开悟比赛-节奏战队技术整理与分享

高华泽 西南交通大学计算机系

胡虎 西南交通大学计算机系

杨玲 西南交通大学计算机系

叶小康 西南交通大学计算机系

冯靖婷 西南交通大学计算机系

指导老师:邢焕来

一、 简介

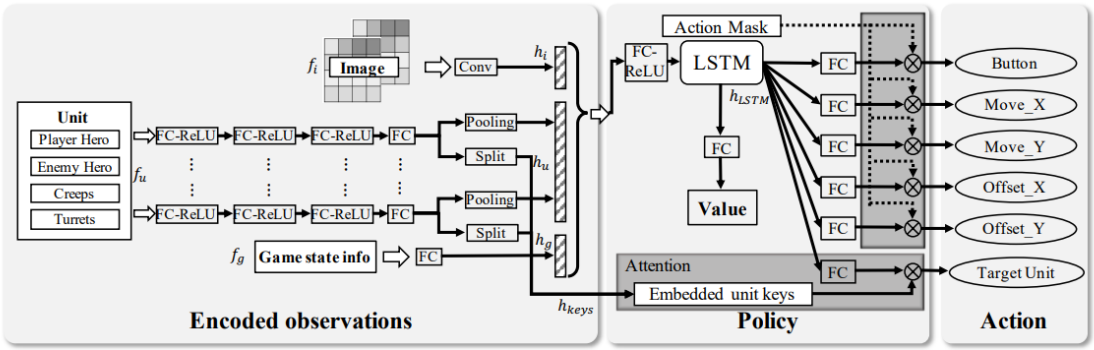
在 2022 年 9 月-2023 年 4 月举办的腾讯第三届开悟 MOBA 多智能体强化学习大赛中，我们队伍在初赛中有幸获得了第 12 名的成绩。这得益于实验室老师与腾讯官方的大力支持， 也得益于我们队伍本身的技术探索与积累。本文首先会简单叙述比赛的基本情况，随后从各关键模块出发，简要介绍本队伍在开悟比赛中的探索历程与心得体会。

二、 参赛概况

本次比赛是通过邢老师与李老师介绍加入的，在参加这次比赛之前我们已经听说过开悟比赛，并且观看了上一届的决赛，因此，对于比赛有着浓厚的兴趣，这次参赛成员由两位老师的学生组成，在比赛期间，我们为每位同学合理分配工作并定期开会讨论。

三、网络设计

在本次开悟大赛中，我们的神经网络模型整体框架都与论文 Mastering Complex Control in MOBA Games with Deep Reinforcement Learning 中的神经网络模型保持一致。在此基础上做了以下的一些研究和探索。



1、增加网络规模

我们在原有的网络规模基础上将 lstm 前后的连接单元从 512 维增加到了 768 维，因为王者荣耀是一个十分复杂的环境，我们想要通过更多的神经网络单元来更好的拟合王者荣耀环境，通过更大规模的 lstm 层让 agent 做出更好的动作以及学习更好的技能连招。通多我们做的对比实验，增加网络规模确实能够让我们的模型得到提升。

2、尝试残差连接

在特征提取层中，模型大多都是采用多层 mlp 进行特征提取的，我们考虑到过度深层的 mlp 不一定能够很好的提取特征信息，因为我们尝试在特征提取层中增加残差网络连接，但是对比实验后并未发现有更好的效果，所以放弃该策略。

3、尝试建立分支策略网络

由于初赛的目的是同一个模型要能够耦合 5 个不同射手因英雄，因此我们考虑在策略层之后根据英雄 id 来进行不同的策略网络分支，具体的是在 lstm 层之后使用分支网络，使得不同的英雄有不同得策略输出。但是改了之后模型超出了规定大小，也因为模型过大导致训练速度十分慢，并没有很好得效果，故放弃该策略。

4、尝试引入噪声网络

在复盘对战视频时我们发现英雄会偶尔出现原地来回走动的情况，于是我们也鼓励智能体加强探索，并在网络中加入高斯噪声，由于是在训练中期时加入，破坏了原有的网络结构，可能导致参数出现过大的偏差，在实际应用中没有起到很好的效果。

