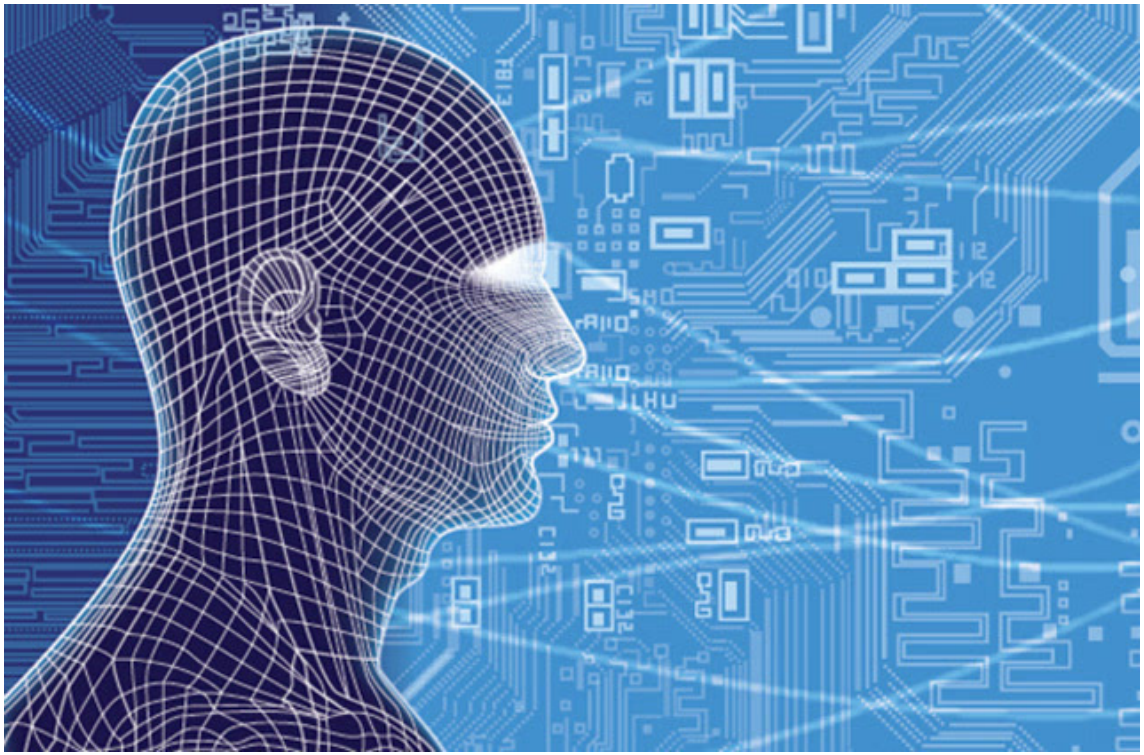
36大数据

2736 文章 | 1482万 总阅读

查看TA的文章>

深度增强学习暑期学校 PPT讲解

2016-09-13 09:41



译者：Flood Sung ·

这是2016年机器学习暑期学校的课程，OpenAI的John Schulman做了4次Deep Reinforcement Learning的讲座，讲座的具体链接如下：

Lecture 1: intro, derivative free optimizationLecture 2: score function gradient estimation and policy gradientsLecture 3: actor critic methodsLecture 4: trust region and natural gradient methods, open problems

本文基于其讲座的ppt及视频，对讲座的内容按照ppt的顺序做一些翻译，分析和理解。

课程的实验材料在这里：https://goo.gl/5wsgbJ

Learning: Pong from Pixels上说，Policy Gradient是目前更侧重的深度增强学习的方法。那么本身John Schulman最重要的工作就是TPRO，一个基于Policy Gradient的改进算法。这个算法目前在OpenAI Gym的表现非常好。

首先是整体介绍部分

什么是增强学习?

https://zhuanlan.zhihu.com/p/21262246?refer=intelligentunit

增强学习是机器学习的一个分支，不同于监督学习和无监督学习，它主要侧重于如何输出序列动作。也就是增强学习关注决策与控制。那么最基本的问题描述就是一个智能体(Agent)在未知的环境中，如何根据观察(Observation)和反馈(Reward)来调整自己行为(Action)，从而使累积的反馈(Reward)最大化。注意这里说的是累积的反馈，而不是单一行为之后的反馈。这很好理解。举炒股票为例，我们一个买入卖出动作之后，当天的收益也就是Reward，但是我们看重的是比如一周之后的收益是怎样的。这就需要累积每一天的Reward。我们的目标就是希望未来的最终收益最大化。这就是增强学习要研究的问题。

