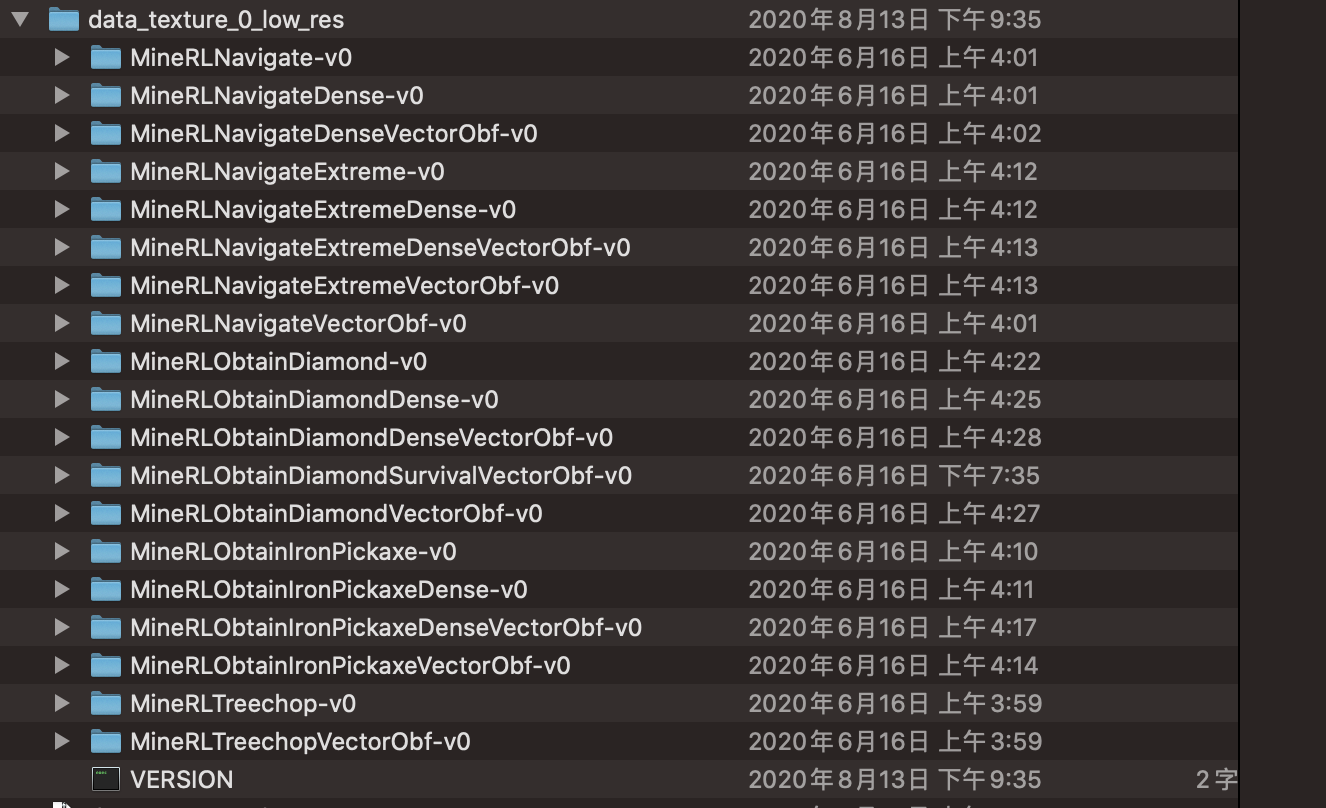
机器学习大作业作业报告。

1. 通过阅读参赛规则我了解到整个比赛的要求是通过机器学中的强化学习来拿官方给到的样本来训练使得在mc内的人物可以成功挖到钻石（我暂时不考虑代价）
2. 比赛需要训练的数据为MineRLObtainDiamondVectorObf-v0



Observation Space为

Dict({

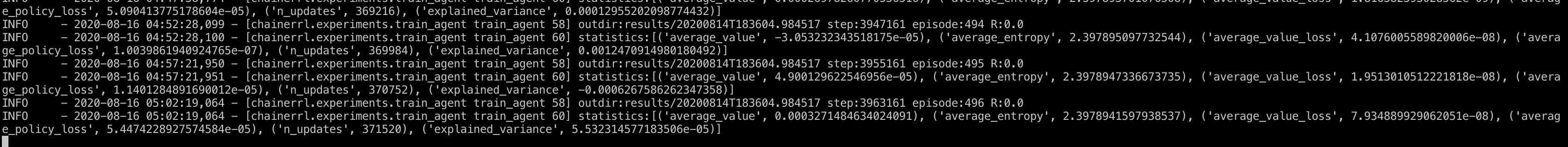
"[pov](https://minerl.io/docs/environments/handlers.html" \l "pov)": "Box(low=0, high=255, shape=(64, 64, 3))",

