实验 3. 强化学习实践

MG1733098，周华平，zhp@smail.nju.edu.cn

2017 年 12 月 30 日

综述

实验二.

实验三.

Deep Q-network(DQN) 实现

在本实验中，我选择使用 PyTorch 来实现 DQN。在定义 Q 值网络时我使用了 MLP，其中网络结构由 3 层 Linear 构成；激活函数使用 PReLU，同时在每个隐层中增加 Batch Norm 来对相应的 activation 做规范化操作。另外，在 DQN 中我采用optim.Adam作为优化函数，用nn.MSELoss()来计算均方误差。

Q 值网络的定义如下所示：

```
class DQN(nn.Module):
    def __init__(self, input_dim, output_dim, hidden_dim):
        super(DQN, self).__init__()
        self.layer1 = nn.Sequential(
            nn.Linear(input_dim, hidden_dim),
            nn.BatchNorm1d(hidden_dim),
            nn.PReLU(),
        )
        self.layer2 = nn.Sequential(
            nn.Linear(hidden_dim, hidden_dim),
            nn.BatchNorm1d(hidden_dim),
            nn.PReLU(),
        )
        self.out = nn.Linear(hidden_dim, output_dim)

    def forward(self, x):
        x = self.layer1(x)
        x = self.layer2(x)
```

```
        return self.out(x)
```

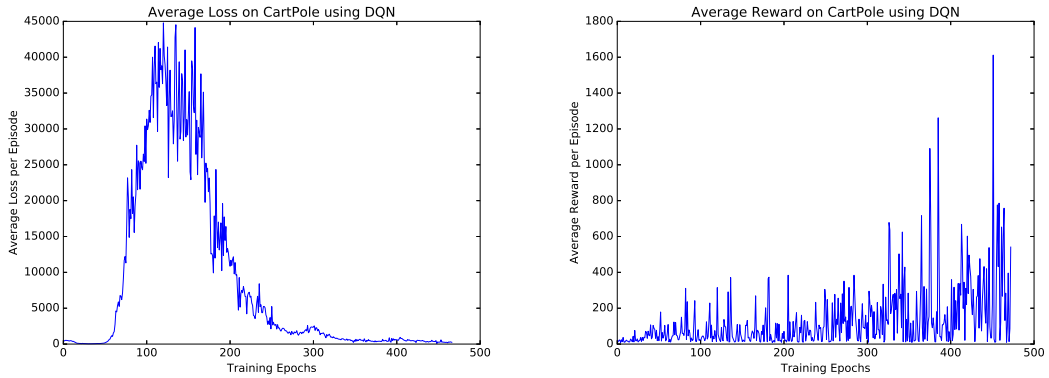
CartPole

针对 CartPole，DQN 的超参数设置如表 1所示。其中 ϵ 的衰减公式同公式 (abcd)。

超参数	参数意义	参数值
memory_size	Replay Memory 的大小	10000
batch_size	mini-batch 的大小	128
hidden_dim	DQN 的隐层维度	50
discount	DQN 算法中的 γ	0.99
learning_rate	DQN 算法中的 α	0.001
eps_start	ϵ 的初始值	0.9
eps_end	ϵ 的结束值	0.05
eps_decay	ϵ 的衰减权重	200

表 1: DQN 超参数设置 (CartPole)

DQN 在 CartPole 上的实验结果如图 1所示。可以观察到 Loss 在超过 450 轮后达到收敛的状态。由于 ϵ 的最小值被设置为 0.05，因此即使 Training 了较多轮数，DQN 依旧会以 5% 的概率随机选择 Action。而 CartPole 似乎对于错误的 Action 比较敏感，当随机到错误的 Action 时，可能会导致该轨迹提前结束。因此在 Training 阶段 Reward 似乎并没有收敛到一个固定值，然而我们可以观察到随着 Training 轮数的增加，Reward 的上限也在不断提高，这也从侧面体现出了训练是有效果的。



