# 使AlphaGo零点工作的三个技巧

2017年有深度学习和人工智能，但很少有人能像DeepMind那样引起公众的关注和兴趣。这项计划确实是一个令人震惊的突破：它不仅击败了先前版本的AlphaGo -  - 该计划仅在一年半前击败了17届世界冠军Lee Sedol - 100–0，而且它的训练没有任何来自真实人类游戏的数据。称之为机器学习。

那么DeepMind是怎么做到的呢？在这篇文章中，我将尝试给出一个直观的想法，AlphaGoZero使用的技术，是什么让它们工作，以及对未来人工智能研究的影响。让我们从AlphaGo和AlphaGo Zero玩围棋的一般方法开始。

### DeepMind的一般方法

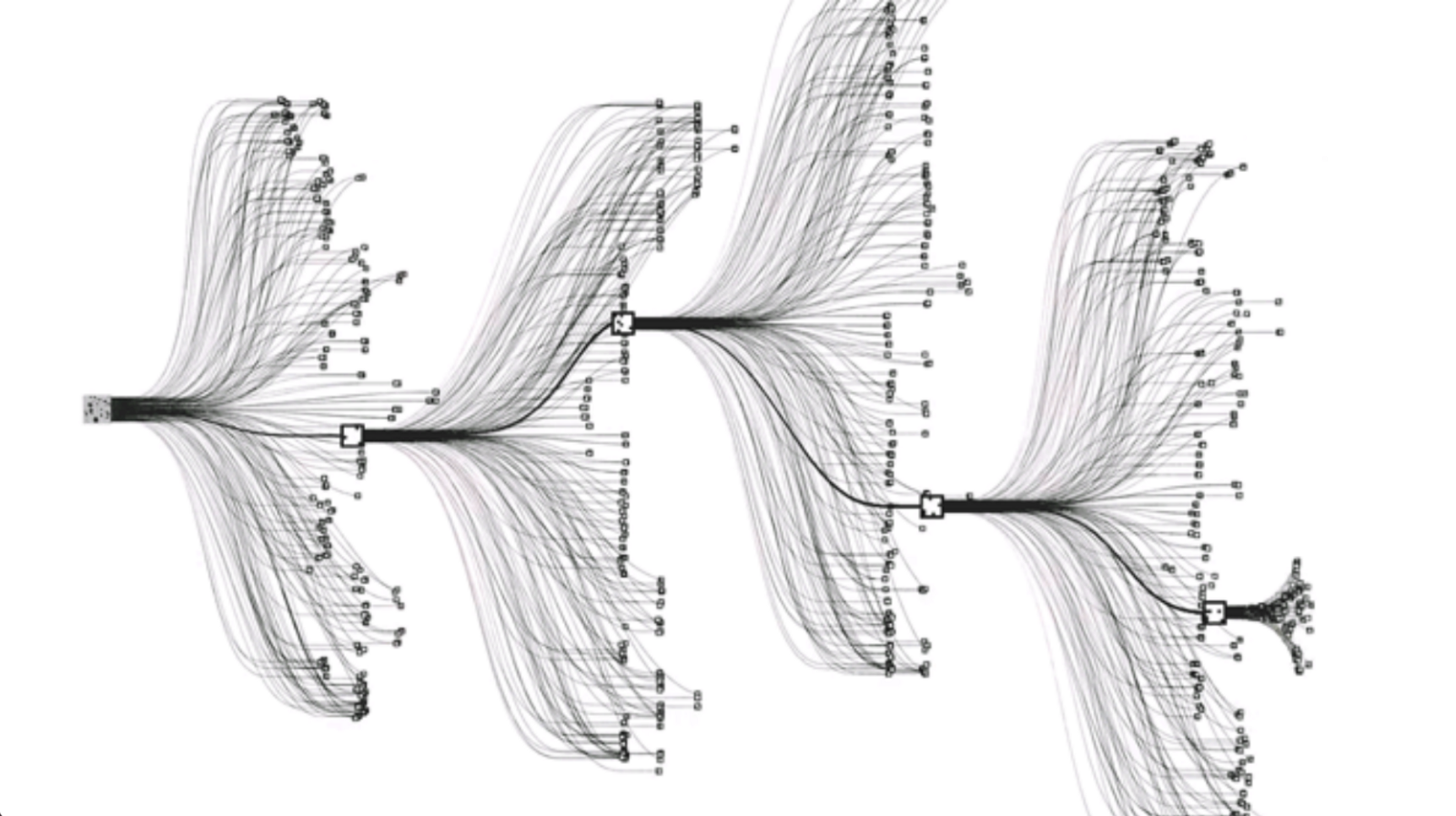
AlphaGo和AlphaGo Zero都评估了围棋板，并使用两种方法的组合来选择移动：

1. 执行“向前看”搜索：通过模拟游戏向前看几步，从而确定当前的哪一步最有可能在未来获得“好”的位置。
2. 根据“直觉”来评估仓位，即仓位是“好”还是“坏”，也就是说，很可能导致输赢。

