深度增强学习

2017 年 5 月 18 日

有部分内容摘抄自网上

机器学习包括目标，表示和优化。目标是指学习到的模型应该达到什么目的，强化学习得到的模型使的奖赏值最大，与我们的很多状态转移决策任务的最终目标一致。表示是指原始的数据应该表示成什么样子能达到最好的学习结果，深度学习主要就是在学习特征表示。那么如果将两者结合起来，深度学习得到较好的数据表示，来表示决策过程中的状态，强化学习控制学习的方向，便得到深度强化学习。

增强学习简单的讲，就是根据当前的状态，选择一个 action，根据 action 会给一个 reward，智能体再根据给的 reward 调整自己的 action，大概率选择当前状态 s 下得分大的 action。不断这种调整，面对 s 越来越能找到 reward 大的 a。在之前的一个 qlearning 的 ppt 中解释了 qlearning 的查表法学习过程，但是那个过程要求状态和动作必须全部列举出来，只适用于状态数目可接受的情况，如果输入的是个 84*84 的图片，每个图片的灰度是 256，那么状态总数 256^{84*84}，这个状态数是无法穷举的，这时候可以考虑用一个带参数的函数 $Q(s,a,\theta)$ 来近似 Q 函数，可以使用神经网络来拟合 Q 函数，这个神经网络称为 Q-network，神经网络接收输入状态，如一副图片，输出每个 action 的 Q 值，参数 θ 就是神经网络要学习的权值，那么如何求解 θ 是一个有监督深度学习问题，这并不意味着和强化学习没有关系，因为这里的 target 也就是常说的 label 并不是人工标注的，而是与环境互动得到的。

这里的目标是估计 Q 值，是一个回归问题，那么目标函数可

以选择最小化均方差，

$$J = min(r + \lambda max_a Q(s, a) - Q(s, a, w))^2$$

但是有个问题，神经网络要求数据独立同分布，数据分布是固定的，但是增强学习中，不同时刻的状态高度相关，且数据分布可能会改变，因为根据 state 选择 action，action 决定下一个 state，如果把一批连续的 $(action, reward.state$ 作为一个 batch 给神经网络训练，那么可能不同 batch 上得到的模型不一致。(后来有一片文章，用旧的神经网络得到目标值，

$$J = min(r + \lambda max_a Q(s, a, w_{old}) - Q(s, a, w))^2$$

, 这样做可以更好的打破数据之间的关联关系)

有一个比较直观的解决方法,将训练过程中的样本 (s_t, a_t, r_t, s_{t+1}) 放入一个集合 D 中，每次迭代更新时，从 D 中随机选择一个 min_batch, 这样可以消除样本间的相关性，并且使数据的分布尽量不变

为了对于任意的状态，都能给出最优的 action, 那么应该尽力丰富 $(s_t, a_t, r_t, s_t + 1)$，在训练开始阶段，并不是最优的网络，如果读到输入的状态，并利用这个 action，然后转移到下一个状态，并不合理，所以在训练开始时，增加样本的随机性，使用随机策略选择 action，后期减少使用随机策略，用训练出的 Q-network 选择 action。