



SEARCH

2017年十月 (3)
最新文章列表

我爱机器学习

机器学习干货站



MIND如何用REINFORCEMENT LEARNING玩游戏

NIPS 2016—Daily
Highlights

我爱机器学习(52ml.net) 2016年12月13日

NIPS 2016—Day 1 Highlights

NIPS 2016—Day 2 Highlights: ...
欢迎关注我爱机器学习微信公众
号:



2016深度
十大指数级
12月12日
学习零基础



2016年12月11日

公告栏

Yoshua



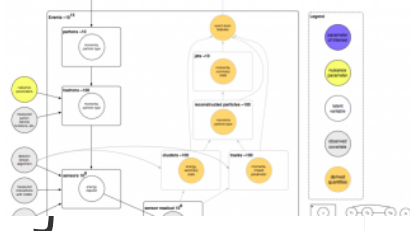
期待您的加入

2016年12月11日

2017年十月 (3)

最新文章列表

结构之美



NIPS 2016—Daily Highlights

我爱机器学习(52ml.net)

13 璵

NIPS 2016—Day 1 Highlights

NIPS 2016—Day 2 Highlights:

欢迎关注我爱机器学习微信公众平台

号:



2016年度
十大指数级
12月12日
学习零基础

2016年12月11日

公告栏

Yoshua

期待您的加入

2010/07/04

, 非Deep learning和Reinforcement learning莫属 (以下分别简

在实际应用中表现的很酷，在机器学习理论中也有不俗的表

兩者之精髓，在Stella模擬機上讓機器自己玩了7個Atari 2600

美洲，走向世界，超越了物种的局限。不仅战胜了其他机器

超越了人类游戏专家。噢，忘记说了，Atari 2600是80年代风

现在肯定不会喜欢了。长成什么样子？玩玩当下最火的flappy



SEARCH

走……由Stella倾情打造了模拟机，甚至还有为学术界专门贡献的Arcade Learning

2017年十月 (3)

最新文章列表

目的呢，当然是赢得游戏，分数多多益善。



很酷的科学家的，肯定不会亲手玩游戏咯，当然一方面也是玩游戏，先得想清楚人类是怎么玩游戏的：

在初始时刻。然后，游戏场景开始变换，玩家眼睛捕捉到画面信号传递回脑皮层进行处理。

NIPS 2016—Daily Highlights

我爱机器学习(52ml.net) 2016年12月13日

觉信号转换为游戏的语义信息，通过经验指导，将语义信息与应景进入下一帧，玩家得到一定的回报，如越过关隘，或者吃到

NIPS 2016—Day 1 Highlights [到游戏结束。

NIPS 2016—Day 2 Highlights:... 欢迎关注我爱机器学习微信公众号:



!016深度
十大指数级
:12月12日
学习零基础



2016年12月11日

公告栏

Yoshua



期待您的加入

2016年12月11日



Q SEARCH

semantics unknown

2017年十月 (3)

最新文章列表

image formation

deep encoder

feature vector

semantics unknown

reward

function approximator

q-values

action selection

action

agent

NIPS 2016—Daily Highlights

我爱机器学习(52ml.net) 2016年12月13日

在游戏中的那些事情是玩家所不用考虑的，玩家能够覆盖的。即输入视觉信号，输出手指动作。而手指动作到下一帧场游戏内部的过程。

NIPS 2016—Day 1 Highlights

NIPS 2016—Day 2 Highlights

欢迎关注我爱机器学习微信公众号

既...，并分解出实际需要玩家的部分内容，下一步就是让机器替代人类玩家了。

2016深度学习十大指数级

12月12日

学习到玩游戏的知识，即经验（什么场景下需要什么操作）

理解（降维，高层特征抽取）

学习零基础

视觉特征，选择合理的经验（动作）

2016年12月11日

玩家手动的部分

公告栏

Yoshua

期待您的加入

2017年12月11日

2016深度学习十大指数级

12月12日

学习到玩游戏的知识，即经验（什么场景下需要什么操作）

理解（降维，高层特征抽取）

学习零基础

视觉特征，选择合理的经验（动作）

2016年12月11日

玩家手动的部分

公告栏

Yoshua

期待您的加入

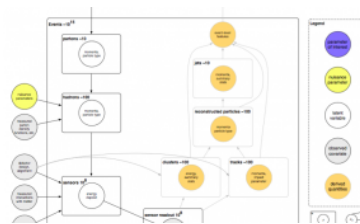
2017年12月11日



看agent游戏玩的到底如何。总结涉及对RL的升华。

2017年十月 (3)

