Supplementary Material for Continual Reinforcement Learning with Complex Synapses

1 Experimental details

Tables of parameters for both the tabular and deep Q-learning experiments are shown below.

Table 1: Parameter values for Tabular Q-learning experiments

Parameter	Value
# Epochs	24
# Episodes/Epoch	10000
Max # steps per Episode	20000
γ	0.9
λ	0.9
ϵ	0.05
Learning rate	0.1
Grid size	10x10
# Benna-Fusi variables	3
Benna-Fusi $g_{1,2}$	10^{-5}
Elig. trace scale factor*	10

**Multiple of eligibility trace that flow between beakers is scaled by in modified Benna-Fusi model*

Table 2: Parameter values for Deep RL experiments

Parameter	Multi-task	Single task
# Epochs	40	1
# Episodes/Epoch	20000	100000
Max # time steps / episode	500	500
Cart-Pole γ	0.95	0.95
Catcher γ	0.99	0.99
Initial ϵ (Epoch start)	1	1
ϵ-decay / episode	0.9995	0.9995
Minimum ϵ	0	0
Neuron type	ReLU	ReLU
Width hidden layer 1	400	100
Width hidden layer 2	200	50
Optimiser	Adam	Adam
Learning rate	10^{-3} to 10^{-6}	10^{-3} to 10^{-6}
Adam β_1	0.9	0.9
Adam β_2	0.999	0.999
Experience replay size	2000	1
Replay batch size*	64	1
Soft target update τ	0.01	0.01
Soft Q-learning α	0.01	0.01
# Benna-Fusi variables	30	30
Benna-Fusi $g_{1,2}$	0.001625	0.01
Test Frequency (Episodes)	10	10

**Updates were made sequentially as in stochastic gradient descent, not all in one go as a minibatch.*

2 Additional Experiments

2.1 Varying Epoch Lengths

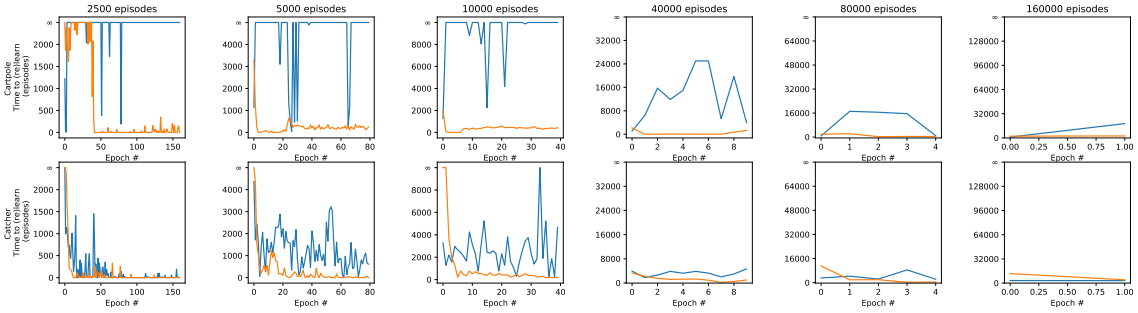

Figure 1: Comparison of time to (re)learn each task in the control agent (blue) and the Benna-Fusi agent (orange) for different epoch lengths. Both agents had a learning rate of 0.001 and the runs with longer epochs were run for fewer epochs. In all cases the Benna-Fusi agent becomes quicker (or in a couple of instances equally quick) at relearning each task than the control agent, demonstrating the Benna-Fusi model's ability to improve memory at a range of timescales.

2.2 Three-task experiments

In order to ensure that the benefits of the Benna-Fusi model were not limited to the two-task setting, we introduced a new task and ran experiments where training was rotated over the three tasks. The new task was a modified version of Cart-Pole where the length of the pole is doubled (dubbed Cart-PoleLong); our criterion for judging that this task was different enough to Cart-Pole to be considered a new task was that when trained sequentially after Cart-Pole in a control agent, it subsequently led to catastrophic forgetting of its policy for the Cart-Pole task.

Figure 2 shows the remembering times for each task for a control agent and a Benna-Fusi agent when training was rotated over the three tasks (Cart-PoleLong $->$ Catcher $->$ Cart-Pole) over a total of 24 epochs. The results indicate that the Benna-Fusi model exhibits the same benefits as in the two-task setting.

