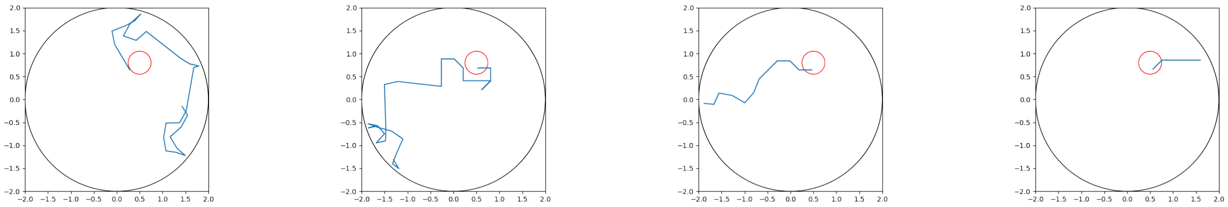

Fig. 5. A sample trial where the animat found the escape platform quickly without exploring most of the maze. On the second run from an unfamiliar location, it has to explore again to find the goal.

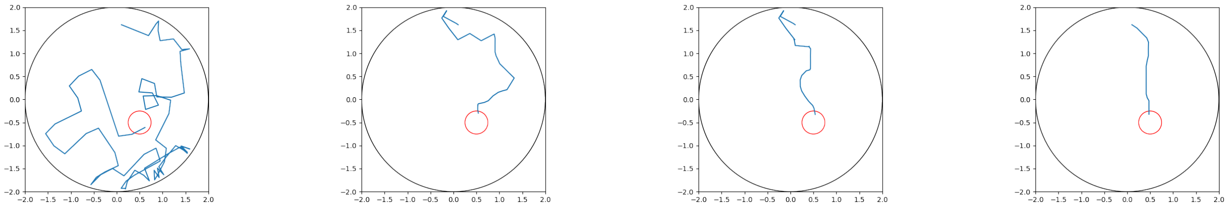

Fig. 6. Optimization of the path taken to the goal from the same starting point is shown over four runs. On the initial run, the goal location is unknown and the animat explores randomly till it discovers it. On subsequent runs, it exploits the previously learnt reward map and gradually improves it to generate a more efficient path.

in the first episode, and was thus able to build a value map with broad coverage. This enabled it to find the goal efficiently from new start positions on the subsequent episodes.

Figure 5 shows a trial in which the animat found the escape platform quickly in the first episode before exploring most of the environment. In the second episode from an unfamiliar starting location, it had to explore further to find the goal and then update the value map with this experience. This value map was then exploited on the third and fourth runs when it started near previously seen locations.

As the place fields have some overlap, i.e., place cells with place fields adjacent to the maximally active place cell also have some level of activation, the value map learned by the animat is not confined only to the locations it has explored; it also interpolates naturally to locations in between. Thus, after sufficient exploration, the animat is also able to infer completely novel shortcuts. This is shown in Fig. 6 where the animat begins from the same start location on each of the four runs. After finding the goal on the first run, it follows an inefficient path that reflects its experience. However, it is able to gradually improve this path over the next runs as it learns that the adjacent place fields are more direct.

VI. Discussion

Our model succeeds in learning the explored portion of the maze in one shot with the intrinsically generated replay, which effectively approximates dynamic programming. While this is not a particularly novel concept in computation, it has previously been unclear how this might occur in biologically-plausible neural networks. While we do not claim our model be true to the biology in its details, the previously discussed Markovian nature of place cells and the backward replay observed at reward sites indicate that a similar process might occur in the hippocampus.

An interesting aspect of the model, and one that distinguishes it from previous models, is that replay occurs in *all directions* from the reward location. What has typically been observed in experimental studies is replay in the direction from which the animal approached the goal, i.e., only replay of the most recent path. This could imply a more complex mechanism that preferentially selects recently activated place cells or is perhaps more likely to be an artifact of the testing environments that usually feature one-dimensional tracks and, as a result, directional place cells. Also, in our model, even in the direction of approach, the replay might not follow exactly the same sequence as that taken to reach the goal. That is, once an environment has been learned sufficiently, regardless of the path taken to find the reward in the current episode, place cells will replay the best sequence of known positions backwards from the reward. This is similar to what has been observed experimentally in humans [54] showing replay reorganized to reflect learned rules over the immediately preceding experience.

