「重磅」AlphaZero炼成最强通用棋类AI，DeepMind强化学习算法8小时完爆人类棋类游戏

由于是通用棋类AI，因此去掉了代表围棋的英文“Go”，没有使用人类知识，从零开始训练，所以用Zero，两相结合得到“AlphaZero”，这个新AI强在哪里?

世界最强围棋AI AlphaGo Zero带给世人的震撼并没有想象中那么久——不是因为大家都去看谁（没）跟谁吃饭了，而是DeepMind再次迅速超越了他们自己，超越了我们剩下所有人的想象。

12月5日，距离发布AlphaGo Zero论文后不到两个月，他们在arXiv上传最新论文《用通用强化学习算法自我对弈，掌握国际象棋和将棋》（Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm），用看似平淡的标题，平淡地抛出一个炸弹。

其中，DeepMind团队描述了一个**通用棋类AI“AlphaZero”**，在不同棋类游戏中，战胜了所有对手，而这些对手都是各自领域的顶级AI：

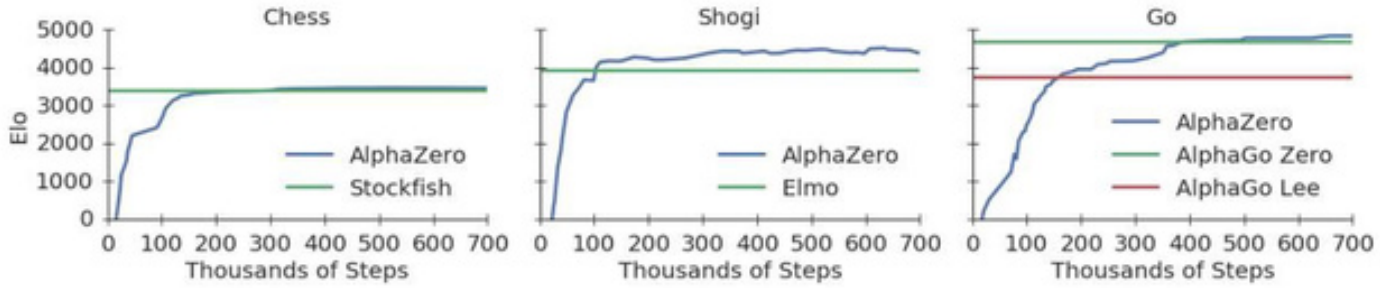
战胜最强国际象棋AI Stockfish：28胜，0负，72平；

战胜最强将棋AI Elmo：90胜，2平，8负；

战胜最强围棋AI AlphaGo Zero：60胜，40负

其中，Stockfish是世界上最强的国际象棋引擎之一，它比最好的人类国际象棋大师还要强大得多。与大多数国际象棋引擎不同，Stockfish是开源的（GPL license）。用户可以阅读代码，进行修改，回馈，甚至在自己的项目中使用它，而这也是它强大的一个原因。

将棋AI Elmo的开发者是日本人泷泽城，在第27届世界计算机将棋选手权赛中获得优胜。Elmo的策略是在对战中搜索落子在哪个位置胜率更高，判断对战形势，进而调整策略。Elmo名字的由来是electric monkey（电动猴子，越来越强大之意），根据作者的说法也有elastic monkey（橡皮猴子，愈挫愈勇）之意。



而AlphaGo Zero更是不必介绍，相信“阿法元”之名已经传遍中国大江南北。而AlphaZero在训练34小时后，也胜过了训练72小时的AlphaGo Zero。

AlphaZero横空出世，网上已经炸开了锅，Reddit网友纷纷评论：AlphaZero已经不是机器的棋了，是神仙棋，非常优美，富有策略性，更能深刻地谋划（maneuver），完全是在调戏Stockfish。

> 看着AlphaZero赢，简直太不可思议了！这根本就不是计算机，这压根儿就是人啊！Holy fu*ck，第9场比赛太特么疯狂了！**DeepMind太神了！我的神啊！它竟然只玩d4/c4。总体上来看，它似乎比我们训练的要少得多。这条消息太疯狂了。**

