



新闻

体育

汽车

房产

旅游

教育

时尚

科技

财经

娱乐

更多

登录狐友

400 XMM - 11111 - 11111

0

分享到

新智元

1770  
文章1182万  
总阅读

查看TA的文章&gt;

# 【强化学习实战】基于gym和tensorflow的强化学习算法实现

2017-09-13 06:04

程序设计 / 搜狐

1新智元推荐

编辑：熊笑

【新智元导读】知乎专栏强化学习大讲堂作者郭宪博士开讲《强化学习从入门到进阶》，我们为您节选了其中的第二节《基于gym和tensorflow的强化学习算法实现》，希望对您有所帮助。同时，由郭宪博士等担任授课教师的深度强化学习国庆集训营也将于10月2日—6日在北京举办。

## 基于gym和tensorflow的强化学习算法实现

上一讲已经深入剖析了 gym 环境的构建强化学习实战《第一讲 gym学习及二次开发 - 知乎专栏》。这一讲，我们将利用gym和tensorflow来实现两个最经典的强化学习算法qleanring和基于策略梯度的方法。本节课参考了莫烦的部分代码（见知乎问答《强化学习（reinforcement learning）有什么好的开源项目、网站、文章推荐一下？》），在此对其表示感谢。这一讲分为两个小节，2.1小节讲讲用qlearning的方法解决机器人找金币（该环境已经在上一节给出）；2.2 小节以小车倒立摆为例子，详细讲解基于策略梯度的强化学习方

24小时热文

大家都在搜：iPhone8屏幕

1

叙利亚对  
卧场面

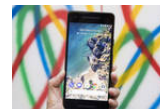
2

基于安  
6界面曝

3

10月16  
机的下-

货到付款

我敢发誓  
你没穿过这么谷歌亲  
看完拍iPhone  
果粉们

热门图集



老照片再现前苏联核试验 场面震撼





新闻

体育

汽车

房产

旅游

教育

时尚

科技

财经

娱乐

更多

登录狐友

0

分享到

## 第1小节 qlearning算法实现

### 1.1 qlearning伪代码

qlearning算法是off-policy的基于值函数的TD(0)强化学习算法。基于值函数的强化学习算法的本质是更新值函数。其理论和伪代码已经在第四讲给出。现在我们回顾一下：

1. 初始化  $Q(s, a), \forall s \in S, a \in A(s)$ , 给定参数  $\alpha, \gamma$
2. Repeat:
  - 给定起始状态  $s$ , 并根据  $\epsilon$  贪婪策略在状态  $s$  选择动作  $a$
  - Repeat (对于一幕的每一步)
    - (a) 根据  $\epsilon$  贪婪策略在状态  $s_t$  选择动作  $a_t$ , 得到回报  $r_t$  和下一个状态  $s_{t+1}$
    - (b)  $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha [r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)]$  ← 目标策略为贪婪策略
    - (c)  $s = s', a = a'$
  - Until  $s$  是终止状态
  - Until 所有的  $Q(s, a)$  收敛
3. 输出最终策略:  $\pi(s) = \operatorname{argmax}_a Q(s, a)$

图2.1 qlearning 算法伪代码

从图2.1中我们看到，qlearning算法的实现可以分为以下关键点：行为值函数的表示，探索环境的策略，epsilon贪婪策略，值函数更新时选择动作的贪婪策略，值函数更新。下面，我就逐个讲解一下。

### 1.2 qlearning的行为值函数表示



太有才！中秋的月亮都快被玩儿坏了

## 24小时热文

1

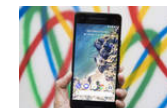
叙利亚卧场面

2

基于安6界面曝

3

10月16机的下-



谷歌亲看完拍



iPhone果粉们

## 24小时热文

1

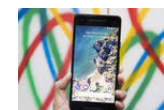
携程订票100亿？官方回应；iPhone 8...

2

发生在国家级贫困县的一幕

3

10月16日，Mate 10或将开启手机的下一个时代！



谷歌亲儿子Pixel 2值不值得买？看完拍摄样张再说



新闻

体育

汽车

房产

旅游

教育

时尚

科技

财经

娱乐

更多

登录狐友

0

分享到

表，其中一维为状态，另外一维为动作。对于机器人找金币的例子：

状态空间为：[1,2,3,4,5,6,7,8]

动作空间为：['n', 'e', 's', 'w']

行为值函数可以用字典数据类型来表示，其中字典的索引由状态-动作对来表示。因此行为值函数的初始化为：

```
qfunc = dict() #行为值函数为qfun
```

```
for s in states:
```

```
for a in actions:
```

```
key = "d%_s%"%(s,a)
```

```
qfun[key] = 0.0
```

### 1.3 探索环境的策略：epsilon贪婪策略

智能体通过epsilon贪婪策略来探索环境，epsilon贪婪策略的数学表达式为：



果粉们或继续等待至12月份

#### 24小时热文

1

叙利亚  
卧场面

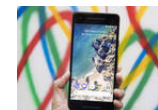
2

基于安  
6界面曝

3

10月16  
机的下-

¥148.00 ¥296



谷歌亲  
看完拍

#### 搜狐号推荐



IT之家  
IT之家是业内领先  
网站。IT之家快速



iPhone  
果粉们



搜狐科技视界  
搜狐科技官方原创...  
件、大趋势和新变化，用我们的视角观...



