DOI: 10.7641/CTA.2016.60526

解读AlphaGo背后的人工智能技术

刘知青†, 吴修竹

(北京邮电大学 软件学院, 北京 100876)

摘要: 随着人工智能在各个领域的应用, 越来越多的问题通过人工智能得到更优的解决, 但是围棋因其本身的复杂度一直是人工智能领域的难解之题. AlphaGo团队利用了人工智能中的一个重要分支—深度学习训练了一款围棋人工智能程序, 并在2016年3月与职业九段选手李世石的对弈中以4:1的比分获胜, 受到了大众的广泛关注. 本文介绍了AlphaGo这一程序背后的复杂的网络构造以及不同网络的优缺点.

关键词: AlphaGo; 深度学习; 价值网络; 策略网络

中图分类号: TP273　　**文献标识码:** A

Interpretation of the artificial intelligence technology behind Alphago

LIU Zhi-qing†, WU Xiu-zhu

(College of Software, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: With the application of artificial intelligence in various fields, more and more problems have been solved. But computer Go has been a difficult problem in the field of artificial intelligence, because of the complexity of the game. AlphaGo team has trained a Go AI program which took advantage of an important branch of artificial intelligence – deep learning. In March 2016 AlphaGo won 4–1 the game with professional Go player Lee se-dol (9P), received extensive attention of the public.

Key words: AlphaGo; deep learning; value network; policy network

1 引言(Introduction)

计算机博弈是人工智能领域的一个重要分支, 计算机围棋在计算机博弈中处于核心的地位. 将深度学习方法应用到计算机围棋中是AlphaGo团队的主要方法, AlphaGo利用"价值网络"去计算局面, 用"策略网络"去选择下子, 这两个不同的神经网络像两个大脑一样合作来改进下棋[1].

本文简要讲述了 AlphaGo 背后的人工智能技术, 第2节简要讲述了 AlphaGo 的发展阶段, 第3节介绍了AlphaGo中运用到的人工智能技术, 第4节对全文进行了总结并对人工智能进行了展望.

2 AlphaGo发展历程(Development of AlphaGo)

大约在2014年3月, Google旗下 DeepMind 公司的AlphaGo团队开始了以"测试是否能用深度学习实现围棋的理解与对弈"为主题的研究, 到2015年8月, AlphaGo已经可以全面超越当时顶级的计算机围棋程序. 2015年10月AlphaGo(版本13)以5:0战胜樊麾二段, 而3个月之后AlphaGo(版本18)让4子与AlphaGo(版本13)对弈取得胜利, 并达到4500等级分, 超出了顶级围棋职业选手1000等级分. 2016年3月, AlphaGo(版本18)以4:1战胜了李世石九段, 引起了大众的关注. 3月24日AlphaGo的David Silver在 UCL 报告了AlphaGo技术, 从这份报告中可以发现AlphaGo的水平仍在不断的提升, 尤其是其在价值网络上的提升.

AlphaGo今后的目标着眼点是可以改变世界的应用, 比如精准医疗、家用机器人以及智能手机助手等.

3 AlphaGo背后的人工智能方法(Artificial intelligence method behind AlphaGo)

围棋一直以来就因其具有的天文数字的状态空间和决策空间被认为是最复杂的智力游戏, 它涉及逻辑推理、形象思维以及优化选择等多种人类智能, 是公认的人工智能领域长期以来的重大挑战, 国际学术界曾经普遍认为解决围棋问题需要10~20年时间. 正是由于这种天文数字的状态空间和决策空间, 普通的蛮力计算无法解决围棋的问题.

围棋职业选手得以取胜的主要原因是多年培养的棋感直觉以及对当前盘面可能产生的变化进行搜索

收稿日期: 2016−07−19; 录用日期: 2016−11−16.

†通信作者. E-mail: linxiaomo1992@163.com; Tel.: +86 13269570900.

本文责任编委: 胡跃明.

验证, AlphaGo的人工智能正是将这两种方法模拟出来: 落子与胜负的棋感直觉以及落子与胜负的搜索验证, 利用这两个方法, AlphaGo解决了围棋的复杂性问题.

AlphaGo主要包括4个部分, 它们分别是策略网络、快速走子网络、价值网络、以及蒙特卡罗树搜索. 其中策略网络主要用于预测下一步的走棋, 快速走子网络形成自身对弈的棋谱, 价值网络则用于估计当前局势, 而蒙特卡罗树[2–3]搜索则用于把3个部分联系起来, 形成一个完整的系统.

3.1 棋感直觉(Intuition)

棋感直觉一直以来就是高水平围棋对弈的要素之一, 它反映了一名职业棋手长期的学习、训练、以及对弈所积累的经验. AlphaGo通过深度神经网络机器学习的方法, 使程序获得了围棋棋感直觉, 并且对该网络的训练强度远超过任何棋手的个人能力, 因此才能实现AlphaGo的超人棋力.