A notable limitation in our model is that set by n_{rs} on the number of replays or how far out the value map spreads from the reward location. Outside the laboratory, many mammals navigate and learn on scales from centimeters to tens of kilometers [55]. Future work on this model would include devising a method to spread out the value map to the range required of the task.

An aspect of our model which is biologically implausible is the directionality of the place cell to place cell synapses. Rather than having a single synapse connecting each place cell pair, we represent this using eight synapses, one for each of the encoded head directions. This is unlikely to be the case in the hippocampus but was necessitated to enable evaluation of the value of the possible actions. In a more biologically plausible model, this could be replaced by a model of the entorhinal cortex, which has been suggested to underlie path integration. Such a model would likely require recurrent connectivity between the hippocampus and entorhinal cortex, as has been found in anatomical studies [56]. The location encoded in the hippocampus, as well as the action under consideration, would be transmitted to the entorhinal cortex where the expected new location from taking the selected action would be determined and fed back to the hippocampus.

In this work, we have only focused on online replay that happens while the animal is awake. However, offline replay that occurs while the animal is asleep has also been observed experimentally [29]–[31]. This could further increase the resolution of the model and spread out the learnt value map beyond the range of the initial online replays.

VII. Conclusions

In this paper, we presented a hippocampally-inspired model of goal-seeking navigation using replay and reinforcement learning. A simple version of the model implemented on a similated robot showed that the model could rapidly build a value map in an environment, and use it to find the goal very efficiently – including discovering shortcuts that had not be experienced during learning. In future work, this model will be extended to include grid cells of the entorhinal cortex, to deal with obstacles, and to work in environments of varying size and complexity.

References

[1] Edvard I. Moser, Emilio Kropff, and May-Britt Moser. Place cells, grid cells, and the brain's spatial representation system. *Annu. Rev. Neurosci.*, 31:69–89, 2008.

[2] John O'Keefe and Lynn Nadel. *The hippocampus as a cognitive map*. Oxford: Clarendon Press, 1978.

[3] Nachum Ulanovsky and Cynthia F. Moss. Hippocampal cellular and network activity in freely moving echolocating bats. *Nature neuroscience*, 10(2):224–233, 2007.

[4] Robert U. Muller, John L. Kubie, and James B. Ranck. Spatial firing patterns of hippocampal complex-spike cells in a fixed environment. *Journal of Neuroscience*, 7(7):1935–1950, 1987.

[5] Robert U. Muller and John L. Kubie. The firing of hippocampal place cells predicts the future position of freely moving rats. *Journal of Neuroscience*, 9(12):4101–4110, 1989.

[6] Torkel Hafting, Marianne Fyhn, Sturla Molden, May-Britt Moser, and Edvard I. Moser. Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052):801–806, 2005.

[7] Marianne Fyhn, Torkel Hafting, Alessandro Treves, May-Britt Moser, and Edvard I. Moser. Hippocampal remapping and grid realignment in entorhinal cortex. *Nature*, 446(7132):190–194, 2007.

[8] David C. Rowland, Yasser Roudi, May-Britt Moser, and Edvard I. Moser. Ten Years of Grid Cells. *Annual Review of Neuroscience*, 39(1):annurev–neuro–070815–013824, 2016.

[9] A. David Redish and David S. Touretzky. Cognitive maps beyond the hippocampus. *Hippocampus*, 7(1):15–35, 1997.

[10] Bruce L. McNaughton, Francesco P. Battaglia, Ole Jensen, Edvard I. Moser, and May-Britt Moser. Path integration and the neural basis of the 'cognitive map'. *Nature reviews. Neuroscience*, 7(8):663–678, 2006.

[11] Caswell Barry, Robin Hayman, Neil Burgess, and Kathryn J Jeffery. Experience-dependent rescaling of entorhinal grids. *Nature neuroscience*, 10(6):682, 2007.

[12] Daniel Bush, Caswell Barry, Daniel Manson, and Neil Burgess. Using Grid Cells for Navigation. *Neuron*, 87(3):507–520, 2015.

[13] Tim V. P. Bliss and Terje Lømo. Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *The Journal of physiology*, 232(2):331–356, 1973.

[14] Jill K. Leutgeb, Emily A. Mankin, and Stefan Leutgeb. Population coding by place cells and grid cells. pages 300–317, 2013.

