1. Learning Architecture
    a) Algorithm

Algorithm 1 DDPG algorithm

Randomly initialize critic network $Q(s,a|\theta^Q)$ and actor $\mu(s|\theta^\mu)$ with weights θ^Q and θ^μ.
Initialize target network Q' and μ' with weights $\theta^{Q'} \leftarrow \theta^Q$, $\theta^{\mu'} \leftarrow \theta^\mu$
Initialize replay buffer R
for episode = 1, M **do**
  Initialize a random process $\mathcal{N}$ for action exploration
  Receive initial observation state s_1
  for t = 1, T **do**
    Select action $a_t = \mu(s_t|\theta^\mu) + \mathcal{N}_t$ according to the current policy and exploration noise
    Execute action a_t and observe reward r_t and observe new state s_{t+1}
    Store transition (s_t, a_t, r_t, s_{t+1}) in R
    Sample a random minibatch of N transitions (s_i, a_i, r_i, s_{i+1}) from R
    Set $y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'})$
    Update critic by minimizing the loss: $L = \frac{1}{N}\sum_i(y_i - Q(s_i, a_i|\theta^Q))^2$
    Update the actor policy using the sampled policy gradient:

$$\nabla_{\theta^\mu} J \approx \frac{1}{N}\sum_i \nabla_a Q(s,a|\theta^Q)|_{s=s_i,a=\mu(s_i)} \nabla_{\theta^\mu}\mu(s|\theta^\mu)|_{s_i}$$

    Update the target networks:

$$\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}$$
$$\theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}$$

  end for
end for

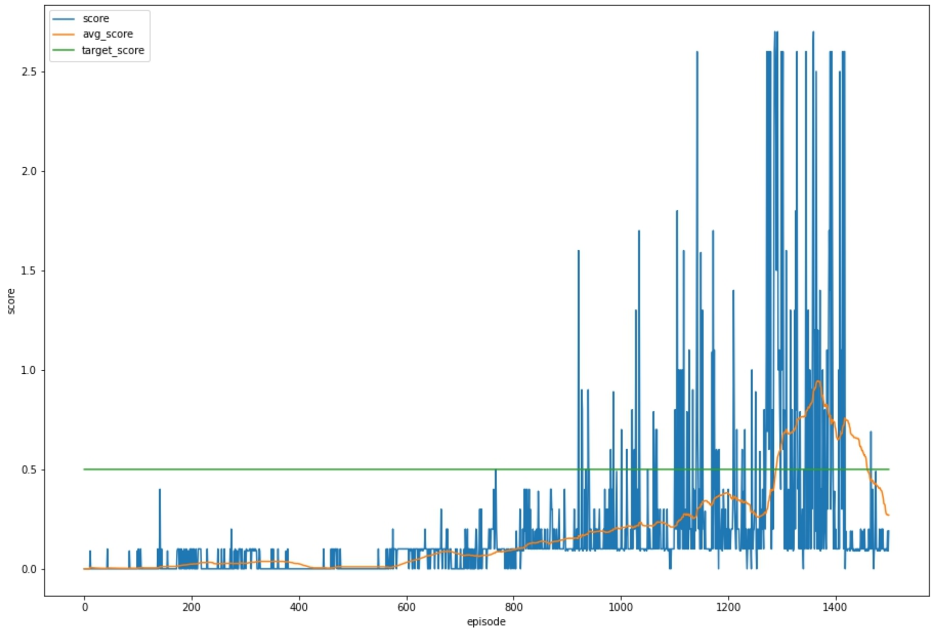
---

    b) Model Structure
        i. Actor:
            1. Input size = State size = 24
            2. Hidden layers(2)
                a) Fully connected with 128 batch-normalized rectifiers
                b) Fully connected with 256 rectifiers
            3. Output size = Action size = 2
        ii. Critic:
            1. Input 1 size = State size = 24 at 1st layer
            2. Input 2 size = Action size = 2 concat to 2nd layer
            3. Hidden layers(2)
                a) Fully connected with 128 batch-normalized rectifiers
                b) Fully connected with 256+4 rectifiers
            4. Output size = Value function = 1
    c) Hyperparameters(Final)

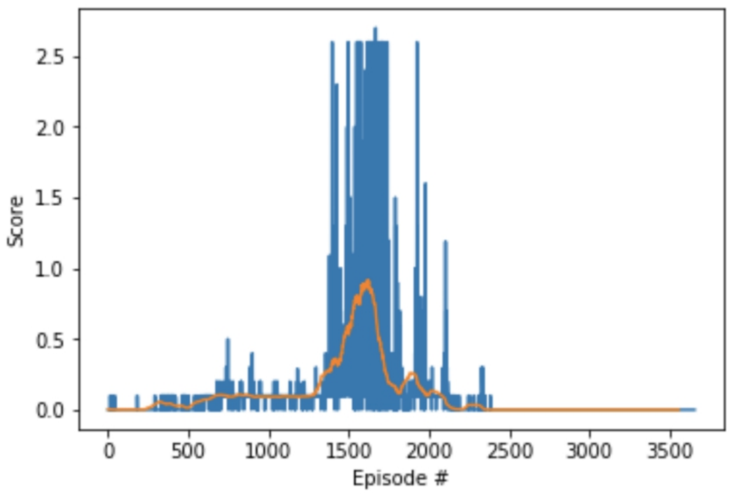
i.	Batch size	40
ii.	Memory buffer size	1e6
iii.	Number of episodes	1500
iv.	Target score	30.0
v.	Discount factor gamma	1e-3
vi.	Learning rate for Actor	1e-4
vii.	Learning rate for Critic	1e-3

viii. Update Period 5
ix. Update Times per update 10
x. Weight Decay 0
xi. Agent number 20
xii. Alpha(prioritized exp replay) 0.7
xiii. Beta(prioritized exp replay) 0.8

2. Results
   a) DDPG with prioritized experience replay
      i. Folder: ddpg_result_2023_05_21_22_08_57
      ii. Result:



3. Conclusions
   a) The prioritized exp replay method do have accelerations on training(<1300 episodes), comparing to the original DDPG baseline provided by course instructions below(>1500 episodes).



4. Future Improvements
    a) Will try to implement n-step bootstrapping
    b) Will try to use array to represent replay buffer’s binary tree
    c) Will try to implement other algorithms like Reinforce and TRPO