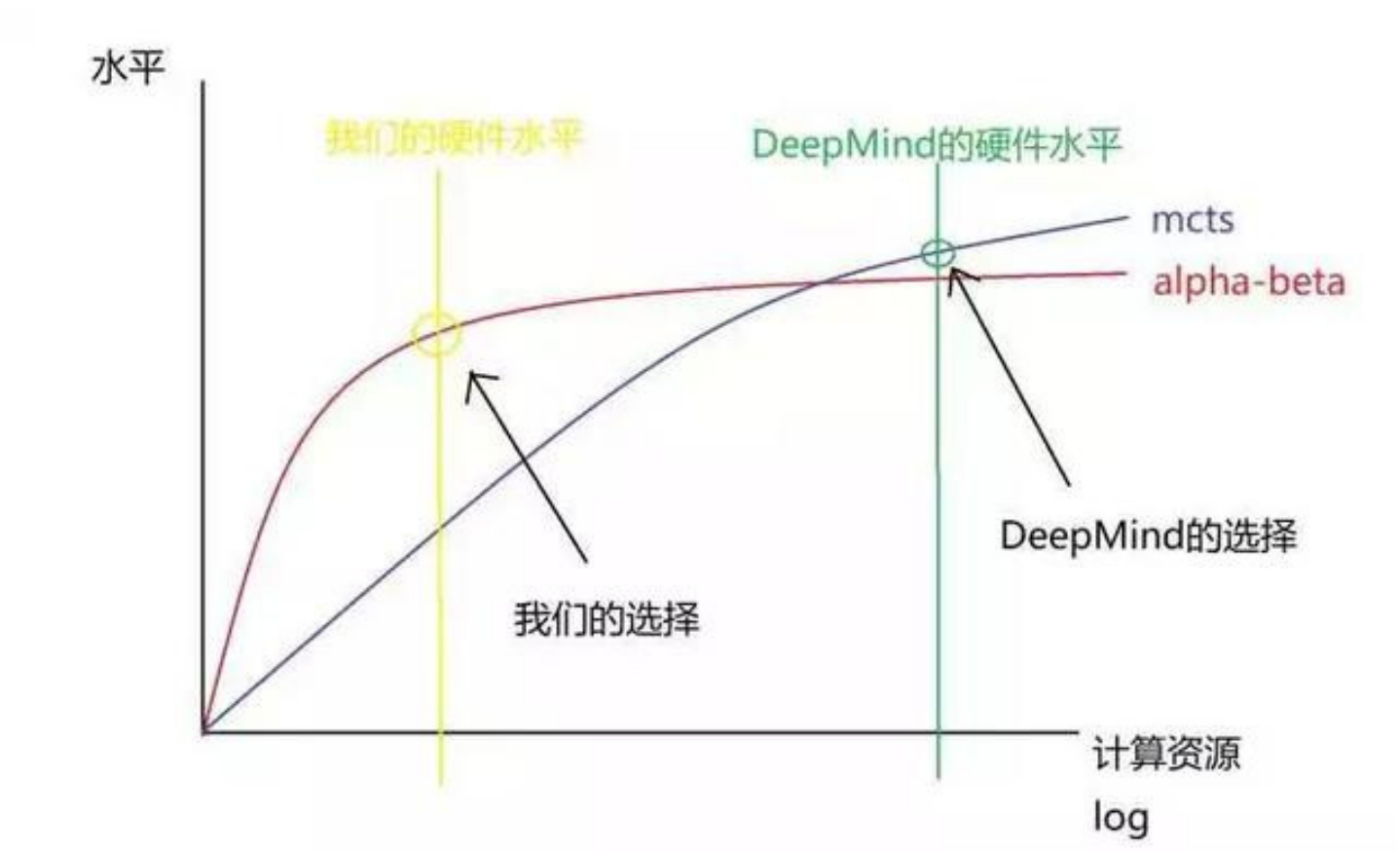
而知乎上，短短几小时内也有很多评论：

知乎用户fffasttime：专治各种不服的DeepMind又出师了，但这次的主攻的内容不再是围棋了，而是所有的棋类游戏。……之前AlphaGo把围棋界打得心态崩了，而现在AlphaZero赢的不光是人类棋手，还包括各路象棋的AI作者。

