



🏠 > 人工智能 > 文章详情

# 人工智能进行连续决策的关键——强化学习入门指南

2017-06-12 阅读 15

分享到

搜索

管理

## 订阅热词

- 互联网 智能家居 VR AR 人工智能 智能出行 智能硬件

## 最新头条

朱啸虎：没有独角兽潜力的企业我们不投

2017-08-20 21:45

女神要和我做回朋友，原因竟然是因为它 | 钛空舱

2017-08-20 15:45

ofo多个公众号用车功能被封；高铁“复兴号”将扩大开行范围；苹果在印度推出热点专题

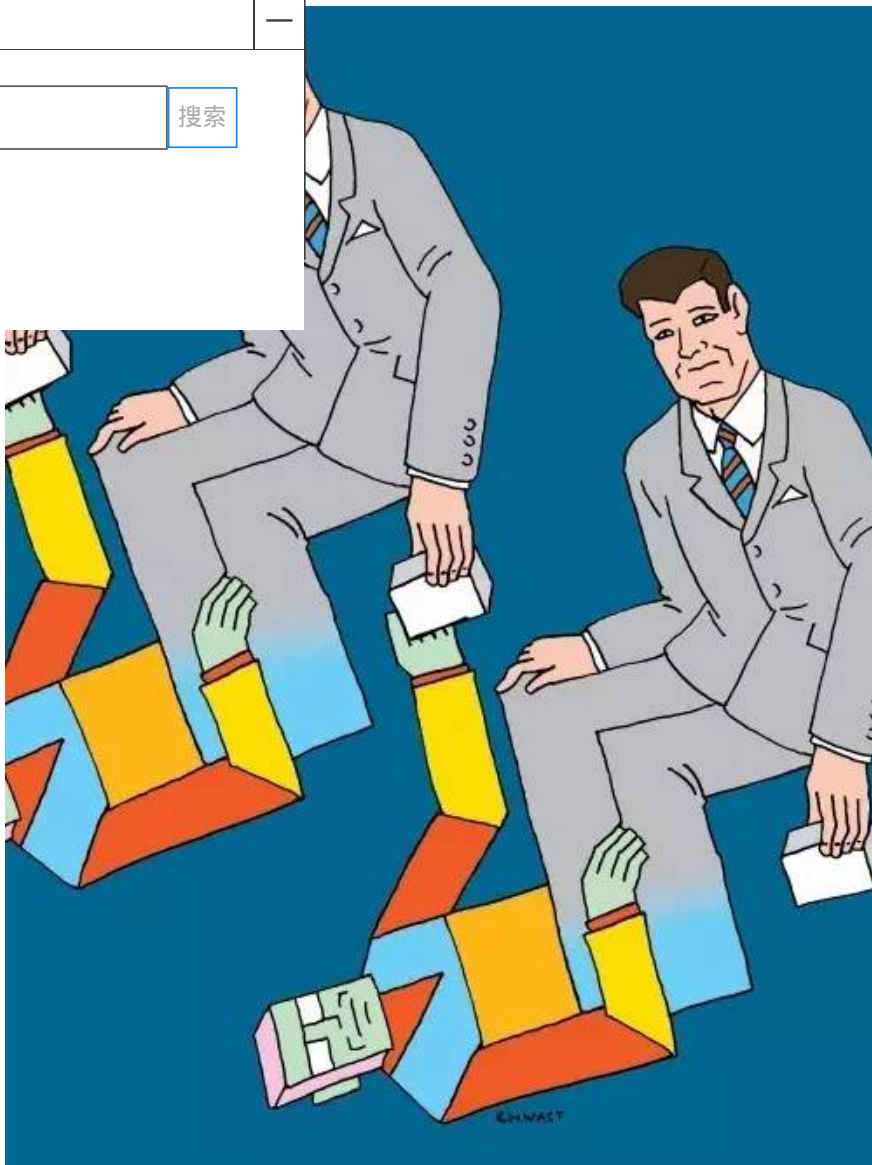
2017-08-20 08:25

反馈建议

在本站内搜索:

请输入关键词

搜索



跳出命运的怪圈，陈年走出他的“陈年”  
价值封面人物

2017-08-19 21:50

七夕宠溺好礼力荐，让Ta把无与伦比的美丽带回家 | 钛空舱

2017-08-19 15:45

支付宝就抄袭微信小程序代码致歉；全国首家互联网法院成立；摩托罗拉在研..

