# 开悟比赛-四川大学队技术整理与分享

陈建明 四川大学计算机系

薛明峰 四川大学计算机系

苏裕杰 四川大学计算机系

李知谦 四川大学计算机系

陈泽西 四川大学计算机系

指导老师: 吕建成、周吉喆、汤臣薇、李茂

## 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(你说就是你队)在初赛中有幸获得了第 8 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。

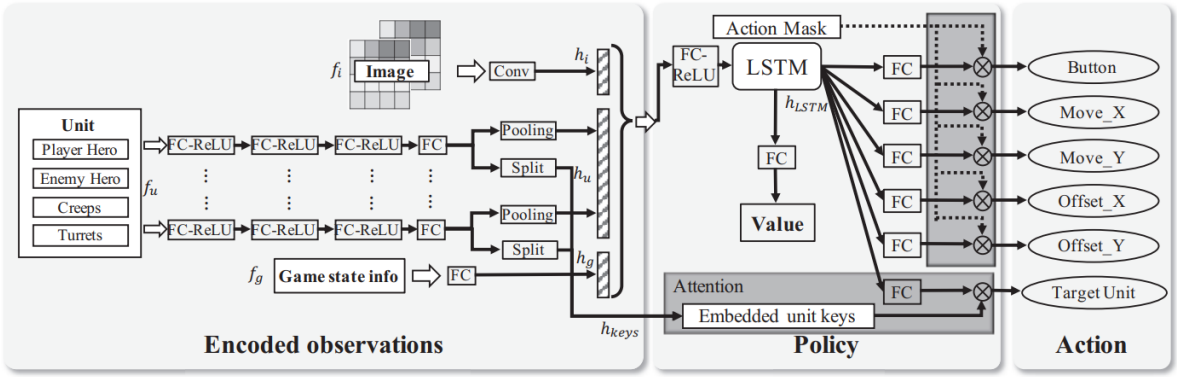
## 二、 参赛概况

在第三届开悟多智能体强化学习大赛中， 每支参赛队伍被分配了 32 台 CPU 服务器与 1 块 GPU 显卡的计算资源，需要采用强化学习的算法，在给定的资源下训练并提交一个模型。在初赛中，控制狄仁杰、公孙离、后羿、鲁班七号、马可波罗五位英雄进行墨家机关道 1v1 对战。初赛主要考查单智能体解决方案，模型结构设计，强化学习算法设计和训练方式探索。重点探索模型泛化性和通用性。

我们队伍 5 位同学有一位参加过去年的比赛，在本届主要作为顾问。其余同学均是首次接触强化学习和游戏 AI 设计，希望能够通过本次比赛入门强化学习，为以后进行游戏相关的工作打下基础。

## 三、网络设计

网络部分，首先以官方提供的默认网络模型为基础，在上面进行一些调整。初赛中我们对于网络修改较少，工作主要集中于奖励体系上。网络模型大致上如[1]：




网络首先是一个特征编码模块，对游戏内的所有信息进行特征提取。然后由 LSTM 模块接收特征产生行动预测。在此基础之上，前期我们尝试了别的网络作为特征编码的 backbone，以及不同的深度宽度。由于观察到实际效果不佳或无变化，遂暂时作罢。

## 四、奖励体系

在奖励部分，我们首先参照了[1]，做了如下设置：

| Reward | Weight | Type | Description |
|---|---|---|---|
| hp_point | 2.0 | dense | the health point of hero |
| tower_hp_point | 10.0 | sparse | the health point of turrets and base |
| money (gold) | 0.008 | dense | the gold gained |
| ep_rate | 0.8 | dense | the rate of mana |
| death | -1.0 | sparse | being killed |
| kill | -0.5 | sparse | kill an enemy hero |
| exp | 0.008 | dense | the experience gained |
| last_hit | 0.5 | sparse | last hitting to enemy units |

然后，通过观察录像，我们做了进一步调整。我们初赛的英雄全是射手，所以可以根据射手的特点进行一些微调，首先是蓝量（ep rate）：射手大多数时候都靠的是普攻，所以可以降低蓝量的 Reward。其次是塔（tower hp point）：射手英雄推塔比其他英雄容易很多，我们在录像里发现类似如下情况：推塔就赢了却还在补兵，补完兵就回去了，所以塔的 Reward 也可以进行一定程度的上调 。

后面是一些直觉上的改动，修改的数值相对较小（不一定能提升，但是应该不会负收益）。射手补兵更加容易，所以 last_hit 可以稍微提升，在 1V1 中击杀敌方英雄往往能更容易获得游戏的胜利，只要不是把击杀奖励调的特别高，导致它不推塔就一直堵着对面杀就行。所以可以考虑把 kill 提高，同样把死亡惩罚提高。每个射手基本都会出吸血刀，带狂暴，回血能力都很强，所以消耗血量带来的收益相对比较低，可以小幅度降低，作为提升击杀死亡 Reward 的平衡。

## 五、召唤师技能

射手的召唤师技能建议优先狂暴，比如公孙离鲁班这两个英雄一定要带狂暴，非常适合英雄 solo 对线。这个改了之后应该可以提升不少胜率。另外我们发现马可英雄本身不太适合狂暴，我们去了解了职业选手和主播 SOLO 时马可的召唤师技能，基本带的是眩晕，另外看回放我们眩晕马可还打赢了对面的狂暴马可，感觉马克眩晕就是最优解了。

## 六、系统工程架构

在强化学习训练过程中，我们也使用了一些训练强化学习常见的逐步调参手段，比如随着训练时间推进逐步熵损失函数的权重，进行学习率褪火。通过调整对局样本比重，提高我们弱势英雄的对局样本生成，进行针对性的训练。

## 七、 模型迭代过程

另外在比赛的前期，需要从头训练模型，在后期基本都是在已有的预训练模型基础上 resume 恢复训练进行迭代。因此我们在比赛前期稍微调高了样本生成/消耗的比例，使模型的学习更具多样性；在比赛后期，有了预训练模型，仅需要进行细微调整或者进一步迭代，我们稍微降低了样本生成/消耗的比例，保证模型稳定。

## 八、 总结与展望

在本次比赛中，在赛程前期，得益于我们从去年比赛中以及论文中汲取的经验，我们很快进入了较好的比赛状态，前期势头较好，每周的排行往往能够名列前茅。其中，奖励体系和召唤师技能的设置，实现简单，仅需要较好的游戏理解和对录像的查看总结，改动正确即能获得较大提升。但是在后期模型性能提升十分有限。由于对强化学习算法了解比较有限，未能对算法做出改进，所以很快达到了瓶颈，比较遗憾。

## 参考文献

[1] Mastering complex control in moba games with deep reinforcement learning