Deep Q-Networks

CMU 10703 Deep Reinforcement Learning & Control

by Prof. Russ Salakhutdinov

翻译贡献者：

邱迪聪，CMU，Robotics (1-2)

李柏杨，电子科技大学(3-7,26-31)

name，institute, xx-xx

组长&校对：邱迪聪

「机器人学家」授权翻译

/\*\*\*\*\*\*\*\*\*\* 邱迪聪 (1-2) \*\*\*\*\*\*\*\*\*\*/

# 所使用的教学材料

提要：本章节所使用的很多材料和幻灯片是从 Richard S. Sutton 的课程，以及 David Silver 的深度增强学习指南中借鉴过来的：

* [Reinforcement Learning: An Introduction](http://incompleteideas.net/sutton/book/the-book.html) by Richard S. Sutton and Andrew G. Barto
* [Richard S. Sutton’s Home Page](http://incompleteideas.net/sutton/)
* [UCL Course on RL](http://www0.cs.ucl.ac.uk/staff/d.silver/web/Teaching.html) by David Silver

/\*\*\*\*\*\*\*\*\*\* name (xx-xx) \*\*\*\*\*\*\*\*\*\*/

content...

/\*\*\*\*\*\*\*\*\*\* 李柏杨，电子科技大学 (3-7) \*\*\*\*\*\*\*\*\*\*/

**增强学习中Agent的基本要素**

●　增强学习中，一个Agent由Policy(策略)、Value function(价值函数)、Model(模型)三部分组成，但这三部分不是必须同时存在的。

　——　策略(Policy)：定义了学习Agent在给定时间内的行为方式

　——　价值函数(Value function)：在长期运行层面界定什么才是好的选择

　——　模型(Model)：预测环境下一步会做出什么样的改变，从而预测Agent接收到的状态或者奖励(reward)是什么。[有两种类型的模型：一种是预测下一个状态(state)的过渡模型(transition model)，即**Pass’=p[St+1=s’|St=s,At=a]Pss’a=p[St+1=s’|St=s,At=a]**；另一种是预测下一次 奖励的奖励模型(reward model)，即**Ras=E[Rt+1|St=s,At=a]Rsa=E[Rt+1|St=s,At=a]**，译者注]

●　策略就是Agent从环境感知的状态到在这些状态中Agent可采用动作的一个映射

●　策略是从状态到动作的映射。有两种表达形式：

　——　确定性策略(Deterministic policy)：a = π(s)

　——　随机策略(Stochastic policy)：π(a|s) = P[a|s]

**回顾：价值函数**

●　价值函数预测了当前状态下未来可能获得的奖励的期望

　——　执行动作 a 的我能从状态 s 中获得多少奖励？

●　Q值函数(Q-value function)给出了预测的总奖励

　——　状态 s 和动作 a

　——　策略 π

　——　奖励折扣 γ

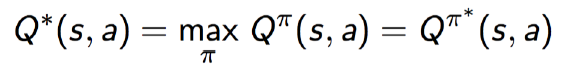
Lec10_fig1.png

●　价值函数可被分解成Bellman方程

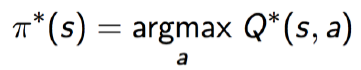
Lec10_fig2.png

**最优价值函数(Optimal Value Function)**

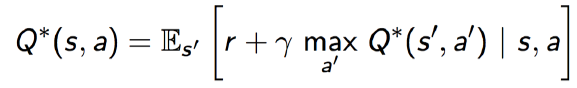
●　最优价值函数是所有策略下的价值函数的最大值



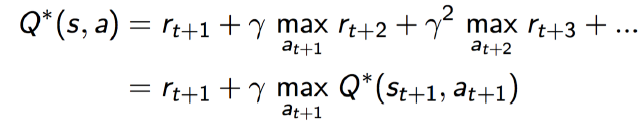
●　当我们得到 Q 的最大值Q\*，则Agent可由最优策略 π\* 执行动作 a



