Neurally Plausible Reinforcement Learning of Working Memory Tasks[1]

Jaldert O. Rombouts [a] Sander M. Bohte [a] Pieter R. Roelfsema [b]

[a] *CWI, Life Sciences, Amsterdam*
[b] *Netherlands Institute for Neuroscience, Amsterdam*

Abstract. By giving reward at the right times, animals like monkeys can be trained to perform complex tasks that require the mapping of sensory stimuli onto responses, the storage of information in working memory and the integration of uncertain sensory evidence. While significant progress has been made in reinforcement learning theory [2, 5], a generic learning rule for neural networks that is biologically plausible and also accounts for the versatility of animal learning has yet to be described. We propose a simple biologically plausible neural network model that can solve a variety of working memory tasks. The network is illustrated in Figure 1: as output, the network predicts action-values (Q-values) for different possible actions [2], and it learns to minimize SARSA [2] temporal difference (TD) prediction errors by stochastic gradient descent. In the hidden layer, the model has both standard neural units as well as memory units. Memory units are inspired by neurons in lateral intraparietal (LIP) cortex and prefrontal cortex that exhibit persistent activations for task related cues in visual working memory tasks [1]. In the model, the memory units learn to represent an internal state that allows the network to solve working memory tasks by transforming POMDPs into MDPs. The updates for synaptic weights have two components. The first is a synaptic tag that arises from an interaction between feedforward and feedback activations. Tags form on those synapses that are responsible for the chosen actions by an attentional feedback process [6]. The second factor is a global neuromodulatory signal that reflects the TD error, and this signal interacts with the tags to yield synaptic plasticity. TD-errors are represented by dopamine neurons [5]. The persistence of tags permits learning if time passes between synaptic activity and the animals choice, for example if information is stored in working memory or evidence accumulates before a decision is made. The learning rules are biologically plausible because the information required for computing the synaptic updates is available at the synapse. We call the new learning scheme AuGMEnT (Attention-Gated MEmory Tagging).

We show that AuGMEnt explains how neurons in association cortex learn to temporarily store task-relevant information in non-linear stimulus-response mapping tasks (Shown in Figure 2, [1, 3]). The memory saccade/anti-saccade task (Fig. 2A) is based on [3]. This task requires a non-linear transformation and cannot be solved by a direct mapping from sensory units to Q-value units. Trials started with an empty screen, shown for one time step. Then either a black or white fixation mark was shown indicating a pro-saccade or anti-saccade trial, respectively. The model had to fixate on the fixation mark within ten time-steps, or the trial was terminated. After fixating for two timesteps, a cue was presented on the left or right. The cue was shown for one time-step, and then only the fixation mark was visible for two time-steps before turning off. In the pro-saccade condition, the offset of the fixation mark indicated that the model should make an eye-movement towards the cue location to collect a reward. In the anti-saccade condition, the model had to make an eye-movement away from the cue location. The input to the model (Fig. 2B) consisted of four binary variables representing the information on the virtual screen; two for the fixation marks and two for

[1] Appeared as "Neurally Plausible Reinforcement Learning of Working Memory Tasks", J.O. Rombouts, S.M. Bohte & P.R. Roelfsema, *Advances in Neural Information Processing (NIPS) 25*, 2012, pp 1880–1888.

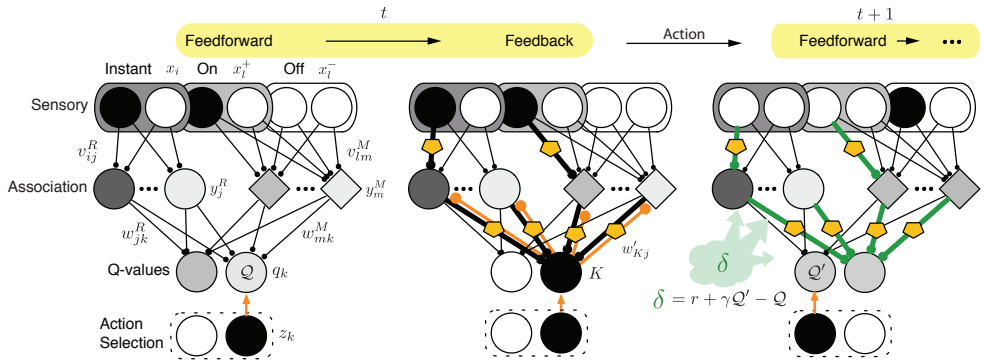

Figure 1: AuGMEnT model and learning. Pentagons represent synaptic tags.