最新文章列表 Reinforcement Learning



是
NIPS 2016—Daily Highlights

我爱机器学习(52ml.net) 2016年12月13日

NIPS 2016—Day 1 Highlights

NIPS 2016—Day 2 Highlights: ...
欢迎关注我爱机器学习微信公众号:



2016深度
十大指数级
12月12日
学习零基础



公告栏

2016年12月11日

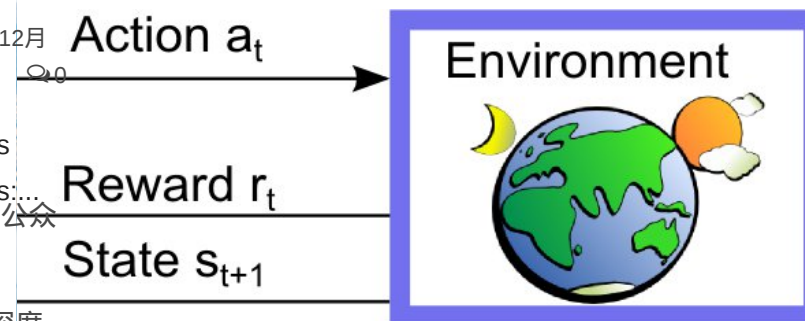
Yoshua



期待您的加入

2016年12月11日

程。传统的机器学习中的supervised learning就是给定一个supervisor，学习一个好的函数，来对未知数据作出很好的决策。是什么，即一开始不知道什么是“好”的结果，所以RL不是监督学习，这个回报函数决定当前状态得到什么样的结果（“好”还是一个马尔科夫决策过程。最终的目的是决策过程中整体的回报



Reinforcement Learning Setup

一个状态集合，其中每一个元素都代表一个状态。在游戏的场景时刻采集到的视觉信号。

包含所有合法操作。如flappy bird中点击一下屏幕，temple run中



SEARCH

- 回报函数 R 。 R 是一个映射，跟状态转移概率 P 有点联系， R 说明的是，在当前状态 s_t 将会得到怎样的回报。需要注意的是，这里的回报不一定是即时回报，如棋牌游戏中，棋子移动一次可能会立刻吃掉对方的棋子，也可能在好多步

→ 产生作用



NIPS 2016—Daily Highlights

我爱机器学习(52ml.net) 2016年12月13日

NIPS 2016—Day 1 Highlights

NIPS 2016—Day 2 Highlights: ...
欢迎关注我爱机器学习微信公众

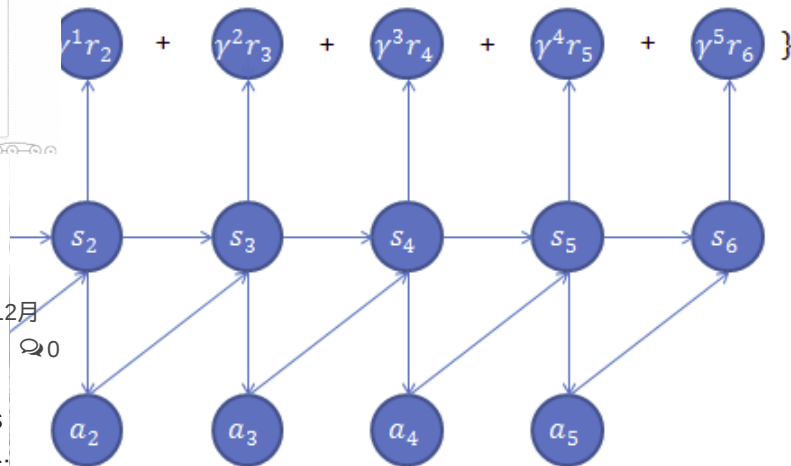
号:



公告栏

Yoshua
期待您的加入

2016年12月11日



ky。

!016深度

十大指数

12月12日

学习零基础

学习零基础

学习零基础

学习零基础

学习零基础

学习零基础

学习零基础

学习零基础

学习零基础

过程，意即整个决策的过程都是有概率特性的，每一步的选择都是一个概率分布中采样出来的结果。因此，整个回报函数是一种积分。依据贝叶斯定理，开局时刻不确定性是最大的，开局基于先验知识。随着游戏的不断进行接近终点，局势会逐渐晴朗，深蓝对战国际象棋大师卡斯帕罗夫的时候，开局就是一些经典规则，多考虑战略优势，局势逐渐明朗，因此这时候一般会出现正常就是一些战术上的考量，如何更快的将军等。类似地，在积分中，每一步的回报都会乘上一个decay量，即回报随



Q SEARCH

2017年十月 (3) 最新文章列表



NIPS 2016—Daily Highlights

我爱机器学习(52ml.net) 2016年12月13日

Ex NIPS 2016—Day 1 Highlights

的 NIPS 2016—Day 2 Highlights: ...
欢迎关注我爱机器学习微信公众
号



2016深度
十大指数级
2016年12月12日
学习零基础



2016年12月11日

公告栏

Yoshua



期待您的加入

2016年12月11日

了，剩下的事情就是寻找一种最优策略（policy）。所谓策略，我们的目的是，找到一种最优策略，使得遵循这种策略进行的决策过程，得到的全局回报最大。所以，RL的本质就是在这些信号下找到这个最佳策

理论基石就来自Bellman公式。Bellman公式告诉我们，在一个解的路径是最优路径，那么其中的每个分片都是当前解合起来就是全局最优解。回报函数的最大化就服从

的性质，表示着我们可以不断迭代求解问题。旅行商问题就不是NP-hard问题。

方面，两方面相互交织，最终得到结果。这是一种典型的算法的过程。EM算法在机器学习中是相当经典的算法，大量的方法。

求解RL的示例：



SEARCH

while($\Delta > \theta$)

2017年十月 (3)

最新文章列表

$$V(s) := \sum_{s' \in S} T(s, \pi(s), s')(R(s, \pi(s), s') + \lambda V(s'))$$

$$V(s) := \sum_{s' \in S} T(s, \pi(s), s')(R(s, \pi(s), s') + \lambda V(s'))$$



NIPS 2016—Daily Highlights

该: 我爱机器学习(52ml.net) 2016年12月13日

优: NIPS 2016—Day 1 Highlights

NIPS 2016—Day 2 Highlights

欢迎关注我爱机器学习微信公众号

号:



2016年12月11日

公告栏

Yoshua
期待您的加入

$$\sum_{s' \in S} T(s, a, s')(R(s, a, s') + \lambda V(s'))$$

Policy Evaluation

commit 2013, 由Adobe的Nedim Lipka介绍了RL在市场策略(网
这里抛开具体的应用语义, 以及分布式算法, 来简单分析RL
过程。

函数以当前状态s为参数, 返回一个动作a, 这个动作是一个概
率: 态s下, 转移到任意另外一个状态的概率是多少。假设我们有三

可能是这个样子的:

| | | |
|--------|-----|-----|
| 12月12日 | 2 | 3 |
| 学习零基础 | 0.1 | 0.6 |

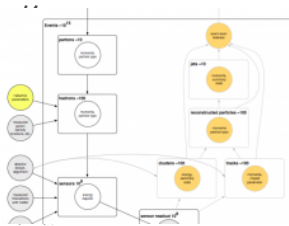


SEARCH

即时 (immediate) 回报函数，即从状态s出发，经过a这个动作的作用，走到这个状态获

2017年十月 (3) 用户在某个页面上浏览，点了一个广告，到了广告商的页面，广

最新文章列表



下面的转移概率，表示当前动作下的即时回报函数，是从s期望价值。

是和都是未知的，而这两个量是相互纠缠的，计算需要最行积分。所以这是个典型的Expectation-Maximization算

法 NIPS 2016—Daily Highlights

EM算法中的Expectation，第二部分就是EM算法的

我爱机器学习(52ml.net) 2016年12月13日

迭代呢？那是因为大家记得随机游走，都不是游走一次就能结

束！ NIPS 2016—Day 1 Highlights

稳定状态，需要多次迭代才可以。这就类似于Gibbs sampling

算， NIPS 2016—Day 2 Highlights...

