CS294-112 Assignment 3: Q-Learning and Actor-Critic

Fan Zhang

October 2018

1 Part 1: Q-Learning

1.1 Question 1: basic Q-learning performance

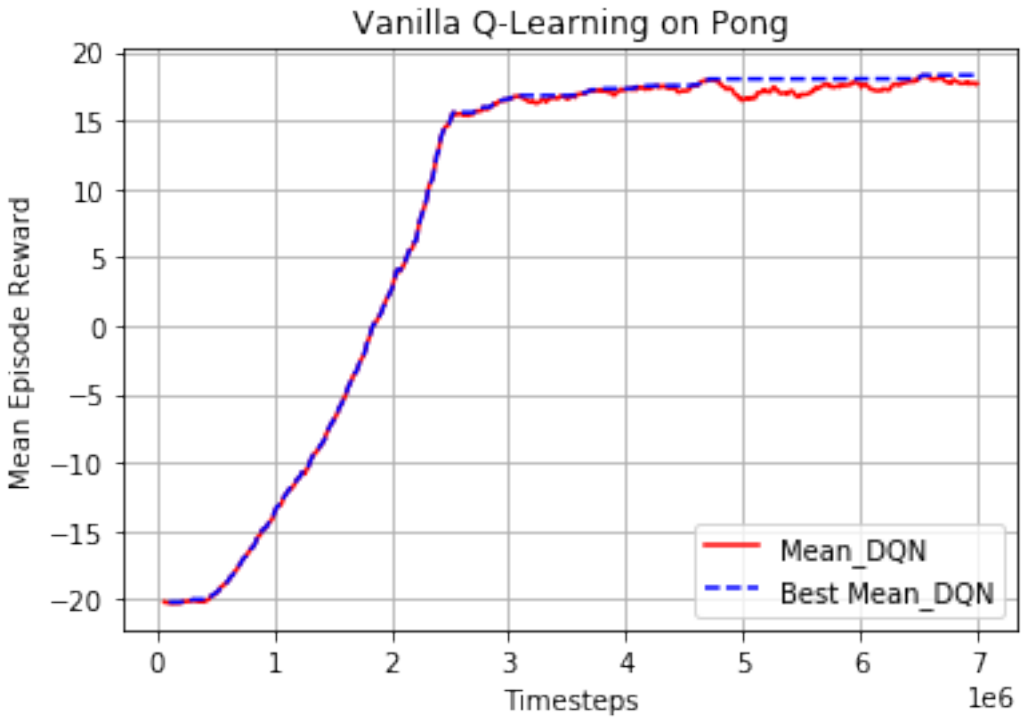

Figure 1: Learning curves (mean 100-episode reward and best mean reward with timesteps to 7m) for the experiments with Pong in Atari

