**本论文摘自机器之心，欢迎大家关注机器之心微信公众号。**

计算机国际象棋和计算机科学本身一样古老。查尔斯·巴贝奇、艾伦·图灵、克劳德·香农和冯诺依曼都曾设计硬件、算法以及理论来让计算机分析和玩国际象棋。国际象棋随后成为了一代人工智能研究者努力希望克服的挑战，最终，我们也实现了超越人类水平的国际象棋程序。然而，这些程序高度局限于它们所处的领域，在没有人类大幅度修改的情况下，无法被泛化去处理其他任务。

创造可以以简单规则为基础不断自我学习的程序一直是人工智能领域的重要目标。最近，AlphaGo Zero 算法在围棋上实现了超过人类水平的成绩，而背后使用的是卷积神经网络，只通过强化学习进行自我对弈训练。在本论文中，DeepMind 实现了类似但完全泛化的算法（fully generic algorithm）——在未输入游戏规则以外任何知识的情况下，其推出的全新算法 AlphaZero 在国际象棋和日本将棋上实现了和围棋同样的高水平。DeepMind 宣称该研究证明了 AlphaZero 作为一个通用性强化学习算法可以从零开始，在多种具有挑战性的任务上实现超越人类的水平。

人工智能领域的一个里程碑事件是 1997 年「深蓝」击败了人类世界冠军卡斯帕罗夫。在随后的 20 年里，计算机程序的国际象棋水平一直稳定处于人类之上。这些程序使用人类大师仔细调整的权重来评估落子步骤，同时结合了高性能的 α-β 搜索技术，通过大量启发式机制和对特定领域的适应而扩展出大的搜索树。这些程序包括 2016 年 Top Chess Engine Championship（TCEC）世界冠军 Stockfish；其他强大的国际象棋程序，包括「深蓝」，也使用了非常相似的架构。

在计算复杂性方面，日本将棋（Shogi）要比国际象棋复杂得多：前者有一个更大的棋盘，任何被吃的棋子都可以改变阵营重新上场，被放置在棋盘的大多数位置。此前最强大的将棋程序，如 Computer Shogi Association（CSA）世界冠军 Elmo 直到 2017 年才击败了人类世界冠军。这些程序和计算机国际象棋程序使用了类似的算法，同样基于高度优化的 α-β 搜索引擎和很多对特定域的适应性调整。

围棋非常适合 AlphaGo 中的神经网络体系结构，因为游戏规则是转移不变的（与卷积神经网络的权重共享结构相对应），是根据棋盘上相邻点位的自由度来定义的（与卷积神经网络局部结构相对应），而且是旋转和镜像对称的（这允许数据增强和数据合成）。此外，围棋

的动作空间很简单（一个子可能被落在每一个可能的位置上），游戏的结果仅限于二元的输或赢，而两者都有助于神经网络进行训练。

国际象棋和日本将棋可以说相对不适用于 AlphaGo 的神经网络架构。因为其规则是依赖于棋盘位置的（如两种棋类的棋子都可以通过移动到棋盘的某个位置而升级）而且不对称（如一些旗子只能向前移动，而另一些如王和后可以更自由的移动）。这些规则包含了远程互动（例如，后可以一步穿越整个棋盘，从远距离对王将军）。国际象棋的动作空间包含两名棋手棋盘上棋子的所有合法落子位置；而日本将棋甚至还允许被吃掉的棋子重返棋盘（加入另一方）。国际象棋和日本将棋都允许胜负之外的其他结果；事实上，人们相信国际象棋的最优解是平局。

AlphaZero 算法是 AlphaGo Zero 的通用化版本，后者首先被应用在了围棋任务上。它使用深度神经网络和从零开始的强化学习代替了手工编入的知识和特定领域的增强信息。

AlphaZero 不使用手动编写的评估函数和移动排序启发式算法，转而使用深度神经网络 (p, v) = f_θ(s) 和参数 θ。该神经网络将棋盘位置 s 作为输入，输出一个针对每个动作 a 的分量 p_a = P_r（a | s）的移动概率 p 的向量，以及从位置 s 估计期望结果 z 的标量值 v ≈E [z | s]。AlphaZero 完全从自我对弈中学习这些步的获胜概率；这些结果随后用于指导程序的搜索。

