西安交通大学王者荣耀比赛经验分享

网络结构：

多智能体强化学习的一个问题是；The question of whether independent agents are better at learning cooperative behaviour than multiple complete observing agents ,那经过简单的对比实验和已发表论文结果分析后，我们将 n 个智能体多多智能体强化学习分解成 n 个单独的单智能体问题，训练过程中，不同智能体共享模型权重。

此外在模型设计部分，我们还尝试了 DenseNet，但是参数量增加，效果也并不明显。尝试了对比学习，训练拖慢，效果也不明显。大火的 Gated TransformerGTrXL：试错成本太高，训练过慢，因此在训练几天后提升不大就放弃了没有做进一步的实验。

交通与通信：

这个方法的一个缺点也很明显，它的学习策略是仅仅基于每个智能体当前的观察，智能体之间缺乏通信交流，面对这样一个需要团队协作的任务，显然是有些不合理的。因此我们借鉴 openai five 加入 cross-hero pool 的模块。也就是在特征提取部分后做了一次所有英雄的特征融合。max-pool 综合了所有英雄特征的前 1/4 个分量，和剩余的分量 concat 到一起，来代表游戏中各自英雄关于游戏的全部认知。

探索：

面向这么一个有着高维观察空间和动作空间的，游戏规则十分复杂的环境，智能体在这个广阔的空间进行充分有效的探索并不容易。因此为了鼓励探索我们尝试了 noisynet、RND 等探索方法，但是在 reward 设置没差别的情况下。在比赛场景中，效果与 noisynet 比差一点，不过 rnd，icm 收敛速度会比 noisynet。

reward 设计：

总体来说，我们的奖励函数设计分为 4 个阶段，按照课程学习的思想，在前期，我们希望英雄能够先学习打野对线，将赵云的打野参数提高，为了让李元芳学会对线，不让他打野让他更快清兵线，将它的打野调到 0.01。第二个阶段是让英雄学会合作，这部分主要是通过 team spirit 来体现，在整个训练过程中，我们会逐步提升 teamsprit 的比重。在训练初期小一些的 team spirit 有利于各个英雄独立的快速学习，之后逐步增加的 ts 使得他们慢慢学会合作，朝着共同的目标努力。我们训练的第三个阶段是提升英雄的攻击欲望，与别的队伍对比发现，有可能是奖励函数设置的原因，我们的英雄攻击欲望很小，很习惯拉长战线，因此我们提升 kill 和 hp 的权重，之后 kda 上升，但是经济不行。因此第四个阶段，我们提升金钱和经验的比重到 0.008。

有个小 tip 是：每个英雄 reward 差异大有可能会学崩，因此我们同质化 reward 只有打野和塔对于每个英雄设置不一样，其他都接近。