1.2 Question 2: double Q-learning

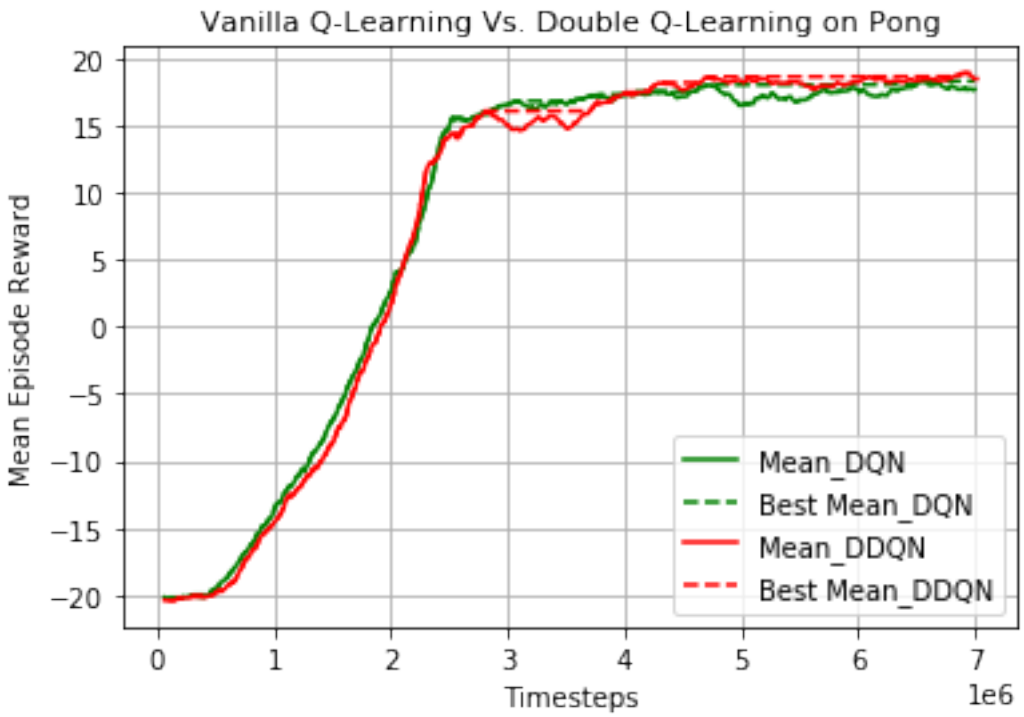

Figure 2: Vanilla Q-learning and Double Q-learning on Pong in Atari. After 7m timesteps, the best mean reward of Double Q-learning is 18.93 while the best mean reward of Vanilla Q-learning is 18.31. The double Q-learning is a little bit good than Vanilla Q-learning and the result is consistent with the result of [1]

