第二届开悟大赛清水河畔混分王队技术分享

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 中提到的 dual clip PPO 方法，并且在 value 网络计算时也用了 clip 技巧，算法方面我们队伍探究的不多。

我们队伍对网络模型改动很多。其中我们网络模型中策略神经网络和价值神经网络的输入并不是由默认的 LSTM 构成，我们用了 GRU，原因在于相比 LSTM 更节省参数。虽然在论文中 OpenAIFive 和腾讯绝悟都只用了 LSTM，在少量计算资源下可能效果并不好。我们做了实验发现将 GRU 和 MLP 拼接在一起作为策略神经网络和价值神经网络的输入效果会更好，拼接的比例我们是 GRU 占 1/4，MLP 占 3/4。

为了实现多智能体之间的通信，我们网络中创新性的设计了一个基于注意力机制的通信模块，方法类似于 Actor-Attention-Critic for Multi-Agent Reinforcement Learning 中的办法，我们将其改进到了开悟网络中，并且做了 OpenAI Five 提到的 max-pool fuse 通信的对比实验，我们的效果会比该方法和不考虑通信效果更好，这一部分我们放在了待投的小论文中，细节屏蔽一下。

考虑到过大和过深的神经网络并不适合强化学习，而王者荣耀游戏状态空间巨大特征复杂，又需要深的深度神经网络去拟合状态信息特征，我们提出了幂级连接网络来解决该问题，并取得了不错的效果，这一部分我们也放在了待投的小论文中，细节屏蔽一下。

在奖励函数方面，我们设计了额外的场景奖励，其中重要的动机就是想让三个英雄一起打暴君，因为三条暴君对于胜利的加成太大，而仅仅通过常规的自我博弈训练很难学到三英雄一起打暴君的场景。我们的具体做法如下：在 actor.log 中我们可以拿到对局中每一帧英雄的小项 reward 信息，比如 money，kill，exp 等等。我们通过找规律发现击杀暴君时会出现三个英雄同时获得固定值大小的 money exp 奖励项，并且击杀暴君的英雄会有 atk_monster 的奖励，于是我们在 reward manager 中就去判断当这一帧英雄击杀暴君时，我们就在该样本的 final_reward 上加额外设计的奖励，这样渐渐我们训练的 AI 会越来越关注暴君的争夺。大家可以按照类似的思路去设计更多的额外场景奖励。

我们还用了 OpenAI Five 中提到的 dynamic sampling system 用于替代目前过去的随机的历史模型池，这一块代码实现比较复杂，当时动机是防止模型陷入局部最优，不忘记过去的策略，越来越强，但实际效果的话并没有特别好，比默认的方法提高不多。

关于训练的经验。Reward 调参的大体方向在前期我们会把调高 money exp 的 reward，训练中期我们会把 money exp reward 降下来，提高 hurt_to_hero 的 reward，后期我们会把 money exp reward 降下来，提高 Kill dead assist 的 reward。而实际上我们队伍的奖励项在整个训练过程中是需要不断调整的，我们观察录像发现英雄行为达不到预期就会调整。还有一些重要的参数，我们前期 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。