DOI:10.12132/ISSN.1673-5048.2019.0070

AlphaZero 原理与启示

唐　川，陶业荣*，麻曰亮

（中国洛阳电子装备试验中心，河南 洛阳 471000）

摘　要：近几年计算机围棋的成功引发了又一轮的人工智能热潮，从计算机围棋中发展出来的 AlphaZero 框架成功地应用在了其他完全信息条件下的二人有限零和博弈问题，进而展示出了深度学习和强化学习在智能决策领域的优异性能。本文首先介绍了 AlphaZero 框架中三个核心技术：深度学习、强化学习以及蒙特卡罗树搜索。然后详细说明了 AlphaZero 框架两个关键阶段——AlphaGo 和 AlphaGo Zero——的基本原理。最后，本文对 AlphaZero 框架提出了自己的思考，并基于对 AlphaZero 原理的剖析讨论了其对军事决策智能化的启示。

关键词：深度学习；强化学习；蒙特卡罗树搜索；AlphaZero；军事决策智能化

中图分类号：TJ 760　　文献标识码：A　　文章编号：1673-5048（2020）XX-XXXX-XX

0　引　　言

象棋、围棋、日本将棋等棋类博弈游戏一直是人工智能关注和研究的热门领域，一般将其抽象为完全信息条件下的二人有限零和博弈模型，该模型的含义是指在任意时刻，双方玩家（“二人”）都知道游戏的全部状态（“完全信息”），并且有限步（“有限”）之后游戏的结果非胜即负（“零和”），至多加上平局。而双方在游戏中对抗，目的是自己获得尽可能好的结果（“博弈”）。1997 年发布的国际象棋人工智能“深蓝”轰动一时。“深蓝”依赖强大的计算能力对国际象棋的所有状态空间进行穷尽式暴力搜索，用确定性算法求解国际象棋的复杂决策问题，体现了一种“机器思维”，然而这一方法并不能适用于围棋。围棋复杂的盘面局势评估和巨大的状态搜索空间，成为学界面临的巨大挑战。国际象棋每一步可供选择的走法平均为 35 种（即空间搜索宽度约为 35），每盘棋平均需要 80 步决出胜负（即空间搜索深度约为 80），所以如果要遍历完整下棋过程，整个搜索空间大约为 35^{80}；而对于围棋，其搜索宽度平均为 250，搜索深度平均为 150，整个搜索空间为 250^{150}，超过了可观测宇宙中的原子数目，因此无法采用暴力搜索方式。

为实现高智能的计算机围棋，早期的研究通过专家系统和模糊匹配来控制搜索空间规模，但一方面算法效果一般，此外当时的计算资源和硬件能力也捉襟见肘，所以效果并不明显。2006 年，蒙特卡罗树搜索（MCTS）的应用引领着计算机围棋进入了新的阶段[1]。现代计算机围棋的主要算法是基于蒙特卡罗树的优化搜索进行决策，该算法推动了计算机围棋的发展，对战规模从九路（9×9 的棋盘）围棋发展到 19 路(19×19 的全尺寸棋盘)围棋，棋力从普通段位水平发展到业余六七段的实力，但是业界依然认为计算机围棋需要数十年的时间才能达到职业水准。然而，随着深度学习和蒙特卡罗树搜索方法的结合，这一观点开始受到挑战。2015 年，Facebook 人工智能研究院的 Tian 结合深度卷积神经网络和蒙特卡罗树搜索开发出的计算机围棋 Dark Forest 表现出了与人类相似的下棋风格和惊人的实力[2]。2016 年，基于深度强化学习和蒙特卡罗树搜索的 AlphaGo 击败了人类顶尖职业棋手，引起了全世界的关注[3]。2017 年，Deep Mind 在《Nature》上公布了最新版 AlphaGo 论文，介绍了迄今为止最强的围棋人工智能 AlphaGo Zero[4]。AlphaGo Zero 不需要人类专家知识，只使用纯粹的深度强化学习技术和蒙特卡罗树搜索，经过 3 天自我对弈就以 100 比 0 的成绩完败了上一版本的 AlphaGo。AlphaGo Zero 证明了深度强化学习的强大能力，也推动了人

收稿日期：2019-04-25

作者简介：唐川（1988-），男，河南开封人，博士，助理研究员，研究方向是人工智能芯片设计。
通讯作者：陶业荣，（1976-），男，河南太康人，学士，高级工程师，研究方向人工智能技术试验与评估。E-mail：taoyerong@126.com

引用格式：唐川，陶业荣，麻曰亮. AlphaZero 原理与启示[J]. 航空兵器，2020，27
Tang Chuan, Tao Yerong, Ma Yueliang. The Principle and Enlightenment of AlphaZero[J]. Aero Weaponry, 2020, 27

工智能领域的进一步发展。2017 年年底，Deep Mind 使用了完全类似 AlphaGo Zero 的同一套算法框架，在完全没有人类对弈数据的情况下，解决了诸多困难的棋类问题，在国际象棋和将棋方面轻松碾压当前最强的算法框架[5]，证明了其深度强化学习加蒙特卡罗树搜索的方法不仅适用于围棋，也适用于大多人类可以掌控玩法的棋类，乃至适用于所有完全信息条件下的二人有限零和博弈问题，并将这一框架命名为 AlphaZero。

以 AlphaZero 为代表和标志的技术突破，预示着一种具有直觉、认知和自我进化能力的新的人工智能时代的到来，也预示着智能化决策、智能化武器装备的发展以及智能化战争的到来。针对 AlphaZero 智能化方法框架的研究可以启发人工智能在智能指挥决策、智能化武器装备等军事领域的应用，为解决复杂军事指挥和智能决策问题指明方向[[6]-[9]]。特别是航空领域，一直以来都是前沿技术实践应用的倡导者和先锋，加上航空设备工作环境的特殊性，许多任务无法人工完成，对智能化具有更强的依赖性。本文将对 AlphaZero 框架的两个主要发展阶段 AlphaGo 和 AlphaGo Zero 的技术原理进行深入剖析，并以通俗易懂的类比方式进行说明，最后基于对 AlphaZero 的剖析，谈谈对于 AlphaZero 以及军事决策智能化的思考与启示。

1 核心技术

1.1 深度学习

深度学习起源于传统的神经网络，是基于深度神经网络的一种学习方法，是机器学习的一个特定分支。它通过建立多个隐含层模拟人脑分析学习的机制，吸收大量数据的经验建立规则（网络参数），实现特征的自主学习[10]，主要适用于无法编制程序、需求经常改变、有大量数据且无需精确求解的一类问题。神经网络早期的灵感来源是神经科学，通过模仿大脑的神经元之间传递、处理信息的模式进行学习。但随着技术的发展，神经网络以及深度神经网络都不再以神经科学为主要指导了。基于神经网络的机器学习方法的组成主要包括输入、神经元单元、神经网络、成本函数和算法。

传统神经网络一般层数较少，深度神经网络可以理解为有多层神经元的神经网络，它的基本组成和神经网络基本一致。

深度学习能够从原始数据中逐层提炼出更高级更抽象的特征属性，每层神经元的处理机制可看作是在对输入信号进行逐层加工，从而把初始的、与输出目标之间联系不太密切的输入表示转化成与输出目标联系更密切的表示，使得传统神经网络仅基于最后一层输出映射难以完成的任务成为可能[11]。换言之，通过多层处理，逐渐将初始的“低层”特征表示转化为“高层”特征表示后，即可用“简单模型”完成复杂的学习任务，而且网络层数越多，意味着能够提取到的特征越丰富，越抽象，越具有语义特征。

为了便于理解，以傅里叶展开与深度学习模型及其训练做一个类比（两者思想上有相似之处，可以借鉴，但目的、方法、机理等方面还是存在明显差别），傅里叶展开利用一系列正弦函数项来拟合任意函数（深度学习的任意目标任务），其中每一项都只是一个周期不同的简单正弦函数（深度神经网络中的一个网络层），使用多个正弦函数项的加权（类比于网络权重）叠加（网络间的连接）去拟合目标函数。使用的正弦函数项越多，傅里叶展开的描述能力越强越精细（网络层数越多深度神经网络的表达能力越强），叠加后的函数越接近目标函数。当项数趋向无穷时，傅里叶展开就可以精确表示任意函数，而进行傅里叶展开的过程就是计算每一项的系数（训练过程中的权重调节），最终计算得到的展开式就是目标函数的一个近似表达（权重调节好的深度神经网络就可以近似处理目标任务）。

1.2 强化学习

强化学习（Reinforcement Learning，RL）又叫做增强学习，是近年来机器学习和智能控制领域的主要方法之一，它关注的是智能体如何在目标环境中采取一系列行为从而获得最大的价值回报。强化学习是机器学习中一个非常活跃且有趣的领域，相比其他学习方法，强化学习更接近生物学习的本质，因此有望获得更高的智能，这一点在棋类游戏中已经得到体现。

