n Cognitive Sciences

CellP
F

v

iforcement Learning, Fast and Slow

Botvinick,[1,2,*] Sam Ritter,[1,3] Jane X. Wang,[1] Zeb Kurth-Nelson,[1,2] Charles Blundell,[1]
ıssabis[1,2]

【强化学习 64】Fast and Slow

张楚珩
清华大学 交叉信息院博士在读

这里的Fast and Slow指的是快速的学习或者慢速的学习。

原文传送门

Botvinick, Mathew, et al. "Reinforcement Learning, Fast and Slow." Trends in cognitive sciences (2019).

特色

这是关于强化学习的一篇很有见地的review，这个月才发表的。它还从心理学和神经科学的角度来看目前强化学习的发展对其带来的影响。

过程

1. 强化学习的痛点和难点

强化学习中最大的痛点在于需要大量的样本来学习（sample inefficiency）。文章认为该痛点的来源有两处：

- 一是incremental parameter adjustment，不管是策略还是值函数的学习，目前最为强大的深度强化学习都使用的是神经网络。神经网络的训练必须是逐步进行的，如果训练的过快（比如选取较大的步长），就会导致前面学习到的部分被破坏（即产生了catastrophic interference）。如果使用神经网络，就要求了算法学习不能太快，由此限制了其sample efficiency；
- 二是weak inductive bias，即强化学习的前提假设（或者说是归纳偏置）非常少。如果假设空间大，那么自然需要更多的样本来帮助算法找到一个好的策略；如果假设空间能通过某种方式缩小（可以是人为的，也可以是算法自己学到的），并且最优解还在我们的假设空间里面，那么算法效率自然就会变得更高。

由此，文章称目前的两大类算法更好能够分别针对该痛点的两大来源来解决sample inefficiency。其中，episodic deep RL解决了incremental parameter adjustment的问题；meta RL解决了weak inductive bias的问题。

2. Episodic deep RL

incremental parameter adjustment产生的原因是因为我们对于价值函数或者策略的拟合都是使用了神经网络。神经网络是一种parametric的函数拟合方法，它通过逐步调整其中的参数来达到函数拟合的目的。要想解决此问题，我们可以使用non-parametric的方法（在心理学中也叫做instance/exemplar-based的方法）。

以价值函数的拟合为例，通常的function approximation的做法是利用过去的样本训练一个神经网络，使其能够在给定一个新样本的时能够输出其对应的价值函数。然而，这里提出的episodic deep RL则采用non-parametric，即把过去的样本（状态和对应的收益）都存储下来，给定一个新的状态的时候，将它和存储下来的样本进行比较，返回和它比较相似的历史上状态的收益作为该状态价值函数值的估计。如下图所示。

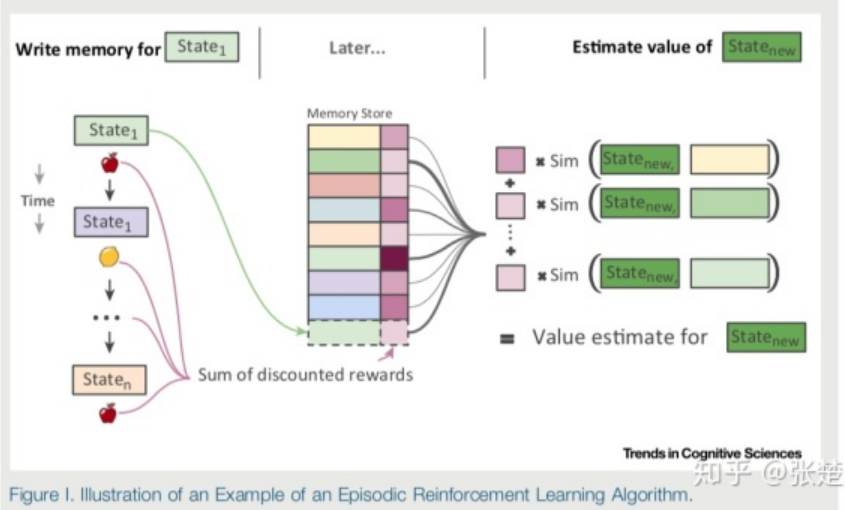

Figure I. Illustration of an Example of an Episodic Reinforcement Learning Algorithm.