[15] Alexander Mathis, Andreas V. M. Herz, and Martin B. Stemmler. Multiscale codes in the nervous system: the problem of noise correlations and the ambiguity of periodic scales. *Physical Review E*, 88(2):022713, 2013.

[16] Martin Stemmler, Alexander Mathis, and Andreas V. M. Herz. Connecting multiple spatial scales to decode the population activity of grid cells. *Science Advances*, 1(11):e1500816, 2015.

[17] Hanne Stensola, Tor Stensola, Trygve Solstad, Kristian Frøland, May-Britt Moser, and Edvard I. Moser. The entorhinal grid map is discretized. *Nature*, 492(7427):72, 2012.

[18] Trygve Solstad, Edvard I. Moser, and Gaute T. Einevoll. From grid cells to place cells: a mathematical model. *Hippocampus*, 16(12):1026–1031, 2006.

[19] Daniel Bush, Caswell Barry, and Neil Burgess. What do grid cells contribute to place cell firing? *Trends in neurosciences*, 37(3):136–145, 2014.

[20] Michael J. Milford, Gordon F. Wyeth, and David Prasser. RatSLAM: a hippocampal model for simultaneous localization and mapping. In *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*, volume 1, pages 403–408. IEEE, 2004.

[21] Gordon Wyeth and Michael Milford. Spatial cognition for robots. *Robotics & Automation Magazine, IEEE*, 16(3):24–32, 2009.

[22] Michael J. Milford and Gordon F. Wyeth. Mapping a suburb with a single camera using a biologically inspired slam system. *IEEE Transactions on Robotics*, 24(5):1038–1053, 2008.

[23] Jean Kumagai. Special report: Can we copy the brain?-navigate like a rat. *IEEE Spectrum*, 54(6):58–63, 2017.

[24] Michael Milford and Gordon Wyeth. Persistent navigation and mapping using a biologically inspired slam system. *The International Journal of Robotics Research*, 29(9):1131–1153, 2010.

[25] Michael J. Milford. *Robot navigation from nature: Simultaneous localisation, mapping, and path planning based on hippocampal models*, volume 41. Springer Science & Business Media, 2008.

[26] Rafael Berkvens, Maarten Weyn, and Herbert Peremans. Asynchronous, electromagnetic sensor fusion in RatSLAM. *2015 IEEE SENSORS - Proceedings*, 2015.

[27] Adam Jacobson, Zetao Chen, and Michael Milford. SLAM, Autonomous Multisensor Calibration and Closed-loop Fusion for SlM. *Journal of Field Robotics*, 32(1):85–122, 2015.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

[28] Jan Steckel, Andre Boen, and Herbert Peremans. Broadband 3-D sonar system using a sparse array for indoor navigation. *IEEE Transactions on Robotics*, 29(1):161–171, 2013.

[29] Nicolas Maingret, Gabrielle Girardeau, Ralitsa Todorova, Marie Goutierre, and Michaël Zugaro. Hippocampo-cortical coupling mediates memory consolidation during sleep. *Nature neuroscience*, 19(7):959, 2016.

[30] Jens G. Klinzing, Niels Niethard, and Jan Born. Mechanisms of systems memory consolidation during sleep. *Nature neuroscience*, pages 1–13, 2019.

[31] Susanne Diekelmann and Jan Born. The memory function of sleep. *Nature Reviews Neuroscience*, 11(2):114, 2010.

[32] Adam Johnson and A. David Redish. Hippocampal replay contributes to within session learning in a temporal difference reinforcement learning model. *Neural Networks*, 18(9):1163–1171, 2005.

[33] David J. Foster and James J. Knierim. Sequence learning and the role of the hippocampus in rodent navigation. *Current opinion in neurobiology*, 22(2):294–300, 2012.

[34] David J. Foster, Richard G. M. Morris, and Peter Dayan. A model of hippocampally dependent navigation, using the temporal difference learning rule. *Hippocampus*, 10(1):1–16, 2000.

[35] Robert U. Muller and John L. Kubie. The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells. *Journal of Neuroscience*, 7(7):1951–1968, 1987.

[36] David J. Foster and Matthew A. Wilson. Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature*, 440(7084):680, 2006.