收敛。这里也是，计算Expectation需要让整体的网络达到稳定

欢迎关注我爱机器学习微信公众号

看前后两次迭代差距是否足够小，因此判断是否收敛。

号:



!016深度
十大指数级
:12月12日
学习零基础



2016年12月11日

公告栏

Yoshua



期待您的加入

2017年十月 (3)

最新文章列表

网络相依，类似与随机游走)

¹²是一个supervised random walk (可以参考斯坦福大学Jure

ervised Random Walk)。传统的random walk是按照固定的转

S, RL就是在随机游走的每一步, 都选择一个能使回报函数最大

当前状态下最好的action。而RL游走的这个网络，是由状态S为

转移概率P为边权重的有向无环图（DAG）。状态转移概率P

深度在这个网络中的步进，不断变的更加正确，符合现实世界的

十大指标级混沌的网络状态。

2日

是一次action得到的payoff，Return是一序列reward的函数，如

基础 而下面的value function是要学习的函数 Value

来预测在给定状态（或者给

年12月11日

能表现多好。有多好，表明的是往这点的expected reward，即

大期望收益。Approximator, 关键是泛化能力, 在有限的状态-

扩展到全部的状态和动作上？使用动态规划这种“查找表”的



2017年十月 (3) ng in RL

最新文章列表

Deep在何处？换句话说，因为DL参与的RL与传统的RL有何不同，从而要引入DL？我们



戏:

NIPS 2016—Daily Highlights

我爱机器学习(52ml.net) 2016年12月13日

类型:

NIPS 2016—Day 1 Highlights

NIPS 2016—Day 2 Highlights:...

欢迎关注我爱机器学习微信公众

号:



2016深度学习界十大指数级

12月12日

学习零基础



2016年12月11日

公告栏

Yoshua



期待您的加入

2016年12月11日

目的是状态。而实际上，很多时候状态是连续的、复杂的、14个状态就可以了。实际上，假设我们有128*128的画面，的，即有 $2^{(128*128)}$ 中可能存在的状态，这个数字是数字！游戏画面连续存在，就算按照每秒30帧来算，一局游戏，处理数据的速度根本跟不上游戏画面变化的速度，更不用说，DeepMind现在也就能玩玩Atari这种爸爸辈的游戏吧。

意，在此之前有很多人工特征处理，但很明显，一旦引入了人种集成性的系统了，只能成为实验室的二维画面玩具。人类为人脑非常善于处理高维数据，并飞快的从中抽取模式。现在

由Hinton、LeCun和Yoshida等人（原谅我不能一一列举大牛深度学习界的明星方法，早已耳熟能详。但是兹事体大，还是稍首先说明的是，两种形式在深层架构上很类似，但是在每层多种神经网络之不同，DL分类如下：

神经网络的不同，分为Auto-encoder和Restricted Boltzmann

之不同，如何安排深层架构，是直接堆叠，还是通过卷积神经网络



Q SEARCH

- 第四是不同的激活函数选择，常见的是sigmoid函数，但也有通过Rectified Linear

2017年十月 (3)

最新文章列表

所谓Q-learning



Q-L的起始参数，例如episode（其表述一种天然存在分割的局。一个episode就是这样一个天然的分割。）设置为零，replay memory。

NIPS 2016—Daily Highlights

我爱机器学习(52ml.net)
2016年12月13日探

NIPS 2016—Day 1 Highlights

接
NIPS 2016—Day 2 Highlights:
欢迎关注我爱机器学习微信公
众号:
别号:
(-)



2016深度
十大指数级
12月12日
学习零基础



公告栏

2016年12月11日

Yoshua



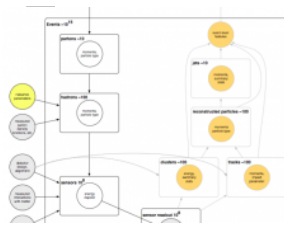
期待您的加入

2016年12月11日



SEARCH

候，不需要和环境进行交互。而上文中提到的动态规划方法，是需要跟环境交互才能计
 2017年十月 (3) 一个称作Q-function的函数，可以完全避免计算最优回报的时候
 最新文章列表 Q-function通常又被称作function approximator.



