初赛的算法基础

Q-learing

基本原理：

Q-learing算法通过不断更新Q值学习最优策略，Q值表示在某一状态下，采取某一动作能够获得的预期累计的奖励。Q值用于衡量当前状态下采取某个动作，最终带来的回报。Q-learing更新公式：

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t) \right]$$

其中：

- $Q(s_t, a_t)$是当前状态s_t采取动作a_t的价值
- α是学习率，控制Q值更新的步长
- r_{t+1}是执行动作后的即时奖励
- s_{t+1}是执行动作后的下一个状态
- γ是折扣因子，衡量当前奖励和未来奖励的重要性
- $max_{a'}Q(s_{t+1}, a')$表示在状态s_{t+1}下选择的最佳动作的Q值

传统的Q-learing中，Q表被用来存储每个状态-动作对的Q值，Q表是一个二维的表格，行表示状态，列表示动作。每个单元格存储了对应状态和动作的Q值。

维护Q表：

1. 在开始时，Q表中的所有Q值通常都被初始化为0。初始化时，agent并不知道任何状态-动作对的值。
2. 学习过程：agent与环境交互，通过采取动作并根据反馈更新Q表。每次代理选择一个动作并执行后，都会根据上述公式更新对应的Q值。
3. 贪婪策略：agent根据其当前的Q表选择动作，通常使用ε-greedy策略，即大部分时间选择当前Q值最大的动作，但有一定概率选择随机动作，以保证探索新状态。
4. Q值更新：每次代理执行一个动作并获得奖励后，Q值会根据贝尔曼方程进行更新，以反映新的估计。随着学习的进行，Q表逐渐收敛，Q值变得更精确

具体示例：

假设环境为一个5×5的格子，智能体起始位置在左上角$(0, 0)$，目标位置是右下角$(4, 4)$，智能体可以选择四种动作之一：

- Up
- Down
- Left
- Right

每个动作都有一个奖励：

- 到达目标时，奖励为+10
- 没移动一步奖励-1
- 碰到墙壁或者超出边界，智能体停留在原地，奖励-1

环境与Q表实现

将环境状态设置为$(0,0)$到$(4,4)$的每个格子，用(x,y)来表示状态，并将其转换为对应的状态编号

In [1]:
```python
import numpy as np
import random

class GridWorld:
    def __init__(self, grid_size=5):
        self.grid_size = grid_size
        self.state_space = grid_size * grid_size
        self.action_space = 4
        self.start_state = (0, 0)
        self.goal_state = (4, 4)

    def state_to_id(self, state):
        return state[0] * self.grid_size + state[1]

    # 0:Up, 1:Down, 2:Left, 3:Right
    def take_action(self, state, action):
        x, y = state
        if action == 0:
            x = max(x - 1, 0)
        elif action == 1:
            x = min(x + 1, self.grid_size - 1)
        elif action == 2:
            y = max(y - 1, 0)
        elif action == 3:
            y = min(y + 1, self.grid_size - 1)
        return (x, y)

    def get_reward(self, state):
        if state == self.goal_state:
            return 10
        else:
            return -1

    def reset(self):
        return self.start_state
```

Q-learing算法实现

1. **参数:**

- `epsilon`：探索率决定智能体是利用已有的知识还是进行随机探索，当`epsilon`较高时，智能体更多探索未知状态，`epsilon`较低时，智能体更多利用已有知识
- `q_table`：Q表是二维矩阵，用来存储每个状态-动作对应的Q值，`q_table[state_id, action]`存储的是从`state_id`状态下采取的`action`动作的预期回报

2. **动作选择：**

如果一个随机值小于当前的`epsilon`，智能体随机选择一个动作，否则智能体会选择当前状态下`Q`值最大的动作

3. **Q值更新**

- 当前Q值：`self.q_table[current_state_id, action]`
- 智能体从当前状态到一下状态的预期回报：`target = reward + self.gamma * self.q_table[next_state_id, best_next_action]`，其中`self.q_table[next_state_id, best_next_action]`是下一状态的最大Q值
- 当前Q值与目标Q值之间的差距：`error = target - self.q_table[current_state_id, action]`
- Q值更新：self.q_table[current_state_id, action] += self.alpha * error

4. **epsilon衰减**

为了逐渐减少探索行为并增加利用行为，`epsilon`会随着训练的进行逐渐减少，`epsilon_decay`控制衰减速率，`min_epsilon`确保探索率不会降的过低