PingWest品玩  
有品好玩的科技，一切与你有关。



IT观察  
洞穿繁杂现象，呈现绝不流俗的观点。有理，  
有据，有温度。



零镜网  
由多位资深科技媒体人建立，专注于研究科技  
新物种的精品阅读平台；我们只向...



新闻

体育

汽车

房产

旅游

教育

时尚

科技

财经

娱乐

更多

登录狐友

0

分享到

该式子的python代码实现为：

```
def epsilon_greedy(qfunc, state, epsilon):
```

```
#先找到最大动作
```

```
amax = 0
```

```
key = "%d_%s"%(state, actions[0])
```

```
qmax = qfunc[key]
```

```
for i in range(len(actions)): #扫描动作空间得到最大动作值函数
```

```
key = "%d_%s"%(state, actions[i])
```

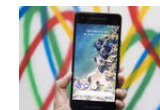
```
q = qfunc[key]
```

```
if qmax < q:
```

```
qmax = q
```



联系我们

谷歌亲  
看完拍iPhone  
果粉们

0

分享到

```
#概率部分

pro = [0.0 for i in range(len(actions))]

pro[amax] += 1-epsilon

for i in range(len(actions)):

    pro[i] += epsilon/len(actions)

##根据上面的概率分布选择动作

r = random.random()

s = 0.0

for i in range(len(actions)):

    s += pro[i]

if s>= r: return actions[i]

return actions[len(actions)-1]
```

24小时热文

- 1

叙利亚对卧场面
- 2

基于安...  
6界面曝
- 3

10月16机的下-
- 

谷歌亲...  
看完拍
- 

iPhone果粉们



新闻

体育

汽车

房产

旅游

教育

时尚

科技

财经

娱乐

更多

登录狐友

0

分享到

作；第2小段给每个动作分配概率；第3小段是根据概率分布采样一个动作。

#### 1.4 值函数更新时，选择动作的贪婪策略

选择动作的贪婪策略就是选择状态为s'时，值函数最大的动作。其python实现为：

```
def greedy(qfunc, state):
```

```
    amax = 0
```

```
    key = "%d_%s" % (state, actions[0])
```

```
    qmax = qfunc[key]
```

```
    for i in range(len(actions)): # 扫描动作空间得到最大动作值函数
```

```
        key = "%d_%s" % (state, actions[i])
```

```
        q = qfunc[key]
```

```
        if qmax < q:
```

```
            qmax = q
```

```
    amax = i
```

#### 24小时热文

1

叙利亚对  
卧场面

2

基于安  
6界面曝

3

10月16  
机的下-谷歌亲  
看完拍iPhone  
果粉们



新闻

体育

汽车

房产

旅游

教育

时尚

科技

财经

娱乐

更多

登录狐友

0

分享到

该段代码与上段代码几乎一样，不同的是所取的状态值不一样。该段代码的状态是当前状态s的下一个状态s'。另外，DQN所做的改变是用来选择行为的值函数网络称为目标值函数网络，跟当前值函数网络不同。

## 1.5 值函数更新

值函数更新公式为：

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha [r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)]$$

代码实现为：

```
key = "%d_%s"%(s, a)
```

```
#与环境进行一次交互，从环境中得到新的状态及回报
```

```
s1, r, t1, i = grid.step(a)
```

```
key1 = ""
```

```
#s1处的最大动作
```

```
a1 = greedy(qfunc, s1)
```

## 24小时热文

1

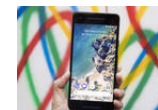
叙利亚对  
卧场面

2

基于安  
6界面曝

3

10月16  
机的下-



谷歌亲  
看完拍



iPhone  
果粉们



新闻

体育

汽车

房产

旅游

教育

时尚

科技

财经

娱乐

更多

登录狐友

0

分享到

## #利用qlearning方法更新值函数

```
qfunc[key] = qfunc[key] + alpha*(r + gamma * qfunc[key1]-qfunc[key])
```

评论：对于表格型值函数更新过程，我们看到每次更新只影响表格中的一个值，而不会影响行为值函数的其他值，这与行为值函数逼近方法不同。表格型强化学习算法效率很高，一般经过几次迭代后便能收敛。全部代码请参看github.gxnk中的qlearning。qlearning 算法的测试在文件learning\_and\_test.py中

## 第2小节：基于策略梯度算法实现详解

该部分需要用到tensorflow和画图库，所以大家先安装一下cpu版的tensorflow。

## 2.1 Tensorflow的安装：

Step1: 在终端激活虚拟环境（如何安装在上一讲）：source activate gymlab

Step2: 安装的tensorflow版本为1.0.0，python=3.5如下命令：

```
pip install --ignore-installed --upgrade https://storage.googleapis.com/tensorflow/linux/cpu/tensorflow-1.0.0-cp35-cp35m-linux_x86_64.whl
```

根据该命令所安装的tensorflow是无gpu的，无gpu的tensorflow对于学习毫无障碍。当然，如果大家做项目，建议安装gpu版的tensorflow.