1.3 Question 3: experimenting with hyperparameters

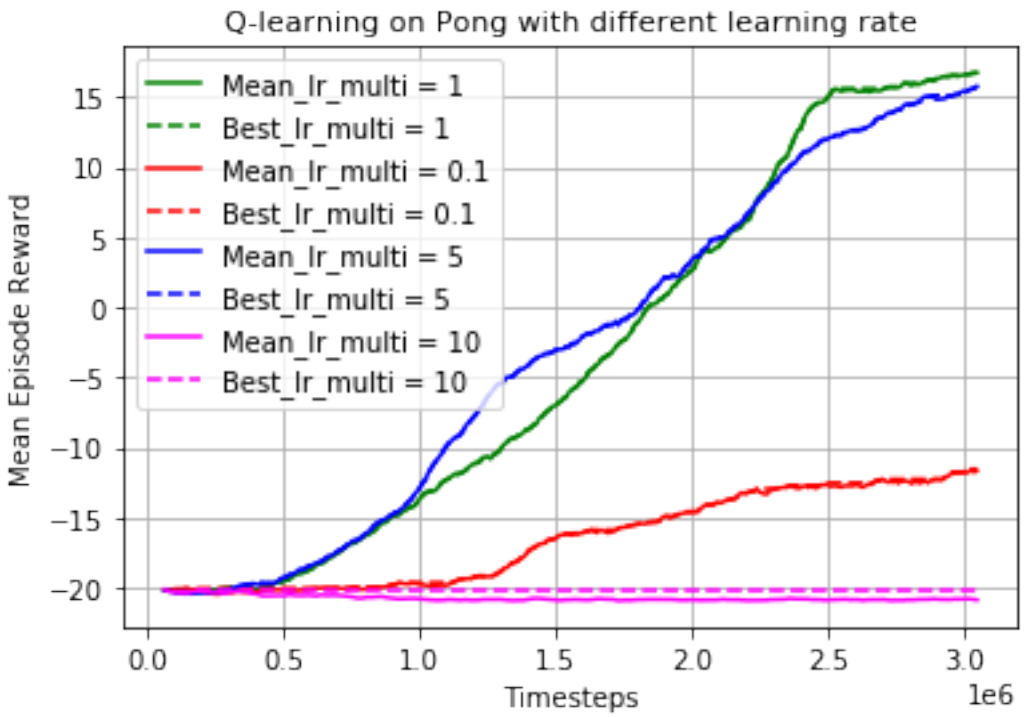

Figure 3: The learning curve with hyperparameter learning rate. Since the learning rate of the Pong is defined by learning multiplier, the default *lr_multiplier*is 1 and here I tested 0.1, 5 and 10. From the plot that when learning multiplier is 10, the learning rate is too big that the training can't converge; When the learning multiplier is 5, the converge speed is faster than default 1, and when learning multiplier is 0.1, the converge rate is slow so that when timesteps arrive 3m, the reward is stll around -10