Motor Control and Robotics

Robotics:

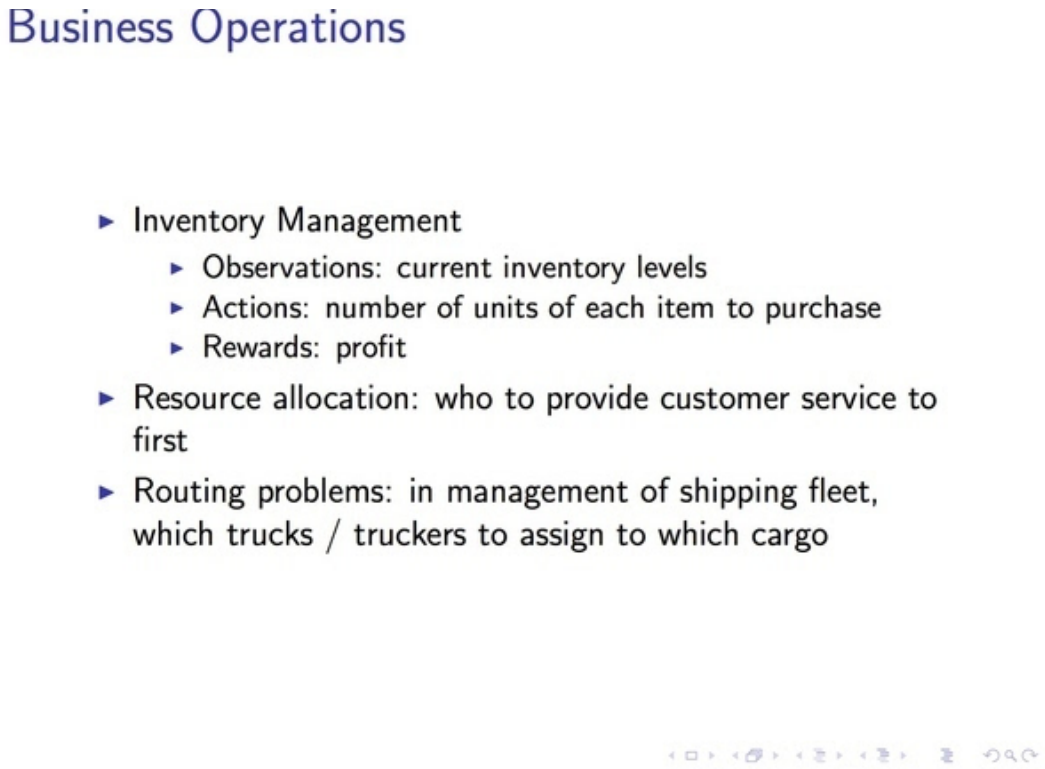
- Observations: camera images, joint angles
- Actions: joint torques
- Rewards: stay balanced, navigate to target locations, serve and protect humans

因为增强学习的重要在于解决决策与控制，因此机器人的控制显然是首当其冲了。那么怎么用增强学习的模型来描述机器人控制的问题呢

观察Observations：摄像头的图像，机械臂关节的角度，，

动作Actions：就是每个电机的输出扭矩，如果是机械臂那么就是对应每个关节的扭矩啦

反馈Rewards：有多种形式，主要就是看是什么任务了。比如保持平衡，移动到某个特定位置，或者服务及保护人类等等



增强学习在商业上的问题

既然涉及到决策，那么商业决策是显然值得研究的。

库存管理。对于一个超市来说，如何管理库存是很需要研究的，最好的情况就是所有的货物都不会积压，都正常卖出，不出现损失。这就需要考虑每个时间段如何根据当前库存购买新的货物的问题。在这个问题上，观察就是当前的库存情况，动作就是购买的每一种货物的数量，反馈就是最后的收益。

收益最大化呢?观察就是当前的投资及资金分配情况，动作就是分配资金，反馈就是收益。

运输问题。比如滴滴的算法大赛问题，就是希望能够估计出不同地区的顾客需求，从而更好的调配出租车的位置。比赛的问题只是一个预测问题。但是如何直接调配出租车那么就是一个增强学习问题了。

