# 实验3. 强化学习实践

MF1733071，严德美， 1312480794@qq.com

2018 年 1 月 3 日

## 综述

本次实验是实现课本中的QLearning算法，后来出现的DQN(deep Q-networks)和DQN的改进版ImprovedDQN（deep Q-learning with experience replay），在实现的基础上进行调参，逐步优化，使得在具体任务上达到最优。各任务修改的reward如下：

```
        # CartPole-v0 reward
        x, x_, theta, theta_ = obv
        r1 = (self.env.x_threshold - abs(x))
         / self.env.x_threshold - 0.8
        r2 = (self.env.theta_threshold_radians - abs(theta))
         / self.env.theta_threshold_radians - 0.5
        reward = r1 + r2
        if x > 4 or x < -4:
            reward = reward - 0.05

        # MountainCar-v0 reward
        s_, r, done, info = self.mydqn.env.step(action)
        position, velocity = s_
        r = np.abs(position-(-0.5))

        # Acrobot-v1 reward
        s_, r, done, info = self.mydqn.env.step(action)
        x1, _, x2, _, _, _ = s_
        r = 1 - x1 + x2
        if done and t < 500 :
            if t < 200:
                r += 1000
            if t < 100:
                r += 10000
            r += 500
```

# 实验二.

## 离散化方法

本次实验使用的离散化的方法主要是将连续的区间投影到一个指定区间离散的整数点，具体叙述如下：假设obv为原状态空间某个变量的取值，obv_min为原状态空间对应变量的最小值，origin_length为原状态空间中对应变量的区间长度（max-min），target_length为所要投影的离散区间的区间长度，那么将原连续区间中的状态变量obv投影到离散区间中的状态变量state：

$$state = Round\left(\frac{(obv - obv\_min)target\_length}{origin\_length}\right) \tag{1}$$

### Q-learning 算法实现思路

从随机的状态开始，起初处于探索阶段，使用$\epsilon - greedy$策略开始探索环境，以$\epsilon$概率随机选择动作，以1-$\epsilon$概率选择当前状态Q值最大的action,环境根据action返回reward，并进入到下一个状态，并计算下一个状态最大的Q值，最后根据下列公式更新当前的Q值

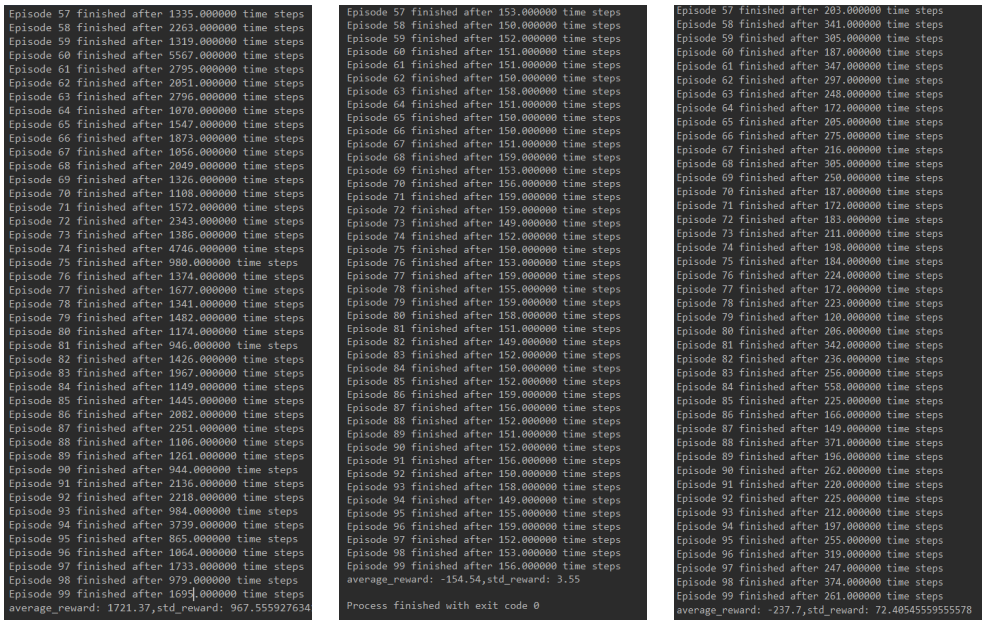
$$Q(s,a) = Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s^{'}, a^{'}) - Q(s,a)] \tag{2}$$

