作业4实验报告

231240002余孟凡 231240002@smail.nju.edu.cn
南京大学计算机科学与技术系, 南京 210093

摘要

关键词: Alpha GO

1 阅读论文

公式其实不是太懂，但大概流程是：

1. 首先，AlphaGo训练了一个监督学习（SL）策略网络，目标是预测人类专家在特定棋盘状态下的最佳移动。这个网络由多个卷积层组成，输入为棋盘状态的图像表示，输出为每个合法移动的概率分布。

2. 在监督学习之后，AlphaGo通过自我对弈来进一步优化策略网络。使用强化学习（RL）策略网络，AlphaGo与之前版本的策略网络进行对弈，优化目标是赢得比赛而非仅仅提高预测准确性。

3. AlphaGo还训练了一个价值网络，用于评估棋盘状态的胜率。这个网络的结构与策略网络相似，但输出的是一个标量值，表示当前状态的预期结果。

4. AlphaGo结合了策略网络和价值网络与MCTS算法。MCTS通过模拟多次游戏来评估每个状态的价值，并在搜索树中选择最优动作。

2 阅读代码

从最外层代码开始看起：

rl_loop.py ：创建两个随机策略的代理 agents，使用 RandomAgent 类。然后进行对弈，在每局对弈中，重置环境 env.reset()，获取初始状态 time_step。在每一步中，调用对应代理的 step 方法选择动作 action_list。环境执行该动作 env.step(action_list)，返回新的状态 time_step。

参考文献

（Mastering the game of Go with deep neural networks and tree search）
（Technion – Israel Institute of Technology Project: Mini Alpha Go ）