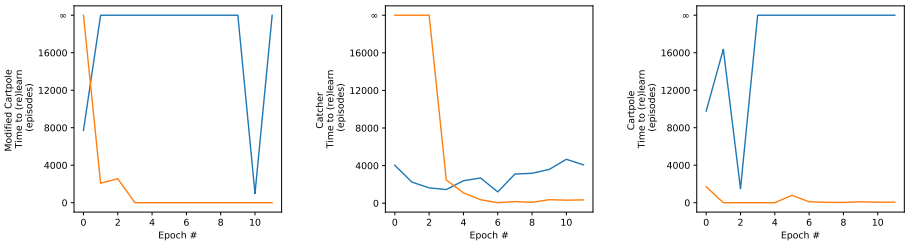

Figure 2: Comparison of time to (re)learn each task in the control agent (blue) and the Benna-Fusi agent (orange) for the three different tasks. Each epoch was run for 20000 episodes and both agents had a learning rate of 0.001. While the Benna-Fusi agent took a little longer to learn Catcher than the control agent, by the end of the simulation the Benna-Fusi agent could learn to recall each task much faster than the control.

2.3 Varying size of replay database

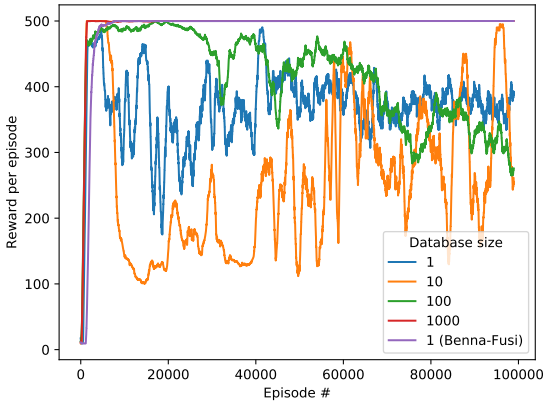

Figure 3: 100 test-episode moving average of reward in Cart-Pole for control agents (all with $\eta = 0.001$) with different sized experience replay databases and the Benna-Fusi agent in just the online setting. For these experiments, 1 experience was sampled for training from the database after every time step. In the control cases, when the database is too small, the agent can not attain a stable performance on the task while the Benna-Fusi agent can.

2.4 Catcher single task

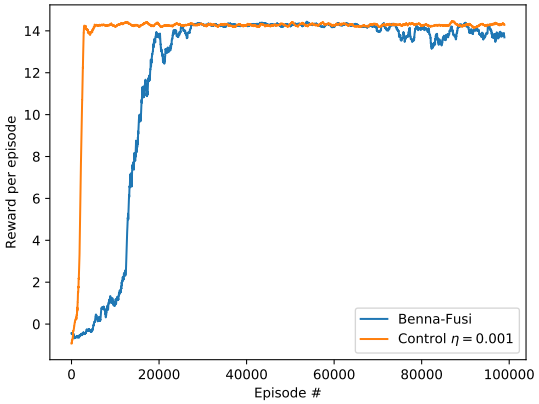

Figure 4: The 100 test-episode moving average of reward per episode in Catcher for the Benna-Fusi agent and the best control agent. The control agent learns faster but both end up learning a good policy.

2.5 Varying final exploration value

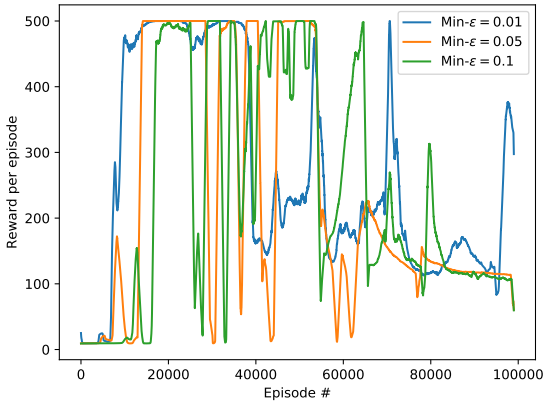

Figure 5: The 100 test-episode moving average of reward per episode in Cart-Pole for control agents where epsilon was not allowed to decay below different minimum values. None of the runs yielded a good stable performance.