作业 10

说明：

参考作业中给出的源码以及 AlphaZero 的伪代码实现。

算法描述：

1) 自对弈：使用 mcts 在每次对弈的时候搜索，并且记录下每次搜索获得的下一步行为的分布，作为策略输出的训练目标。在每一局对弈结束后，记录下每一步时的玩家是胜利还是失败，作为价值输出的训练目标。将每一步的状态保存入缓存以便训练。

2) 模型训练：从缓存中获取一定 batch 的数据，使用均方误差衡量价值输出的 loss，使用 kl 距离衡量策略输出的 loss，将两者的 loss 相加作为最终的 loss 训练模型。

结果测试：

我使用自己训练的模型之间对弈，总计训练了 2000 局。在测试时随机选择先手，并且前两步带有一定的随机性，所以对弈结果存在一定的随机性。从结果看训练 1500 局时的模型最强，2000 局时的模型有一定退化。

- net_99.model VS net_1999.model

模型	胜利	失败	平局
99	15	85	0
1999	85	15	0

- net_499.model VS net_1999.model

模型	胜利	失败	平局
499	29	62	9
1999	62	29	9

- net_999.model VS net_1999.model

模型	胜利	失败	平局
999	42	53	5
1999	53	42	5

- net_1499.model VS net_1999.model

模型	胜利	失败	平局
1499	58	38	4
1999	38	58	4

- net_1999.model VS net_1999.model

模型	胜利	失败	平局
1999	49	45	6
1999	45	49	6