## 24小时热文

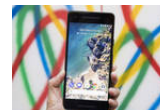
1

叙利亚  
卧场面

2

基于安  
6界面曝

3

10月16  
机的下-谷歌亲  
看完拍iPhone  
果粉们





新闻

体育

汽车

房产

旅游

教育

时尚

科技

财经

娱乐

更多

登录狐友

0

pip3 install matplotlib

分享到

## 2.2 策略梯度算法理论基础

本专栏的第六讲已经给出了策略梯度的理论推导，策略梯度理论表明随机策略的梯度由下式给出：

$$\nabla_{\theta} J(\pi_{\theta}) = E_{s \sim \rho^{\pi}, a \sim \pi_{\theta}} [\nabla_{\theta} \log \pi_{\theta}(a|s) Q^w(s, a)]$$

当随机策略是高斯策略的时候，第六讲已经给出了随机梯度的计算公式。当随机策略并非高斯策略时，如何优化参数？

对于小车倒立摆系统如下图2.2所示。

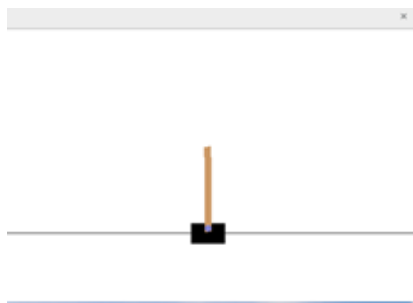


图2.2 小车倒立摆系统

## 24小时热文

1

叙利亚对  
卧场面

2

基于安  
6界面曝

3

10月16  
机的下-谷歌亲  
看完拍iPhone  
果粉们



新闻

体育

汽车

房产

旅游

教育

时尚

科技

财经

娱乐

更多

登录狐友

小车倒立摆的状态空间为

0

分享到

$$\begin{bmatrix} x, \dot{x}, \theta, \dot{\theta} \end{bmatrix}$$

, 动作空间为

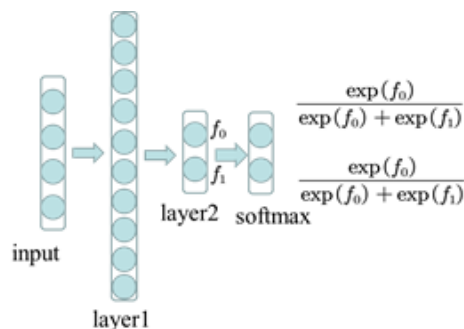
$$\{0, 1\}$$

, 当动作为1时, 施加正向的力10N;当动作为0时, 施加负向的力-10N。

因为动作空间是离散的, 因此我们设计随机策略为softmax策略。Softmax策略如何构建, 以及如何构建损失函数, 从而将强化学习问题变成一个优化问题。

### 2.3 softmax策略及其损失函数

我们设计一个前向神经网络策略, 如图2.3所示。



### 24小时热文

1

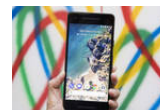
叙利亚对  
卧场面

2

基于安  
6界面曝

3

10月16  
机的下-



谷歌亲  
看完拍



iPhone  
果粉们



新闻

体育

汽车

房产

旅游

教育

时尚

科技

财经

娱乐

更多

登录狐友

0

分享到

该神经softmax策略的输入层是小车倒立摆的状态，维数为4；最后一层是softmax层，维数为2。有机器学习同学都很清楚，softmax常常作为多分类器的最后一层。

一个最基本的概念是何为softmax层？

如图2.3，设layer2的输出为z，所谓softmax层是指对z作用一个softmax函数。即：

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}}, \text{ for } j = 1, \dots, K$$

对于softmax策略，策略梯度理论中的随机策略为：

$$\pi_{\theta}(a|s) = \frac{e^{f_a}}{\sum_{k=1}^K e^{f_k}}$$

如图2.3所示，对应着 layer2 的输出。

$$e^{f_a}$$

表示动作 a 所对应的softmax输出。上面的式子便给出了智能体在状态s处采用动作a的概率。该式是关于的函数，可直接对其求对数，然后求导带入到策略梯度公式，利用策略梯度