AlphaGo和AlphaGo Zero都是通过巧妙地结合这两种方法工作的。让我们依次来看每一个：

#### 围棋方法1：“向前看”

围棋是一个非常复杂的游戏，计算机不能简单地搜索所有可能的移动使用暴力方法来找到最好的一个（事实上）。



AlphaGo之前的最佳围棋程序通过使用“”或MCT克服了这个问题。在高层次上，这种方法首先涉及探索董事会中许多可能的举措，然后随着时间的推移，集中精力进行这种探索，因为某些举措比其他举措更有可能导致胜利。

AlphaGo和AlphaGo Zero都使用相对简单的MCT版本进行“展望”，只需使用中列出的许多最佳实践来正确地管理探索新移动序列或更深入地探索已探索序列之间的权衡（更多信息，请参阅中“方法”下的“搜索”部分中的详细信息。

尽管在AlphaGo之前，MCT是所有成功围棋程序的核心，但正是DeepMind巧妙地将这一技术与基于神经网络的“直觉”结合起来，使其超越了人类的表现。

#### 玩法2：“直觉”

DeepMind对AlphaGo的主要创新是使用深层神经网络来了解游戏状态，然后使用这种理解来智能地指导MCT的搜索。更具体地说：他们训练的网络可以

1. 当前董事会职位
2. 哪个球员在比赛，
3. 最近移动的顺序（必须排除某些移动是非法的）

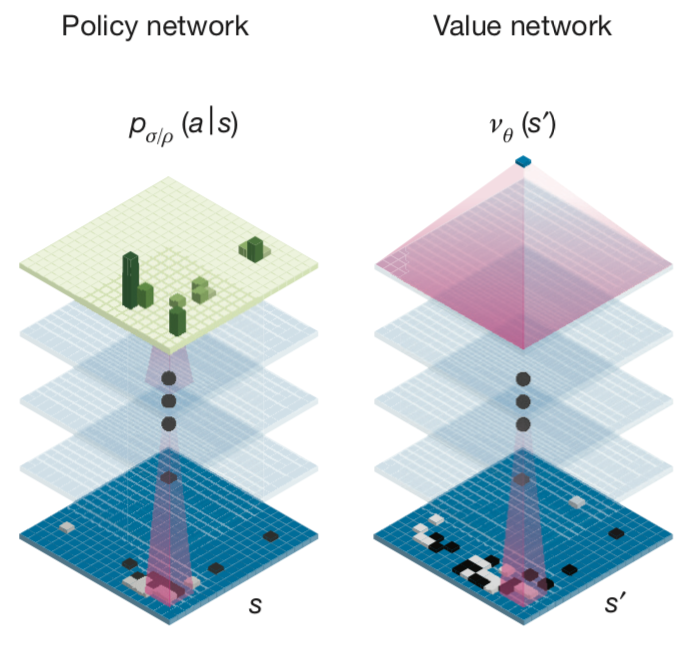
根据这些信息，神经网络可以建议：

1. 哪一步该打
2. 当前玩家是否可能获胜。

DeepMind是如何训练神经网络来做到这一点的？在这里，AlphaGo和AlphaGo Zero使用了非常不同的方法；我们首先从AlphaGo开始：

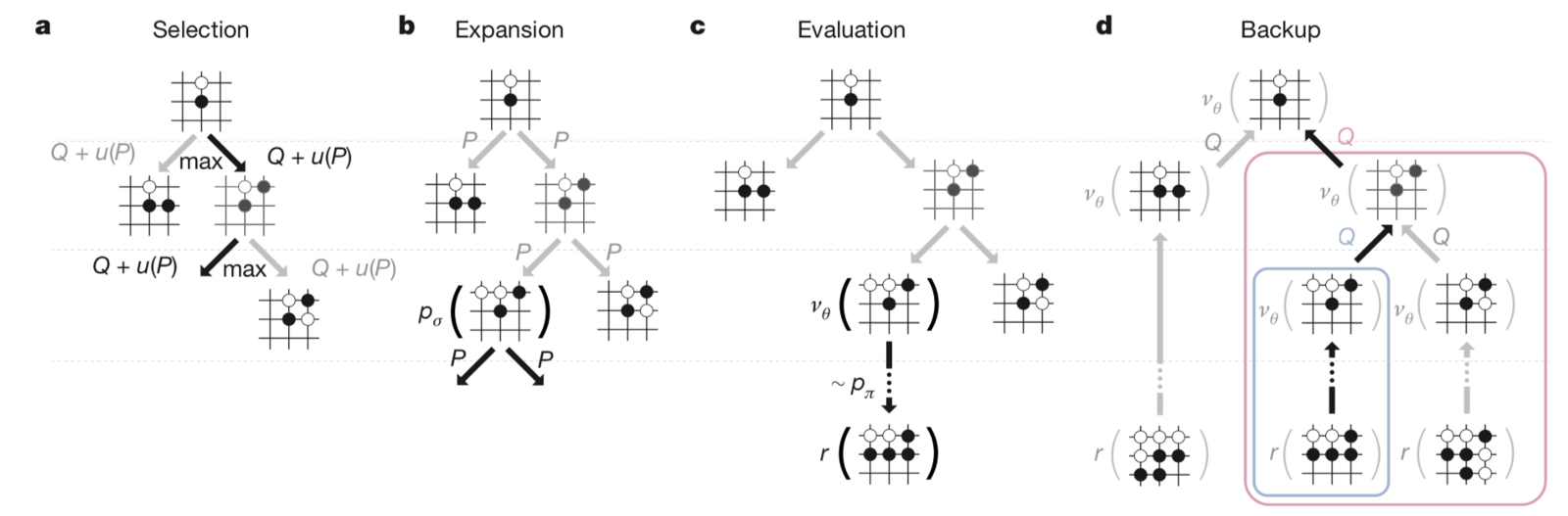
#### AlphaGo的“直觉”：政策网络和价值网络

AlphaGo有两个单独训练的神经网络。



1. 第一个神经网络（随机初始化）被训练成模拟人类专家的游戏，从一个大型真实游戏数据库中显示出3000万个动作。解决这一问题是一个困难但直截了当的模式识别问题；事实上，一旦经过训练，这个网络确实学会了推荐类似于它观察到的人类专家在游戏中的动作。
2. DeepMind不仅仅希望AlphaGo能模仿人类玩家：他们也希望AlphaGo能获胜。为了学习更容易导致胜利而不是失败的动作，这个网络已经被训练成像人类专家一样对自己进行游戏。然后从这些“自玩”游戏中随机抽取动作；如果某个给定的动作发生在当前玩家最终获胜的游戏中，则该网络被训练成更可能在未来玩这种动作，反之亦然。

然后，DeepMind将这两个神经网络与MCTS - 也就是说，程序的“直觉”和它的强力“前瞻”搜索结合起来，以一种非常聪明的方式：它使用经过训练的网络来预测移动，以指导游戏树的哪个分支进行搜索，并使用经过训练的网络来预测一个职位正在“获胜”，以评估它在搜索过程中遇到的职位。这使得AlphaGo能够聪明地搜索即将到来的动作，并最终击败了Lee Sedol。



然而，AlphaGoZero将这一点提升到了一个全新的高度。

### 使AlphaGo零点工作的三个技巧

在较高的层次上，AlphaGo Zero的工作方式与AlphaGo相同：具体来说，它使用基于MCTS的前瞻搜索来播放Go，并由神经网络智能地引导。

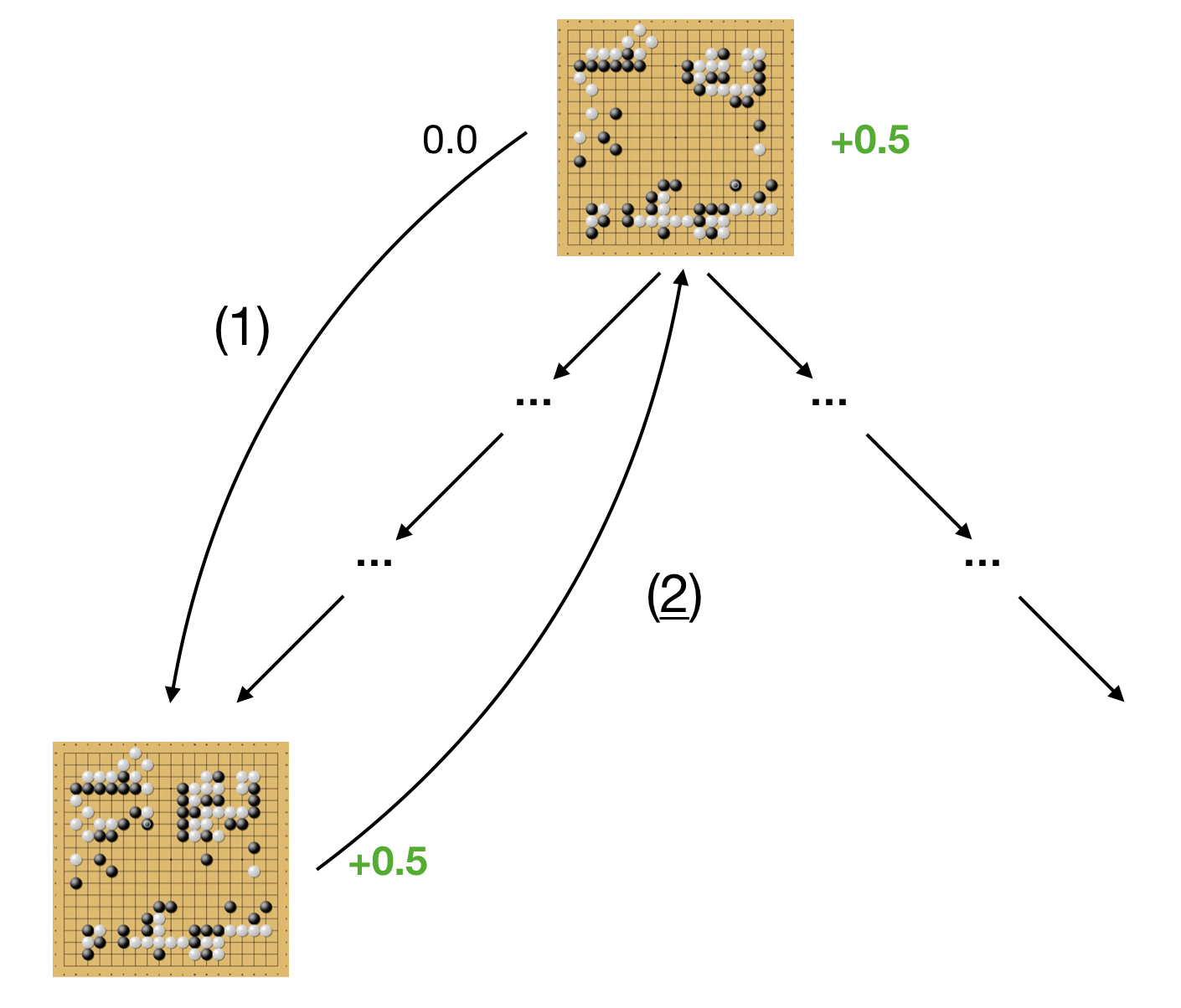
然而，AlphaGo Zero的神经网络“直觉”的训练与AlphaGo完全不同：

#### 诀窍1：如何训练你的字母零，第一部分

假设你有一个试图“理解”围棋游戏的神经网络：也就是说，对于每一个棋盘位置，它都使用一个深度神经网络来生成对最佳动作的评估。DeepMind意识到，不管这个神经网络有多聪明，不管它是完全不知道的还是一个围棋高手，它的评估总是可以通过MCT得到更好的结果。

基本上，MCTS执行的是一种前瞻性搜索，我们可以想象，如果有足够的时间，人类主人会执行这种搜索：它智能地猜测哪些变化——未来移动的序列——最有希望，模拟这些变化，评估它们的实际性能，并更新其对当前最佳移动的评估因此。

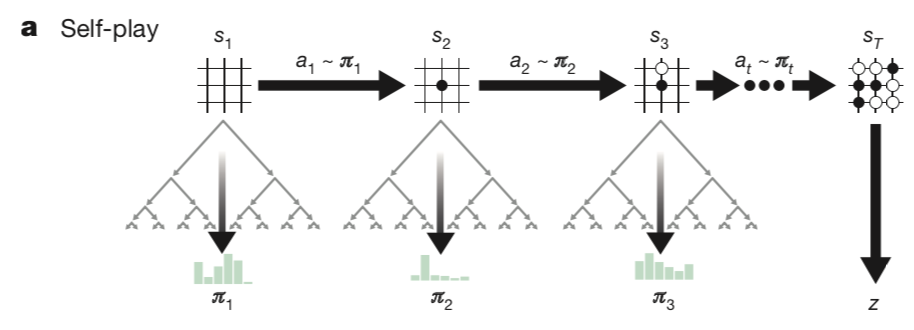
下面是一个例子。假设我们有一个神经网络，它正在读取棋盘，并确定给定的移动结果在游戏中是均匀的，评估值为0.0。然后，网络智能地向前看几步，并找到一系列可以从当前位置强制执行的移动，最终得到0.5的评估。然后，它可以更新对当前董事会职位的评估，以反映它在未来会带来更有利的职位。



因此，这种前瞻性的搜索总是能给我们提供改进的数据，说明在神经网络正在评估的当前位置上，各种移动有多好。无论我们的神经网络是在业余水平还是在专家水平上运行，这都是正确的：我们总是可以通过向前看，看看它当前的哪些选项实际上会导致更好的位置，从而为它生成改进的评估。

#### 诀窍1（续）：如何训练你的字母零，第2部分

此外，正如在AlphaGo中一样，我们也希望我们的神经网络能够学习哪些动作可能导致胜利。所以，和以前一样，我们的代理使用其MCTS改进的评估和其神经网络的当前状态 - 可以对自己玩游戏，赢一些，输一些。



这些数据，纯粹通过前瞻和自我播放生成，是DeepMind用来训练AlphaGo Zero的工具。更具体地说：

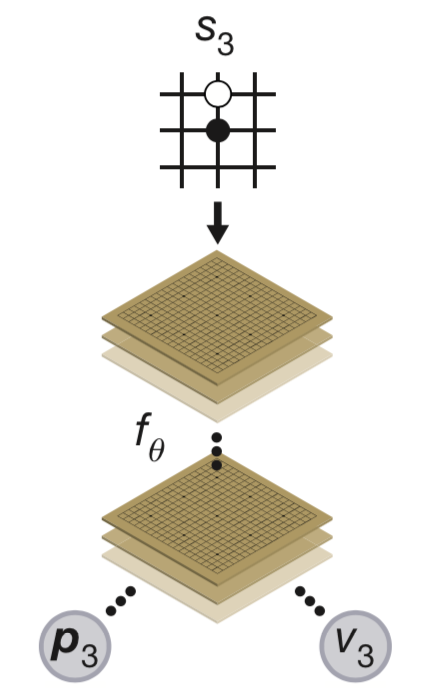
1. 该神经网络被训练来播放反映执行“向前看”搜索改进后的评价的动作。
2. 对神经网络进行了调整，使其更有可能进行类似于导致获胜的动作，而不太可能进行类似于在自玩游戏期间导致失败的动作。

很多人都认为没有人之间的游戏被用来训练AlphaGo Zero，而这第一个“诀窍”就是原因：对于一个Go代理的给定状态，通过执行基于MCTS的lookahead并使用该lookahead的结果来改进代理，它总是可以变得更聪明。这就是AlphaGo Zero能够不断改进的原因，从它是一个业余选手一直到它比最好的人类玩家更好。

第二个技巧是一个新的神经网络结构，我称之为“双头怪物”。

#### 诀窍2：双头怪物

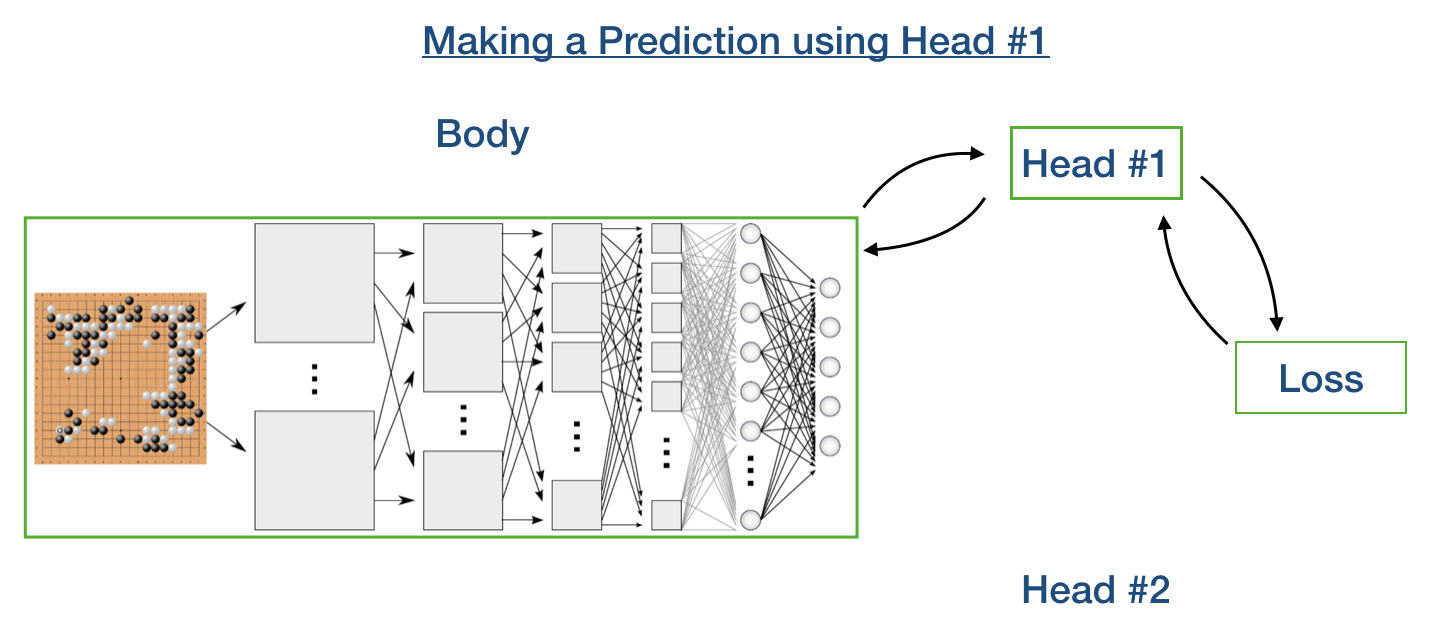
AlphaGoZero是它的神经网络架构，一种“双头”架构。它的前20层左右是一种在现代神经网络架构中常见的类型的层“块”。这些层之后是两个“头部”：一个头部获取前20层的输出并生成围棋代理做出某些动作的概率，另一个头部获取前20层的输出并输出当前玩家获胜的概率。