2017-08-19 08:30

周鸿祎：创业者少谈概念，低调做事

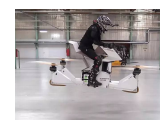
2017-08-18 21:15

## 最热收藏



看片 | 盲人要如何使用智能手表？

1 人收藏



看片 | 战斗民族发明的摩托车，骑上就能飞

1 人收藏

## 最热文章

文 | 不会停的蜗牛

CSDN AI专栏作家



在本站内搜索:

请输入关键词

搜索

更在于其在制造业、库存、电商、广告、推荐、金融、医疗等与我们生活息

## 1 定义

强化学习是机器学习的一个重要分支，是多学科多领域交叉的一个产物，它的本质是解决 **decision making 问题**，即自动进行决策，并且可以做连续决策。

它主要包含四个元素，**agent**，**环境状态**，**行动**，**奖励**，强化学习的目标就是获得最多的累计奖励。

让我们以小孩学习走路来做个形象的例子：

小孩想要走路，但在这之前，他需要先站起来，站起来之后还要保持平衡，接下来还要先迈出一条腿，是左腿还是右腿，迈出一步后还要迈出下一步。

小孩就是 agent，他试图通过采取行动（即行走）来操纵环境（行走的表面），并且从一个状态转变到另一个状态（即他走的每一步），当他完成任务的子任务（即走了几步）时，孩子得到奖励（给巧克力吃），并且当他不能走路时，就不会给巧克力。



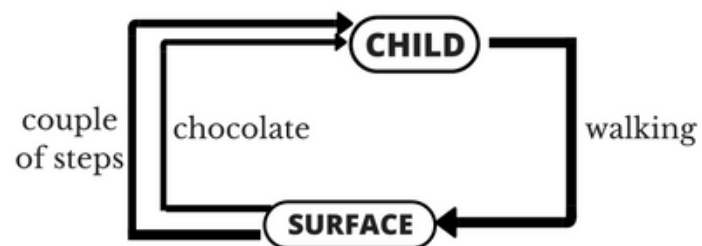
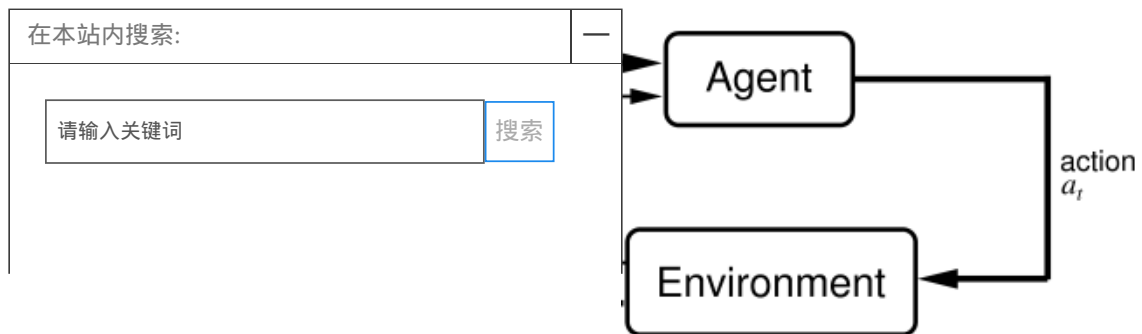
我只想让人生，有一点点不平凡



那个让实习生拿外卖的咪蒙又要招人



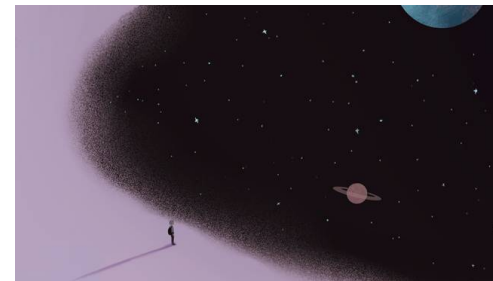
我的普通话，只有自己听得懂



— 2 —

## 强化学习与监督式、非监督式学习的区别

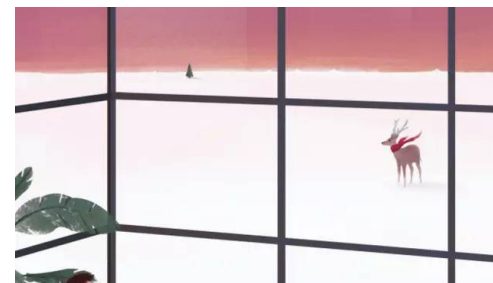
在机器学习中，我们比较熟知的是监督式学习，非监督学习，此外还有一个大类就是强化学习：



