Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 13 (2012) 198 – 204

Procedia
Computer Science

Proceedings of the International Neural Network Society Winter Conference (INNS-WC 2012)

# An Autonomous Distal Reward Learning Architecture for Embodied Agents

Shawn E. Taylor, Michael J. Healy, and Thomas P. Caudell *

*Dept of Electrical & Computer Eng., University of New Mexico, Albuquerque, NM 87131-0001 USA*

**Abstract**

Distal reward refers to a class of problems where reward is temporally distal from actions that lead to reward. The difficulty for any biological neural system is that the neural activations that caused an agent to achieve reward may no longer be present when the reward is experienced. Therefore in addition to the usual reward assignment problem, there is the additional complexity of rewarding through time based on neural activations that may no longer be present. Although this problem has been thoroughly studied over the years using methods such as reinforcement learning, we are interested in a more biologically motivated neural architectural approach. This paper introduces one such architecture that exhibits rudimentary distal reward learning based on associations of bottom-up visual sensory sequences with bottom-up proprioceptive motor sequences while an agent explores an environment. After sufficient learning, the agent is able to locate the reward through chaining together of top-down motor command sequences. This paper will briefly discuss the details of the neural architecture, the agent-based modeling system in which it is embodied, a virtual Morris water maze environment used for training and evaluation, and a sampling of numerical experiments characterizing its learning properties.

*Keywords:* autonomous agents, neural architectures, distal reward learning, seld organizing systems

## 1. Introduction

Animals learn much about the world they live in by interacting with it. Trial and error behavior provides them with valuable learning opportunities about cause (action) and effect (reward) relationships. Aside from

* Corresponding author. Tel.:505-277-5637; fax: 505-277-1439.
*E-mail address:* tpc@ece.unm.edu.

1877-0509 © 2012 Published by Elsevier B.V. Selection and/or peer-review under responsibility of Program Committee of INNS-WC 2012
doi:10.1016/j.procs.2012.09.129

reflexive behavior, there is usually a significant time delay between trial and error that confounds standard correlational (i.e. Hebbian [1]) learning approaches. This problem is known as delayed reward [2] or distal reward [3] and has been thoroughly studied over the last two decades under the broad topic of reinforcement learning (RL) [4]. Roughly speaking, RL uses a (possibly stochastic) rule set called a *policy* to map states of perception into actions of an agent. The short-term goal of the agent is defined by a reward function that maps perceptual states into a graded reward value. The long-term expected reward is the *value* of a state and is estimated using learning algorithms. In the RL paradigm, an agent chooses an action that maximizes this expected reward.

From a biological perspective, distal reward describes a class of agent behavior that requires associating sensory observations, motor actions, and rewards through time using only neural mechanisms. Learning these associations involves reinforcement of neuronal activity and synaptic adaptation due to the experienced reward [5]. Often the reward becomes evident to the agent a significant time after the reward-predicting sensory observations and motor planning have occurred. This creates an explanatory conundrum known in the behavioral literature as the “distal reward problem” [3] and in the reinforcement learning literature as the “credit assignment problem” [6]. In order to learn to attain reward in the future, the agent must assign credit (or penalty) to the preceding actions that participated in the attainment of the initial reward. Its neural system must determine which out of all possible sensory-motor neural activation sequences determined those actions and adapt the appropriate synapses. The key problem is that the responsible sensory-motor sequence may no longer be active when the reward arrives.

The Morris Water Maze [7] is a good example of this effect, where the agent but must perform a sequence of behaviors that will, over time, lead to the hidden reward. Typically a rat is placed in a round tank that is partially filled with murky water and allowed to freely swim. A small invisible platform is submerged in the tank just below the surface of the water providing an escape from drowning (Fig. 1). Over a number of trials, rats tend to learn the location of the platform, allowing them to shorten swim times. Through distal learning processes, the reward arriving at a safe perch is associated with a temporal sequence of motor plans in the past that moved the rat from its release point to the hidden platform.