In [2]:
```
class Agent:
    def __init__(self, env, alpha=0.1, gamma=0.9, epsilon=1.0, epsilon_dacay=0.9
        self.env = env
        self.alpha = alpha
        self.gamma = gamma
        self.epsilon = epsilon
        self.epsilon_decay = epsilon_dacay
        self.min_epsilon = min_epsilon
        self.q_table = np.zeros((env.state_space, env.action_space))
        self.state = env.reset()

    def choose_action(self):
        if random.uniform(0,1) < self.epsilon:
            return random.randint(0, self.env.action_space - 1)
        else:
            return np.argmax(self.q_table[self.env.state_to_id(self.state)])

    def update_q_value(self, next_state, action, reward):
        current_state_id = self.env.state_to_id(self.state)
        next_state_id = self.env.state_to_id(next_state)

        best_next_action = np.argmax(self.q_table[next_state_id])
        target = reward + self.gamma * self.q_table[next_state_id, best_next_act
        error = target - self.q_table[current_state_id, action]
        self.q_table[current_state_id, action] += self.alpha * error

    def update_epsilon(self):
        if self.epsilon > self.min_epsilon:
            self.epsilon *= self.epsilon_decay

    def train(self, episodes=1000):
        for episode in range(episodes):
            self.state = self.env.reset()
            total_reward = 0
            done = False
```

```
            while not done:
                action = self.choose_action()
                next_state = self.env.take_action(self.state, action)
                reward = self.env.get_reward(next_state)

                self.update_q_value(next_state, action, reward)
                self.state = next_state
                total_reward += reward

                if next_state == self.env.goal_state:
                    done = True

            self.update_epsilon()
            print(f"Episode {episode + 1}/{episodes}, Total Reward: {total_rewar

        print(self.q_table)


env = GridWorld()
agent = Agent(env)

agent.train(episodes=1000)
```

```
Episode 1/1000, Total Reward: -293, Epsilon: 0.995
Episode 2/1000, Total Reward: -57, Epsilon: 0.990
Episode 3/1000, Total Reward: -27, Epsilon: 0.985
Episode 4/1000, Total Reward: -34, Epsilon: 0.980
Episode 5/1000, Total Reward: -173, Epsilon: 0.975
Episode 6/1000, Total Reward: -121, Epsilon: 0.970
Episode 7/1000, Total Reward: -8, Epsilon: 0.966
Episode 8/1000, Total Reward: -65, Epsilon: 0.961
Episode 9/1000, Total Reward: -32, Epsilon: 0.956
Episode 10/1000, Total Reward: -65, Epsilon: 0.951
Episode 11/1000, Total Reward: -98, Epsilon: 0.946
Episode 12/1000, Total Reward: -88, Epsilon: 0.942
Episode 13/1000, Total Reward: -12, Epsilon: 0.937
Episode 14/1000, Total Reward: -97, Epsilon: 0.932
Episode 15/1000, Total Reward: -12, Epsilon: 0.928
Episode 16/1000, Total Reward: -29, Epsilon: 0.923
Episode 17/1000, Total Reward: -147, Epsilon: 0.918
Episode 18/1000, Total Reward: -9, Epsilon: 0.914
Episode 19/1000, Total Reward: -28, Epsilon: 0.909
Episode 20/1000, Total Reward: -47, Epsilon: 0.905
Episode 21/1000, Total Reward: -9, Epsilon: 0.900
Episode 22/1000, Total Reward: -17, Epsilon: 0.896
Episode 23/1000, Total Reward: -59, Epsilon: 0.891
Episode 24/1000, Total Reward: -14, Epsilon: 0.887
Episode 25/1000, Total Reward: -82, Epsilon: 0.882
Episode 26/1000, Total Reward: -41, Epsilon: 0.878
Episode 27/1000, Total Reward: -24, Epsilon: 0.873
Episode 28/1000, Total Reward: -91, Epsilon: 0.869
Episode 29/1000, Total Reward: -24, Epsilon: 0.865
Episode 30/1000, Total Reward: -9, Epsilon: 0.860
Episode 31/1000, Total Reward: -57, Epsilon: 0.856
Episode 32/1000, Total Reward: -17, Epsilon: 0.852
Episode 33/1000, Total Reward: -4, Epsilon: 0.848
Episode 34/1000, Total Reward: -15, Epsilon: 0.843
Episode 35/1000, Total Reward: -13, Epsilon: 0.839
Episode 36/1000, Total Reward: 3, Epsilon: 0.835
Episode 37/1000, Total Reward: -14, Epsilon: 0.831
Episode 38/1000, Total Reward: -2, Epsilon: 0.827
Episode 39/1000, Total Reward: -31, Epsilon: 0.822
Episode 40/1000, Total Reward: -1, Epsilon: 0.818
Episode 41/1000, Total Reward: -96, Epsilon: 0.814
Episode 42/1000, Total Reward: -14, Epsilon: 0.810
Episode 43/1000, Total Reward: -58, Epsilon: 0.806
Episode 44/1000, Total Reward: -10, Epsilon: 0.802
Episode 45/1000, Total Reward: -24, Epsilon: 0.798
Episode 46/1000, Total Reward: -18, Epsilon: 0.794
Episode 47/1000, Total Reward: -44, Epsilon: 0.790
Episode 48/1000, Total Reward: -30, Epsilon: 0.786
Episode 49/1000, Total Reward: -11, Epsilon: 0.782
Episode 50/1000, Total Reward: -15, Epsilon: 0.778
Episode 51/1000, Total Reward: -23, Epsilon: 0.774
Episode 52/1000, Total Reward: -10, Epsilon: 0.771
Episode 53/1000, Total Reward: -10, Epsilon: 0.767
Episode 54/1000, Total Reward: -12, Epsilon: 0.763
Episode 55/1000, Total Reward: -12, Epsilon: 0.759
Episode 56/1000, Total Reward: -9, Epsilon: 0.755
Episode 57/1000, Total Reward: -36, Epsilon: 0.751
Episode 58/1000, Total Reward: -6, Epsilon: 0.748
Episode 59/1000, Total Reward: -22, Epsilon: 0.744
Episode 60/1000, Total Reward: 1, Epsilon: 0.740
```

```
Episode 61/1000, Total Reward: -6, Epsilon: 0.737
Episode 62/1000, Total Reward: -34, Epsilon: 0.733
Episode 63/1000, Total Reward: -1, Epsilon: 0.729
Episode 64/1000, Total Reward: -11, Epsilon: 0.726
Episode 65/1000, Total Reward: -18, Epsilon: 0.722
Episode 66/1000, Total Reward: -14, Epsilon: 0.718
Episode 67/1000, Total Reward: -49, Epsilon: 0.715
Episode 68/1000, Total Reward: -3, Epsilon: 0.711
Episode 69/1000, Total Reward: -8, Epsilon: 0.708
Episode 70/1000, Total Reward: -9, Epsilon: 0.704
Episode 71/1000, Total Reward: -2, Epsilon: 0.701
Episode 72/1000, Total Reward: -33, Epsilon: 0.697
Episode 73/1000, Total Reward: 2, Epsilon: 0.694
Episode 74/1000, Total Reward: 1, Epsilon: 0.690
Episode 75/1000, Total Reward: -3, Epsilon: 0.687
Episode 76/1000, Total Reward: -2, Epsilon: 0.683
Episode 77/1000, Total Reward: -4, Epsilon: 0.680
Episode 78/1000, Total Reward: -17, Epsilon: 0.676
Episode 79/1000, Total Reward: 1, Epsilon: 0.673
Episode 80/1000, Total Reward: -67, Epsilon: 0.670
Episode 81/1000, Total Reward: -23, Epsilon: 0.666
Episode 82/1000, Total Reward: -36, Epsilon: 0.663
Episode 83/1000, Total Reward: -7, Epsilon: 0.660
Episode 84/1000, Total Reward: -15, Epsilon: 0.656
Episode 85/1000, Total Reward: -12, Epsilon: 0.653
Episode 86/1000, Total Reward: -10, Epsilon: 0.650
Episode 87/1000, Total Reward: 2, Epsilon: 0.647
Episode 88/1000, Total Reward: -7, Epsilon: 0.643
Episode 89/1000, Total Reward: -3, Epsilon: 0.640
Episode 90/1000, Total Reward: -15, Epsilon: 0.637
Episode 91/1000, Total Reward: -41, Epsilon: 0.634
Episode 92/1000, Total Reward: -13, Epsilon: 0.631
Episode 93/1000, Total Reward: -14, Epsilon: 0.627
Episode 94/1000, Total Reward: -6, Epsilon: 0.624
Episode 95/1000, Total Reward: -5, Epsilon: 0.621
Episode 96/1000, Total Reward: 1, Epsilon: 0.618
Episode 97/1000, Total Reward: 3, Epsilon: 0.615
Episode 98/1000, Total Reward: -9, Epsilon: 0.612
Episode 99/1000, Total Reward: -10, Epsilon: 0.609
Episode 100/1000, Total Reward: -15, Epsilon: 0.606
Episode 101/1000, Total Reward: -23, Epsilon: 0.603
Episode 102/1000, Total Reward: 1, Epsilon: 0.600
Episode 103/1000, Total Reward: 1, Epsilon: 0.597
Episode 104/1000, Total Reward: -16, Epsilon: 0.594
Episode 105/1000, Total Reward: -4, Epsilon: 0.591
Episode 106/1000, Total Reward: 1, Epsilon: 0.588
Episode 107/1000, Total Reward: 3, Epsilon: 0.585
Episode 108/1000, Total Reward: -2, Epsilon: 0.582
Episode 109/1000, Total Reward: -9, Epsilon: 0.579
Episode 110/1000, Total Reward: -3, Epsilon: 0.576
Episode 111/1000, Total Reward: -1, Epsilon: 0.573
Episode 112/1000, Total Reward: -9, Epsilon: 0.570
Episode 113/1000, Total Reward: 2, Epsilon: 0.568
Episode 114/1000, Total Reward: -3, Epsilon: 0.565
Episode 115/1000, Total Reward: -9, Epsilon: 0.562
Episode 116/1000, Total Reward: 0, Epsilon: 0.559
Episode 117/1000, Total Reward: 1, Epsilon: 0.556
Episode 118/1000, Total Reward: -4, Epsilon: 0.554
Episode 119/1000, Total Reward: -17, Epsilon: 0.551
Episode 120/1000, Total Reward: -5, Epsilon: 0.548
```

```
Episode 121/1000, Total Reward: 2, Epsilon: 0.545
Episode 122/1000, Total Reward: -22, Epsilon: 0.543
Episode 123/1000, Total Reward: 1, Epsilon: 0.540
Episode 124/1000, Total Reward: -2, Epsilon: 0.537
Episode 125/1000, Total Reward: 0, Epsilon: 0.534
Episode 126/1000, Total Reward: -12, Epsilon: 0.532
Episode 127/1000, Total Reward: -6, Epsilon: 0.529
Episode 128/1000, Total Reward: -3, Epsilon: 0.526
Episode 129/1000, Total Reward: -8, Epsilon: 0.524
Episode 130/1000, Total Reward: -11, Epsilon: 0.521
Episode 131/1000, Total Reward: 2, Epsilon: 0.519
Episode 132/1000, Total Reward: -4, Epsilon: 0.516
Episode 133/1000, Total Reward: 1, Epsilon: 0.513
Episode 134/1000, Total Reward: -7, Epsilon: 0.511
Episode 135/1000, Total Reward: -1, Epsilon: 0.508
Episode 136/1000, Total Reward: 2, Epsilon: 0.506
Episode 137/1000, Total Reward: 0, Epsilon: 0.503
Episode 138/1000, Total Reward: -3, Epsilon: 0.501
Episode 139/1000, Total Reward: -5, Epsilon: 0.498
Episode 140/1000, Total Reward: -3, Epsilon: 0.496
Episode 141/1000, Total Reward: -13, Epsilon: 0.493
Episode 142/1000, Total Reward: 0, Epsilon: 0.491
Episode 143/1000, Total Reward: 3, Epsilon: 0.488
Episode 144/1000, Total Reward: 1, Epsilon: 0.486
Episode 145/1000, Total Reward: 3, Epsilon: 0.483
Episode 146/1000, Total Reward: 3, Epsilon: 0.481
Episode 147/1000, Total Reward: -2, Epsilon: 0.479
Episode 148/1000, Total Reward: 3, Epsilon: 0.476
Episode 149/1000, Total Reward: 1, Epsilon: 0.474
Episode 150/1000, Total Reward: -32, Epsilon: 0.471
Episode 151/1000, Total Reward: -17, Epsilon: 0.469
Episode 152/1000, Total Reward: -6, Epsilon: 0.467
Episode 153/1000, Total Reward: 0, Epsilon: 0.464
Episode 154/1000, Total Reward: -9, Epsilon: 0.462
Episode 155/1000, Total Reward: -13, Epsilon: 0.460
Episode 156/1000, Total Reward: -1, Epsilon: 0.458
Episode 157/1000, Total Reward: -9, Epsilon: 0.455
Episode 158/1000, Total Reward: -1, Epsilon: 0.453
Episode 159/1000, Total Reward: 3, Epsilon: 0.451
Episode 160/1000, Total Reward: 1, Epsilon: 0.448
Episode 161/1000, Total Reward: 3, Epsilon: 0.446
Episode 162/1000, Total Reward: -22, Epsilon: 0.444
Episode 163/1000, Total Reward: -4, Epsilon: 0.442
Episode 164/1000, Total Reward: -9, Epsilon: 0.440
Episode 165/1000, Total Reward: -1, Epsilon: 0.437
Episode 166/1000, Total Reward: -6, Epsilon: 0.435
Episode 167/1000, Total Reward: -18, Epsilon: 0.433
Episode 168/1000, Total Reward: 2, Epsilon: 0.431
Episode 169/1000, Total Reward: -5, Epsilon: 0.429
Episode 170/1000, Total Reward: -2, Epsilon: 0.427
Episode 171/1000, Total Reward: -8, Epsilon: 0.424
Episode 172/1000, Total Reward: 1, Epsilon: 0.422
Episode 173/1000, Total Reward: -1, Epsilon: 0.420
Episode 174/1000, Total Reward: 0, Epsilon: 0.418
Episode 175/1000, Total Reward: -4, Epsilon: 0.416
Episode 176/1000, Total Reward: -7, Epsilon: 0.414
Episode 177/1000, Total Reward: 1, Epsilon: 0.412
Episode 178/1000, Total Reward: 3, Epsilon: 0.410
Episode 179/1000, Total Reward: 2, Epsilon: 0.408
Episode 180/1000, Total Reward: -1, Epsilon: 0.406
```

```
Episode 181/1000, Total Reward: 3, Epsilon: 0.404
Episode 182/1000, Total Reward: -7, Epsilon: 0.402
Episode 183/1000, Total Reward: 2, Epsilon: 0.400
Episode 184/1000, Total Reward: -1, Epsilon: 0.398
Episode 185/1000, Total Reward: -13, Epsilon: 0.396
Episode 186/1000, Total Reward: 0, Epsilon: 0.394
Episode 187/1000, Total Reward: 0, Epsilon: 0.392
Episode 188/1000, Total Reward: -6, Epsilon: 0.390
Episode 189/1000, Total Reward: 0, Epsilon: 0.388
Episode 190/1000, Total Reward: -1, Epsilon: 0.386
Episode 191/1000, Total Reward: 3, Epsilon: 0.384
Episode 192/1000, Total Reward: -1, Epsilon: 0.382
Episode 193/1000, Total Reward: 2, Epsilon: 0.380
Episode 194/1000, Total Reward: -9, Epsilon: 0.378
Episode 195/1000, Total Reward: -1, Epsilon: 0.376
Episode 196/1000, Total Reward: -2, Epsilon: 0.374
Episode 197/1000, Total Reward: 1, Epsilon: 0.373
Episode 198/1000, Total Reward: -5, Epsilon: 0.371
Episode 199/1000, Total Reward: -16, Epsilon: 0.369
Episode 200/1000, Total Reward: -1, Epsilon: 0.367
Episode 201/1000, Total Reward: 1, Epsilon: 0.365
Episode 202/1000, Total Reward: -2, Epsilon: 0.363
Episode 203/1000, Total Reward: -1, Epsilon: 0.361
Episode 204/1000, Total Reward: 3, Epsilon: 0.360
Episode 205/1000, Total Reward: 3, Epsilon: 0.358
Episode 206/1000, Total Reward: -1, Epsilon: 0.356
Episode 207/1000, Total Reward: 3, Epsilon: 0.354
Episode 208/1000, Total Reward: 0, Epsilon: 0.353
Episode 209/1000, Total Reward: 2, Epsilon: 0.351
Episode 210/1000, Total Reward: 3, Epsilon: 0.349
Episode 211/1000, Total Reward: 1, Epsilon: 0.347
Episode 212/1000, Total Reward: -4, Epsilon: 0.346
Episode 213/1000, Total Reward: 3, Epsilon: 0.344
Episode 214/1000, Total Reward: -1, Epsilon: 0.342
Episode 215/1000, Total Reward: 3, Epsilon: 0.340
Episode 216/1000, Total Reward: 3, Epsilon: 0.339
Episode 217/1000, Total Reward: -1, Epsilon: 0.337
Episode 218/1000, Total Reward: -14, Epsilon: 0.335
Episode 219/1000, Total Reward: 0, Epsilon: 0.334
Episode 220/1000, Total Reward: 2, Epsilon: 0.332
Episode 221/1000, Total Reward: -4, Epsilon: 0.330
Episode 222/1000, Total Reward: 0, Epsilon: 0.329
Episode 223/1000, Total Reward: 2, Epsilon: 0.327
Episode 224/1000, Total Reward: -10, Epsilon: 0.325