实验二在每个任务上测试了100条轨迹，对于每个任务计算出对应的reward之和的均值和标准差

对于CartPole-v0任务最优参数为 $\epsilon = 0.1, \alpha = 0.1, \gamma = 0.90, episodes = 2000$ ,一条轨迹的长度设置为20000，测试的reward之和的均值和标准差如图1（a）所示,均值为7612.63，标准差为6043.95

对于MountainCar-v0任务最优参数为 $\epsilon = 0.1, \alpha = 0.1, \gamma = 0.99, episodes = 2000$ 测试的reward之和的均值和标准差如图1（b）所示,均值为-154.54，标准差为3.55

对于Acrobot-v1任务最优参数为 $\epsilon = 0.1, \alpha = 0.1, \gamma = 0.95, episodes = 2000$ 测试的reward之和的均值和标准差如图1（c）所示,均值为-237.7，标准差为72.41




(a) Q-Learning CartPole-v0 (b) Q-Learning MountainCar-v0 (c) Q-Learning Acrobot-v1

图 1: Q-Learning test

# 实验三.

## DQN实现

Q-Learning只能适用于状态和动作空间是离散的情况，因为其中的Q值是用表格的形式存储的，连续的值用表格无法穷尽的枚举，然而使用值函数近似能解决这个问题，将值函数近似模型选取为一个深度神经网络，本次实验我选取的是三层神经网络，输入状态空间的值，输出每个动作对应的Q值，从环境中采样状态和动作存储到memory中，然后就可以使用memory中的数据当做训练样本，给神经网络进行训练，训练好神经网络后，对于测试中的环境状态，根据神经网络输出的Q值选择Q值最大的对应的action，实验三在每个任务上测试了100条轨迹，对于每个任务计算出对应的reward之和的均值和标准差，并用数据图记录下网络训练误差随训练轮数的变化关系以及每轮训练reward之和随训练轮数的变化关系。

## CartPole-v0 任务

对于CartPole-v0任务最优参数为$\epsilon = 0.1, learning_rate = 0.0035, \gamma = 0.90, episodes = 500$ ,一条轨迹的长度设置为20000，测试的reward之和的均值和标准差如图2(a),(b)所示，相应的训练的loss均值和reward之和随episode变化关系如图3(a),(b),(c),(d)所示