[37] Matthijs A. A. van der Meer, Adam Johnson, Neil C. Schmitzer-Torbert, and A. David Redish. Triple dissociation of information processing in dorsal striatum, ventral striatum, and hippocampus on a learned spatial decision task. *Neuron*, 67(1):25–32, 2010.

[38] Wolfram Schultz, Paul Apicella, Eugenio Scarnati, and Tomas Ljungberg. Neuronal activity in monkey ventral striatum related to the expectation of reward. *Journal of neuroscience*, 12(12):4595–4610, 1992.

[39] Matthijs A. A. van der Meer and A. David Redish. Theta phase precession in rat ventral striatum links place and reward information. *Journal of Neuroscience*, 31(8):2843–2854, 2011.

[40] Francesca Sargolini, Marianne Fyhn, Torkel Hafting, Bruce L. McNaughton, Menno P. Witter, May-Britt Moser, and Edvard I. Moser. Conjunctive representation of position, direction, and velocity in entorhinal cortex. *Science*, 312(5774):758–762, 2006.

[41] Jeffrey S. Taube, Robert U. Muller, and James B. Ranck. Head-direction cells recorded from the postsubiculum in freely moving rats. i. description and quantitative analysis. *Journal of Neuroscience*, 10(2):420–435, 1990.

[42] Uğur M Erdem and Michael E Hasselmo. A biologically inspired hierarchical goal directed navigation model. *Journal of Physiology-Paris*, 108(1):28–37, 2014.

[43] Caswell Barry, Colin Lever, Robin Hayman, Tom Hartley, Stephen Burton, John O'Keefe, Kate Jeffery, and Neil Burgess. The boundary vector cell model of place cell firing and spatial memory. *Reviews in the Neurosciences*, 17(1-2):71–98, 2006.

[44] György Buzsáki. Hippocampal sharp wave-ripple: A cognitive biomarker for episodic memory and planning. *Hippocampus*, 25(10):1073–1188, 2015.

[45] Gene V. Wallenstein and Michael E. Hasselmo. Gabaergic modulation of hippocampal population activity: Sequence learning, place field development, and the phase precession effect. *Journal of Neurophysiology*, 78(1):393–408, 1997. PMID: 9242288.

[46] Karl F. Muenzinger. Vicarious trial and error at a point of choice: I. a general survey of its relation to learning efficiency. *The Pedagogical Seminary and Journal of Genetic Psychology*, 53(1):75–86, 1938.

[47] Dan Hu and Abram Amsel. A simple test of the vicarious trial-and-error hypothesis of hippocampal function. *Proceedings of the National Academy of Sciences*, 92(12):5506–5509, 1995.

[48] Dan Hu, Xiaojuan Xu, and Francisco Gonzalez-Lima. Vicarious trial-and-error behavior and hippocampal cytochrome oxidase activity during y-maze discrimination learning in the rat. *International Journal of Neuroscience*, 116(3):265–280, 2006.

[49] Adam Johnson and A. David Redish. Neural ensembles in ca3 transiently encode paths forward of the animal at a decision point. *Journal of Neuroscience*, 27(45):12176–12189, 2007.

[50] A. David Redish. Vicarious trial and error. *Nature Reviews Neuroscience*, 17(3):147, 2016.

[51] Webots. http://www.cyberbotics.com. Commercial Mobile Robot Simulation Software.

[52] Richard G. M. Morris. Spatial localization does not require the presence of local cues. *Learning and motivation*, 12(2):239–260, 1981.

[53] Richard G. M. Morris, Paul Garrud, John N. P. Rawlins, and John O'Keefe. Place navigation impaired in rats with hippocampal lesions. *Nature*, 297(5868):681–683, 1982.

[54] Yunzhe Liu, Raymond J. Dolan, Zeb Kurth-Nelson, and Timothy E. J. Behrens. Human replay spontaneously reorganizes experience. *Cell*, 178(3):640–652, 2019.

[55] Talbot H. Waterman. *Animal Navigation*. Scientific American Library, 1989.

[56] Pieterke A. Naber, Fernando H. Lopes da Silva, and Menno P. Witter. Reciprocal connections between the entorhinal cortex and hippocampal fields ca1 and the subiculum are in register with the projections from ca1 to the subiculum. *Hippocampus*, 11(2):99–104, 2001.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE