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

研究现状

本文是Alpha Zero的论文。

当前的棋类智能算法，包括之前的AlphaGo、AlphaGo Zero，打过依赖于其领域知识，难以泛化到其他领域。

Alpha Zero比AlphaGo Zero更加通用，本文把Alpha Zero分别运用到围棋、国际象棋和日本将棋，而算法只使用同一个深度神经网络。

研究方法

算法跟Alpha Zero基本一样，使用神经网络进行Self-Play，通过MCTS进行走子搜索。

神经网络输入输出：$(p, v) = f_\theta(s)$

损失函数：$l = (z - v)^2 - \pi^T log p + c||\theta||^2$

跟AlphaGo Zero区别如下：

1. AlphaGo Zero只处理输/赢两种结果，Alpha Zero会处理输、赢和平局，或者其他对战结果；
2. AlphaGo Zero对训练集做数据增强，对棋盘上的位置做随机对称变换，如旋转、翻转棋盘等，但因为国际象棋和日本将棋的规则是不对称的，所以Alpha Zero不做这样的数据增强

3. AlphaGo Zero用当前最优的神经网络进行Self-Play，Self-Play后进行网络迭代，迭代后的新网络会跟当前最优的网络再进行对战，当胜过当前最优网络55%的时候，会用最新迭代的网络代替最优网络进行下一步的Self-Play，如此循环。但Alpah Zero只会对当前网络进行不断迭代更新，Self-Play也只会用最新的网络参数进行，不会进行网络性能评估和用最优网络替代下一轮的Self-Play过程。
4. 对于每种对战游戏，Alpha Zero会复用网络的参数，不会针对特定游戏进行网络微调。

神经网络的输入特征也有所不同。Alpha Zero输入是N x N x (MT+ L)个棋盘图像特征：

- N x N为棋盘大小，如国际象棋为8 x 8；
- M是双方各自的棋局特征，譬如对于围棋，M为2：我方棋子布局（格子上是我方则为1，否则为0，类似对方也是如此）
- L为其他特征，如围棋即为当前哪方走子，我方在棋盘格子全为1，对方则全为0
- T为最近T个时间步的棋盘状态

具体输入特征描述如下：

Go		Chess		Shogi	
Feature	Planes	Feature	Planes	Feature	Planes
P1 stone	1	P1 piece	6	P1 piece	14
P2 stone	1	P2 piece	6	P2 piece	14
		Repetitions	2	Repetitions	3
				P1 prisoner count	7
				P2 prisoner count	7
Colour	1	Colour	1	Colour	1
		Total move count	1	Total move count	1
		P1 castling	2		
		P2 castling	2		
		No-progress count	1		
Total	17	Total	119	Total	362

另外每次走子的动作也有所区别，分为两步：选子和落子位置。为此，论文将策略$\pi(a|s)$表示为（如对于国际象棋）8 x 8 x 73个特征：

- 8 x 8为选子的位置
- 73为：56个特征是类似与皇后类型（如皇后、相、车）的走法：[1..7]为落子的位置，8个走的方位{N, NE, E, SE, S, SW, W, NW}；另外8个特征是国王的走法；还有9个是卒的走法。

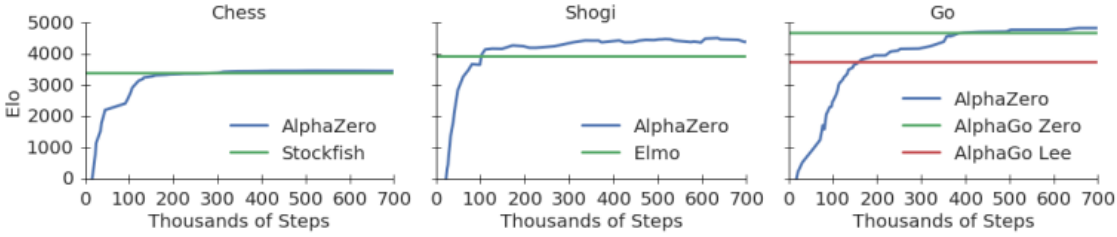
具体描述如下：

Chess		Shogi	
Feature	Planes	Feature	Planes
Queen moves	56	Queen moves	64
Knight moves	8	Knight moves	2
Underpromotions	9	Promoting queen moves	64
		Promoting knight moves	2
		Drop	7
Total	73	Total	139

研究结论

论文采用如下公式衡量算法的（Elo分值）：玩家a打败玩家b的概率为$p(adefeatsb) = \frac{1}{1+exp(c_{elo}(e(b)-e(a)))}$， 其中e(·)是贝叶斯逻辑回归，$c_{elo} = \frac{1}{400}$ ，

各算法效果如下：


Chess
Elo
5000
4000
3000
2000
1000
0
AlphaZero
Stockfish
0 100 200 300 400 500 600 700
Thousands of Steps
Shogi
AlphaZero
Elmo
0 100 200 300 400 500 600 700
Thousands of Steps
Go
AlphaZero
AlphaGo Zero
AlphaGo Lee
0 100 200 300 400 500 600 700
Thousands of Steps