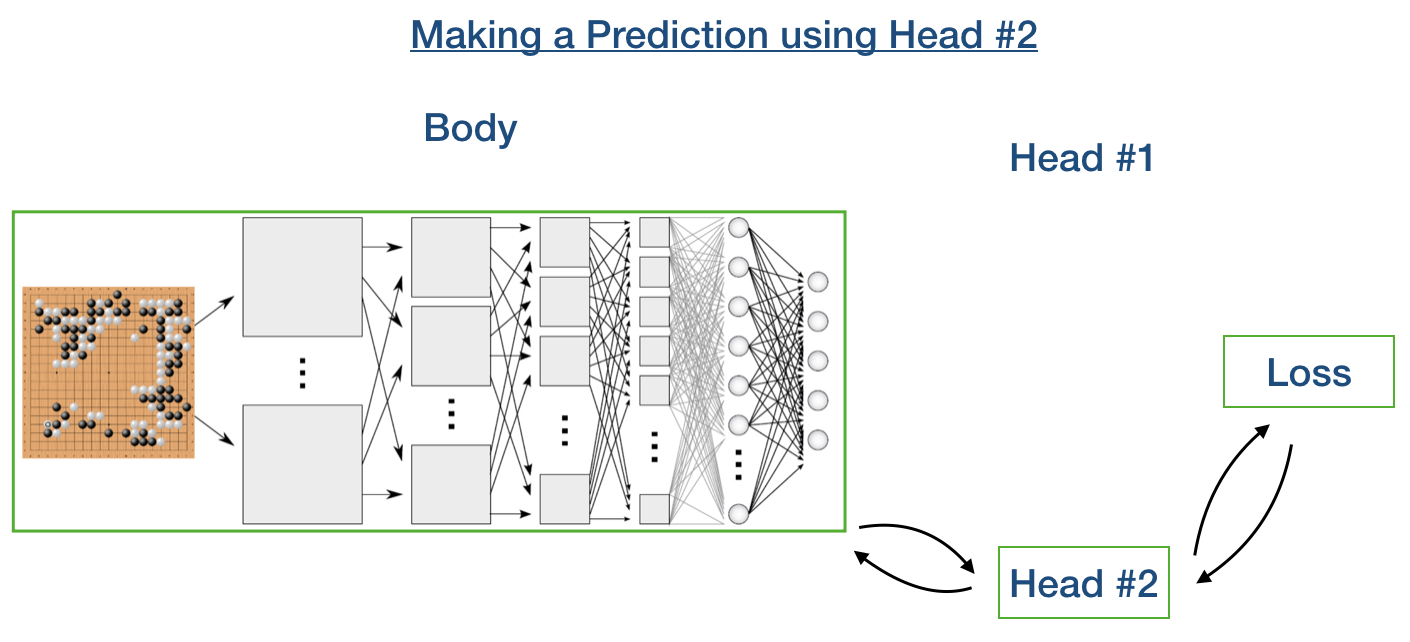


这很不寻常。在神经网络中，输出一个单一的、固定的输出 - ，例如包含狗的图像的概率，或包含包含包含10种对象之一的图像的概率的向量。如果一个网络接收到两组信号：一组是关于它对董事会的评价有多好，另一组是关于它所选择的具体行动有多好，那么它如何学习呢？

答案很简单：记住，神经网络从根本上来说只是一个数学函数，有一堆参数决定它们的预测；我们通过反复向它们显示“正确答案”，并让它们更新参数，使它们产生的答案更接近这些正确答案来“教”它们。

因此，当我们使用双头神经网络使用Head#1进行预测时，我们只需更新导致进行预测的参数，即“Body”和“Head#1”中的参数。类似地，当我们使用Head#2进行预测时，我们更新“Body”和“Head#2”中的参数。





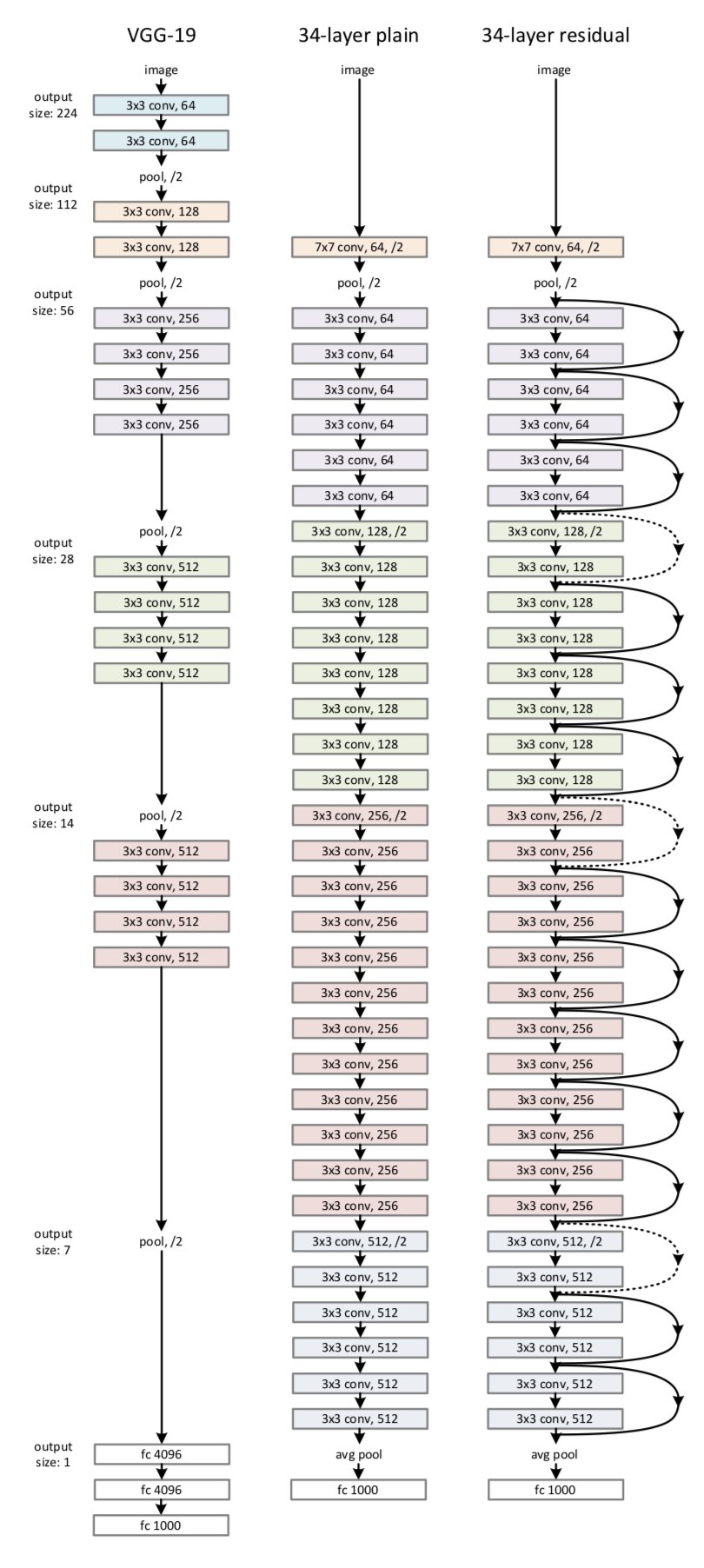
这就是DeepMind如何训练它的单头“双头”神经网络，就像AlphaGo使用两个独立的神经网络一样，它在搜索过程中用来指导MCT。这个技巧占了AlphaGo Zero游戏强度比AlphaGo增加一半的原因。