和 α-β 搜索使用领域特定的增强信息不同，AlphaZero 使用了一个通用的蒙特卡罗树搜索（MCTS）算法。每一次搜索由一系列的自我对弈的模拟比赛组成，遍历了从根 s_root 到叶的整个树。每一次模拟通过在每个状态 s 中选择一个动作 a，a 具有低访问次数、高走棋概率（通过遍历从 s 选择了 a 的模拟的叶状态取平均得到）和根据当前神经网络 f_θ 决定的高价值。搜索会返回一个向量 π 表示走棋的概率分布，通常相对于根状态的访问次数是成比例的或贪婪的。

从随机初始化的参数 θ 开始，AlphaZero 中的深度神经网络参数 θ 通过自我对弈强化学习来训练。双方玩家通过 MCTS 选择游戏动作为 a_t ~ π_t。在游戏结束时，根据游戏规则对终端位置 s_T 进行评分，以计算游戏结果 z：-1 为输，0 为平局，+1 为赢。更新神经网络参数 θ 以使预测结果 v_t 和游戏结果 z 之间的误差最小化，并使策略向量 p_t 与搜索概率 π_t 的相似度最大化。具体而言，参数 θ 通过梯度下降分别在均方误差和交叉熵损失之和上的损失函数 l 进行调整。

$$(\mathbf{p}, v) = f_\theta(s), \qquad l = (z - v)^2 - \boldsymbol{\pi}^\top \log \mathbf{p} + c||\theta||^2 \tag{1}$$

其中 c 为控制 L2 权重正则化程度的参数，更新的参数将用于自我对弈子序列对弈。

本论文描述的 AlphaZero 算法主要在以下一些方面与原版的 AlphaGo Zero 算法有不同。

若假设一局对弈的结果为胜利或失败两个状态，AlphaGo Zero 会估计并最优化胜利的概率。而 AlphaZero 反而会估计和优化结果的期望值，它会同时考虑平局或其它潜在的可能结果。

无论使用怎样的旋转和镜像映射，围棋的规则都是不变的。AlphaGo 和 AlphaGo Zero 都利用了这一事实。首先，通过为每个位置生成 8 次对称，来增强训练数据。其次，在 MCTS（蒙特卡罗树搜索）中，在神经网络进行评估之前使用随机选择的旋转或反射转换棋盘局势，以使蒙特卡罗评估在不同的偏差中进行平均。象棋和将棋的规则是不对称的，且通常无法假设对称。在 MCTS 中，AlphaZero 不增强训练数据，也不转换棋盘局势。

在 AlphaGo Zero 中，自我对弈是由前面所有迭代步中最优玩家生成的。在每次训练的迭代结束后，新玩家的性能通过与最优玩家的对抗而衡量。如果新玩家能以 55% 的优势胜出，那么它就替代当前最优的玩家，而自我对弈的结果将由该新玩家产生。相反，AlphaZero 只是简单地维护单个神经网络以连续更新最优解，而不需要等待一次迭代的完成。

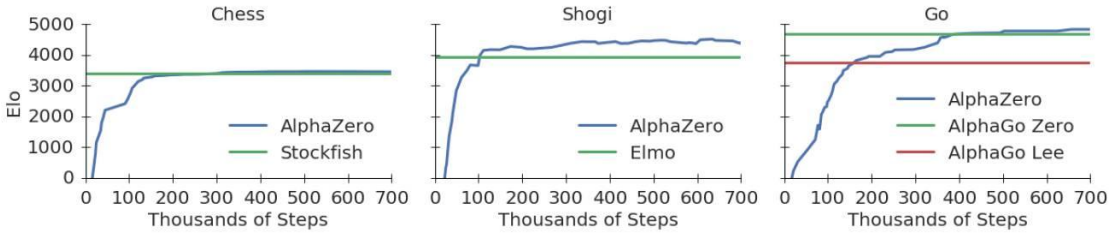


*图 1：将 AlphaZero 训练 700,000 步。假设每手棋用时一秒，利用不同棋手之间的评估游戏计算国际等级分（Elo rating）。a. AlphaZero 在象棋中的表现，与 2016 TCEC 世界冠军 Stockfish 进行对比。b. AlphaZero 在将棋中的表现，与 2017 CSA 世界冠军 Elmo 进行对比。c. AlphaZero 在围棋中的表现，与 AlphaGo Lee 和 AlphaGo Zero 进行对比（20 block / 3 day）（29）。*