NIPS 2016—Daily Highlights

我爱机器学习(52ml.net) 2016年12月13日

NIPS 2016—Day 1 Highlights

NIPS 2016—Day 2 Highlights:...
 欢迎关注我爱机器学习微信公众
 号:



2016深度
 十大指数级
 12月12日
 学习零基础



公告栏

2016年12月11日

Yoshua



期待您的加入

2017年10月4日

h Experience Replay

capacity N

Q with random weights

ϕ_1 and preprocessed sequenced $\phi_1 = \phi(s_1)$

select a random action a_t

$= \max_a Q^*(\phi(s_t), a; \theta)$

in emulator and observe reward r_t and image x_{t+1}

x_{t+1} and preprocess $\phi_{t+1} = \phi(s_{t+1})$

a_t, r_t, ϕ_{t+1} in \mathcal{D}

batch of transitions $(\phi_j, a_j, r_j, \phi_{j+1})$ from \mathcal{D}

Q_0 for terminal ϕ_{j+1}

$\max_{a'} Q(\phi_{j+1}, a'; \theta)$ for non-terminal ϕ_{j+1}

descent step on $(y_j - Q(\phi_j, a_j; \theta))^2$ according to equation 3

最新文章列表



我爱机器学习(52ml.net)

NIBS 2016—Day 1 Highlights

NIPS 2016—Day 2 Highlights:

欢迎关注我爱机器学习微信公众



公告栏

Yoshua

期待您的加入

算法都是有数据分布独立性的假设的，IID是一个很重要的性
关联的，那么计算出来的模型就是有偏向的。但是RL中的数据通
度
深度
数级
2
2
当前情况下的状态会影响下一次的动作选择，而下一次动作
画面，下一帧画面又会影响下下次动作的选择。犹如一个长长
基础
理不清。怎么破IID的问题？DeepMind学习Long-Ji Lin 93年用
1
白
通过使用replay memory，存储过去一段时间内的“状态-动作-
行随机采样以打破依赖，以及用过去的动作做平滑。

的能力，局限性嘛，就是眼光看不到未来，正如当年葡王曼努埃尔，而西班牙女王伊莎贝拉则是拿出自己的首饰珠宝让哥伦布

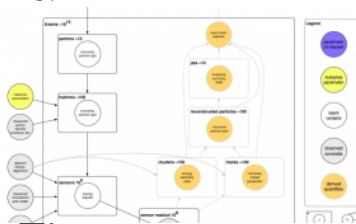


Q SEARCH

向一定的画面，而这种画面会使得agent在这个偏向上持续增强。例如，当前时刻最大化

2017年十月 (3)

最新文章列表



NIPS 2016—Daily Highlights

我爱机器学习(52ml.net)

2016年12月13日

NIPS 2016—Day 1 Highlights

NIPS 2016—Day 2 Highlights

欢迎关注我爱机器学习微信公众号

号、



公告栏

Yoshua



期待您的加入

2017年12月11日

因此agent选择向左移动，所以左侧的画面会被更多的看到，左侧的画面质量训练样本席位，从而控制进一步的学习。这种情况下，强烈的正反馈的循环会让agent迅速陷入局部最优值，甚至直接发散开。（John和Benjamin在97年的论述。）通过replay memory会让更多的历史样本参与训练影响。

根据前文叙述的RL用法，我们可以很happy的看到求解未规划的过程，因此Bellman公式大杀四方，可以快速得到最大未来，这种计算看似很好的解决问题，实则不然。这种情况下预测单条路径的情况进行计算，不具备泛化能力。比如对当前数据达到100%的正确性，但是这个100%的分类器用在其他数据果。这种情况的解决办法是，使用一个自定义的function来模拟数就可以任意选择了，例如有些人选用简单的线型函数，有些如这里使用的卷积神经网络。之前的做法是，给我当前的策略、择作为输入，通过动态规划计算出未来的回报。现在则是给定!016深度学习网络中计算出未来的回报。