## 24小时热文

1

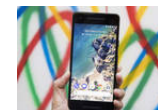
叙利亚对  
卧场面

2

基于安  
6界面曝

3

10月16  
机的下-



谷歌亲  
看完拍



iPhone  
果粉们



新闻

体育

汽车

房产

旅游

教育

时尚

科技

财经

娱乐

更多

登录狐友

步更新，其实是对损失函数为

0

分享到

的一步更新。

而损失函数可写为：

$$L = -E_{s \sim \rho^\pi, a \sim \pi_\theta} [\log \pi_\theta(a|s) Q^w(s, a)]$$

$$L = -E_{s \sim \rho^\pi, a \sim \pi_\theta} [\log \pi_\theta(a|s) Q^w(s, a)] = - \int p_{\pi_{\theta_{old}}} \log q_{\pi_\theta} Q^w(s, a)$$

其中

$$- \int p_{\pi_{\theta_{old}}} \log q_{\pi_\theta}$$

为交叉熵。

在实际计算中，

$$p_{\pi_{\theta_{old}}}$$

由未更新的参数策略网络进行采样，

$$\log q_{\pi_\theta}$$

## 24小时热文

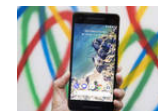
1

叙利亚对  
卧场面

2

基于安  
6界面曝

3

10月16  
机的下-谷歌亲  
看完拍iPhone  
果粉们

0

分享到

, 产生为 a=1; 则

$$\pi_{\theta_{old}}(s) \quad p = [0, 1]$$

$$p = [0, 1]$$

,

$$q = \left[ \frac{\exp(f_0)}{\exp(f_0) + \exp(f_1)}, \frac{\exp(f_1)}{\exp(f_0) + \exp(f_1)} \right]$$

, 则

$$p_{\pi_{\theta_{old}}} \log q_{\pi_{\theta}} = \log \frac{\exp(f_1)}{\exp(f_0) + \exp(f_1)}$$

这是从信息论中交叉熵的角度来理解softmax层。理论部分就暂时介绍到这，接下来我们关心的是如何将理论变成代码。

上面，我们已经将策略梯度方法转化为一个分类问题的训练过程，其中损失函数为：

$$L = -E_{s \sim \rho^{\pi}, a \sim \pi_{\theta}} [\log \pi_{\theta}(a|s) Q^w(s, a)] = - \int p_{\pi_{\theta_{old}}} \log q_{\pi_{\theta}} Q^w(s, a)$$

24小时热文

1

叙利亚对  
卧场面

2

基于安  
6界面曝

3

10月16  
机的下-



谷歌亲  
看完拍



iPhone  
果粉们



新闻

体育

汽车

房产

旅游

教育

时尚

科技

财经

娱乐

更多

登录狐友

0

分享到

输入数据有三个：

第一：小车倒立摆的状态 $s$

第二：作用在小车上的动作 $a$

第三：每个动作对应的累积回报 $v$

我们一一解释，这些输入如何获得。

首先，小车倒立摆的状态 $s$ ，这是与环境交互得到的；其次，作用在小车上的动作 $a$ ，是由采样网络得到的，在训练过程中充当标签的作用；最后，每个动作的累积回报是由该动作后的累积回报累积并进行归一化处理得到的。

因此，该代码可以分为几个关键的函数：**策略神经网络的构建**，**动作选择函数**，**损失函数的构建**，**累积回报函数 $v$ 的处理**。下面我们一一介绍如何实现。

## 2.4 基于 tensorflow 的策略梯度算法实现

### 策略网络的构建

构建一个神经网络，最简单的方法就是利用现有的深度学习软件，由于兼容性和通用性，这里我们选择了tensorflow。我们要构建的策略网络结构为如图2.4：

### 24小时热文

1

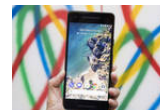
叙利亚对  
卧场面

2

基于安  
6界面曝

3

10月16  
机的下-



谷歌亲  
看完拍



iPhone  
果粉们



新闻

体育

汽车

房产

旅游

教育

时尚

科技

财经

娱乐

更多

登录狐友

0

分享到

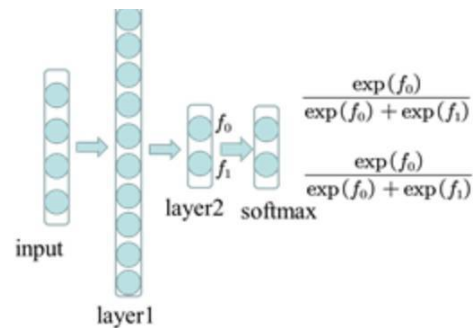


图2.4 策略神经网络

该神经网络是最简单的前向神经网络，输入层为状态s，共4个神经元，第一个隐藏层包括为10个神经元，激活函数为relu。因为输出为动作的概率，而动作有两个，因此第二层为2个神经元，没有激活函数，最后一层为softmax层。