自我对弈通过使用这个神经网络最新的参数而生成，且省略了评估的步骤和最佳玩家的选择。

AlphaGo Zero 通过贝叶斯优化搜索超参数，而 Alpha Zero 对于所有的对弈使用相同的超参数，而不会使用特定的超参数调整方法。唯一的例外是为了保证探索（29）而添加到先前策略的噪声，这与符合（对弈类型）规则的典型移动数成正比。

如同 AlphaGo Zero 一样，棋盘状态仅基于每个对弈的基本规则空间进行编码。这些动作是由其它空间平面或平面向量进行编码，且仅仅基于每个游戏的基本规则。

我们把 AlphaZero 算法应用到了国际象棋、日本将棋和围棋上。除非另做说明，这三种棋类游戏使用的都是同样的算法设置、网络架构和超参数。我们为每一种棋类游戏训练了独立的 AlphaZero 实例。训练进行了 70 万步（批尺寸为 4096），从随机初始化参数开始，使用 5000 个第一代 TPU 生成自我对弈棋局和 64 个第二代 TPU 训练神经网络。关于训练过程的更多细节在 Method 中。

图 1 展示了 AlphaZero 在自我对弈强化学习中的性能，作为训练步的函数，以 Elo Scale 表示（10）。在国际象棋中，AlphaZero 仅仅经过 4 小时（30 万步）就超越了 Stockfish；在日本将棋中，AlphaZero 仅仅经过不到 2 小时（11 万步）就超过了 Elmo；而在围棋中，AlphaZero 经过 8 小时（16.5 万步）就超过了 AlphaGo Lee（29）。

我们评估了经过充分训练的 AlphaZero 在国际象棋、日本将棋和围棋上分别和 Stockfish、Elmo 以及经过 3 天训练的 AlphaGo Zero 的 100 场竞标赛的结果（从 AlphaZero 角度的赢/平/输），每个程序都是一步一分钟的思考时间。AlphaZero 和 AlphaGo Zero 使用 4 个 TPU 的单个机器进行比赛。Stockfish 和 Elmo 使用 64 个线程和 1GB 的哈希表进行比赛。AlphaZero 令人信服地打败了所有的对手，未输给 Stockfish 任何一场比赛，只输给了 Elmo 八场（补充材料理由几场比赛的示例），见表 1。

| Game | White | Black | Win | Draw | Loss |
|---|---|---|---|---|---|
| Chess | *AlphaZero* | *Stockfish* | 25 | 25 | 0 |
| | *Stockfish* | *AlphaZero* | 3 | 47 | 0 |
| Shogi | *AlphaZero* | *Elmo* | 43 | 2 | 5 |
| | *Elmo* | *AlphaZero* | 47 | 0 | 3 |
| Go | *AlphaZero* | *AG0 3-day* | 31 | – | 19 |
| | *AG0 3-day* | *AlphaZero* | 29 | – | 21 |

*表 1：AlphaZero 在国际象棋、日本将棋和围棋上分别和 Stockfish、Elmo 以及经过 3 天训练的 AlphaGo Zero 的 100 场比赛的结果（从 AlphaZero 角度的赢/平/输），每个程序都是一步一分钟的思考时间。*

我们还分析了 AlphaZero 的蒙特卡罗树搜索（MCTS）和 Stochfish、Elmo 使用的当前最佳 α-β 搜索引擎的性能对比。AlphaZero 在国际象棋中每秒搜索了 8 万个位置，在日本将棋中每秒搜索了 4 万个位置，而 Stockfish 每秒需要搜索 7000 万个位置，Elmo 每秒需要搜索 3500 万个位置。AlphaZero 通过使用深度神经网络重点聚焦于最具潜在价值的走法（可以认为这是一种更加类似人类思考方式的搜索方法，由香农首次提出（27））。图 2 展示了每个玩家关于 Elo scale 的思考时间的可扩展性。AlphaZero 的 MCTS 相比 Stochfish 和 Elmo 能更有效地伸缩思考时间，这使我们对人们广泛接受的 α-β 搜索在这些领域的内在优势提出了质疑。

