开悟比赛-五杀蔡文姬队技术整理与分享

陈华玉 清华大学计算机系
沈晓腾 清华大学自动化系
严谕梓 清华大学电子工程系
黄彬 清华大学计算机系
周浩天 清华大学自动化系
指导老师：阎栋 （朱军教授课题组）

一、 简介

在 2021 年 9 月-2022 年 4 月举办的腾讯开悟 MOBA 多智能体强化学习大赛中，我们队伍（五杀蔡文姬）有幸获得了第一名的成绩。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，而后将从神经网络架构，奖励函数设置，强化学习算法设计，系统工程架构四个方面简要介绍本队伍在开悟比赛中的探索历程与心得体会。

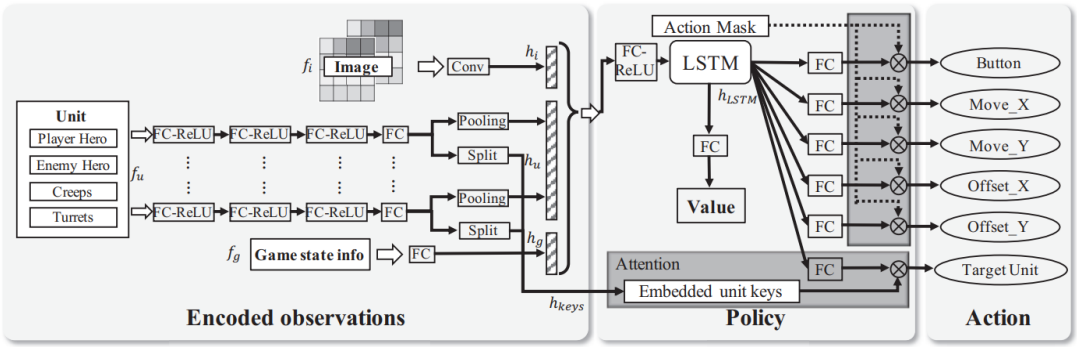
二、 比赛概况

在第二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（1024-2048 核）与 1-2 块 GPU 显卡的计算资源，并需要利用这些计算资源，采用强化学习的算法，在没有任何人类专家数据引导的前提下通过自博弈训练一支 3v3 的英雄队伍。比赛双方的英雄与资源完全对称，三个英雄分别为李云芳，赵云和貂蝉，每个英雄有维度完全相同的动作空间和状态空间。

由于队伍的计算资源有限，因此在比赛全程中机会没有任何机会对单一的更改进行对照试验，因此下文中所有技术的探索以及好坏的比较，都掺杂了极大主观性因素且在有强时间限制下得出的结论。需要辩证看待。

三、神经网络架构

在论文[1]中给出了 MOBA 游戏神经网络架构的推荐形式：



在本次开悟大赛中，我们队伍的网络模型无论是从宏观架构，总参数量上还是从微观超参数选择上，都与如上神经网络模型保持高度一致。我们在此模型基础之上出发，进行了五点探索，分别是增加交流机制，降低模型动作空间维度，尝试增加注意力网络结构，引入残差网络，增加模型参数。下面将会分别阐述

1. 增加交流机制

上述网络模型值考虑了一个英雄，在考虑三个英雄的情况下，我们希望三个英雄在网络模型层面不要互相鼓励，应该有一定结构使得他们可以交换信息。具体地，我们在 LSTM 结构输出了 512 维度信息（下文称“全局融合信息”）之后，分别将三个英雄的全局融合信息通过一层 MLP 下降到 128 维度（下文称“英雄交流信息”），并将三个英雄共享信息按照维度取最大值，糅合成一个 128 维度向量（下文称“英雄共享信息”）。英雄共享信息会被连接到各个英雄全局融合信息的尾部，形成一个 640 维度的向量，再进行后续操作。实验证实这种结果有助于训练速度的提升以及模型能力的增强。

