AlphaGo 与深度强化学习

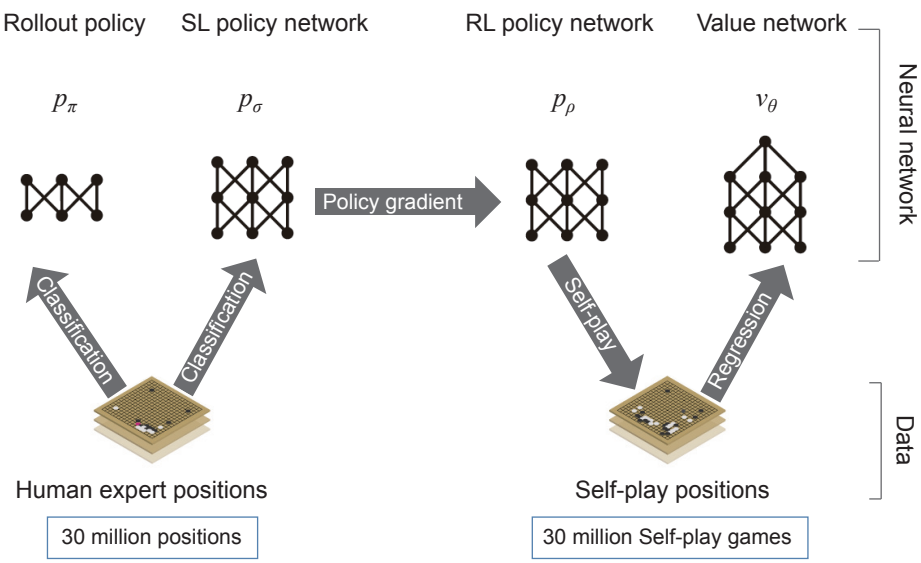
给大家讲讲比较有趣的深度学习和强化学习中的概率统计。之所以选取这个题材，是因为 AlphaGo 现在是个热门话题，计算机居然战胜了世界围棋冠军，先前战胜了韩国的李世石，前不久又战胜了我们中国的柯洁。

AlphaGo 与韩国选手李世石对弈（图片取自网络）

当然，AlphaGo 这么成功，有很多的技术，包括芯片、计算机等等。我这里不是要讲技术和围棋，而是要讲讲 AlphaGo 用到的数学，谈 AlphaGo 算法用到的深度强化学习和蒙特卡罗树搜索，在这里面用到了很多的概率统计知识。

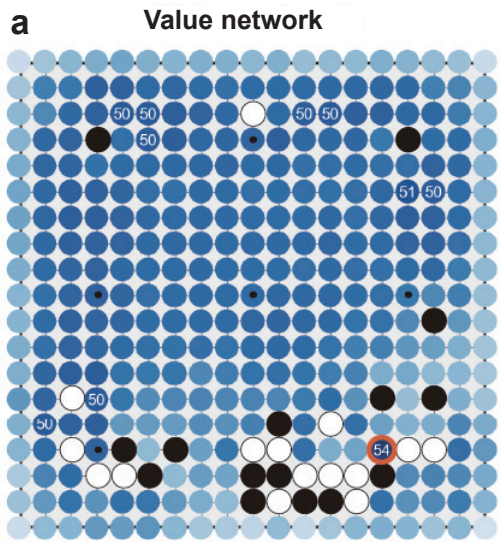
在讲述之前，我公开申明：我要感谢微软亚洲研究院的贺迪。起因是中国科学院大学的一二年级大学生做科创计划，他们选择了学习 AlphaGo，研究 AlphaGo 的概率统计原理，希望我做他们的导师。我就通过我在微软工作的过去的学生陈薇邀请到贺迪，请他给我们作报告介绍 AlphaGo 的原理。下面介绍的内容部分取自贺迪的报告，部分取自查阅互联网获得的资料，不一一注明知识产权的出处。

人工智能下棋已经有很长历史，过去 IBM 有一个深蓝团队，用“深蓝”计算机下国际象棋。国际象棋所有棋局穷尽了大概是 10^{47}，而围棋的所有棋局的可能性大约是 10^{170}。要知道我们整个地球的原子总数也只有 10^{80}，因此围棋的棋局总数远比地球所有原子数目多，这真是一个大数据。过去 IBM 团队用“深蓝”同人类下国际象棋时，用的方法是穷尽，把所有国际象棋的棋谱都让计算机学了。但是，对于围棋做不到，目前的计算机不可能穷尽 10^{170} 这个天文数字。因此设计围棋的人工智能时必须用随机的方法，用概率统计的方法，在具体设计算法时还要有很多智慧和技巧。