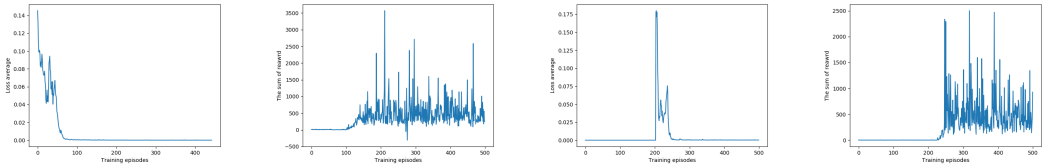
```
Episode 38 finished after 2815 time steps
Episode 39 finished after 1917 time steps
Episode 40 finished after 2503 time steps
Episode 41 finished after 2939 time steps
Episode 42 finished after 2157 time steps
Episode 43 finished after 2497 time steps
Episode 44 finished after 3253 time steps
Episode 45 finished after 2871 time steps
Episode 46 finished after 3047 time steps
Episode 47 finished after 1881 time steps
Episode 48 finished after 2005 time steps
Episode 49 finished after 2163 time steps
Episode 50 finished after 2795 time steps
Episode 51 finished after 3912 time steps
Episode 52 finished after 2350 time steps
Episode 53 finished after 1823 time steps
Episode 54 finished after 2249 time steps
Episode 55 finished after 2979 time steps
Episode 56 finished after 2711 time steps
Episode 57 finished after 2429 time steps
Episode 58 finished after 3402 time steps
Episode 59 finished after 2671 time steps
Episode 60 finished after 2801 time steps
Episode 61 finished after 2645 time steps
Episode 62 finished after 2845 time steps
Episode 63 finished after 2715 time steps
Episode 64 finished after 2047 time steps
Episode 65 finished after 4808 time steps
Episode 66 finished after 2213 time steps
Episode 67 finished after 2319 time steps
Episode 68 finished after 5944 time steps
Episode 69 finished after 2148 time steps
Episode 70 finished after 3019 time steps
Episode 71 finished after 3019 time steps
Episode 72 finished after 2265 time steps
Episode 73 finished after 3125 time steps
Episode 74 finished after 2631 time steps
Episode 75 finished after 2693 time steps
Episode 76 finished after 2297 time steps
Episode 77 finished after 1813 time steps
Episode 78 finished after 2195 time steps
Episode 79 finished after 1831 time steps
Episode 80 finished after 2221 time steps
Episode 81 finished after 1967 time steps
Episode 82 finished after 2355 time steps
Episode 83 finished after 2559 time steps
Episode 84 finished after 2035 time steps
Episode 85 finished after 3594 time steps
Episode 86 finished after 1904 time steps
Episode 87 finished after 2539 time steps
Episode 88 finished after 2665 time steps
Episode 89 finished after 2597 time steps
Episode 90 finished after 2333 time steps
Episode 91 finished after 2845 time steps
Episode 92 finished after 1852 time steps
Episode 93 finished after 2313 time steps
Episode 94 finished after 2129 time steps
Episode 95 finished after 2735 time steps
Episode 96 finished after 3979 time steps
Episode 97 finished after 2317 time steps
Episode 98 finished after 2577 time steps
Episode 99 finished after 2790 time steps
average_reward: 2646.11,std_reward: 633.697
```

(a) CartPole-v0 test-1

```
Episode 38 finished after 3737 time steps
Episode 39 finished after 4025 time steps
Episode 40 finished after 3521 time steps
Episode 41 finished after 3251 time steps
Episode 42 finished after 3575 time steps
Episode 43 finished after 3143 time steps
Episode 44 finished after 3529 time steps
Episode 45 finished after 3379 time steps
Episode 46 finished after 3176 time steps
Episode 47 finished after 3497 time steps
Episode 48 finished after 1236 time steps
Episode 49 finished after 1273 time steps
Episode 50 finished after 3343 time steps
Episode 51 finished after 1172 time steps
Episode 52 finished after 3669 time steps
Episode 53 finished after 3489 time steps
Episode 54 finished after 1268 time steps
Episode 55 finished after 3833 time steps
Episode 56 finished after 3615 time steps
Episode 57 finished after 3797 time steps
Episode 58 finished after 3910 time steps
Episode 59 finished after 917 time steps
Episode 60 finished after 948 time steps
Episode 61 finished after 3105 time steps
Episode 62 finished after 1090 time steps
Episode 63 finished after 3664 time steps
Episode 64 finished after 1468 time steps
Episode 65 finished after 3309 time steps
Episode 66 finished after 2833 time steps
Episode 67 finished after 3861 time steps
Episode 68 finished after 3083 time steps
Episode 69 finished after 3937 time steps
Episode 70 finished after 1965 time steps
Episode 71 finished after 1494 time steps
Episode 72 finished after 1180 time steps
Episode 73 finished after 2890 time steps
Episode 74 finished after 1361 time steps
Episode 75 finished after 3670 time steps
Episode 76 finished after 3381 time steps
Episode 77 finished after 3569 time steps
Episode 78 finished after 3305 time steps
Episode 79 finished after 3591 time steps
Episode 80 finished after 3291 time steps
Episode 81 finished after 3599 time steps
Episode 82 finished after 3400 time steps
Episode 83 finished after 1078 time steps
Episode 84 finished after 1400 time steps
Episode 85 finished after 3259 time steps
Episode 86 finished after 3487 time steps
Episode 87 finished after 4109 time steps
Episode 88 finished after 3245 time steps
Episode 89 finished after 4069 time steps
Episode 90 finished after 3463 time steps
Episode 91 finished after 3299 time steps
Episode 92 finished after 3527 time steps
Episode 93 finished after 3224 time steps
Episode 94 finished after 4129 time steps
Episode 95 finished after 4059 time steps
Episode 96 finished after 4069 time steps
Episode 97 finished after 3121 time steps
Episode 98 finished after 3643 time steps
Episode 99 finished after 827 time steps
average_reward: 2996.46,std_reward: 994.978
```

