Learning Algorithm

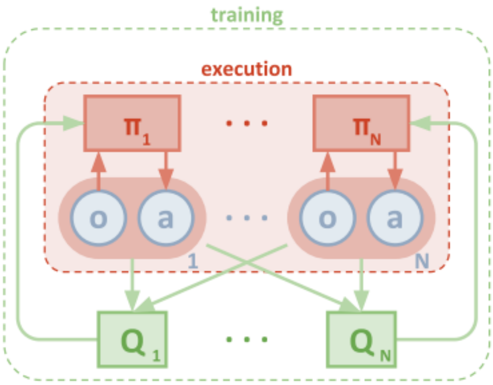


Figure 1 Overview of multi-agent de-centralized actor and centralized critic approach. R Lowe et al 2017 [1]

For tennis project, multi-agent actor-critic (MADDPG) has been implemented with two variations:

Normal replay buffer and prioritized replay buffer method as replay buffer is unable to achieve convergence of average reward +0.5 over 100 episodes.

Main idea behind this MADDPG in comparison with DDPG is with multiple agent sharing the same critic network (shared policy between multiple agent) and localized agent update on actor for predicted action as shown in Figure 1.

	Actor			**Critic**		
	Activation	**In node**	**Out Node**	**Activation**	**In node**	**Out Node**
Input Layer	**Relu**	48	256	**Relu**	48	260
Hidden Layer	**Relu**	256	128	**Relu**	260	128
Output Layer	**Tanh**	128	2		128	1

Table 1 Actor and Critic setup for MADDPG

Actor and critic network setup as Table 1 above.

Results

Initial setup was using replay buffer with above setting to run for 2000 episodes. The last 100 episodes as shows in Figure 2 does not converge to more than 0.1 reward even after 1900 episodes of training.

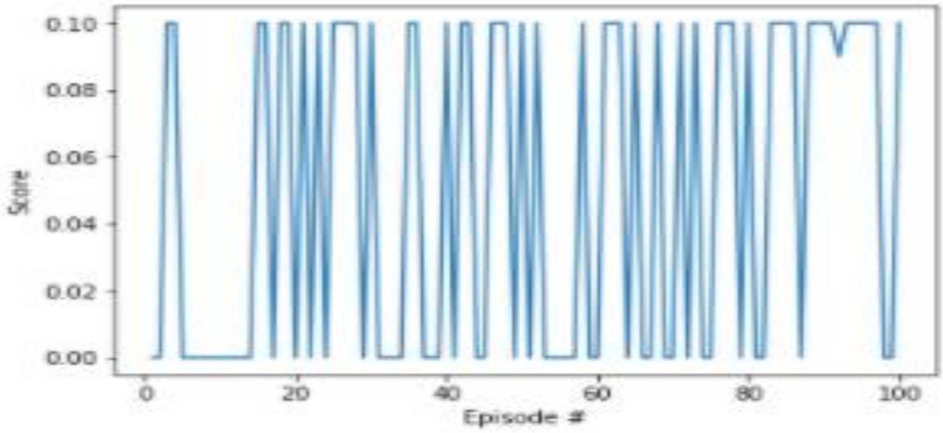
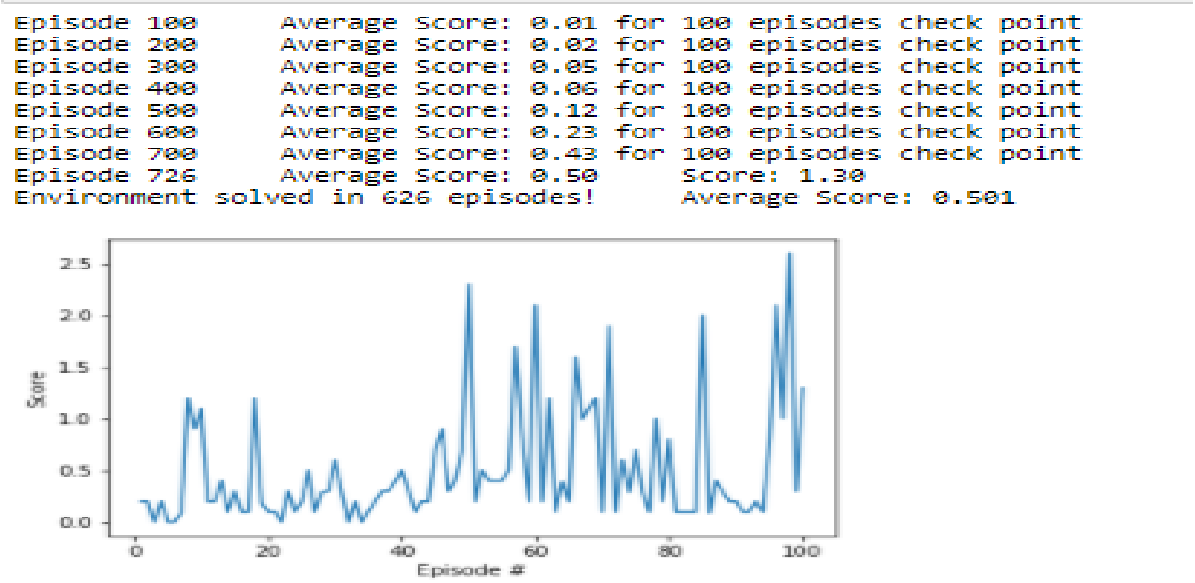


