强化学习

1. 序

2016年，AlphaGo Master击败了李世石，而AlphaGo Zero花了40年就击败了它的前辈AlphaGo Master。

2019年1月，AlphaStar在《星际争霸2》中已10:1击败了顶级职业玩家。

2019年，OpenAI在《Dota2》的比赛中战胜了世界冠军。

做到这些的正是强化学习。

什么是强化学习：

强化学习相较于其他机器学习的算法而言，就是在不断尝试的过程中学习到在特定的情形下选择那种行动能够得到最大的回报。

强化学习的特点

最重要的3个特点在于：

基本是以一种闭环的形式；

不会直接指示选择哪种行动（actions）；

一系列的actions和奖励信号（reward signals）都会影响之后较长的时间。

强化学习的典型算法：

1. Q-Learning
2. Sarsa
3. DQN算法
4. Policy Grandients
5. Actor-Critic
6. DDPG算法
7. AC3算法
8. Monte Carlo
9. 强化学习举例

样例一

在Flappy bird这款游戏中：

小鸟就相当于主体(Agent)；

小鸟某一时刻的飞行状态是(State)；

让小鸟躲避的水管相当于环境(Environment)；

点击让小鸟飞行就相当于行动(Action)；

小鸟飞得越远获得越多的积分就相当于奖励(Reward)。

强化学习比起监督学习和无监督学习二者，相差最多的就是可以不断地尝试来学习某些技能。

样例二

如图例子中，每个节点表示一个房间，箭头表示通路，可以从一个房间走到另一个房间，节点5表示房间外。

加入让一个AI从房间中走到房间外，那么就要给进入每一个房间设置奖励，相当于路径的权值。通向房间5的可以获得100积分，其余的都是0。在房间5徘徊也可以获得100点奖励。

假设AI在房间1，要走到房间5。那么对于一个对环境一无所知的AI，我们就要让他学会如何到达房间5。我们把每一个节点叫做state，从一个节点到达另一个节点叫做action。

房间1可以到达3,5。进入房间3可以到达房间2,4或者返回1。房间2只能返回房间3，房间4可以到达0,3,5。到达房间0只能返回房间4。

那么我们就可以建立一个简单的关联矩阵R，定义矩阵Q为AI学到的经验，它的列表示当前状态，行表示到下一个状态可能的行动。那么：

Q(state,action)=R(state,action)+Gamma\*Max[Q(next state,all actions)].

使用Q-learing算法，设置gamma=0.8，初始状态房间为1。那么Q矩阵为：

假设AI选择了进入房间5，则R：

Q(1, 5) = R(1, 5) + 0.8 \* Max[Q(5, 1), Q(5, 4), Q(5, 5)] = 100 + 0.8 \* 0 = 100

因为这立刻的奖赏R（1，5），下一个状态是房间5，也即是当前的状态，因为5是目标状态，我们这一次训练季便结束了。得到Q：

假设AI在房间3，选择进入房间1，那么此时有房间3,5可以选择：Q(3, 1) = R(3, 1) + 0.8 \* Max[Q(1, 3), Q(1, 5)] = 0 + 0.8 \* Max(0, 100) = 80。Q为：

轨迹图：

不断训练，Q矩阵会最终收敛成：

除以最大值正规化得到：

轨迹图就变成了：

这样路径就生成了，训练结束。

1. 强化学习缺点

强化学习的一大缺点就是比较慢。

导致这种情况的原因之一是：强化学习是一种增量式的参数更新。从最早的算法开始，不管是周围环境的输入，还是AI动作的输出，他们都是通过梯度下降来完成映射的。在这个过程中，必须得让每个增量都很小才能保证学到新的信息，并且把之前的经验给覆盖。这样就会致了它速度的缓慢。

强化学习速度慢的原因之二就是：弱归纳偏置，它会尽可能的限定跟多的可能性，让AI从中选择想要的，这就会导致学习编码。此外，通用神经网络的系统偏见都很低，有数目及其庞大的参数，可以用来拟合许多大范围的数据。

首先要解决的问题是参数增量：在强化学习的过程中，如果遇到需要作出决策的新事件时，就可以将当前事件已有的内部表征和过去的事件对比一下，将匹配分数高的选中。这样可以使过去学到的信息都可以排上用场。

其次得解决的问题是归纳偏置：借鉴过去的经验，以限定一个下狭窄的范围去学习。这在深度学习中也被称作元强化学习。使用情节性元强化学习，可以在从前的经验中恢复一些因那隐藏的活动，从而让之前学过的信息排上用场，来影响当前的策略。

1. 应用场景
2. 机器人、自动化
3. 保健、医学
4. 教育、培训
5. 数据科学、机器学习
6. 传媒、广告

机器人和工业自动化

强化学习在机器人和自动化等中的应用已经是学术界和工业界共同的研究课题。同时初创公司也开始使用强化学习来打造机器人产品。

工业自动化是另一个有前景的领域。

保健和医学

强化学习的智能体和环境进行交互并基于所采取的行动接收反馈的场景和医学里学习治疗策略有相似之处。

教育和培训

在线平台已经开始尝试使用机器学习来创建个性化的体验。

无人驾驶

无人驾驶方面，它运用到了众多技术，世界上最先进的无人驾驶汽车可以在无人干预的情况下行驶八万公里。国内首量无人驾驶汽车于2005年，在上海交大研制成功。

保健和医学

强化学习的智能体和环境进行交互并基于所采取的行动接收反馈的场景和医学里学习治疗策略有相似之处。

文字，语音和对话系统

自然语言处理方面，科大讯飞的语音合成技术也在诸多世界赛事中屡次刷新世界纪录，连续7年夺得了暴风雪测评杯大赛的第一，备受人们的关注，合成出来的语音字正腔圆、抑扬顿挫，已经达到了以假乱真的地步，此项技术应用于语音唤醒、语音识别、智能家居等多个方面。

传媒和广告

微软最近的一篇论文里介绍了一个名为Decision Service（决策服务）的内部系统，这个系统已经在Azure上开放。论文里描述了决策服务在内容推荐和广告中的应用。强化学习的其他应用包括优化跨渠道营销和实时投标在线广告系统。

参考文献

1. https://zhuanlan.zhihu.com/p/36669905

2. https://easyai.tech/ai-definition/reinforcement-learning/

3. https://www.zhihu.com/question/31140846

4. https://www.cnblogs.com/yeyuan111/p/10320050.html