(b) CartPole-v0 test-2

图 2: DQN CartPole-v0 test



(a) test-1 loss (b) test-1 reward (c) test-2 loss (d) test-2 reward

图 3: DQN CartPole-v0 test

## MountainCar-v0 任务

对于MountainCar-v0任务最优参数为 $batch\_size = 32, \epsilon = 0.1, epsilon\_decay = 0.995, learning\_rate = 0.0025, \gamma = 0.9, episodes = 400$ 对MountainCar-v0任务进行了三次测试，每次进行100轮，reward之和的均值与标准差如图4(a),(b),(c)所示

```
Episode 57 finished after 107 time steps
Episode 58 finished after 89 time steps
Episode 59 finished after 147 time steps
Episode 60 finished after 94 time steps
Episode 61 finished after 138 time steps
Episode 62 finished after 149 time steps
Episode 63 finished after 139 time steps
Episode 64 finished after 87 time steps
Episode 65 finished after 107 time steps
Episode 66 finished after 148 time steps
Episode 67 finished after 107 time steps
Episode 68 finished after 108 time steps
Episode 69 finished after 148 time steps
Episode 70 finished after 89 time steps
Episode 71 finished after 92 time steps
Episode 72 finished after 89 time steps
Episode 73 finished after 141 time steps
Episode 74 finished after 138 time steps
Episode 75 finished after 91 time steps
Episode 76 finished after 94 time steps
Episode 77 finished after 154 time steps
Episode 78 finished after 107 time steps
Episode 79 finished after 89 time steps
Episode 80 finished after 138 time steps
Episode 81 finished after 95 time steps
Episode 82 finished after 140 time steps
Episode 83 finished after 138 time steps
Episode 84 finished after 107 time steps
Episode 85 finished after 107 time steps
Episode 86 finished after 107 time steps
Episode 87 finished after 107 time steps
Episode 88 finished after 141 time steps
Episode 89 finished after 90 time steps
Episode 90 finished after 148 time steps
Episode 91 finished after 89 time steps
Episode 92 finished after 143 time steps
Episode 93 finished after 144 time steps
Episode 94 finished after 140 time steps
Episode 95 finished after 138 time steps
Episode 96 finished after 88 time steps
Episode 97 finished after 139 time steps
Episode 98 finished after 96 time steps
Episode 99 finished after 149 time steps
average_reward: -125.22,std_reward: 23.79436
```

