AlphaGo Zero 设计思路及应用实践（上）

前言

AlphaGo Zero 的胜利是启发式搜索（这里是 MCTS）和深度强化学习的胜利，在文章的开始我先大言不惭的立论，具体我说的是不是权威不敢说，但是从这两个角度能够让我们理解（至少自认为理解）AlphaGo Zero 的设计思路。

还有一点需要在前言里说明的是，本次的分享文章会分为上下两篇，上篇主要是讲 AlphaGo Zero 的设计思路，下篇是讲应用实践。分成上下两篇的原因有两个。一是在设立主题的时候没有准确预估到这个话题的复杂性，一篇文章很难讲得清楚。二是这个应用实践的代码目前还没有写好，我又不想直接从 GitHub 上 fork 一个别人参照 AlphaGo Zero 代码实现的棋类应用。闲话少说，接下来我们来聊聊 AlphaGo Zero 的设计思路。

先从人类进化历史说起

人类简史

很早就听说了《人类简史》这本书，也很多次听高晓松在《晓说》中提起。去年的某个周末晚饭后偶然翻起这本书，然后4个小时读完，在读的过程中一直在思考人类是如何学会打猎、学会用火、学会造船、学会生存的？是上帝教的？还是人类本身的基因就是这样，生来全能？显然，这些从历史和科学的考证来看都不是的，当然也不是纯达尔文的进化论所描述的那样。在书中，作者给出了一个很有说服力的答案，是因为人类不光有思维和智力，有语言和文字（知识），在人类长久的与自然环境作斗争的过程中，人类本身的思维能力和智力水平让人类有概率战胜面临的小范围的自然环境。我们假设一个场景，在1000次被野兽即将被吃掉过程中，有一次的概率人类使用了石头战胜了野兽，在原始状态下这种概率几乎等同于随机概率。在岁月的长河中，这种即将被野兽吃掉的事件发生很多很多次，慢慢的人类的思维中有一个方式是可以战胜野兽的，然后就继续发展继续发展，每次遇到相似的困境后相似的方法就会被应用，慢慢的人类就有知识，这些知识是靠语言和文字的方式表达和记录下来。后人在学习这些知识后，然后随着实践的发展和环境的变化又逐步的完善了这些知识，并继续将其保存下来供后人学习借鉴。这就是一个人类逐步解决自己与自然环境之间的关系时所进行的逐步的进步，当然书中的观点对于这种进步是否真的进步有另外的看法，我们这里不讨论人类进步的意义和问题。前面的这些描述，想让大家先对人类这个智能体的进化有一种感觉。因为整个现在 AI 都在思维和意识上来模仿人类，以期能够取得类人类的智能体，AlphaGo Zero 的设计是这种思路的实践。

人类大脑的思维和神经网络

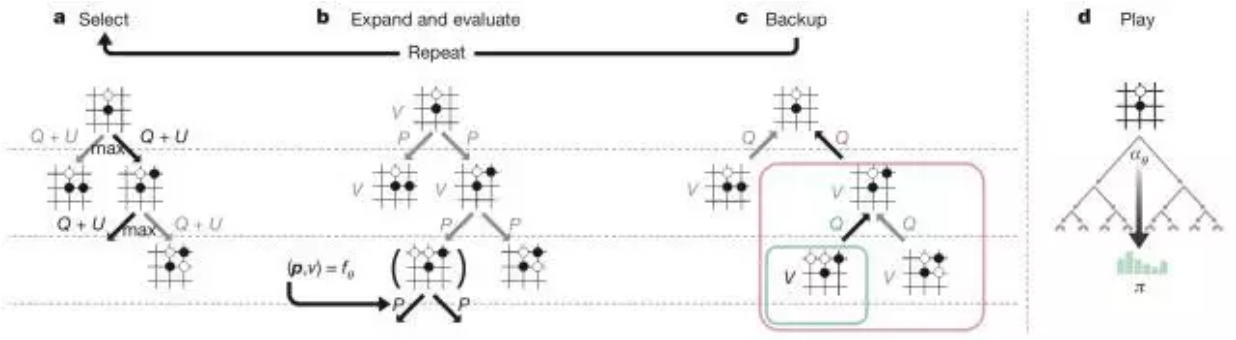
我不研究生物科学，所以这一自然段不是和大家讨论人类大脑的生物构造和神奇的结构，而是从外部行为和自我认知的角度，来非常概要的聊一下我们所能理解的思维和神经网络。这一自然段的目的也是想让大家对思维和神经网络有一个感性的认识，以便更好的理解后续的 AlphaGo Zero 的设计思路。在行为心理学领域有一个非常著名的理论叫 ABC 理论，是从行为心理的角度描述人类从接受到外部环境信号到作出相应反应动作的过程。比如，你走在大街上，过来一个壮汉打了你一拳，这个时候你收到了外界环境信号，有人无缘无故的打了你，这个过程是 A。然后你在作出反击之前，其实我们的大脑已经做各种可能性的评估，比如立即打回去，或者先和他理论为什么打架或者打电话报警，然后你的大脑会作出评估，哪个是对你当前最有利的。如果你更加强壮，你会选择同样打回去，如果你相信法制的力量你会打电话报警，如果你既没有强壮的身体又不想增加太多麻烦，你会忍气吞声不了了之。这个过程就是 B，我们称之为评估。最后的结果是打回去还是打电话报警，就是 C。其实我们的大脑的思维和神经网络的运行从宏观上就是这种 ABC 理论所描述的，而我们进行评估的依据就是我们的经验以及最大利我的这种**意识**，我特地把这个意识给加粗，因为在后面的过程我们会发现这个很有意思，尤其是用神经网络去表示的时候。而这种意识的形成，是人类大脑学习外界环境以及先前知识的过程中形成，AlphaGo Zero 的训练和决策过程和人类的这个过程有同工之妙。

我不确定经过上述两自然段的描述大家能否隐约感受到这种微妙而又神奇的联系和机制，因为我越写我觉得文字限制了我的表达。这部分我们先聊这些，如果有没有找到感觉的可以在微信群分享的时候交流。接下来我们主要来看一下 MCTS 和深度强化学习，大家可以先从网上找点相关资料做个铺垫，因为对于这两神奇事物的介绍网上真的是太多了，我讲 MCTS 和深度强化学习主要是从理解其设计思路和运行机制的角度来讲，让大家能够意会到存在其中的奥妙。

MCTS（蒙特克洛树搜索）

MCTS的作用是什么？

我们这里先说 MCTS 是干什么的，再说 MCTS 是什么，因为这样比较好理解。在现代的围棋程序中，MCTS 都是围棋的核心，因为 MCTS 是用来模拟棋局的。比如再用上面提到的你走在路上被一个壮汉打了一拳来举例，你在进行评估前先在自己大脑里模拟不同反应带来的后果，比如你立马打回去，由于你没有壮汉强壮你就被按在地上摩擦摩擦摩擦……，你大脑里的无意识的联想（模拟）就是相当于进行了 MCTS，然后你根据前一步模拟的状态然后模拟到你被打进了医院，然后有多惨等等，这就是进行多次的 MCTS，最终模拟的结果是你会很惨，然后模拟结束。但是因为围棋的状态很多，在不考虑规则的情况下，第一步之后有360！可能的状态，这是一个不可想象的大树，用超算都要算几十年，所以一般取前30步，然后评估胜率，就返回结果。这是在训练完成之后的状态，在训练状态中，返回评估结果后还会更新 MCTS 的参数以及神经网络的参数。在训练过程中的状态我们来看一下 DeepMind 论文中的介绍图。



关于这个图的详细解释，可以参照很有价值的一篇博客《深入浅出看懂 AlphaGo Zero》。在 AlphaGo Master 以及之前的版本中，都是先用人类的历史对弈棋谱来初始化蒙特克洛树，每个节点上代表着该节点的概率分布和胜率估计。而在 AlphaGo Zero 的版本中，这里不再有人类棋谱作为树节点的初始化，而是完全进行随机的下棋模拟，因为没有先验概率指导，所以初始时基本上都是很臭的棋局，在这个过程中，神经网络会根据棋局结果来学习预测每个落子点概率，在一个棋局状态输入后，首先由神经网络来评估出每个可能落子点的概率 P（P是一个向量），然后将 P 向量作为 MCTS 的先验概率，然后执行 MCTS，这样的话就会通过 MCTS 来模拟每个落子点的胜率，整个网络表达式如下图。注意这里不是 MCTS 的表达式，

$$f_\theta(\vec{\mathbf{s}}) = (\mathbf{p}, v)。$$

总的来讲，我们可以把 MCTS 类比为我们进行联想的过程，这个过程中会用到我们以往的经验和记忆，然后有利于我们能够做出稳定准确的评估。不知道这样说大家能否明白，就是我们知道我们被壮汉按在地上摩擦后会受伤，然后有概率去医院以及不去医院，这个是根据被按在地上摩擦了多久而定的。而我们受伤了需要去医院，这个是我们在以往的生活中累积下来的经验或者叫认知。MCTS 每个节点上的概率和胜率就是我们认为的经验和认知，这不是一个非常严谨的类比，但是能够帮助我们理解这个 MCTS。