最后，我们分析了由 AlphaZero 发现的象棋知识。表 2 分析了 12 个最常见的人类国际象棋开局分析（在线数据集记录出现超过了 10 万次）。每一个开局都由 AlphaZero 在自我对抗训练过程中独立发现并频繁使用。从每一个人类国际象棋开局开始，AlphaZero 都能击败 Stockfish，这表明它确实掌握了大量的国际象棋棋谱知识。

使用国际象棋比赛展示 AI 研究的前沿进展已经有几十年的历史。当前最佳的程序都是基于能搜索几百万个位置、利用人工编入的领域专业知识和复杂的领域适应性的引擎。AlphaZero 是一个通用的强化学习算法（最初为围棋而设计），可以在数小时内达到优越的结果，其需要搜索的位置少了几千倍，除了国际象棋的规则外不需要任何的领域知识。此外，同样的算法不需要修改就可以应用到更具挑战性的日本将棋上，同样在数小时内超过了当前最佳结果。

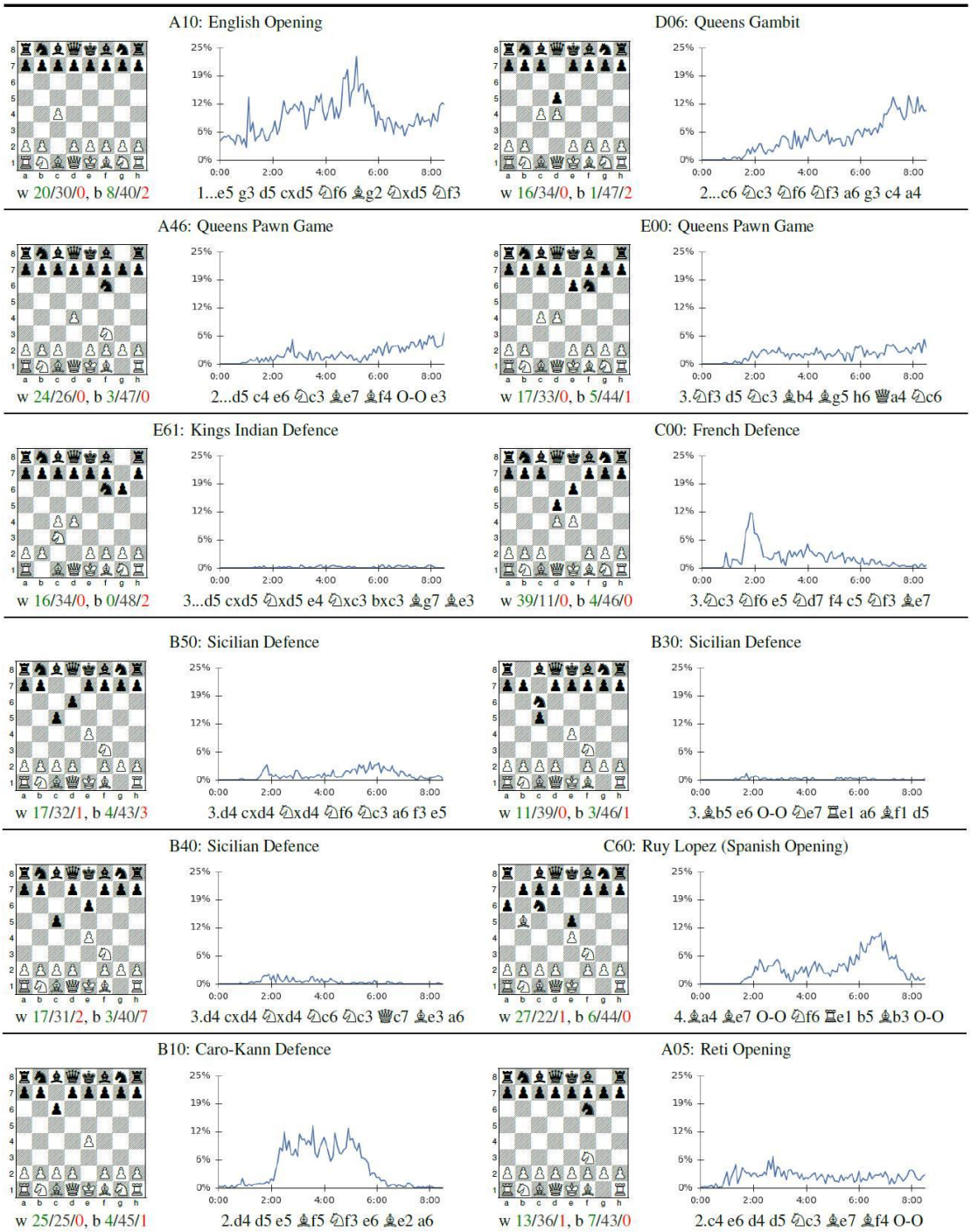


*表 2：12 个最常见的人类国际象棋开场局的分析（在线数据集记录出现超过了 10 万次）。每一个开局由其 ECO 码和常用名标记。这些图展示了 AlphaZero 在自我对抗训练棋局中使用这种开场局的比例随训练时间的变化。我们还报告了 100 场 AlphaZero vs. Stockfish 每个比赛的开局和结果（从 AlphaZero 角度的赢/平/输，无论作为白方还是黑方）。最后，还报告了 AlphaZero 每个开局后的整体棋局主要变化。*