游戏上的应用

因为AlphaGo的出现，大家对此就比较了解了。很多游戏都是RL的问题。

围棋。完全的信息，确定性决策(就是每次只有一个确定性的选择)。Backgammon, 西洋双陆棋。

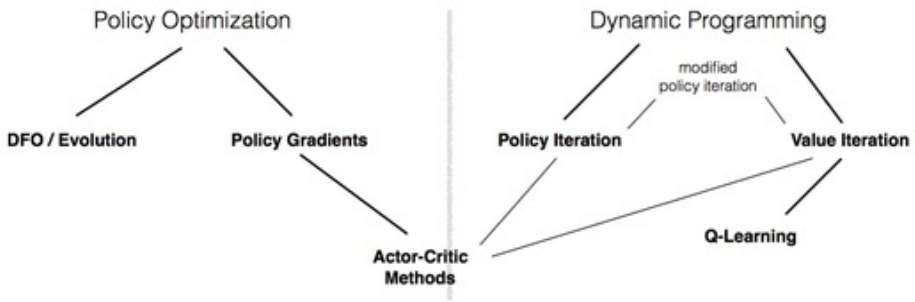
这种游戏也是完全信息的，只不过需要掷骰子，因此存在随机性Stochastic。这个游戏有个著名的算法叫TD-Gammon，很早了，用了简单的神经网络实现。也就是深度增强学习其实也很有历史啦。

Stragego 西洋陆军棋

这种棋和我们的军棋差不多，也是下暗棋，所以是不完全信息博弈，但它的每一步是确定性的。

Poker 扑克，特别是德州扑克

对于这种游戏，显然是不完全信息博弈，并且每一个回合都存在随机性，大家都不知道发的什么牌。



增强学习的解决方法

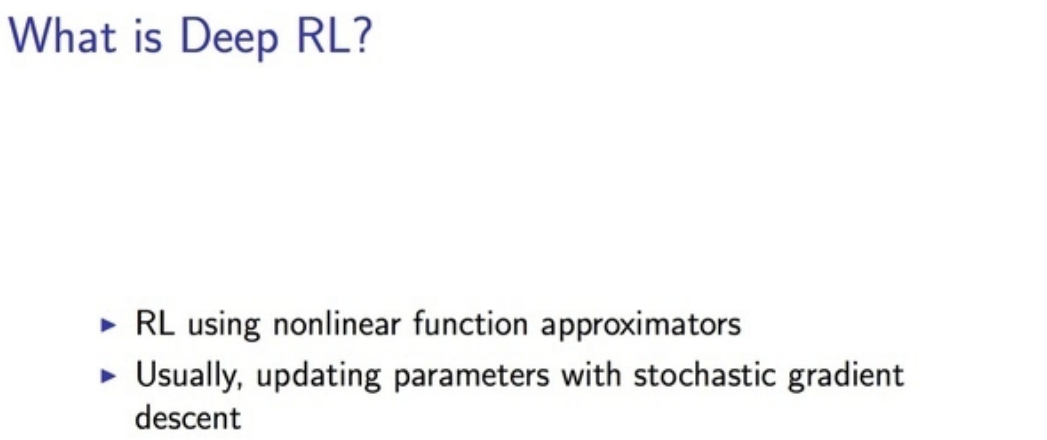
那么从上面的ppt中可以看到基于RL的解决办法可以分成两大类：

1)策略优化Policy Optimization。就是直接优化策略（πs）。这种方法又分为两种：一种称为进化Evolution方法，就是不断调整策略的参数，选择更好的参数。另一种就是策略梯度Policy Gradient。通过计算策略的梯度方向，通过梯度下降的方式来优化策略。

2)动态规划Dynamic Programming。这个方法就是利用价值函数Value Function来实现曲线救国。有策略迭代Policy Iteration和值迭代Value Iteration。最有名的就是Q-Learning了，

除了上面的两种方法，第三种就综合上面两种方法形成所谓的Actor-Critic方法，也就是行动-监督方法。

其实，对于增强学习的问题，还有一个方法就是只使用监督学习，只是这种方法需要有现有的方法提供监督学习的样本。UC Berkerley其实这方面的成果是最多的，Guided Policy Search就属于这类监督学习的方法。如果你不是很明白，但是了解AlphaGo。AlphaGo一开始的监督学习就是利用人类棋谱做训练，这也可以达到不错的效果。



深度增强学习就是将增强学习中的策略Policy或者价值函数Value Function用非线性函数来近似，然后显然用深度神经网络是最好的非线性函数。那么通常就是使用随机梯度下降的方式来更新里面的参数。

这里将Deep RL研究的问题与人类大脑皮层的功能做对比。Deep RL类比为人类的前半部分大脑皮层，用于处理决策与控制。而大脑的后半部分则处理感知类比计算机视觉处理的功能。从这里的类比可见Deep RL研究的重要性。