MCTS（蒙特卡洛树搜索）是什么？

MCTS（蒙特卡洛书搜索）是什么，或者说蒙特卡洛方法是什么，关于这个问题，现在网上真的有很多解析。因为我不是搞学术研究的，所以要是在这里进行数学表达的推导，说真的我还真做不到。不过这不影响我来讲 MCTS 是什么，我们先说蒙特卡洛方法，蒙特卡洛方法其实是一种随机概率估计在实际过程中的应用。有这样一个非常有名例子来说明蒙特卡洛方法，就是比如有一个完全不规则的图形，我们怎么去计算这个图形的面积那，我们可以将这个不规则图形放到一个规则的矩形中，然后在这个矩形中进行抛色子游戏（完全随机），进行足够多的抛色子，最后我们统计落在不规则图形内色子的概率，然后这个概率与矩形面积的乘积就是这个不规则图形的面积。在生活中其实有很多这样的实例，比如下雨了，你抬起头往上看的过程中不知道大家有没有想过在一定区域内落的雨点是和区域面积有关的，这个也是进行降雨量统计的思路。MCTS 是蒙特卡洛方法在树这种结构上的应用，为什么要应用在树架构上，因为下围棋的过程本身就是一个树。因为围棋状态多到无法计算，就像一个完全不规则图没有办法进行计算一样，我们采用随时在树架构上下棋的方法，来评估每个点的概率和胜率。当然，MCTS实际应用过程中会做一些变种和更新策略来完成所需的概率统计，类比到抛色子计算不规则图像面

积的问题上，就是在抛色子之后总进行当前概率的统计吧，然后会更新落在不规则图形内的概率。

深度强化学习

深度神经网络与 MCTS 完美的形成一对对抗体

在前面介绍 MCTS 的时候，我们知道在训练状态，MCTS 是进行自我更新的，因为随着访问次数和模拟次数的增加每个点的概率分布和胜率是有变化的，那么相应的 MCTS 也会进行更新。类比到人类这里，这就相当于吃一堑长一智一样的道理。而深度学习网络也会进行更新的，因为每次神经网络的输出都是作为 MCTS 的先验概率，而在进行 MCTS 更新的时候，在返回胜率的情况下神经网络也会更新。比如说，在一次神经网络的预测中，在（3，3）点的概率最大，意味着当前状态最应该在（3，3）点落子，然后执行 MCTS 后发现在（3，3）落子的最后是输了，那么预测结果和 MCTS 搜索结果就形成了梯度，然后就进行 backward 来更新深度学习网络。这样的话大家有没有发现，深度学习网络和 MCTS 之间就会一直进行对抗，双方都会根据对抗结果进行更新，深度神经网络进行策略的升级，而 MCTS 进行策略价值的评估。当然，说到这里大家就会联想到 GAN，其实他们的对抗思想是一样的，但是在生成这一块是有点区别的，就是 GAN 是完全随机，是无监督的，而 AlphaGo Zero 是有监督的，它的胜负判断和规则就是最强的监督。

深度强化学习概述

随着 AlphaGo 的横空出世、横扫一切围棋棋局，也同时将人们的视线引入到了深度强化学习上，这也呼应前言中立论中说的 AlphaGo Zero 是启发式搜索和深度强化学习的胜利。强化学习和有监督学习，不一样的地方是强化学习是执行完操作之后才能得到这次操作的反馈或者标记，然后基于这个标记来推动神经网络的更新。而有监督学习是在之前就标好了各个操作的反馈，然后让神经网络去拟合这些结果。所以说，深度强化学习是一个脱离人类知识经验的一个深度机器学习方式，这种方式的好处就是不被人类的认知所误导，而是以实际的结果为导向进行全局的学习和更新。深度强化学习的好处，就是可以打破人类认知的局限以及学习无法进行大量标记的问题求解，它是一个全局搜索且随机的一个过程，不好的地方就是我们没有办法去预测学习的结果以及需要大量的算力（因为是全局进行随机尝试的，但是会根据反馈把对的给更新到神经网络中保存下来）。

应用实践的计划

如前言中所说，这次分享包含上下篇，上篇是讲设计思路，下篇是应用实践。应用实践的初步想法是这样的，我会尝试将 AlphaGo Zero 的设计思路应用到一个类似的实际问题求解上，比如城市道路规划等。我不确定上面讲的大家能否理解或者体会到，当然里面为了能够讲的更加容易理解有一些不严谨的类比，只是想让大家感觉这种神奇思路。