Episode 225/1000, Total Reward: 3, Epsilon: 0.324
Episode 226/1000, Total Reward: 3, Epsilon: 0.322
Episode 227/1000, Total Reward: 1, Epsilon: 0.321
Episode 228/1000, Total Reward: -1, Epsilon: 0.319
Episode 229/1000, Total Reward: -4, Epsilon: 0.317
Episode 230/1000, Total Reward: -1, Epsilon: 0.316
Episode 231/1000, Total Reward: -4, Epsilon: 0.314
Episode 232/1000, Total Reward: -2, Epsilon: 0.313
Episode 233/1000, Total Reward: -7, Epsilon: 0.311
Episode 234/1000, Total Reward: 2, Epsilon: 0.309
Episode 235/1000, Total Reward: 1, Epsilon: 0.308
Episode 236/1000, Total Reward: -1, Epsilon: 0.306
Episode 237/1000, Total Reward: -2, Epsilon: 0.305
Episode 238/1000, Total Reward: -14, Epsilon: 0.303
Episode 239/1000, Total Reward: 1, Epsilon: 0.302
Episode 240/1000, Total Reward: 3, Epsilon: 0.300
```

```
Episode 241/1000, Total Reward: 1, Epsilon: 0.299
Episode 242/1000, Total Reward: 1, Epsilon: 0.297
Episode 243/1000, Total Reward: 3, Epsilon: 0.296
Episode 244/1000, Total Reward: -2, Epsilon: 0.294
Episode 245/1000, Total Reward: 1, Epsilon: 0.293
Episode 246/1000, Total Reward: 3, Epsilon: 0.291
Episode 247/1000, Total Reward: 3, Epsilon: 0.290
Episode 248/1000, Total Reward: 1, Epsilon: 0.288
Episode 249/1000, Total Reward: -1, Epsilon: 0.287
Episode 250/1000, Total Reward: 1, Epsilon: 0.286
Episode 251/1000, Total Reward: -2, Epsilon: 0.284
Episode 252/1000, Total Reward: 3, Epsilon: 0.283
Episode 253/1000, Total Reward: 1, Epsilon: 0.281
Episode 254/1000, Total Reward: 3, Epsilon: 0.280
Episode 255/1000, Total Reward: 0, Epsilon: 0.279
Episode 256/1000, Total Reward: 1, Epsilon: 0.277
Episode 257/1000, Total Reward: -1, Epsilon: 0.276
Episode 258/1000, Total Reward: 1, Epsilon: 0.274
Episode 259/1000, Total Reward: -5, Epsilon: 0.273
Episode 260/1000, Total Reward: -2, Epsilon: 0.272
Episode 261/1000, Total Reward: 1, Epsilon: 0.270
Episode 262/1000, Total Reward: 1, Epsilon: 0.269
Episode 263/1000, Total Reward: -11, Epsilon: 0.268
Episode 264/1000, Total Reward: 0, Epsilon: 0.266
Episode 265/1000, Total Reward: -1, Epsilon: 0.265
Episode 266/1000, Total Reward: -1, Epsilon: 0.264
Episode 267/1000, Total Reward: 1, Epsilon: 0.262
Episode 268/1000, Total Reward: 0, Epsilon: 0.261
Episode 269/1000, Total Reward: 3, Epsilon: 0.260
Episode 270/1000, Total Reward: -1, Epsilon: 0.258
Episode 271/1000, Total Reward: 3, Epsilon: 0.257
Episode 272/1000, Total Reward: 1, Epsilon: 0.256
Episode 273/1000, Total Reward: 3, Epsilon: 0.255
Episode 274/1000, Total Reward: 1, Epsilon: 0.253
Episode 275/1000, Total Reward: 3, Epsilon: 0.252
Episode 276/1000, Total Reward: 3, Epsilon: 0.251
Episode 277/1000, Total Reward: -3, Epsilon: 0.249
Episode 278/1000, Total Reward: 3, Epsilon: 0.248
Episode 279/1000, Total Reward: -2, Epsilon: 0.247
Episode 280/1000, Total Reward: 3, Epsilon: 0.246
Episode 281/1000, Total Reward: -3, Epsilon: 0.245
Episode 282/1000, Total Reward: -3, Epsilon: 0.243
Episode 283/1000, Total Reward: 1, Epsilon: 0.242
Episode 284/1000, Total Reward: 1, Epsilon: 0.241
Episode 285/1000, Total Reward: -5, Epsilon: 0.240
Episode 286/1000, Total Reward: -4, Epsilon: 0.238
Episode 287/1000, Total Reward: -5, Epsilon: 0.237
Episode 288/1000, Total Reward: -4, Epsilon: 0.236
Episode 289/1000, Total Reward: 3, Epsilon: 0.235
Episode 290/1000, Total Reward: 0, Epsilon: 0.234
Episode 291/1000, Total Reward: 1, Epsilon: 0.233
Episode 292/1000, Total Reward: 0, Epsilon: 0.231
Episode 293/1000, Total Reward: 3, Epsilon: 0.230
Episode 294/1000, Total Reward: -7, Epsilon: 0.229
Episode 295/1000, Total Reward: 1, Epsilon: 0.228
Episode 296/1000, Total Reward: 3, Epsilon: 0.227
Episode 297/1000, Total Reward: 3, Epsilon: 0.226
Episode 298/1000, Total Reward: -1, Epsilon: 0.225
Episode 299/1000, Total Reward: 3, Epsilon: 0.223
Episode 300/1000, Total Reward: 1, Epsilon: 0.222
```

```
Episode 301/1000, Total Reward: -2, Epsilon: 0.221
Episode 302/1000, Total Reward: -1, Epsilon: 0.220
Episode 303/1000, Total Reward: 1, Epsilon: 0.219
Episode 304/1000, Total Reward: 1, Epsilon: 0.218
Episode 305/1000, Total Reward: 3, Epsilon: 0.217
Episode 306/1000, Total Reward: 1, Epsilon: 0.216
Episode 307/1000, Total Reward: -2, Epsilon: 0.215
Episode 308/1000, Total Reward: 3, Epsilon: 0.214
Episode 309/1000, Total Reward: 1, Epsilon: 0.212
Episode 310/1000, Total Reward: 3, Epsilon: 0.211
Episode 311/1000, Total Reward: 3, Epsilon: 0.210
Episode 312/1000, Total Reward: 3, Epsilon: 0.209
Episode 313/1000, Total Reward: 3, Epsilon: 0.208
Episode 314/1000, Total Reward: 2, Epsilon: 0.207
Episode 315/1000, Total Reward: 1, Epsilon: 0.206
Episode 316/1000, Total Reward: -3, Epsilon: 0.205
Episode 317/1000, Total Reward: 3, Epsilon: 0.204
Episode 318/1000, Total Reward: 2, Epsilon: 0.203
Episode 319/1000, Total Reward: 2, Epsilon: 0.202
Episode 320/1000, Total Reward: 3, Epsilon: 0.201
Episode 321/1000, Total Reward: 3, Epsilon: 0.200
Episode 322/1000, Total Reward: 0, Epsilon: 0.199
Episode 323/1000, Total Reward: 1, Epsilon: 0.198
Episode 324/1000, Total Reward: 1, Epsilon: 0.197
Episode 325/1000, Total Reward: 1, Epsilon: 0.196
Episode 326/1000, Total Reward: 0, Epsilon: 0.195
Episode 327/1000, Total Reward: 3, Epsilon: 0.194
Episode 328/1000, Total Reward: -1, Epsilon: 0.193
Episode 329/1000, Total Reward: -3, Epsilon: 0.192
Episode 330/1000, Total Reward: 3, Epsilon: 0.191
Episode 331/1000, Total Reward: 3, Epsilon: 0.190
Episode 332/1000, Total Reward: 3, Epsilon: 0.189
Episode 333/1000, Total Reward: -3, Epsilon: 0.188
Episode 334/1000, Total Reward: 1, Epsilon: 0.187
Episode 335/1000, Total Reward: 1, Epsilon: 0.187
Episode 336/1000, Total Reward: -1, Epsilon: 0.186
Episode 337/1000, Total Reward: 1, Epsilon: 0.185
Episode 338/1000, Total Reward: 3, Epsilon: 0.184
Episode 339/1000, Total Reward: 2, Epsilon: 0.183
Episode 340/1000, Total Reward: 0, Epsilon: 0.182
Episode 341/1000, Total Reward: 3, Epsilon: 0.181
Episode 342/1000, Total Reward: -6, Epsilon: 0.180
Episode 343/1000, Total Reward: 3, Epsilon: 0.179
Episode 344/1000, Total Reward: 1, Epsilon: 0.178
Episode 345/1000, Total Reward: 3, Epsilon: 0.177
Episode 346/1000, Total Reward: 0, Epsilon: 0.177
Episode 347/1000, Total Reward: -1, Epsilon: 0.176
Episode 348/1000, Total Reward: -3, Epsilon: 0.175
Episode 349/1000, Total Reward: 1, Epsilon: 0.174
Episode 350/1000, Total Reward: 2, Epsilon: 0.173
Episode 351/1000, Total Reward: 1, Epsilon: 0.172
Episode 352/1000, Total Reward: -2, Epsilon: 0.171
Episode 353/1000, Total Reward: -1, Epsilon: 0.170
Episode 354/1000, Total Reward: 3, Epsilon: 0.170
Episode 355/1000, Total Reward: 2, Epsilon: 0.169
Episode 356/1000, Total Reward: 1, Epsilon: 0.168
Episode 357/1000, Total Reward: 2, Epsilon: 0.167
Episode 358/1000, Total Reward: 3, Epsilon: 0.166
Episode 359/1000, Total Reward: -1, Epsilon: 0.165
Episode 360/1000, Total Reward: 3, Epsilon: 0.165
```

```
Episode 361/1000, Total Reward: 3, Epsilon: 0.164
Episode 362/1000, Total Reward: 2, Epsilon: 0.163
Episode 363/1000, Total Reward: 3, Epsilon: 0.162
Episode 364/1000, Total Reward: -1, Epsilon: 0.161
Episode 365/1000, Total Reward: -1, Epsilon: 0.160
Episode 366/1000, Total Reward: 3, Epsilon: 0.160
Episode 367/1000, Total Reward: 3, Epsilon: 0.159
Episode 368/1000, Total Reward: 1, Epsilon: 0.158
Episode 369/1000, Total Reward: 1, Epsilon: 0.157
Episode 370/1000, Total Reward: -5, Epsilon: 0.157
Episode 371/1000, Total Reward: 1, Epsilon: 0.156
Episode 372/1000, Total Reward: 3, Epsilon: 0.155
Episode 373/1000, Total Reward: -3, Epsilon: 0.154
Episode 374/1000, Total Reward: -1, Epsilon: 0.153
Episode 375/1000, Total Reward: 3, Epsilon: 0.153
Episode 376/1000, Total Reward: 1, Epsilon: 0.152
Episode 377/1000, Total Reward: 1, Epsilon: 0.151
Episode 378/1000, Total Reward: 3, Epsilon: 0.150
Episode 379/1000, Total Reward: 3, Epsilon: 0.150
Episode 380/1000, Total Reward: 3, Epsilon: 0.149
Episode 381/1000, Total Reward: 3, Epsilon: 0.148
Episode 382/1000, Total Reward: -3, Epsilon: 0.147
Episode 383/1000, Total Reward: 3, Epsilon: 0.147
Episode 384/1000, Total Reward: 3, Epsilon: 0.146
Episode 385/1000, Total Reward: 2, Epsilon: 0.145
Episode 386/1000, Total Reward: 3, Epsilon: 0.144
Episode 387/1000, Total Reward: 1, Epsilon: 0.144
Episode 388/1000, Total Reward: 3, Epsilon: 0.143
Episode 389/1000, Total Reward: 3, Epsilon: 0.142
Episode 390/1000, Total Reward: 2, Epsilon: 0.142
Episode 391/1000, Total Reward: 1, Epsilon: 0.141
Episode 392/1000, Total Reward: 1, Epsilon: 0.140
Episode 393/1000, Total Reward: 3, Epsilon: 0.139
Episode 394/1000, Total Reward: 3, Epsilon: 0.139
Episode 395/1000, Total Reward: 3, Epsilon: 0.138
Episode 396/1000, Total Reward: 3, Epsilon: 0.137
Episode 397/1000, Total Reward: 0, Epsilon: 0.137
Episode 398/1000, Total Reward: 3, Epsilon: 0.136
Episode 399/1000, Total Reward: 1, Epsilon: 0.135
Episode 400/1000, Total Reward: -1, Epsilon: 0.135
Episode 401/1000, Total Reward: 3, Epsilon: 0.134
Episode 402/1000, Total Reward: -4, Epsilon: 0.133
Episode 403/1000, Total Reward: 3, Epsilon: 0.133
Episode 404/1000, Total Reward: 3, Epsilon: 0.132
Episode 405/1000, Total Reward: 2, Epsilon: 0.131
Episode 406/1000, Total Reward: -1, Epsilon: 0.131
Episode 407/1000, Total Reward: 3, Epsilon: 0.130
Episode 408/1000, Total Reward: 2, Epsilon: 0.129
Episode 409/1000, Total Reward: 3, Epsilon: 0.129
Episode 410/1000, Total Reward: 3, Epsilon: 0.128
Episode 411/1000, Total Reward: 1, Epsilon: 0.127
Episode 412/1000, Total Reward: 2, Epsilon: 0.127
Episode 413/1000, Total Reward: 3, Epsilon: 0.126
Episode 414/1000, Total Reward: 2, Epsilon: 0.126
Episode 415/1000, Total Reward: 1, Epsilon: 0.125
Episode 416/1000, Total Reward: 3, Epsilon: 0.124
Episode 417/1000, Total Reward: 3, Epsilon: 0.124
Episode 418/1000, Total Reward: 3, Epsilon: 0.123
Episode 419/1000, Total Reward: 3, Epsilon: 0.122
Episode 420/1000, Total Reward: 2, Epsilon: 0.122
```

```
Episode 421/1000, Total Reward: 3, Epsilon: 0.121
Episode 422/1000, Total Reward: 3, Epsilon: 0.121
Episode 423/1000, Total Reward: -3, Epsilon: 0.120
Episode 424/1000, Total Reward: 3, Epsilon: 0.119
Episode 425/1000, Total Reward: -3, Epsilon: 0.119
Episode 426/1000, Total Reward: 1, Epsilon: 0.118
Episode 427/1000, Total Reward: 3, Epsilon: 0.118
Episode 428/1000, Total Reward: 3, Epsilon: 0.117
Episode 429/1000, Total Reward: 3, Epsilon: 0.116
Episode 430/1000, Total Reward: 3, Epsilon: 0.116
Episode 431/1000, Total Reward: 1, Epsilon: 0.115
Episode 432/1000, Total Reward: 3, Epsilon: 0.115
Episode 433/1000, Total Reward: 3, Epsilon: 0.114
Episode 434/1000, Total Reward: 3, Epsilon: 0.114
Episode 435/1000, Total Reward: 3, Epsilon: 0.113
Episode 436/1000, Total Reward: 3, Epsilon: 0.112
Episode 437/1000, Total Reward: 3, Epsilon: 0.112
Episode 438/1000, Total Reward: 0, Epsilon: 0.111
Episode 439/1000, Total Reward: 2, Epsilon: 0.111
Episode 440/1000, Total Reward: 3, Epsilon: 0.110
Episode 441/1000, Total Reward: 1, Epsilon: 0.110
Episode 442/1000, Total Reward: 3, Epsilon: 0.109
Episode 443/1000, Total Reward: 3, Epsilon: 0.109
Episode 444/1000, Total Reward: 1, Epsilon: 0.108
Episode 445/1000, Total Reward: 2, Epsilon: 0.107
Episode 446/1000, Total Reward: 3, Epsilon: 0.107
Episode 447/1000, Total Reward: 1, Epsilon: 0.106
Episode 448/1000, Total Reward: 3, Epsilon: 0.106
Episode 449/1000, Total Reward: 1, Epsilon: 0.105
Episode 450/1000, Total Reward: 1, Epsilon: 0.105
Episode 451/1000, Total Reward: -1, Epsilon: 0.104
Episode 452/1000, Total Reward: 1, Epsilon: 0.104
Episode 453/1000, Total Reward: 3, Epsilon: 0.103
Episode 454/1000, Total Reward: 3, Epsilon: 0.103
Episode 455/1000, Total Reward: 3, Epsilon: 0.102
Episode 456/1000, Total Reward: 3, Epsilon: 0.102
Episode 457/1000, Total Reward: 3, Epsilon: 0.101
Episode 458/1000, Total Reward: 3, Epsilon: 0.101
Episode 459/1000, Total Reward: 1, Epsilon: 0.100
Episode 460/1000, Total Reward: -1, Epsilon: 0.100
Episode 461/1000, Total Reward: 3, Epsilon: 0.099
Episode 462/1000, Total Reward: 3, Epsilon: 0.099
Episode 463/1000, Total Reward: 1, Epsilon: 0.098
Episode 464/1000, Total Reward: 0, Epsilon: 0.098
Episode 465/1000, Total Reward: 2, Epsilon: 0.097
Episode 466/1000, Total Reward: 3, Epsilon: 0.097
Episode 467/1000, Total Reward: 3, Epsilon: 0.096
Episode 468/1000, Total Reward: 3, Epsilon: 0.096
Episode 469/1000, Total Reward: 2, Epsilon: 0.095
Episode 470/1000, Total Reward: 3, Epsilon: 0.095
Episode 471/1000, Total Reward: 1, Epsilon: 0.094
Episode 472/1000, Total Reward: 1, Epsilon: 0.094
Episode 473/1000, Total Reward: 3, Epsilon: 0.093
Episode 474/1000, Total Reward: 1, Epsilon: 0.093
Episode 475/1000, Total Reward: 3, Epsilon: 0.092
Episode 476/1000, Total Reward: 3, Epsilon: 0.092
Episode 477/1000, Total Reward: 3, Epsilon: 0.092
Episode 478/1000, Total Reward: 3, Epsilon: 0.091
Episode 479/1000, Total Reward: 3, Epsilon: 0.091
Episode 480/1000, Total Reward: 1, Epsilon: 0.090
```

```
Episode 481/1000, Total Reward: 1, Epsilon: 0.090
Episode 482/1000, Total Reward: 3, Epsilon: 0.089
Episode 483/1000, Total Reward: 3, Epsilon: 0.089
Episode 484/1000, Total Reward: 3, Epsilon: 0.088
Episode 485/1000, Total Reward: -1, Epsilon: 0.088
Episode 486/1000, Total Reward: 3, Epsilon: 0.088
Episode 487/1000, Total Reward: 3, Epsilon: 0.087
Episode 488/1000, Total Reward: 0, Epsilon: 0.087
Episode 489/1000, Total Reward: 3, Epsilon: 0.086
Episode 490/1000, Total Reward: 3, Epsilon: 0.086