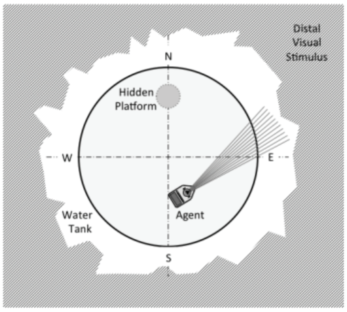


Fig. 1. An illustration of the Morris Water Maze to evaluate rodent learning. The rays emanating from the agent represent its visual field. The compass headings are classical rat release points in the tank.

Associating observations, actions, and rewards through time differs from learning problems that involve instantaneous decisions that can be thought of as sense-react behavior. In distal reward tasks, immediate cues at the time of action do not tell an agent a priori what action might lead to reward. In order to address this, episodic memories that encode temporal sequences of sensory and motor activity are required to link actions with temporally distant consequences.

Once the agent reaches the target, what neural process detects that a rewarding event has occurred? Nitz, Kargo, and Fleischer [8] present an experiment which links dopamine signaling with distal reward behavior. Dopamine receptor deficient rats exhibit impaired decision making at the choice points that are temporally distal to the reward, while normal control rats do not. This result leads to consideration of dopaminergic action as an important component of distal reward behavior.

This paper introduces a biologically motivated fully embodied self-organizing neural architecture that exhibits distal reward learning in a rudimentary form in the Morris Water Maze. By full embodiment, we mean that an agent receives information only from sensory observation of its environment and itself, and that all actions are based on internally generated motor commands. The neural architecture's operation is based on associating multiple learned sequences of bottom-up visual and somatosensory inputs, and the top-down recall of the somatosensory sequences as motor plans, reminiscent of Common Code Theory [9] notions. This recall "drives" the agent to the ultimate reward based upon a chaining-together of several previously learned motor plans. Although functionally similar to RL approaches, delayed reward behavior is accomplished here with no explicit function approximation or stochastic modeling, using only the autonomous interaction between multiple self-organizing neural modules.

The following section briefly describes the neural architecture. Section 3 presents the experimental method for testing it in a virtual water maze. Following that, Section 4 presents the results of the experiments, briefly discusses these results in the context of statistical analysis, and concludes the paper.

## 2. The Agent and Neural Architecture

A schematic of the architecture is given in Fig. 2. It is composed of a small number of subnetwork components including Fuzzy ART (FA) [10], Fuzzy LAPART [11], Temporal Integrators (TI) [12], gating logic, sensors, and actuators. FA is a self-organizing neural module that learns to classify input patterns. Fuzzy LAPART is a self-organizing neural module that learns associations between two FAs; i.e. classes of inputs are learned as well as associations between them. A TI module is a layer of leaky temporal integration nodes.

The agent has a visual sensor composed of an array of optical detectors, proprioceptive sensors that measure motor activity, touch sensors that detect contact with objects in the world, and finally motor actuators that control the speed and direction of the agent in the world. The neural architecture is divided into five major components: 1) *V*, the visual temporal sequence component, 2) *P*, the proprioceptive temporal sequence component, 3) *L*, the Fuzzy LAPART association component, 4) *D*, the dopaminergic component, and 5) *M*, the default motor plans component. The latter causes the agent to execute random path motor plans in the absence of a learned motor plan recall. The *V* and *P* components are composed of FA ↔ TI ↔ FA "sandwiches" interconnected with reciprocal connections, while the *L* component contains two laterally connected FA modules. Note that *V*, *P* and *L* are loosely analogous to visual cortex, motor cortex, and hippocampus [13] respectively. These components are interconnected with connection bundles, some of which are modulated by switches. Details of the agent and the architecture may be found in Taylor [14].

