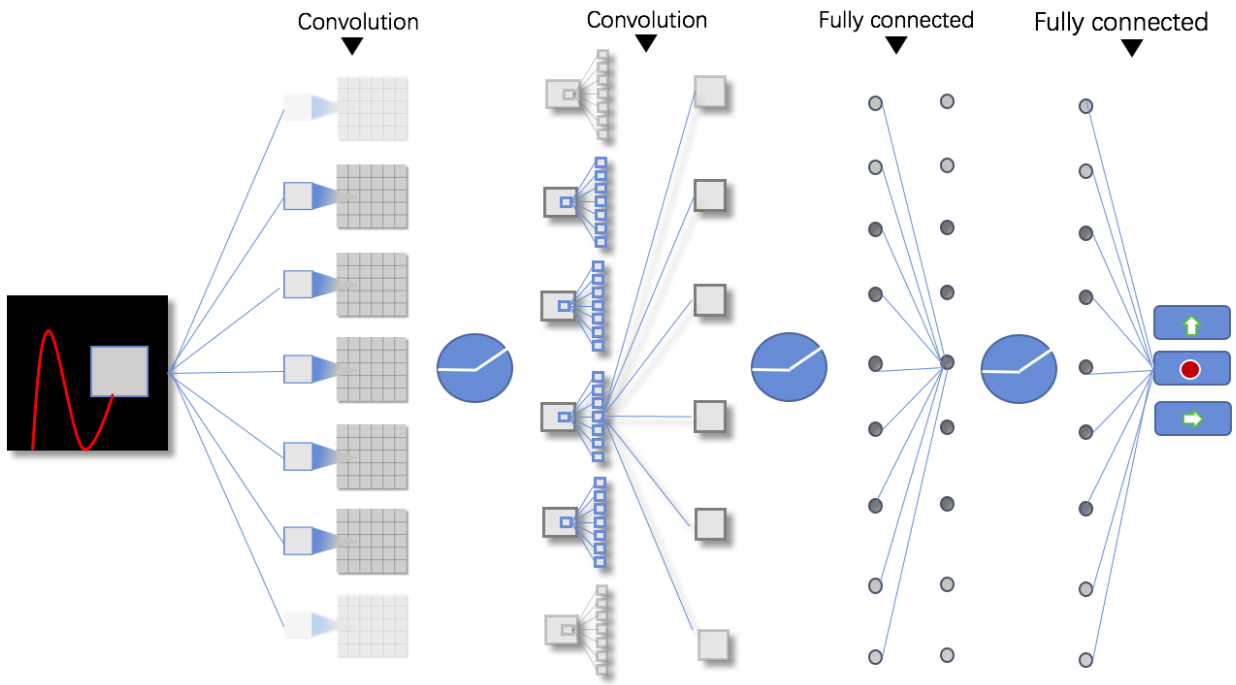


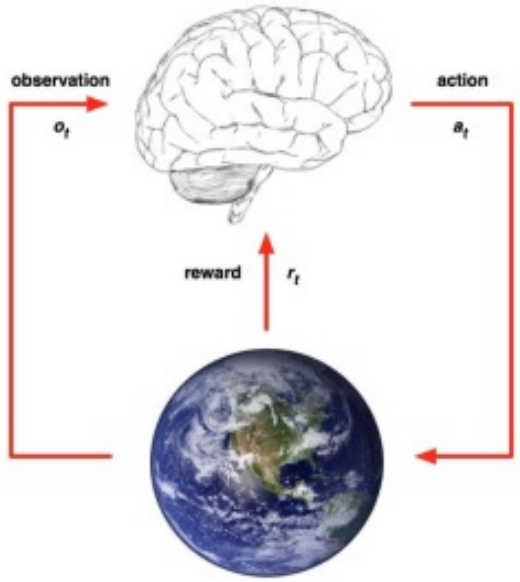
深度增强学习-1：Black-Box Optimization

庄学坤 · 6 个月前

本系列将基于John Schulman 老师在MLSS-2016上的PPT来讲解deep reinforcement learning

什么是增强学习（reinforcement learning）？

在人工智能领域，通常定义一个学习到某些知识并具有了决策能力的个体为Agent，Agent会从环境（environment）获取一些观察（observation），从而执行某种动作（action），最后这个动作会获的某种奖励（reward）。强化学习就是一类可以训练Agent做出正确决策的机器学习算法。这里以训练一个Agent学习打乒乓球为例，乒乓球从Agent左前方30度位置飞来（观察），Agent做出了决策向左移动15cm挥拍（动作），结果球扑空了（奖励），如果经历过N次这个过程，增强学习就会根据每种情形下的动作获取的奖励来进行训练，让Agent学习到如何做出正确的挥拍动作，提高成功率。