Episode 491/1000, Total Reward: 3, Epsilon: 0.085
Episode 492/1000, Total Reward: 1, Epsilon: 0.085
Episode 493/1000, Total Reward: 3, Epsilon: 0.084
Episode 494/1000, Total Reward: 3, Epsilon: 0.084
Episode 495/1000, Total Reward: 3, Epsilon: 0.084
Episode 496/1000, Total Reward: 1, Epsilon: 0.083
Episode 497/1000, Total Reward: 3, Epsilon: 0.083
Episode 498/1000, Total Reward: 3, Epsilon: 0.082
Episode 499/1000, Total Reward: 1, Epsilon: 0.082
Episode 500/1000, Total Reward: 3, Epsilon: 0.082
Episode 501/1000, Total Reward: 1, Epsilon: 0.081
Episode 502/1000, Total Reward: 1, Epsilon: 0.081
Episode 503/1000, Total Reward: 3, Epsilon: 0.080
Episode 504/1000, Total Reward: 3, Epsilon: 0.080
Episode 505/1000, Total Reward: 3, Epsilon: 0.080
Episode 506/1000, Total Reward: 1, Epsilon: 0.079
Episode 507/1000, Total Reward: 3, Epsilon: 0.079
Episode 508/1000, Total Reward: 1, Epsilon: 0.078
Episode 509/1000, Total Reward: 3, Epsilon: 0.078
Episode 510/1000, Total Reward: 3, Epsilon: 0.078
Episode 511/1000, Total Reward: 3, Epsilon: 0.077
Episode 512/1000, Total Reward: 0, Epsilon: 0.077
Episode 513/1000, Total Reward: 1, Epsilon: 0.076
Episode 514/1000, Total Reward: 3, Epsilon: 0.076
Episode 515/1000, Total Reward: 3, Epsilon: 0.076
Episode 516/1000, Total Reward: 3, Epsilon: 0.075
Episode 517/1000, Total Reward: 3, Epsilon: 0.075
Episode 518/1000, Total Reward: 3, Epsilon: 0.075
Episode 519/1000, Total Reward: 2, Epsilon: 0.074
Episode 520/1000, Total Reward: 3, Epsilon: 0.074
Episode 521/1000, Total Reward: 1, Epsilon: 0.073
Episode 522/1000, Total Reward: 3, Epsilon: 0.073
Episode 523/1000, Total Reward: 3, Epsilon: 0.073
Episode 524/1000, Total Reward: 3, Epsilon: 0.072
Episode 525/1000, Total Reward: 3, Epsilon: 0.072
Episode 526/1000, Total Reward: 3, Epsilon: 0.072
Episode 527/1000, Total Reward: 3, Epsilon: 0.071
Episode 528/1000, Total Reward: 3, Epsilon: 0.071
Episode 529/1000, Total Reward: 3, Epsilon: 0.071
Episode 530/1000, Total Reward: 3, Epsilon: 0.070
Episode 531/1000, Total Reward: 3, Epsilon: 0.070
Episode 532/1000, Total Reward: 2, Epsilon: 0.069
Episode 533/1000, Total Reward: 3, Epsilon: 0.069
Episode 534/1000, Total Reward: 3, Epsilon: 0.069
Episode 535/1000, Total Reward: 3, Epsilon: 0.068
Episode 536/1000, Total Reward: 1, Epsilon: 0.068
Episode 537/1000, Total Reward: 1, Epsilon: 0.068
Episode 538/1000, Total Reward: 3, Epsilon: 0.067
Episode 539/1000, Total Reward: 1, Epsilon: 0.067
Episode 540/1000, Total Reward: 3, Epsilon: 0.067
```

```
Episode 541/1000, Total Reward: 3, Epsilon: 0.066
Episode 542/1000, Total Reward: 1, Epsilon: 0.066
Episode 543/1000, Total Reward: 3, Epsilon: 0.066
Episode 544/1000, Total Reward: 3, Epsilon: 0.065
Episode 545/1000, Total Reward: 3, Epsilon: 0.065
Episode 546/1000, Total Reward: 3, Epsilon: 0.065
Episode 547/1000, Total Reward: 1, Epsilon: 0.064
Episode 548/1000, Total Reward: 2, Epsilon: 0.064
Episode 549/1000, Total Reward: -1, Epsilon: 0.064
Episode 550/1000, Total Reward: 3, Epsilon: 0.063
Episode 551/1000, Total Reward: 3, Epsilon: 0.063
Episode 552/1000, Total Reward: 1, Epsilon: 0.063
Episode 553/1000, Total Reward: 3, Epsilon: 0.063
Episode 554/1000, Total Reward: 3, Epsilon: 0.062
Episode 555/1000, Total Reward: 3, Epsilon: 0.062
Episode 556/1000, Total Reward: 3, Epsilon: 0.062
Episode 557/1000, Total Reward: 3, Epsilon: 0.061
Episode 558/1000, Total Reward: 3, Epsilon: 0.061
Episode 559/1000, Total Reward: 3, Epsilon: 0.061
Episode 560/1000, Total Reward: 1, Epsilon: 0.060
Episode 561/1000, Total Reward: 1, Epsilon: 0.060
Episode 562/1000, Total Reward: 3, Epsilon: 0.060
Episode 563/1000, Total Reward: 3, Epsilon: 0.059
Episode 564/1000, Total Reward: 3, Epsilon: 0.059
Episode 565/1000, Total Reward: 3, Epsilon: 0.059
Episode 566/1000, Total Reward: -1, Epsilon: 0.059
Episode 567/1000, Total Reward: 3, Epsilon: 0.058
Episode 568/1000, Total Reward: 3, Epsilon: 0.058
Episode 569/1000, Total Reward: 3, Epsilon: 0.058
Episode 570/1000, Total Reward: 3, Epsilon: 0.057
Episode 571/1000, Total Reward: 3, Epsilon: 0.057
Episode 572/1000, Total Reward: 3, Epsilon: 0.057
Episode 573/1000, Total Reward: 2, Epsilon: 0.057
Episode 574/1000, Total Reward: 2, Epsilon: 0.056
Episode 575/1000, Total Reward: 3, Epsilon: 0.056
Episode 576/1000, Total Reward: 1, Epsilon: 0.056
Episode 577/1000, Total Reward: 3, Epsilon: 0.055
Episode 578/1000, Total Reward: 3, Epsilon: 0.055
Episode 579/1000, Total Reward: -1, Epsilon: 0.055
Episode 580/1000, Total Reward: 1, Epsilon: 0.055
Episode 581/1000, Total Reward: 3, Epsilon: 0.054
Episode 582/1000, Total Reward: 3, Epsilon: 0.054
Episode 583/1000, Total Reward: 3, Epsilon: 0.054
Episode 584/1000, Total Reward: 3, Epsilon: 0.054
Episode 585/1000, Total Reward: 2, Epsilon: 0.053
Episode 586/1000, Total Reward: 3, Epsilon: 0.053
Episode 587/1000, Total Reward: 3, Epsilon: 0.053
Episode 588/1000, Total Reward: -1, Epsilon: 0.052
Episode 589/1000, Total Reward: 3, Epsilon: 0.052
Episode 590/1000, Total Reward: 1, Epsilon: 0.052
Episode 591/1000, Total Reward: 3, Epsilon: 0.052
Episode 592/1000, Total Reward: 2, Epsilon: 0.051
Episode 593/1000, Total Reward: 3, Epsilon: 0.051
Episode 594/1000, Total Reward: 3, Epsilon: 0.051
Episode 595/1000, Total Reward: 2, Epsilon: 0.051
Episode 596/1000, Total Reward: 0, Epsilon: 0.050
Episode 597/1000, Total Reward: 3, Epsilon: 0.050
Episode 598/1000, Total Reward: 3, Epsilon: 0.050
Episode 599/1000, Total Reward: 3, Epsilon: 0.050
Episode 600/1000, Total Reward: 3, Epsilon: 0.049
```

```
Episode 601/1000, Total Reward: 3, Epsilon: 0.049
Episode 602/1000, Total Reward: 1, Epsilon: 0.049
Episode 603/1000, Total Reward: 3, Epsilon: 0.049
Episode 604/1000, Total Reward: 3, Epsilon: 0.048
Episode 605/1000, Total Reward: 3, Epsilon: 0.048
Episode 606/1000, Total Reward: 3, Epsilon: 0.048
Episode 607/1000, Total Reward: 3, Epsilon: 0.048
Episode 608/1000, Total Reward: 3, Epsilon: 0.047
Episode 609/1000, Total Reward: 3, Epsilon: 0.047
Episode 610/1000, Total Reward: 3, Epsilon: 0.047
Episode 611/1000, Total Reward: 3, Epsilon: 0.047
Episode 612/1000, Total Reward: 3, Epsilon: 0.047
Episode 613/1000, Total Reward: 3, Epsilon: 0.046
Episode 614/1000, Total Reward: 1, Epsilon: 0.046
Episode 615/1000, Total Reward: 3, Epsilon: 0.046
Episode 616/1000, Total Reward: 3, Epsilon: 0.046
Episode 617/1000, Total Reward: 3, Epsilon: 0.045
Episode 618/1000, Total Reward: 3, Epsilon: 0.045
Episode 619/1000, Total Reward: 3, Epsilon: 0.045
Episode 620/1000, Total Reward: 3, Epsilon: 0.045
Episode 621/1000, Total Reward: 3, Epsilon: 0.044
Episode 622/1000, Total Reward: 3, Epsilon: 0.044
Episode 623/1000, Total Reward: 3, Epsilon: 0.044
Episode 624/1000, Total Reward: 3, Epsilon: 0.044
Episode 625/1000, Total Reward: 3, Epsilon: 0.044
Episode 626/1000, Total Reward: 3, Epsilon: 0.043
Episode 627/1000, Total Reward: 3, Epsilon: 0.043
Episode 628/1000, Total Reward: 3, Epsilon: 0.043
Episode 629/1000, Total Reward: 3, Epsilon: 0.043
Episode 630/1000, Total Reward: 3, Epsilon: 0.043
Episode 631/1000, Total Reward: 3, Epsilon: 0.042
Episode 632/1000, Total Reward: 3, Epsilon: 0.042
Episode 633/1000, Total Reward: 3, Epsilon: 0.042
Episode 634/1000, Total Reward: 3, Epsilon: 0.042
Episode 635/1000, Total Reward: 3, Epsilon: 0.041
Episode 636/1000, Total Reward: 3, Epsilon: 0.041
Episode 637/1000, Total Reward: 3, Epsilon: 0.041
Episode 638/1000, Total Reward: 3, Epsilon: 0.041
Episode 639/1000, Total Reward: 3, Epsilon: 0.041
Episode 640/1000, Total Reward: 3, Epsilon: 0.040
Episode 641/1000, Total Reward: 3, Epsilon: 0.040
Episode 642/1000, Total Reward: 3, Epsilon: 0.040
Episode 643/1000, Total Reward: 3, Epsilon: 0.040
Episode 644/1000, Total Reward: 1, Epsilon: 0.040
Episode 645/1000, Total Reward: 2, Epsilon: 0.039
Episode 646/1000, Total Reward: 3, Epsilon: 0.039
Episode 647/1000, Total Reward: 3, Epsilon: 0.039
Episode 648/1000, Total Reward: 1, Epsilon: 0.039
Episode 649/1000, Total Reward: 3, Epsilon: 0.039
Episode 650/1000, Total Reward: 3, Epsilon: 0.038
Episode 651/1000, Total Reward: 1, Epsilon: 0.038
Episode 652/1000, Total Reward: 3, Epsilon: 0.038
Episode 653/1000, Total Reward: 3, Epsilon: 0.038
Episode 654/1000, Total Reward: 3, Epsilon: 0.038
Episode 655/1000, Total Reward: 3, Epsilon: 0.038
Episode 656/1000, Total Reward: 3, Epsilon: 0.037
Episode 657/1000, Total Reward: 3, Epsilon: 0.037
Episode 658/1000, Total Reward: 1, Epsilon: 0.037
Episode 659/1000, Total Reward: 3, Epsilon: 0.037
Episode 660/1000, Total Reward: 3, Epsilon: 0.037
```

```
Episode 661/1000, Total Reward: 3, Epsilon: 0.036
Episode 662/1000, Total Reward: 3, Epsilon: 0.036
Episode 663/1000, Total Reward: 3, Epsilon: 0.036
Episode 664/1000, Total Reward: 1, Epsilon: 0.036
Episode 665/1000, Total Reward: 1, Epsilon: 0.036
Episode 666/1000, Total Reward: 3, Epsilon: 0.035
Episode 667/1000, Total Reward: 1, Epsilon: 0.035
Episode 668/1000, Total Reward: 3, Epsilon: 0.035
Episode 669/1000, Total Reward: 3, Epsilon: 0.035
Episode 670/1000, Total Reward: 1, Epsilon: 0.035
Episode 671/1000, Total Reward: 3, Epsilon: 0.035
Episode 672/1000, Total Reward: 3, Epsilon: 0.034
Episode 673/1000, Total Reward: 3, Epsilon: 0.034
Episode 674/1000, Total Reward: 3, Epsilon: 0.034
Episode 675/1000, Total Reward: 3, Epsilon: 0.034
Episode 676/1000, Total Reward: 3, Epsilon: 0.034
Episode 677/1000, Total Reward: 3, Epsilon: 0.034
Episode 678/1000, Total Reward: 3, Epsilon: 0.033
Episode 679/1000, Total Reward: 1, Epsilon: 0.033
Episode 680/1000, Total Reward: 3, Epsilon: 0.033
Episode 681/1000, Total Reward: 3, Epsilon: 0.033
Episode 682/1000, Total Reward: 3, Epsilon: 0.033
Episode 683/1000, Total Reward: 1, Epsilon: 0.033
Episode 684/1000, Total Reward: 3, Epsilon: 0.032
Episode 685/1000, Total Reward: 3, Epsilon: 0.032
Episode 686/1000, Total Reward: 3, Epsilon: 0.032
Episode 687/1000, Total Reward: 3, Epsilon: 0.032
Episode 688/1000, Total Reward: 3, Epsilon: 0.032
Episode 689/1000, Total Reward: 3, Epsilon: 0.032
Episode 690/1000, Total Reward: 3, Epsilon: 0.031
Episode 691/1000, Total Reward: 3, Epsilon: 0.031
Episode 692/1000, Total Reward: 3, Epsilon: 0.031
Episode 693/1000, Total Reward: 3, Epsilon: 0.031
Episode 694/1000, Total Reward: 3, Epsilon: 0.031
Episode 695/1000, Total Reward: 3, Epsilon: 0.031
Episode 696/1000, Total Reward: 3, Epsilon: 0.031
Episode 697/1000, Total Reward: 3, Epsilon: 0.030
Episode 698/1000, Total Reward: 3, Epsilon: 0.030
Episode 699/1000, Total Reward: 2, Epsilon: 0.030
Episode 700/1000, Total Reward: 1, Epsilon: 0.030
Episode 701/1000, Total Reward: 1, Epsilon: 0.030
Episode 702/1000, Total Reward: 2, Epsilon: 0.030
Episode 703/1000, Total Reward: 3, Epsilon: 0.029
Episode 704/1000, Total Reward: 3, Epsilon: 0.029
Episode 705/1000, Total Reward: 1, Epsilon: 0.029
Episode 706/1000, Total Reward: 3, Epsilon: 0.029
Episode 707/1000, Total Reward: 1, Epsilon: 0.029
Episode 708/1000, Total Reward: 3, Epsilon: 0.029
Episode 709/1000, Total Reward: 3, Epsilon: 0.029
Episode 710/1000, Total Reward: 3, Epsilon: 0.028
Episode 711/1000, Total Reward: 3, Epsilon: 0.028
Episode 712/1000, Total Reward: 3, Epsilon: 0.028
Episode 713/1000, Total Reward: 3, Epsilon: 0.028
Episode 714/1000, Total Reward: 3, Epsilon: 0.028
Episode 715/1000, Total Reward: 3, Epsilon: 0.028
Episode 716/1000, Total Reward: 3, Epsilon: 0.028
Episode 717/1000, Total Reward: 3, Epsilon: 0.027
Episode 718/1000, Total Reward: 3, Epsilon: 0.027
Episode 719/1000, Total Reward: 3, Epsilon: 0.027
Episode 720/1000, Total Reward: 2, Epsilon: 0.027
```

```
Episode 721/1000, Total Reward: 3, Epsilon: 0.027
Episode 722/1000, Total Reward: 3, Epsilon: 0.027
Episode 723/1000, Total Reward: 3, Epsilon: 0.027
Episode 724/1000, Total Reward: 3, Epsilon: 0.027
Episode 725/1000, Total Reward: 3, Epsilon: 0.026
Episode 726/1000, Total Reward: 3, Epsilon: 0.026
Episode 727/1000, Total Reward: 3, Epsilon: 0.026
Episode 728/1000, Total Reward: 3, Epsilon: 0.026
Episode 729/1000, Total Reward: 3, Epsilon: 0.026
Episode 730/1000, Total Reward: 3, Epsilon: 0.026
Episode 731/1000, Total Reward: 3, Epsilon: 0.026
Episode 732/1000, Total Reward: 3, Epsilon: 0.025
Episode 733/1000, Total Reward: 3, Epsilon: 0.025
Episode 734/1000, Total Reward: 3, Epsilon: 0.025
Episode 735/1000, Total Reward: 3, Epsilon: 0.025
Episode 736/1000, Total Reward: 3, Epsilon: 0.025
Episode 737/1000, Total Reward: 3, Epsilon: 0.025
Episode 738/1000, Total Reward: 3, Epsilon: 0.025
Episode 739/1000, Total Reward: 3, Epsilon: 0.025
Episode 740/1000, Total Reward: 3, Epsilon: 0.024
Episode 741/1000, Total Reward: 3, Epsilon: 0.024
Episode 742/1000, Total Reward: 3, Epsilon: 0.024
Episode 743/1000, Total Reward: 3, Epsilon: 0.024
Episode 744/1000, Total Reward: 3, Epsilon: 0.024
Episode 745/1000, Total Reward: 3, Epsilon: 0.024
Episode 746/1000, Total Reward: 3, Epsilon: 0.024
Episode 747/1000, Total Reward: 3, Epsilon: 0.024
Episode 748/1000, Total Reward: 3, Epsilon: 0.024
Episode 749/1000, Total Reward: 3, Epsilon: 0.023
Episode 750/1000, Total Reward: 3, Epsilon: 0.023
Episode 751/1000, Total Reward: 3, Epsilon: 0.023
Episode 752/1000, Total Reward: 3, Epsilon: 0.023
Episode 753/1000, Total Reward: 1, Epsilon: 0.023
Episode 754/1000, Total Reward: 3, Epsilon: 0.023
Episode 755/1000, Total Reward: 3, Epsilon: 0.023
Episode 756/1000, Total Reward: 3, Epsilon: 0.023
Episode 757/1000, Total Reward: 3, Epsilon: 0.022
Episode 758/1000, Total Reward: 3, Epsilon: 0.022