Markov Decision Processes

马尔科夫决策过程最基本的定义就是(S,A,P)，其中

S：状态空间A：动作空间P(r,s'|s,a)：状态转移的概率分布。也就是在当前s和动作a下，下一个时间片的状态s和反馈r会是怎样的，这通常用概率分布来表示。对于一些完全可观察的问题比如围棋，那么下一步的情况则是确定性的。

根据问题的设定常常会定义额外的对象：

u：初始的状态分布λ：衰减系数。就是反馈Reward一般根据时间作用越来越小，用该系数来表示。

篇章式设定

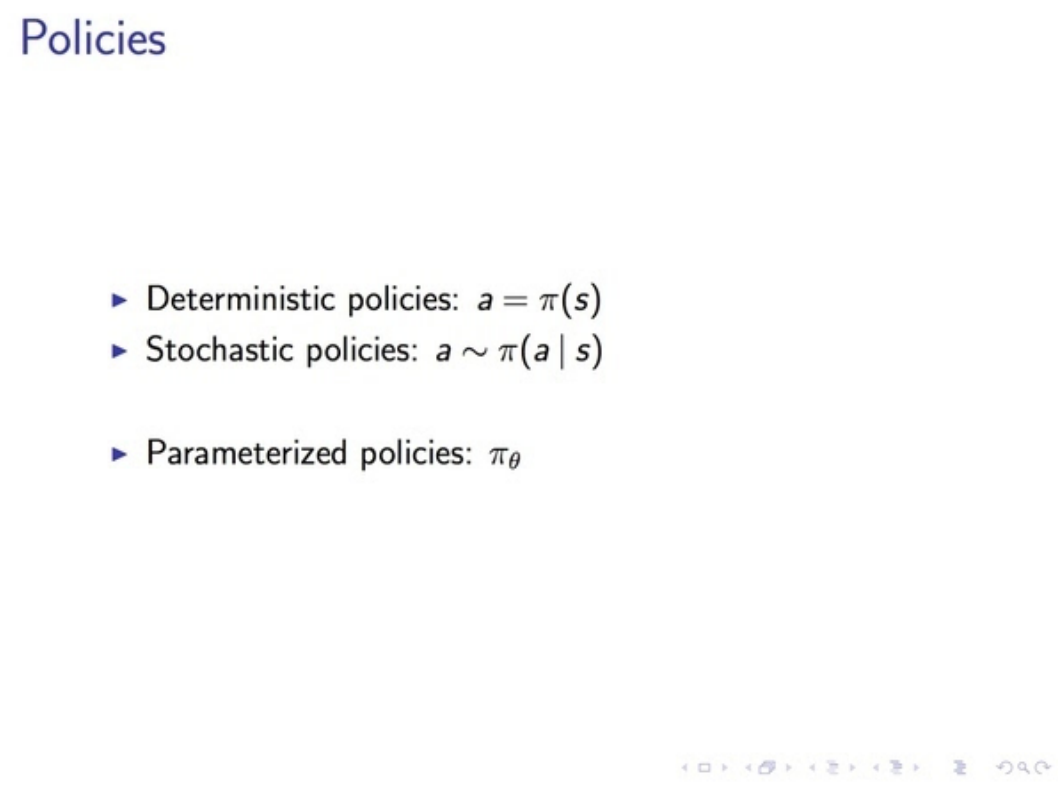
在很多RL问题中，问题有一个欲达成的目标，也就是一个任务的时间长度是有限制的，是可以结束的，对于这种可以结束的问题，就是每运行一次实验就称为一次episode，依旧是从初始状态开始，直到最终的结束状态到达。比如：

的士机器人到达目的地(结束状态是好的)侍者机器人完成一个移动(在有限时间内)移动机器人摔倒(结束状态是坏的)

还有其他的比如围棋就是到一盘棋下完，而玩Atari游戏就是到游戏结束。

那么这种问题的目标是比较好确定的，就是到任务结束时得到最大的累加反馈值。那么有没有一些问题是没有结束的呢?只能说可以有，但都可以转化为可以结束的情况。比如说机器

走到某个特定位置。



策略

策略有两种方式表示，这个我们在之前的专栏文章中也说明过：

一种就是确定性策略：a=π(s)，也就是有一个状态s，就对应一个动作a另一种就是随机分布：a~π(a|s)，也就是即使面对同一个状态s，也存在多种动作选择可能都是较优的，因此用概率来表示选择某一个动作的可能性。

