开悟比赛-清水河畔混分王队技术整理与分享

曾达 电子科技大学计算机科学与工程学院

滕达 电子科技大学计算机科学与工程学院

李政 电子科技大学计算机科学与工程学院

陈鑫凯 电子科技大学信息与通信工程学院

李嘉铭 电子科技大学计算机科学与工程学院

指导老师:谢宁

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(清水河畔混分王)在初赛中有幸获得了第 11 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

请在本节中描述本队参加比赛的概况，包括成员信息、参赛动机、时间投入等

本队成员来自电子科技大学计算机科学与工程学院以及信息与通信工程学院，由 2 名研三的同学（曾达，腾达）和 3 名研二的同学（李政，陈鑫凯，李嘉铭）组成。

参赛动机是想通过实践来更深入地学习和理解前沿的强化学习技术，了解和体验强化学习技术的现实应用，同时积累科研以及工程经验，更好地提升自我。

时间投入上，我们队每周会定期进行至少 2 小时的技术交流会，每个队员在完成自身学业和科研任务之余，每人每天至少投入 2-3 小时用于开悟比赛的研究。

三、网络设计

请在本节中描述神经网络架构的设计思路以及具体架构

网络结构由特征编码模块和前向推理模块组成。我们尝试过的探索及改进效果如下：

特征网络，在处理英雄特征的时候，根据官方给出的关于特征的信息，我们区分开了英雄的公共特征和私有特征，英雄私有特征是与英雄技能相关的稀疏特征，133 维的私有特征中仅有几个数值是与当前英雄有关的，其他都是 0，因此我们用一层 FC 对私有特征进行降维，再将降维后的私有特征与公共特征拼接，组成了新的英雄特征。

Target_embedding，target_embedding 作为目标选择时被查询的列表，存储着对应目标降维后的特征。a. 我们尝试过将 split 改为 FC 得到目标特征；b. 将对应 None 目标的常数特征

改为对应我方小兵的目标特征；c. 调整我方英雄和敌方英雄目标特征的前后顺序，使之与 target 输出顺序相对应。以上改动并没有带来明显的提升，并且改动 b 似乎使得英雄过于跟随我方小兵。

分类网络，我们通过解析不同英雄的私有特征，手工构造相应的 weight&bias，得到了能以 onehot 形式表示不同英雄的 hero_index 特征。使用 hero_index 特征，我们分别尝试了 2 分类、3 分类、5 分类这几种结构，以及软分类（通过模型训练学习分类权重）和硬分类（使用不可训练的常数作为分类权重），但均未达到理想的效果。理论上说，良好的分类网络可以帮助模型在后期训练时分别学习各个英雄的策略，而不会因为不同英雄策略不同产生震荡。但是我们使用分类网络的模型始终没有达到较高的水平，推测是分类导致各网络的训练样本太少，模型收敛过慢。