(a) Average Loss　　(b) Average Reward

图 1: Training Result of CartPole using DQN

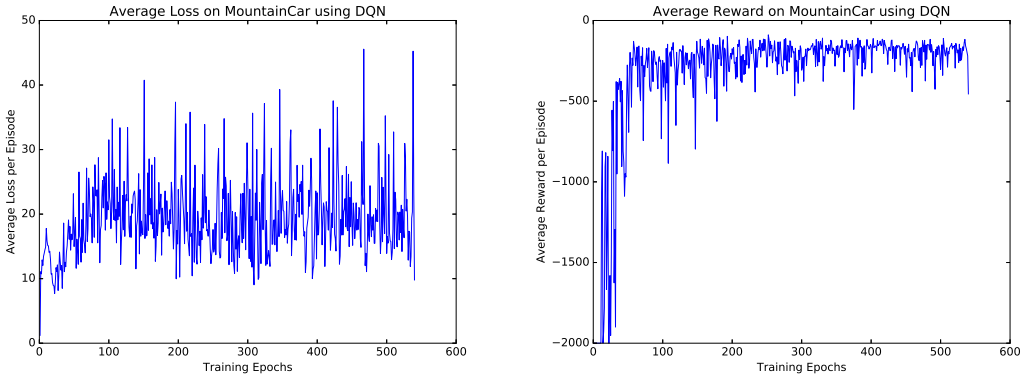
MountainCar

针对 MountainCar，DQN 的超参数设置如表 2所示。

超参数	参数意义	参数值
memory_size	Replay Memory 的大小	10000
batch_size	mini-batch 的大小	128
hidden_dim	DQN 的隐层维度	50
discount	DQN 算法中的 γ	0.99
learning_rate	DQN 算法中的 α	0.001
eps_start	ϵ 的初始值	0.9
eps_end	ϵ 的结束值	0.05
eps_decay	ϵ 的衰减权重	50

表 2: DQN 超参数设置 (MountainCar)

DQN 在 MountainCar 上的实验结果如图 2所示。其中 Reward 在超过 200 轮之后达到收敛的状态，而 Loss 也在超过 200 轮之后达到了基本稳定的状态。



(a) Average Loss

(b) Average Reward

图 2: Training Result of MountainCar using DQN

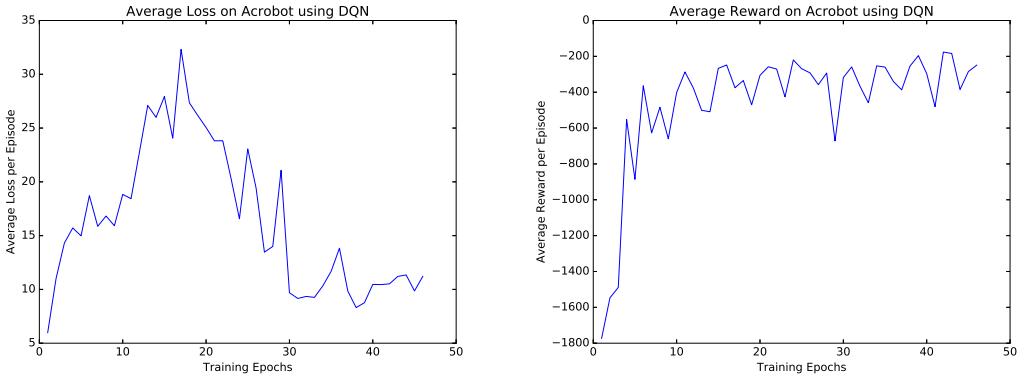
Acrobot

针对 Acrobot，DQN 的超参数设置如表 3所示。

超参数	参数意义	参数值
memory_size	Replay Memory 的大小	5000
batch_size	mini-batch 的大小	128
hidden_dim	DQN 的隐层维度	50
discount	DQN 算法中的 γ	0.99
learning_rate	DQN 算法中的 α	0.001
eps_start	ϵ 的初始值	0.9
eps_end	ϵ 的结束值	0.05
eps_decay	ϵ 的衰减权重	200

表 3: DQN 超参数设置 (Acrobot)

DQN 在 Acrobot 上的实验结果如图 3所示。在超过 40 轮之后，Loss 达到了较低的水平，并且 Reward 也趋近于收敛。



(a) Average Loss

(b) Average Reward

图 3: Training Result of Acrobot using DQN

实验四.

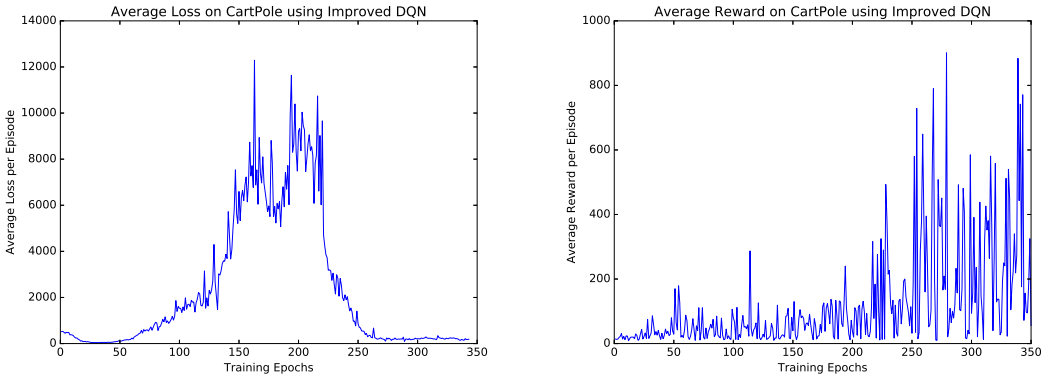
CartPole

针对 CartPole，Improved DQN 的超参数设置如表 4所示。

超参数	参数意义	参数值
memory_size	Replay Memory 的大小	5000
batch_size	mini-batch 的大小	128
hidden_dim	DQN 的隐层维度	50
target_c	$\hat{Q}$ 的更新频率	10
discount	DQN 算法中的 γ	0.99
learning_rate	DQN 算法中的 α	0.001
eps_start	ϵ 的初始值	0.9
eps_end	ϵ 的结束值	0.05
eps_decay	ϵ 的衰减权重	200