●　形式上，最优价值函数分解成Bellman方程



●　直观上，最优价值函数遍历所有决策后取得的最大值为

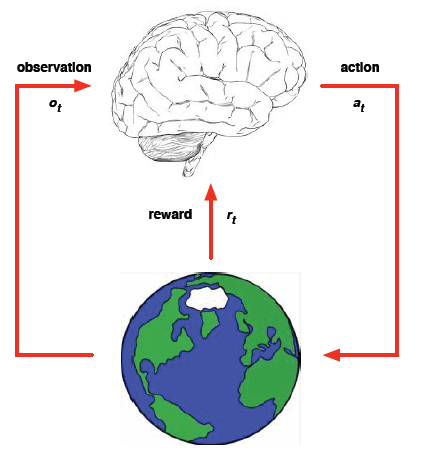
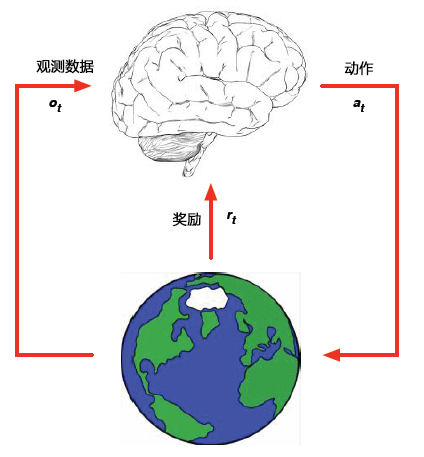


**模型(Model)**

●　模型是从经验中进行学习的

●　模型被视为环境的代理

●　模型与其建立者之间可进行交互，例如，使用前向搜索(look-ahead search)

**P.S.　沫宏，我是柏杨，3-7的内容我是在翻译任务下发之前翻译完成的，接下来的翻译任务就麻烦你了！另，若认为我有翻译不当的地方，敬请指正，辛苦了！**

/\*\*\*\*\*\*\*\*\*\* 李柏杨，电子科技大学 (26-31) \*\*\*\*\*\*\*\*\*\*/

**Dueling Networks（部分文献译作“竞争架构网络”，本部分译者作不译处理，译者注）**

　　[Dueling Networks这一神经网络架构用于无模型的强化学习中，该网络架构包含两类独立的用于估计的规则：一类用于状态值函数(state value function)，一类用于状态依存的动作优势函数(state-dependent action advantage function)。这样分类的主要优势在于泛化了跨行为学习，而不对底层强化学习算法施加任何改变。Dueling Networks在多种相似价值的动作面前能够得出更好的策略评估，译者注]

●　将Q-网络分解成两类函数表示的通道

●　非动作依存的价值函数　V(s,v)

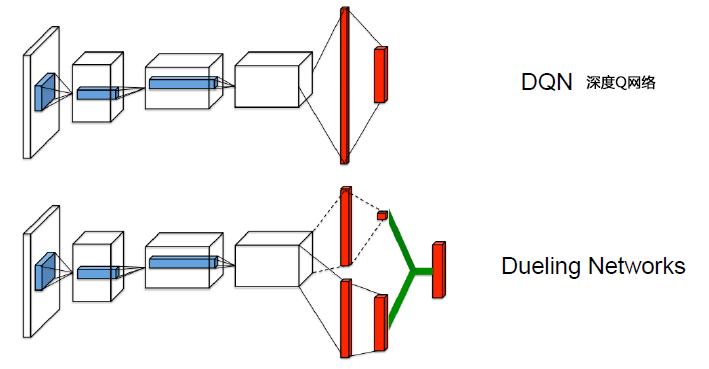
●　动作依存的优势函数　A(s, a, w)

Lec10_fig8.png

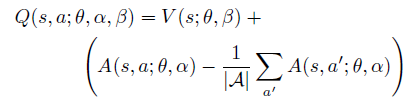
●　不妨给出优势函数的定义式

Lec10_fig9.png

**Dueling Networks Vs 深度Q网络(DQNs)**



**[流行的单流 Q 网络（图中顶部，此处以深度 Q 网络为例）和Dueling Networks（图中底部）。后者具有两类流以分别估计出（标量）状态值和每个动作的优势，绿色的输出模块基于下式**



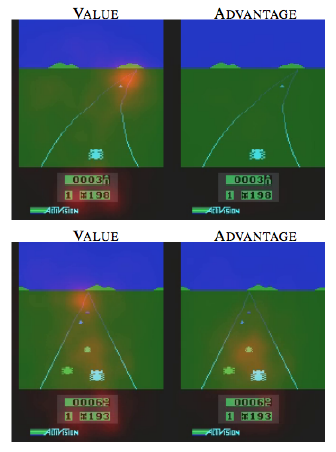
**将上述两类流进行合并。图中所示的两种网络架构针对每一个动作输出相应 Q 值]**