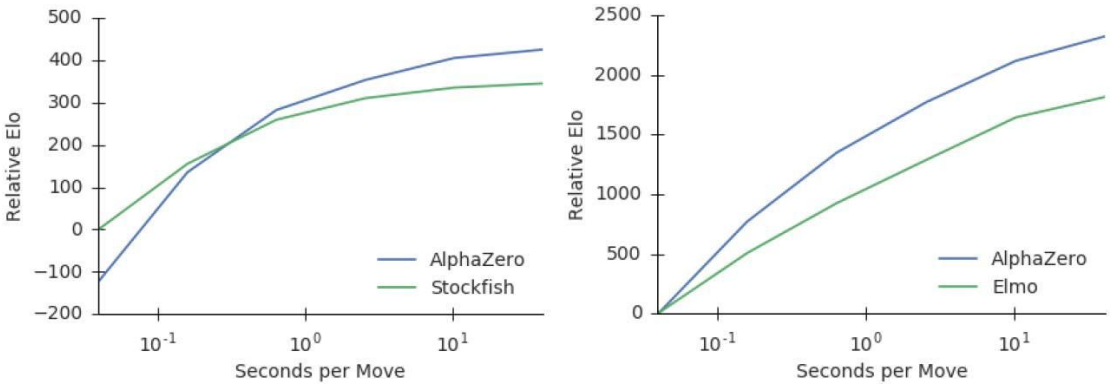


*图 2：AlphaZero 的思考时间的可扩展性。a. AlphaZero 和 Stockfish 在象棋上的 Relative Elo 对比，横坐标为每一步的思考时间。b. AlphaZero 和 Elmo 在日本将棋上的 Relative Elo 对比，横坐标为每一步的思考时间。*

**论文：Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm**

论文链接：https://arxiv.org/abs/1712.01815

摘要：国际象棋是人工智能史上被研究最为广泛的领域。解决国际象棋问题最为强大的技术是通过复杂搜索技术、特定领域的适应性调整以及人类专家几十年来不断手动编写改进的评估函数。相比之下，AlphaGo Zero 程序最近在围棋项目中实现了超过人类的表现，而且它是完全从零开始进行自我强化学习的。在本论文的研究中，我们泛化了这个方法而得到了单个 AlphaZero 算法，使其可以从零开始自我学习，并在很多种具有挑战性的领域里超越人类的性能。模型从随机动作开始初始化，除了游戏规则，我们未向程序输入任何知识，而 AlphaZero 在 24 小时内像围棋一样掌握了游戏，达到了超越人类的国际象棋和日本将棋水平，并令人信服地在每个项目中击败了目前业内顶级的各类程序。