知乎用户陆君慨：棋类的解决框架一直都是基于 minimax + heuristic。以前围棋难是因为minimax在有着很大分支的游戏上无法产生足够的深度，并且heuristic难以设计。Alphago Zero时候就已经证明了cnn很适合做heuristic，而mcts也可以解决深度问题。那为什么别人不做呢?

因为贫穷限制了我们的想象力。

有钱真的是可以为所欲为。



比AlphaGo Zero更强的AlphaZero来了！8小时解决一切棋类！

知乎用户PENG Bo迅速就发表了感慨，我们取得了他的授权，转载如下（知乎链接见文末）：

读过AlphaGo Zero论文的同学，可能都惊讶于它的方法的简单。另一方面，深度神经网络，是否能适用于国际象棋这样的与围棋存在诸多差异的棋类？MCTS（蒙特卡洛树搜索）能比得上alpha-beta搜索吗？许多研究者都曾对此表示怀疑。

但今天AlphaZero来了（https://arxiv.org/pdf/1712.01815.pdf），它破除了一切怀疑，通过使用与AlphaGo Zero一模一样的方法（同样是MCTS+深度网络，实际还做了一些简化），它从零开始训练：

4小时就打败了国际象棋的最强程序Stockfish！

2小时就打败了日本将棋的最强程序Elmo！

8小时就打败了与李世石对战的AlphaGo v18！

在训练后，它面对Stockfish取得100盘不败的恐怖战绩，而且比之前的AlphaGo Zero也更为强大（根据论文后面的表格，训练34小时的AlphaZero胜过训练72小时的AlphaGo Zero）。

这令人震惊，因为此前大家都认为Stockfish已趋于完美，它的代码中有无数人类精心构造的算法技巧。

然而，现在Stockfish就像一位武术大师，碰上了用枪的AlphaZero，被一枪毙命。

喜欢国象的同学注意了：AlphaZero不喜欢西西里防御。

训练过程极其简单粗暴。超参数，网络架构都不需要调整。无脑上算力，就能解决一切问题。

Stockfish和Elmo，每秒种需要搜索高达几千万个局面。

AlphaZero每秒种仅需搜索几万个局面，就将他们碾压。深度网络真是狂拽炫酷。

Program	Chess	Shogi	Go
AlphaZero	80k	40k	16k
Stockfish	70,000k		
Elmo		35,000k	

Table S4: Evaluation speed (positions/second) of *AlphaZero*, *Stockfish*, and *Elmo* in chess, shogi and Go.

当然，训练AlphaZero所需的计算资源也是海量的。这次DeepMind直接说了，需要5000个TPU v1作为生成自对弈棋谱。

不过，随着硬件的发展，这样的计算资源会越来越普及。未来的AI会有多强大，确实值得思考。

个人一直认为，MCTS+深度网络是非常强的组合，因为MCTS可为深度网络补充逻辑性。我预测，这个组合未来会在更多场合显示威力，例如有可能真正实现自动写代码，自动数学证明。

为什么说编程和数学，因为这两个领域和下棋一样，都有明确的规则和目标，有可模拟的环境。（在此之前，深度学习的调参党和架构党估计会先被干掉..... 目前很多灌水论文，电脑以后自己都可以写出来。）

也许在5到20年内，我们会看到《Mastering Programming and Mathematics by General Reinforcement Learning》。然后许多人都要自谋出路了.....

一个通用强化学习算法，横跨多个高难度领域，实现超人性能

David Silver曾经说过，强化学习+深度学习=人工智能（RL+DL=AI）。而深度强化学习也是DeepMind一直以来致力探索的方向。AlphaZero论文也体现了这个思路。论文题目是《用通用强化学习自我对弈，掌握国际象棋和将棋》。可以看见，AlphaGo Zero的作者Julian Schrittwieser也在其中。

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

David Silver,[1*] Thomas Hubert,[1*] Julian Schrittwieser,[1*]
Ioannis Antonoglou,[1] Matthew Lai,[1] Arthur Guez,[1] Marc Lanctot,[1]
Laurent Sifre,[1] Dharshan Kumaran,[1] Thore Graepel,[1]
Timothy Lillicrap,[1] Karen Simonyan,[1] Demis Hassabis[1]

[1]DeepMind, 6 Pancras Square, London N1C 4AG.
*These authors contributed equally to this work.

摘要：国际象棋是人工智能史上最被广泛研究的领域。最强大的象棋程序是基于复杂的搜索技术、特定领域适应性以及人工评估函数的结合，这些函数在过去几十年里由人类专家不断完善改进。相比之下，AlphaGo Zero最近在围棋中取得了超越人类的成绩，利用的是自我下棋的“白板”强化学习（译注：tabula rasa，意为“白板”，指所有知识均由感官和经验而来，即从零开始的学习）。在这篇论文中，我们将这种方法推广到一个单一的AlphaZero算法，它可以在多个具有挑战性的领域实现超越人类的性能，同样是以“白板”的学习方式。从随机下棋开始，除了游戏规则之外，没有给它任何专门领域的知识，AlphaZero在24小时内实现了在国际象棋、日本将棋和围棋上超越人类水平的表现，并且在这三种棋都以令人信服的成绩击败了当前世界冠军的程序。

对计算机国际象棋的研究和计算机科学一样古老。巴贝奇、图灵、香农和冯·诺依曼都设计过硬件、算法和理论来分析国际象棋，以及下国际象棋。象棋后来成为了一代人工智能研究者的挑战性任务，在高性能的计算机的助力下，象棋程序达到了顶峰，超越了人类的水平。然而，这些系统高度适应它们的领域，如果没有大量的人力投入，就不能归纳到其他问题。

人工智能的长期目标是创造出可以从最初的原则自我学习的程序。最近，AlphaGo Zero算法通过使用深度卷积神经网络来表示围棋知识，仅通过自我对弈的强化学习来训练，在围棋中实现了超越人类的表现。在本文中，除了游戏规则之外，我们还应用了一个类似的但是**完全通用的算法，我们把这个算法称为AlphaZero，除了游戏规则之外，没有给它任何额外的领域知识，这个算法证明了一个通用的强化学习算法可以跨越多个具有挑战性的领域实现超越人类的性能，并且是以“白板”（tabula rasa）的方式。**

1997年，“深蓝”在国际象棋上击败了人类世界冠军，这是人工智能的一个里程碑。计算机国际象棋程序在自那以后的20多年里继续稳步超越人类水平。这些程序使用人类象棋大师的知识和仔细调整的权重来评估落子位置，并结合高性能的alpha-beta搜索函数，利用大量的启发式和领域特定的适应性来扩展巨大的搜索树。我们描述了这些增强方法，重点关注2016年TCEC世界冠军Stockfish；其他强大的国际象棋程序，包括深蓝，使用的是非常相似的架构。

在计算复杂性方面，将棋比国际象棋更难：在更大的棋盘上进行比赛，任何被俘的对手棋子都会改变方向，随后可能会掉到棋盘的任何位置。计算机将棋协会（CSA）的世界冠军Elmo等最强大的将棋程序，直到最近才击败了人类冠军。这些程序使用与计算机国际象棋程序类似的算法，再次基于高度优化的alpha-beta搜索引擎，并具有许多特定领域的适应性。

围棋非常适合AlphaGo中使用的神经网络架构，因为游戏规则是平移不变的（匹配卷积网络的权重共享结构），是根据棋盘上落子点之间的相邻点的自由度来定义的，并且是旋转和反射对称的（允许数据增加和合成）。此外，动作空间很简单（可以在每个可能的位置放置一个棋子），而且游戏结果仅限于二元输赢，这两者都可能有助于神经网络的训练。