（此技巧在技术上称为具有硬参数共享的多任务学习。）.

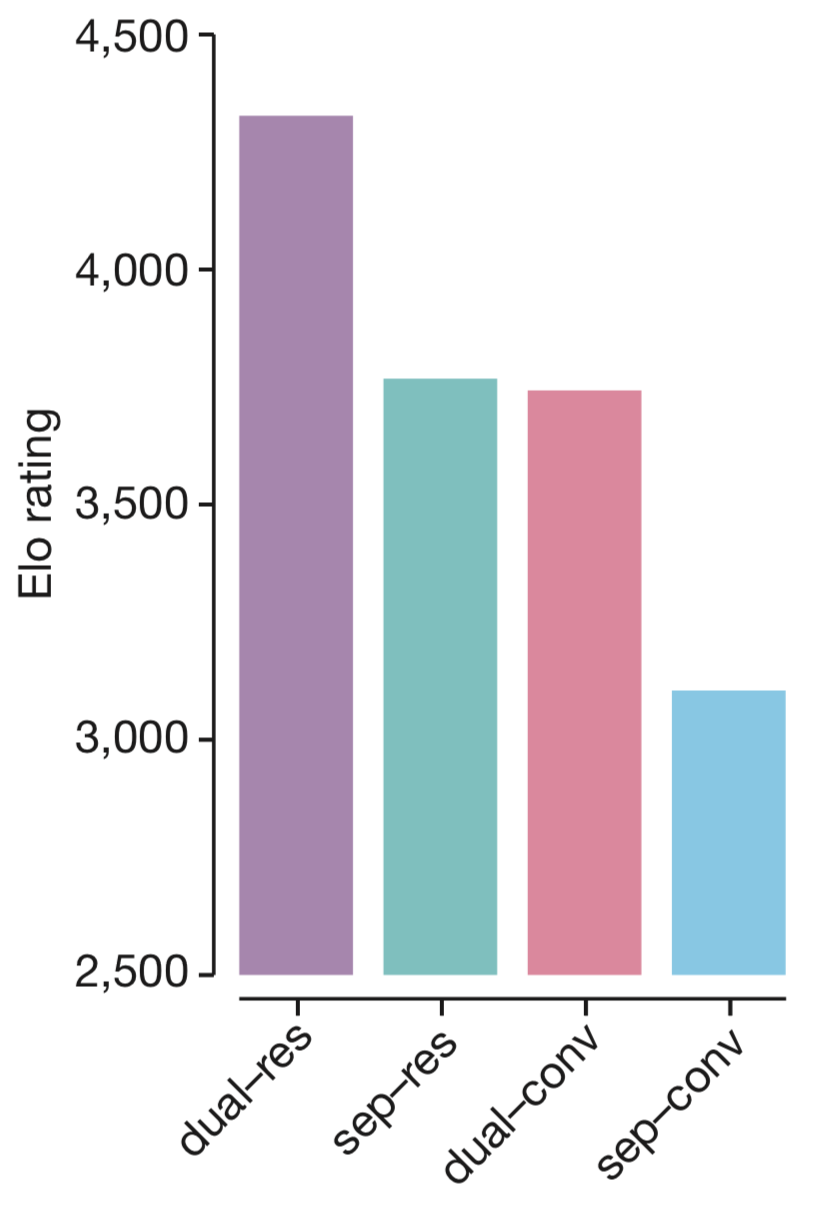
游戏强度的另一半增长仅仅来自于使神经网络架构跟上该领域的最新进展：

#### 诀窍3：“剩余”网

与AlphaGo相比，AlphaGo Zero使用了一种更“尖端”的神经网络架构。具体来说，他们使用的是“残差”神经网络架构，而不是纯粹的“卷积”架构。残留的网络是，就在第一个版本的AlphaGo工作结束的时候，所以DeepMind没有在最初的AlphaGo程序中使用它们是可以理解的。



有趣的是，如下图所示，这两种与神经网络相关的技巧从卷积结构转换到剩余结构，并使用“双头怪物”神经网络结构而不是单独的神经网络结构，这两种技巧的每一种都会导致大约一半的游戏强度的增加，这是当两者结合在一起时实现的。