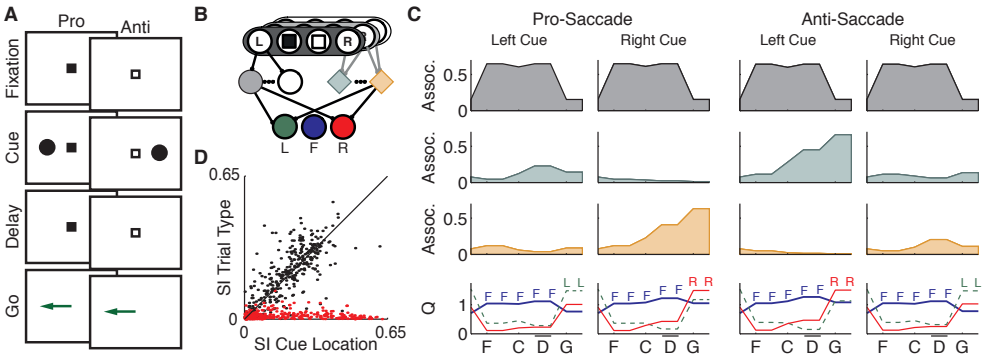

Figure 2: A Memory saccade/antisaccade task. B Model network. In the association layer, a regular unit and two memory units are color coded gray, green and orange, respectively. Output units L,F,R are colored green, blue and red, respectively. C Unit activation traces for a sample trained network. Symbols in bottom graph indicate highest valued action. F, fixation onset; C, cue onset; D, delay; G, fixation offset (Go signal). Thick blue: fixate, dashed green: left, red: right. D Selectivity indices of memory units in saccade/antisaccade task (black) and in pro-saccade only task (red).

the cue location, and $+/-$ derivative cells that encode positive and negative changes in sensory inputs as input to the memory units. The activity of a trained network is illustrated in Fig. 2C. The Q-unit for fixating at the center had strongest activity at fixation onset and throughout the fixation and memory delays, whereas the Q-unit for the appropriate eye movement became more active after the go-signal. This activity derives from memory units in the association layer that maintain a trace of the cue as persistent elevation of their activity and are also tuned to the difference between pro- and antisaccade trials (Fig. 2D). Additionally, AuGMent can also learn to integrate probabilistic evidence for perceptual decision making [4] (see full paper). Together, these experiments show that by including memory units as an integral part of the neural network, AuGMEnt can learn to solve difficult POMDP tasks that require learning "information states" in the neural network's hidden layer.

References

[1] Gnadt, J. and Andersen, R. A. Memory Related motor planning activity in posterior parietal cortex of macaque. Experimental Brain Research, 70(1):216220, 1988.

[2] Sutton, R. S. and Barto, A. G. Reinforcement learning. MIT Press, Cambridge, MA, 1998.

[3] Gottlieb, J. and Goldberg, M. E. Activity of neurons in the lateral intraparietal area of the monkey during an antisaccade task. Nature neuroscience, 2(10):90612, 1999.

[4] Yang, T. and Shadlen, M. N. Probabilistic reasoning by neurons. Nature, 447:107580, 2007.

[5] Montague, P. R., et al. Computational roles for dopamine in behavioural control. Nature, 431:7607, 2004.

[6] Roelfsema, P. R. and van Ooyen, A. Attention-Gated Reinforcement Learning of Internal Representations for Classification. Neural Computation, 2214(17):21762214, 2005.