Introduction of a Popular Multi-Agent Reinforcement Learning Environment —StarCraft II

--孟令辉

星际争霸是由暴雪游戏开发的实时策略对抗游戏（real-time strategy, RTS）。在其中人类玩家相互对战或与规则 AI 对战以收集资源、建造建筑来构建武器打败他们的对手。星际争霸包含两种 RTS 的对战模式：macromanagement 和 micromanagement

- Macromanagement (macro) 指的是高层级的策略对弈如经济、资源管理
- Micromanagement (micro) 指的是独立单元的细粒度的控制 （学术上基本在研究 Micromanagement 下的问题）

(c) SMAC corridor

(d) SMAC 2c_vs_64zg

Micromanagement（微操）

星际争霸已被用作 AI 的研究平台，最近也被用作强化学习 RL。通常游戏被视为竞争问题：一个智能体扮演人类玩家的角色做出宏观上的决策和执行微观层面的动作。为适应多智能体强化学习框架，StarCraft 将每个单元交给一个独立的智能体分布式地控制，而该智能体只依赖于其受限于一定区域的局部可观的观测。整个由这些智能体组成的团体需要被训练以解决具有挑战的对战场景下的问题，目的是击败由系统规则 AI 控制的对手。

在战斗中，适当的单位数量会使对敌方单位的伤害最大化，同时将受到的伤害降至最低，并且需要一系列技能。 例如，一种重要的技术是集中射击，即命令部队一次又一次地联合攻击并杀死敌方部队。 集中火力时，避免 overkill 很重要：对单位造成的伤害超过杀死单位所需的伤害。其他常见的 Micromanagement 技术包括：根据装甲类型将单位组装成编队；在保持足够的距离的情况下使敌方单位追逐，以至于几乎不造成伤害（风筝敌人）；协调单位的位置以从不同方向进攻或利用 击败敌人的阵型。

场景（Scenarios）：

这里介绍微操场景（Micromanagement Scenarios），均为了评估每个独立的智能体可以很好地协作以解决复杂的任务。这些场景被仔细设计且划分好等级不同的难度，每个场景均为两方多个单元的对弈。不同的场景可操作单元具有不同的初始位置、数量、类型等。一部分势力是由可学习的我方智能体控制，另一部分势力则由规则 AI 控制（使用精心设计的非学习的启发式方法）。在每一场对局开始时，规则 AI 会使用预先设定的策略引导其控制的单元攻击我方智能体。当两方势力任意一方的单体个数为 0 或游戏时间超过预设的时长则游戏结束。每个场景的最终目标均为最大化可学习的我方智能体能够达到的胜率（对局中获胜对局数的占比）。此外为了加速，规则 AI 会在对局开始时攻击我方的出生点。

这些场景中包括同构 *homogeneous* 的或者异构 *heterogeneous* 的场景，其中同构的表示为每个单元均属于同意类型（如 Marines），这类设置下的获胜策略只要集中在开火、保证己方单元存活；异构则代表我方势力中包括不少于一种类型的单元（如 Stalkers 和 Zealots），在这种设置下我方智能体必须降低角色间互相冲突的性质以保护队友不受攻击。下面时一些场景的概览

Name	**Ally Units**	**Enemy Units**	**Type**
3m	3 Marines	3 Marines	homogeneous & symmetric
8m	8 Marines	8 Marines	homogeneous & symmetric
25m	25 Marines	25 Marines	homogeneous & symmetric
2s3z	2 Stalkers & 3 Zealots	2 Stalkers & 3 Zealots	heterogeneous & symmetric
3s5z	3 Stalkers & 5 Zealots	3 Stalkers & 5 Zealots	heterogeneous & symmetric
MMM	1 Medivac, 2 Marauders & 7 Marines	1 Medivac, 2 Marauders & 7 Marines	heterogeneous & symmetric
5m_vs_6m	5 Marines	6 Marines	homogeneous & asymmetric
8m_vs_9m	8 Marines	9 Marines	homogeneous &