- At each step t the agent:
  - Executes action a_t
  - Receives observation o_t
  - Receives scalar reward r_t
- The environment:
  - Receives action a_t
  - Emits observation o_{t+1}
  - Emits scalar reward r_{t+1}

让机器学会打乒乓球的一个实例，感兴趣的朋友可以看下：[Deep Reinforcement Learning: Pong from Pixels]

深度增强学习是将增强学习与深度学习相结合产生的一类新的机器学习算法。它实现了从Perception感知到Action动作的端对端学习，其真正成功的开端是DeepMind在NIPS 2013上发表的Playing Atari with Deep Reinforcement Learning一文，之后DeepMind在Nature上发表了改进版的DQN文章，引起了广泛的关注。尤其是AlphaGo在围棋大赛战胜李世乭以后，深度增强学习方法被认为是通往通用学习机器的一把有力钥匙。

提到深度增强学习，大家可能首先想到的是著名的DQN算法，它是Q-Learning结合深度神经网络形成的。传统的强化学习算法在面临高维输入问题的时候，其Q矩阵往往面临维度爆炸，无法进行有效训练，而如果使用神经网络来代替Q矩阵的作用，就诞生了一种新的深度增强方法-DQN。DQN是“价值迭代”类算法的一种，当然还有一个重要的分支“策略迭代”，其代表算法是策略梯度算法（Policy Gradients）。

因此，深度增强学习是传统的增强学习结合深度神经网络之后，诞生的一类通用性很强的增强学习算法。

下面我们回到本文的正题，看一下深度增强学习的黑盒优化方法。

黑盒优化

所谓黑盒优化，就是将强化学习的决策网络当作是一个黑盒来进行优化，仅关心输入和输出，而忽略其内部机制。我们知道增强学习的监督信号是**Agent**的一个动作（**action**）所引起的奖励（**reward**），其学习目标便是不断优化决策网络，使得所有动作的奖励最大化。决策网络通常可以是Q矩阵（如DQN），或者神经网络（如PG），也就是说决策网络通常是可以参数化的，这时候我们进行优化首先要考虑的是**收敛性。**以神经网络为例，它是一个参数化的多层网络，常用的优化方法可以是误差反向传播（BP），我们可以将激励目标作为调控信号，在参数网络上进行梯度下降，使得网络的输出目标尽可能获得最大激励。这其中的核心便是我们

黑盒优化面临最大的问题就是“**收敛性**”问题，它是不依赖梯度的（*Derivative Free Optimization Approach*），只关心输入的动作和输出的奖励，如何保证收敛性？讨论收敛性问题显然超出了本文的讨论范畴，庆幸的是我们生活的宇宙恰好能够使这类算法**几乎收敛**，这种优化方法跟进化算法（EA）十分相似，而且 works **embarrassingly** well 。

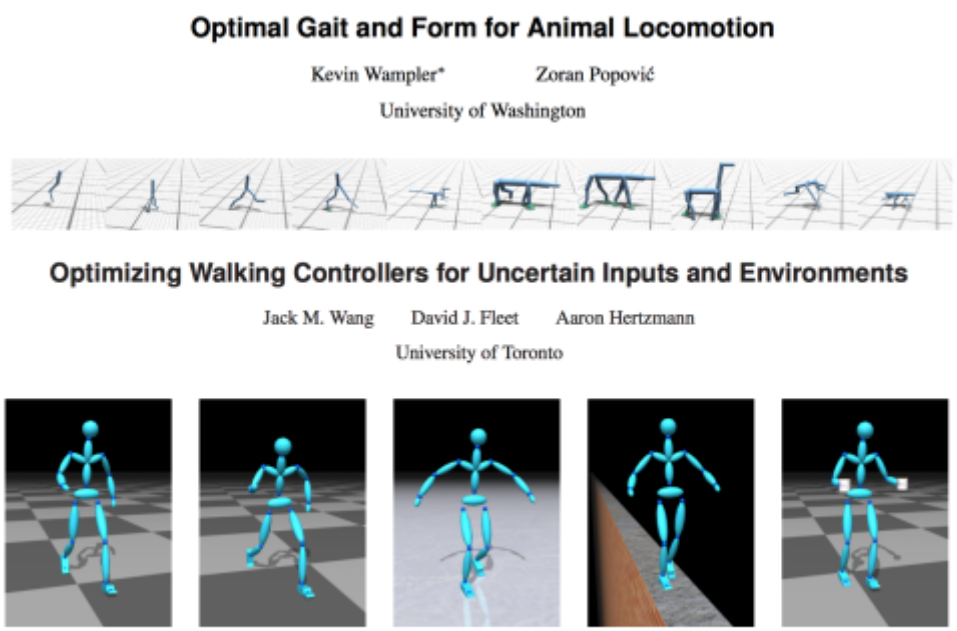
OK，到这里我们可以总结下黑盒优化的**目标函数**了。

- Objective:

$$\text{maximize } E[R \mid \pi(\cdot, \theta)]$$

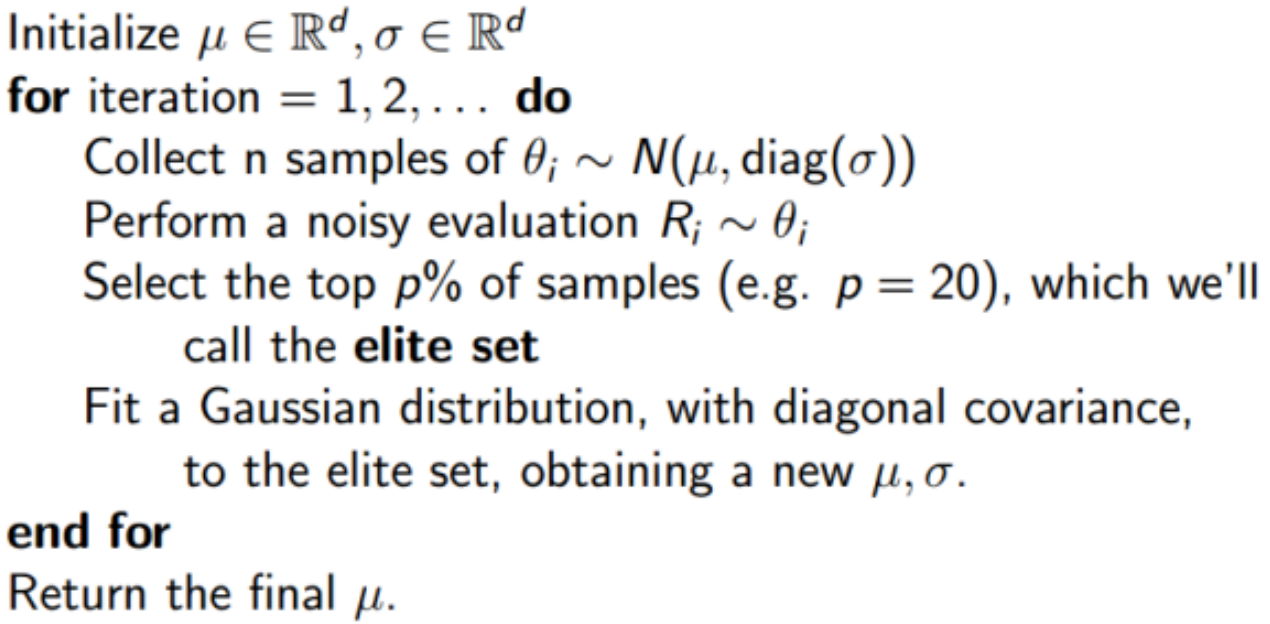
- View $\theta \rightarrow \blacksquare \rightarrow R$ as a black box
- Ignore all other information other than R collected during episode

CMA-ES 算法的一个有趣的应用



交叉熵方法（Cross-Entropy Method）

本文重点讨论的黑盒优化方法是交叉熵方法，它的工作机制跟进化算法里面的**CMA-ES**算法十分相似。基本流程如下：



其基本思路是从一个概率分布中采样出我们的策略网络参数θ,进行奖励评估，选择得分高的参数，重新估计概率分布，最终返回最优的θ。这里将策略参数看成是染色体的话，其实跟进化算法如出一辙了，就是在**进化网络参数**。

一个实例

我在GitHub上给出了John Schulman老师的一个交叉熵方法作业题的实现，仅供参考（*通常公开作业题答案并不好，这里为了给大家提供一个例子*）。

rl-gym-doc.s3-website-us-west-2.amazonaws.com...

代码：stiger104/DeepRL

下一讲我们将一起看下策略梯度算法（PG）。

更多的学习资源

墙裂推荐UC Berkeley的CS-294课程，有精力的同学可以修这门课

[CS 294 Deep Reinforcement Learning, Spring 2017]

open-ai 提供的深度增强学习练兵场，各种算法的效果都可以使用它们的接口来检验!

[OpenAI Gym: A toolkit for developing and comparing reinforcement learning algorithms]

必读经典书籍

[ufal.mff.cuni.cz/~strak...]

Sutton书上的很多DRL实例的代码

dennybritz/reinforcement-learning

[ShangtongZhang/reinforcement-learning-an-introduction]

参考资料：

rl-gym-doc.s3-website-us-west-2.amazonaws.com...

LearningTetris using the noisy cross-entropy method . In: Neural computation 18.12 (2006)

Approximate Dynamic Programming Finally Performs Well in the Game of Tetris . In: Advances in Neural Information Processing Systems. 2013

声明：本文大部分内容来自John Schulman的 PPT讲座，版权归原作者所有，仅供学习参考之用。转载请注明出处。

机器学习

强化学习 (Reinforcement Learning)

深度学习（ Deep Learning ）

还没有评论

写下你的评论...