2 Part 2: Actor-Critic

2.1 Question 1: Sanity check with Cartpole

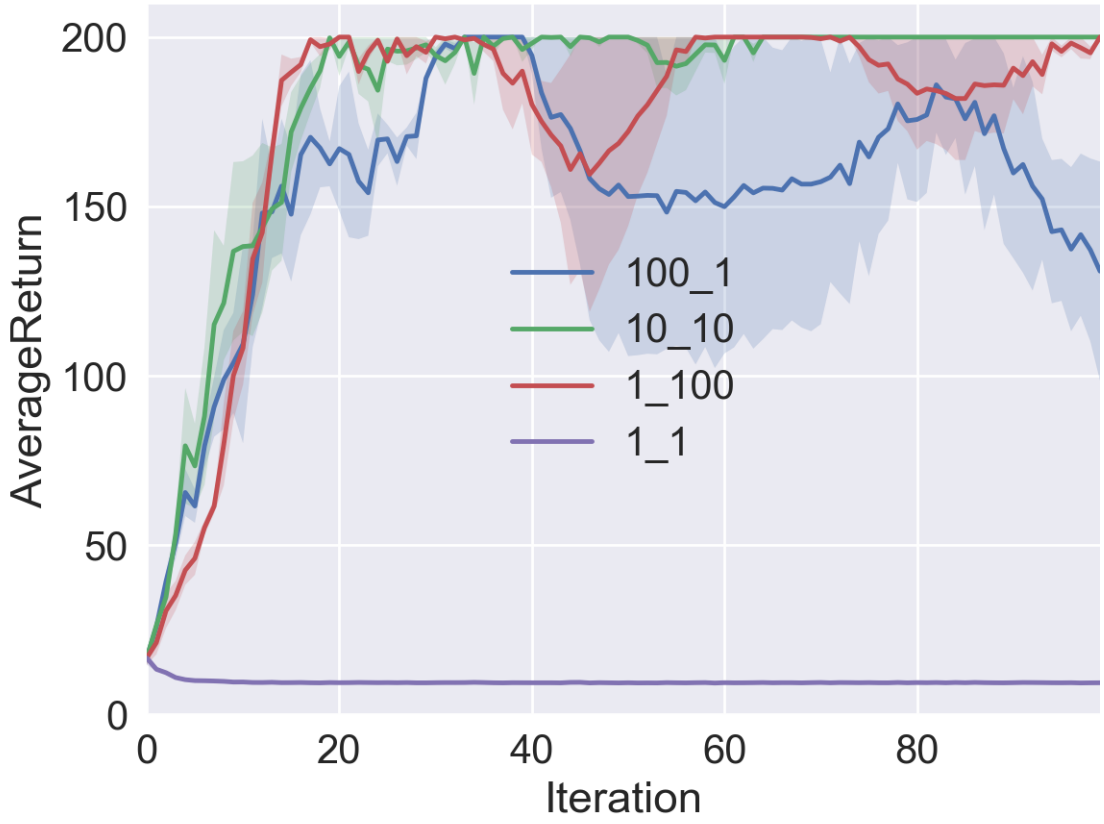

Figure 4: Implement the actor-critic to Cartpole-v0 with different target updates steps and gradient updates steps. And the 10_10 works best. With only 1_1 updating step there is no effect on it because the critic network and actor network only update once. And if 1_100 and 100_1 then the variance is large. So an suitable updating steps number of 10_10 works well with critic network and actor network

```
python train_ac_f18.py CartPole-v0 -n 100 -b 1000 -e 3 --exp_name 1_1 -ntu 1 -
    ngsptu 1
python train_ac_f18.py CartPole-v0 -n 100 -b 1000 -e 3 --exp_name 1_100 -ntu 1
     -ngsptu 100
python train_ac_f18.py CartPole-v0 -n 100 -b 1000 -e 3 --exp_name 100_1 -ntu
    100 -ngsptu 1
python train_ac_f18.py CartPole-v0 -n 100 -b 1000 -e 3 --exp_name 10_10 -ntu
    10 -ngsptu 10
```

2.2 Question 2: Run actor-critic with more difficult tasks

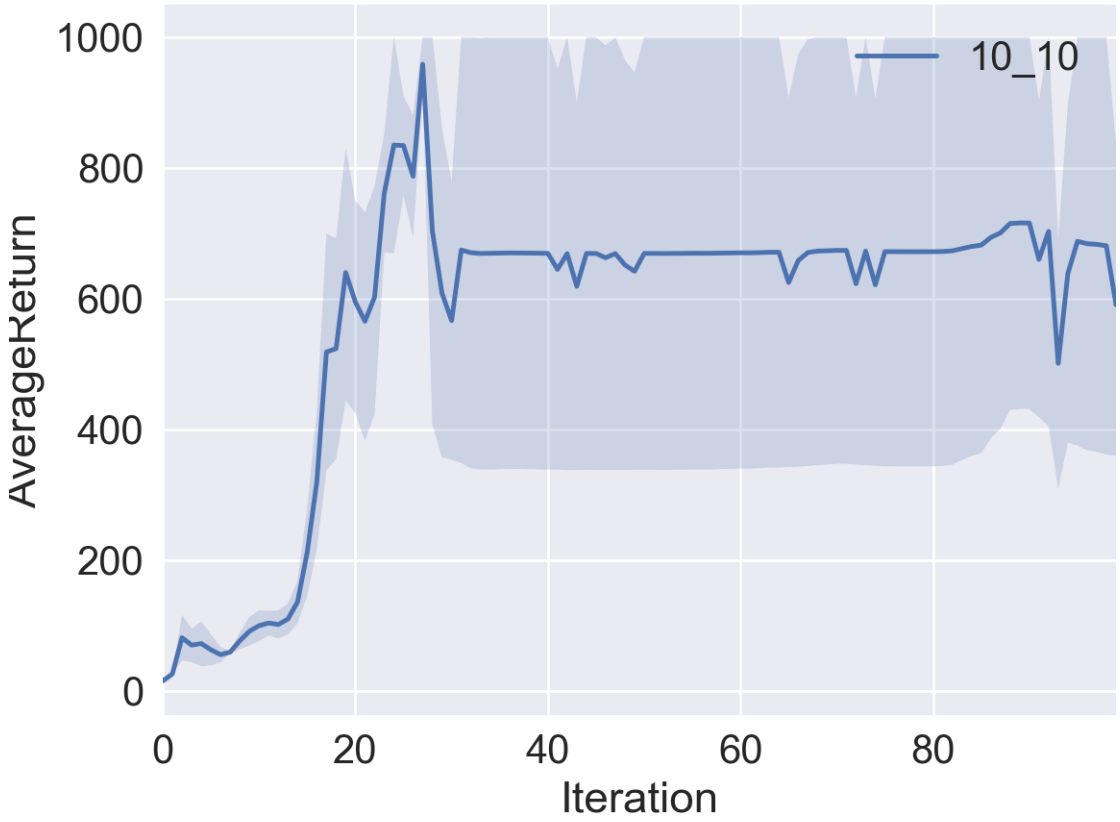

Figure 5: Implement actor-critic to InvertedPendulum task

```
python train_ac_f18.py InvertedPendulum-v2 -ep 1000 --discount 0.95 -n 100 -e
    3 -l 2 -s 64 -b 5000 -lr 0.01 --exp_name 10_10 -ntu 10 -ngsptu 10
```