更抽象地，可以对强化学习所要解决的问题进行如下描述：在目标环境（E）中存在多种状态（S，状态空间集合）阶段，通过行动（A，动作空间集合）可以使得状态发生转移（P，状态转移的条件概率矩阵），状态的变迁会带来奖励（R，价值函数），而目标就是学得一种策略（π）使奖励最大化[12]。因此强化学习中的目标环境对应一个四元组 $E=<S, A, P, R>$，目标就是学会策略 π。策略 π 可以表示成一个函数，如果 π 属于确定性策略，其可以表示为 $\pi: S \rightarrow A$，即输入当前状态 $s \in S$ 策略 π 输出自己建议的动作 $a \in A$；如果 π 属于概率性策略，其可以表示为 $S \times A \rightarrow R$，即已知当前状态 $s \in S$ 时，策略输出采用动作 $a \in A$ 的可能性是多少（通常是 0~1 的实数）。

通常情况，根据环境四元组 $E=<S，A，P，R>$ 是否完全已知，强化学习可以分为有模型学习和免模型学习。

有模型学习表示四元组 $E=<S，A，P，R>$已知，即机器可以对环境进行完整建模，能在机器内部模拟出与环境相同或近似的状况，可以通过模拟推算计算出来不同策略带来的价值回报，通过不断的模拟计算，总能找出一个（可能存在多个最优策略）最优的策略来得到最大的回报，因此在模型已知时强化学习任务能够归结为基于动态规划的寻优问题。

在实际的强化学习任务中，环境中状态的转移概率 P、价值函数 R 通常很难得到，甚至很难知道环境中一共有多少状态。因此将学习算法不依赖于环境建模的方法称为免模型学习，这比有模型学习更困难也更实用。由于模型未知，没法通过计算的方式得到准确的最终奖励，并以此来评估当前策略的好坏；因此只能通过在环境中执行选择的动作来观察状态的转移情况以及得到的奖励，并利用蒙特卡罗思想，用多次“采样”的平均值来近似表示实际的价值函数，同时在多次“采样”过程中，发现存在的状态集合和状态之间的转移关系。换言之，通过不断的尝试，去近似估计未知参数；然后再通过对不同策略的尝试与评估，总结归纳并优化策略。

然而在实际任务处理过程中，由于资源、实时性、处理能力等方面的限制，“尝试”的机会往往是有限的，在这有限的尝试中，既需要通过探索去发现更多的选择并提高参数估计的准确性，另一方面还希望利用现有的最佳策略尽可能得到更多的奖励（类似于有限次数多摇臂老虎机赌博问题）。因此如何在探索和利用之间进行权衡是强化学习的一个关键任务。可以看出探索的过程就是一个“试错”的过程，如果机器有一定的经验，可以有选择性地探索，加快探索效率；如果机器没有任何经验，也可以从随机开始，在不断试错的过程中成长，基于成长后的策略进行选择性的探索同样可以加快探索效率。所以强化学习可以不依赖任何人类知识而学习到目标知识，类似于人类探索未知事物的学习方式。

1.3 蒙特卡罗树搜索

蒙特卡罗树搜索(Monte Carlo Tree Search)，一种通过随机游戏推演来逐渐建立一棵不对称搜索树的过程，它是人工智能领域中寻找最优决策的一种方法。

蒙特卡罗树搜索采用树状结构表征围棋博弈问题，初始阶段棋盘为空，这构成博弈树的根节点，此时可以选择的动作有 361 种，因此根节点就有 361 个分支，随机选择一个分支，并以此类推可以使得分支逐步生长，直到终结点（Terminal Node）游戏结束，这一过程就是一次遍历过程。如果通过足够多次的尝试将每一个节点都遍历到，就能生长出一颗完整的博弈树。基于这棵完整的博弈树，可以在任何状态下规划下一步的最优决策以走向胜利（在完整决策树已知的情况下，博弈游戏的胜负完全由猜先决定，即先手必胜或先手必败）。

若假设完整的博弈树已知，接下来就要规划下棋的策略。在规划的过程中，每一状态的动作选择依赖于对动作的价值评估或者说胜负评估，不仅要规划自己的策略，同时还要考虑对手的决策。在规划过程中，不确定对手的决策能力，但为了使得决策规划更具实用性，只能假定对手会全力追求胜利，因此规划的过程是一个基于价值评估的极小极大交替选择过程，也可以说是价值评估的传递过程[13]。

然而，这一过程的探索规模随搜索宽度和深度的增加成指数速度扩大，对于围棋这样的游戏，遍历得到整个博弈树是不现实的。人类棋手并不会对全部空间进行暴力搜索，而是先通过宏观的“势”，或者是所谓的“棋感”选出几个感觉较好的落子方案，再对每个方案进行“深思熟虑”的多步推演，然后比较得出最好的落子位置。人类棋手凭经验和“直觉”确定候选方案，是在降低搜索的宽度，一些明显不好的落子方案不再进行深入的搜索。人类棋手的“深思熟虑”也不是推演到棋局的最后一步，往往是推演几步最多十几步后就对盘面进行综合评估判断局势好坏。这种综合评估，降低了搜索的“深度”。

因此，在大空间博弈问题中，设计者往往采用低复杂度搜索算法，如蒙特卡罗树搜索算法。蒙特卡罗树搜索减少了搜索的宽度和深度，并在有限的遍历过程中，寻找到最有潜力的下一步行动，即形成决策。其主要思想是：在宽度方面，通过一定次数的遍历后，部分分支会表现出更高的胜率，将有限的遍历集中在这类更有潜力的分支上，以减少搜索的宽度；但与此同时，基于潜力的倾向性遍历会增加纵深方向单步搜索的计算复杂度，使得深度方向的搜索时间更长；因此在深度方面，为了避免复杂搜索算法导致的搜索代价增加，可以在搜索到某一中间节点时停止搜索，用基于简单算法（常见算法是均匀随机算法）的模拟过程执行到终结点，又或者在停止搜索后利用评估函数直接预测当前中间节点盘面的胜负。

蒙特卡罗树搜索的主要概念是搜索，即沿着博弈树向下的一组循环遍历过程。单次遍历的路径会

从根节点（当前博弈状态）延伸到没有完全展开的节点。未完全展开的节点意味着其子节点至少有一个未访问到。遇到未完全展开的节点时，它的一个未访问子节点将会作为单次模拟的根节点并推演到终盘，随后模拟的结果将会反向传播回当前树的根节点并更新博弈树的节点统计数据。一旦循环遍历过程受限于时间或算力而终止，下一步行动将基于收集到的统计数据进行决策。因此可以将蒙特卡罗树搜索划分为选择、扩展、模拟评估和反向更新四个步骤，如图 1[14]。

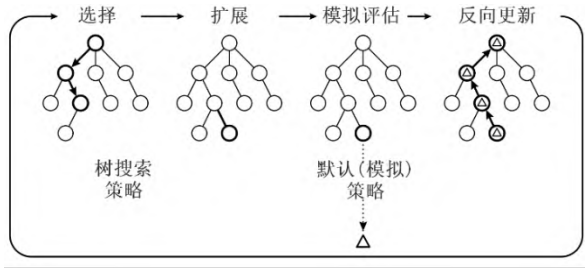

图 1 蒙特卡罗树搜索步骤

Fig. 1 Procedures of Monte Carlo Tree Search

由于蒙特卡罗树搜索采用了倾向性搜索算法以减少不必要的探索过程，但是这也增加了陷入局部最优的可能性，因此与强化学习类似，蒙特卡罗树搜索算法也存在探索和利用的权衡问题。

2 AlphaGo

2.1 AlphaGo 的结构组成

AlphaGo 由监督学习策略网络（Supervised Learning of Policy Networks，简称 SL 策略网络）、强化学习策略网络（Reinforcement Learning of Policy Networks，简称 RL 策略网络）、快速走棋策略网络（Rollout Policy Networks）和价值网络（Value Networks）组成，其中策略网络用于模拟人类的“棋感”，而价值网络用于模拟人类对盘面的综合评估，即盘面胜负评估。

SL 策略网络是一个 13 层的深度卷积神经网络，该网络的输入是棋盘特征，也叫做盘面，其表现形式是一个 19×19×48 二值平面，19×19 是围棋的棋盘布局，48 个平面对应不同的盘面特征信息[15]（如棋子颜色、轮次、气、打吃数目等）。输入经过 13 层深度卷积神经网络的逐层理解和分析，最终输出一个走棋策略$p_\sigma(a|s)$，表示当前状态 s 下所有合法动作 a 的概率分布，其中 σ 表示该网络的权重参数也表示策略的名称。SL 策略网络的决策计算速度是 3 ms/步，即输入当前盘面后，3 ms 可得到下一步的决策。该策略网络主要用于在蒙特卡罗树搜索的选择阶段提供先验概率信息。

RL 策略网络本质上是以训练好的 SL 策略网络为初始状态，通过强化学习过程优化 SL 策略网络中的网络权重参数后得到的新策略网络。因此其结构以及输入、输出的形式同 SL 策略网络一样，输出的 RL 策略标记为$p_\rho(\mathrm{a}|s)$，该策略网络用于产生自对弈棋谱供价值网络进行训练。

快速走棋策略网络的学习模型是一个简单线性模型，输入也更简单且包含了一些人工加工的特征，这使得快速走棋策略的输出$p_\pi(\mathrm{a}|s)$的棋力较差，但其计算速度非常快，每步决策仅需要 2 μs，因此该策略网络用于在蒙特卡罗树搜索的模拟评估阶段执行快速模拟。