Lec10_fig11.png

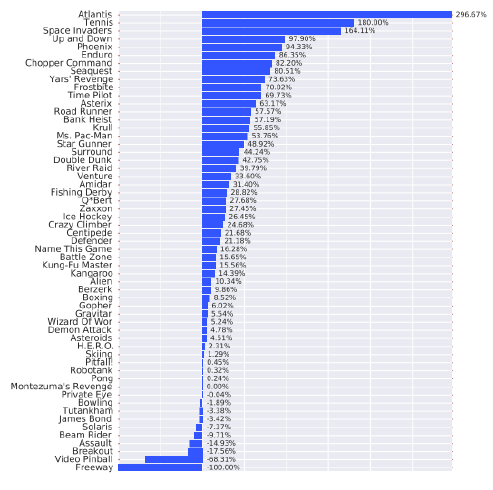
　　[在上一节译者的注解中曾提到，Dueling Networks包含两类表示价值和优势函数的流(stream)，同时共享着公共卷积特征学习模块。上述两类流经由专门的聚合层(aggregating layer)组合以生成对状态-动作值函数(state-action value function) Q 的估计，如上图所示。该网络架构可理解成具有两类流的单个 Q 网络，其取代了在现有算法【如深度 Q 网络(DQN)】中流行的单流 Q 网络(single-stream Q network)。Dueling Networks能够在没有任何额外的监督下自主地生成状态值函数和优势函数的独立估计，译者注]

●　价值流(value stream)：学会对路面投以注意

●　优势流(advantage stream)：学会只在前方有车辆时投以注意，以此避免碰撞



**[观察、注意和驾驶：一个训练完的竞争架构在ATARI游戏 Enduro 的价值显著性地图与优势显著性地图（红色标注）。价值流学会对路面投以注意。优势流学会只在前方有车辆时投以注意，以此避免碰撞。上面两张图表示，前方无车时，选择什么动作并不会太影响行车状态。这个时候系统关注状态的价值，而对影响动作优势不是很关心。下面两张图表示，前方有车时，选择动作至关重要。这个时候系统需要关心优势了。这个例子说明，Q -网络分解为价值和优势这两类函数更能突出强化学习的过程，译者注]**



**[对 Dueling Networks 架构在挑战 Atari 2600 试验台中获得的结果进行评估，一个有相同结构和超参数的强化学习Agent，必须能仅通过观察像素和游戏分数，就能玩 57 种不同的游戏。该结果例证了对 van Hasselt 等人(2015)的单流基线(baseline Single network)的大幅度改进。上图使用了下式**

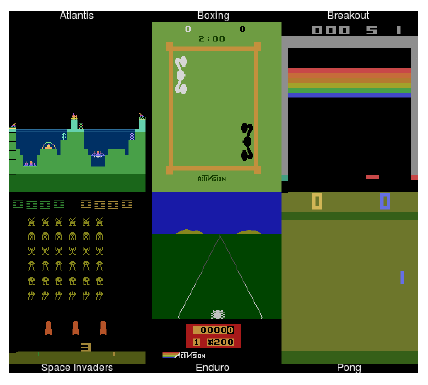
Lec10_fig15.png

**描述的度量。右侧的条状线表示 Dueling Networks 优于单流网络(single-stream network)的程度，译者注]**

**多任务深度Q网络(Multitask DQNs)**

　　[多任务学习是在多个学习任务之间实现信息的共享、转移的一种学习方法，**目前多任务学习方法大致可以总结为两类：一是不同任务之间共享相同的参数(common parameter)，二是挖掘不同任务之间潜在的共有数据特征(latent feature)**，译者注]

