Udacity Continuous Control project

Prepared for: Udacity Deep Reinforcement Learning Nano Degree
Prepared by: Divyanshu Kalra
18 March 2019

SUMMARY

Algorithm

The algorithm used to solve the problem is Twin Delayed DDPG this is because it overcomes the limitations of vanilla DDPG.

Explanation:

Unlike DDPG TD3 simultaneously learns two Q-functions, Q_ϕ1 and Q_ϕ2, by mean square Bellman error minimisation. The Q- functions learnt by comparing MSE on the same target as shown Equation 1 and this target is decided by choosing minimum between the two as shown in Equation 2 but the policy is learnt by maximising Q_ϕ1 similar that of DDPG as shown in Equation 3.

$$y(r, s', d) = r + \gamma(1 - d) \min_{i=1,2} Q_{\phi_{i,\text{targ}}}(s', a'(s')),$$

Equation 1.

$$L(\phi_1, \mathcal{D}) = \underset{(s,a,r,s',d)\sim\mathcal{D}}{\mathrm{E}} \left[\left(Q_{\phi_1}(s, a) - y(r, s', d) \right)^2 \right]$$

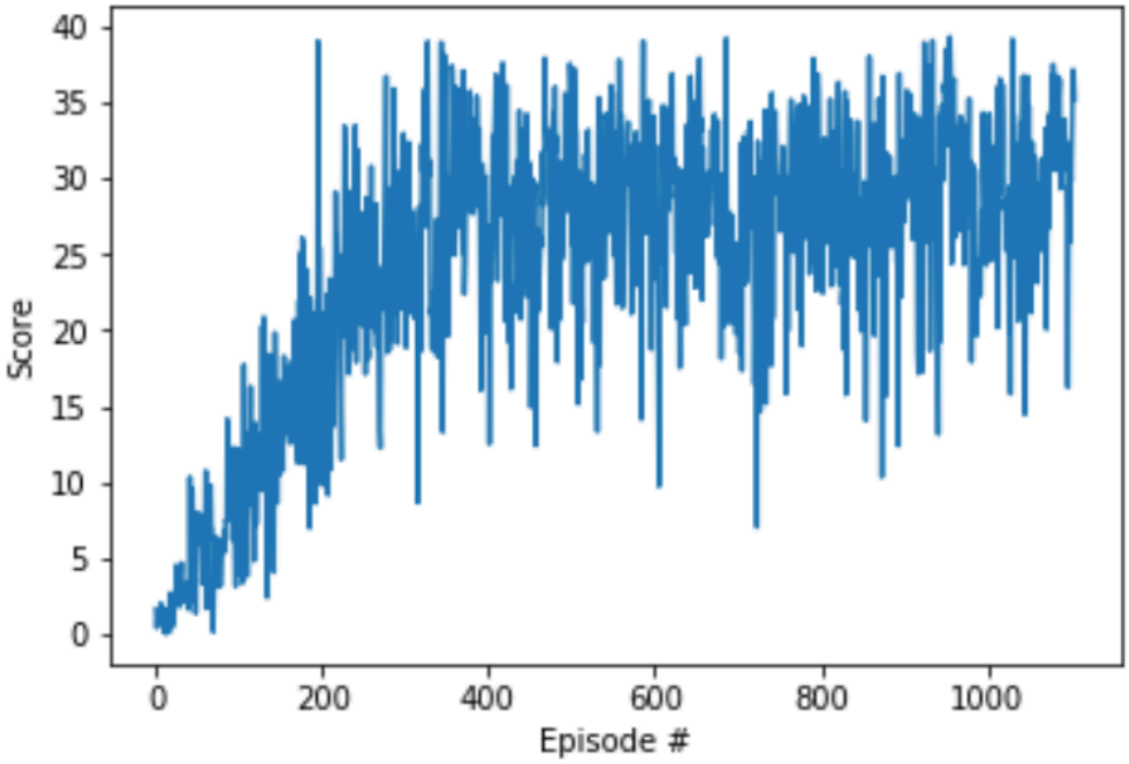
$$L(\phi_2, \mathcal{D}) = \underset{(s,a,r,s',d)\sim\mathcal{D}}{\mathrm{E}} \left[\left(Q_{\phi_2}(s, a) - y(r, s', d) \right)^2 \right]$$

Equation 2.

$$\max_{\theta} \underset{s\sim\mathcal{D}}{\mathrm{E}} \left[Q_{\phi_1}(s, \mu_\theta(s)) \right]$$

Equation 3.

REWARD PLOT



As it can be seen in about 400 epochs the average reward was more than +20, hence the environment was solved.

FUTURE IDEAS

1) Implementing Soft Actor Critic
2) Implementing MultiAgent DDPG