价值网络依然是一个 13 层的深度卷积神经网络，其输入和 SL 以及 RL 策略网络一样，都是当前的盘面信息，区别仅在于其输出为对当前盘面结局（输赢的期望）的预测$v_\theta(s)$，价值网络的输出同样用于蒙特卡罗树搜索的模拟评估阶段，以直接提供对叶节点盘面的结局预测。

可以看出 SL 策略网络、RL 策略网络以及价值网络具有相似甚至相同的结构，这是因为深度神经网络的不同层的目标是为了提取输入信息的隐含特征，这些隐含特征对于预测输赢或者决策下子应该都具有相关性。由于策略网络与价值网络的学习目标不同，在最后的输出映射阶段两者的结构会有区别。另外，即使是目标相同的 SL 策略网络和 RL 策略网络，由于训练方法和数据的差别，同样的隐藏特征在对结果的影响力上会表现出差别，这一差别可以通过网络中权重的差别体现出来，进而导致产生了不同的策略。

2.2 离线训练过程

SL 策略网络的训练数据来自于棋圣堂围棋服务器（Kiseido Go server，KGS）上 3 000 万个专业棋手对弈棋谱的落子数据。基于专业棋手的棋谱数据，采用随机梯度上升法更新网络参数，以提高模仿的准确性。因此 SL 策略网络学习的目标是模拟专业棋手的下棋风格，最终 SL 策略模拟的准确度达到了 55.7%。

快速走棋网络使用与 SL 策略网络相同的训练数据，只是提取的数据特征较简单，并且使用线性回归方法进行训练。快速走棋网络在牺牲了部分准确度的情况下极大地提高了走棋的速率。快速走棋网络与 SL 策略网络一样属于监督学习，类似于人类学习过程中背棋谱的学习阶段。

RL 策略网络采用强化学习方法，因此训练时不需要额外的训练数据。第一步，先使用 SL 策略网络对 RL 策略网络进行初始化；第二步，将当前的 RL 策略网络与对手池（在第四步中生成）中之

前的某个随机版本进行对局，得到棋局结果（输赢）；第三步，根据棋局结果利用强化学习中的策略梯度算法更新网络权重以最大化期望结果（赢）；第四步，每 500 次迭代就复制当前网络参数到对手池中用于第二步的随机版本对局。重复上述四个步骤直到参数收敛稳定既得到最终的 RL 策略网络。其中第四步记录的 RL 策略网络的历史版本是为了防止训练过程中出现过拟合现象，第二步的对局本质上是和“历史自我”进行的“自我对弈”。同时也能看出 RL 策略网络训练追求的目标是胜利，与 SL 策略网络追求的目标（尽可能的模仿专业棋手）是不同的，两者对弈结果统计，RL 策略网络的胜率达到 80%。类比人类学习过程，RL 策略网络的训练近似于有一定基础的棋手通过与高手对弈不断提高棋力，追求制胜之道。

价值网络的训练数据来自 RL 策略网络“自我对弈”过程中产生的棋谱，它根据产生棋谱的最终胜负结果，使用随机梯度下降法来最小化预测值 $v_\theta(s)$与实际对弈结果 z（赢为+1，输为-1）间的差值。训练好的价值网络可以对棋局进行评估，预测当前盘面的胜负期望，也即胜负的概率。类比人类棋手，该训练过程近似于观摩大量高手的比赛后使自身具备了丰富的经验，结合当前盘面和过往经验能预测棋局的胜负。

2.3 在线对弈过程

AlphaGo 在线对弈过程以蒙特卡罗树搜索为主要框架，并结合 SL 策略网络、快速走棋网络和价值网络以提高蒙特卡罗树搜索的效率。在介绍对弈过程前，首先介绍下每个蒙特卡罗树搜索节点（即盘面 s）的统计信息。每一个节点 s 包含多条边连接着 s 与其子节点，每一条边对应一个合法的状态-动作对(s,a)，每一条边对应一个六元组统计信息：$\{P(s,a), N_v(s,a), N_r(s,a), W_v(s,a), W_r(s,a), Q(s,a)\}$，并将其记录在节点 s 处。$P(s,a)$是树搜索策略中需要使用的先验概率，在 AlphaGo 中 $P(s,a)$是 SL 策略网络的输出。$N_v(s,a)$是遍历经过该边并利用价值网络评估的次数，而 $N_r(s,a)$则是遍历经过该边并利用快速走棋网络评估的次数。$W_v(s,a)$表示 $N_v(s,a)$次价值网络评估结果的累加值，$W_r(s,a)$表示 $N_r(s,a)$次快速走棋评估结果的累加值。所以 $W_v(s,a)/N_v(s,a)$和 $W_r(s,a)/N_r(s,a)$分别表示价值网络和快速走棋网络模拟对盘面胜负的平均估计。$Q(s,a)$是价值网络和快速走棋网络评估均值的加权平均，表示对应边的联合平均胜负估值，如式（1），除了 $P(s,a)$初始化为 SL 策略网络的输出，其余统计信息初始化值为零。

$$Q(s,a)=(1-\lambda)\frac{W_v(s,\mathrm{a})}{N_v(s,\mathrm{a})}+\lambda\frac{W_r(s,\mathrm{a})}{N_r(s,\mathrm{a})} \quad (1)$$

在线对弈过程主要包括四个步骤，如图 2。

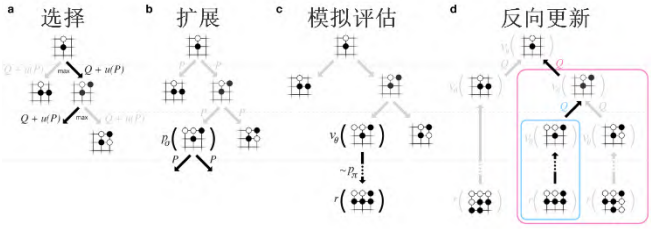

图 2 AlphaGo 在线对弈步骤[3]

Fig. 2 Procedures of AlphaGo online Game

选择：选择阶段，从根节点开始执行树搜索策略进行分支选择，搜索执行到叶节点 L 为止。树搜索策略基于式（2）进行动作决策，其中 $u(s_t,a)$是控制参数用于鼓励探索，如式（3）：

$$\mathrm{a}_t=\arg\max_{\mathrm{a}}\big(Q(s_t,\mathrm{a})+u(s_t,\mathrm{a})\big), t<L \quad (2)$$

$$u(s_t,\mathrm{a})=C\frac{P(s,\mathrm{a})}{1+N_r(s,\mathrm{a})} \quad (3)$$

式（3）中 C 可以简单认为是常数，从式（3）能看出，当(s,a)是新展开边时，$N_r(s,a)$和 $Q(s,a)$均为 0，此时树搜索决策主要依赖于先验概率 $P(s,a)$即 SL 策略网络的策略。当经过几次模拟后，树搜索决策由 $Q(s,a)$与 $u(s,a)$共同决定，由于 $u(s,a)$随分母部分 $1+N_r(s,a)$的增大而减小，使得决策倾向于模拟次数少的分支，进而鼓励了探索。当模拟次数进一步增多时，遍历模拟得到的 $Q(s,a)$值越来越准确，而 $u(s,a)$由于分母的增大趋向于 0，此时决策主要依赖于 $Q(s,a)$值。通俗地讲，在选择模拟阶段，为了减少搜索宽度，AlphaGo 倾向于胜率高的分支，但由于模拟次数少的时候胜负估计不够准确，所以基于以往的经验进行指导；同时为了鼓励探索避免陷入局部最优策略，AlphaGo 鼓励探索模拟次数少的分支，最终，伴随胜负估计的逐步精确，后续决策基本仅取决于模拟的结果。

扩展：扩展阶段会使博弈树生长出新的叶节点。在 AlphaGo 中，当某条边的访问次数大于阈值 n_{thr}（动态阈值，默认 40）后，该边指向的节点 s′ 将被加入到博弈树中，并进行统计信息初始化。

模拟评估：当到达叶节点 s_L 时，若 s_L 之前没有使用价值网络评估过，则将 s_L 节点加入价值网络评估队列以得到$v_\theta(s_L)$；若 s_L 之前访问并使用价值网络评估过，则不再进行价值网络评估，即每个节点只进行一次价值网络评估。于此同时，快速走棋网络则以 s_L 节点为起点，基于快速走棋策略$(a_t\sim p_\pi(\cdot|s_t), t>L)$模拟执行到终盘，得到最终的胜负结果 z_T，T 为终盘时刻。