国际象棋和将棋可能不太适合AlphaGo的神经网络架构。这些规则是依赖于位置的（例如棋子可以从第二级向前移动两步，在第八级晋级）和不对称的（例如棋子只向前移动，而王翼和后翼易位则不同）。规则包括远程互动（例如，女王可能在一步之内穿过棋盘，或者从棋盘的远侧将死国王）。国际象棋的行动空间包括棋盘上所有棋手的所有符合规则的目的地；将棋也可以将被吃掉的棋子放回棋盘上。国际象棋和将棋都可能造成胜负之外的平局；实际上，人们认为国际象棋的最佳解决方案就是平局。

AlphaZero：更通用的AlphaGo Zero

AlphaZero算法是AlphaGo Zero算法更通用的版本。它用深度神经网络和白板（tabula rasa）强化学习算法，替代传统游戏程序中所使用的手工编码知识和领域特定增强。

AlphaZero不使用手工编码评估函数和移动排序启发式算法，而是利用参数为θ的深度神经网络（p，v）=fθ（s）。这个神经网络把棋盘的位置作为输入，输出一个落子移动概率矢量p，其中每个动作a的分量为pa = Pr（a | s），标量值v根据位置s估计预期结果z，v ≈E [L| S]。**AlphaZero完全从自我对弈中学习这些移动概率和价值估计，然后用学到的东西来指导其搜索。**

AlphaZero使用通用的蒙特卡洛树搜索（MCTS）算法。每个搜索都包含一系列自我对弈模拟，模拟时会从根节点到叶节点将一棵树遍历。每次模拟都是通过在每个状态s下，根据当前的神经网络fθ，选择一步棋的走法移动a，这一步具有低访问次数、高移动概率和高的价值（这些值是从s中选择a的模拟叶节点状态上做了平均的）。搜索返回一个向量π，表示移动的概率分布。

AlphaZero中的深度神经网络参数θ通过自我对弈强化学习（self-play reinforcement learning）来训练，从随机初始化参数θ开始。游戏中，MCTS轮流为双方选择下哪步棋，atπt。游戏结束时，根据游戏规则，按照最终的位置sT进行评分，计算结果z：z为-1为输，0为平局，+1为赢。在反复自我对弈过程中，不断更新神经网络的参数θ，让预测结果vt和游戏结果z之间的误差最小化，同时使策略向量pt与搜索概率πt的相似度最大化。具体说，参数θ通过在损失函数l上做梯度下降进行调整，这个损失函数l是均方误差和交叉熵损失之和。

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

其中，c是控制L2权重正则化水平的参数。更新的参数将被用于之后的自我对弈当中。

AlphaZero与AlphaGo Zero的4大不同

AlphaZero算法与原始的AlphaGo Zero算法有以下几大不同：

1、AlphaGo Zero是在假设结果为赢/输二元的情况下，对获胜概率进行估计和优化。而**AlphaZero会将平局或其他潜在结果也纳入考虑，对结果进行估计和优化**。

2、AlphaGo和AlphaGo Zero会转变棋盘位置进行数据增强，而AlphaZero不会。根据围棋的规则，棋盘发生旋转和反转结果都不会发生变化。对此，AlphaGo和AlphaGo Zero使用两种方式应对。首先，为每个位置生成8个对称图像来增强训练数据。其次，在MCTS期间，棋盘位置在被神经网络评估前，会使用随机选择的旋转或反转进行转换，以便MonteCarlo评估在不同的偏差上进行平均。而在国际象棋和将棋中，棋盘是不对称的，一般来说对称也是不可能的。因此，AlphaZero不会增强训练数据，也不会在MCTS期间转换棋盘位置。