Shallow 网络，用于保存各个单位的浅层特征。

target 网络，我们使用了 attention 模块，并且将 target 网络的输出拼接到后续网络的输入上。

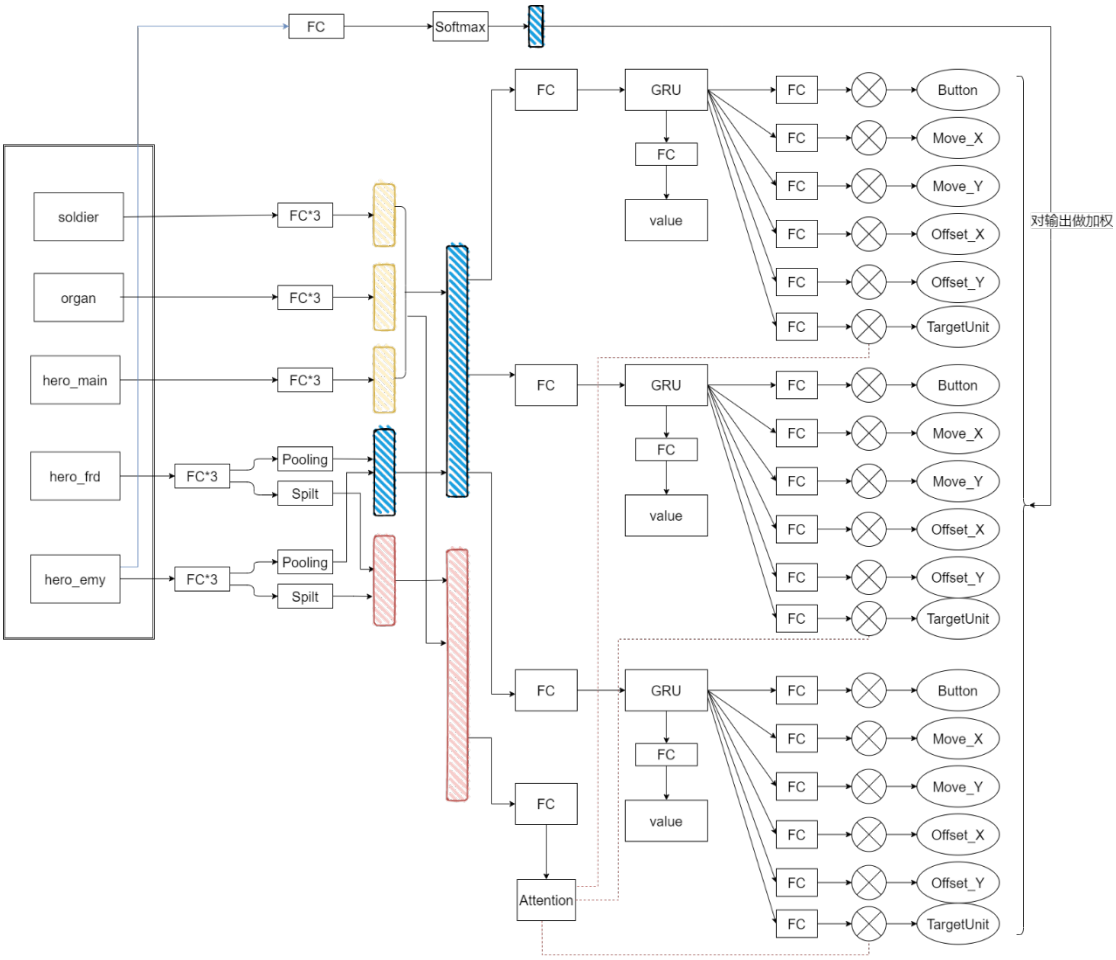
action 网络，可以使用特别小的负常数项来作为 none0，recall，skill4 和 equipment_skill 等动作的输出，因为这些是在 1v1 中绝对用不上的 action，在网络上 mask 掉可以减小网络尺寸。

value 网络，可以由直接输出 value 改为训练两个网络分别学习 total_value 和 enemy_value，使用 total_value-enemy_value 来作为自己的 value 输出。

网络尺寸，网络并非越大越好，越深越好，较小的网络结构已经足以完成对战任务，过大的网络结构反而难以收敛。

整体网络结构，最后的三分类网络结构大致如下图，其中，图上方是对英雄进行分类，这部分我们是直接通过推测私有特征获取的准确的英雄类别，分类后得到一个概率 tensor，类似 [1,0,0]，之后会和每个子网络进行加权和，以此给英雄网络分类。

图中间部分蓝色部分是经过 FC 层后 contact 之后的特征，上面的特征过 GRU 后，用于做动作选择；红色部分是作为 embedding 的特征，经过 attention 后，用于做目标选择：



四、奖励体系

请在本节中描述奖励相关的设计思路以及具体改动

奖励退火：奖励退火是非常有必要的，游戏时长越长，后面获得的奖励越低，否则智能体可能会为了获得更多奖励而恶意延长游戏。我们目前采用的是随游戏时长线性退火，如按照往届或论文中的非线性退火可能效果会更好。