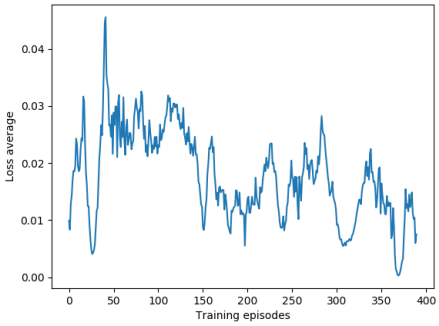
```
Episode 57 finished after 136 time steps
Episode 58 finished after 149 time steps
Episode 59 finished after 138 time steps
Episode 60 finished after 144 time steps
Episode 61 finished after 116 time steps
Episode 62 finished after 158 time steps
Episode 63 finished after 155 time steps
Episode 64 finished after 139 time steps
Episode 65 finished after 152 time steps
Episode 66 finished after 144 time steps
Episode 67 finished after 116 time steps
Episode 68 finished after 116 time steps
Episode 69 finished after 152 time steps
Episode 70 finished after 146 time steps
Episode 71 finished after 153 time steps
Episode 72 finished after 139 time steps
Episode 73 finished after 139 time steps
Episode 74 finished after 154 time steps
Episode 75 finished after 145 time steps
Episode 76 finished after 115 time steps
Episode 77 finished after 152 time steps
Episode 78 finished after 142 time steps
Episode 79 finished after 149 time steps
Episode 80 finished after 155 time steps
Episode 81 finished after 115 time steps
Episode 82 finished after 139 time steps
Episode 83 finished after 142 time steps
Episode 84 finished after 144 time steps
Episode 85 finished after 115 time steps
Episode 86 finished after 116 time steps
Episode 87 finished after 181 time steps
Episode 88 finished after 151 time steps
Episode 89 finished after 152 time steps
Episode 90 finished after 177 time steps
Episode 91 finished after 116 time steps
Episode 92 finished after 154 time steps
Episode 93 finished after 116 time steps
Episode 94 finished after 104 time steps
Episode 95 finished after 141 time steps
Episode 96 finished after 139 time steps
Episode 97 finished after 154 time steps
Episode 98 finished after 138 time steps
Episode 99 finished after 155 time steps
average_reward: -139.79,std_reward: 18.098229195145034
```

```
Episode 57 finished after 164 time steps
Episode 58 finished after 165 time steps
Episode 59 finished after 166 time steps
Episode 60 finished after 153 time steps
Episode 61 finished after 161 time steps
Episode 62 finished after 166 time steps
Episode 63 finished after 164 time steps
Episode 64 finished after 157 time steps
Episode 65 finished after 160 time steps
Episode 66 finished after 159 time steps
Episode 67 finished after 155 time steps
Episode 68 finished after 164 time steps
Episode 69 finished after 158 time steps
Episode 70 finished after 156 time steps
Episode 71 finished after 156 time steps
Episode 72 finished after 159 time steps
Episode 73 finished after 159 time steps
Episode 74 finished after 160 time steps
Episode 75 finished after 161 time steps
Episode 76 finished after 159 time steps
Episode 77 finished after 166 time steps
Episode 78 finished after 161 time steps
Episode 79 finished after 164 time steps
Episode 80 finished after 156 time steps
Episode 81 finished after 155 time steps
Episode 82 finished after 155 time steps
Episode 83 finished after 155 time steps
Episode 84 finished after 153 time steps
Episode 85 finished after 166 time steps
Episode 86 finished after 160 time steps
Episode 87 finished after 157 time steps
Episode 88 finished after 157 time steps
Episode 89 finished after 163 time steps
Episode 90 finished after 157 time steps
Episode 91 finished after 158 time steps
Episode 92 finished after 157 time steps
Episode 93 finished after 156 time steps
Episode 94 finished after 166 time steps
Episode 95 finished after 157 time steps
Episode 96 finished after 159 time steps
Episode 97 finished after 159 time steps
Episode 98 finished after 158 time steps
Episode 99 finished after 155 time steps
average_reward: -159.44,std_reward: 3.4909
```

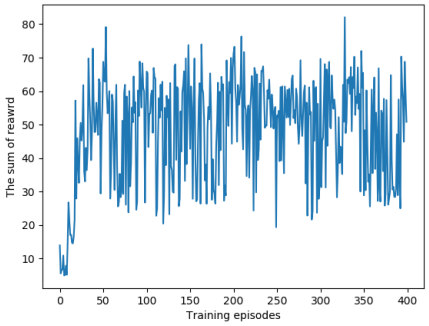
(a) DQN MountainCar-v0 test-1 (b) DQN MountainCar-v0 test-2 (c) DQN MountainCar-v0 test-3