Name	Ally Units	Enemy Units	Type
			asymmetric
10m_vs_11m	10 Marines	11 Marines	homogeneous & asymmetric
27m_vs_30m	27 Marines	30 Marines	homogeneous & asymmetric
3s5z_vs_3s6z	3 Stalkers & 5 Zealots	3 Stalkers & 6 Zealots	heterogeneous & asymmetric
MMM2	1 Medivac, 2 Marauders & 7 Marines	1 Medivac, 3 Marauders & 8 Marines	heterogeneous & asymmetric
2m_vs_1z	2 Marines	1 Zealot	micro-trick: alternating fire
2s_vs_1sc	2 Stalkers	1 Spine Crawler	micro-trick: alternating fire
3s_vs_3z	3 Stalkers	3 Zealots	micro-trick: kiting
3s_vs_4z	3 Stalkers	4 Zealots	micro-trick: kiting
3s_vs_5z	3 Stalkers	5 Zealots	micro-trick: kiting
6h_vs_8z	6 Hydralisks	8 Zealots	micro-trick: focus fire
corridor	6 Zealots	24 Zerglings	micro-trick: wall off
bane_vs_bane	20 Zerglings & 4 Banelings	20 Zerglings & 4 Banelings	micro-trick: positioning
so_many_banelings	7 Zealots	32 Banelings	micro-trick:

Name	Ally Units	Enemy Units	Type
			positioning
2c_vs_64zg	2 Colossi	64 Zerglings	micro-trick: positioning
1c3s5z	1 Colossi & 3 Stalkers & 5 Zealots	1 Colossi & 3 Stalkers & 5 Zealots	heterogeneous & symmetric

状态和观测（State and Observation）

在每一个时刻下，智能体都会收到在其视野范围内的局部观测，这包括有关每个单元周围的圆形区域内的地图信息，其半径等于视线范围。从每个智能体的角度看“视线范围”使环境部分可观。若智能体还活着并且位于视线范围内，他们也只能观察其他智能体。因此每个智能体无法确定他们的队友是在视线范围外还是已经死亡，这对合作造成巨大难度。

对于我方和敌方通常 feature vector 包含以下属性：位置坐标、相对距离 x、相对距离 y、血量、盾、单元类型。盾为额外的保护手段，在对单体的健康值造成伤害前，需必须将盾卸下才可产生伤害。神族的所有单位都有盾牌，如果不造成新的伤害盾牌可以重新生成（其他两个种族不具有该属性）。此外智能体可以获取视野内盟军上一时刻的 action。智能体还可以观察周边的地形特征，特别是固定半径的 8 个点的高度和可行军度（walkability）。

全局状态只有在集中式训练（centralized training）过程中可获取到，其包含了地图内所有单体的信息。特别地 state vector 包括所有智能体相对地图中心的坐标，以及当前所有智能体的 observations。此外 state 还包含了 Medivacs 掠夺者的 energy 以及其他单体的冷却值，其代表了两次攻击间最小的延迟时间。Central state 还会存储所有智能体上一时刻的动作。所有 state 或 observation 向量都被经过归一化，且所有智能体的视野范围为 9。

动作空间（Action Space）

离散空间的动作集合包括：*move[direction] (four directions: north, south, east, or west), attack[enemy_id], stop and no-op*。已经死亡的智能体只能做 *no-op* 动作，但活着的智能体不能做 *no-op*。作为抢劫单元，Medivacs 掠夺者必须使用 *heal[agent_id]*动作而不是 *attack[enemy_id]*。根据场景的不同，智能体可以采取的动作有 7~70 个。

为了保证分布式的任务，agent 只能攻击可射击范围（shooting range）内的敌人采取 *attack[enemy_id]*动作。此外这还会限制部队对远处的敌人使用 attack-move 的宏观策略。可射击范围被设置为 6，拥有比射击范围更大的视野会迫使智能体在射击之前利用移动的指令。

奖励（Rewards）

对于战斗场景的最终目标为获得最高的胜率。SMAC（星际的开源框架）提供了稀疏奖励（sparse rewards）的选项，在该条件下环境会在赢的对局返回+1 输的对局返回-1 作为一局的奖励。此外该框架还提供了默认选项，用于根据智能体造成和收到的生命值伤害计算出的 shaped rewards，在杀死敌方（盟军）单位后获得一些正（负）奖励。