将这段代码翻译成tensorflow语言则为：

```
def _build_net(self):
```

```
    with tf.name_scope('input'):
```

```
        #创建占位符作为输入
```

```
        self.tf_obs = tf.placeholder(tf.float32, [None, self.n_features], name="observations")
```

```
        self.tf_acts = tf.placeholder(tf.int32, [None, ], name="actions_num")
```

## 24小时热文

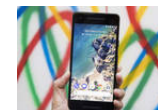
1

叙利亚对  
卧场面

2

基于安  
6界面曝

3

10月16  
机的下-谷歌亲  
看完拍iPhone  
果粉们

0

#第一层

分享到

```
layer = tf.layers.dense(  
  
    inputs=self.tf_obs,  
  
    units=10,  
  
    activation=tf.nn.tanh,  
  
    kernel_initializer=tf.random_normal_initializer(mean=0, stddev=0.3),  
  
    bias_initializer=tf.constant_initializer(0.1),  
  
    name='fc1',  
  
)
```

#第二层

```
all_act = tf.layers.dense(  
  
    inputs=layer,
```

24小时热文

- 1

叙利亚对  
卧场面
- 2

基于安  
6界面曝
- 3

10月16  
机的下-
- 

谷歌亲  
看完拍
- 

iPhone  
果粉们





新闻

体育

汽车

房产

旅游

教育

时尚

科技

财经

娱乐

更多

登录狐友

0

分享到

`activation=None,``kernel_initializer=tf.random_normal_initializer(mean=0, stddev=0.3),``bias_initializer=tf.constant_initializer(0.1),``name='fc2'``)`

#利用softmax函数得到每个动作的概率

`self.all_act_prob = tf.nn.softmax(all_act, name='act_prob')`

全部代码可去github上看，在policynet.py文件中。

### 动作选择函数：

动作选择函数是根据采样网络生成概率分布，利用该概率分布去采样动作，具体代码为：

#定义如何选择行为，即状态 s 处的行为采样.根据当前的行为概率分布进行采样

`def choose_action(self, observation):`

### 24小时热文

1

叙利亚对  
卧场面

2

基于安  
6界面曝

3

10月16  
机的下-谷歌亲  
看完拍iPhone  
果粉们



新闻

体育

汽车

房产

旅游

教育

时尚

科技

财经

娱乐

更多

登录狐友

0

分享到

```
{self.tf_obs:observation[np.newaxis,:])
```

```
#按照给定的概率采样
```

```
action = np.random.choice(range(prob_weights.shape[1]), p=prob_weights.ravel())
```

```
return action
```

其中函数np.random.choice是按照概率分布p=prob\_weights.ravel()进行采样的函数。

**损失函数的构建：**

在理论部分我们已经说明了损失函数为

即交叉熵乘以累积回报函数。以下为代码部分：

```
#定义损失函数
```

```
with tf.name_scope('loss'):
```

```
neg_log_prob =
```

```
tf.nn.sparse_softmax_cross_entropy_with_logits(logits=all_act,labels=self.tf_acts)
```

```
loss = tf.reduce_mean(neg_log_prob*self.tf_vt)
```

## 24小时热文

1

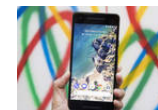
叙利亚对  
卧场面

2

基于安  
6界面曝

3

10月16  
机的下-



谷歌亲  
看完拍



iPhone  
果粉们



新闻

体育

汽车

房产

旅游

教育

时尚

科技

财经

娱乐

更多

登录狐友

0

分享到

```
def _discount_and_norm_rewards(self):
```

```
#折扣回报和
```

```
discounted_ep_rs = np.zeros_like(self.ep_rs)
```

```
running_add = 0
```

```
for t in reversed(range(0, len(self.ep_rs))):
```

```
running_add = running_add * self.gamma + self.ep_rs[t]
```

```
discounted_ep_rs[t] = running_add
```

```
#归一化
```

```
discounted_ep_rs -= np.mean(discounted_ep_rs)
```

```
discounted_ep_rs /= np.std(discounted_ep_rs)
```

```
return discounted_ep_rs
```

有了策略神经网络，动作选择函数，损失函数，累积回报函数之后，学习的过程就简单了，只需要调用一个语句即可：

## 24小时热文

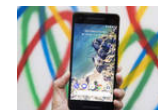
1

叙利亚对  
卧场面

2

基于安  
6界面曝

3

10月16  
机的下-谷歌亲  
看完拍iPhone  
果粉们



新闻

体育

汽车

房产

旅游

教育

时尚

科技

财经

娱乐

更多

登录狐友

0

分享到

```
with tf.name_scope('train'):
```

```
self.train_op = tf.train.AdamOptimizer(self.lr).minimize(loss)
```