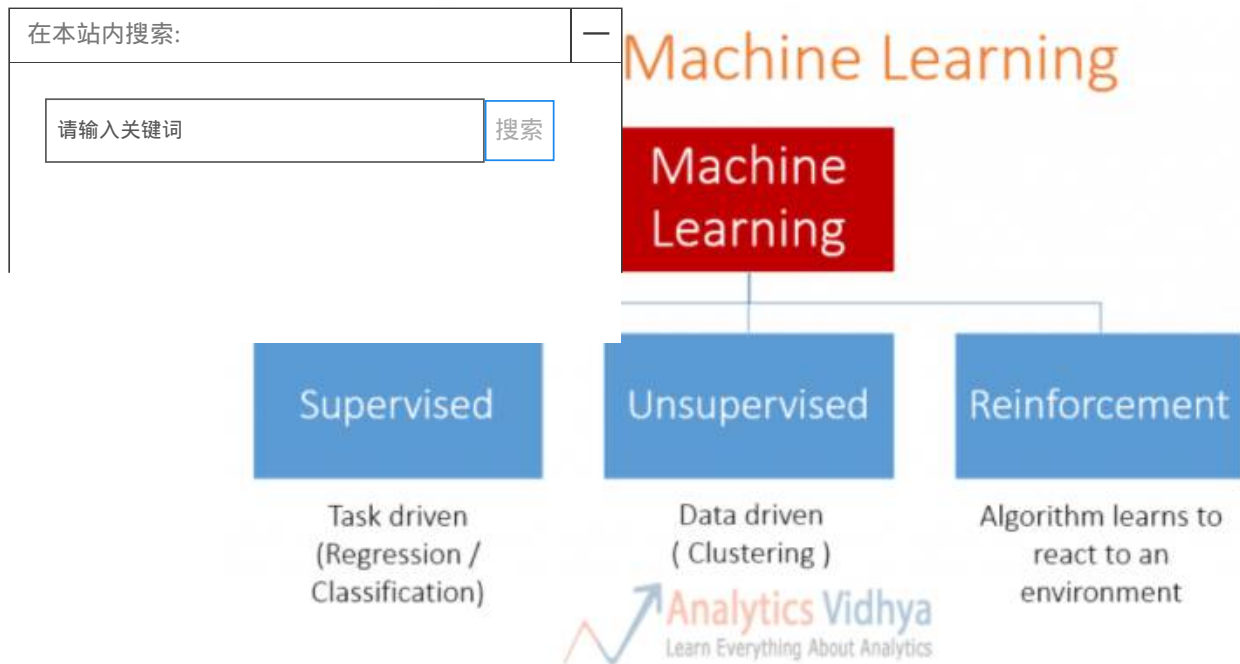
今天4月1日，我却想说件严肃的事



因为长得丑，我亏了150万



我青春里的两个男人和一本黄书



最可怕的是，为人父母不需要考试！

## 最新文章



合作活动 | CiGA开发者大会嘉宾揭露

2017-10-01 10:56



后续报道 | 中加国际电影节中的VR

2017-09-30 10:00



干货 | 虚拟（增强）现实白皮书（2017年）...

2017-09-29 13:22



9月末的热潮，TGS中的VR

2017-09-25 10:00



活动推荐 | WePlay想要承包你2017年10月的...

2017-09-19 12:01

## 强化学习和监督式学习的区别：

监督式学习就好比你在学习的时候，有一个导师在旁边指点，他知道怎么是对的怎么是错的，但在很多实际问题中，例如 chess，go，这种有成千上万种组合方式的情况，不可能有一个导师知道所有可能的结果。

而这时，强化学习会在没有任何标签的情况下，通过先尝试做出一些行为得到一个结果，通过这个结果是对还是错的反馈，调整之前的行为，就这样不断的调整，算法能够学习到在什么样的情况下选择什么样的行为可以得到最好的结果。

