**Mastering the game of Go without human knowledge**

***David Silver1\*, Julian Schrittwieser1\*, Karen Simonyan1\****

A long-standing goal of artificial intelligence is an algorithm that learns, tabula rasa, superhuman proficiency in challenging domains. Recently, AlphaGo became the first program to defeat a world champion in the game of Go. The tree search in AlphaGo evaluated positions and selected moves using deep neural networks. These neural networks were trained by supervised learning from human expert moves, and by reinforcement learning from self-play. Here we introduce an algorithm based solely on reinforcement learning, without human data, guidance or domain knowledge beyond game rules. AlphaGo becomes its own teacher: a neural network is trained to predict AlphaGo’s own move selections and also the winner of AlphaGo’s games. This neural network improves the strength of the tree search, resulting in higher quality move selection and stronger self-play in the next iteration. Starting tabula rasa, our new program AlphaGo Zero achieved superhuman performance, winning 100–0 against the previously published, champion-defeating AlphaGo.

人工智能领域一个长期的目标就是，在极具挑战的领域，算法自我学习能力的不断提升，超越人类的熟练度。最近Alpha Go成为了第一次在围棋游戏中打败了世界冠军的计算机程序。AlphaGo采取树搜索，和深度神经网络评估棋局和选择落子位置。神经网络通过监督学习学习人类落子的位置，并且使用强化学习完成自我博弈。AlphaGo Zero 只采取强化学习的方法，没有利用任何的人类数据，它对游戏的理解超越了游戏规则的束缚、和人类对围棋领域的理解。Alpha Go无师自通，训练一个神经网络去预测自己的落子，并且赢得与自己的这场比赛。这个神经网络改善了树搜索的能力，使得下一回合能有更高的落子水平和更加强大的自我博弈对手。新的AlphaGo Zero获得了一个超前的性能水平，以100-0赢得了之前的AlphaGo。

AlphaGo Fan used two deep neural networks: a policy network that outputs move probabilities and a value network that outputs a position evaluation. The policy network was trained initially by supervised learning to accurately predict human expert moves, and was subsequently refined by policy­gradient reinforcement learning. The value network was trained to predict the winner of games played by the policy network against itself.

Alpha Go采用两个深度神经网络：一个策略网络，输出落子的概率；一个价值网络，输出对当前棋盘的评估。策略网络最初的训练是通过监督学习的方式，以达到精确预测人类专家的落子为目的进行学习。之后通过强化学习的方法使其改进。价值网络被训练用来预测棋盘的最终的胜利。

Once trained, these networks were combined with a Monte Carlo tree search (MCTS)13–15 to provide a lookahead search, using the policy network to narrow down the search to high­probability moves, and using the value network (in conjunction with Monte Carlo rollouts using a fast rollout policy) to evaluate positions in the tree.

训练完成以后，这些网络就与蒙特卡洛树搜索结合起来提供前向搜索。使用策略网络缩小搜索范围，使得他搜索高概率的落子情况。使用价值网络去评估蒙特卡洛搜索中产生的树的状态(棋盘状态)。

Our program, AlphaGo Zero, differs from AlphaGo Fan and AlphaGo Lee12 in several important aspects.

AlphaGo Zero的程序,与之前版本的AlphaGo在下面几个方面是不一样的：

First and foremost, it is trained solely by self­play reinforcement learning, starting from random play, without any supervision or use of human data.

首要的是，它完全是依据强化学习算法，初始参数随机给定，通过自我对弈从而进行改进，而不需任何的人类下棋数据。

Second, it uses only the black and white stones from the board as input features.

第二，它只采用棋盘中的黑子、白子去构建输入特征(19\*19\*17的数据，表示的是17张大小为19\*19)。

Third, it uses a single neural network, rather than separate policy and value networks.

第三，它使用单个的神经网络，而不是分开的策略网络和价值网络。

Finally, it uses a simpler tree search that relies upon this single neural network to evaluate positions and sample moves, without performing any Monte Carlo rollouts.

最后，它采用一个简单的树搜索结合单个神经网络去评估当前局势和决定落子位置，而没有用任何的蒙特卡洛快速落子方式。

To achieve these results, we introduce a new reinforcement learning algorithm that incorporates lookahead search inside the training loop, resulting in rapid improvement and precise and stable learning.

为了去达到上述结果，AlphaGo Zero中提出了一种新的强化学习算法，在训练过程中结合了前向搜索，使得强化学习能够学习地更加精确和稳定。

**Reinforcement learning in AlphaGo Zero**

Our new method uses a deep neural network fθ with parameters θ. This neural network takes as an input the raw board representation s of the position and its history, and outputs both move probabilities and a value, (p, v) = fθ(s).

AlphaGo Zero只采取一个深度神经网络，这个深度神经网络的输入是原始的棋盘，代表状态s，反应当前以及过去的棋局。输出落子的概率p，以及对当前棋局的评估v。

The neural network consists of many residual blocks4 of convolutional layers16,17 with batch normalization18 and rectifier nonlinearities19 (see Methods).

这个神经网络是由许多残差卷积神经网络所组成的，结合了批标准化，和ReLu等方法防止其过拟合。

The neural network in AlphaGo Zero is trained from games of selfplay by a novel reinforcement learning algorithm. In each position s, an MCTS search is executed, guided by the neural network fθ.

AlphaGo Zero中的训练是通过一种新奇的强化学习算法自我博弈而达到的。在每一个状态s(某个时刻的棋局)，都执行蒙特卡洛树搜索算法。并且蒙特卡洛树搜索算法是由强化学习神经网络所指导的。

The MCTS search outputs probabilities π of playing each move. These search probabilities usually select much stronger moves than the raw move probabilities p of the neural network fθ(s);

蒙特卡洛树搜索算法输出落子概率，通过这种方式搜索出来的落子概率通常是比强化学习神经网络输出的落子概率是要好的。其工作过程如下图所示：

|  |
| --- |
|  |

The main idea of our reinforcement learning algorithm is to use these search operators repeatedly in a policy iteration procedure.

强化学习的主要思想就是在迭代的时候反复利用这些搜索算法。

the neural network’s parameters are updated to make the move probabilities and value (p, v) = fθ(s) more closely match the improved search probabilities and selfplay winner (π, z);

神经网络参数的更新就是使得神经网络的输出值p、v与蒙特卡洛树自我博弈时的输出π，z一样。如下图所示：

|  |
| --- |
|  |

The neural network parameters θ are updated to maximize the similarity of the policy vector pt to the search probabilities πt, and to minimize the error between the predicted winner vt and the game winner z。The MCTS uses the neural network fθ to guide its simulations

神经网络的参数更新是以最大化强化学习输出的策略p和蒙特卡洛输出算法π，以及最小化最终奖励z和强化学习输出的评估值v为目标。蒙特卡洛树以强化学习算法作为指导进行仿真。两者浑然天成，精妙绝伦—建议阅读原论文体会其精妙所在。神经网络损失函数如下图所示：

|  |
| --- |
|  |