Since learning temporal sequences is central to this architecture, a more detailed explanation of the "sandwich" mentioned in the previous paragraph is in order. The dynamics of this component are as follows. As a temporal sequence of complemented coded real-valued input patterns is presented to the input of the lower FA, designated S in Fig. 2, it performs unsupervised pattern classification resulting in a sequence of activations of S-$F_2$ nodes. Note that the $F_2$ layer of an FA is winner-takes-all. The resultant S-$F_2$ activations (unit strength) are directed as input into the TI layer. When a node in the TI module receives a unit level signal, it rapidly integrates to its maximum value in one time step. When a TI node receives a zero level input, its output decays with an exponential time constant. When its output reaches a minimum noise threshold level, it is reset to zero. Each TI neuron may be thought of as an RC circuit with different rise and fall time constants. As a result, a recency gradient of activations will form across the TI layer, encoding the order of input stimulus – the most recent S class input will have the largest TI node activation, with decreasing magnitude representing the reverse order of previous S classes. At each time step, the current complement coded real-valued recency gradient is directed from TI to the input of the upper FA, designated L in Fig. 2, which again performs unsupervised pattern classification. Note that the classification process is the same in both networks, resulting in learned template patterns in S and L based on their respective input sequences. Through this hierarchical process, L-$F_2$ layer produces codes for short temporal sequences of patterns presented to S. The two FA units' vigilance parameters $\rho_S$ and $\rho_L$ control the granularity of their respective classifications.

This network may also recall short temporal sequences through a top-down process. During recall mode, an L-$F_2$ node is activated by higher-level circuitry, reading out a recency gradient stored in one of its templates $T_{Lh}$ into the L-$F_1$ layer. This gradient template is transferred through reciprocal connections from the L-$F_1$ layer to the TI layer, initializing its nodes. The TI nodes are then allowed to naturally decay over time. Through reciprocal connections from TI to the S-$F_2$ layer, S-$F_2$ nodes are stimulated in the sequence order, reading out a sequence of templates $T_{Sk}$ into the S-F1 layer. This "movie" of templates constitutes a short-term episodic recall.

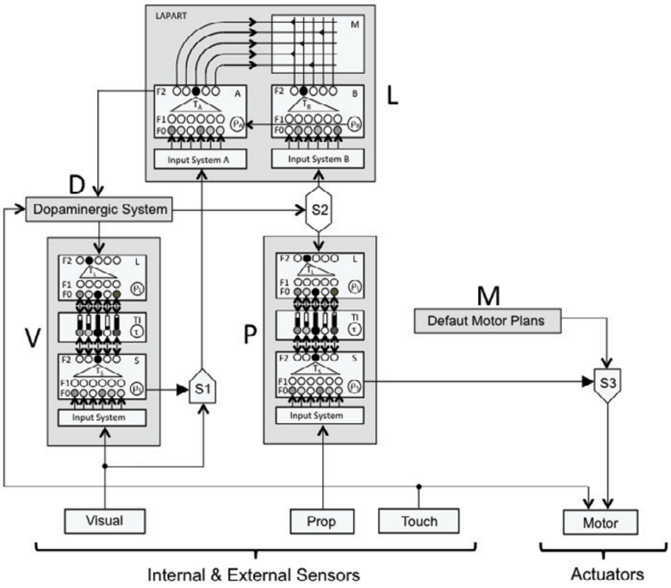


Fig. 2. The complete neural architecture. The boxes along the bottom represent sensors and actuators available to the agent. The visual sequence learning component(V), proprioceptive sequence learning component (P), dopaminergic system component (D), association component (L), and motor system (M) are connected with connection bundles indicated by the solid lines. The polygons S1, S2, and S3 are switches that modulate connection bundles. The boxes labeled *S, L, A*, and *B* are Fuzzy ART modules, *TI* are temporal integrator modules, and *M* is an associative memory modules.

Given this background, we are prepared to discuss how this architecture learns to find its target through distal reward learning. For this paper, we will use the Morris Water Maze [7] as an example scenario (Fig. 3) where the hidden platform is located at position $X_N$. We will use two notional learning trials to illustrate the processing: 1) In Trial 1, the agent is released at a point $X_{N-1}$ near the platform and moves by random chance to $X_N$, receiving the reward. 2) In Trial 2, the agent is released form a point $X_{N-2}$ a little further away from the platform and by random chance encounters the first release point $X_{N-1}$.

