20233508 DoHyoung Lee

Term project 3

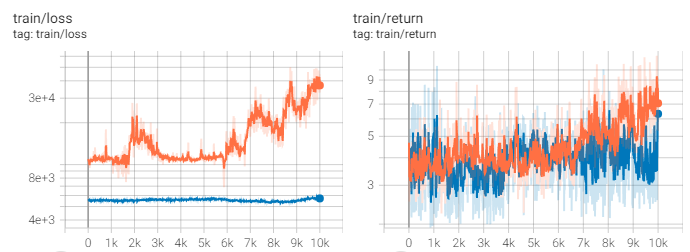
1. **Explanation of Algorithm**

Q Actor-Critic algorhtm (QAC, A. Barto, R. Sutton, 1983) evaluates the action agent did using q(s, a) function, expected return when the agent choose action a in state s. This q value role as a critic when the policy network update its parameter through gradient descent method, and q value also be evaluated and updated through TD(0) method.

In this project, there are two networks: policy network and critic network. Both have same number of layers, and same number of neurons in each layer. In one epoch, 1000 steps are operated and 24-point observation, 6-point action and reward information is saved in trajectory. Policy network receives observation information(1000 \* 24), and returns action tensor(1000 \* 6). The policy network and critic network has only varies on the initialization, but policy network weights are updated based on the gradient of absolute value of (log probability \* critic network output), where the weights of critic network are updated by mean square loss of TD error(reward + discount \* q values of next observation – q values of current observation).

1. **Plot of Return / Loss**

Figure Train Loss(Left) and Return(Right) of REINFORCE(Orange) and Q-based Actor Critic(Blue) Agent

Compare with the sample REINFORCE algorithm, QAC agent showed higher variance and lower performance during training. In contrast, the amount of loss was lower than that of REINFORCE algorithm, as was the variance of loss.

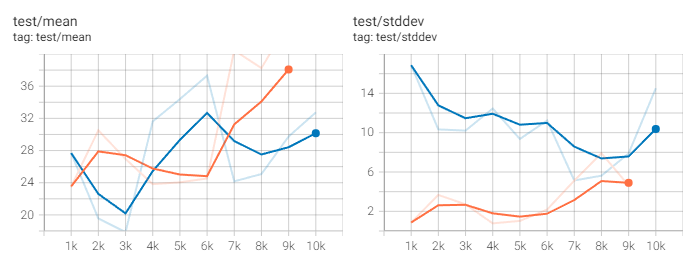
In test cases, however, the mean return of QAC showed increasing tendency. Also, the standard deviation of QAC showed decreasing tendency even though the absolute value was larger than that of REINFORCE.

Figure Test Mean(Left) and Standard Deviation(Right) of REINFORCE(Orange) and Q-based Actor Critic(Blue) Agent

1. **Analysis**

The main advantage of actor-critic reinforcement learning is that it shows low variance. It was possible by critic can evaluate how good the action is, and q value expectation is not directly by the policy updates.

In the training sequence, the return was increased very slowly. I guess this result is due to the insufficient fine tuning and how the actor network works. First, the standard deviation, which is related with variance, tends to decrease. Hence, we can see low variance tendency from here even though it is higher than that of REINFORCE algorithm.

In this model, the number of output layer is same with how many points the action values occur, and this is because each action is non-discrete unlike gridworld, select one discrete action among action candidates. Actor network updating is related to q value, and this q value can be calculated from two methods: using critic network and computing from episode using Monte Carlo method. In this model, updating based on q value acquired from critic network, but to make q expectation not related by policy update, MC based q value expectation would be better.

1. **Hyperparameters**

Discount gamma: 0.99 / Learning rate of actor and critic network: 0.001. During training, random seed value of environment and two networks are fixed as zero.