Episode 759/1000, Total Reward: 3, Epsilon: 0.022
Episode 760/1000, Total Reward: 3, Epsilon: 0.022
Episode 761/1000, Total Reward: 3, Epsilon: 0.022
Episode 762/1000, Total Reward: 1, Epsilon: 0.022
Episode 763/1000, Total Reward: 3, Epsilon: 0.022
Episode 764/1000, Total Reward: 3, Epsilon: 0.022
Episode 765/1000, Total Reward: 3, Epsilon: 0.022
Episode 766/1000, Total Reward: 3, Epsilon: 0.022
Episode 767/1000, Total Reward: 3, Epsilon: 0.021
Episode 768/1000, Total Reward: 3, Epsilon: 0.021
Episode 769/1000, Total Reward: 3, Epsilon: 0.021
Episode 770/1000, Total Reward: 3, Epsilon: 0.021
Episode 771/1000, Total Reward: 3, Epsilon: 0.021
Episode 772/1000, Total Reward: 3, Epsilon: 0.021
Episode 773/1000, Total Reward: 3, Epsilon: 0.021
Episode 774/1000, Total Reward: 3, Epsilon: 0.021
Episode 775/1000, Total Reward: 3, Epsilon: 0.021
Episode 776/1000, Total Reward: 3, Epsilon: 0.020
Episode 777/1000, Total Reward: 3, Epsilon: 0.020
Episode 778/1000, Total Reward: 3, Epsilon: 0.020
Episode 779/1000, Total Reward: 3, Epsilon: 0.020
Episode 780/1000, Total Reward: 3, Epsilon: 0.020
```

```
Episode 781/1000, Total Reward: 3, Epsilon: 0.020
Episode 782/1000, Total Reward: 3, Epsilon: 0.020
Episode 783/1000, Total Reward: 3, Epsilon: 0.020
Episode 784/1000, Total Reward: 3, Epsilon: 0.020
Episode 785/1000, Total Reward: 3, Epsilon: 0.020
Episode 786/1000, Total Reward: 3, Epsilon: 0.019
Episode 787/1000, Total Reward: 3, Epsilon: 0.019
Episode 788/1000, Total Reward: 3, Epsilon: 0.019
Episode 789/1000, Total Reward: 3, Epsilon: 0.019
Episode 790/1000, Total Reward: 3, Epsilon: 0.019
Episode 791/1000, Total Reward: 3, Epsilon: 0.019
Episode 792/1000, Total Reward: 3, Epsilon: 0.019
Episode 793/1000, Total Reward: 3, Epsilon: 0.019
Episode 794/1000, Total Reward: 3, Epsilon: 0.019
Episode 795/1000, Total Reward: 3, Epsilon: 0.019
Episode 796/1000, Total Reward: 3, Epsilon: 0.019
Episode 797/1000, Total Reward: 3, Epsilon: 0.018
Episode 798/1000, Total Reward: 3, Epsilon: 0.018
Episode 799/1000, Total Reward: 3, Epsilon: 0.018
Episode 800/1000, Total Reward: 3, Epsilon: 0.018
Episode 801/1000, Total Reward: 3, Epsilon: 0.018
Episode 802/1000, Total Reward: 0, Epsilon: 0.018
Episode 803/1000, Total Reward: 3, Epsilon: 0.018
Episode 804/1000, Total Reward: 3, Epsilon: 0.018
Episode 805/1000, Total Reward: 3, Epsilon: 0.018
Episode 806/1000, Total Reward: 3, Epsilon: 0.018
Episode 807/1000, Total Reward: 3, Epsilon: 0.018
Episode 808/1000, Total Reward: 3, Epsilon: 0.017
Episode 809/1000, Total Reward: 3, Epsilon: 0.017
Episode 810/1000, Total Reward: 3, Epsilon: 0.017
Episode 811/1000, Total Reward: 3, Epsilon: 0.017
Episode 812/1000, Total Reward: 3, Epsilon: 0.017
Episode 813/1000, Total Reward: 3, Epsilon: 0.017
Episode 814/1000, Total Reward: 3, Epsilon: 0.017
Episode 815/1000, Total Reward: 3, Epsilon: 0.017
Episode 816/1000, Total Reward: 3, Epsilon: 0.017
Episode 817/1000, Total Reward: 3, Epsilon: 0.017
Episode 818/1000, Total Reward: 3, Epsilon: 0.017
Episode 819/1000, Total Reward: 3, Epsilon: 0.016
Episode 820/1000, Total Reward: 3, Epsilon: 0.016
Episode 821/1000, Total Reward: 3, Epsilon: 0.016
Episode 822/1000, Total Reward: 3, Epsilon: 0.016
Episode 823/1000, Total Reward: 3, Epsilon: 0.016
Episode 824/1000, Total Reward: 3, Epsilon: 0.016
Episode 825/1000, Total Reward: 3, Epsilon: 0.016
Episode 826/1000, Total Reward: 3, Epsilon: 0.016
Episode 827/1000, Total Reward: 3, Epsilon: 0.016
Episode 828/1000, Total Reward: 3, Epsilon: 0.016
Episode 829/1000, Total Reward: 3, Epsilon: 0.016
Episode 830/1000, Total Reward: 3, Epsilon: 0.016
Episode 831/1000, Total Reward: 3, Epsilon: 0.016
Episode 832/1000, Total Reward: 3, Epsilon: 0.015
Episode 833/1000, Total Reward: 3, Epsilon: 0.015
Episode 834/1000, Total Reward: 3, Epsilon: 0.015
Episode 835/1000, Total Reward: 3, Epsilon: 0.015
Episode 836/1000, Total Reward: 3, Epsilon: 0.015
Episode 837/1000, Total Reward: 3, Epsilon: 0.015
Episode 838/1000, Total Reward: 3, Epsilon: 0.015
Episode 839/1000, Total Reward: 3, Epsilon: 0.015
Episode 840/1000, Total Reward: 3, Epsilon: 0.015
```

```
Episode 841/1000, Total Reward: 3, Epsilon: 0.015
Episode 842/1000, Total Reward: 3, Epsilon: 0.015
Episode 843/1000, Total Reward: 3, Epsilon: 0.015
Episode 844/1000, Total Reward: 3, Epsilon: 0.015
Episode 845/1000, Total Reward: 3, Epsilon: 0.014
Episode 846/1000, Total Reward: 3, Epsilon: 0.014
Episode 847/1000, Total Reward: 3, Epsilon: 0.014
Episode 848/1000, Total Reward: 3, Epsilon: 0.014
Episode 849/1000, Total Reward: 3, Epsilon: 0.014
Episode 850/1000, Total Reward: 3, Epsilon: 0.014
Episode 851/1000, Total Reward: 3, Epsilon: 0.014
Episode 852/1000, Total Reward: 3, Epsilon: 0.014
Episode 853/1000, Total Reward: 3, Epsilon: 0.014
Episode 854/1000, Total Reward: 3, Epsilon: 0.014
Episode 855/1000, Total Reward: 3, Epsilon: 0.014
Episode 856/1000, Total Reward: 3, Epsilon: 0.014
Episode 857/1000, Total Reward: 3, Epsilon: 0.014
Episode 858/1000, Total Reward: 3, Epsilon: 0.014
Episode 859/1000, Total Reward: 3, Epsilon: 0.013
Episode 860/1000, Total Reward: 3, Epsilon: 0.013
Episode 861/1000, Total Reward: 3, Epsilon: 0.013
Episode 862/1000, Total Reward: 3, Epsilon: 0.013
Episode 863/1000, Total Reward: 3, Epsilon: 0.013
Episode 864/1000, Total Reward: 3, Epsilon: 0.013
Episode 865/1000, Total Reward: 3, Epsilon: 0.013
Episode 866/1000, Total Reward: 3, Epsilon: 0.013
Episode 867/1000, Total Reward: 3, Epsilon: 0.013
Episode 868/1000, Total Reward: 3, Epsilon: 0.013
Episode 869/1000, Total Reward: 3, Epsilon: 0.013
Episode 870/1000, Total Reward: 3, Epsilon: 0.013
Episode 871/1000, Total Reward: 3, Epsilon: 0.013
Episode 872/1000, Total Reward: 3, Epsilon: 0.013
Episode 873/1000, Total Reward: 3, Epsilon: 0.013
Episode 874/1000, Total Reward: 3, Epsilon: 0.013
Episode 875/1000, Total Reward: 3, Epsilon: 0.012
Episode 876/1000, Total Reward: 3, Epsilon: 0.012
Episode 877/1000, Total Reward: 3, Epsilon: 0.012
Episode 878/1000, Total Reward: 3, Epsilon: 0.012
Episode 879/1000, Total Reward: 3, Epsilon: 0.012
Episode 880/1000, Total Reward: 3, Epsilon: 0.012
Episode 881/1000, Total Reward: 3, Epsilon: 0.012
Episode 882/1000, Total Reward: 3, Epsilon: 0.012
Episode 883/1000, Total Reward: 3, Epsilon: 0.012
Episode 884/1000, Total Reward: 3, Epsilon: 0.012
Episode 885/1000, Total Reward: 3, Epsilon: 0.012
Episode 886/1000, Total Reward: 3, Epsilon: 0.012
Episode 887/1000, Total Reward: 3, Epsilon: 0.012
Episode 888/1000, Total Reward: 3, Epsilon: 0.012
Episode 889/1000, Total Reward: 3, Epsilon: 0.012
Episode 890/1000, Total Reward: 3, Epsilon: 0.012
Episode 891/1000, Total Reward: 3, Epsilon: 0.011
Episode 892/1000, Total Reward: 3, Epsilon: 0.011
Episode 893/1000, Total Reward: 3, Epsilon: 0.011
Episode 894/1000, Total Reward: 3, Epsilon: 0.011
Episode 895/1000, Total Reward: 3, Epsilon: 0.011
Episode 896/1000, Total Reward: 3, Epsilon: 0.011
Episode 897/1000, Total Reward: 3, Epsilon: 0.011
Episode 898/1000, Total Reward: 3, Epsilon: 0.011
Episode 899/1000, Total Reward: 3, Epsilon: 0.011
Episode 900/1000, Total Reward: 3, Epsilon: 0.011
```

```
Episode 901/1000, Total Reward: 3, Epsilon: 0.011
Episode 902/1000, Total Reward: 3, Epsilon: 0.011
Episode 903/1000, Total Reward: 3, Epsilon: 0.011
Episode 904/1000, Total Reward: 3, Epsilon: 0.011
Episode 905/1000, Total Reward: 3, Epsilon: 0.011
Episode 906/1000, Total Reward: 3, Epsilon: 0.011
Episode 907/1000, Total Reward: 3, Epsilon: 0.011
Episode 908/1000, Total Reward: 3, Epsilon: 0.011
Episode 909/1000, Total Reward: 3, Epsilon: 0.010
Episode 910/1000, Total Reward: 3, Epsilon: 0.010
Episode 911/1000, Total Reward: 1, Epsilon: 0.010
Episode 912/1000, Total Reward: 3, Epsilon: 0.010
Episode 913/1000, Total Reward: 1, Epsilon: 0.010
Episode 914/1000, Total Reward: 3, Epsilon: 0.010
Episode 915/1000, Total Reward: 3, Epsilon: 0.010
Episode 916/1000, Total Reward: 3, Epsilon: 0.010
Episode 917/1000, Total Reward: 3, Epsilon: 0.010
Episode 918/1000, Total Reward: 3, Epsilon: 0.010
Episode 919/1000, Total Reward: 2, Epsilon: 0.010
Episode 920/1000, Total Reward: 3, Epsilon: 0.010
Episode 921/1000, Total Reward: 1, Epsilon: 0.010
Episode 922/1000, Total Reward: 3, Epsilon: 0.010
Episode 923/1000, Total Reward: 3, Epsilon: 0.010
Episode 924/1000, Total Reward: 3, Epsilon: 0.010
Episode 925/1000, Total Reward: 3, Epsilon: 0.010
Episode 926/1000, Total Reward: 3, Epsilon: 0.010
Episode 927/1000, Total Reward: 3, Epsilon: 0.010
Episode 928/1000, Total Reward: 3, Epsilon: 0.010
Episode 929/1000, Total Reward: 3, Epsilon: 0.010
Episode 930/1000, Total Reward: 3, Epsilon: 0.010
Episode 931/1000, Total Reward: 3, Epsilon: 0.010
Episode 932/1000, Total Reward: 3, Epsilon: 0.010
Episode 933/1000, Total Reward: 3, Epsilon: 0.010
Episode 934/1000, Total Reward: 3, Epsilon: 0.010
Episode 935/1000, Total Reward: 3, Epsilon: 0.010
Episode 936/1000, Total Reward: 3, Epsilon: 0.010
Episode 937/1000, Total Reward: 3, Epsilon: 0.010
Episode 938/1000, Total Reward: 1, Epsilon: 0.010
Episode 939/1000, Total Reward: 3, Epsilon: 0.010
Episode 940/1000, Total Reward: 3, Epsilon: 0.010
Episode 941/1000, Total Reward: 3, Epsilon: 0.010
Episode 942/1000, Total Reward: 3, Epsilon: 0.010
Episode 943/1000, Total Reward: 3, Epsilon: 0.010
Episode 944/1000, Total Reward: 3, Epsilon: 0.010
Episode 945/1000, Total Reward: 3, Epsilon: 0.010
Episode 946/1000, Total Reward: 3, Epsilon: 0.010
Episode 947/1000, Total Reward: 3, Epsilon: 0.010
Episode 948/1000, Total Reward: 3, Epsilon: 0.010
Episode 949/1000, Total Reward: 3, Epsilon: 0.010
Episode 950/1000, Total Reward: 3, Epsilon: 0.010
Episode 951/1000, Total Reward: 1, Epsilon: 0.010
Episode 952/1000, Total Reward: 3, Epsilon: 0.010
Episode 953/1000, Total Reward: 3, Epsilon: 0.010
Episode 954/1000, Total Reward: 3, Epsilon: 0.010
Episode 955/1000, Total Reward: 3, Epsilon: 0.010
Episode 956/1000, Total Reward: 3, Epsilon: 0.010
Episode 957/1000, Total Reward: 3, Epsilon: 0.010
Episode 958/1000, Total Reward: 3, Epsilon: 0.010
Episode 959/1000, Total Reward: 3, Epsilon: 0.010
Episode 960/1000, Total Reward: 3, Epsilon: 0.010
```

```
Episode 961/1000, Total Reward: 3, Epsilon: 0.010
Episode 962/1000, Total Reward: 3, Epsilon: 0.010
Episode 963/1000, Total Reward: 3, Epsilon: 0.010
Episode 964/1000, Total Reward: 3, Epsilon: 0.010
Episode 965/1000, Total Reward: 3, Epsilon: 0.010
Episode 966/1000, Total Reward: 3, Epsilon: 0.010
Episode 967/1000, Total Reward: 3, Epsilon: 0.010
Episode 968/1000, Total Reward: 3, Epsilon: 0.010
Episode 969/1000, Total Reward: 3, Epsilon: 0.010
Episode 970/1000, Total Reward: 3, Epsilon: 0.010
Episode 971/1000, Total Reward: 3, Epsilon: 0.010
Episode 972/1000, Total Reward: 3, Epsilon: 0.010
Episode 973/1000, Total Reward: 3, Epsilon: 0.010
Episode 974/1000, Total Reward: 3, Epsilon: 0.010
Episode 975/1000, Total Reward: 3, Epsilon: 0.010
Episode 976/1000, Total Reward: 3, Epsilon: 0.010
Episode 977/1000, Total Reward: 3, Epsilon: 0.010
Episode 978/1000, Total Reward: 3, Epsilon: 0.010
Episode 979/1000, Total Reward: 3, Epsilon: 0.010
Episode 980/1000, Total Reward: 3, Epsilon: 0.010
Episode 981/1000, Total Reward: 1, Epsilon: 0.010
Episode 982/1000, Total Reward: 2, Epsilon: 0.010
Episode 983/1000, Total Reward: 3, Epsilon: 0.010
Episode 984/1000, Total Reward: 3, Epsilon: 0.010
Episode 985/1000, Total Reward: 3, Epsilon: 0.010
Episode 986/1000, Total Reward: 3, Epsilon: 0.010
Episode 987/1000, Total Reward: 3, Epsilon: 0.010
Episode 988/1000, Total Reward: 3, Epsilon: 0.010
Episode 989/1000, Total Reward: 3, Epsilon: 0.010
Episode 990/1000, Total Reward: 3, Epsilon: 0.010
Episode 991/1000, Total Reward: 3, Epsilon: 0.010
Episode 992/1000, Total Reward: 3, Epsilon: 0.010
Episode 993/1000, Total Reward: 2, Epsilon: 0.010
Episode 994/1000, Total Reward: 3, Epsilon: 0.010
Episode 995/1000, Total Reward: 3, Epsilon: 0.010
Episode 996/1000, Total Reward: 3, Epsilon: 0.010
Episode 997/1000, Total Reward: 3, Epsilon: 0.010
Episode 998/1000, Total Reward: 3, Epsilon: 0.010
Episode 999/1000, Total Reward: 3, Epsilon: 0.010
Episode 1000/1000, Total Reward: 3, Epsilon: 0.010
[[-1.4163967  -0.434062   -1.52561672 -0.55010143]
 [-2.52703448  0.62121926 -2.9101423  -1.30928724]
 [-1.46736703  1.76695103 -2.86260561 -1.71526663]
 [-1.80222112  0.65998242 -2.34803275 -1.82011844]
 [-1.7298786   0.30536656 -1.55011864 -1.67812335]
 [-1.51815257  0.18806647 -0.51843195  0.62882   ]
 [-0.77574664  1.76268646 -0.72830364  1.8098    ]
 [ 0.41994899  3.122       0.58045585  2.53404508]
 [-1.78416781  4.44234124 -0.41086203  1.6839014 ]
 [-1.50439835  5.004562   -0.73545128  0.18728279]
 [-1.83585111 -1.34724944 -1.68649366  1.70197393]
 [-0.72827102  1.68308454 -1.07059182  3.12189374]
 [ 1.72236102  4.58        1.72870357  4.52922463]
 [ 0.92679633  6.19808874  1.50113465  4.73371843]
 [ 1.22653625  7.89988244  1.96937222  3.6392814 ]
 [-2.2006532  -1.35558963 -1.69383116  1.35329462]
 [ 0.25959803  2.14371451 -1.68697865  4.57308614]
 [ 3.1069141   6.2         3.0623283   6.16612433]
 [ 3.02034656  7.66635358  3.50378617  7.99999377]
 [ 5.22365064  9.99999957  5.18642772  7.4978404 ]
```

```
[-1.64187712 -1.35935714 -1.41385867  2.11133279]
[ 0.87784478  1.30891422 -0.72460857  6.19012049]
[ 4.56826726  6.16176901  4.51768745  8.        ]
[ 6.13078295  7.94714063  6.15273795 10.        ]
[ 0.          0.          0.          0.        ]]
```