英雄生命值：原生奖励中，在视野外的敌方英雄血量会认为是 1，在视野内的敌方英雄血量则是真实值，当敌方英雄进入视野或离开视野时，就会因为敌方英雄血量的变化给一个很大的惩罚或奖励，这是不合理的。工蜂论坛上同样有人反映了这个问题。于是我们根据前后决策帧敌方英雄的血量和状态重新计算了敌方英雄生命值的变化率，并且拆分了我方英雄血量和敌方英雄血量的奖励。

生命值奖励曲线：根据游戏经验，角色血量比率越高，越不需要看重自身血量；角色血量比率越低，越应该关注自身血量。因此，我们设计了一个血量越低奖励系数越高的 N 次方曲线，并将该形式应用在我方英雄、我方防御塔、敌方英雄、敌方防御塔的血量奖励上。

离散奖励：我们根据获得的帧状态设计了若干离散奖励，离散奖励在训练前期可以帮助智能体收敛，但是在中后期可能会限制智能体的探索，应该适时取消离散奖励。

零和奖励：在我们的奖励设计中没法严格遵循零和奖励，但是还是应该尽量按照零和进行设计，例如设计我方英雄闲置有额外的惩罚，就应该设计敌方英雄闲置时我方能够得到额外的奖励，以避免智能体找到一种双赢的策略。

奖励降低：将奖励系数总体设计得比较低，影响到的训练效果会更加精确，有利于降低value_loss，也有利于模型训练。

奖励验证：设计任何新奖励之前，应该通过本地的对战 log 进行逐帧验证，确保没有奇怪的大量奖励或惩罚出现，且胜负方总奖励接近零和，且游戏时长越长总奖励的绝对值越小。

五、特征与规则

请在本节中描述对原始特征的主要改动，以及基于规则实现的前后处理等操作

自定义 feature：通过对英雄公共特征的解析，我们可以对英雄攻击、暴击率、暴击效果等特征项进行处理得到期望暴击伤害等难以直接学到的交叉特征，然后把这些特征嵌入原本的特征当中。该特征预期能够提升模型对暴击状况的判断。

基于帧状态，我们编写了若干基于规则的策略，并且以一定概率在对战中启用这些策略。理论上，基于规则的 AI 可以帮助智能体学会一些实用的连招、mask 一些无用的操作，但是实践上，rule_AI 很难编写出时序场景下的复杂策略，且效果尚不明显。

六、 强化学习算法

请在本节中描述关于强化学习算法的优化思路及实现情况

算法仍采用开悟原有的算法，在算法流程上基本没有变动，且 loss 计算上没有做更多改进。

七、 系统工程架构

请在本节中描述系统或工程层面的的调优、改进及二次开发情况

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，但是当训练模型达到 baseline_level 0 以上，对战 common_AI 的胜率及各项数据就没有参考意义了，都是完全碾压的胜局。因此，后期我们将 common_AI 替换成当前训练网络下的阶段性模型，就可以通过胜率曲线看出模型是否提升。

并且，我们可以选取当前训练网络下的阶段性模型作为训练对战的对手，用于生成对战样本。我们可以通过不同的 reward 设计得到同一网络不同对战风格的模型，然后训练中以一定概率选择这些模型进行对战，从而避免当前训练的模型陷入对自己战斗风格的过拟合。

八、 模型迭代过程

请在本节中描述模型逐步迭代的过程及课程学习相关设计（如有）

模型的迭代主要通过调节超参来控制，每次确定好网络的改进后，我们会进行一轮的训练，主要调节的超参数如下：

learning_rate：学习率，前期采用 1e-4，后期从 5e-5 逐渐降到 1e-5。前期学习率应该还能更大一些，训练后期模型停止进步，表现不断震荡的时候，降低学习率和 beta 可以进一步提升模型表现。

beta：entropy_loss 相关系数，前期采用 0.015，后期采用 0.01 逐渐降到 0.0005。由于 entropy_loss 反映了当前策略与历史策略的相似程度，随着 beta 的降低，entropy_loss 的绝对值也会减小，即当前策略更接近历史策略，模型的策略在进一步收敛。反则是鼓励模型进行与历史策略不同的探索。我们最终模型的 entropy_loss 大约从-10 左右到最后降到了-7 左右，胜率提升也很明显。

batch_size：每批训练数据尺寸，默认值为 4096，根据上届经验，采用 512 或 1024 即可。理论上，batch_size 越大，更新越平均越稳定，反之更新越随机。较小的 batch_size 有利于快速收敛和突破局部最优，训练后期可以考虑增大。

gamma：折扣系数，默认为 0.995，折扣系数越大则考虑越多步长的奖励，根据上届经验，模型停止进步后可以考虑提高 gamma 和 lambda，但是会对 value_loss 造成负面影响。