#### 小结

这三个诀窍使AlphaGo Zero实现了其令人难以置信的性能，甚至让AlphaGo大吃一惊：

1. 利用蒙特卡罗树搜索提供的评价——“智能展望” - 不断改进神经网络对董事会职位的评价，而不是使用人类游戏。
2. 用一个“双头怪物”神经网络同时学习“智能展望”推荐的动作和可能导致胜利的动作，而不是两个独立的神经网络。
3. 使用更尖端的神经网络架构 - 一个“剩余”架构，而不是“卷积”架构。

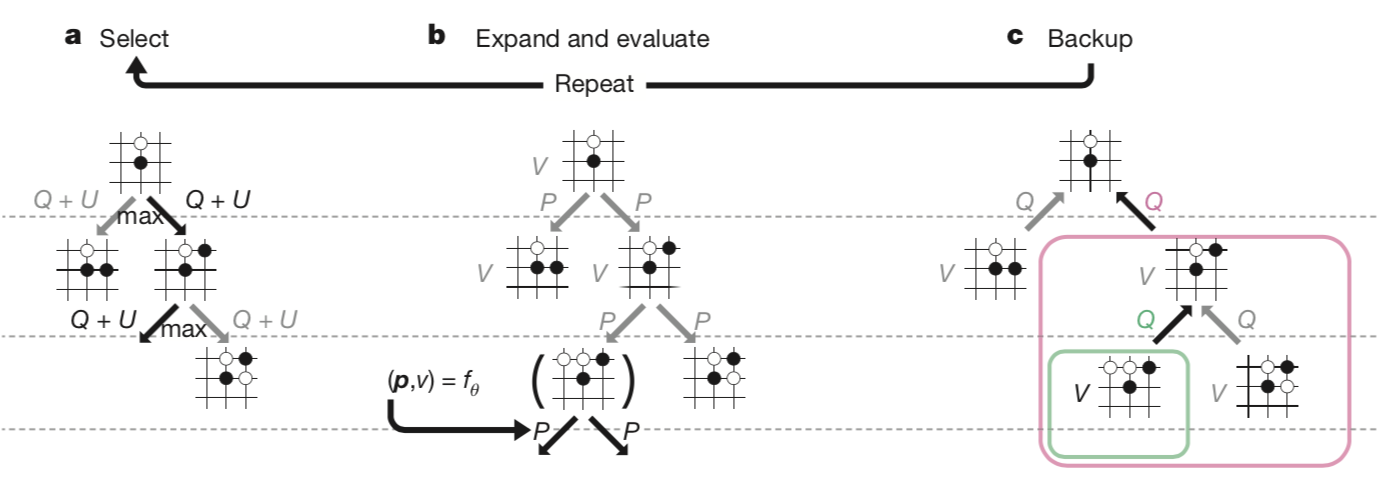
#### 一句话

值得注意的是，AlphaGo没有使用任何经典的、甚至是“尖端的”强化学习概念，没有深度Q学习、异步的参与者-批评者代理，或者我们通常与强化学习相关的任何其他概念。它只是简单地使用模拟来生成训练数据，供其神经网络以有监督的方式学习。在AlphaGo Zero发布后的这篇推文中，我们很好地总结了这个想法：

#### 数字：训练字母归零，循序渐进

以下是AlphaGo Zero培训的“逐步”时间表：

1. 初始化神经网络。
2. 玩自玩游戏，每次移动使用1600个MCT模拟（大约需要0.4秒）。



1. 随着这些自玩游戏的开展，从最近50万场游戏中抽取2048个位置，以及游戏是赢了还是输了。对于每一个移动，记录A）MCTS对这些位置的评估结果 - 这些位置的各种移动的“好”程度取决于向前看 - ；和B）当前玩家是否赢或输了游戏。
2. 训练神经网络，使用A）由MCTS lookahead搜索生成的移动评估和B）当前玩家是赢是输。
3. 最后，每1000次步骤3-4的迭代，对照之前的最佳版本评估当前的神经网络；如果它至少赢得了55%的游戏，开始使用它来生成自玩游戏，而不是先前的版本。

重复第3-4步700000次，三天后，你将拥有一个AlphaGo零！

### 对其他人工智能的影响

DeepMind令人难以置信的成就对人工智能研究的未来有很多启示。以下是几个关键点：

首先，模拟生成的自玩数据“足够好”能够训练网络，这一事实表明，模拟的自玩数据可以训练代理在极其复杂的任务中超越人类的性能，甚至完全从零开始 - 可能不需要人类专家生成的数据。

第二，“双头怪物”技巧似乎能显著地帮助代理学习在许多领域执行一些相关的任务，因为它似乎能防止代理将其行为过度适应任何单独的任务。DeepMind似乎真的很喜欢这个技巧，并使用它和它的更高级版本来构建可以在中学习多个任务的代理。



机器人学中的许多项目，特别是新兴的使用仿真来教机器人代理使用他们的肢体来完成任务的领域，正在使用这两个技巧取得巨大的效果。强调了许多令人印象深刻的新结果，这些技巧与许多前沿强化学习技术一起使用。实际上，移动似乎是“双头怪物”技巧的完美用例：例如，可以同时训练机器人代理使用球棒打棒球和用拳头打移动目标，因为这两项任务需要学习一些常见技能（如平衡、躯干旋转）。



DeepMind’s AlphaGo Zero was one of the most intriguing advancements in AI and Deep Learning in 2017. I can’t wait to see what 2018 brings!