DQN算法

基本内容

DQN是结合Q-learing和深度学习神经网络的强化学习算法，通过神经网络来近似Q函数。神经网络接受状态s_t作为输入，输出每个可能动作a_t的Q值。同时使用了下面几种改进技术：

1. 经验回放

- 在Q-learing中，智能体根据当前状态和动作更新Q值，每次更新依赖于当前的状态，导致了样本之间的相关性
- 经验回放通过存储智能体的经历(state, action, reward, next_state)，智能体可以随即地从中采样

2. 目标网络

- Q-learing中，Q值更新依赖于最大化下一个状态的Q值，$max_{a'}Q(s_{t+1},a')$
- DQN引入了目标网络，是当前Q网络的一个副本，目标网络的参数更新慢

算法流程

1. 初始化
    - 初始化Q网络和目标网络，Q网络用于估计Q值，目标网络用于计算目标Q值
    - 初始化经验回放缓冲区
2. 探索和训练
    - 在每个时间步，选择一个动作，使用贪心算法
    - 执行动作，获取下一个状态和奖励
    - 存储经历到经验回放缓冲区
    - 从经验回放缓冲区中随机采样一小批经历
    - 使用神经网络和目标网络来计算目标Q值，更新Q网络的权重
3. 更新目标网络
    - 每隔一定步数，将Q网络的权重复制到目标网络中
    - 进行重复训练迭代