反向更新：由于价值网络在搜索到叶节点 s_L 就开始执行评估，所以价值网络评估完成后就会异步地对遍历过程 $t<L$ 的每一步经过的边进行统计信息

更新，如式（4）。另一方面，快速走棋网络完成模拟后，即可更新 $t<L$ 过程中的每一条边，如式（5）。同时，每条边的 $Q(s_t,a_t)$值也会按照式（1）进行更新。

$$\begin{aligned} N_v(s_t,\mathrm{a}_t) &= N_v(s_t,\mathrm{a}_t)+1 \\ W_v(s_t,\mathrm{a}_t) &= W_v(s_t,\mathrm{a}_t)+v_\theta(s_L) \end{aligned} \tag{4}$$

$$\begin{aligned} N_r(s_t,\mathrm{a}_t) &= N_r(s_t,\mathrm{a}_t)+1 \\ W_r(s_t,\mathrm{a}_t) &= W_r(s_t,\mathrm{a}_t)+\mathrm{z}_T \end{aligned} \tag{5}$$

反复进行上述四步过程达到一定次数后搜索完成，算法选取从根节点出发访问次数最多的那条边落子，完成单步落子决策。该条边对应的子树也将保留下来作为下一步棋决策的初始状态，然后重复执行蒙特卡罗树搜索过程进行单步决策，最终走到终盘完成比赛。

AlphaGo 算法的训练和对弈流程图如图 3。上半部分表示离线训练的过程，下半部分是基于蒙特卡罗树搜索的在线对弈过程。

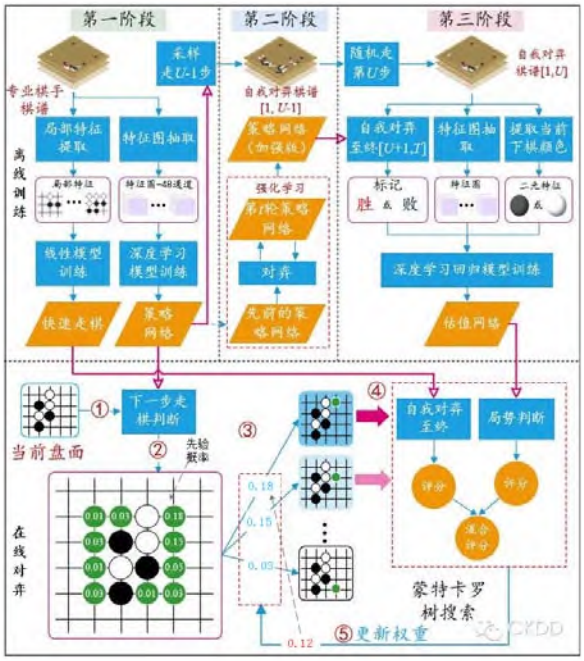

图 3 AlphaGo 训练和对弈流程图[16]

Fig. 3 Flow-process diagram of the train and game of AlphaGo

2.4 AlphaGo 中的特殊现象

- **策略网络选择**

基于强化学习的 RL 策略网络在与 SL 策略网络对弈时，胜率可以达到 80%，然而在线对弈过程中 RL 网络并没有直接参与决策，仅作为了价值网络训练数据的提供者。为什么不采用 RL 策略网络的输出作为蒙特卡罗树搜索的先验概率？对于这一问题，最直接的原因是：实际实验中，其他同等条件下，基于 SL 策略网络的对弈效果更好。对于这一特别现象的解释，有一种说法是 SL 策略的探索更具有多样性。SL 策略在模仿专业棋手的棋风时，学到了“大局棋”概念，即跳出当前的局部布局在其他位置提前进行布局的一种策略。这使得 SL 策略网络棋风更具有多样性，而 RL 策略网络在这一方面不如 SL 策略网络。

- **价值网络不使用人类数据训练**

AlphaGo 的价值网络用于评估盘面的胜负，然而供其训练的数据是强化学习网络自己产生的，而不是直接使用专业棋谱。对于这一问题，一个直接原因是 RL 策略网络是 3 个策略网络中的最强策略。但是从另一个角度考虑，SL 策略网络不如 RL 策略网络是因为 SL 策略网络模拟专业棋手的相似度只有 55%，如果直接使用专业棋谱数据（100%相似度）是否会达到更好的效果？对此学者认为人类数据其实并不适合价值评估。很多人类的棋局都是因为中间偶然的失误导致了全盘覆灭（所谓“一着不慎满盘皆输”），其中的偶然性非常大，盘面的估值瞬息万变，所以棋局的结果离理想的估值差距较大。不如让 AlphaGo 培养自己的“感觉”，自己的“胜负观”，而不是轻易被人类棋局的胜负所左右。

- **价值网络与快速走棋网络**

按常理揣测，基于强大的 RL 策略网络训练出来的价值网络，在评估方面应该超越快速走棋网络。然而，实际实验当中，同等条件下单纯基于价值网络评估的效果并不如单纯基于快速走棋网络评估的效果，而两者的结合使得效果有进一步的飞跃。对此现象，可以理解为 AlphaGo 自己产生的“胜负观”和人类经验形成的“胜负观”具有一定的互补作用。从后面对 AlphaGo Zero 的介绍可以推测，两者确实具有互补效果，而价值网络的不足主要是由于网络本身的表达能力不够。

- **自暴自弃**

当 AlphaGo 在判断自己胜算不足的时候就会自暴自弃，走棋具有随机性。笔者推测，在胜算不足时，各个分支的 $Q(s,a)$值都不高（必输情况下所有 $Q(s,a)$值均为零），此时为了增加探索性的一些扰动机制会使得基于 $Q(s,a)$值的倾向性搜索失去作用，搜索过程呈现扰动机制的随机性。对于此问题，有人建议在胜算不足的情况下，将模拟对弈的对手替换为棋力较弱的模型，以保持系统的“战斗意志”。但这种方式间接将胜利寄托在了对方的失误。

- **“神之一手”**

在 AlphaGo 和李世石比赛的第四盘中，李世石第 78 手成为了棋局的点睛之笔，使其获得了比赛的唯一一场胜利，这一手棋被称为“神之一手”。赛后，AlphaGo 的设计团队多次分析实战数据，结论都是“人类棋手几乎不会下的一手”，“人类棋手下这步棋的概率不到万分之一”。由于基于人类训

练数据产生的 SL 策略网络的相似度仅有 55%，所以我们无法评论 AlphaGo 忽略这“万分之一”可能性的原因是自身的不足（相似度不够高）还是这一步真的出乎意料（其他专业棋手也想不到）。但不管是哪种原因，究其本质还是在探索和利用的天平太偏向于利用，忽视了小概率走法。

3 AlphaGo Zero

AlphaGo Zero 是 AlphaZero 框架围棋系列的最后一款产品，是 AlphaZero 框架设计思路的具体表现形式，。因此通过解析 AlphaGo Zero 就可以了解 AlphaZero 框架。AlphaGo Zero 摆脱了人类知识的约束，能够在没有人类知识做指导和训练的条件下学得围棋的下法和人类棋谱中的“定式”，并且发现人类未知的新“定式”，创作了知识，也印证了强化学习的强大。

3.1 AlphaGo 的不足

（1）结构复杂

AlphaGo 由 4 个网络构成，3 个策略网络，1 个价值网络。策略网络功能相同，却无法互相替代。价值网络和快速走棋网络用途相同，但功能互补无法舍弃。这既浪费了有限的平台算力（间接影响了棋力），也暗示了 AlphaGo 的网络并不完美。

（2）人类经验的羁绊

“尽信书，不如无书。”以往的人类经验可以减少搜索空间，并使得算法快速稳定的地收敛到更优策略，但同时它也局限了我们的探索范围。AlphaGo 中的强化学习网络就尝试摆脱人类经验的束缚，但其初始状态仍然是人类经验的体现。

（3）RL 策略网络仍然存在性能瓶颈

强化学习利用策略模拟、策略改进、策略再模拟的迭代过程来优化网络结构，其效果固然强大，但策略改进的效率决定了其最终效果，目前 AlphaGo 简单的通过自我对弈还无法达到最佳的效果。好比两个幼年孩童不断的随意对弈真的就能达到职业 9 段水平么？即使达到了，需要多长的时间？因此从现有策略如何提高是一个关键问题。

（4）探索与利用

探索与利用的权衡对于强化学习以及蒙特卡罗树搜索方法的性能都具有显著的影响。尽管 AlphaGo 中加入了丰富探索多样性的机制，但目前并没有理论可以证明怎样的平衡才能达到最佳。式（2）中的红利 $u(s,a)$虽然鼓励探索，但是式（2）本身属于确定性决策方式（决策时动作选择不是概率性的采样），使得某一分支占优后很难跳出去探索其他分支。“神之一手”的出现进一步印证了 AlphaGo 探索不足的问题。

以上不足为 AlphaGo Zero 的设计指引了方向，将在 AlphaGo Zero 的设计思想中看到针对以上问题的处理。

3.2 AlphaGo Zero 的结构组成

AlphaGo Zero 将原先两个结构独立的策略网络（SL 策略网络和快速走棋网络）和价值网络合为一体，合并成一个深度神经网络（改善了 AlphaGo 的不足（1））。在该神经网络中，从输入层到中间层的权重是完全共享的（AlphaGo 中 SL 策略网络和价值网络结构共享，权重独立），最后的输出阶段分成了策略函数输出和价值函数输出。此外，在 AlphaGo 中采用的 13 个卷积层网络被替换为了 19（扩展版为 39）个残差模块（或叫残差网络），形成了深度残差神经网络 $f_\theta(s)$，它通过实现更深的神经网络以提取到更丰富且更抽象的输入特征并具有了更强的表达能力。AlphaGo Zero 的网络结构示意图如图 4。