图 4: DQN MountainCar-v0 test

网络训练误差随训练轮数的变化关系如图5(a)所示，每轮训练reward 之和随训练轮数的变化关系如图5(b)所示




(a) DQN MountainCar-v0 loss

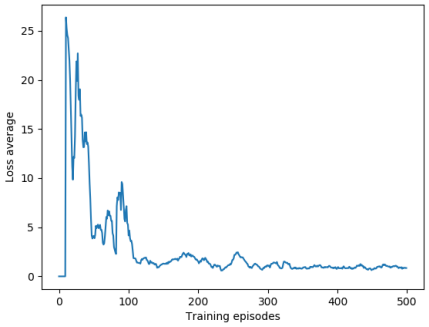



(b) DQN MountainCar-v0 reward

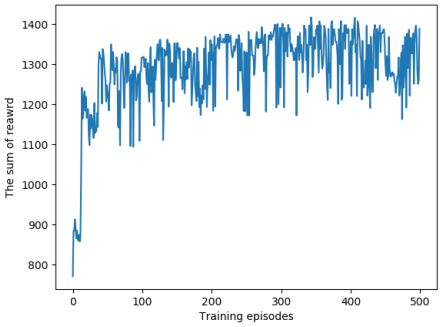
图 5: DQN

## Acrobot-v1 任务

对于Acrobot-v1任务最优参数为 $\epsilon = 0.1, learning\_rate = 0.0035, \gamma = 0.90, epsilon\_decay = 0.995, episodes = 500$ ,一条轨迹的长度设置为2000，测试的reward之和的均值和标准差如图7（a）所示,均值为-78.65，标准差为11.1,相应的训练的loss均值和reward之和随episode变化关系如图6(a),(b)所示




(a) Acrobot-v1 loss




(b) Acrobot-v1 reward

图 6: DQN Acrobot-v1


(a) DQN Acrobot-v1 (b) InprovedDQN CartPole-v0

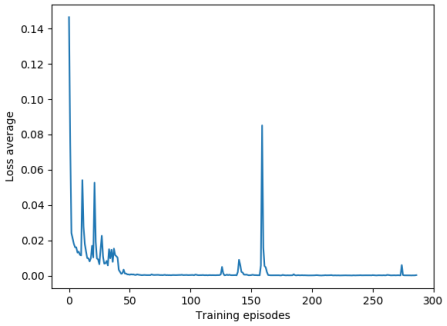
图 7: test

# 实验四.

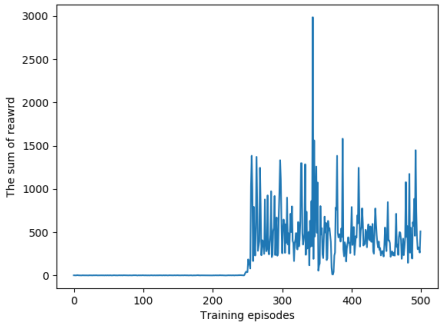
在DQN的基础上增加一个网络t_net,把Q值函数的学习任务交给当前网络e_net,并不是每次都去更新t_net,而是每C步将当前网络e_net的参数复制给t_net网络，延迟更新，减少了目标计算与当前值的相关性，从而增加网络的稳定性。

## CartPole-v0 任务

对于CartPole-v0任务最优参数为 $batch\_size = 128, \epsilon = 0.05, learning\_rate = 0.0025, \gamma = 0.99, epsilon\_decay = 0.95, episodes = 500$ ,一条轨迹的长度设置为20000，测试的reward之和的均值和标准差如图7（b）所示,均值为2407.49，标准差为504.13,相应的训练的loss均值和reward之和随episode变化关系如图8(a),(b)所示




(a) CartPole-v0 loss




(b) CartPole-v0 reward

图 8: InprovedDQN CartPole-v0

## MountainCar-v0 任务

对于MountainCar-v0任务最优参数为 $\epsilon = 0.1, learning\_rate = 0.01, \gamma = 0.9, episodes = 400, batch\_size = 32$ 进行了两次测试，每次一百轮，两次的测试的reward之和的均值和标准差分别为：-150.62 37.49,-194.99 11.49，两次的测试结果和图像如下：

