**算法阶段实验报告**

修订记录

| 日期 | 修订版本 | 修改章节 | 修改描述 | 作者 |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

# 实验目的

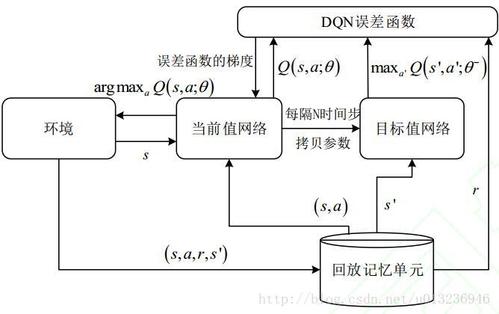
在两个Atari游戏上复现DQN，验证DQN可以使得机器在和环境的交互中学习最优策略，获得最大收益。

# 算法原理和流程

DQN算法是将Q-learning通过神经网络近似值函数的一种方法。Q-learning用于找到一个最优策略，即从当前状态开始，所有连续步骤的总回报的期望值最大。最开始，Q被初始化为一个可能的固定任意值，然后在每个时间t，agent会根据贪婪策略选择一个动作，得到一个奖励，进入一个新的状态, Q值更新，即：



总的过程如下：



回放记忆单元（Experience Replay）：将系统探索环境得到的数据存储起来，随机采样样本更新神经网络的参数。作用是，满足数据独立分布，打破数据之间的关联性。

手机屏幕截图

描述已自动生成

其中，push用于将数据存入，sample用于将数据取出用于训练。

当前网络与目标网络：

模型通过当前时刻的回报和下一个时刻的估值进行更新，因为数据存在不稳定性，每一轮迭代都会产生一些波动，这些波动会反应到下一个迭代的计算中，使得模型不稳定。为解决这个问题，要尽可能的使两个部分解开关联，引入了目标网络。

1. 开始时，两个网络使用完全相同的参数；手机屏幕截图

   描述已自动生成
2. 训练过程中，当前网络负责与环境交互，得到交互样本；

手机屏幕截图

描述已自动生成

手机屏幕截图

描述已自动生成

1. 由Q-learning得到的目标价值由目标网络得到，目标网络与当前网络的估计值比较得出目标值，更新当前网络；

手机屏幕截图

描述已自动生成

1. 完成一定次数的迭代，当前网络的参数会同步给目标网络，就可以进行下一个阶段的学习。

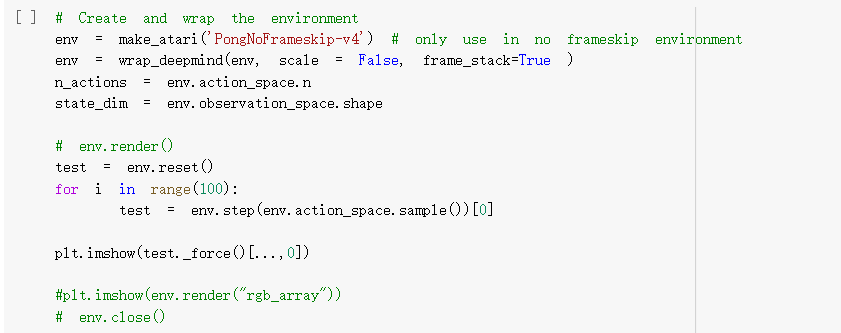
手机屏幕截图

描述已自动生成

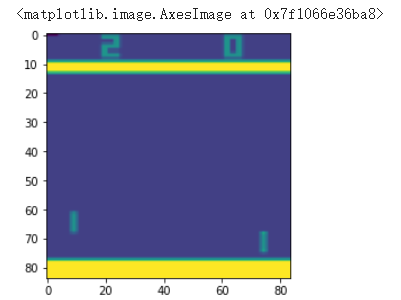
# 实验环境/实验数据

## 环境配置基本情况

安装导入gym库，配置atari游戏的环境；这里游戏选择的是pong，玩家和电脑各拿一个板子，接对方弹来的球，没接住，对方得一分，先得到21分获胜。利用baselines包中的wrap\_deepmind函数，将环境输出的frame转换成84\*84的array，方便训练。



Pong：



Space invaders:

电脑萤幕

描述已自动生成

## 训练/测试数据说明

输入是atari游戏的一帧图像，输出是每个action的价值。

手机屏幕截图

描述已自动生成

## 算法默认参数情况说明

手机屏幕截图

描述已自动生成

# 实验结果

## 游戏选择pong，结果如下：

图片包含 游戏机, 文字

描述已自动生成

图片包含 游戏机

描述已自动生成x轴是训练frame的数量，y轴是近10个episodes的平均rewards。

## 游戏选择space invaders，结果如下：

图片包含 游戏机

描述已自动生成

图片包含 游戏机

描述已自动生成

# 实验结果分析和总结

通过实验结果说明，DQN确实可以从环境的交互中学到一个较优的策略，在游戏中得到一个比较高的分数。