核心原理

假设用$Q(s,a;\theta)$表示Q网络，其中θ是Q网络的参数，希望最小化损失函数：

$$L(\theta) = \mathbb{E}_{(s_t,a_t,r_t,s_{t+1})\sim\mathcal{D}}\left[\left(r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a'; \theta^-) - Q(s_t, a_t; \theta)\right)^2\right]$$

其中：

- $Q(s_t, a_t; \theta)$是当前Q网络的Q值
- $Q(s_{t+1}, a'; \theta^-)$是目标网络的Q值
- D是经验回放池

训练过程中，神经网络通过优化损失函数$L(\theta)$来更新网络的参数θ，从而使得Q网络的输出尽可能接近期望的目标Q值。损失函数的最小化使用反向传播算法完成，计算损失函数的梯度并利用梯度下降更新参数，同时损失函数中利用了$\mathbb{E}$期望，表示计算目标时要从经验回放池中随机采样一批经历

DQN示例

针对环境**CartPole-v1**实现一个DQN算法，利用深度Q网络学习最优策略。**CartPole-v1**任务为agent控制小车的左右运动，使杆子竖直不倒下。每一个回合`max_step=500`，杆子倒下或者超过最大步数，回合结束。agent在每一个回合内的目标是保持杆子尽可能长的时间

1. 状态空间：
   - x：推车的位置
   - x_dot：推车的速度
   - theta：杆子的角度
   - theta_dot：杆子的角速度
2. 动作空间：
   - 0：推车向左移动
   - 1：推车向右移动
3. 奖励：
   - agent每一步成功使杆子不倒下，reward+1
   - 当杆子倒下或推车移出界限时，回合结束，奖励为0，环境进入结束状态

In [3]:
```
import gym
import numpy as np
import torch
import torch.nn as nn
import torch.optim as optim
import random
from collections import deque


class QNetwork(nn.Module):
    def __init__(self, state_dim, action_dim):
        super(QNetwork, self).__init__()
        self.fc1 = nn.Linear(state_dim, 64)
        self.fc2 = nn.Linear(64, 64)
        self.fc3 = nn.Linear(64, action_dim)

    def forward(self, state):
        x = torch.relu(self.fc1(state))
        x = torch.relu(self.fc2(x))
```

```
        return self.fc3(x)

# env = gym.make('CartPole-v1')

class DQNAgent:
    def __init__(self, env, gamma=0.99, epsilon=1.0, epsilon_decay=0.995, min_ep
        self.env = env
        self.gamma = gamma
        self.epsilon = epsilon
        self.epsilon_decay = epsilon_decay
        self.min_epsilon = min_epsilon
        self.alpha = alpha
        self.batch_size = batch_size
        self.memory = deque(maxlen=memory_size)

        device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
        self.device = device

        # 当前Q网络与目标Q网络
        self.q_network = QNetwork(env.observation_space.shape[0], env.action_spa
        self.target_network = QNetwork(env.observation_space.shape[0], env.actio
        self.optimizer = optim.Adam(self.q_network.parameters(), lr=alpha)
        self.update_target_network()

    def choose_action(self, state):
        state = np.array(state)
        state = torch.tensor(state, dtype=torch.float32).to(self.device)
        if random.random() < self.epsilon:
            return self.env.action_space.sample()
        q_values = self.q_network(state)
        return torch.argmax(q_values).item()

    def store_experience(self, state, action, reward, next_state, done):
        self.memory.append((state, action, reward, next_state, done))

    def update_target_network(self):
        self.target_network.load_state_dict(self.q_network.state_dict())

    def sample_batch(self):
        return random.sample(self.memory, self.batch_size)

    def train(self):
        if len(self.memory) < self.batch_size:
            return

        batch = self.sample_batch()
        states, actions, rewards, next_states, dones = zip(*batch)

        states = np.array(states)
        next_states = np.array(next_states)

        states = torch.tensor(states, dtype=torch.float32).to(self.device)
        actions = torch.tensor(actions, dtype=torch.long).to(self.device)
        rewards = torch.tensor(rewards, dtype=torch.float32).to(self.device)
        next_states = torch.tensor(next_states, dtype=torch.float32).to(self.dev
        dones = torch.tensor(dones, dtype=torch.float32).to(self.device)

        # 计算当前Q网络输出的Q值
        q_values = self.q_network(states).gather(1, actions.view(-1, 1)).squeeze
```

```
        # 计算目标Q网络的输出Q值
        next_q_values = self.target_network(next_states).max(1)[0]
        target_q_values = rewards + self.gamma * next_q_values * (1 - dones)

        # Loss
        loss = nn.MSELoss()(q_values, target_q_values)

        self.optimizer.zero_grad()
        loss.backward()
        self.optimizer.step()

        if self.epsilon > self.min_epsilon:
            self.epsilon *= self.epsilon_decay

def save_model(agent, filename="dqn_cartpole_v1.pth"):
    torch.save({
        'q_network': agent.q_network.state_dict(),
        'target_network': agent.target_network.state_dict(),
        'epsilon': agent.epsilon,
    }, filename)
    print(f"Model saved to {filename}")


def train_dqn(episodes=1000):
    env = gym.make('CartPole-v1')
    agent = DQNAgent(env)

    for episode in range(episodes):
        state, info = env.reset()
        done = False
        total_reward = 0

        while not done:
            action = agent.choose_action(state)
            next_state, reward, terminated, truncated, info = env.step(action)
            done = terminated or truncated

            agent.store_experience(state, action, reward, next_state, done)
            agent.train()
            state = next_state
            total_reward += reward

        if episode % 10 == 0:
            agent.update_target_network()
            print(f"Episode {episode + 1}/{episodes}, Total Reward: {total_rewar

    save_model(agent)

# 训练
train_dqn(1000)


def load_model(agent, filename="dqn_cartpole_v1.pth"):
    checkpoint = torch.load(filename)
    agent.q_network.load_state_dict(checkpoint['q_network'])
    agent.target_network.load_state_dict(checkpoint['target_network'])
    agent.epsilon = checkpoint['epsilon']
    print(f"Model loaded from {filename}")

def run_inference(agent, env):
```

```
    state, info = env.reset()
    done = False
    total_reward = 0

    while not done:
        action = agent.choose_action(state)
        next_state, reward, terminated, truncated, info = env.step(action)
        done = terminated or truncated

        total_reward += reward
        state = next_state
        env.render()

    print(f"Total reward in the test run: {total_reward}")
    env.close()


def test_dqn():
    env = gym.make('CartPole-v1')
    agent = DQNAgent(env)

    load_model(agent)

    run_inference(agent, env)

test_dqn()
```

```
d:\Anaconda\envs\transformer\lib\site-packages\gym\utils\passive_env_checker.py:2
33: DeprecationWarning: `np.bool8` is a deprecated alias for `np.bool_`.  (Deprec
ated NumPy 1.24)
  if not isinstance(terminated, (bool, np.bool8)):
```

```
Episode 1/1000, Total Reward: 11.0, Epsilon: 1.000
Episode 11/1000, Total Reward: 14.0, Epsilon: 0.506
Episode 21/1000, Total Reward: 14.0, Epsilon: 0.243
Episode 31/1000, Total Reward: 10.0, Epsilon: 0.132
Episode 41/1000, Total Reward: 24.0, Epsilon: 0.052
Episode 51/1000, Total Reward: 393.0, Epsilon: 0.010
Episode 61/1000, Total Reward: 122.0, Epsilon: 0.010
Episode 71/1000, Total Reward: 131.0, Epsilon: 0.010
Episode 81/1000, Total Reward: 152.0, Epsilon: 0.010
Episode 91/1000, Total Reward: 175.0, Epsilon: 0.010
Episode 101/1000, Total Reward: 219.0, Epsilon: 0.010
Episode 111/1000, Total Reward: 126.0, Epsilon: 0.010
Episode 121/1000, Total Reward: 197.0, Epsilon: 0.010
Episode 131/1000, Total Reward: 179.0, Epsilon: 0.010
Episode 141/1000, Total Reward: 249.0, Epsilon: 0.010
Episode 151/1000, Total Reward: 214.0, Epsilon: 0.010
Episode 161/1000, Total Reward: 247.0, Epsilon: 0.010
Episode 171/1000, Total Reward: 500.0, Epsilon: 0.010
Episode 181/1000, Total Reward: 373.0, Epsilon: 0.010
Episode 191/1000, Total Reward: 279.0, Epsilon: 0.010
Episode 201/1000, Total Reward: 30.0, Epsilon: 0.010
Episode 211/1000, Total Reward: 272.0, Epsilon: 0.010
Episode 221/1000, Total Reward: 225.0, Epsilon: 0.010
Episode 231/1000, Total Reward: 171.0, Epsilon: 0.010
Episode 241/1000, Total Reward: 209.0, Epsilon: 0.010
Episode 251/1000, Total Reward: 207.0, Epsilon: 0.010
Episode 261/1000, Total Reward: 197.0, Epsilon: 0.010
Episode 271/1000, Total Reward: 206.0, Epsilon: 0.010
Episode 281/1000, Total Reward: 406.0, Epsilon: 0.010
Episode 291/1000, Total Reward: 210.0, Epsilon: 0.010
Episode 301/1000, Total Reward: 258.0, Epsilon: 0.010
Episode 311/1000, Total Reward: 158.0, Epsilon: 0.010
Episode 321/1000, Total Reward: 303.0, Epsilon: 0.010
Episode 331/1000, Total Reward: 300.0, Epsilon: 0.010
Episode 341/1000, Total Reward: 241.0, Epsilon: 0.010
Episode 351/1000, Total Reward: 169.0, Epsilon: 0.010
Episode 361/1000, Total Reward: 160.0, Epsilon: 0.010
Episode 371/1000, Total Reward: 212.0, Epsilon: 0.010
Episode 381/1000, Total Reward: 208.0, Epsilon: 0.010
Episode 391/1000, Total Reward: 176.0, Epsilon: 0.010
Episode 401/1000, Total Reward: 263.0, Epsilon: 0.010
Episode 411/1000, Total Reward: 255.0, Epsilon: 0.010
Episode 421/1000, Total Reward: 339.0, Epsilon: 0.010
Episode 431/1000, Total Reward: 274.0, Epsilon: 0.010
Episode 441/1000, Total Reward: 188.0, Epsilon: 0.010
Episode 451/1000, Total Reward: 147.0, Epsilon: 0.010
Episode 461/1000, Total Reward: 165.0, Epsilon: 0.010
Episode 471/1000, Total Reward: 186.0, Epsilon: 0.010
Episode 481/1000, Total Reward: 183.0, Epsilon: 0.010
Episode 491/1000, Total Reward: 151.0, Epsilon: 0.010
Episode 501/1000, Total Reward: 163.0, Epsilon: 0.010
Episode 511/1000, Total Reward: 163.0, Epsilon: 0.010
Episode 521/1000, Total Reward: 283.0, Epsilon: 0.010
Episode 531/1000, Total Reward: 151.0, Epsilon: 0.010
Episode 541/1000, Total Reward: 139.0, Epsilon: 0.010
Episode 551/1000, Total Reward: 107.0, Epsilon: 0.010
Episode 561/1000, Total Reward: 37.0, Epsilon: 0.010
Episode 571/1000, Total Reward: 13.0, Epsilon: 0.010
Episode 581/1000, Total Reward: 157.0, Epsilon: 0.010
Episode 591/1000, Total Reward: 177.0, Epsilon: 0.010
```

```
Episode 601/1000, Total Reward: 222.0, Epsilon: 0.010
Episode 611/1000, Total Reward: 141.0, Epsilon: 0.010
Episode 621/1000, Total Reward: 155.0, Epsilon: 0.010
Episode 631/1000, Total Reward: 302.0, Epsilon: 0.010
Episode 641/1000, Total Reward: 258.0, Epsilon: 0.010
Episode 651/1000, Total Reward: 284.0, Epsilon: 0.010
Episode 661/1000, Total Reward: 128.0, Epsilon: 0.010
Episode 671/1000, Total Reward: 148.0, Epsilon: 0.010
Episode 681/1000, Total Reward: 191.0, Epsilon: 0.010
Episode 691/1000, Total Reward: 469.0, Epsilon: 0.010
Episode 701/1000, Total Reward: 77.0, Epsilon: 0.010
Episode 711/1000, Total Reward: 65.0, Epsilon: 0.010
Episode 721/1000, Total Reward: 497.0, Epsilon: 0.010
Episode 731/1000, Total Reward: 211.0, Epsilon: 0.010
Episode 741/1000, Total Reward: 500.0, Epsilon: 0.010
Episode 751/1000, Total Reward: 500.0, Epsilon: 0.010
Episode 761/1000, Total Reward: 500.0, Epsilon: 0.010
Episode 771/1000, Total Reward: 114.0, Epsilon: 0.010
Episode 781/1000, Total Reward: 121.0, Epsilon: 0.010
Episode 791/1000, Total Reward: 125.0, Epsilon: 0.010
Episode 801/1000, Total Reward: 137.0, Epsilon: 0.010
Episode 811/1000, Total Reward: 115.0, Epsilon: 0.010
Episode 821/1000, Total Reward: 27.0, Epsilon: 0.010
Episode 831/1000, Total Reward: 121.0, Epsilon: 0.010
Episode 841/1000, Total Reward: 197.0, Epsilon: 0.010
Episode 851/1000, Total Reward: 139.0, Epsilon: 0.010
Episode 861/1000, Total Reward: 195.0, Epsilon: 0.010
Episode 871/1000, Total Reward: 114.0, Epsilon: 0.010
Episode 881/1000, Total Reward: 320.0, Epsilon: 0.010
Episode 891/1000, Total Reward: 60.0, Epsilon: 0.010
Episode 901/1000, Total Reward: 117.0, Epsilon: 0.010
Episode 911/1000, Total Reward: 33.0, Epsilon: 0.010
Episode 921/1000, Total Reward: 500.0, Epsilon: 0.010
Episode 931/1000, Total Reward: 419.0, Epsilon: 0.010
Episode 941/1000, Total Reward: 500.0, Epsilon: 0.010
Episode 951/1000, Total Reward: 500.0, Epsilon: 0.010
Episode 961/1000, Total Reward: 500.0, Epsilon: 0.010
Episode 971/1000, Total Reward: 15.0, Epsilon: 0.010
Episode 981/1000, Total Reward: 500.0, Epsilon: 0.010
Episode 991/1000, Total Reward: 500.0, Epsilon: 0.010
Model saved to dqn_cartpole_v1.pth
Model loaded from dqn_cartpole_v1.pth
Total reward in the test run: 500.0
```