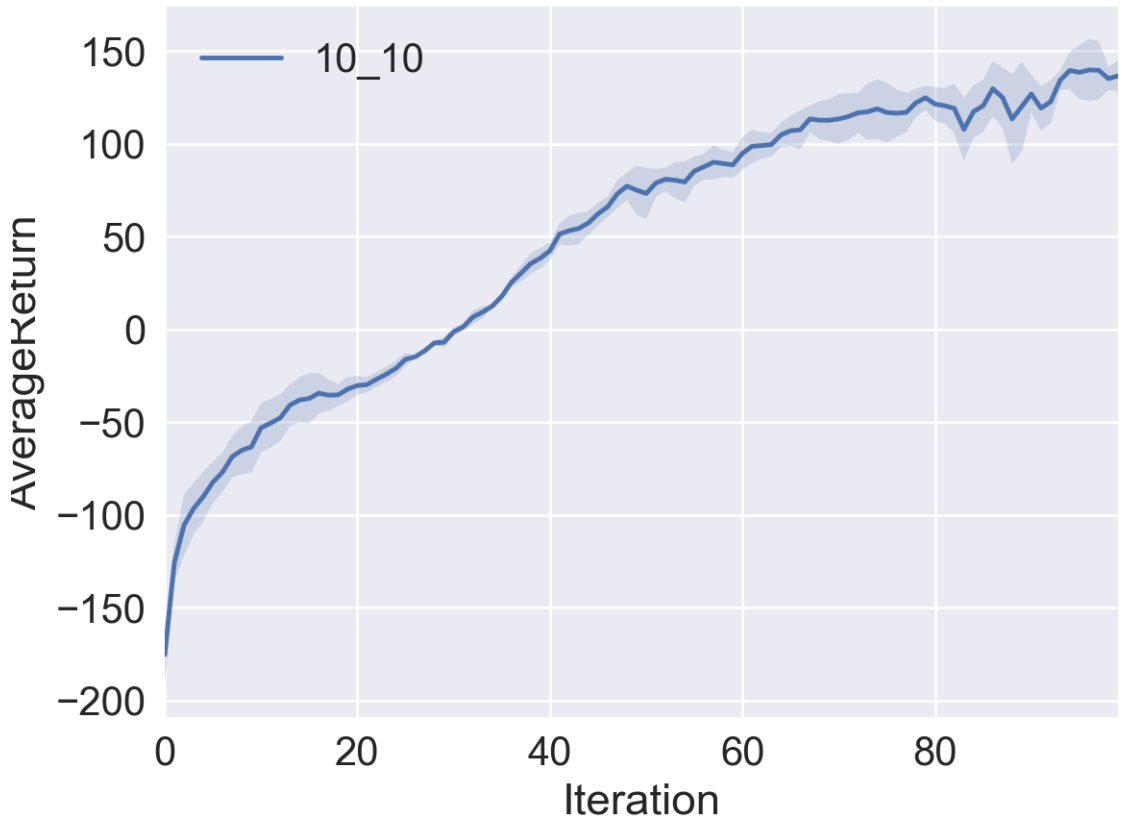

Figure 6: Implement actor-critic to HalfCheetah task and the average return achieves 150

```
python train_ac_f18.py HalfCheetah-v2 -ep 150 --discount 0.90 -n 100 -e 3 -l 2
    -s 32 -b 30000 -lr 0.02 --exp_name 10_10 -ntu 10 -ngsptu 10
```

References

[1] Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *AAAI*, volume 2, page 5. Phoenix, AZ, 2016.