In Trial 1, assuming no previous learning, the agent is released from point $X_{N-1}$ and executes a random walk default motor plan gated through switch S3 that causes exploratory behavior. Fig 3 illustrates a successful random walk starting at point $X_{N-1}$ and ending at reward point $X_N$. During this walk, the *V* and *P* components are actively capturing visual and motor perceptual sequences encoded as *V* L-$F_2$ and *P* L-$F_2$ node activations.

When the agent encounters the reward at location $X_N$, the touch sensor activates the *D* component indicating the presence of a reward in the environment. The *D* component then may initiate two possible processes: 1) retrospective learning [15], and 2) prospective action. In Trial 1, the agent has encountered the reward for the first time and therefore the *D* component enables only the retrospective learning process since not action is required.

Retrospective learning process occurs in the *L* component of the architecture as follows. Due to *D* component activation, the *V* component undergoes episodic recall using the most recent *V* L-$F_2$ active node. This recall would, if allowed to completely unfold, play back a sequence of visual templates in the *V* S-$F_1$ layer representing the visual sequence from $X_{N-1}$ to $X_N$. For the purposes of learning a distal reward, only the first (oldest) S visual template is recalled and forwarded to the A-side of the *L* component through switch S1 (left center of Fig. 2). In parallel with this action, the *D* component gates the current *P* L-F2 layer code into the B-side of the *L* component through switch S2. The A and B-side FAs in the *L* component perform their self-organizing classifications, and if a lateral reset does not occur, a new lateral association in M is learned between *L* A-$F_2$ and *L* B-$F_2$. This creates an association between the visual scene at location $X_{N-1}$ and the proprioceptive motor sequence that leads from $X_{N-1}$ to $X_N$.

In Trial 2, the agent starts further away at point $X_{N-2}$ (Fig. 3) and executes a random walk motor plan that happens to move the agent to location $X_{N-1}$. Like in Trial 1, the *V* and *P* components are actively capturing visual and motor perceptual sequences during this walk. Note that the *D* component is not active during this walk, allowing raw visual patterns to be simultaneously directed to the A-side of the *L* component through switch S1 as well as to the *V* component. During most of the walk, none of these visual patterns resonates with an existing *L* A-$F_2$ template, leaving the A FA module of the *L* component inactive. However, when the walk reaches the previously visited location $X_{N-1}$, the current visual pattern will likely resonant with a template learned during the previous walk, initiating two actions: a) the *D* component is activated through a connection from the *L* A-$F_2$ layer indicating the presence of a "pseudo reward", and b) through previously learned lateral priming connections, the B FA of the *L* component will read out a learned association with a P $F_2$ node. (Note that P $F_2$ nodes encode temporal sequences of motor commands.) We refer to the action a) as detecting a pseudo reward in the sense that the agent is at least partially aroused as if an actual reward has occurred, because it now knows how to get to the reward location from this familiar location. As before, given an active *D* component, as with Trial 1, two possible processes may occur: 1) retrospective learning, and 2) prospective action.

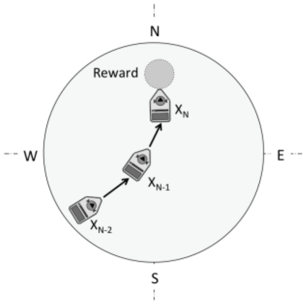


Fig. 3. An example of the agent moving between points to arrive at the reward platform. Either a learned or a default motor plan moves the agent from one point to another.