就好比有一只还没有训练好的小狗，每当它把屋子弄乱后，就减少美味食物的数量（惩罚），每次表现不错时，就加倍美味食物的数量（奖励），那么小狗最终会学到一个知识，就是把客厅弄乱是不好的行为。

两种学习方式都会学习出输入到输出的一个映射，监督式学习出的是之间的关系，可以告诉算法什么样的输入对应着什么样的输出，强化学习出的是给机器的反馈 reward function，即用来判断这个行为是好是坏。

另外强化学习的结果反馈有延时，有时候可能需要走了很多步以后才知道以前的某一步的选择是好还是坏，而监督学习做了比较坏的选择会立刻反馈给算法。

而且强化学习面对的输入总是在变化，每当算法做出一个行为，它影响下一次决策的输入，而监督学习的输入是独立同分布的。

通过强化学习，一个 agent 可以在探索和开发（exploration and exploitation）之间做权衡，并且选择一个最大的回报。

exploration 会尝试很多不同的事情，看它们是否比以前尝试过的更好。

itation 会尝试过去经验中最有效的行为。



在本站内搜索:

请输入关键词


搜索

— pitative。


和在向用户推荐新闻文章的任务中，非监督式会找到用户先前已经阅读过类似  
先推荐少量的新闻，并不断获得来自用户的反馈，最后构建用户可能会喜欢的

3

主要算法和分类



教程 | 360°全景图的绘制分析  
2017-09-14 10:00



最新虚幻引擎4 (UE4)  
AR游戏大作The...  
2017-09-13 18:53

从强化学习的几个元素的角度划分的话，方法主要有下面几类：

Policy based, 关注点是找到最优策略。

Value based, 关注点是找到最优奖励总和。

Action based, 关注点是每一步的最优行动。

我们可以用一个最熟知的旅行商例子来看，

我们要从 A 走到 F，每两点之间表示这条路的成本，我们要选择路径让成本越低越好：

```
graph LR; A ---|30| B; A ---|15| C; A ---|20| E; B ---|-10| C; B ---|10| D; C ---|50| D; C ---|300| E; D ---|-200| E; D ---|-100| F; E ---|1| F; C ---|-200| F;
```

大元素分别是：

<http://www.weiot.net/article-107610.html>

热点  
专题  
反馈  
建议

6/17

在本站内搜索:

请输入关键词

搜索

y，是一种 **Policy based** 的方法，当然了这个路径并不是最优的走法。

此外还可以从不同角度使分类更细一些：

如下图所示的四种分类方式，分别对应着相应的主要算法：

	Model-free	Model-based	Policy based	Value based	Monte-carlo update	Temporal-difference update	On-policy	Off-policy
Qlearning	✓	✓		✓		✓		✓
Sarsa	✓	✓		✓		✓	✓	
Policy Gradients	✓	✓	✓		✓			
actor-critic			✓	✓				
升级版的 policy gradients						✓		
Monte-carlo learning					✓			
sarsa lambda							✓	
Deep-Q-Network								✓

**Model-free**：不尝试去理解环境，环境给什么就是什么，一步一步等待真实世界的反馈，再根据反馈采取下一步行动。

**Model-based**：先理解真实世界是怎样的，并建立一个模型来模拟现实世界的反馈，通过想象来预判断接下来将要发生的所有情况，然后选择这些想象情况中最好的那种，并依据这种情况来采取下一步的策略。它比 Model-free 多出了一个虚拟环境，还有想象力。

**Policy based**：通过感官分析所处的环境，直接输出下一步要采取的各种动作的概率，然后根据概率采取行动。

**Value based**：输出的是所有动作的价值，根据最高价值来选动作，这类方法不能选取连续的动作。

**Monte-carlo update**：游戏开始后，要等待游戏结束，然后再总结这一回合中的所有转折点，再更新行为准则。

**Temporal-difference update**：在游戏进行中每一步都在更新，不用等待游戏的结束，这样就能边玩边学习了。

**On-policy**：必须本人在场，并且一定是本人边玩边学习。

**Off-policy**：可以选择自己玩，也可以选择看着别人玩，通过看别人玩来学习别人的行为准则。

主要算法有下面几种，今天先只是简述：



热点  
专题  
反馈  
建议

在本站内搜索:

请输入关键词

搜索

policy derived from  $Q$  (e.g.,  $\epsilon$ -greedy)  
 f episode):