"vector": "Box(low=-1.2000000476837158, high=1.2000000476837158, shape=(64,))"

})

1. 在了解到这些基本知识点后，我在本机环境下下载了基线baselines包并完成了配置。

运行了ppo.sh 有关MineRLObtainDiamond-v0部分。

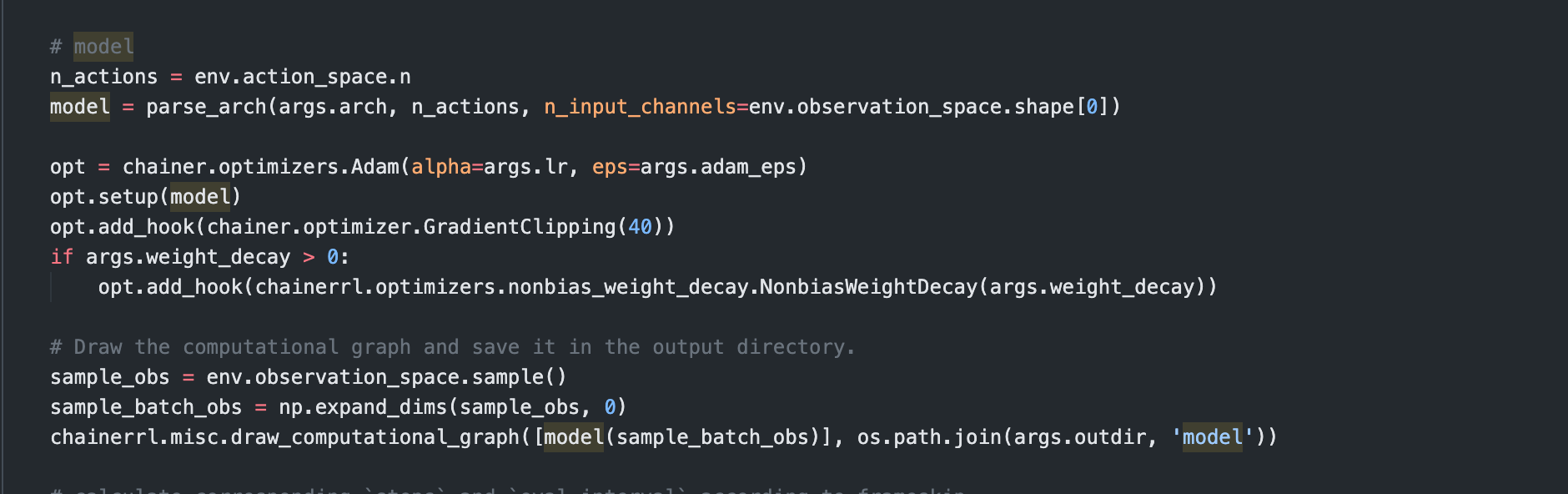


由于本机在运行我讲gpu设为-1

同时可以看到下图中的图形界

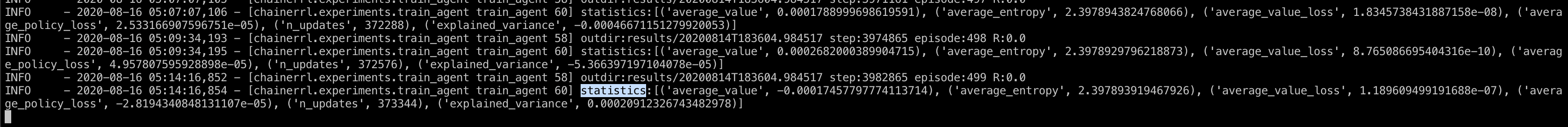


从ppo的定义中我可可以知道，策略pi接受状态s，输出动作概率分布，在动作概率分布中采样动作，执行动作，得到回报，跳到下一个状态。在这样的步骤下，我们可以使用策略pi收集一批样本，然后使用梯度下降算法学习这些样本，但是当策略pi的参数更新后，这些样本不能继续被使用，还要重新使用策略pi与环境互动收集数据，真的非常耗时。于是PPO拿出杀手锏--重要性采样，使这些样本可以被重复使用。



从baseline中可以看出

在执行500个episode的结果



可以看出相邻的episode的所得参数更新慢，跳变幅度大，但是与刚开始训练来对比的话，反而变化比较大，其实这个问题是他在对同一个样本进行了多次重复的采样导致的，