Lecture 10: Proximal Policy Optimzation

Object Manipulation and Task Planning

Victor Risager

March 12, 2024

1 Introduction

- Not necessarily the state of the art.

Comparrison to standard deep reinforcement learning.

- The distribution can be different in different times during training and execution. This is called dynamic datasets, and you can also use offline reinforcement learning, which utilises a prerecorded dataset.
- Uses sthochastic gradient descent that does not require a great dataset beforehand.
- The exercise only needs to train for 10 minutes.
- It records the dataset during the training.
- DQN stores the action taken in each state.

Challenges:

- Training instabilities. The everchanging data environment can make the tareining process unstable and unpredictable.
-

PPO offers great solutions to the epsilon-greedy tradeoff. Exploration/explotation
Policy Gradient Method

- Instead of predicting Q-values, we can then output the action directly, so it essentially can work as a continuous action space.

PPO has online learning, so it takes the transitions and makes the action directly.
Proporties

- It has a trust region.
- If the gradient is in a good area, then we update the policy a lot, and vice versa.
- Calculate the policy gradient loss
  - This function includes the policy
  - Advantage is indicting how much better this action is compared to the typical action taken in that state.
  - Expectation
- Advantage estimate
  - Discounted sum of rewards - baseline estimate. uses the value function $V(s)$
  - If the advantage is possitive, the gradient will be positive. This will increase the action policy.
- Conservative gradient methods.

  - Gradient descent on sampled data may move the policy too for away from good regions.

- We are looking at ratios between the current policy relative to the previous policy. If the ratio is > 1 then it was better than the previous. Therfore the action is in greater favor relative to the previous policy.
- Clipped surrogate function $\rightarrow$ we have to set the epsilon value. usually 0.2

Actor critic. The actor is the neural network that computes the action. The critic neural network tells the actor how it is doing. This can also be done on PPO.