那么对于篇章式设定的问题，目标是非常简单可以设定的，就是累加反馈的期望值，如上面的公式表示。

那么上面这个图可以更清楚的表示这个过程。大家要记住这一点。就是在MDP的框架下，时间可分割，因此就是分割成一个一个的时间片。因此可以形成{s0，a0，s1，r0，a1, s2,r1 ……..}

这样的时间序列数据。

参数化策略

就是用参数来表示一个策略，比如用一个线性方程来表示策略，那么线性方程的参数就影响了这个策略。如果用θ来表示参数，那么参数化的策略就如下表示：

确定性策略：a=π(s,θ)随机性策略：π(a|s,θ)

很多不是非常理解这个带参数的策略应该是如何构建，特别是使用神经网络。神经网络的输出可是分类Classification也可以是回归Regression，取决于动作是连续还是离散。比如Atari游戏，输出就是几个离散的动作，那么就可以使用分类的神经网络。输入的是图像，输出的是几个动作的概率，这就可以用softmax来输出。那么如果是连续的动作空间，比如机器人控制基本都是连续动作，那么可以使用regression回归神经网络。网络输出可以是高斯分布的平均值及对角协方差。关于这一点在后面的CEM算法中可以有清晰的理解。

接下来就介绍通过黑盒优化来求解增强学习问题

说到优化我们首先会想到梯度下降，但是其实有很多其他的优化的方法并不使用梯度下降。比如模拟退火，爬山法，遗传算法等等。那么这里我们就把策略Policy看做是一个黑盒，我们的目标很明确，就是让期望的反馈最大化。除了R之外，其他的信息我们都不管。

交叉熵方法Cross Entropy Method

一种进化算法。但是却取得了很不错的效果。ppt列出了较早的将CEM方法应用到增强学习的论文。用一种带噪声的交叉熵方法。

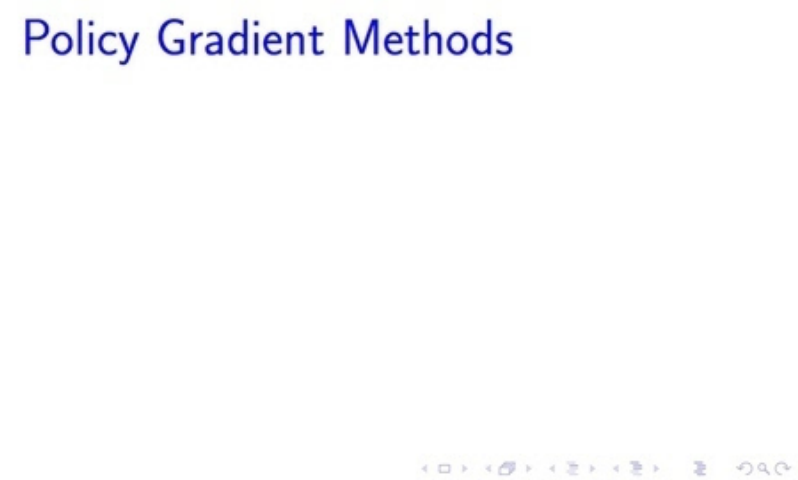
一个和交叉熵类似的方法叫做Covariance Matrix Adaption，简称CMA。在图形学上已经成为了一种标准算法。

交叉熵方法

这里具体介绍交叉熵的方法。可以用很简单的一句话来说明：就是每次根据当前的策略参数采样多组带噪声的新的参数，然后进行试验，选择反馈R最好的前几组参数，然后取平均作为新的参数。这种方法其实就是一种贪婪算法，通过纯随机的方式来寻找最佳的参数。按道理这种方法不应该和梯度下降的方法相提并论，但没想到CEM方法出奇的好。就是即使这个策略用巨大的神经网络来表示，也同样能够取得非常不错的效果。

有些知友可能看不明白上面的算法中的高斯分布是啥意思，其实就是选择中的几组参数的平均值就是高斯分布的平均值，而方差就是几组参数的方差。方差和平均值都用于之后再产生新的参数。

这边作者也没有深入的分析，只是提了几句相关的算法。



策略梯度方法

策略梯度方法

策略梯度方法的目标是和上面的CEM方法是一样的，就是最大化期望的反馈R。那么一个很直观的做法就是不断试验，采集各种状态动作数据，也就是所谓的轨迹trajectory。那么有

第一种是让好的轨迹出现可能性增大第二种是让好的动作出现可能性增大(actor-critic方法，GAE方法)第三种是让动作趋向于好的动作(DPG，SVG方法)

那么这三种有什么不一样呢?第一种就是要求一系列的动作都好，比如某一次实验比较好，那么这次实验的每一个动作的可能性都增大。第二种就是只针对单一的动作。如果某一个动作的评估比较好，那么让其出现可能性增大。第三种的目标是让动作趋向好的动作，也就是有一个好的动作的目标，比如DPG使用Q值，目标就是让动作的Q值增大。那就得到好的动作了。

第一部分先到这里。

原文>>>

End.