Take action  $A$ , observe  $R, S'$

Choose  $A'$  from  $S'$  using policy derived from  $Q$  (e.g.,  $\epsilon$ -greedy)

$$Q(S, A) \leftarrow Q(S, A) + \alpha[R + \gamma Q(S', A') - Q(S, A)]$$

$$S \leftarrow S'; A \leftarrow A';$$

until  $S$  is terminal

$Q$  为动作效用函数 (action-utility function)，用于评价在特定状态下采取某个动作的优劣，可以将之理解为智能体 (Agent) 的大脑。

SARSA 利用马尔科夫性质，只利用了下一步信息，让系统按照策略指引进行探索，在探索每一步都进行状态价值的更新，更新公式如下所示：

$$q(s, a) = q(s, a) + \alpha(r + \gamma q(s', a') - q(s, a))$$

$s$  为当前状态， $a$  是当前采取的动作， $s'$  为下一步状态， $a'$  是下一个状态采取的动作， $r$  是系统获得的奖励， $\alpha$  是学习率， $\gamma$  是衰减因子。

## 2. Q learning



在本站内搜索:

请输入关键词

搜索

episode):

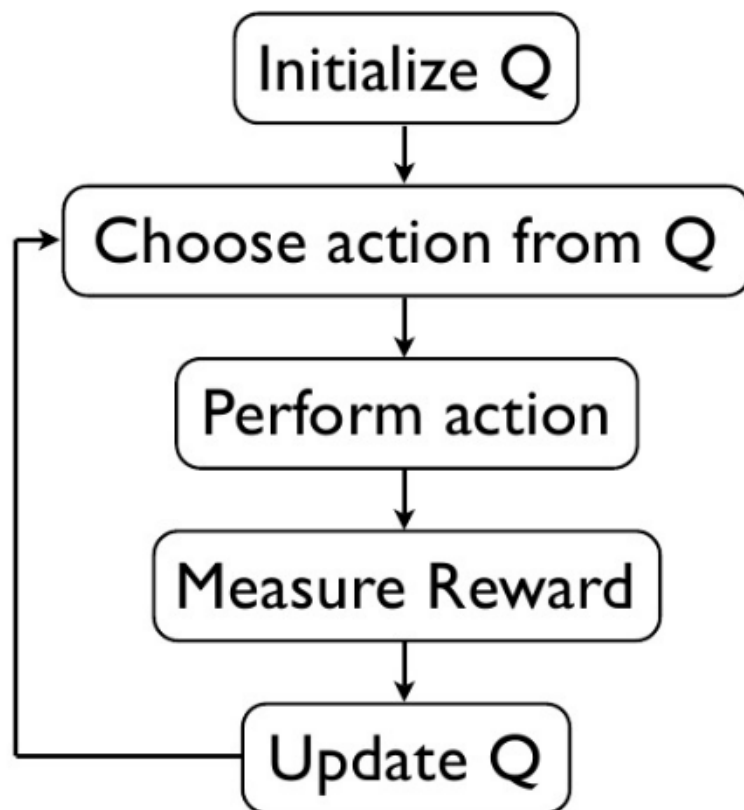
ing policy derived from  $Q$  (e.g.,  $\epsilon$ -greedy)

Take action  $A$ , observe  $R, S'$

$$Q(S, A) \leftarrow Q(S, A) + \alpha[R + \gamma \max_a Q(S', a) - Q(S, A)]$$

$S \leftarrow S'$ ;

until  $S$  is terminal



在本站内搜索:

请输入关键词

搜索

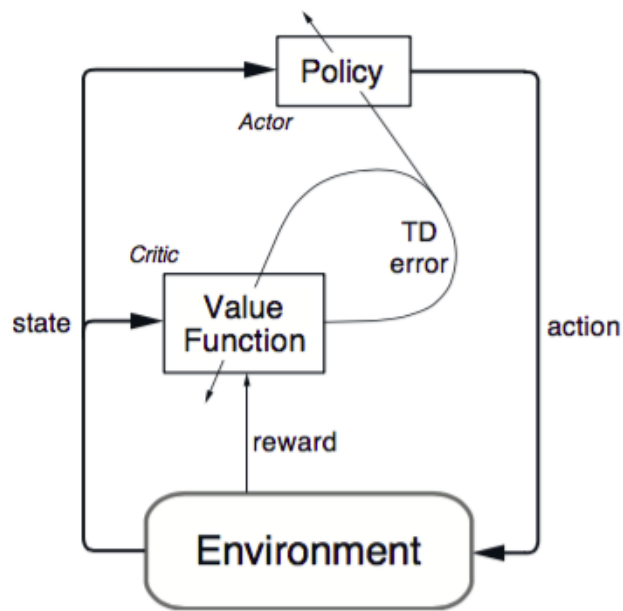
按照策略指引进行探索，在探索每一步都进行状态价值的更新。关键在于 Q

g 的更新公式如下：

$$q(s,a) \leftarrow r + \max_{a'} \{ \gamma q(s',a') \} - q(s,a)$$

在 Actor-Critic 模型中，Actor 负责在系统探索环境，生成一个从起始状态到终止状态的状态-动作-奖励序列， $s_1, a_1, r_1, \dots, s_T, a_T, r_T$ ，在第  $t$  时刻，我们让  $g_t = r_t + \gamma V(s_{t+1}) - V(s_t)$  等于  $q(s_t, a_t)$ ，从而求解策略梯度优化问题。

#### 4. Actor-Critic



算法分为两个部分：Actor 和 Critic。Actor 更新策略，Critic 更新价值。Critic 就可以用之前介绍的 SARSA 或者 Q Learning 算法。

#### 5 Monte-carlo learning

在本站内搜索:

请输入关键词

搜索

$s_1, a_1, r_1, \dots, s_k, a_k, r_k \in \pi$

其衰减奖励:

$+ \gamma r_{t+1} + \dots + \gamma^{k-t} r_k$

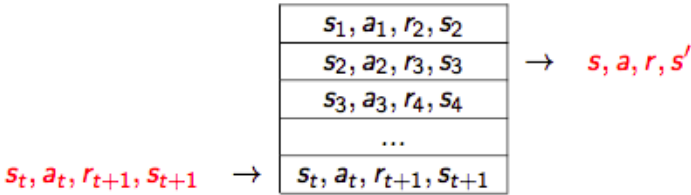
$i) = S(s) + g_s$

$N(s) = N(s) + 1$

$v(s) = \frac{S(s)}{N(s)}$

6. Deep-Q-Network

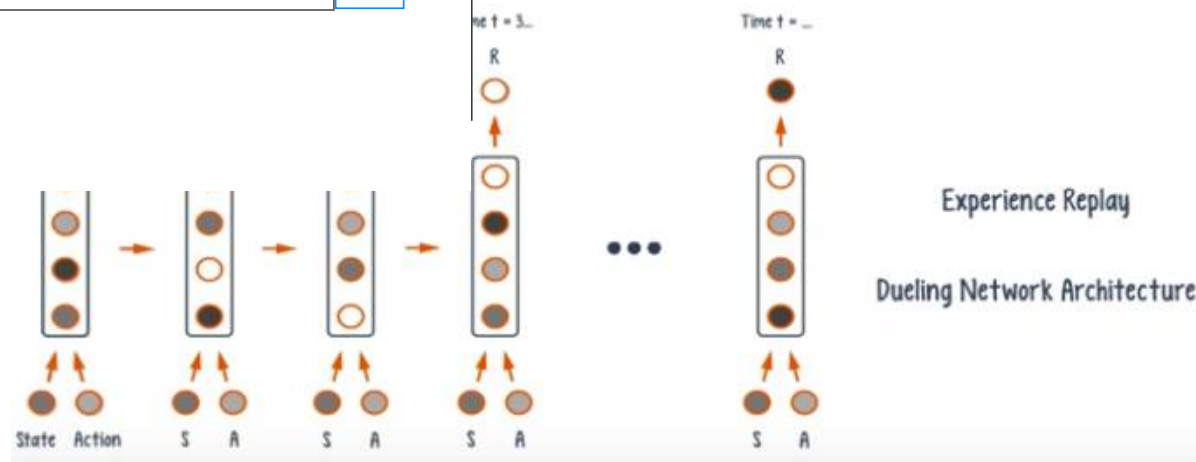
DQN 算法的主要做法是 Experience Replay，将系统探索环境得到的数据储存起来，然后随机采样样本更新深度神经网络的参数。它也是在每个 action 和 environment state 下达到最大回报，不同的是加了一些改进，加入了经验回放和决斗网络架构。