3、在AlphaGo Zero中，自我对弈是由以前所有迭代中最好的玩家生成的。而这个“最好的玩家”是这样选择出来的：每次训练结束后，都会比较新玩家与最佳玩家；如果新玩家以55%的优势获胜，那么它将成为新的最佳玩家，自我对弈也将由这个新玩家产生的。**AlphaZero只维护单一的一个神经网络，这个神经网络不断更新，而不是等待迭代完成。**自我对弈是通过使用这个神经网络的最新参数生成的，省略了评估步骤和选择最佳玩家的过程。

4、使用的超参数不同：AlphaGo Zero通过贝叶斯优化调整搜索的超参数；**AlphaZero中，所有对弈都重复使用相同的超参数，因此无需进行针对特定某种游戏的调整。**唯一的例外是为保证探索而添加到先验策略中的噪音；这与棋局类型典型移动数量成比例。

奢华的计算资源：5000个第一代TPU，64个第二代TPU，碾压其他棋类AI

像AlphaGo Zero一样，棋盘状态仅由基于每个游戏的基本规则的空间平面编码。下棋的行动则是由空间平面或平面矢量编码，也是仅基于每种游戏的基本规则。

作者将AlphaZero应用在国际象棋、将棋和围棋中，都使用同样的算法设置、网络架构和超参数。他们为每一种棋都单独训练了一个AlphaZero。训练进行了700,000步（minibatch大小为4096），**从随机初始化的参数开始，使用5000个第一代TPU生成自我对弈，使用64个第二代TPU训练神经网络**。

下面的图1展示了AlphaZero在自我对弈强化学习中的性能。**下国际象棋，AlphaZero仅用了4小时（300k步）就超越了Stockfish；下将棋，AlphaZero仅用了不到2小时（110k步）就超越了Elmo；下围棋，AlphaZero不到8小时（165k步）就超越了李世石版的AlphaGo。**

图1：训练AlphaZero 70万步。Elo 等级分是根据不同玩家之间的比赛评估计算得出的，每一步棋有1秒的思考时间。a. AlphaZero在国际象棋上的表现，与2016 TCEC世界冠军程序Stockfish对局；b. AlphaZero在将棋上的表现，与2017 CSA世界冠军程序Elmo对局；c. AlphaZero在围棋上的表现，与AlphaGo Lee和AlphaGo Zero（20 block / 3 天）对战。

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视角下，在比赛中赢，平局或输的局数。经过3天的训练，AlphaZero分别与Stockfish，Elmo以及之前发布的AlphaGo Zero在国际象棋、将棋和围棋分别进行100场比赛。每个AI每步棋都有1分钟的思考时间。

他们还使用完全训练好的AlphaZero与Stockfish、Elmo和AlphaGo Zero（训练了3天）分别在国际象棋、将棋和围棋中对比，对局100回，每下一步的时长控制在1分钟。AlphaZero和前一版AlphaGo Zero使用一台带有4个TPU的机器训练。Stockfish和Elmo都使用最强版本，使用64线1GB hash的机器。AlphaZero击败了所有选手，与Stockfish对战全胜，与Elmo对战输了8局。

此外，他们还比较了Stockfish和Elmo使用的state-of-the-art alpha-beta搜索引擎，分析了AlphaZero的MCTS搜索的相对性能。AlphaZero在国际象棋中每秒搜索8万个局面（position），在将棋中搜索到4万个。相比之下，Stockfish每秒搜索7000万个，Elmo每秒能搜索3500万个局面。AlphaZero通过使用深度神经网络，更有选择性地聚焦在最有希望的变化上来补偿较低数量的评估，就像香农最初提出的那样，是一种更“人性化”的搜索方法。图2显示了每个玩家相对于思考时间的可扩展性，通过Elom量表衡量，相对于Stockfish或者Elmo 40ms的思考时间。AlphaZero的MCTS的思维时间比Stockfish或Elmo更有效，这对人们普遍持有的观点，也即认为alpha-beta搜索在这些领域本质上具有优越性，提出了质疑。