Figure 2 Replay buffer agent unable to converge

Without successful run of achieve +0.5 rewards , prioritized replay buffer was implemented to improve the MADDPG further. In Figure 3 , with prioritized replay buffer in MADDPG is able to achieve +0.5 rewards over 100 episodes with 726 episodes played by tennis agent.



```
Episode 100      Average Score: 0.01 for 100 episodes check point
Episode 200      Average Score: 0.02 for 100 episodes check point
Episode 300      Average Score: 0.05 for 100 episodes check point
Episode 400      Average Score: 0.06 for 100 episodes check point
Episode 500      Average Score: 0.12 for 100 episodes check point
Episode 600      Average Score: 0.23 for 100 episodes check point
Episode 700      Average Score: 0.43 for 100 episodes check point
Episode 726      Average Score: 0.50     Score: 1.30
Environment solved in 626 episodes!      Average Score: 0.501
```

Figure 3 Prioritized Replay Buffer in MADDPG

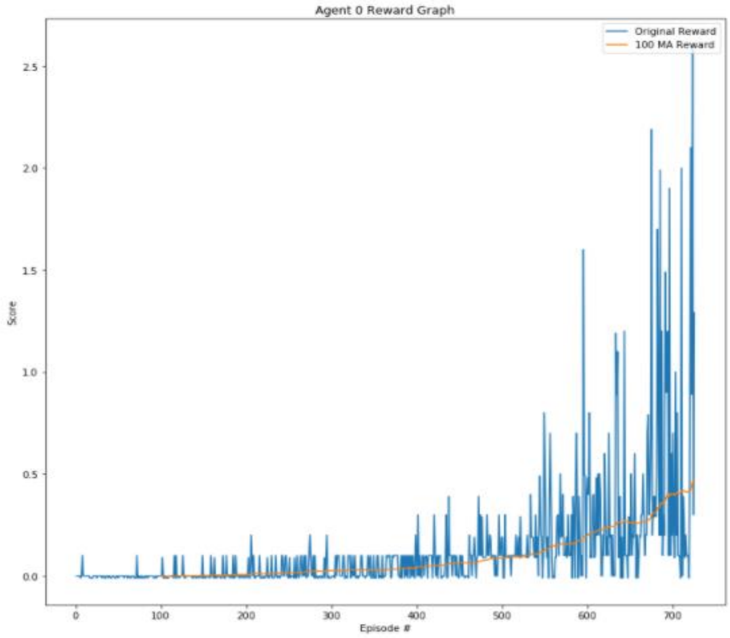


Figure 4 Agent 0 on tennis game reward graph over 726 episodes

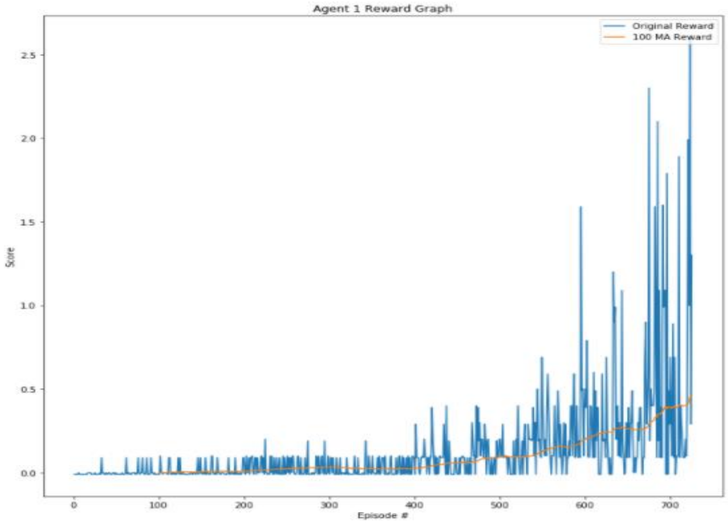


Figure 5 Agent 1 tennis game reward graph over 726 episodes

Figure 4 and Figure 5 separately show the reward over episodes played by agent. With the centralized critic method from MADDPG, agent able to improve with similar performance shown from 726 episodes plotted above.

Objective of the project has been achieved by hitting +0.5 rewards over average of 100 episodes using MADDPG implemented.

Future Work

Self-play using monte-carlo tree search that was implemented in Tic-Tac-Toe example can be future implementation method that able to achieve target more effectively. In comparison with MADDPG , MCTS could be a more stable solution.

In a recent article of self-play in maze from Baker et al [2] , agent is able to achieve superior performance in hiding from the seeker. Using method suggested in paper [2] could be potential future work to further improve self -play of agent.

References

[1] Lowe, R., Wu, Y.I., Tamar, A., Harb, J., Pieter Abbeel, O. and Mordatch, I., 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, *30*, pp.6379-639

[2] Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B. and Mordatch, I., 2019. Emergent tool use from multi-agent autocurricula. *arXiv preprint arXiv:1909.0752*