3.1.1 策略网络: 落子棋感(Policy network)

策略网络是深度神经网络中的有监督学习, 即直接对已有的棋手对弈棋谱进行不断的学习, 从而获得棋手的棋感. 策略网络学习的对象是利用围棋知识从当前盘面中提取的多个特征矩阵, 这其中包括棋盘各点上的棋子颜色、现在各点气的情况、所有合法点的位置等等, 除此之外每个盘面都有一个标示, 这个标示标记了该盘面下一步棋手是如何落子的. 在训练过程中, 训练者建立好13层的策略网络模型之后, 将特征矩阵作为输入投入到计算中, 经过每一层的计算最终得到程序预测的该盘面下一步走棋的位置, 在这之后程序将预测到的位置与该盘面的标示进行比较, 再根据比较结果对每一层网络的参数进行更新, 如此周而复始的进行多次计算后最终得到的策略网络便具有了一般棋手的棋感直觉.

策略网络的网络结构中用到的是卷积神经网络的算法, 这是一种近年来广泛应用于模式识别、图像处理等领域的一种高效识别算法. 卷积神经网络具有结构简单、训练参数少和适应性强等特点, 所以在计算机围棋这种计算数量级庞大的问题上卷积神经网络具有其他算法不具有的优势. 策略网络学习的训练集是由职业棋手和业余高端棋手的棋谱组成的, 这包括十几万份棋手的棋谱, 也就是上千万数量级的落子方式. 在经过这一层网络的训练之后, AlphaGo也就获取了围棋盘面下的落子棋感.

值得注意的是, 在这样的模型下, 策略网络预测的正确率可以达到57%, 不过运行的时间略长, 然而正确率的一小点提升都会使策略网络的棋力有很大的进步, 只不过消耗的计算时间也就越长.

除此之外, AlphaGo又训练了一个快速走子网络. 这个网络与策略网络具有基本相似的结构, 不同的是该网络输入的从盘面提取的特征矩阵更少, 并且将策略网络中一些复杂的网络结构简化为简单的线性结构, 使网络更加简单, 计算也更加快速. 快速走子网络的准确率只有24.2%, 但是它的计算时间比策略网络小三个数量级.

3.1.2 策略网络的强化学习(Reinforcement learning of policy network)

为了提高策略网络的正确率, AlphaGo对策略网络进行了一次强化学习[4], 与策略网络的初次学习不同的是, 强化学习选择了累积奖赏回报训练网络更新参数的方法: 利用现有的策略网络与策略网络中随机选择的一起迭代进行对弈, 将对弈结束后的结果进行记录, 并最终根据记录的结果对网络进行更新, 从而形成更可靠的策略网络. 强化后的策略网络在80%的比赛中都可以打败初次学习后的策略网络, 在与其他的计算机围棋软件的对弈中也比初次学习的策略网络成绩有明显的提高.

3.1.3 价值网络: 胜负棋感(Value network)

价值网络模型是AlphaGo训练通道的最后一个步骤. 这个网络的数据集是利用强化后的策略网络进行三千万盘的自我博弈形成的棋谱, 通过学习三千万盘棋的胜负结果进行学习和更新, 从而获取在围棋盘面的胜负棋感. 价值网络的输入与策略网络基本相同, 不过在策略网络的基础上又添加了现在要走子的颜色这一特征矩阵, 价值网络的网络结构也是在策略网络的基础上加入了一个卷积网络层与两个全连接层, 在经过15层的网络训练之后价值网络将输出对当前该盘面胜负的预测.

3.2 蒙特卡罗树搜索: 搜索验证(Monte Carlo tree search)

没有棋感直觉是不行的, 但是完全依赖棋感直觉也是不可靠的, 想要得到最可信的结果, 就要通过严格的数学模型和计算方法对棋感直觉进行验证. AlphaGo使用了蒙特卡罗树搜索, 对落子棋感和胜负棋感进行了计算验证, 最终得到最为有效的结果.

蒙特卡罗树是在树搜索的过程中采用蒙特卡罗方法对树的节点进行动态评估, 通过评估的结果来指引对搜索树的选择, 一般的蒙特卡罗树在节点的选择上采用了UCT(upper confidence tree)公式进行选择, AlphaGo则结合了策略网络和价值网络对节点进行评估并选择最优点.

为了高效的结合蒙特卡罗树与深度神经网络, AlphaGo利用了异步多线程搜索在CPU上执行模拟建树, 在GPU上并行计算策略网络和价值网络. 在训练了普通版本的AlphaGo的同时, 研究团队还训练了分布式版本的AlphaGo, 这一版本在配置上比普通版本

强化了很多, 并且比一般版本的ELO评分高300分左右.

3.2.1 快速模拟采样: 胜负棋感验证(Fast rollout)

快速模拟采样是基于数学期望的胜负评估模型. 这个模型基于蒙特卡罗模拟新型胜负结果的采样, 并根据这一采样的结果验证盘面胜负的数学期望. 这一模型的可靠程度与采样的规模有直接的关系.