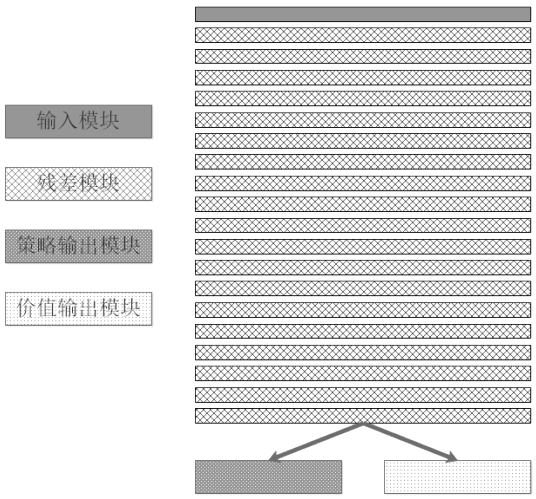

图 4 AlphaGo Zero 网络结构示意图

Fig. 4 Network structure diagram of AlphaGo Zero

深度残差网络输入的盘面状态 s 是 19×19×17 的二值平面，相比于 AlphaGo 的策略网络更加简洁。主要有 3 部分内容：己方棋面，对方棋面，当前执棋颜色。输入信息经过深度残差网络的处理，得到了盘面的深层次特征，基于这些特征分别利用策略输出模块和价值输出模块得到下棋策略 $\boldsymbol{p}$ 和盘面胜负评估 v，其中 $\boldsymbol{p}$ 为 361 维向量，表明当前盘面下，不同动作选择的概率（在 AlphaGo Zero 中策略以向量的形式进行数学描述替代了 AlphaGo 中的概率分布形式，但本质上两者策略表示是相同的）。在 AlphaGo Zero 中没有采用快速走棋网络，其蒙特卡罗树搜索的模拟评估阶段完全依赖于深度残差网络的价值输出 v。

3.3 AlphaGo Zero 训练与对弈过程

由于 AlphaGo Zero 没有参考人类知识（改善了 AlphaGo 的不足（2）），其网络的训练主要依赖于强化学习和蒙特卡罗树搜索，并且 AlphaGo Zero 的离线训练过程蕴含了在线对弈的经过，所以本节以离线训练过程介绍为主，并细致介绍其中的蒙特卡罗树搜索算法。

3.3.1 离线训练过程

AlphaGo Zero 仅含有一个深度残差网络 $f_\theta(s)$ 输出为$(\boldsymbol{p},v)$，其训练的目标即为优化深度残差网络的权重参数 θ，使得策略 $\boldsymbol{p}$ 棋力更强，而胜负评估 v 更准确。

初始状态时，由于没有人类知识的介入，网络的权重参数 θ 以随机值进行初始化，得到初始深度残差网络。将初始深度残差网络作为当前的最优策略，迭代进行自我对弈、训练优化以及对决评估步骤，最终实现 AlphaGo Zero 的离线训练过程，如图 5。

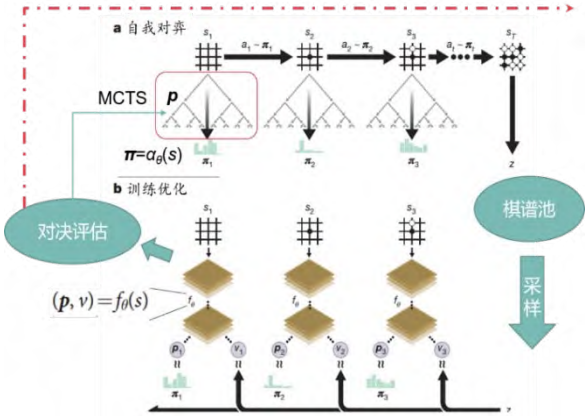

图 5 AlphaGo Zero 离线训练过程示意图[4]

Fig. 5 Schematic diagram of AlphaGo Zero offline training process

自我对弈：使用基于当前最优策略的蒙特卡罗树搜索进行自我对弈的单步决策。每次单步决策需要经过 1 600 次蒙特卡罗树搜索模拟，得到并记录下当前局面 s_t（t 表示自我对弈的第 t 个单步）的策略 $\boldsymbol{\pi}_t$。策略 $\boldsymbol{\pi}_t$ 是深思熟虑（1 600 次模拟）后的新策略，相比于当前最优策略是一个更好的策略，因此蒙特卡罗树搜索进一步提升了强化学习的策略改进速度（改善了 AlphaGo 的不足（3））。根据策略 $\boldsymbol{\pi}_t$，系统采样进行当前盘面的动作决策，得到动作$a_t{\sim}\pi_t$。因此单步决策是一个概率性决策过程，每个动作都有选择的可能性，选择概率服从策略 $\boldsymbol{\pi}_t$，这增加了探索的丰富性（改善了 AlphaGo 的不足（4））。持续执行单步决策过程，直到进行到终盘 T 时刻，得到结果 z，并将该过程记录下的每一个$(s_t,\boldsymbol{\pi}_t,z)$存入棋谱池，用于为后面的训练优化提供数据。重复进行自我对弈过程，丰富棋谱池，达到一定次数后，进行参数的训练优化过程。

训练优化：棋谱池中有大量的数据，从最近的 500 000 盘对弈中进行均匀随机地盘面采样，采样出来的数据$(\mathrm{s},\boldsymbol{\pi},z)$用以优化深度残差网络的参数。在已知$f_\theta(s)=(\boldsymbol{p},v)$的情况下，优化目标包括两方面，一方面希望胜负评估 v 与实际结果 z 尽可能一致；另一方面希望策略 $\boldsymbol{p}$ 能尽可能接近策略 $\boldsymbol{\pi}$。参数优化过程基于损失函数梯度下降方法，由于深度残差神经网络同时输出策略和胜负评估，因此损失函数同时考虑胜负评估值和落子概率，其形式如式（6）。

$$loss=(z-v)^2+(-\boldsymbol{\pi}^T\log\boldsymbol{p})+c\|\theta\|^2 \qquad (6)$$

其中，式第一部分考虑的是胜负评估结果 v 与实际结果 z 的方差；第二部分是输出策略 $\boldsymbol{p}$ 和策略 $\boldsymbol{\pi}$ 的交叉信息熵，交叉信息熵越小两个策略就越相似；第三部分则是用来防止过拟合现象，其中 c 是一个常数。训练优化过程持续进行，每完成 1 000 次训练步骤就产生一个记录点，记录该次训练后的新参数。该参数对应的策略将用在对决评估阶段，与当前最优策略竞争，确定新的当前最优策略。

对决评估：为了保证数据质量越来越好，需要评估新的记录点对应策略和当前最优策略的优劣，择优作为接下来的当前最优策略进行自我对弈，因此需要对两种策略进行对决评估。对决过程和自我对弈过程基本一样，区别只在于当前最优策略的对手不再是自己而是当前评估的记录点对应策略。对决过程中，双方依次使用蒙特卡罗树搜索进行单步决策，每次单步决策执行 1 600 次模拟，直到比赛结束；400 场比赛后，若记录点对应策略的胜率达到 55%以上，则用其替换当前最优策略，并基于新的最优策略通过自我对弈继续产生更好的数据；否则，放弃该记录点，仍采用当前最优策略进行自我对弈。可以看出对决评估过程本质就是在线对弈的过程，所以 AlphaGo Zero 的在线对弈过程和对决评估以及自我对弈过程相近，本文不再重复介绍。

重复以上三个步骤，深度残差网络的棋力就会不断提升，图 6 就是 AlphaGo Zero 随着训练时间的增加棋力的变化曲线。图中的实线表示 AlphaGo Zero 的棋力随训练时间增加的变化情况。绿色虚线是 4-1 打败李世石的 AlphaGo Lee（AlphaGo 的升级版，但在结构和算法思路上没有区别），棋力值 3 739（棋力值计算方法参见文献[3]）。蓝色虚线对应的是 AlphaGo Master，棋力值为 4 858，它和 AlphaGo Zero 结构、训练以及对弈方法完全一样，区别在于 AlphaGo Master 的输入是和 AlphaGo 同样的 19×19×48 的二值平面，蒙特卡罗树搜索的模拟评估阶段保留了快速走棋模拟评估，并且以基于

人类知识的监督学习网络作为深度残差网络的初始状态。如图 6，尽管早期的 AlphaGo Zero 弱小的像一个孩子，但其经过 3 天的学习就可以超过 AlphaGo Lee，不到一个月就达到了人类知识辅助 AlphaGo Master 的棋力水平，并且最终超越了 AlphaGo Master 达到了 5 185 的棋力值。

