彭钰驯+1170300916+AI＆ML（1）

评分：

1. alphago的原理以及大体上用的是什么学习机制

在架构上，AlphaGo可以说是拥有两个大脑，两个神经网络结构几乎相同的两个独立网络：策略网络与评价网络，这两个网络基本上是个13层的卷积神经网络所构成，卷积核大小为5\*5，所以基本上与存取固定长宽像素的图像识别神经网络一样，只不过我们将矩阵的输入值换成了棋盘上各个坐标点的落子状况。

第一个大脑「策略网络」基本上就是一个单纯的监督式学习，用来判断对手最可能的落子位置，本质是一个多分类问题。他的做法是大量的输入这个世界上职业棋手的棋谱，用来预测对手最有可能的落子位置。AlphaGo的策略网络有做了两个层面增强，第一个层面是利用了名为增强策略网络(reinforced-learning (RL) policynetwork)的技术，他先使用部分样本训练出一个基础版本的策略网络，以及使用完整样本建立出来的进阶版策略网络，然后让两个网络对弈，后者进阶版策略网络等于是站在基础版前的「高手」，因此可以让基础网络可以快速的熟即到高手可能落子的位置数据，进而又产生一个增强版，这个增强版又变成原有进阶版的「高手」，以此循环修正，就可以不断的提升对于对手(高手)落子的预测。第二个层面则是现在的策略网络不再需要在19\*19的方格中找出最可能落子位置，改良过的策略网络可以先透过卷积核排除掉一些区域不去进行计算，然后再根据剩余区域找出最可能位置。

第二个大脑是评价网络。在评价网络中则是关注在目前局势的状况下，每个落子位置的「最后」胜率(这也是我所谓的整体棋局)，而非是短期的攻城略地。也就是说策略网络是分类问题(对方会下在哪)，评价网络是评估问题(我下在这的胜率是多少)。评价网络并不是一个精确解的评价机制，因为如果要算出精确解可能会耗费极大量的计算能力，因此它只是一个近似解的网络，而且透过卷积神经网络的方式来计算出卷积核范围的平均胜率(这个做法的目的主要是要将评价函数平滑化，同时避免过度学习的问题)，最终答案他会留到最后的蒙利卡罗搜索树中解决。

可以看到，alphago的原理是一种强化学习方法。

1. 强化学习的定义，并且举一个例子

强化学习：强化学习是智能体（Agent）以“试错”的方式进行学习，通过与环境进行交互获得的奖赏指导行为，目标是使智能体获得最大的奖赏，强化学习不同于连接主义学习中的监督学习，主要表现在强化信号上，强化学习中由环境提供的强化信号是对产生动作的好坏作一种评价(通常为标量信号)，而不是告诉强化学习系统RLS(reinforcement learning system)如何去产生正确的动作。由于外部环境提供的信息很少，RLS必须靠自身的经历进行学习。通过这种方式，RLS在行动-评价的环境中获得知识，改进行动方案以适应环境。

简而言之，强化学习就是不断评价再改进强化的一个过程。

强化学习的实例：alphago，阿里小蜜，淘宝电商搜索