ML2022-2023 Spring HW12 Report

Report Questions

Question 1 Implement Advanced RL algorithm

Q1(a)

Choose one algorithm from Actor-Critic、REINFORCE with baseline、Q Actor-Critic、A2C, A3C or other advance RL algorithms and implement it.

Answer:
The algorithm I choose: Actor-Crtic, and the algorithm pseudocode is as follows:

Algorithm 2 Actor-Critic

function REINFORCE WITH BASELINE
  Initialize policy parameters θ
  Initialize baseline function parameters ϕ
  for each episode $\{s_1, a_1, r_1, \ldots, s_T, a_T, r_T\} \sim \pi_\theta$ **do**
    for $t = 1$ to T **do**
      Calculate discounted reward $R_t = \sum_{i=t}^{T} \gamma^{i-t} r_i$
      Estimate advantage $A_t = R_t - b_\phi(s_t)$
      Re-fit the baseline by minimizing $\|b_\phi(s_t) - R_t\|^2$
      $\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi_\theta(a_t|s_t) A_t$
    end for
  end for
  return θ
end function

Q1(b)

Please explain the difference between your implementation and Policy Gradient

Answer:
Compared with Policy Gradient, the Actor to Critic model backend has two branches, one predicts actions and one predicts rewards. The loss function also needs to add the loss of predicted rewards.

Q1(c)

Please describe your implementation explicitly (If TAs can't understand your description, we will check your code directly.

Answer:

This Actor-Critic model processes the state through a fully connected network and outputs the probability distribution and state value of the action respectively. By storing the state value and log probability at each step, and using these values to calculate the loss during the learning process, the model is able to optimize parameters through backpropagation and learn how to choose appropriate actions in different states.

Quesion 2 Answer Questions based on the InstructGPT paper

Q2(a)

How does the objective function of "PPO-ptx" differ from the “PPO” during RL training as used in the InstructGPT paper?

Answer:

According to the paper, The objective function of PPO-ptx adds the penalty term of the pre-training data on the basis of PPO. The object function is as follows:

$$\text{objective}(\phi) = E_{(x,y)\sim D_{\pi_\phi^{\text{RL}}}}\left[r_\theta(x,y) - \beta \log\left(\pi_\phi^{\text{RL}}(y \mid x)/\pi^{\text{SFT}}(y \mid x)\right)\right] + \gamma E_{x\sim D_{\text{pretrain}}}\left[\log(\pi_\phi^{\text{RL}}(x))\right]$$

Since in PPO, γ is set as 0, so we do not need to take the pretraining data into consideration.

Q2(b)

Also, what is the potential advantage of using "PPO-ptx" over “PPO” in the InstructGPT paper?

Answer: PPO-ptx is optimized by combining pre-training data with reinforcement learning data, which enables the model to be trained more consistently and efficiently, and thus may achieve better performance in real applications.