十大指数级

12月12日

学习零基础

2016年12月11日

用这个DRL玩了7个Atari游戏，分别是激光骑士（Beam

out），摩托大战（Enduro），乓（Pong），波特Q精灵

quest），太空侵略者（Space Invaders）。玩这些游戏的



SEARCH

一手好牌。)当然,有一点肯定是把不同的游戏修改了的,那就是得分。不同的游戏得
2017年十月 (3)] , 导致处理起来很麻烦。因此,玩游戏的过程中,每得到一个
最新文章列表 一个负分(滚粗)就给个减一。通过这种做法让不同的游戏都融合在
一个框架内,不会因为奇怪的得分、给分方法导致出现计算上的困难。



NIPS 2016—Daily Highlights

我爱机器学习(52ml.net) 2016年12月13日

NIPS 2016—Day 1 Highlights

NIPS 2016—Day 2 Highlights
欢迎关注我爱机器学习微信公众号
大I号:



2016深度
十大指数级
12月12日
学习零基础



公告栏

2016年12月11日

Yoshua



期待您的加入

2016年12月11日

Environment模拟器,跟agent配合起来会有一些问题,因
算出来很快,超过了agent的计算判决时间,所以导致游戏
(棋牌类游戏,可以给出思考时间),因为设置ALE出k帧才
能保证玩起来不是那么的卡。在本组实验中,k通常设置为4。

,评测是简单确定的,给定了测试集,就可以对现有模型给出
测是很困难的。最自然的评测莫过于计算游戏的结果,或者几
是训练过程中周期性的分数统计。但是,这种做法会有很大的
微小扰动可能造成策略扫过的状态大不相同(回顾一下,状态
作选择会导致下一帧画面的变化,这个效应累计起来变化是巨
ind选择了更加稳定的评价策略,即直接使用动作的价值函数,
人得到的折扣回报。

少的,虽然论文本身标榜基本无预处理。但是显然,DeepMind
用现成的Deep Neural Network (Hinton 2012年做ImageNet分
并使用了GPU加速),而不是自己从头开始。正所谓“做像罗马
了直接使用“罗马人”开发的DL,首先做的是降维处理,将RGB
做了一些裁剪,将原图像由210×160采样成110×84的图
像。最终是每4帧图像合在一起当作一次训练的样本。

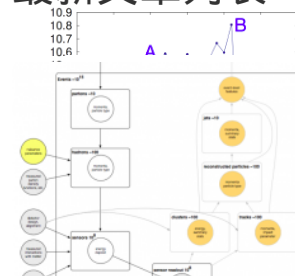


Q SEARCH

下层的至连接的线性函数。DeepMind称这种与RL结合使用的卷积神经网络为Deep Q-

2017年十月 (3)

最新文章列表



NIPS 2016—Daily Highlights

我爱机器学习(52ml.net)

2016年12月

最
个!

NIPS 2016—Day 1 Highlights

NIPS 2016—Day 2 Highlights

欢迎关注我爱机器学习微信公众号

号人:



2016深度学习十大指数级增长! 2016深度学习十大指数级增长! 2016深度学习十大指数级增长! 2016深度学习十大指数级增长! 2016深度学习十大指数级增长! 2016深度学习十大指数级增长! 2016深度学习十大指数级增长! 2016深度学习十大指数级增长! 2016深度学习十大指数级增长! 2016深度学习十大指数级增长!

学习零基础

2016年12月11日

公告栏

Yoshua



期待您的加入

2017年十月 (3) [▼](#) 来自三个方面，一是微观结构的易变性，稀疏性导致缺少显著最新文章列表是复杂动力系统的混沌性，简单的微扰会带来巨大的变化；三是人类行为的因变性，导致数据分布改变影响预测模型。而不同的目的导向也导致了不同的不同



，而是某种程度上可预测的随机。因为依据状态的不同，分布。所谓一花一世界，一叶一菩提，RL正如现实世界的一世界和人类高度的拟真性，笔者才感觉这俩是机器学习中最识你自己”，尼采也有言“离每个人最远的，就是他自己”，RL人类认识自我，认识环境的道路上渐行渐远。

NIPS 2016—Daily Highlights

13卷:

NIPS 2016—Day 1 Highlights

NIPSS 2016—Day 2 Highlights

欢迎关注我爱机器学习微信公众



十大指数这些结构都是美不胜收的。分别是“模型的结构”“数据的结

2016年12月11日

公告栏

Yoshua

期待您的加入



SEARCH

2017年十月 (3)
最新文章列表



NIPS 2016—Daily
Highlights

我爱机器学习(52ml.net) 2016年12月13日

NIPS 2016—Day 1 Highlights

NIPS 2016—Day 2 Highlights:
欢迎关注我爱机器学习微信公众
号:



2016深度学习
十大指数级
12月12日
学习零基础



2016年12月11日

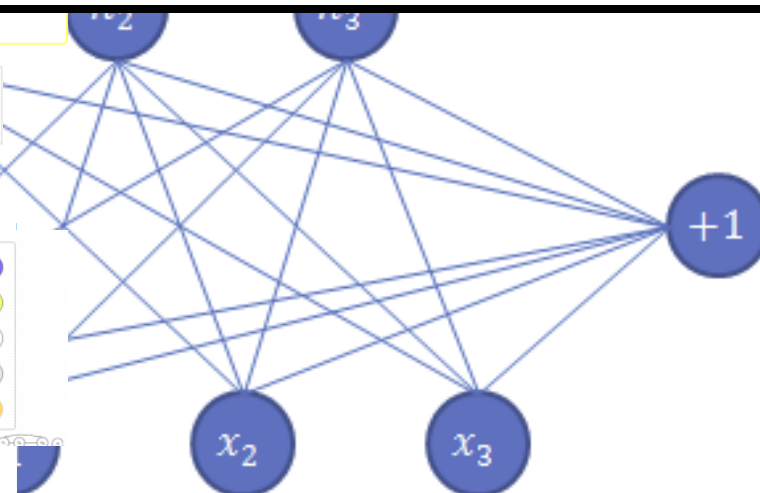
公告栏

Yoshua (obs sampling的网络相依, 节点为隐含变量和观测变量)



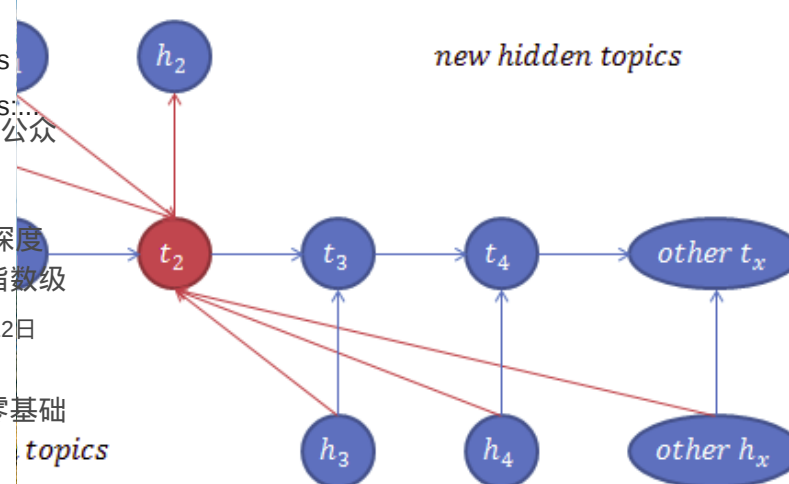
期待您的加入

2017年12月11日




的受限波尔特兹曼自动机)

Q0





SEARCH

1. Playing Atari with Deep Reinforcement Learning
- 2017年十月 (3) 最新文章列表
2. Reinforcement Learning with Function Approximation
3. Bayesian Learning of Recursively Factored Environments
- 
- Environment: An Evaluation Platform for General Agents
- Reinforcement Learning and Control
- Improve Restricted Boltzmann Machines
- Self-Difference Learning with Function Approximation

NIPS 2016—Daily Highlights

我爱机器学习(52ml.net) 2016年12月13日

1 NIPS 2016—Day 1 Highlights

NIPS 2016—Day 2 Highlights:...
欢迎关注我爱机器学习微信公众号!



2016深度学习并热爱机器学习相关内容，对自然语言处理、推荐系统等有十大指数级学习算法并行、凸优化层面的算法优化问题，以及大数据平台Hout、GraphLab等开源项目有所尝试和理解，并希望从优化层算法及平台做出贡献。



公告栏

2016年12月11日

Yoshua



期待您的加入

