**实验三报告**

**窦雨杉**

**2017202104**

**实验环境：**

Windows anaconda3 python3.7 tensorflow2.0

**实验内容：**

Gym中的CartPole游戏由一个小车及小棍组成，玩家需要左右移动小车，使小车不碰壁且小棍不倾倒。每次观察结果observation是一个四维向量，分别表示小车位置、小车速度、小棒倾角正弦、小棒角度变化率。玩家的决策action是一个整数，0代表向左，1代表向右。

**实验方法：**

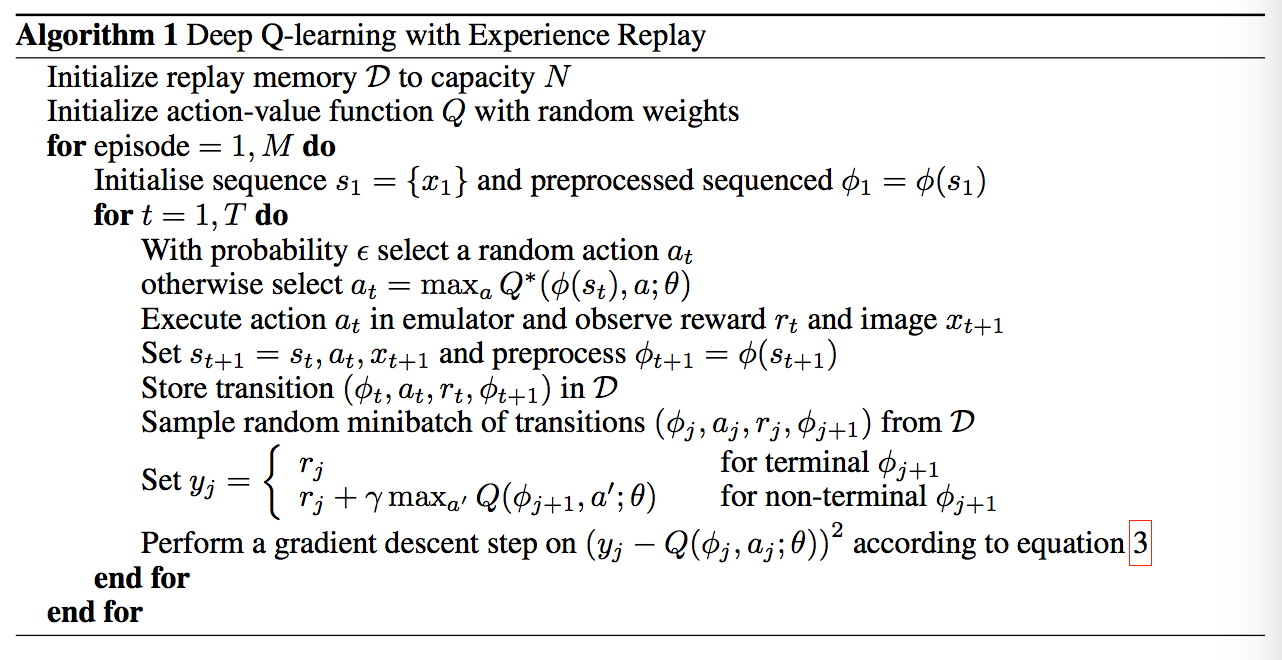
在本实验中，我们采用DQN模型。

强化学习是一个反复迭代的过程，每一次迭代要解决两个问题：给定一个策略求值函数，和根据值函数来更新策略。

DQN使用神经网络来近似值函数，即神经网络的输入是state s,输出是Q(s,a),∀a∈AQ(s,a),∀a∈A (action space)。通过神经网络计算出值函数后，DQN使用ϵ−greedy策略来输出action。值函数网络与ϵ−greedy策略之间的联系是这样的：首先环境会给出一个observation，智能体根据值函数网络得到关于这个obs的所有Q(s,a)Q(s,a)，然后利用ϵ−greedy选择action并做出决策，环境接收到此action后会给出一个奖励Reward及下一个obs。这是一个step。此时我们根据Reward去更新值函数网络的参数。接着进入下一个step。如此循环下去，直到我们训练出了一个好的值函数网络。