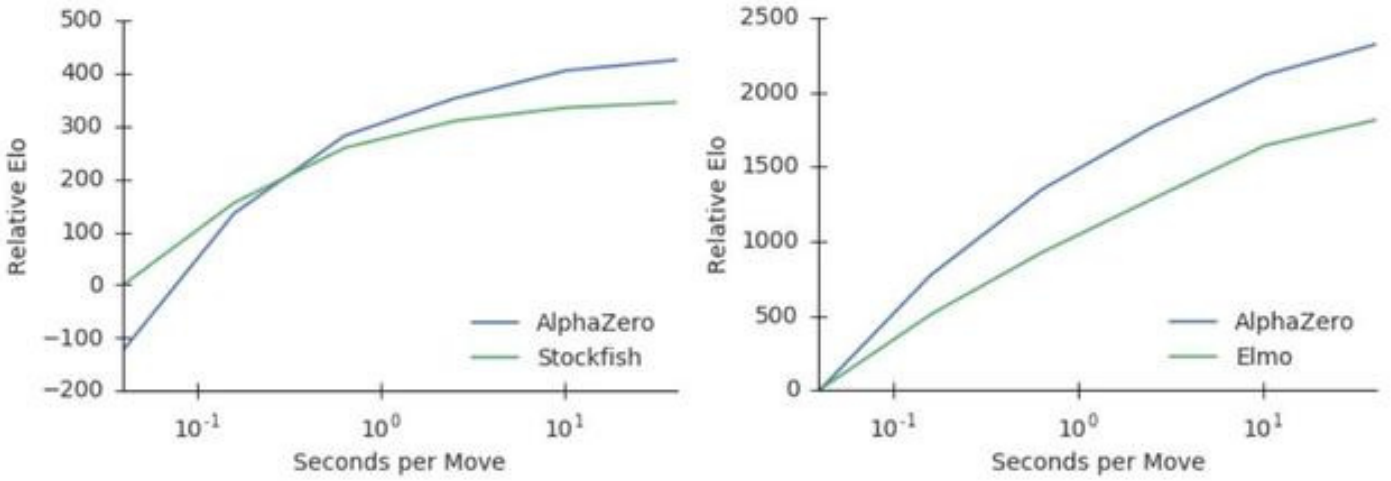


图2：用每步棋的思考时间来衡量AlphaZero的可扩展性，以Elo作为衡量标准。a. 在国际象棋中，AlphaZero和Stockfish的表现，横轴表示每步棋的思考时间。b. 在将棋中，AlphaZero和Elmo的表现，横轴表示每步棋的思考时间。

分析10万+人类开局，AlphaZero确实掌握了国际象棋，alpha-beta搜索并非不可超越

最后，我们分析了AlphaZero发现的国际象棋知识。表2分析了人类最常用的开局方式（在人类国际象棋游戏在线数据库中玩过超过10万次的opening）。在自我训练期间，这些开局方式被AlphaZero独立地发现和对弈。以每个人类开局方式为开始，AlphaZero彻底击败Stockfish，表明它确实掌握了广泛的国际象棋知识。