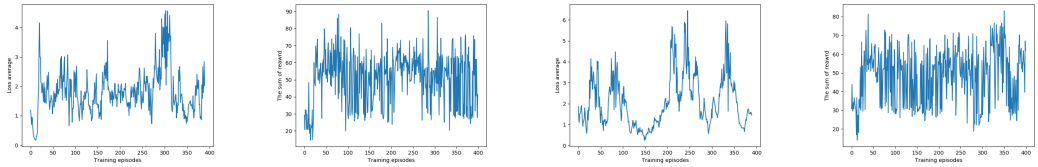
```
Episode 69 finished after 180.000000 time steps
Episode 70 finished after 184.000000 time steps
Episode 71 finished after 167.000000 time steps
Episode 72 finished after 191.000000 time steps
Episode 73 finished after 169.000000 time steps
Episode 74 finished after 127.000000 time steps
Episode 75 finished after 168.000000 time steps
Episode 76 finished after 168.000000 time steps
Episode 77 finished after 168.000000 time steps
Episode 78 finished after 168.000000 time steps
Episode 79 finished after 100.000000 time steps
Episode 80 finished after 168.000000 time steps
Episode 81 finished after 169.000000 time steps
Episode 82 finished after 86.000000 time steps
Episode 83 finished after 87.000000 time steps
Episode 84 finished after 101.000000 time steps
Episode 85 finished after 89.000000 time steps
Episode 86 finished after 88.000000 time steps
Episode 87 finished after 169.000000 time steps
Episode 88 finished after 167.000000 time steps
Episode 89 finished after 184.000000 time steps
Episode 90 finished after 94.000000 time steps
Episode 91 finished after 169.000000 time steps
Episode 92 finished after 88.000000 time steps
Episode 93 finished after 90.000000 time steps
Episode 94 finished after 170.000000 time steps
Episode 95 finished after 91.000000 time steps
Episode 96 finished after 168.000000 time steps
Episode 97 finished after 191.000000 time steps
Episode 98 finished after 173.000000 time steps
Episode 99 finished after 189.000000 time steps
average_reward: -150.62,std_reward: 37.493940843
```

(a) InprovedDQN MountainCar-v0 test-1

```
Episode 69 finished after 178.000000 time steps
Episode 70 finished after 165.000000 time steps
Episode 71 finished after 131.000000 time steps
Episode 72 finished after 130.000000 time steps
Episode 73 finished after 164.000000 time steps
Episode 74 finished after 165.000000 time steps
Episode 75 finished after 165.000000 time steps
Episode 76 finished after 165.000000 time steps
Episode 77 finished after 165.000000 time steps
Episode 78 finished after 163.000000 time steps
Episode 79 finished after 130.000000 time steps
Episode 80 finished after 169.000000 time steps
Episode 81 finished after 165.000000 time steps
Episode 82 finished after 173.000000 time steps
Episode 83 finished after 132.000000 time steps
Episode 84 finished after 168.000000 time steps
Episode 85 finished after 164.000000 time steps
Episode 86 finished after 131.000000 time steps
Episode 87 finished after 165.000000 time steps
Episode 88 finished after 175.000000 time steps
Episode 89 finished after 165.000000 time steps
Episode 90 finished after 170.000000 time steps
Episode 91 finished after 128.000000 time steps
Episode 92 finished after 124.000000 time steps
Episode 93 finished after 175.000000 time steps
Episode 94 finished after 130.000000 time steps
Episode 95 finished after 179.000000 time steps
Episode 96 finished after 127.000000 time steps
Episode 97 finished after 171.000000 time steps
Episode 98 finished after 170.000000 time steps
Episode 99 finished after 169.000000 time steps
average_reward: -154.74,std_reward: 18.653482248
```

