**Lab5: Deep Q-Network and Deep Deterministic Policy Gradient**

**系級:智能系統 學號:312581006 姓名:張宸瑋**

1. Experimental Results
2. LunarLander-v2 using deep Q-network DQN

|  |
| --- |
|  |
| Testing Results |
|  |
| Tensorboard   1. LunarLanderContinuous-v2 DDPG  |  | | --- | |  | | Testing Results | |  | | Tensorboard | |

1. BreakoutNoFrameskip-v4 using deep Q-network (DQN)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | |  | | Testing Results | |  | | Tensorboard | |

1. Experimental Results of bonus parts (DDQN, TD3) (15%)

|  |
| --- |
| LunarLander-v2 using deep Q-network DDQN |
|  |
| Testing Results |
|  |
| Tensorboard |

1. Questions
   * 1. **Describe your major implementation of both DQN and DDPG in detail.Your description should at least contain three parts**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **(1)** Your implementation of Q network updating in DQN.  在DQN當中，我使用了Experience Replay跟Fixed Target Network，而Experience Replay的實現，我使用了一個緩衝區，用來儲存環境代理，在與環境交互中觀察的狀態，透過這個緩衝區，我會隨機抽樣一批經驗，透過這樣的操作來減少相關性，而透過Fixed Target Network來提高穩定性，當前網路用於計算Q值，而固定目標網路，來用於計算目標Q值，並且定期更新，將當前網路的參數複製到固定目標網路。   |  | | --- | |  | | 更新參數實現 |   **(2)** Your implementation and the gradient of actor updating in DDPG.   |  | | --- | | 我們會根據當前狀態，使用演員網路生成一個行動，透過這個行動，計算演員網路參數的梯度，這個梯度表示如何調整演員網路的參數，並且透過行動梯度來更新演員網路的參數，以最大化期望獎勵。 | |  | | 更新網路參數實現 |   **(3)** Your implementation and the gradient of critic updating in DDPG.   |  | | --- | | 評論家網路用於估計狀態-行動對的Q值，以指導演員網路的訓練，主要的更新步驟如下，第一，計算Q值誤差，第二，計算梯度，第三，更新評論家網路。 | |  | | 更新網路參數實現 | |

* + 1. **Explain effects of the discount factor**

γ是一個強化學習很重要的參數，它影響著代理學習的和策略的行為，而較低的γ則更關注即時獎勵。這對於探索與利用的權衡具有關鍵性影響，較高的γ鼓勵探索，而較低的γ可能導致更保守的策略。γ還影響學習的收斂速度，較高的γ可能需要更長的時間來收斂，因為代理需要考慮複雜的長期影響。折扣因子的選擇應根據具體的任務和環境來調整，不同的應用場景可能需要不同的γ值。

* + 1. **Explain benefits of epsilon-greedy in comparison to greedy to action selection**

epsilon-greedy策略在強化學習中的優勢在於它的探索能力和適應性，面對不確定性或噪聲的情況下，表現更穩定它可以幫助代理更好地學習和適應不同的環境，同時避免陷入次優解中。

|  |
| --- |
| * + 1. **Explain the necessity of the target network**   目標網路的作用主要是在穩定性、波動的減少，以及處理估算偏差上。目標網路的存在使得訓練更為穩定，有助於更快地達到良好的策略，使智能體能夠更好地處理複雜的環境和學習任務。   * + 1. **Describe the tricks you used in Breakout and their effects,and how they differ from those used in LunarLander** |

在這次的實作上，我們使用了在講義中提到的frame sracking，透過將環境的幀堆疊再一起，我們能夠捕捉動態訊息，減少過去訊息的丟失，以及減少噪音，而在Breakout方面，frame sracking可以幫助代理更好的捕捉球的運動，以及彈跳版的移動，連續的幀能夠提供關於球的速度，和方向的訊息，也能夠幫助代理更容易了解磚塊的位置跟分數情況，而相對於Breakout來說LunarLander，比較不需要太多的frame sracking因為遊戲的動作相對緩慢，而且幀之間的關聯性相對較高。因此，在這種情況下，通常可以使用較小的幀堆疊窗口。