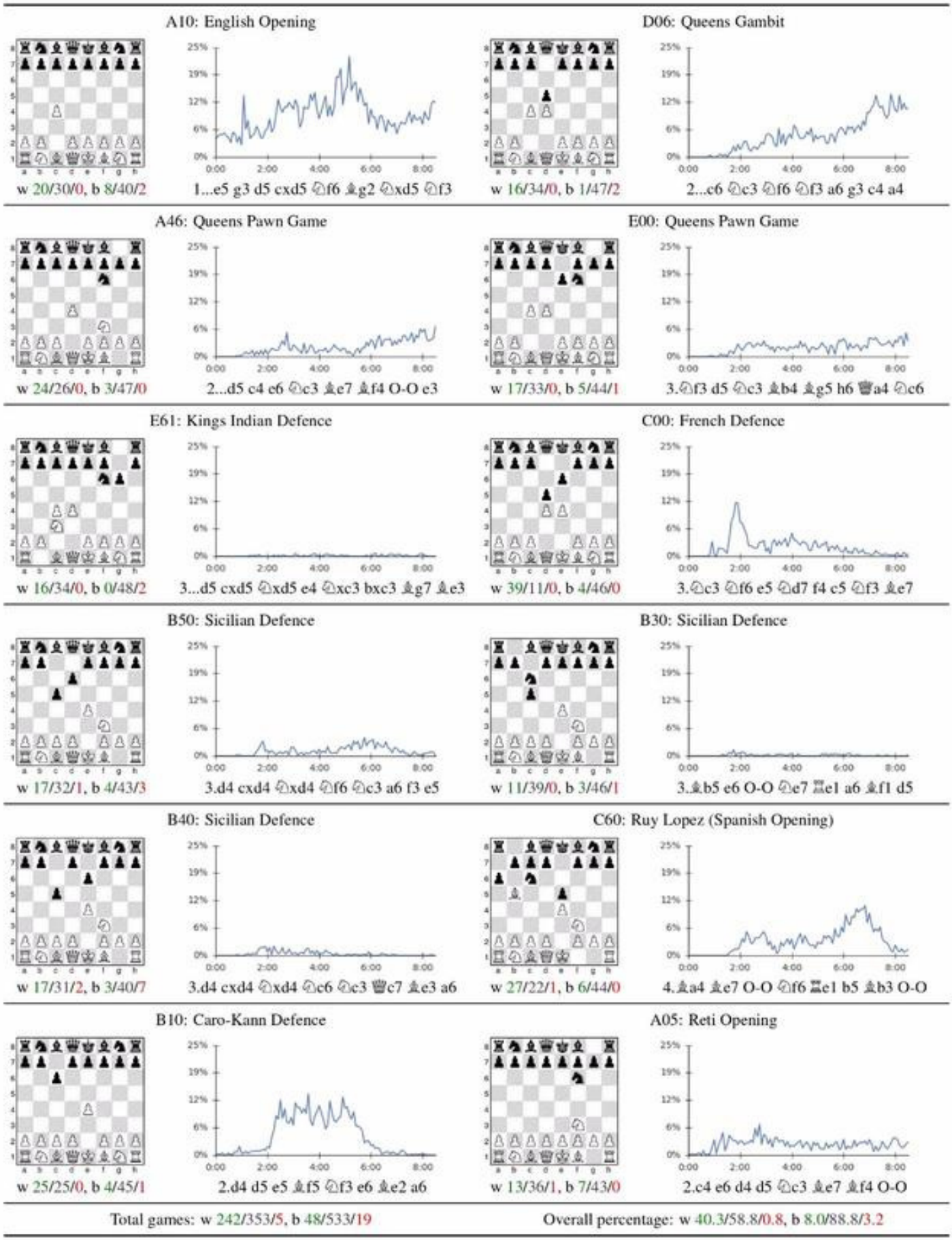


表2：对12种最受欢迎的人类的开局（在一个在线数据库的出现次数超过10万次）的分析。每个开局都用ECO代码和通用名称标记。这张图显示了自我对弈的比例，其中AlphaZero都是先手。

在过去的几十年里，国际象棋代表了人工智能研究的顶峰。State-of-the-art的程序是建立在强大的engine的基础上的，这些engine可以搜索数以百万计的位置，利用人工的特定领域的专业知识和复杂的领域适应性。

AlphaZero是一种通用的强化学习算法，最初是为了围棋而设计的，它在几小时内取得了优异的成绩，搜索次数减少了1000倍，而且除了国际象棋的规则外，不需要任何领域知识。此外，**同样的算法在没有修改的情况下，也适用于更有挑战性的游戏，在几小时内再次超越了当前最先进的水平**。

参考资料

[1] **Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm**：https://arxiv.org/pdf/1712.01815.pdf

[2] **PENG Bo的知乎专栏**：https://zhuanlan.zhihu.com/p/31749249

[3] **陆君慨的知乎回答**：https://www.zhihu.com/question/263681009/answer/271873812

[4] 更多知乎讨论：https://www.zhihu.com/question/263681009/answer/271834015