转载请注明来自36大数据（36dsj.com)：36大数据» 深度增强学习暑期学校 PPT讲解

返回搜狐，查看更多

24小时热文

1 叙利亚
卧场面

2 由钻石
55E

3 一文看
机、音
化


谷歌Pix
835+4G


浅谈：

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
谷歌
Pixel
引力波
iOS
VR
OPPO
魅族
天猫

推荐阅读

谷歌发布最强拍照手机Pixel 2 秒杀苹果iPhone8




搜狐科技视界 · 今天 06:38 7

“向上帝借眼睛”，成就无数科学家的冷冻电镜终于征服2017年诺贝尔化学奖！

24小时热文

1 叙利亚
卧场面

2 由钻石
55E

3 一文看
机、音
化


谷歌Pix
835+4G


浅谈：

DeepTech深科技 · 昨天 19:03 1

DeepMind成立了一个新的秘密小组，还是研究AI与道德

量子位 · 今天 05:04

可靠消息称，Waymo很可能于今年底在凤凰城推出全自动驾驶出租车

DeepTech深科技 · 昨天 19:03

谷歌发布Google Buds耳机：可以实时翻译

IT之家 · 今天 08:03

哈佛、MIT研发新型纹身墨水，能根据健康情况变色

DeepTech深科技 · 昨天 19:03

连失两项诺奖背后，CRISPR基因编辑技术让全人类恐惧改造自身了吗？

24小时热文

1 叙利亚
卧场面

2 由钻石
55E

3 一文看
机、音
化


谷歌Pix
835+4G


浅谈：




造就 · 昨天 19:11

IEEE：全球超一半大公司正在研究块链，但是你需要区块链吗？

雷锋网 · 昨天 16:54

经典老茶，10年陈皮普洱198元/桶，卖爆了

广告 · 今天 10:50

这18句话决定了埃隆·马斯克为什么能成功




资本实验室 · 昨天 16:28

ofo和摩拜又传合并 难盈利始终是痛点




威锋网 · 昨天 14:53 2

24小时热文

1 叙利亚

卧场面

2 由钻石

55E

3 一文看

机、音

化

谷歌Pix

835+4G

浅谈：

PingWest品玩 · 今天 09:07

史上最大规模黑客事件，30亿账号信息泄露，而4年后真相还没出现

PingWest品玩 · 今天 09:33

红茶卖疯了，夏日办公必备红茶，陈皮普洱198元一桶，还送原装木桶！

广告 · 今天 10:50

苹果也得为它打工！iPhone X大卖，这家公司将入账143亿美元

每日经济新闻 · 昨天 20:01 5

假期荐书 | 2017年出版的最好的商业书籍

36氪 · 今天 07:21

24小时热文

1 叙利亚

2 由钻石

3 一文看

谷歌Pix 835+4G

浅谈：

小黄车和小橙车真要合并了？摩拜这句回应，让ofo扎心了

每日经济新闻 · 昨天 13:14 8

冷冻电镜牛X在哪里？5位科学家说：它彻底“消灭”了结构生物学

果壳网 · 昨天 21:47 1

餐桌上缺了这一口，那你一定过了个假中秋 | 钛空舱

钛媒体APP · 今天 10:00

FinTech从妄想变成现实，谁是控制论的下一个预言？|未来十年变革

全天候科技 · 今天 08:47

中秋夜，你朋友圈里的月亮都不太圆？你没瞎！后天才是满月！

24小时热文

1 叙利亚

卧场面

2 由钻石

55E

3 一文看

机、音

化

谷歌Pix

835+4G

浅谈：

果壳网 · 昨天 23:28 1

全民“公敌”五仁月饼犯了何罪

钛媒体APP · 昨天 18:23 27

加载更多

24小时热文

1 叙利亚
卧场面

2 由钻石
55E

3 一文看
机、音
化

谷歌Pix
835+4G

浅谈：