九、 训练效果分析

请在本节中评估关键 feature 开发，及模型训练迭代过程的开销和收益，并分析最终提交模型的能力水平和技术特点

整体训练的周期分成了前期的探索、中期的优化与最后的冲刺。前期时间花费最多，大概占总赛程的一半还多，中期的优化大概占总赛程的 1/4，最后冲刺为最后两周。

前期的探索较为分散，花的训练时间也是最多的。训练的探索包括网络结构探索、特征处理探索、系统工程的优化等等，以上已经提到，在此不再赘述。中期的优化主要是挑选了几个表现较为稳定且优秀的网络，集中进行整合并开始调参训练。比赛最后两周我们选取了当前表现较好的模型进行调参，经过调参模型实力得到了明显提升。

总体来说，前期训练的探索投入多，收益低，中期优化的收益较为稳定，后期冲刺收益最高。

最终提交模型在控制不同英雄时有着不同的作战倾向，在控制鲁班、狄仁杰、后羿等站桩英雄时，模型的进攻性更强；在控制公孙离、马可波罗等机动英雄时，模型更加保守。值得高兴的是，模型学会了马可波罗的眩晕加大招的连招，在使用马可波罗时取得了一定的胜率。最终提交模型大约达到了开悟提供的 baseline level 3 的水平。

十、 总结与展望

请在本节对上述工作进行总结并简单介绍后续可以提升的思路, 比如描述由于时间或其他原因导致的未能实现的功能以及预期效果评估等

本次初赛，我们队伍在整体的节奏上还是有点拖沓，赛后也进行了反思，主要分析总结了以下的经验与教训。

应当尽早确定训练新模型时使用的 RL 超参，而不是使用默认的 RL 超参；往届对 RL 超参和 loss 有着较深的研究，而我们到比赛后期才对 RL 超参和 loss 开始重视。如果能尽早探究清楚 RL 超参对模型收敛速度的影响，可以加快我们测试新网络结构的过程。

应当明确衡量网络结构好坏的标准，不要在表现明显不好的网络上浪费训练资源；时间和算力在比赛中是非常宝贵的，我们没有明确对于一个需要测试的新网络来说，应当取它训练多少时长的模型表现来进行评判，训练 12h、24h、36h 还是 60h？对于一个新网络，它初步收敛时的表现已经足以看出它是否优于其他网络，但是我们对于前期表现不好的模型还是心存侥幸，投入了更多的资源进行训练，事实证明它训练更多时长表现也是不好的，拖慢了我们探究网络结构的进度。

前期一直在探索网络，尤其是分类网络，由于训练中发现三分类网络表现更好，我们的最终模型还是是一个三分类网络，并不是五分类网络，且这部分由于时间问题也没有再进行探索。也许分类子网络的数目存在一定的最优解，这部分后续可以继续研究。

参考文献

[1] Deheng Ye et al. “Mastering Complex Control in MOBA Games with Deep Reinforcement Learning” National Conference on Artificial Intelligence (2020).C

[2] Schulman J， Wolski F， Dhariwal P , et al. Proximal Policy Optimization Algorithms[J]. 2017.

[3] Yu C, Velu A, Vinitsky E. The surprising effectiveness of MAPPO in cooperative[J]. Multi Agent Games, 2021

[4] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Advances in Neural Information Processing Systems. 2017: 5998-6008.

[5] Marcin Andrychowicz et al. “What Matters for On-Policy Deep Actor-Critic Methods? A Large-Scale Study” International Conference on Learning Representations (2021).