2. 降低模型动作空间维度

这部分主要针对的是动作中 skillx 和 skillz 两个动作维度，这两个动作都是长度为 42 维度的离散数字序列。按照我们队伍队员对于王者荣耀游戏本身的理解，判断英雄不应当需要那么精细的技能操作控制，因此相关动作标签的输出仅有 14 维度，并通过差值算法将 14 维度扩展为 42 维度，以降低探索空间。

3. 增加注意力网络结构

假设全图有 20 个野怪，每个野怪经过神经网络编码后得到了一个 32 维度的向量数据，[1]中的处理方法是将 32*20 维度的向量数据经过 maxpooling 降维成为 32*1 维度的数据。我们认为这种 maxpooling 主观上不见得有效，因此尝试将原特征处理处理部分替换为 transformer 架构，用注意力机制的方式实现数据降维。类比野怪，同时在全局所有使用 maxpooling 的地方都使用注意力机制替代。后续经试验观察，这种试验结构短期会拖慢训练速度（0.5 天），中期对于英雄能力提升不大（2 天），因此放弃此尝试。

4. 尝试残差网络

我们也尝试了使用残差网络结构，在 LSTM 架构的前端和后端进行残差连接，但并未观察到明显效果提升。

5. 在比赛前期，我们在原有神经网络结构的基础上，把核心部分的网络宽度提升了一倍，因观察到训练速度大大降低而放弃。

三、奖励函数设置

王者荣耀有着巨大的观察空间（observation space）与动作空间（action space），这使得智能体随机探索得到正反馈十分困难。对此我们设计了课程学习的方法（curriculum learning）,针对智能体不同阶段的首要学习目标设计了不同奖励机制，逐步迭代智能体的作战风格。本队伍在整个训练过程中的奖励函数调整本着先着重调高稠密奖励项（dense reward），再着重调高离散奖励项（sparse reward）的原则。

具体地，在训练早期，我们在[1]论文中奖励函数权重所声明的数量级基础上，首先讲 money 的权重设置成为 0.003，随后又逐步调高到 0.005，其间也一定程度上调高了血量奖励等稠密奖励项。在训练进入中后期之后，我们将会把这些引导性的稠密奖励逐步下降，并使用击杀等离散奖励进行代替。

根据我们对于对局的分析，前期多采用稠密奖励有利于智能体快速学习和探索，而后期提高击杀奖励则有助于提高 KDA，引导英雄多抓人，多追击。

对于团队合作因子，我们在比赛全程设置为 0.2，未做大量的实验与更改。

注意到智能体在比赛后期获得相同奖励的时间要比前期少很多，我们也将所有奖励统一乘以 0.6 的指数衰减因子，但对此我们并未进行任何详尽的对比试验，无法保证超参最优。

最后为了更好地引导智能体产生良好的特征，我们修改了对战框架，类似于[2]复现了其多头价值网络输出的训练方法。这并不会使得训练速度加快，但有利于提高智能体水平上限，避免训练过早收敛。

四、 强化学习算法设计

王者荣耀 3v3 比赛的一个特点是团队内队友配合与团队间队伍对抗结合。传统上队伍内部配合是依靠各个英雄独立训练，引入团队因子（team spirit）参数实现的，队伍间对抗依靠零和博弈机制实现。但是现实场景下，设置团队因子过高或者零和博弈会让学习过程变得十分不稳定。对这种问题，我们对于比赛常用的 PPO 算法进行了优化，增强了其数值稳定性，使得高团队因子的设置成为可能，从而提高了智能体的团队合作能力。对于这部分算法的修改细节，可以参考实验室未来可能公开的相关论文。