Retrospective learning in Trial 2 is identical to Trial 1. Episodic recall of the oldest *V* S-F2 template is recalled and forwarded to the *L* A-side. In parallel with this action, the *D* component gates the current *P* L-F2 layer code into the B-side of the *L* component. The A and B-side FAs in the *L* component perform their self-organizing classifications, and if a lateral reset does not occur, a lateral priming connection is learned between

$L$ A-F$_2$ and $L$ B-F$_2$. This creates an association between the visual scene at location $X_{N-2}$ and the proprioceptive plan representing the motion from $X_{N-2}$ to the pseudo reward at location $X_{N-1}$.

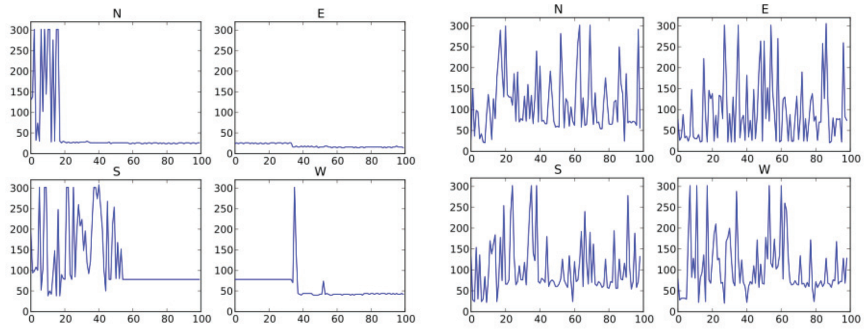
When this is complete, the architecture initiates the second process: prospective action, moving the agent from $X_{N-1}$ to $X_N$. Activation of the $P$ L-F$_2$ layer by the $L$ B module initiates episodic recall of the proprioceptive sequence that was stored when the agent traversed from $X_{N-1}$ to $X_N$ in Trial 1. A unique aspect of this approach is the use of proprioceptive recall to replay motor function, reminiscent of Common Coding Theory [9]. As this occurs, $P$ S-F2 templates are channeled to the motor actuators in the same sequence that they were learned, reproducing the motion from location $X_{N-1}$ to $X_N$, thus walking to the reward. By recursive application of Trial 2 notions with multiple agent release locations and orientations, the agent will learn to chain together trees of proprioceptive sequences to use as motor plans to move to the reward location from possibly any location in the tank.

## 3. Experiment in a virtual Morris Water Maze

There are three objects in the virtual Morris Water Maze environment in addition to the agent: i) the rim of water tank, ii) the virtual "room" surrounding the tank, and iii) the hidden platform (Fig. 1). The virtual tank is a circular region that confines the agent's range of movement as well as providing proximal visual stimulus. Three main classes of experiment were performed: a) distal reward learning, b) persistence to goal, and c) rapid transfer learning. Only the first experiment will be briefly discussed here, while the details of the other two may be found in Taylor [14].

For the experiment, one trial consists of agent release from the four release points N, E, S & W (illustrated in Fig. 3) in a block design. One hundred trials were performed for each agent. The distal reward learning experiment characterized the performance of this embodied architecture under two conditions: 1) fixed platform location and 2) randomly located platform. Moving the platform to a location chosen "at random" at each release creates an experimental condition in which little distal reward learning is expected. This provides a baseline condition for comparison.

In this experiment, the measure of performance is the sum of distance an agent travels from each of the four release point for a trial. If an individual agent is benefiting from the distal reward learning, then the distance it travels should decrease as a function of trial number. For this study, individual agents differ only in the choice of the pseudo-randomly chosen parameters associated with their default random walk motor plans. The null hypothesis is for this experiment is that agent's performance will be the same under the two experimental conditions: 1) fixed platform location and 2) randomly located platform.



(a)

(b)

Fig. 4. a) Example for one agent from fixed platform condition one showing path length data for each of the four release points verses learning trail number. Path lengths are in arbitrary units fixed by the diameter of the virtual water tank. b) Example of an agent from varying platform location condition two showing path length data for each of the four release points verses learning trail number.