最后附上最新 SOTA 算法在 StarCraft 不同微操环境下的表现和难度划分

Maps	Difficulty	MAPPO	IPPO	QMix	RODE	MAPPO(cut)	QMix(cut)
2m vs. 1z	Easy	**100.0(0.0)**	**100.0(0.0)**	95.3(5.2)	/	100.0(0.0)	100.0(2.9)
3m	Easy	**100.0(0.0)**	**100.0(0.0)**	96.9(1.3)	/	100.0(0.0)	93.8(1.9)
2s vs. 1sc	Easy	**100.0(0.0)**	**100.0(1.5)**	96.9(2.9)	**100(0.0)**	100.0(0.0)	96.9(0.7)
2s3z	Easy	**100.0(0.7)**	**100.0(0.0)**	95.3(2.5)	**100(0.0)**	100.0(1.5)	92.2(3.9)
3s vs. 3z	Easy	**100.0(0.0)**	**100.0(0.0)**	96.9(12.5)	/	100.0(0.0)	100.0(1.5)
3s vs. 4z	Easy	**100.0(0.9)**	99.2(1.5)	97.7(1.9)	/	100.0(1.9)	84.4(3.7)
so many baneling	Easy	**100.0(0.0)**	**100.0(1.5)**	96.9(2.3)	/	100.0(1.5)	81.2(2.6)
8m	Easy	**100.0(0.0)**	**100.0(0.7)**	97.7(1.9)	/	100.0(0.0)	93.8(2.7)
MMM	Easy	**96.9(2.6)**	**96.9(0.0)**	95.3(2.5)	/	93.8(2.6)	92.2(3.9)
1c3s5z	Easy	**100.0(0.0)**	**100.0(0.0)**	96.1(1.7)	**100(0.0)**	100.0(0.0)	95.3(1.5)
bane vs. bane	Easy	**100.0(0.0)**	**100.0(0.0)**	**100.0(0.9)**	**100(46.4)**	100.0(0.0)	100.0(0.0)
3s vs. 5z	Hard	96.9(37.5)	97.7(1.7)	**98.4(2.4)**	78.9(4.2)	96.9(41.7)	56.2(6.4)
2c vs. 64zg	Hard	**100.0(0.0)**	98.4(1.3)	92.2(4.0)	**100(0.0)**	96.9(2.5)	71.9(3.9)
8m vs. 9m	Hard	87.5(4.0)	89.8(4.5)	**92.2(2.0)**	/	78.1(15.1)	89.1(1.9)
25m	Hard	**100.0(1.5)**	**100.0(0.0)**	90.6(3.8)	/	98.4(3.3)	89.1(3.9)
5m vs. 6m	Hard	**75.0(18.2)**	57.0(15.6)	**75.0(6.9)**	71.1(9.2)	29.7(16.3)	54.7(2.2)
3s5z	Hard	**96.9(0.7)**	**96.9(1.5)**	88.3(2.9)	93.75(1.95)	71.9(15.6)	85.9(5.2)
10m vs. 11m	Hard	**96.9(4.8)**	93.0(7.4)	95.3(1.0)	95.3(2.2)	84.4(7.9)	82.8(4.8)
MMM2	Super Hard	**90.6(2.8)**	86.7(7.3)	87.5(2.6)	89.8(6.7)	46.9(23.0)	81.2(4.5)
3s5z vs. 3s6z	Super Hard	84.4(34.0)	82.8(19.1)	82.8(5.3)	**96.8(25.11)**	73.4(36.3)	50.0(7.8)
27m vs. 30m	Super Hard	93.8(2.4)	82.0(10.3)	50.0(10.5)	**96.8(1.5)**	93.8(3.8)	34.4(5.1)
6h vs. 8z	Super Hard	**86.7(11.8)**	84.4(33.3)	9.4(2.0)	78.1(37.0)	78.1(14.5)	3.1(1.5)
corridor	Super Hard	**100.0(1.2)**	98.4(3.1)	84.4(2.5)	90.6(18.3)	93.8(4.9)	70.3(13.6)

Table 1. Median evaluation win rate and standard deviation on all the SMAC maps for different methods, using at most 10M training timesteps. Columns with "cut" display results using the same number of timesteps as RODE.