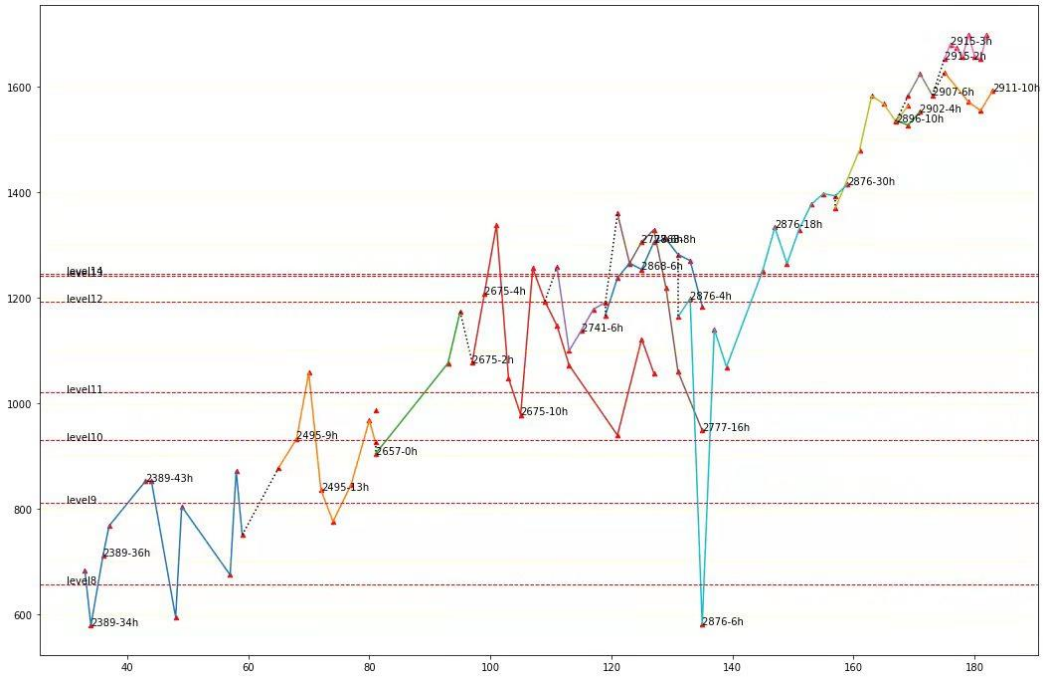
五、 系统工程架构

在第二届开悟比赛中，队员们需要手工将对局产生的英雄模型拉取到开发机中再上传至托管平台。且需要手工选择模型进行对战，人工判断模型的好坏与克制关系，并利于这些知识开展下一阶段的实验。这其间需要消耗大量的人力资源，降低了实验的周转效率。对此我们实现了多风格智能体水平评估与托管对战的自动化部署，较大程度上节省了人力，增加了实验周转效率。具体来说，我们可以全自动化地将正在进行的实验模型定期拉到开发机并上传托管平台。相关爬虫程序可以自动从托管平台选择模型进行对战并把对战数据保存至本地。在本地我们也部署了相关算法可以实时计算每个模型的 ELO 分数，并将分数排序反馈给爬虫程序作为其选择对战模型的参考。

在这样一套托管系统的基础上，我们引入了集成学习的思想，把 ELO 分数相近但是风格不同且可能相互克制的模型，通过模型集成的方法整合在一起，上传到托管系统进行对战。并将集成模型和原模型混合排序，以方便进行下一个阶段的模型集成工作，循环往复。

根据观察，高频率的模型集成能小程度提高模型能力，可以大程度提高模型鲁棒性，避免其被单一模型克制。

最后，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。又比如在观察到数据离线程度（offpolicy）较高的时候同步调整 GPU 速率与数据缓存器大小以减少数据复用率，和使用延迟。再比如逐步降低或归零附加损失函数的权重以减少 PPO 算法损失函数的偏差等等。由于每次网络参数的调整都是通过停止实验，更改参数再重启实验完成的，我们也制作了专门的可视化工具，结合自动的 ELO 评估系统对于实验增长情况进行了可视化，这样可以帮助我们更好地分析实验结果，降低人力成本，提高实验周转效率。



相关文献:

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

[2] Towards Playing Full MOBA Games with Deep Reinforcement Learning