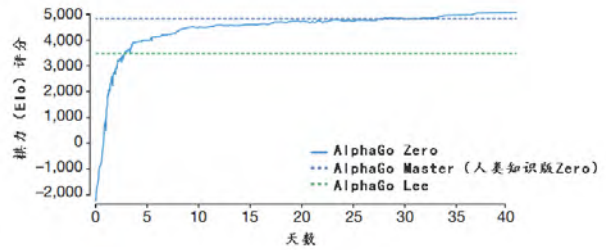

图 6 AlphaGo Zero 棋力变化曲线[4]
Fig. 6 Performance of AlphaGo Zero

3.3.2 AlphaGo Zero 中的蒙特卡罗树搜索

在 AlphaGo Zero 中，蒙特卡罗树搜索算法贯穿了离线训练和在线对弈的整个过程。并且，相比于 AlphaGo 中的蒙特卡罗树搜索算法，AlphaGo Zero 进行了改进优化，使得其最终得到了更好的性能。本小节将对 AlphaGo Zero 中的蒙特卡罗树搜索算法进行详细介绍。

AlphaGo Zero 中的蒙特卡罗树搜索算法总共有 3 个步骤：选择、扩展与评估以及反向更新，如图 7。相比于 AlphaGo，AlphaGo Zero 将扩展和模拟评估两个步骤合并为一个；另外由于删除了快速走棋网络，博弈树的每条边(s,a)的统计信息简化为$\{N(s,a)，W(s,a)，Q(s,a)，P(s,a)\}$，其中 $N(s,a)$表示该边的模拟次数，$W(s,a)$是该边所有模拟过程胜负评估值的总和，$Q(s,a)=W(s,a)/N(s,a)$是胜负评估均值，$P(s,a)$是执行树搜索策略时的先验概率。

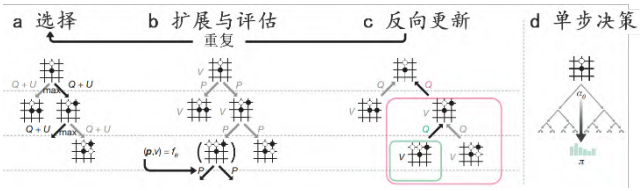

图 7 AlphaGo Zero 蒙特卡罗树搜索流程[4]
Fig. 7 Procedures of Monte Carlo Tree Search for AlphaGo Zero

选择：该阶段和 AlphaGo 的选择阶段基本一样，从根节点 s_0（这里的下标是蒙特卡罗树搜索的时刻标志，与自我对弈的单步决策的步骤下标无关）开始，基于树搜索策略选择路径，直到叶节点 s_L。在 $t<L$ 的每一步中，动作选择策略和 AlphaGo 一样，如公式（2）。其中，$u(s_t,\mathrm{a})$仍然表示红利，意义和 AlphaGo 中的相同，为了方便阅读，树搜索策略和红利的表达式如式（7）。在 AlphaGo Zero 中，先验概率来自于深度残差网络。但为了增加探索性，在根节点 s_0 进行选择决策时，先验概率进行了改进，加入了服从狄利克雷分布的噪声，其形式如式（8）。其中先验概率 p_{a} 来自于深度残差网络的策略输出，$\eta_{\mathrm{a}}\sim\mathrm{Dir}(0.03)$表示狄利克雷噪声，权重系数 $\varepsilon=0.25$ 用以控制探索程度。噪声的加入使得根节点所有的合法动作都可能被搜索到，改善了 AlphaGo 的不足（4）。

$$\mathrm{a}_t=\arg\max_{\mathrm{a}}\big(Q(s_t,\mathrm{a})+u(s_t,\mathrm{a})\big),t<L$$
$$u(s,\mathrm{a})=C\frac{P(s,\mathrm{a})}{1+N(s,\mathrm{a})} \tag{7}$$
$$P(s_0,\mathrm{a})=(1-\varepsilon)p_{\mathrm{a}}+\varepsilon\eta_{\mathrm{a}} \tag{8}$$

扩展与评估：扩展与评估阶段同时完成扩展以及胜负评估任务。在该阶段，当搜索到达叶节点 s_L 后，盘面 s_L 送入到深度残差网络中进行胜负评估得到 $v(s_L)$；同时将 s_L 进行扩展（在 AlphaGo Zero 中扩展阈值为 1，即每次模拟都会扩展分支，而在 AlphaGo 中扩展的阈值为 40。），扩展后的每条边(s_L,a)的统计信息初始化为$\{N(s_L,\mathrm{a})=0，W(s_L,\mathrm{a})=0，Q(s_L,\mathrm{a})=0，P(s_L,\mathrm{a})=p_{\mathrm{a}}\}$。

反向更新：将深度残差网络的胜负评估 $v(s_L)$ 反向更新 $t<L$ 步骤中每条边的统计信息，更新方式如下：

$$N(s_t,\mathrm{a}_t)=N(s_t,\mathrm{a}_t)+1$$
$$W(s_t,\mathrm{a}_t)=W(s_t,\mathrm{a}_t)+v(s_L) \tag{9}$$
$$Q(s_t,\mathrm{a}_t)=W(s_t,\mathrm{a}_t)/N(s_t,\mathrm{a}_t)$$

重复执行以上三个步骤 1 600 次，此时即可根据统计信息进行单步决策。在 AlphaGo 中，在线对弈时的单步决策完全依赖于动作模拟的次数，而在 AlphaGo Zero 中，为了增加探索性，在单步决策时引入了退火思想。若将策略向量$\boldsymbol{\pi}$表示成概率形式，蒙特卡罗树搜索输出的策略如式（10），表示在盘面 s_0的条件下选择动作 a 的概率。在每盘棋的前 30 步单步决策时，参数 $\tau=1$，此时每个动作 a 的概率就是模拟过程出现的频率，由于对弈过程是基于 $\pi(\mathrm{a}|s_0)$的采样决策，因此在开盘的前 30 步落棋具有丰富的可能性（改善了 AlphaGo 的不足（4））。在 30 步之后，$\tau\to0$，取一个趋近于 0 的极小值，此时式（10）的分布极其尖锐，出现次数最多的动作的概率趋向于 1，其他动作的概率均趋向于 0，尽管此时仍然是基于$\pi(\mathrm{a}|s_0)$的采样决策，但实际效果已转化为确定性决策。这一机制的思想是考虑到开局时未来变化空间大，无论是策略亦或是胜负评估都不甚准确，此时需要增加探索性避免陷入局部最优；而随着盘面推进，局势变化可能性逐步收缩，策略和胜负评估的指导性更准更强，此时则应该遵循蒙特卡罗树搜索的决策，追求更高的胜率。

$$\pi(\mathrm{a}|s_0)=\frac{N(s_0,\mathrm{a})^{1/\tau}}{\sum_b N(s_0,b)^{1/\tau}} \tag{10}$$

3.4 类比小结

通过上面的分析能够发现，AlphaGo Zero 针对 AlphaGo 的不足做出了许多改进，两者技术体系的改进框图如图 8。为了更好地理解 AlphaGo Zero 的设计思想，将 AlphaGo Zero 类比为一个仅知道游戏规则的围棋新手 Go，它希望通过自己摸索学习围棋。

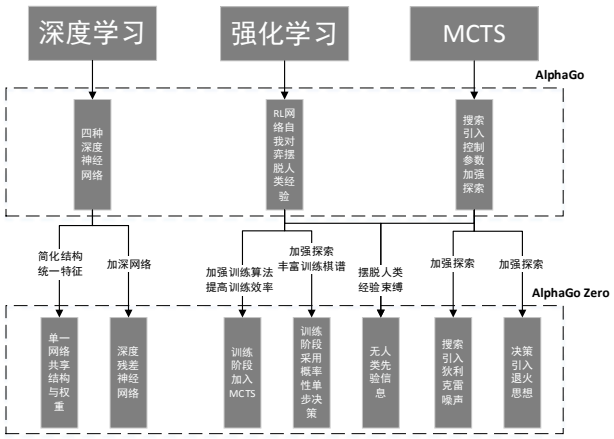

图 8 技术体系改进框图

Fig.8 Block diagram of technical system improvement

首先 Go 知道要想下好围棋，每一步棋的决策非常重要，而要想知道什么是好棋就需要能对盘面有准确的评估。尽管 Go 知道自己的朋友很多是分别独立学习下棋和盘面评估，但 Go 的个人哲学理念认为两者（下棋和盘面评估）在重要特征方面一定具有相通性（单一深度残差网络）。此外 Go 也没有找其他围棋高手学习，因为 Go 觉得每一个人类棋手都会有缺陷，都有“一着不慎满盘皆输”的情况，而 Go 又无法从一个人的棋风中分离出这个人的缺陷，容易陷入局部最优，所以 Go 决定自己摸索（放弃人类知识）尝试探索全局最优的棋风。摸索的方法就是自己和自己下棋（自我对弈），根据每盘棋的输赢，总结教训（训练优化），升级自己的棋风，之后利用新棋风继续自我对弈并总结，迭代执行不断提高自己。Go 的想法很好，但其中存在一些问题。首先对自己棋谱的总结最终也仅能从自己的棋风中去糟存优，但是自己的初始棋力本身就弱小，自我对弈时的好棋可能只是因为对手太弱了，从这样的对决中总结教训，借鉴意义太小，即使能有所提高，效率也太慢。所以 Go 决定只学习有意义的棋谱，为了产生有意义的棋谱，每一步棋都先经过深思熟虑，反复推演（蒙特卡罗树搜索），这一过程相当于剔除了部分比较随意的对局，提高了总结教训的效率。其次，总结教训的结果尽管多数情况提升了自己，但也可能误入歧途走偏了道路，所以每次总结完还要和之前的棋风进行对比（对决评估），确定新的棋风是否好用。此外，既然棋风可能误入歧途，而且 Go 也清楚自己不可能已经步步好棋，所以总按照自己的棋风下棋有可能错过某些妙招。因此在下棋的时候，Go 也不断的尝试一下新的走法（增加探索性），特别是像开局这种局势不明朗难以预料后事的时侯（开局多样，收官谨慎）。基于此，Go 在不断的训练过程当中棋力显著提升，而在与他人进行对弈的时候，Go 吸取训练的经验，不仅仅依赖于当前的棋风，同样每步棋反复推演（在线对弈时的蒙特卡罗树搜索），将每一次与他人的对弈也当作是一次训练，使得 Go 在比赛当中也能收获提升。