在non-parametric的方法中，刚刚遇到的样本立即就能够被用于下一步的预测中（只要新的状态和该样本相似），因此，这样的学习是快速的。但是，该算法中也有慢速的部分，即状态representation/embedding的学习，该部分通常仍然由神经网络来完成。

个人感觉，对于做RL的人来说，该方法有几个值得进一步发掘的地方，比如更好的representation learning；状态之间的similarity度量；历史样本的storage、aggregation等。

3. Meta RL

下面考虑解决weak inductive bias的问题。显然，对于特定的问题，我们可以通过算法的设计产生一些有益的architectural/algorithmic bias来加速问题的学习，比如CNN就是利用了图片中的平移不变性来设计特定的算法框架，以产生这样一种特定的architectural bias。这种思路固然可以使我们

更好地解决特定的问题。

Meta RL则希望利用其它任务让算法学习到一个较好的inductive bias（个人认为，也可以叫做先验知识，prior）。如果觉得meta learning从字面上不好理解，大家可以想象成learn to learn。

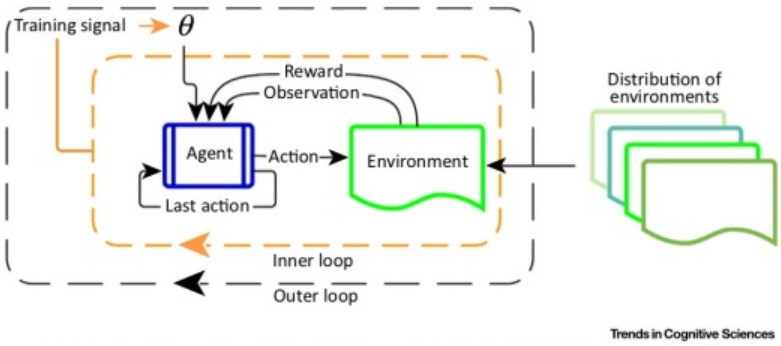

Figure 1. Schematic of Meta-reinforcement Learning, Illustrating the Inner and Outer Loops of Training. The outer loop trains the parameter weights θ, which determine the inner-loop learner ('Agent', instantiated by a recurrent neural network) that interacts with an environment for the duration of the episode. For every cycle of the outer loop, a new environment is sampled from a distribution of environments, which share some common structure.

Meta-reinforcement Learning

上面这幅图展示了一个meta RL的算法逻辑。每一次从环境的分布中采样得到一个环境，对于采样到的这个环境来学习相应的策略。对于特定的任务，内层的算法能够很快地学习到一个好的策略，外层的算法通过对于不同任务的学习来得到一个好的先验知识。

具体地，可以参考下面这幅来自文献[1]的图，策略使用一个RNN来表示，对于同一个环境来说，RNN的hidden state是连续的，即hidden state表示了对于该环境的快速学习；遇到一个新的任务之后，会重新从一个新的hidden state开始，即不同任务之间学习到的只有RNN的权重。在这样的模型下，RNN的权重代表了学习到的先验知识或者inductive bias（slow），hidden state代表了该环境的信息（fast）。

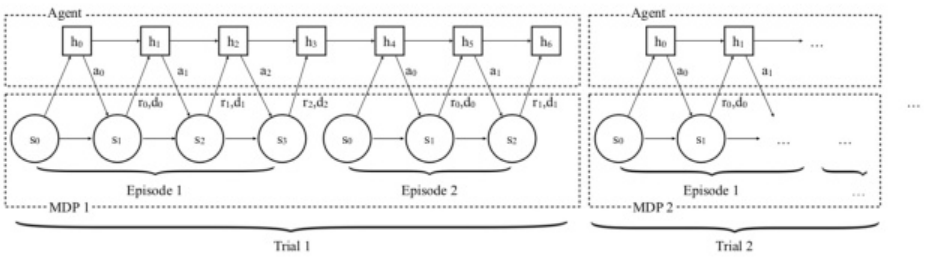

Figure 1: Procedure of agent-environment interaction

4. Episodic meta RL

上述两种方法还可以结合起来，在meta RL中，hidden state是通过RNN一步步产生的，它代表了在某个环境中快速学习获取的信息。在episodic meta RL中，hidden state可以通过把新遇到的state和历史上的state做比较，然后把hidden state直接设置为某个之前遇到的hidden state。这种方法不仅在任务之间学习到先验知识，而且在单个任务上也能很快地学习好。

5. 启发

文章后半部分讲了很多RL发展对于心理学、神经科学、认知科学等的启发，由于我也不是很了解，就不瞎讲了。有一个比较有意思的点是，文章刻画了一副生物是如何获取学习能力的图像。

- 漫长的进化把地球上的环境信息嵌入到了生物中，这对应的就是前面提到的architectural/algorithmic bias；
- 在生物体的终身学习中，奖励预测误差（reward prediction error）促使多巴胺缓慢地改变突触连接，这会使得生物体更倾向于下次做出被强化的行动。这类似于meta RL里面较慢的一层学习，产生多任务通用的先验知识或者说inductive bias。
- 在单一的任务上，生物体的会从记忆中提取相应的神经元激活的pattern到大脑皮层中，从而快速地对于新的任务做出反应。这就类似于episodic RL里面从历史样本中找出相似样本然后做出相应操作。

参考文献

Duan, Yan, et al. "RL $^ 2$: Fast Reinforcement Learning via Slow Reinforcement Learning."*arXiv preprint arXiv:1611.02779*(2016).

发布于 2019-05-29

强化学习 (Reinforcement Learning)