PPO算法

策略梯度

可以考虑直接对策略参数θ进行梯度上升，但是直接按此更新可能，步长过大可能导致策略塌陷

$$\nabla_\theta J(\theta) = \mathbb{E}_{s,a\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(a \mid s)\, Q^{\pi^\theta}(s,a)\right]$$

重要性采样比与无约束目标

为了离线利用旧策略采样的数据，引入重要性采样：

$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}.$$

原始的无约束替代目标：

$$L^{\text{PG}}(\theta) = \mathbb{E}_t \left[r_t(\theta)\, \hat{A}_t \right],$$

其中A_t为估计的优是函数，衡量动作a_t对于基准值多好

裁剪代理目标

PPO的核心是对于$r_t(\theta)$进行裁剪，避免过大更新：

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t \left[\min \left(r_t(\theta)\hat{A}_t,\ \text{clip}\left(r_t(\theta), 1-\epsilon, 1+\epsilon\right) \hat{A}_t \right) \right]$$

- ε裁剪阈值，一般取0.1-0.2
- $clip(x, a, b)$将x裁剪到区间$[a, b]$内：$min(max(x,a),b)

值函数与熵正则化

完整目标为三项之和：

$$L(\theta) = \mathbb{E}_t \left[L_t^{\text{CLIP}}(\theta) - c_1 \left(V_\theta(s_t) - V_t^{\text{target}} \right)^2 + c_2 \mathcal{H}\left[\pi_\theta(\cdot \mid s_t)\right] \right]$$

- $L_t^{\text{CLIP}}(\theta)$为Clip策略梯度目标函数
- $c_1 \left(V_\theta(s_t) - V_t^{\text{target}} \right)^2$值函数回归项，类似于Loss
- $-\sum_a \pi(a|s) \log \pi(a|s)$策略的熵，衡量策略随机性

优势函数估计：GAE

一般会用广义优势估计来平衡偏差于方差

$$\delta_t = r_t + \gamma V_\theta(s_{t+1}) - V_\theta(s_t)$$

$$\hat{A}_t = \sum_{l=0}^{T-t-1} (\gamma\lambda)^l\, \delta_{t+l}$$

结合了TD以及Monte Carlo的方法，当λ接近0时，为TD，当λ接近1时为Monte Carlo

算法流程

初始化策略–价值网络参数 θ
初始化经验缓存 buffer

循环直到训练结束：

1. 与环境交互，收集 BATCH_SIZE 步数据： state, action, reward, done, log_prob_old, value_old
2. 计算最后状态的V_last，用于 GAE 递推

3. 用 GAE 计算优势A_t和目标回报R_t并对A_t做标准化
4. 将所有state/action/log_prob_old/value_old/A_t/R_t转为张量
5. 多轮优化（UPDATE_EPOCHS 轮）：
   对每个 mini-batch：
   A. 计算当前策略
   B. 计算比率
   C. 计算裁剪目标：
   D. 计算价值误差：
   E. 计算熵正则：
   F. 总loss
   G. 反向传播 + 优化器更新
6. 重复步骤 1–5，直到达到最大训练步数或收敛

PPO-Clip示例

环境同样采用前一个的CartPole-v1，下面给出具体代码

In [6]:
```python
import gym
import numpy as np
import torch
import torch.nn as nn
import torch.optim as optim
from torch.distributions import Categorical

# 参数设置
ENV_NAME = "CartPole-v1"
GAMMA = 0.99
LAMBDA = 0.95
CLIP_EPS = 0.2
LR = 3e-4
VF_COEF = 0.5
ENT_COEF = 0.01
BATCH_SIZE = 2048
MINI_BATCH_SIZE = 256
UPDATE_EPOCHS = 10
MAX_TRAIN_STEPS = 200_000
DEVICE = torch.device("cuda" if torch.cuda.is_available() else "cpu")


# 网络定义

class ActorCritic(nn.Module):
    def __init__(self, state_dim, action_dim):
        super().__init__()
        self.shared = nn.Sequential(
            nn.Linear(state_dim, 64), nn.Tanh(),
            nn.Linear(64,64), nn.Tanh()
        )
        self.pi = nn.Linear(64, action_dim)
        self.v = nn.Linear(64,1)

    def forward(self, x):
        x = self.shared(x)
```

```
        return self.pi(x), self.v(x)

    def act(self, s):
        logits, _ = self.forward(s)
        dist = Categorical(logits=logits)
        a = dist.sample()
        # 返回动作索引a，对数概率，以及相应的分布熵
        return a, dist.log_prob(a), dist.entropy()

    def evaluate(self, s, a):
        logits, v = self.forward(s)
        dist = Categorical(logits=logits)
        logp = dist.log_prob(a)
        entropy = dist.entropy()
        return logp, entropy, v.squeeze(-1)


class RolloutBuffer:
    def __init__(self):
        self.states = []
        self.actions = []
        self.rewards = []
        self.dones = []
        self.logprobs = []
        self.values = []

    def add(self, s, a , r, d, logp, v):
        self.states.append(s)
        self.actions.append(a)
        self.rewards.append(r)
        self.dones.append(d)
        self.logprobs.append(logp)
        self.values.append(v)

    def clear(self):
        self.__init__()

    # 优势函数的计算
def gae_advantages(rewards, values, dones, gamma, lam):
    adv = np.zeros_like(rewards, dtype=np.float32)
    gae = 0.0
    for i in reversed(range(len(rewards))):
        mask = 1.0 - dones[i]
        # TD误差
        delta = rewards[i] + gamma * values[i+1] * mask - values[i]
        gae = delta + gamma * lam * mask * gae
        adv[i] = gae
    returns = adv + values[:-1]
    return adv, returns


def train(save_path="ppo_cartpole_v1.pth"):
    env = gym.make(ENV_NAME)
    state_dim = env.observation_space.shape[0]
    action_dim = env.action_space.n

    ac = ActorCritic(state_dim, action_dim).to(DEVICE)
    opt = optim.Adam(ac.parameters(), lr=LR)
    buffer = RolloutBuffer()
```

```
state, _ = env.reset()
total_steps = 0
episode = 0
while total_steps < MAX_TRAIN_STEPS:

    buffer.clear()
    for _ in range(BATCH_SIZE):

        state = np.array(state)
        s_tensor = torch.from_numpy(state).float().to(DEVICE)
        with torch.no_grad():
            a, logp, _ = ac.act(s_tensor)
            v = ac.forward(s_tensor)[1].item()
        act = a.item()
        next_state, reward, terminated, truncated, _ = env.step(act)
        done =terminated or truncated

        buffer.add(state, act, reward, float(done), logp.item(), v)

        state = next_state
        total_steps += 1
        if done:
            episode += 1
            state, _ = env.reset()

    with torch.no_grad():
        state = np.array(state)
        s_tensor = torch.from_numpy(state).float().to(DEVICE)
        last_v = ac.forward(s_tensor)[1].item()


    rewards = np.array(buffer.rewards, dtype=np.float32)
    dones   = np.array(buffer.dones, dtype=np.float32)
    values  = np.array(buffer.values + [last_v], dtype=np.float32)
    adv, rets = gae_advantages(rewards, values, dones, GAMMA, LAMBDA)
    adv = (adv - adv.mean()) / (adv.std() + 1e-8)


    states = torch.tensor(np.array(buffer.states), dtype=torch.float32, devi
    actions = torch.tensor(buffer.actions, dtype=torch.int64, device=DEVICE)
    old_logp = torch.tensor(buffer.logprobs, dtype=torch.float32, device=DEV
    advantages = torch.tensor(adv, dtype=torch.float32, device=DEVICE)
    returns = torch.tensor(rets, dtype=torch.float32, device=DEVICE)


    idx = np.arange(BATCH_SIZE)
    for _ in range(UPDATE_EPOCHS):
        np.random.shuffle(idx)
        for start in range(0,BATCH_SIZE,MINI_BATCH_SIZE):
            end = start + MINI_BATCH_SIZE
            mb_idx = idx[start:end]

            logp, entropy, value = ac.evaluate(states[mb_idx], actions[mb_id
            ratio = torch.exp(logp - old_logp[mb_idx])

            surr1 = ratio * advantages[mb_idx]
            surr2 = torch.clamp(ratio, 1-CLIP_EPS, 1+CLIP_EPS) * advantages[
            policy_loss = -torch.min(surr1, surr2).mean()

            value_loss = (returns[mb_idx] - value).pow(2).mean()
```

```
                entropy_loss = entropy.mean()

                loss = policy_loss + VF_COEF*value_loss - ENT_COEF*entropy_loss

                opt.zero_grad()
                loss.backward()
                opt.step()

        if episode % 10 == 0:
            print(f"Steps:{total_steps} Ep:{episode} Loss:{loss.item():.3f} "
                  f"V:{value_loss.item():.3f} P:{policy_loss.item():.3f}")
    torch.save(ac.state_dict(), save_path)
    print(f"模型已经保存到{save_path}")
    env.close()

def test(model_path="ppo_cartpole.pth", episodes=10, render=False):
    env = gym.make(ENV_NAME, render_mode="human" if render else None)
    ac  = ActorCritic(env.observation_space.shape[0],
                      env.action_space.n).to(DEVICE)

    ac.load_state_dict(torch.load(model_path, map_location=DEVICE))
    ac.eval()

    for ep in range(1, episodes+1):
        state, _ = env.reset()
        done = False
        total_reward = 0.0

        while not done:

            s_tensor = torch.from_numpy(state).float().to(DEVICE)
            with torch.no_grad():
                logits, _ = ac.forward(s_tensor)
                dist = Categorical(logits=logits)
                action = dist.sample().item()

            # 交互
            state, reward, terminated, truncated, _ = env.step(action)
            done = terminated or truncated
            total_reward += reward

            if render:
                env.render()

        print(f"Test Episode {ep}: Reward = {total_reward:.1f}")

    env.close()

if __name__ == "__main__":
    train("ppo_cartpole_v1.pth")

    test("ppo_cartpole_v1.pth", episodes=5, render=True)
```

```
d:\Anaconda\envs\transformer\lib\site-packages\gym\utils\passive_env_checker.py:2
33: DeprecationWarning: `np.bool8` is a deprecated alias for `np.bool_`.  (Deprec
ated NumPy 1.24)
  if not isinstance(terminated, (bool, np.bool8)):
```

```
Steps:6144 Ep:270 Loss:20.457 V:40.953 P:-0.012
Steps:10240 Ep:460 Loss:19.374 V:38.716 P:0.023
Steps:30720 Ep:1180 Loss:33.128 V:66.428 P:-0.080
Steps:57344 Ep:1650 Loss:53.534 V:107.210 P:-0.064
Steps:73728 Ep:1850 Loss:62.941 V:125.948 P:-0.027
Steps:77824 Ep:1900 Loss:71.991 V:143.817 P:0.087
Steps:102400 Ep:2130 Loss:8.177 V:16.445 P:-0.039
Steps:104448 Ep:2160 Loss:54.449 V:108.892 P:0.008
Steps:110592 Ep:2210 Loss:37.766 V:75.218 P:0.163
Steps:126976 Ep:2350 Loss:36.119 V:72.391 P:-0.071
Steps:135168 Ep:2410 Loss:47.573 V:95.080 P:0.039
Steps:182272 Ep:2780 Loss:15.382 V:30.786 P:-0.005
Steps:184320 Ep:2800 Loss:83.269 V:166.566 P:-0.009
Steps:194560 Ep:2870 Loss:61.161 V:122.316 P:0.008
Steps:196608 Ep:2880 Loss:4.267 V:8.638 P:-0.046
模型已经保存到ppo_cartpole_v1.pth
Test Episode 1: Reward = 89.0
Test Episode 2: Reward = 296.0
Test Episode 3: Reward = 154.0
Test Episode 4: Reward = 500.0
Test Episode 5: Reward = 54.0
```

算法总结

Q-learing和DQN是两个类似的算法，其中Q-learing无法解决复杂环境下的问题(state过多导致Q表开销过大)，而DQN解决了这一问题。然后DQN无法很好解决连续高维的动作空间，PPO提供了一个更稳定的策略梯度方案，但是训练成本也相应提升。针对我们初赛的题目，普通的MLP-DQN应该是可以解决的，当前也可也考虑部分优化的DQN算法。同时PPO也当然可以实现，但是鉴于实现的成本与难度，可以作为备选项考虑。