该训练过程为采用自适应动量的优化方法。学习优化的过程如下：

```
#学习，以便更新策略网络参数，一个episode之后学一回
```

```
def learn(self):
```

```
#计算一个episode的折扣回报
```

```
discounted_ep_rs_norm = self._discount_and_norm_rewards()
```

```
#调用训练函数更新参数
```

```
self.sess.run(self.train_op, feed_dict={
```

```
self.tf_obs: np.vstack(self.ep_obs),
```

```
self.tf_acts: np.array(self.ep_as),
```

```
self.tf_vt: discounted_ep_rs_norm,
```

## 24小时热文

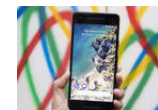
1

叙利亚对  
卧场面

2

基于安  
6界面曝

3

10月16  
机的下-谷歌亲  
看完拍iPhone  
果粉们



新闻

体育

汽车

房产

旅游

教育

时尚

科技

财经

娱乐

更多

登录狐友

0

#清空episode数据

分享到

```
self.ep_obs, self.ep_as, self.ep_rs = [], [], []
```

```
return discounted_ep_rs_norm
```

## 2.5 基于策略梯度算法的小车倒立摆问题

有了策略网络和训练过程，对于解决小车的问题就很简单了。基本的框架为：

1. 创建一个环境

2. 生成一个策略网络

3. 迭代学习

通过与环境交互，学习更新策略网络参数

4. 利用学到的策略网络对小车倒立摆系统进行测试

利用softmax策略定义一个贪婪策略。

具体代码在github上的learning\_cartpole.py文件中。

## 24小时热文

1

叙利亚对  
卧场面

2

基于安  
6界面曝

3

10月16  
机的下-



谷歌亲  
看完拍



iPhone  
果粉们



新闻

体育

汽车

房产

旅游

教育

时尚

科技

财经

娱乐

更多

登录狐友

0

分享到

PS：该部分讲得有点乱，强烈建议大家去github上下载代码，我已经做好了中文注释，大家可以改改参数，亲身体会下。所有代码都在github的第一讲文件夹中gxnk/reinforcement-learning-code。

### 国庆深度强化学习实战特训营

由郭博士和香港理工大学增强学习方向博士 Traffas 担任授课教师的深度强化学习国庆集训营将于 10 月 2 日— 6 日在北京举办。

课程主办方 | 探灵教育科技

# 中秋大趴

## 深度强化学习国庆集训营

授课教师：郭博士 Traffas

郭博士 | 知乎专栏强化学习大讲堂作者《深入浅出强化学习》（即将出版）作者  
Traffas | 香港理工大学 增强学习方向博士

10月2日-6日 北京  
强化学习知识大讲堂 <https://zhuanlan.zhihu.com/sharer1>  
强化学习知识分享群 202570720  
扫描二维码报名

### 24小时热文

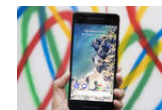
1

叙利亚  
卧场面

2

基于安  
6界面曝

3

10月16  
机的下-谷歌亲  
看完拍iPhone  
果粉们



新闻

体育

汽车

房产

旅游

教育

时尚

科技

财经

娱乐

更多

登录狐友

0

分享到

探灵教育科技有限公司在8月初已经成功举办第一期强化学习入门进阶培训课程，受到学员一致好评。根据学员的反馈以及我们最新的教研成果，我们进一步对课程进行了升级、完善。国庆期间，特别推出为期五天的强化学习特训营活动，通过五天的理论讲解以及编程实战，帮助大家全面、系统的了解、掌握强化学习技术。

面向对象：强化学习的小白、初学者、自己已有一定基础但是没有建立系统知识体系的以及其他对于强化学习感兴趣的人士。有一定的微积分、线线性代数、概率论基础，有python编程基础。学员上课需要自带电脑。

授课时间地点：10.2-10.6日 北京海淀区（具体地点另行通知）

招生人数：精品小班制，上限 30 人，报名15 人以上开班。

学费：7999 早鸟票 7499（9.24日之前报名）

特别声明：凡报名参加本次国庆特训营的学员，一年之内可以免费参加两次由我公司主办的为期两天的线下课程（价值5999元）。

讲师介绍：

**郭宪**，南开大学计算机与控制工程学院博士后。2009年毕业于华中科技大学机械设计制造及自动化专业，同年保送到中国科学院沈阳自动化研究所进行硕博连读，主攻机器人动力学建模与控制，于2016年1月获得工学博士学位，期间在国内外知名杂志和会议发表论文数10篇。2016年以来，郭博士主攻方向为机器人智能感知和智能决策，目前主持两项国家级课