四、奖励体系

在奖励函数设计方面由于 baseline 给的奖励项已经很多了"reward_money" ,"reward_exp"，"reward_hp_point"，"reward_ep_rate"，"reward_kill"， "reward_dead","reward_tower_hp_point"，"reward_last_hit"。我们也主要是用这些奖励来训练我们得模型，具体的参数大小也是根据论文以及官方给的 baseline 进行调整的，并在训练过程中对各个奖励项进行动态调整。

五、 强化学习算法

在算法方面我们队伍也是基于 PPO 算法，使用了 Mastering Complex Control in MOBA Games with Deep Reinforcement Learning 文章中提到的 dual clip PPO 方法，由于我们是第一次参赛所以在算法方面基本准寻了论文中的方法，因此算法方面我们队伍基本没怎么去探究。

六、 模型迭代过程

我们将模型分阶段进行训练，按照课程学习的思想，在前期，我们希望英雄能够先快速学会如何让自己取得更良好的发育，我们将 reward_money 和 reward_exp 设置为 0.009，reward_hp_point 设置为 1.0。其他奖励项暂时都设置为 0，因为其他非连续奖励项在模型前期不能很好的给出奖励信号，因此会大大拖长模型训练时间，经过我们实验也证明，前期更少的奖励项确实能够提高初期模型训练效果。

训练中期，我们发现我们的模型能够让英雄在对线时期正常的发育打钱，但是有些时候会

过分的为了追求发育而使得自己阵亡，这个时期，我们就会将 money 和 exp 奖励项调低，增加 reward_hp_point，以及给 reward_dead 奖励项一个负值。在经过一定时期的训练后，我们的模型就不会再出现送人头的现象。

训练后期，我们会逐步通过降低连续奖励项的系数，通过增加离散奖励项系数的大小，让我们的模型学会补尾刀，击杀敌方英雄。

当然我们在训练过程中也使用了一些常见的训练模型技巧，例如学习率退火，增加折扣回报因子的大小（王者荣耀是一个推塔的游戏，更加注重长期回报，有时候越塔杀人不一定对游戏的胜利有帮助）。前期 GAMMA 较小后期调大，entropy 的系数前期较大后期调小，学习率前期大后期小，PPO clip 的参数前期大后期调小。

七、 训练效果分析

我们所训练模型较初始模型扩大了网络结构，因此增加了一定的训练开销，但通过 reward 的动态调整，一定程度上加速了训练的时间，整个训练过程过后，我们的 ai 表现较为均衡，不会很激进，也不会很畏缩，技能释放也较为合理，但还存在一定的问题，例如会出现特定英雄表现很好，而其余英雄表现则一般的情况。

八、 总结与展望

本次初赛我们主要注重在网络结构以及整个训练过程中，期间会根据视频回放中 ai 的具体表现对代码以及参数进行适当调整，例如，在加强英雄探索方面，我们先后尝试了加入 Dropout 层以及引入噪声网络等方式；在模型泛化的优化中，我们先后调研了知识蒸馏，多任务学习等方案，但在应用过程中，需要学习多个网络，考虑到训练开销大等因素，没能得到实施。本次比赛没能尝试的是在特征提取环节进行适当优化，以及动态奖励权重的设计等，不过在本次比赛中我们也收获了很多经验。

参考文献

[1] Ye D, Liu Z, Sun M, et al. Mastering complex control in moba games with deep reinforcement learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(04): 6672-6679.

[2] Ye D, Chen G, Zhang W, et al. Towards playing full moba games with deep reinforcement learning[J]. arXiv preprint arXiv:2011.12692, 2020.

[3] Rusu A A, Colmenarejo S G, Gulcehre C, et al. Policy distillation[J]. arXiv preprint arXiv:1511.06295, 2015.

[4] C. Berner, G. Brockman, B. Chan, V. Cheung, P. D˛ebiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, et al. Dota 2 with large scale deep reinforcement learning. arXiv preprint arXiv:1912.06680, 2019.

[5] W. M. Czarnecki, R. Pascanu, S. Osindero, S. Jayakumar, G. Swirszcz, and M. Jaderberg. Distilling policy distillation. In The 22nd International Conference on Artificial Intelligence and Statistics, pages 1331–1340, 2019.