3.2.2 最大信心上限搜索: 落子棋感验证 (Upper confidence bound 1 applied to trees)

最大信心上线搜索是在线机器学习的重要方法, 这一方法平衡了机器学习过程中探索与利用之间的矛盾. 搜索最优的落子点, 同时也搜索次数最多的、信心最大的、胜率最高的落子点, 两者结合起来才是最可信的结果.

AlphaGo整个落子过程的搜索结果是双方最佳的落子序列, 这个结果反映了对棋局进程的展望. 在一般情况下, 28步落子序列展望已经超过围棋职业选手的搜索深度, 但是在一些特殊复杂的情况下, 28步的搜索深度仍显不足.

3.3 AlphaGo的核心技术突破 (Core technology breakthrough of AlphaGo)

AlphaGo的核心技术是使用深度神经网络获得围棋棋感直觉, 而在这之中增强型深度学习获得胜负棋感直觉尤为关键. AlphaGo中使用的蒙特卡罗树搜索方法已经是成熟的技术, 但是将深度神经网络与蒙特卡罗树搜索结合在一起也是一种新颖的方法.

AlphaGo 使用的对弈硬件配置普通, 但是整个项目训练的数量大、配置昂贵, 所需的时间也非常长.

4 总结(Summary)

围棋是人工智能的重要目标, 也是衡量人工智能进步的标尺. 目标促使我们寻找人工智能的途径, 而标尺帮助我们衡量人工智能的水平. 围棋的突破表示我们正处于人工智能爆发的重大转折点, 未来几年数据驱动的人工通用智能会井喷式地发展. 未来人工智能的核心也将围绕着3个方面发展, 直觉获取、搜索验证以及优化决策.

所谓直觉就是不经过思考过程, 很快就能出现直接想法、感觉、信念或者偏好, 而对直觉获取最好的理解就是通过深度神经网络和大数据的训练而获得结果. 验证是指为直觉建立真实性、准确性和可靠性的过程, 它是核实直觉不存在偏差的一个充分条件. 由于廉价并行计算和大数据的支持, 直觉可以通过搜索计算来验证, 从而确保准确性. 而优化决策则可用于人类生活中所能面临的方方面面的问题: 照片上的肿瘤是否是良性的、手机里的股票是否继续持有、驾车到交通灯是否继续执行等等. 但是优化决策中优化选择的实现依赖于直觉获取和搜索验证.

乔布斯曾说过:"从机械的角度来说, 秃鹰是地球上效率最高的动物, 但是骑上自行车后人便能把秃鹰甩在身后. 电脑是我们发明出来的非凡工具, 它相当于我们大脑的自行车." 而在人工智能极速发展的今天, 我们可以毫无疑问地肯定人工智能是我们大脑的自行车.

自AlphaGo击败李世石后, 全球各国研究人工智能的团队都活跃了起来, 日本宣布"Deep Zen Go"项目启动, 要超越Google的AlphaGo, 打造世界最强围棋软件, 美国Facebook公司也在采用了和AlphaGo相似的人工智能技术, 研发DarkForest围棋软件, 而中国人工智能将如何应对这些挑战? 让我们拭目以待.

参考文献(References):

[1] SILVER D, HUANG A, MADDISON C, et al. Mastering the game of Go with deep neural networks and tree search [J]. *Nature*, 2016, 529(7587): 484 – 489.

[2] CAFLISCH R E. Monte Carlo and quasi-Monte Carlo methods [M] //*Acta Numerica*. Cambridge, UK: Cambridge University Press, 1998: 1 – 49.

[3] THRUN S. Monte Carlo POMDPs [C] //*Advances in Neural Information Processing Systems*. Denver: MIT Press, 1999, 12: 1064 – 1070.

[4] LITTMAN M L. Reinforcement learning improves behaviour from evaluative feedback [J]. *Nature*, 2015, 521(7553): 445 – 451.

[5] TIAN Yuandong. A simple analysis of AlphaGo [J]. *Acta Automatica*, 2016, 42(5): 671 – 675.

[6] ZHAO Dongbin, SHAO Kun, ZHU Yuanheng, et al. Review of deep reinforcement learning and discussions on the development of computer Go [J]. *Control Theory & Applications*, 2016, 33(6): 701 – 717. (赵冬斌, 邵坤, 朱圆恒, 等. 深度强化学习综述: 兼论计算机围棋的发展 [J]. 控制理论与应用, 2016, 33(6): 701 – 717.)

作者简介:

刘知青　(1966–), 男, 清华大学计算机科学与技术系工科学士, 纽约大学理学硕士、哲学博士, 曾任美国AT&T公司贝尔实验室高级研究员、美国印第安纳大学终身制助理教授、北京邮电大学计算机围棋研究所所长等, 曾参与美国DARPA、NIH的研究项目, 主持开发了国内最强大的计算机围棋软件"本手围棋", 多次获得国内计算机围棋锦标赛冠军, 成为国内多家学术期刊编辑, 在国内外发表学术论文数10篇, 与李文峰合著《现代计算机围棋技术》;

吴修竹　(1992–), 女, 硕士, 研究方向为计算智能, E-mail: linxiaomo1992@163.com.