AlphaGo 训练的四个神经网络

谷歌的研发团队用深度学习和强化深度学习为 AlphaGo 训练了四个神经网络，相当于四个大脑，它们是：快速走子策略，监督学习策略，强化学习策略和估值网络。研发团队先用 KGS 围棋服务器上的 3000 万个棋局有监督地学习出两个神经网络：其一是用 13 层卷积神经网络学出来的监督学习策略，另一个是用逻辑回归学出来的快速走子策略。这两个网络都可以近似理解为基于 3000 万个有标注的数据 (s, a)，评价在当前局面 s 下，棋子落在某一位置 a 的概率，也就是 $p(a|s)$。其中“快速走子策略”可以被看作是“监督学习策略”的轻量级版本，它能够比“监督学习策略”快 1000 倍，但是精确性较差。

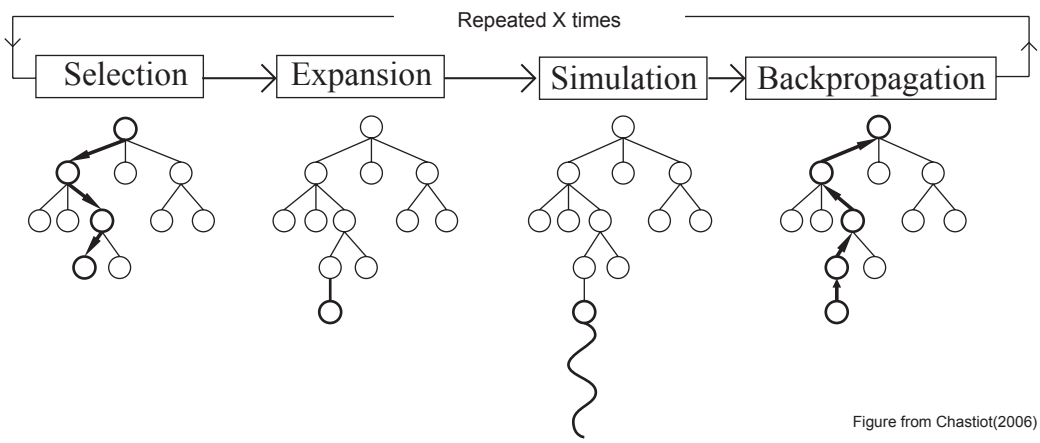


从任意给定棋盘局面去猜测大致的双方赢棋概率。深蓝色表示下一步有利于赢棋的位置。（截图取自 doi:10.1038/nature16961）

AlphaGo 的强大在于它还有自我学习的能力。在监督学习策略的基础上，通过机器和机器自我对弈，又产生多达 3 千万个标注样本，每个样本的局面 s 都来自不同的棋局，它再用大量增加的样本自我学习，训练出一个强化学习策略网络。这个网络也是评价在当前局面 s 下，棋子落在某一位置 a 的概率。而第四个网络，是在策略网络和强化学习网络的基础上训练出来的估值网络，它可以估出在当前棋局下获胜的概率有多大。总体来说，前三个神经网络都以当前围棋对弈局面为输入，经过计算后输出可能的走子选择和对应的概率，概率越大的点意味着神经网络更倾向于在那一点走子，这个概率是针对输入局面下所有可能的走子方法而计算的，也就是每个可能的落子点都有一个概率，当然会有不少的点概率为 0。第四个神经网络是用来进行价值判断的，输入一个对弈局面，它会计算出这个局面下黑棋和白棋的胜率。我的理解，四个网络都是概率，前三个是概率分布，第四个是一个概率值。

这些都是下棋前的准备工作，真正下棋的时候，它用的是蒙特卡罗树搜索 (MCTS) 算法。这个算法用到贝叶斯分析，用到马氏链，还用到其它数学方法。关键的是，它在不断地用蒙特卡罗树搜索的时候，还不断地自我更新它的策略，这就体现了人工智能。

MCTS 算法有不同版本，并且在不断地改进。 在 *Nature* 上发表的 Google



蒙特卡洛树搜索示意图（图片取自 https://www.jianshu.com/p/d011baff6b64）