## 24小时热文

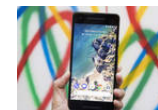
1

叙利亚  
卧场面

2

基于安  
6界面曝

3

10月16  
机的下-谷歌亲  
看完拍iPhone  
果粉们



新闻

体育

汽车

房产

旅游

教育

时尚

科技

财经

娱乐

更多

登录狐友

0

分享到

年3月开始在知乎专栏强化学习知识大讲室，其深入浅出的讲解收到广大知友一致好评。即将出版《强化学习深入浅出完全教程》一书。

知乎链接：<https://zhuanlan.zhihu.com/sharerl>

**Traffas**，于2014年7月在瑞典皇家理工学院获得硕士学位，曾在瑞典Accedo公司做程序开发，现在在香港理工大学计算机系攻读博士学位，任研究助理。Traffas 的研究方向为机器学习、增强学习。目前已发表六篇论文，其中包括中国计算机学会（CCF）推荐的B类论文1篇，C类会议论文1篇。

日程安排：

第一天：授课老师 Traffas

1. 什么是强化学习以及强化学习的方法汇总？

强化学习可以让AlphaGo无需人类的指导，自己‘左右互搏’，就能悟到更佳出奇制胜的围棋技巧；可以让机器人的行动不再需要人类繁杂的编程，自己就可以适应所处的环境。为什么强化学习有如此神奇的功能？到底什么是强化学习？本课将为你娓娓道来....

2. 强化学习领域的基础概念。

解锁强化学习领域的术语。介绍增强学习可以解决的问题。介绍Bellman Equation原理,介绍RL和动态规划的异同点。介绍传统的tubular based RL。

## 24小时热文

1

叙利亚  
卧场面

2

基于安  
6界面曝

3

10月16  
机的下-



谷歌亲  
看完拍



iPhone  
果粉们





新闻

体育

汽车

房产

旅游

教育

时尚

科技

财经

娱乐

更多

登录狐友

0

分享到

4. 动手编写第一个增强学习的python 程序（30分钟），找到玩老虎机的最优策略。

5. 基于蒙特卡罗强化学习介绍、同策略、异策略强化学习算法介绍。

6. 答疑、交流

第二天：授课老师 Traffas

1、 强化学习算法实践，基于强化学习玩21点游戏以及gridworld游戏。

2、 强化学习时间差分算法。介绍同策略Q-learning强化学习方法以及异策略Sara算法。比较和蒙特卡洛算法异同点。介绍eligibility Tree以及TD (lamda) 算法。

3、 Gym环境构建以及强化学习算法实现。包括Gym环境的安装、测试，Gym环境关键函数讲解以及如何创建自定义Gym环境。

4、 学员动手实践

5、 老师答疑、交流。

第三天：授课老师 Traffas

1、 DQN详解

## 24小时热文

1

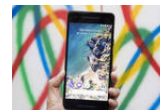
叙利亚对  
卧场面

2

基于安  
6界面曝

3

10月16  
机的下-



谷歌亲  
看完拍



iPhone  
果粉们



新闻

体育

汽车

房产

旅游

教育

时尚

科技

财经

娱乐

更多

登录狐友

0

分享到

电子游戏中的表现超过了人类的顶级玩家。同时，我们会讲解DQN的变种Double DQN、Prioritized Replay，Dual DQN。

## 2、实践准备

介绍深度神经网络DNN以及RNN。Keras安装，动手设计RNN网络，解决分类问题。

3、深度强化学习实战，亲自动手编写一个可以打败游戏高手的AI。

4、Bug调试、老师答疑、指导、交流。

## 第四天：授课老师 郭宪

### 1、策略梯度方法：

教学内容包括：策略梯度方法介绍，似然率策略梯度推导及重要性采样视角推导，似然率策略梯度的直观理解，常见的策略表示，常见的减小方差的方法：引入基函数法，修改估计值函数法

2、编程实践课：基于tensorflow和gym实现小车倒立摆系统、乒乓球游戏

3、TRPO方法介绍及推导：具体包括替代回报函数的构建，单调的改进策略，TRPO实用算法介绍，共轭梯度法搜索可行方向，PPO方法，基于python的TRPO方法实现

4、编程指导、交流、答疑。

## 24小时热文

1

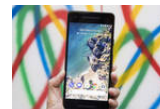
叙利亚对  
卧场面

2

基于安  
6界面曝

3

10月16  
机的下-



谷歌亲  
看完拍



iPhone  
果粉们

0

分享到

1、AC方法，具体内容包括随机策略与确定性策略比较、随机策略 AC 的方法、确定性策略梯度方法、DDPG方法及实现、A3C方法讲解、基于python 的 DDPG 方法实现。

2、AC方法及DDPG、A3C实现。

3、逆向强化学习介绍，包括逆向强化学习分类、学徒学习、MMP 方法、结构化分类方法、神经逆向强化学习、最大熵逆向强化学习、相对熵逆向强化学习、深度逆向强化学习。

4、编程指导、答疑、交流。

报名请扫海报中的二维码。 [返回搜狐，查看更多](#)

声明：本文由入驻搜狐号的作者撰写，除搜狐官方账号外，观点仅代表作者本人，不代表搜狐立场。

阅读 (2184)