在更新神经网络的参数时，与机器学习类似，我们首先定义一个loss function，然后使用梯度下降GD来更新参数。

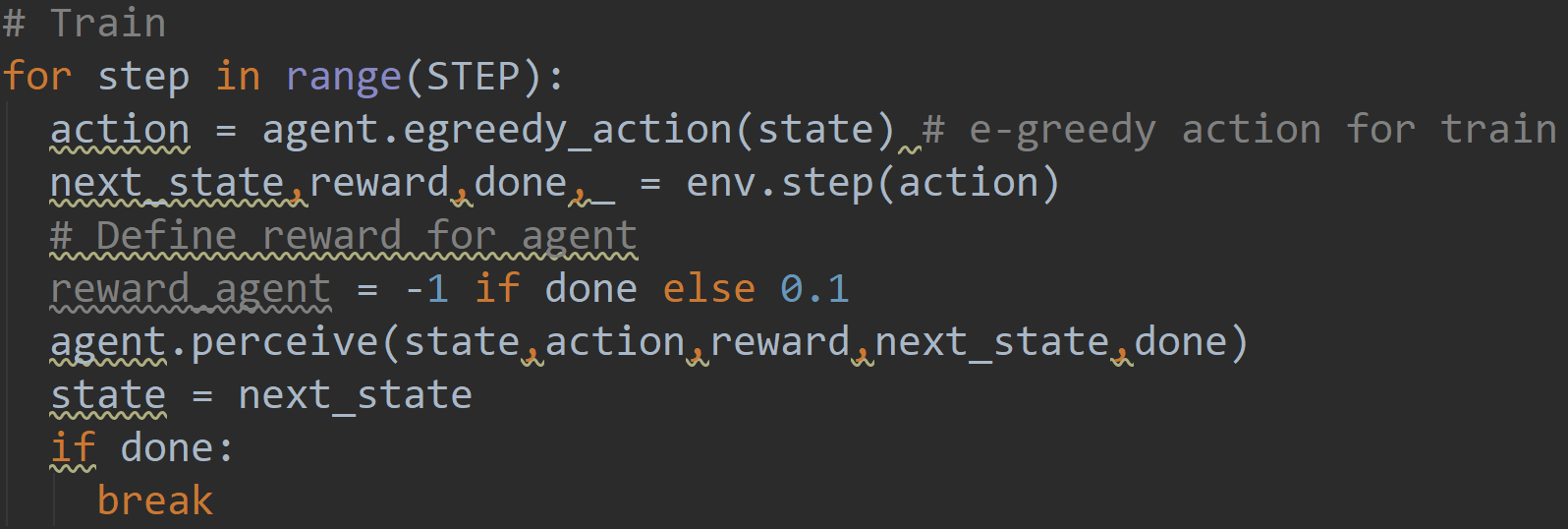
算法流程见下图：



具体实现见源码

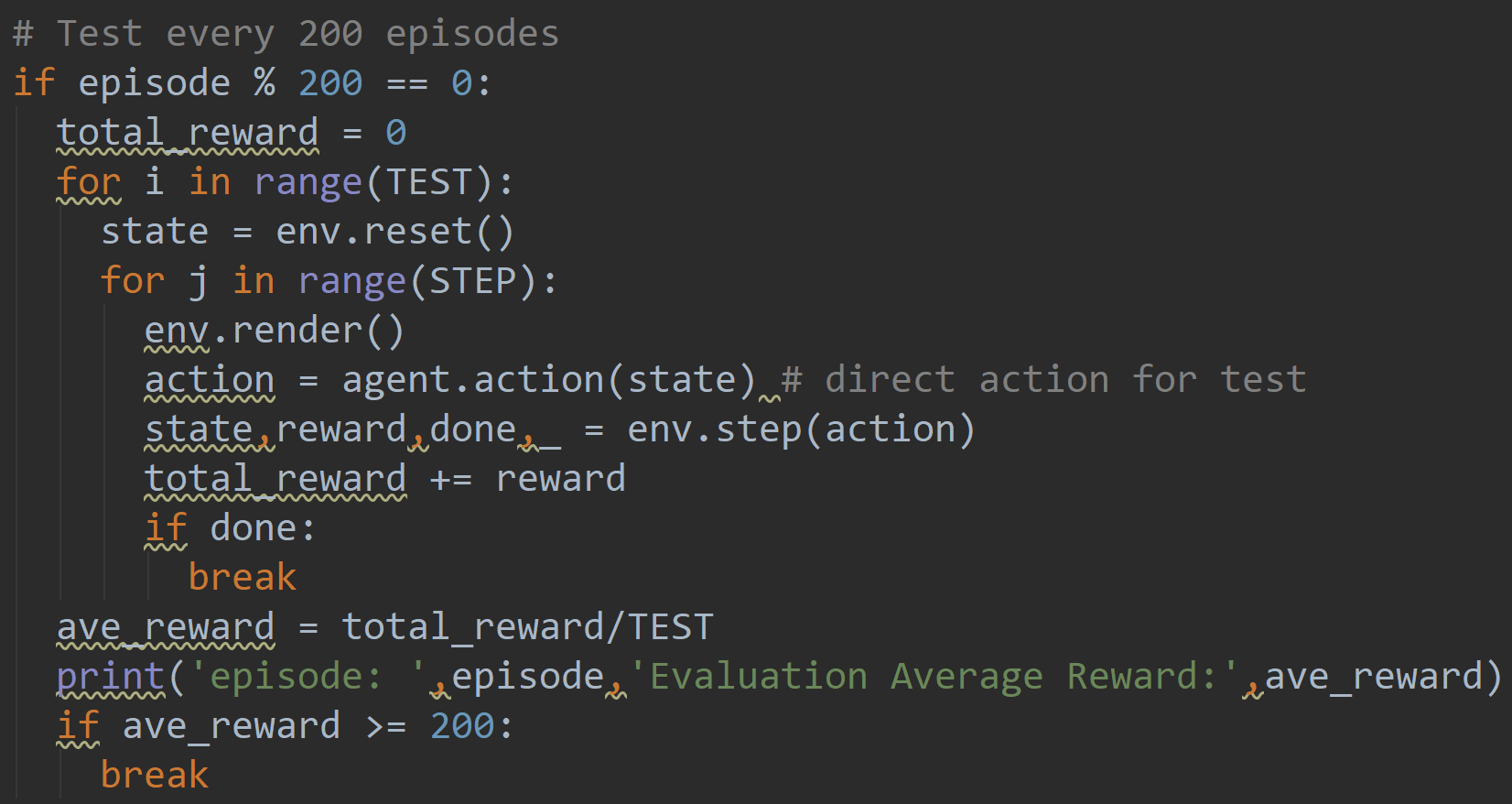
**实验步骤：**

1. 训练模型

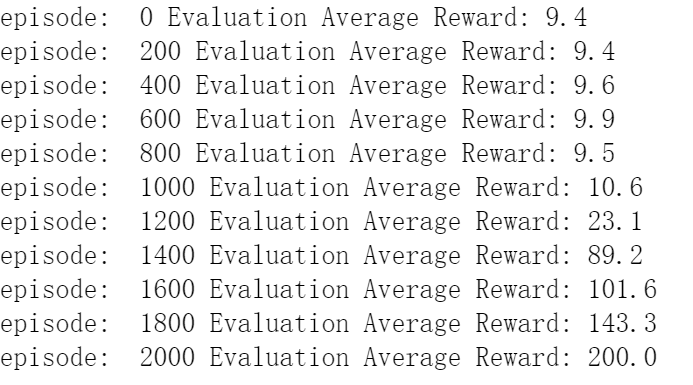
****

1. 测试

模型每迭代200次进行一次测试，输出平均得分



**实验结果：**



（这里pycharm出了一点问题，程序是在notebook上运行的）

模型在约2000次迭代后达到稳定，成功完成这一任务