开悟比赛-DCD 队技术整理与分享

李佳晖 浙江大学计算机学院

李星晨 浙江大学计算机学院

田琪 浙江大学计算机学院

刘宇泽 浙江大学计算机学院

指导老师:况琨

一、 简介

在 2022 年 9 月-2023 年 4 月举办的腾讯第三届开悟 MOBA 多智能体强化学习大赛中，我们队伍 DCD 在初赛中有幸获得了第 3 名的成绩。这得益于实验室老师与腾讯官方的大力支持， 也得益于我们队伍本身的技术探索与积累。本文首先会简单叙述比赛的基本情况，随后从各关键模块出发，简要介绍本队伍在开悟比赛中的探索历程与心得体会。

二、 参赛概况

本队成员来自浙江大学计算机科学与技术院。

四名队员皆为博士研究生，由于报名时间临近截止，仓促组队。

其中李佳晖队长研究方向为多智能体强化学习，比较契合本次比赛。

李星晨同学研究方向为视觉场景理解

田琪同学研究方向为对抗学习

刘宇泽同学研究方向为黑盒优化

队长的参赛动机是想了解强化学习在工业界的应用，

其余队员的参赛动机是想对强化学习相关技术进行学习

本次比赛中，由队长全程操作。

队长每天在比赛中花费 0-10 小时时间，平均每天花费 2.5 小时以上

其余队员在科研之余进行观摩

队长每周在零散的时间内提交训练，最长以 24 小时为周期提交一次训练，最短以 1 小时为周期提交一次训练，提交任务仅有超参数上的改动。

三、网络设计

与腾讯官方论文一致，并未修改，详见论文[1]

四、奖励体系

总体思路上与腾讯官方论文一致，与官方论文不同的是本次训练并不根据训练的前中后期将奖励分为三个阶段 r1，r2，r3

而是使用函数 r=wx，x 是原有奖励，r 是计算后的奖励，w 是系数（0<w<2），随时间改变（或增大或减少）。

每种奖励都设置 lower bound 和 upper bound，根据训练状况调整，仍未发现最佳的超参，奖励有效性未知。

例如，在游戏的前期金钱比较关键，而在后期装备已经成型了，金钱就不那么关键了。所以设置关于金钱的奖励的 w 在 10000 个 step 之内由 1.5 逐步衰减到 0.1

而游戏的最终目的是拆塔，所以关于攻击防御塔的奖励的 w 在 10000 个 step 中由 0.8 增至 1.6

具体的超参数仍未找到最佳，比赛中盲目调整是失败的一大关键因素，建议学习黑盒优化理论。

五、特征与规则

特征和腾讯官方论文完全一致，无任何附加规则

召唤师技能全带狂暴。

六、 强化学习算法

1. 算法和腾讯官方论文论文一致 dual clip PPO

2. 因为 PPO 算法超参数太多，所以寻找一组最合适的超参数才是最关键的。

关键的参数包括奖励，学习率，折扣因子，熵系数 beta，GAE 相关参数

3. 其余技巧包括模型设计，exploration 设计，balance exploration-exploitation 尝试过但没太大用处。

   Exploration 的尝试方法是最常用的 prediction error，即建立另一个网络，输入为当前时刻智能体的 state 以及 acrion，去估计 next step 的 Q 或 state 或 reward. 由于王者荣耀是复杂场景 state 空间过于庞大，所以这种探索策略不奏效。

在模型上，希望在 lstm 或预测层之前，对 concat 的特征加双层 attention，第一层是粗粒度 attention，attention 的每一个纬度对应于不同类别的特征，如敌人状态，自身状态，防御塔状态。第二层是细粒度 attention，attention 每一个纬度对应于特征的每一个纬度，Attention，对应于每个特征的每一个纬度，但是并未实现。

七、 系统工程架构

开悟平台提供了基于规则的 common_AI 用于与训练模型对战生成胜率曲线。

我们将 common_AI 替换成之前训练阶段的 3-5 个模型，随机 load ckpt，以观察模型是否有提升。

八、 模型迭代过程

1. 根据测试情况调整各个奖励大小，在击杀，推塔，赚钱之间平衡。

   若经常死亡，则增加死亡惩罚项，若英雄经常被消耗，则增加生命值相关奖励，若不拆塔，则增加防御塔相关奖励，若难以击杀对面，则增加击杀相关奖励
2. 逐步减小学习率，由 5e-4 逐步减少为 5e-7
3. 逐步减小熵参数 beta，用于控制 exploration，由 default 逐步减少为 1e-6
4. 下一次的训练使用 best checkpoint

总体步骤

(->训练->批量下载 ckpt->批量测试->选择最佳 ckpt->调整超参->)

九、 训练效果分析

每次训练结束，与不同 baseline 对战，看结果中的数据统计，取 vs baseline 胜率高就行。

十、 总结与展望

RL 训练分析：

1. 模型迭代到一定的地步便无法进一步提升。可能受限于 PPO 算法特性，其次可能受限于 RL 本身的 unstable 问题。

2. 已尝试过一些 exploration 算法，不奏效。因为赛题可选择的优化方向太多，显得 exploration 并不是能提升性能的主要原因，放弃尝试。

3. 由于操作繁琐，test 时间过长，tf 代码修改困难等原因，并没有设计精致的模型以及算法，但是精致的模型以及算法确实可能会提升模型的上限。

4. 模型的迭代受到上一次迭代的影响，所以在第一周训练参数的不合适可能会导致一步错步步错的结果，很难修正回来。需要观看官方提供的最佳模型在比赛中的表现，决定训练一开始的超参数，主要是各项奖励的设置。 例如，官方模型的最强模型在前期偏向于补刀而不是压制对面英雄血量，则将金钱和经验相关的奖励设置的高一点。

本队失败总结：

1. 算力给的实在太多了，比赛时间太长，本队的训练迭代方式并不适合比赛，在比赛中后期，模型经常会遇见瓶颈，超参数经常需要调整，难以继续迭代流程。
2. 测试成本以及训练成本使得队伍不愿去 search 以及寻找一个合理的模型架构
3. 第一次参赛毫无经验，每周都在浪费算力，前两周几乎什么都没干，还在熟悉比赛。

参考文献

[1]YeD,LiuZ,SunM,etal.Masteringcomplexcontrolinmobagameswithdeepreinforcementlearning[C]//ProceedingsoftheAAAIConferenceonArtificialIntelligence.2020,34(04):6672-6679.

[2]ZhaoE,YanR,LiJ,etal.AlphaHoldem:High-PerformanceArtificialIntelligenceforHeads-UpNo-LimitPokerviaEnd-to-EndReinforcementLearning[C]//ProceedingsoftheAAAIConferenceonArtificialIntelligence.2022,36(4):4689-4697