不感兴趣

投诉

本文相关推荐

强化学习算法	基于内容推荐算法+实现	tensorflow神经网络算法
深度学习+tensorflow教程	c语言实现贪婪算法	基于密度的聚类算法
基于遗传算法的路径规划	基于dv算法的路由模拟	基于直方图去雾算法
基于分解的多目标进化算法	基于边缘的图像分割算法	基于蚁群算法的无人机航迹

24小时热文

1

叙利亚男  
卧场面

2

基于安  
6界面曝

3

10月16  
机的下-



谷歌亲  
看完拍



iPhone  
果粉们

0

分享到



¥67.00 ¥160



¥476.00 ¥529



¥499.00 ¥649

红米  
note8  
小米  
手机 广告

我来说两句

0人参与，0条评论

来说两句吧.....

登录并发表

搜狐“我来说两句” 用户公约

还没有评论，快来抢沙发吧！

推荐阅读



核心主管纷纷离职：贾跃亭的法拉第未来怎么了？

IT之家 · 今天 06:30

13

时隔半年再来辟谣，携程到底冤不冤？



虎嗅 虎嗅APP · 昨天 22:25

2

24小时热文

1

叙利亚  
卧场面

2

基于安  
6界面曝

3

10月16  
机的下-



谷歌亲  
看完拍



iPhone  
果粉们

推荐

谷歌

今日头条

双11

Pixel

LG

HTC

引力波

iOS



新闻

体育

汽车

房产

旅游

教育

时尚

科技

财经

娱乐

更多

登录狐友

安卓

微信小程序

尼康

“重明环返休”为何成「禁忌话题」？



虎嗅 虎嗅APP · 今天 05:56

18

ofo发布十一假期出行报告：共享单车用户日均出行次数涨超15%



36氪 36氪 · 今天 10:36

持股过节的人注意了！这个微信曝光节后个股行情，不看就亏大了！



广告 · 今天 12:36



苹果首席设计师乔纳森艾维：从未有人像乔布斯那样专注

IT之家 IT之家 · 今天 09:28

前有自动售货机 后有无人便利店，为何偏偏无人货架火了？

## 24小时热文

1

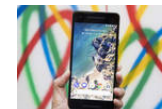
叙利亚对  
卧场面

2

基于安  
6界面曝

3

10月16  
机的下-



谷歌亲...  
看完拍



iPhone...  
果粉们

安卓十周岁了！这些消失的经典App你记得吗



 IT之家 · 今天 07:08

💬 5



为什么总是爱奇艺被传IPO？

 懂懂笔记 · 昨天 22:05

💬 1



任正非：数据是公司的核心资产，要像经营资本一样来“经营”数据

 36氪 · 今天 09:16

💬

《你的名字。》将拍真人版，老美的画风你敢看吗？



 极客视界 · 今天 11:47

💬



年轻人的第一次，惨。

 PingWest品玩 · 今天 10:27


💬


24小时热文

- 1

叙利亚男卧场面曝
- 2

基于安...  
6界面曝
- 3

10月16  
机的下-
- 

谷歌亲...  
看完拍
- 

iPhone  
果粉们



新闻

体育

汽车

房产

旅游

教育

时尚

科技

财经

娱乐

更多

登录狐友



IT观察 · 今天 07:20



蚕丝棉西裤，降价了！一年只降1次，超低价，不能再降了，厂家直销，买一送三！

广告 · 今天 12:36



通用自动驾驶车一个月被撞六次，终于明白国内自动驾驶上路有多难

爱范儿 · 昨天 15:51

9

电子墨水屏诞生 20 年，离取代纸张还有多远？



极客公园 · 今天 09:17



携程订票“陷阱”，一年坑消费者100亿？官方回应；iPhone 8至少5起爆裂 苹果...



## 24小时热文

1

叙利亚  
卧场面

2

基于安  
6界面曝

3

10月16  
机的下-



谷歌亲  
看完拍



iPhone  
果粉们



新闻

体育

汽车

房产

旅游

教育

时尚

科技

财经

娱乐

更多

登录狐友

## 【专访】女生适合玩哪种机械臂和3D打印，她用行动告诉你 | Madeline Gannon



文艺星球 · 昨天 19:00



### 【钛晨报】美国政府正式宣布将重返月球，建立永久性月球基地

钛媒体 钛媒体APP · 今天 07:11

1



### 微软Edge浏览器起全面登陆iOS/Android

品玩 PingWest品玩 · 今天 11:03



### Jony Ive：iPhone X 开发用了 5 年、设计和想法必须等待技术的跟进



PingWest品玩 · 昨天 16:34



### 蓝色起源未来18个月后将游客送上太空：略晚于Space X

IT之家 · 今天 06:50



加载更多

## 24小时热文

1

叙利亚对  
卧场面

2

基于安  
6界面曝

3

10月16  
机的下-



谷歌亲  
看完拍



iPhone  
果粉们