Continuous Control Report

Caijiawen

February 2019

1 Learning Algorithm

1.1 PPO algorithm

We use PPO (proximal policy optimization) to solve the continuous control project in unity environment.

In continuous control context , the parameter-reward surface is usually irregular . We are prone to be stucked in "cliff" when we use common policy iteration algorithm(REINFORCE). The PPO algorithm use clip trick to make the training process safer.

PPO can be represented as optimization problem:

$$\max_{\theta} \sum_{n=1}^{N} \frac{\pi_\theta(a_n|s_n)}{\pi_{\theta_{old}}(a_n|s_n)} A_n$$

And we create loss surrogate L_{clip}:

$$L_{clip}(\theta) = E_t[min(r_t(\theta)A_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t)]$$

where $r_t(\theta)$ represents $\frac{\pi_\theta(a_n|s_n)}{\pi_{\theta_{old}}(a_n|s_n)}$.

1.2 Implementation Detail

1.2.1 Neural Network Models

We use 2 networks : actor network and critic network.

Actor network use state to predict action , its network architecture:

- (fc1): Linear(InFeatures=33, OutFeatures=512, bias=True)
- (fc2): Linear(InFeatures=512, OutFeatures=256, bias=True)
- (fc3): Linear(InFeatures=256, OutFeatures=4, bias=True)

Critic network use state to predict value(expectation of discounted return of future) , its network architecture:

- (fc1): Linear(InFeatures=33, OutFeatures=512, bias=True)
- (fc2): Linear(InFeatures=512, OutFeatures=256, bias=True)
- (fc3): Linear(InFeatures=256, OutFeatures=1, bias=True)

1.2.2 PPO agent

In the agent part , we process data and train agent to get higher score in every episode. We divide every step into three parts: collect trajectories , compute advantages and training.

Collect trajectories First we reset the environment , use latest policy to compute action , gaussian probability and value. then we use action to get next state , reward and done information. We collect every step's information into list at last.

Compute advantages In this part, we use GAE(generalized advantage function) as advantage function. In every time step , the gae equals td error plus gae tau multiply next time step's advantage function.

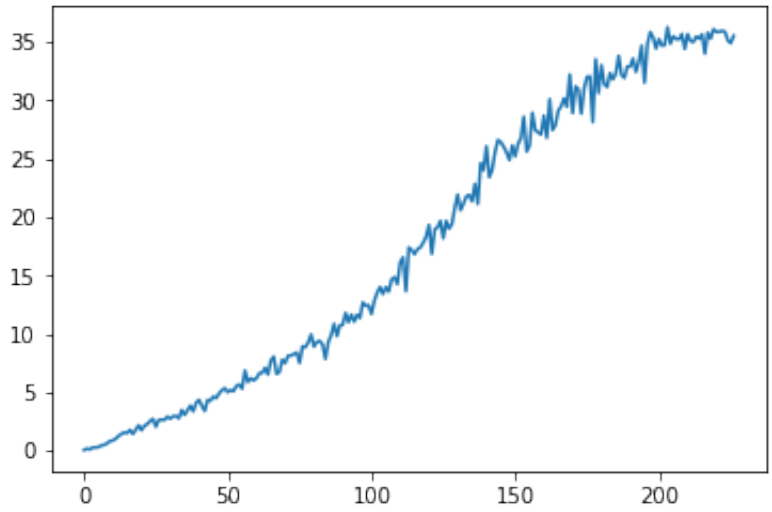
Training We use stochastic gradient descent to train the networks. To avoid "cliff" in training process , we use clip function to control the value of loss.

1.2.3 Hyperparameters

- SGD epochs : 4
- SGD batch size : 64
- discount rate: 0.99
- ratio clip(clip function parameter): 0.1
- ratio clip decay rate(ratio clip parameter decay after every episode): 0.999
- GAE tau: 0.95
- gradient clip parameter: 5

2 Results

Figure 1: Score over Episodes



The environment is solved by our PPO agent in 226 episodes.

3 Conclusion and Future Work

There two main issues in policy iteration algorithm , one is data efficiency, how can we use one trajectory to learn more? Another one is stability , how can we prevent learner fall into "desperate ground"?

In our PPO agent , we use critic network and GAE function to reduce variance of training , thus improve data efficiency. And we use clip function to make training process more stable.

In future , we may try more algorithm such as A3C or DDPG to attack this question and make some comparison. We can design better algorithm to solve the 2 central problem of policy iteration problem.