●　我们可以训练单个深度 Q 网络来同时进行多款游戏

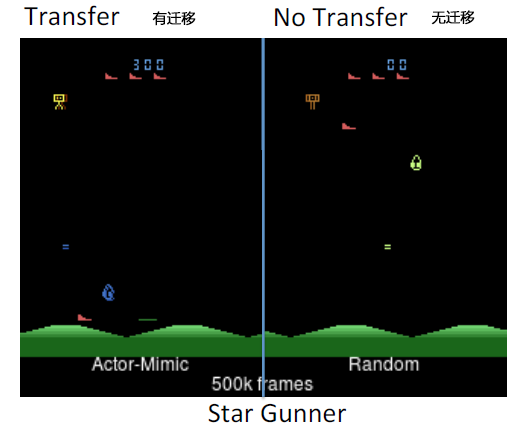


　　[在对多个相似的学习任务进行学习的时候，共享各学习任务的信息并同时对其进行求解的方法，往往比对各个学习任务单独进行求解有更高的学习精度。感兴趣的读者可参阅 **Parisotto, Ba, Salakhutdinov, ICLR 2016 “ACTOR-MIMIC DEEP MULTITASK AND TRANSFER REINFORCEMENT LEARNING”**, <https://arxiv.org/pdf/1511.06342.pdf>，译者注]

**迁移学习(Transfer Learning)**

　　[利用过去学习得到的经验、知识，来提高当前以及将来进行的学习任务的求解精度。像这样灵活应用其它学习任务的信息，使得当前学习任务的求解精度得以提高的方法称为迁移学习。迁移学习的方法是把一个学习任务的信息单方面地提供给另一个学习任务使用，它放宽了传统机器学习中的两个基本假设：**① 用于学习的训练样本与新的测试样本满足独立同分布的条件；② 必须有足够可利用的训练样本才能学习得到一个好的分类模型**，目的是迁移已有的知识来解决目标领域中仅有少量有标签样本数据甚至没有的学习问题，译者注]

●　网络可以利用有关之前学习到的游戏知识来更快地学习新游戏



　　[在本节译者第一次的注解中曾提到，迁移学习放宽了传统机器学习中的两个基本假设，具体地，在传统分类学习中，为了保证训练得到的分类模型具有准确性和高可靠性，通常情况下都有两个基本的假设：**① 用于学习的训练样本与新的测试样本满足独立同分布的条件；② 必须有足够可利用的训练样本才能学习得到一个好的分类模型。**但是，在实际应用中，,这两个条件往往无法满足。首先，随着时间的推移，原先可利用的有标签的样本数据可能变得不可用，与新来的测试样本的分布产生语义、分布上的缺口。例如股票数据就是很有时效性的数据，利用上月份的训练样本学习得到的模型并不能很好地预测本月份的新样本。另外，有标签的样本数据往往很匮乏，而且很难获得。在Web数据挖掘领域,，数据不断涌现，已有的训练样本已经不足以训练得到一个可靠的分类模型，而标注大量的样本又非常费时费力，而且由于人的主观因素容易出错，这就引起了机器学习中另外一个重要问题：如何利用少量的有标签训练样本或者源领域数据，建立一个可靠的模型对目标领域数据进行预测（源领域数据和目标领域数据可以不具有相同的数据分布）。He等人**(QING HE, XIU-RONG ZHAO, and ZHONG-ZHI SHI, “MINIMAL CONSISTENT SUBSET FOR HYPER SURFACE CLASSIFICATION METHOD”, *Int. J. Patt. Recogn. Artif. Intell*. 22, 95 (2008). )**指出数据分类首先要解决训练集样本抽样问题,如何抽到具有代表性的样本集作为训练集是一个值得研究的重要问题。该文献提出极小样本集抽样方法用于基于超曲面分类算法，该方法可感知非结构化数据的分布，并以极小样本集作为代表子集。该文还指出了极小样本集有多少种表达方式，给出了样本缺失情况下准确率的精确估计。文献表明，在实际中保证训练得到的分类模型具有高准确性和可靠性的两个基本假设并不是每种算法都能做到的，因此研究迁移学习变得非常重要，译者注]

/\*\*\*\*\*\*\*\*\*\* name (xx-xx) \*\*\*\*\*\*\*\*\*\*/

content...

/\*\*\*\*\*\*\*\*\*\* name (xx-xx) \*\*\*\*\*\*\*\*\*\*/

content...

/\*\*\*\*\*\*\*\*\*\* name (xx-xx) \*\*\*\*\*\*\*\*\*\*/

content...