## 4. Results, Discussion, and Conclusion

For this experiment, 40 agents were tested, 20 in condition one with the fixed platform location, and 20 in condition two with the platform location placed randomly per release. An example of the resulting data is given in Figure 4. Comparing Fig. 4a to Fig. 4b clear shows learning differences. ANOVA was used to compare the combined results for the two conditions giving a p-value of $1.956x10^{-4}$. This indicates that a significant amount of distal reward learning is present in this experiment. See Taylor [14].

This paper presented a fully autonomous neural architecture designed to implement distal reward behavior in embodied agents. Numerical experiments were conducted to study its performance in a virtual Morris Water Maze. Analysis of these experiments showed statistically significant distal reward behavior resulting from the processing of this architecture. Other more difficult learning tasks are planned for the future to challenge this architecture, as well are comparisons with other forms of reinforcement learning algorithms.

## Acknowledgements

The authors wish to thank Prof. Derek Hamilton at the Univ. of New Mexico, Dept. of Psychology and Dr. Michael L. Bernard of Sandia National Laboratories for their many helpful discussions and suggestions. We also wish to thank the reviewers for excellent suggestions that made the paper clearer and better grounded.

## References

[1] Hebb DO, The Organization of Behavior, John Wiley, New York, NY, 1949.

[2] Watkins CJCH (1989). Learning from Delayed Rewards. PhD Thesis, King's College, http://www.cs.rhul.ac.uk/home/chrisw/new_thesis.pdf

[3] Hull CL (1943). Principles of behavior. New York: Appleton-Century

[4] Barto, Sutton, Anderson, http://webdocs.cs.ualberta.ca/~sutton/book/ebook/the-book.html

[5] Schultz W (2007). Predictive reward signal of dopamine neurons. Neurophysiology , 80, 1-27.

[6] Sutton RS & Barto AG (1998). Reinforcement learning: an introduction. Cambridge, MA: The MIT Press.

[7] Morris RG (1984). Developments of a water-maze procedure for studying spatial learning in the rat. Journal of Neuroscience Methods , 11, 47-60.

[8] Nitz, DA, Kargo WJ, & Fleischer J (2007). Dopamine signaling and the distal reward problem. NeuroReport , 18, 1833-36.

[9] Prinz W (1984). Modes of linkage between perception and action. In W. Prinz & A.-F. Sanders (Eds.), Cognition and motor processes (pp. 185-93). Berlin: Springer.

[10] Carpenter GA, Grossberg S, and Rosen DB (1991) "Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system," *Neural Networks*, vol. 4, pp. 759–71.

[11] Healy MJ, Caudell TP & Smith SD (1997) "Acquiring rule sets as a product of learning in the logical neural architecture LAPART," *IEEE Transactions on Neural Networks*, vol. 8, no. 3, pp. 461-74.

[12] Taylor SE, Bernard ML, Verzi SJ, Morrow JD, Vineyard CM, Healy MJ and Caudell TP (2009), "Temporal Semantics: An Adaptive Resonance Theory Approach*", International Joint Conference on Neural Networks*, Atlanta, Georgia, USA, pp. 3111-17.

[13] Vineyard CM, Bernard ML, Taylor SE, Caudell TP, Watson P, Verzi S, Cohen NJ, Eichenbaum H (2010). A Neurologically Plausible Artificial Neural Network Computational Architecture of Episodic Memory and Recall. In A.V. Samsonovich, K. R. Johannsdottir, A Chella, & B. Goertzel (Eds.), Biologically Inspired Cognitive Architectures (pp. 175-180). Amsterdam, Netherlands: IOS Press BV.

[14] Taylor SE (2012). A New Class of Neural Architectures to Model Episodic Memory: Computational Studies of Distal Reward Learning. http://repository.unm.edu/, UNM Doctoral Dissertation.

[15] Caudell TP, Burch CT, Zengin M, Gauntt N, and Healy MJ (2011), "Retrospective Learning of Spatial Invariants During Object Classification by Embodied Autonomous Neural Agents", International Joint Conference on Neural Networks, San Jose, CA, pp. 2135-42.