以上就是 AlphaGo Zero 的围棋学习道路。回顾其学习之路，尽管并没有背记过人类高手的棋谱，但它通过自己的摸索发现了围棋利器——“定式”，很多“定式”与人类棋谱中记录的“定式”一样，同时它还发现了更多新的围棋“定式”，并且最终战胜了当世的围棋最强者。

4 思考与启示

4.1 对 AlphaZero 的思考

AlphaZero 框架是以 AlphaGo Zero 为基础的深度强化学习框架，它去除了 AlphaGo Zero 中围棋独有的算法特征，保留了普适性的学习思想、方法和技巧，适用于完全信息条件下的二人有限零和博弈模型，而其中关于探索与利用的平衡、策略推演方式、结果评估方法等方面对于更广泛的强化学习领域同样具有借鉴意义。AlphaZero 框架的伟大之处在于第一次让机器可以不通过任何棋谱，不依赖任何人类的经验，在只告诉其规则的前提下，成为一个围棋高手。这种无师自通的学习模式在人工智能发展道路上是非常有里程碑意义的。但是同时这种无师自通在很多人工智能推广应用上也存在一些局限，因为严格的讲，围棋规则和判定棋局输赢也是一种监督信号，所以严格意义上，说人类无用，或者说机器可以自己产生认知都是对 AlphaZero 理解的不精确。此外，目前 AlphaZero 框架仍然需要上百万盘的自我对弈才能真正掌握围棋，而这与人类掌握围棋的过程还有明显的区别，这可能是思考方式上的本质差别，也可能是学习方式上的差别导致的学习效率的差别。因此 AlphaZero 的出现固然伟大，但我们也不要对其过分解读。

通过对 AlphaGo 和 AlphaGo Zero 的分析对比，能够描绘出 AlphaZero 框架形成的发展历程并发现其中的关键点。首先，深度神经网络的结构非常重要，网络的组织形式与层数决定了网络表达的丰富性和能力，可以从 AlphaGo 的 13 层深度卷积神经

网络到 AlphaGo Zero 的 39 层深度残差神经网络之间的性能对比看出结构的重要性。然后，目前深度学习和强化学习的理论基础还很薄弱，许多研究都是基于探索性或启发式的方法，新方法的优劣评估也存在许多定性的经验性解读。例如深度学习、强化学习以及蒙特卡罗树搜索之间的结合之前也有相关的尝试，但是结合方法的不同，或者某些参数的差别导致性能相差甚远，但对此现象却缺少理论的剖析支撑。这一方面指引学者要加强理论方面的研究，增强算法的可解释性，从理论层面阐述方法的优劣，并以理论的指导去探究更优的方法；另一方面，对于“聪明”的科技工作者这也是一个机会，可以在较少的理论基础条件下通过其他领域知识的触类旁通或启发式的探索，在智能决策领域做出突破。其次，探索和利用的平衡问题可以显著影响算法性能，通过 AlphaGo 和 AlphaGo Zero 的对比可以发现，通过加强探索，强化了系统选择的多样性，降低了陷入局部最优解的可能性；但同时探索的加强增加了计算的复杂度，阻碍了算法的收敛，无法满足具有实时性或准实时性的系统要求。最后，算力问题是智能决策发展的关键支撑。本文中并未过多的提及平台计算能力问题（文献[3]和[4]均对计算能力对棋力的影响进行了研究），但在实际应用中平台算力决定了训练速度和在线对弈时蒙特卡罗树搜索的模拟速度，进而决定了“推演模拟”的精度。平台的计算能力主要由处理芯片决定，因此业界的巨头公司均在人工智能芯片领域投入大量人力和财力，如 Google 的 TPU3、NVIDIA 的 GV100 GPU、AMD 的 Vega64 显卡芯片、中科院牵头的寒武纪系列处理器以及许多面向深度学习的 ASIC（领域专用集成电路）芯片，这也将是我国人工智能未来发展的一个重要建设领域。

4.2 AlphaZero 对军事应用的启示

象棋、围棋等博弈类游戏，本身就是对于军事战争的抽象模拟，因此博弈类游戏的智能决策对于军事决策的智能化具有重要借鉴意义。在 2007 年人机国际象棋大赛中，“深蓝”一举击败人类棋手卡斯帕罗夫，在全世界引起轰动，同时也引起美国军方高度关注，提出了“深绿”计划。“深绿”是美国防部高级研究计划局（DARPA）2007 年起支持的一项指挥决策领域研究项目，原计划执行 3 年，至今未完成，且项目内容已大大减少。该计划完成的系统将嵌入美国陆军现有旅级之上 C^4ISR 的战时指挥决策支持系统。“深绿”计划核心思想是借鉴“深蓝”，预判敌人的可能行动，从而提前做出决策[6]，也就是类似 AlphaZero 的一个博弈决策系统。

航空兵器作为未来军事战争的重要作战力量，同样需要面临即将到来的智能化战争考验。目前导弹、飞机中的雷达、制导、目标选取、飞行控制都在向智能化方向发展[17]，在航空兵器智能决策发展早期，通常使用专家系统与数据存储和通信网络技术结合，用于机载预警和控制系统等。专家系统通过模型库、数据库和方法库的信息输入，根据自身的知识进行推理决策，完成飞行控制或帮助判断敌军位置和动机；而从单一功能上升到战斗机等完整武器系统指挥，则需要引入类似 AlphaZero 这类更复杂、更智能的决策技术，特别是在导弹、飞机、无人机这类高速应用场景，人类的反应难以适应战争的“秒杀”节奏，此时智能化决策技术将成为目前可预见的最佳选择，2016 年美国辛辛那提大学研发的“阿尔法”AI 就成功操控 F-15 战机击败了飞行员驾驶的 F-22 战机[18]。更进一步，针对群体装备系统或体系指控装备，还需要兵棋推演这类更宏观的智能决策系统，一方面可更准确地预测战术/战略实施效果，另一方面可通过兵棋推演系统去验证和优化作战方案。这类兵棋推演系统也是 AlphaZero 的重要舞台。

因此，AlphaZero 的出现为“深绿”、“阿尔法”或者类似系统的设计、训练和学习方法提供了新的借鉴。可以分析和理解战场特性构建符合战场态势的深度神经网络结构；然后利用已有的演习和试验数据来构建战场环境模型；之后抛弃已有演习数据，基于战场环境的反馈，通过自我对战的模拟，从零开始逐步学习、理解并认知战场态势，模拟期间合理平衡探索和利用，在有效的时间内得到尽量准确的决策。

然而，AlphaZero 的博弈与实际战争仍然存在着极大的差别。AlphaZero 的目标是处理完全信息条件下的二人有限零和博弈问题，而战场指挥问题的本质是一个态势感知与估计、实时响应、非完全信息博弈和多智能体协同等多个问题构成的复杂性系统问题[9]。

对于态势感知与估计问题，AlphaZero 能够提供较好的借鉴示范，但是对于如何描述战场态势输入、表征和抽象战场模型、构建战场环境，如何选择与战场特性相适应的网络结构等问题仍然需要更进一步的研究。

对于实时响应问题，一方面，AlphaZero 的博弈本质是一个回合制游戏，而战争则是即时战略类游戏，要解决有限状态与战场连续性的矛盾；另一方面，这也对平台计算能力提出要求，尤其在 2018 年 4 月爆出美国制裁中兴事件后，高性能处理芯片将成为一个重要制约因素。

对于非完全信息博弈问题，一方面，敌人不是合作者，永远不会有足够信息，甚至会提供虚假数据信息误导决策。另一方面，演训数据较少缺乏学习样本，如果利用模拟方式生成训练数据，则要对模拟的逼真程度提出严格的要求。

对于多智能体协同问题，实际战场往往是多人或多方的合作通信及竞争关系，AlphaZero 的双人博弈模型明显不足，需要将单一模型扩展为多个智能体之间相互合作、通信及竞争的多智能体深度强化学习系统[19]。

