前沿强化学习问题

张楚珩
清华大学 交叉信息院博士在读

总结一下强化学习领域目前的一些主要问题和已有的一些解决思路，如果有做相关研究找不到课题的同学可以参考一下。现在写的不是很全，这部分内容会经常更新。

一、学习所需样本太多

现在学习一个简单的任务所需的经验数目相比与人类多了好几个数量级，即使我们只比较怼了一堆CPU/GPU的机器的运行时间，学习所需要的时间还是过多。因此，降低学习所需要的样本是一个重要的问题，即采样效率（sample efficiency）。

目前大家用到的方法有

1. **off-policy（experience replay）**：即之前的经验存下来，之后反复使用；
2. **model-based learning**：学习或者构建模型，这样一方面能够更有方向性地探索从而减小盲目探索，另一方面在预测的时候利用模型规划使得行动的质量更高；
3. **prior**：从其他地方获取先验知识，把先验知识结合进来从而能够快速学习；
4. **faster convergence**：算法本身能够更快收敛，个人感觉这部分已经探索的比较充分了；

二、学习的最终效果不好

对于工业界来说，更重要的是希望最后的效果要好，比如Boston Dynamics的机器人、比如AlphaGo下围棋。沮丧的是，如果目的是不管什么手段把这件事情做好，大家可能最后选择不用强化学习。

不过在强化学习内，大家提升最终性能的方式是

1. **针对特定的问题尽量建立针对该问题的模型**：我看到的几篇文章比如本专栏里面提到的NJUStarCraft、h-DQN就利用了这样的技巧。其中NJUStarCraft利用了对抗简单星际AI的小技巧，即先发展，然后兵爆够了一波推平，然后把策略限定在了此空间内，有效缩小了探索空间，在复杂问题上一下子取得了很好的效果。h-DQN针对Montezuma's Revenge游戏，利用了该游戏的子策略可以归纳为"走到某个位置"的特性，直接使用了带mask的特征作为子目标的表示。还有其他的针对性模型则更为隐蔽，比如针对Gym和Mujoco小人的线性模型则其实利用了在平衡态附近的动力学总可以线性化这样的先验。（并不是在批评这些工作啦，如果目标真是解决特定问题，那么这些技巧应该是能用则用）
2. **提高模型的容量和抽象能力**：比如分层强化学习通过提高其时域上的抽象能力来提高最终的性能，再比如有些model-based的方法通过提高模型的容量（比如能够刻画multi-modal分布形态）来提高最终性能。
3. **从专家的示范中学习**：复杂强化学习问题里的状态空间巨大，要让算法直接探索到合适的子空间十分困难，通过学习专家的示范，能够让算法直接从最优策略附近开始探索，即imitation learning。
4. **逐步学习**：先学习较为简单的情形，再学习更复杂的情形，即curriculum learning。

三、奖励设置很困难

大家一般跑实验就在Gym上面跑，Gym对于特定的任务都已经人为定义了比较合适奖励了。然而现实问题中奖励需要大家自己定义，而定义一个好的奖励十分困难，如果奖励定义得不好，可能直接导致整个算法学习不到东西。

如果我们直接把目标总结成奖励，那么定义出来的奖励常常十分**稀疏**（Sparse），这并不利于RL来解决。定义的奖励最好一步步引导agent来解决问题，但是agent常常是愚蠢而偷懒的，它们会想尽办法利用你所定义奖励的漏洞来骗你，所以我们需要定义的奖励最好形态完好而平滑，让agent

除了一步步划到终点别无选择。

目前大家有如下可能的解决方案：

1. 有困难也要上，再去定义更好的解决方案，比如迭代更新的Gym任务；
2. 让它能自己学习到奖励，比如用imitation learning、inverse learning；
3. 直接定义内在奖励，比如本专栏讲到的curiosity、diversity等；

四、对于特定环境的过拟合

看到有人说RL是光明正大地在测试集上训练。强化学习训练出来的agent基本上只能在训练的这个环境或者这个环境分布里面表现比较好，不是太能够泛化到其他的任务上。

目前的一些尝试：

1. 在更广的问题上学习prior，然后在特定问题上尽快学习好，比如transfer learning、meta learning、few-shot learning；
2. 逻辑推理能力，这是整个人工智能里面大家都觉得很有前景的事情了，不过目前没有看得到效果的进展；

五、调参困难

监督学习里面的超参数现在大家还比较能够驾驭了，但是强化学习里面超参数和随机性基本上让做研究的人奔溃。主要原因

1. 随机性大：由于强化学习任务本身随机性更大，监督学习每个样本假设是i.i.d.采样的，状态空间是单样本的状态空间；而强化学习每个样本的采样依赖于之前的历史，可能出现的随机情况更多，状态空间是整个轨迹（多个样本）的状态空间，这极大增加了随机性；
2. 测试一组超参数花费时间长：强化学习算法训练一次花的时间本来就长，随机性这么大吧，就要随机多跑几次降低方差，这样一来，一次调参花费的时间不知道比监督学习多了多少；
3. 算法不稳定：state-of-the-art的算法也不能保证每次跑都能收敛到期望的水平，即使参数调好了也不稳定；

基于此有人提出了可复制性的要求，希望研究人员在做实验的时候要不同随机种子多跑几次，以此来实际衡量算法的真实水平。

六、补充

写到这里的一些值得探索的强化学习问题并不是说它们不重要，主要是由于我个人目前不太了解。

1. **Reality Gap**：模拟环境中训练好的模型，拿到现实环境中用效果会大打折扣（感谢评论区补充）；当然也可以在实体机器人上进行训练，但是现实中也不可能去试验上百万次来学习，那样你扶机器人都要扶到手抽筋；同时，也希望机器人不要直接被摔坏了，因此希望safe exploration。
2. **Multi-agent RL**：多个智能体之间相互竞争或者合作，彼此互为环境的一部分，如何稳定训练？智能体之间如何通讯？

欢迎补充！

编辑于 2018-10-28

机器学习　深度学习（Deep Learning）　强化学习 (Reinforcement Learning)

文章被以下专栏收录

强化学习前沿

读呀读paper

进入专栏