表 4: Improved DQN 超参数设置 (CartPole)

Improved DQN 在 CartPole 上的实验结果如图 4所示。



(a) Average Loss (b) Average Reward

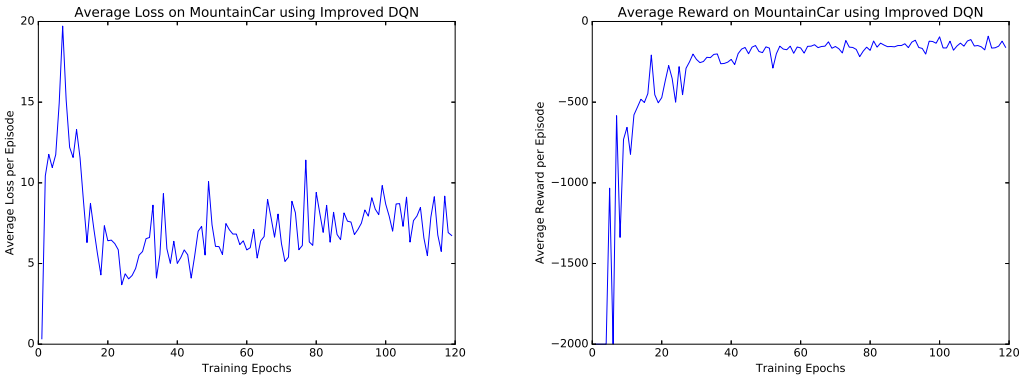
图 4: Training Result of CartPole using Improved DQN

MountainCar

超参数	参数意义	参数值
memory_size	Replay Memory 的大小	5000
batch_size	mini-batch 的大小	128
hidden_dim	DQN 的隐层维度	50
target_c	$\hat{Q}$ 的更新频率	10
discount	DQN 算法中的 γ	0.99
learning_rate	DQN 算法中的 α	0.001
eps_start	ϵ 的初始值	0.9
eps_end	ϵ 的结束值	0.05
eps_decay	ϵ 的衰减权重	50

表 5: Improved DQN 超参数设置 (MountainCar)

Improved DQN 在 MountainCar 上的实验结果如图 5所示。



(a) Average Loss (b) Average Reward

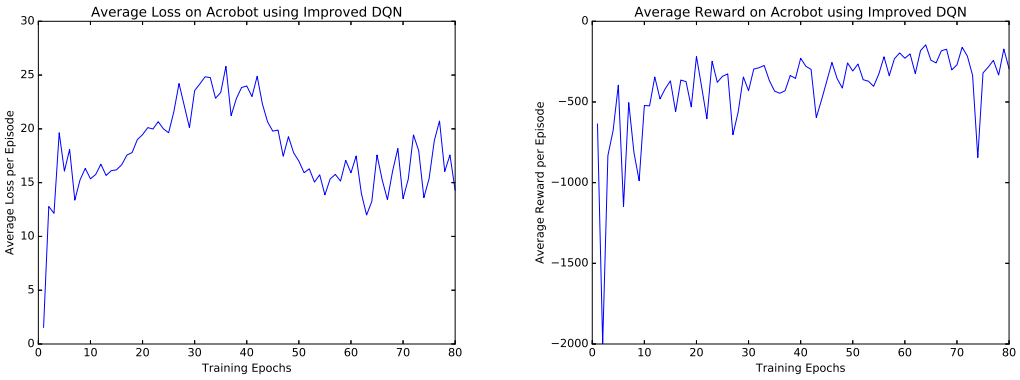
图 5: Training Result of MountainCar using Improved DQN

Acrobot

超参数	参数意义	参数值
memory_size	Replay Memory 的大小	10000
batch_size	mini-batch 的大小	128
hidden_dim	DQN 的隐层维度	50
target_c	$\hat{Q}$ 的更新频率	5
discount	DQN 算法中的 γ	0.99
learning_rate	DQN 算法中的 α	0.001
eps_start	ϵ 的初始值	0.9
eps_end	ϵ 的结束值	0.05
eps_decay	ϵ 的衰减权重	200

表 6: Improved DQN 超参数设置 (Acrobot)

Improved DQN 在 Acrobot 上的实验结果如图 6所示。



(a) Average Loss (b) Average Reward

图 6: Training Result of Acrobot using Improved DQN