需要特别说明，在航空兵器的飞控、制导等具体任务领域（即不考虑航空兵器的整机指挥或多体的兵棋推演任务），对于 AlphaZero 需要有选择地吸收借鉴。常见的专家系统或基于遗传算法的智能决策通常需要提供经过人工模型处理后的信息（如弹道轨迹模型输出、飞行轨迹模型输出、地理信息系统输出、姿态信息等），这类似于 AlphaGo 早期训练时的棋谱学习，这些模型的输出可以理解为信息或知识的提炼，但也可以看作既有知识的约束。对待这一情况，不能简单借鉴 AlphaZero 摒弃人类经验，因为在围棋领域里，由于其规模庞大、价值反馈滞后，人类既有知识归纳和总结存在许多错误，这类知识的继承和学习确实会羁绊和约束学习者；但是航空领域的知识结构成熟且具备一定共识，因此在知识正确的前提下，既有知识反而可以使决策快速收敛，而且经过既有知识“洗涤”过的信息更易处理，实时性好（如基于查询方式的专家系统），因此更适用于航空兵器领域中高速物体的实时决策。鉴于此种情况，在航空兵器具体任务领域，可以结合既有知识和 AlphaZero 的创新学习能力，在实施任务决策时仍然采用基于既有知识模型的专家系统，而知识模型的生成则采用 AlphaZero 的思想进行创造性的学习。此外，基于 AlphaZero 思想的模型学习系统可以直接部署于飞行器，将实际飞行任务作为训练数据提供给它，实现在线学习，使其可以实时更新知识模型。

因此，尽管 AlphaZero 的出现，给予了军事智能决策新的启示，但对于两者之间的差别仍有许多问题等待解决。目前即时战略游戏（如星际争霸 II）的电脑智能研究对于智能决策的实时响应、多智能体协同问题上具有较多的借鉴意义[20]，并且新公布的 AlphaStar 模型已经战胜星际争霸 II 的专业玩家，这将是智能决策技术的又一里程碑[21]；而“一对一无限注德州扑克”作为非完全信息博弈代表，目前也受到广泛关注，基于深度强化学习算法的 Deep Stack 在该游戏中已经具备了职业玩家的水平[22]。未来我国需要加强在相关领域的探索研究，并大力发展人工智能领域的芯片设计及制造行业，推动我国军事决策智能化发展，在未来作战指挥决策中取得致胜先机。

5 结 论

AlphaZero 的出现将人工智能的狂潮推上了一个新的高度，其在智能决策方面表现出来的突出成绩，为相关领域的研究提供了宝贵经验。本文通过对 AlphaZero 框架的两个主要发展阶段 AlphaGo 和 AlphaGo Zero 的技术原理进行深入剖析，阐释了 AlphaZero 框架成功的原因；通过与“深绿”计划的关联对比，揭示了 AlphaZero 对军事决策智能化的启示，也指出了其与实际指挥决策之间的差异。未来我国军事应用，特别是航空兵器方面，要加强深度强化学习领域的基础研究，并深入探索态势感知与估计、实时响应、非完全信息博弈和多智能体协同等具体领域，以推动军事决策智能化的发展。

参考文献：

[1] Kocsis L, Szepesvari C. Bandit based Monte-Carlo Planning [C] //Proceedings of the European Conference on Machine Learning. Berlin: Springer, 2006: 282 – 293.

[2] Tian Y, Zhu Y. Better Computer Go Player with Neural Network and Long-term Prediction[C]// ICLA, 2016.

[3] 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.

[4] Silver D, Schrittwieserh J, Siomonyan K, et al. Mastering the Game of Go without Human Knowledge [J]. Nature, 2017, 550(7676): 354 – 359.

[5] Silver D, Hubert T, Schrittwieser J, et al. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm[EB/OL]. （ 2017-12-05 ）[2019-04-25].https://arxiv.org/pdf/1712.01815.

[6] 胡晓峰, 郭圣明, 贺筱媛. 指挥信息系统的智能化挑战——“深绿”计划及 AlphaGo 带来的启示与思考[J]. 指挥信息系统与技术, 2016, 7(3): 1-7.
Hu Xiaofeng, Guo Shengming, He Xiaoyuan. The Intelligent Challenge of Command Information System——The Enlightenment and Reflection of "Dark Green" Project and AlphaGo[J]. Command Information System and Technology, 2016, 7(3): 1-7. (in Chinese)

[7] 胡晓峰. 军事指挥信息系统中的机器智能:现状与趋势[J]. 人民论坛・学术前沿, 2016(15):22-34.
Hu Xiaofeng. Machine Intelligence in Military Command Information System: Status and Trends[J]. People's Forum •Academic Frontier, 2016(15): 22-34. (in Chinese)

[8] 陶九阳，吴琳，胡晓峰. AlphaGo 技术原理分析及人工智能军事应用展望[J]. 指挥与控制学报，2016，2(2): 114-120.
Tao Jiuyang, Wu Lin, Hu Xiaofeng. Principle Analysis of AlphaGo and Prospect of Military Application of Artificial Intelligence[J]. Journal of Command and Control, 2016, 2(2): 114-120. (in Chinese)

[9] 唐振韬，邵坤，赵冬斌,等. 深度强化学习进展:从 AlphaGo 到 AlphaGo Zero[J]. 控制理论与应用，2017，34(12): 1529-1546.
Tang Zhentao, Shao Kun, Zhao Dongbin, et al. Deep Reinforcement learning progress: from AlphaGo to AlphaGo Zero[J]. Journal of Control Theory and Applications, 2017, 34(12): 1529-1546. (in Chinese)

[10] Lecun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436.

[11] Goodfellow I, Bengio Y, Courville A. Deep Learning[M]. The MIT Press, 2016.

[12] 周志华. 机器学习[M]. 北京：清华大学出版社, 2016.
Zhou Zhihua. Machine Learning[M]. Beijing：Tsinghua University Press, 2016. (in Chinese)

[13] 28 天自制你的 AlphaGo (6)：蒙特卡洛树搜索（MCTS）基础 [EB/OL].[2019-04-25]. https://zhuanlan.zhihu.com/p/25345778.
Make your AlphaGo in 28 Days (6): The Monte Carlo Tree Search Basics. [2019-04-25]. https://zhuanlan.zhihu.com/p/25345778.

[14] Browne C B, Powley E, Whitehouse D, et al. A Survey of Monte Carlo Tree Search Methods[J]. IEEE Transactions on Computational Intelligence and AI in Games, 2012, 4(1): 1-43.

[15] 深度解读 AlphaGo 算法原理 [EB/OL]. [2019-04-25]. https://blog.csdn.net/songrotek/article/details/51065143
Deep Interpretation of the AlphaGo Algorithm[EB/OL]. [2019-04-25].https://blog.csdn.net/songrotek/article/details/51065143.

[16] 一张图解 AlphaGo 原理及弱点 [EB/OL]. http://www.kddchina.org/#/Content/alphago
Illustrating the Principle and weaknesses of AlphaGo in a Picture[EB/OL]. [2019-04-25]. http://www.kddchina.org/#/Content/alphago

[17] 程进，齐航，袁健全，等. 关于导弹武器智能化发展的思考[J]. 航空兵器, 2019, 26(1): 20-24.
Cheng Jin, Qi Hang, Yuan Jianquan, et al. Discussion on the Development of Intelligent Missile Technology[J]. Aero Weaponry, 2019, 26(1): 20-24. (in Chinese)

[18] 石纯民. 当“阿尔法”走上战场[N]. 中国国防报,2016-07-11.
Shi Chunmin. When “Alpha” Goes to the Battlefield[N]. China National Defense News, 2016-07-11. (in Chinese)

[19] 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.
Zhao Dongbin, Shao Kun, Zhu Yuanheng, et al. Review of Deep Reinforcement Learning: Also on the Development of Computer Go[J]. Journal of Control Theory and Applications, 2016, 33(6): 701-717. (in Chinese)

[20] Vinyals O, Ewalds T, Bartunov S, et al. Starcraft Ⅱ: A New Challenge for Reinforcement Learning [EB/OL].[2019-04-25]. site:https://arxiv.org/pdf/1708.04782

[21] AlphaStar: Mastering the Real-Time Strategy Game StarCraft Ⅱ [EB/OL]. [2019-04-25]. https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/.

[22] Moravčík M, Schmid M, Burch N, et al. DeepStack: Expert-level artificial intelligence in heads-up no-limit poker[J]. Science, 2017, 356(6337): 508.

Principle and Enlightenment of AlphaZero

Tang Chuan，Tao Yerong*，Ma Yueliang

（Luoyang Electronic Equipment Test Center，Luoyang 471000，China）

Abstract: In recent years, the success of computer Go has triggered another round of artificial intelligence boom. The AlphaZero framework developed from computer Go has been successfully applied to the other problems which are two-person zero-sum finite game under complete information conditions. The success of AlphaZero shows the excellent performance of deep learning and reinforcement learning in the field of intelligent decision-making. In this article, we first introduce three core technologies in the AlphaZero framework: deep learning, reinforcement learning and Monte-Carlo tree search. Then the basic principles of the two key phases of the AlphaZero framework (AlphaGo and

AlphaGo Zero) is detailed. Finally, we put forward some thoughts on the AlphaZero framework and discusse its enlightenment on the intelligence of military decision.

Keywords: deep Learning；reinforcement learning；Monte-Carlo Tree Search；AlphaZero；intelligence of military decision