(b) InprovedDQN MountainCar-v0 test-2

图 9: InprovedDQN MountainCar-v0



(a) test-1 loss (b) test-1 reward (c) test-2 loss (d) test-2 reward

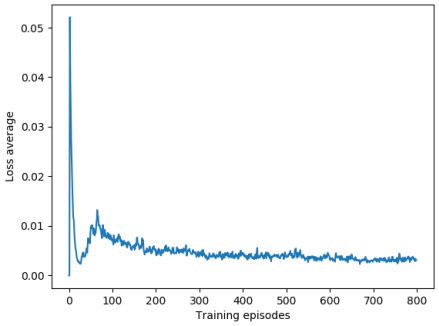
图 10: InprovedDQN MountainCar-v0

## Acrobot-v1 任务

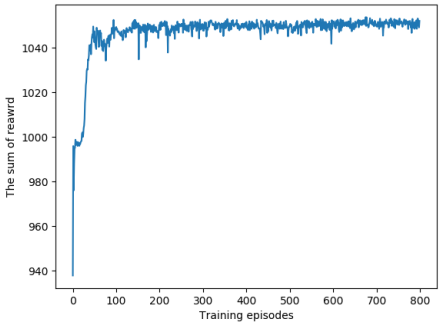
对于Acrobot-v1任务最优参数为 $\epsilon = 0.1, learning\_rate = 0.001, \gamma = 0.90, epsilon\_decay = 0.995, episodes = 800$ ,一条轨迹的长度设置为2000，测试的reward之和的均值和标准差如图11所示,均值为-83.44，标准差为19.68,相应的训练的loss均值和reward之和随episode变化关系如图12(a),(b)所示

```
Episode 55 finished after 90.000000 time steps
Episode 56 finished after 101.000000 time steps
Episode 57 finished after 79.000000 time steps
Episode 58 finished after 81.000000 time steps
Episode 59 finished after 94.000000 time steps
Episode 60 finished after 73.000000 time steps
Episode 61 finished after 210.000000 time steps
Episode 62 finished after 63.000000 time steps
Episode 63 finished after 92.000000 time steps
Episode 64 finished after 90.000000 time steps
Episode 65 finished after 72.000000 time steps
Episode 66 finished after 69.000000 time steps
Episode 67 finished after 71.000000 time steps
Episode 68 finished after 88.000000 time steps
Episode 69 finished after 70.000000 time steps
Episode 70 finished after 79.000000 time steps
Episode 71 finished after 96.000000 time steps
Episode 72 finished after 89.000000 time steps
Episode 73 finished after 93.000000 time steps
Episode 74 finished after 62.000000 time steps
Episode 75 finished after 101.000000 time steps
Episode 76 finished after 91.000000 time steps
Episode 77 finished after 91.000000 time steps
Episode 78 finished after 86.000000 time steps
Episode 79 finished after 80.000000 time steps
Episode 80 finished after 73.000000 time steps
Episode 81 finished after 101.000000 time steps
Episode 82 finished after 92.000000 time steps
Episode 83 finished after 151.000000 time steps
Episode 84 finished after 69.000000 time steps
Episode 85 finished after 73.000000 time steps
Episode 86 finished after 70.000000 time steps
Episode 87 finished after 70.000000 time steps
Episode 88 finished after 73.000000 time steps
Episode 89 finished after 86.000000 time steps
Episode 90 finished after 69.000000 time steps
Episode 91 finished after 69.000000 time steps
Episode 92 finished after 73.000000 time steps
Episode 93 finished after 83.000000 time steps
Episode 94 finished after 86.000000 time steps
Episode 95 finished after 62.000000 time steps
Episode 96 finished after 63.000000 time steps
Episode 97 finished after 69.000000 time steps
Episode 98 finished after 93.000000 time steps
Episode 99 finished after 93.000000 time steps
average_reward: -83.44,std_reward: 19.6821340306
```

图 11: InprovedDQN Acrobot-v1




(a) Acrobot-v1 loss




(b) Acrobot-v1 reward

图 12: InprovedDQN Acrobot-v1