在本站内搜索:

请输入关键词

搜索



#### 4

#### 应用举例

强化学习有很多应用，除了无人驾驶，AlphaGo，玩游戏之外，还有下面这些工程中实用的例子：

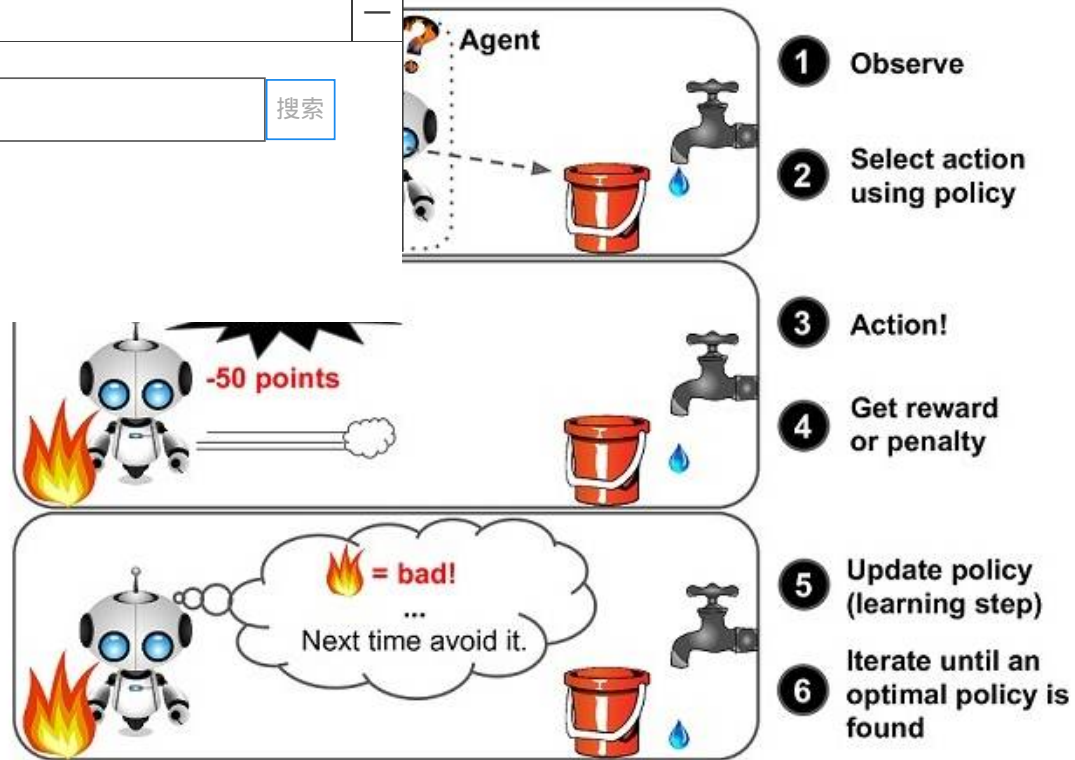
### 1. Manufacturing

例如一家日本公司 Fanuc，工厂机器人在拿起一个物体时，会捕捉这个过程的视频，记住它每次操作的行动，操作成功还是失败了，积累经验，下一次可以更快更准地采取行动。

在本站内搜索:

请输入关键词

搜索



## 2. Inventory Management

在库存管理中，因为库存量大，库存需求波动较大，库存补货速度缓慢等阻碍使得管理是个比较难的问题，可以通过建立强化学习算法来减少库存周转时间，提高空间利用率。

## 3. Dynamic pricing

强化学习中的 Q-learning 可以用来处理动态定价问题。

## 4. Customer Delivery

在本站内搜索:

请输入关键词

搜索

需求的

同时降低车队总成本。通过 multi-agents 系统和 Q-learning，可以降

行为，定制产品和服务以满足客户的个性化需求。

## 6. Ad Serving

例如算法 LinUCB（属于强化学习算法 bandit 的一种算法），会尝试投放更广范围的广告，尽管过去还没有被浏览很多，能够更好地估计真实的点击率。

再如双 11 推荐场景中，阿里巴巴使用了深度强化学习与自适应在线学习，通过持续机器学习和模型优化建立决策引擎，对海量用户行为以及百亿级商品特征进行实时分析，帮助每一个用户迅速发现宝贝，提高人和商品的配对效率。还有，利用强化学习将手机用户点击率提升了 10-20%。

