Continuous control Project

Yunze Han

UDACITY

DEEP REINFORCEMENT LEARNING NANODEGREE
UDACITY

September 14, 2021

Content

1 Project description **1**
1.1 Environment 1
1.2 Learning algorithm 1

2 Plot of rewards **3**

3 Ideas of future works **4**

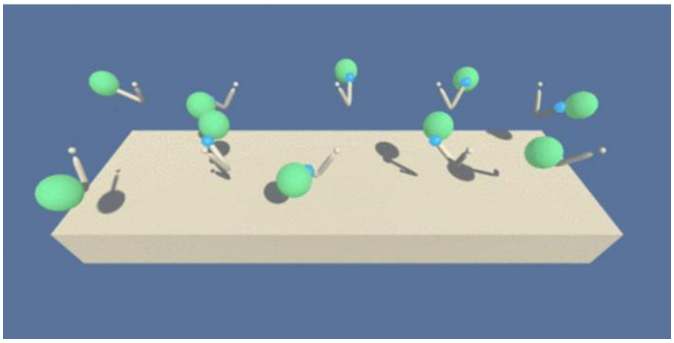
Part 1

Project description

1.1 Environment

In this environment, a double-jointed arm can move to target locations. A reward of +0.1 is provided for each step that the agent's hand is in the goal location. Thus, the goal of your agent is to maintain its position at the target location for as many time steps as possible.

The observation space consists of 33 variables corresponding to position, rotation, velocity, and angular velocities of the arm. Each action is a vector with four numbers, corresponding to torque applicable to two joints. Every entry in the action vector should be a number between -1 and 1.



1.2 Learning algorithm

DDPG (Lillicrap, et al., 2015), short for Deep Deterministic Policy Gradient, is a model-free off-policy actor-critic algorithm, combining DPG with DQN. Recall that DQN (Deep Q-Network) stabilizes the learning of Q-function by experience replay and the frozen target network. The original DQN works in discrete space, and DDPG extends it to continuous space with the actor-critic framework while learning a deterministic policy.

Actor:

1. Fc1: input 33 output 400 activation function: ReLu
2. Fc2: input 400 output 300 activation function: ReLu
3. Fc3: input 300 output 4 activation function: tanh

Critic:

1. Fc1: input 33 output 400 activation function: ReLu
2. Fc2: input 400+action_size output 300 activation function: ReLu
3. Fc3: input 300 output 1

The training hyperparameters:

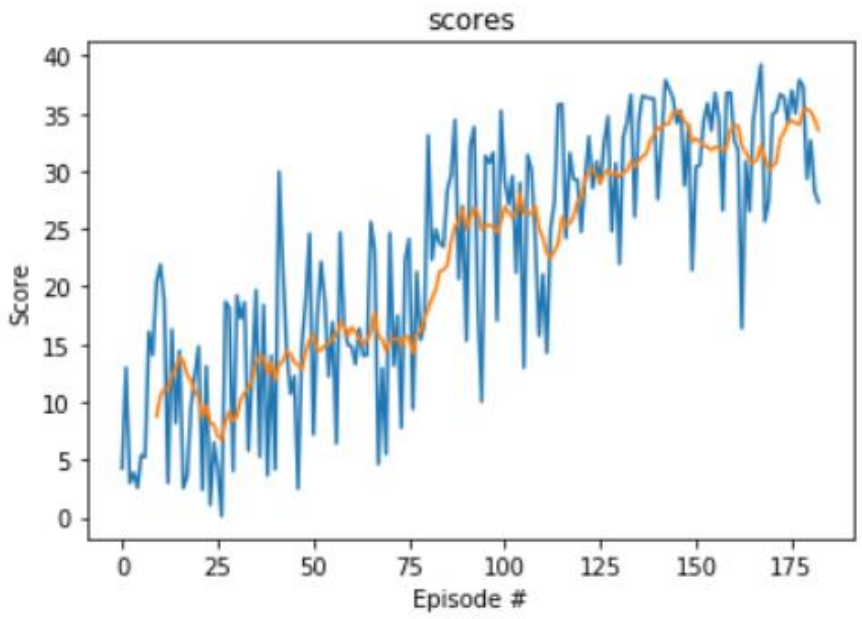
```
BUFFER_SIZE = int(1e6)  # replay buffer size
BATCH_SIZE = 128        # minibatch size
GAMMA = 0.99            # discount factor
TAU = 5e-3              # for soft update of target parameters
LR_ACTOR = 5e-4         # learning rate of the actor
LR_CRITIC = 5e-4        # learning rate of the critic
WEIGHT_DECAY = 0        # L2 weight decay
NOISE_SD = 0.10         # noise scale
```

Part 2

Plot of rewards

- save the model as ‘checkpoint.pth’
- Environment solved in 183 episodes! Average Score: 30.02

Here is the graph of the score evolution:



Part 3

Ideas of future works

To improve the efficiency of experience replay in DDPG method, we can replace the original uniform experience replay with prioritized experience replay.