## 7. Financial Investment Decisions

例如这家公司 Pit.ai，应用强化学习来评价交易策略，可以帮助用户建立交易策略，并帮助他们实现其投资目标。

## 8. Medical Industry

动态治疗方案（DTR）是医学研究的一个主题，是为了给患者找到有效的治疗方法。例如癌症这种需要长期施药的治疗，强化学习算法可以将患者的各种临床指标作为输入 来制定治疗策略。

5

参考资源

上面简单地介绍了强化学习的概念，区别，主要算法，下面是一些学习资源，供参考：



在本站内搜索:

请输入关键词

搜索

— Learning ,

ent Learning: An Introduction 被引用2万多次

bk.pdf

豆人”游戏：Berkeley Pac-Man Project (CS188 Intro to AI)

car Tracking (CS221 AI: Principles and Techniques)

15 CS 294 Deep Reinforcement Learning, Fall 2015。

David Silver强化学习

<http://www0.cs.ucl.ac.uk/staff/d.silver/web/Teaching.html>

相关文章：

TensorFlow-11-策略网络：用 Tensorflow 创建一个基于策略网络的 Agent 来解决 CartPole 问题。

<http://www.jianshu.com/p/14625de78455>

强化学习是什么：简单图解了 DQN

<http://www.jianshu.com/p/2100cc577a46>

参考：

<https://www.marutitech.com/businesses-reinforcement-learning/>

<https://www.analyticsvidhya.com/blog/2017/01/introduction-to-reinforcement-learning-implementation/>

<https://morvanzhou.github.io/tutorials/machine-learning/ML-intro/4-02-RL-methods/>

<https://www.zhihu.com/question/41775291>

<http://www.algorithmdog.com/reinforcement-learning-model-free-learning>



**2017中国人工智能大会 | 7月22-23日 杭州**

由中国人工智能学会、蚂蚁金服主办

CSDN承办

是最专业的年度技术盛宴：

- 40位以上实力讲师

- 8场权威专家主题报告

在本站内搜索:

请输入关键词

搜索

- 4场开放式专题研讨会

- 超过100家媒体报道

- 1000位技术精英和专业人士参会

, 到官网报名：<http://ccai.ccai.cn/>



☆ 收藏

👍 顶(0)

👎 踩(0)

分享到:

## 相关文章

合作活动 | CiGA开发者大会嘉宾揭露 2017-10-01

后续报道 | 中加国际电影节中的VR 2017-09-30

干货 | 虚拟（增强）现实白皮书（2017年）免费下载 2017-09-29

9月末的热潮，TGS中的VR 2017-09-25

活动推荐 | WePlay想要承包你2017年10月的最后几天 2017-09-19

教程 | 360°全景图的绘制分析 2017-09-14

最新虚幻引擎4 (UE4) AR游戏大作The Machines 登上苹果大会演示舞台 2017-09-13

中立 专业 共享 双赢

✉: [online@weiot.net](mailto:online@weiot.net) 岳先生

关于威腾网 什么是威腾网？

[关于威腾网](#) | [威腾网大事记](#) | [联系威腾网](#) | [商务合作](#) | [公司招聘](#) | [寻求报道](#)



微信扫一扫  
关注热点  
专题  
反馈  
建议

在本站内搜索:

请输入关键词

搜索

网站地图

快速找到你想要的

智能硬件展会现场 | 智能硬件展商动态 | 智能清洁设备 | 现场佳丽 | 活动 | 智能照明设备 | 智能眼镜设备 | 智能安防设备 | 智能语音设备 | 智能设备-自行车 | 智能服饰设备 | 融资-互联网创业项目 | 专访-互联网创业项目 | 心得-互联网创业项目 | 智能设备-独轮车 | 智能设备-平衡车 | 手机频道 | 智能手表设备 | 智能手环设备 | O2O | 威腾专栏 | 智能设备-无人驾驶 | 智能设备-扭扭车 | 智能空气净化设备 | 智能设备-电动车

最新动态新闻 | 原创汇集 | 产品地图 | TAG标签 | 网站地图 | 威腾网-智能硬件互动平台 ( 京ICP备-09048584-7 )

站长统计

http://www.weiot.net/article-107610.html

热点  
专题  
反馈  
建议

17/17