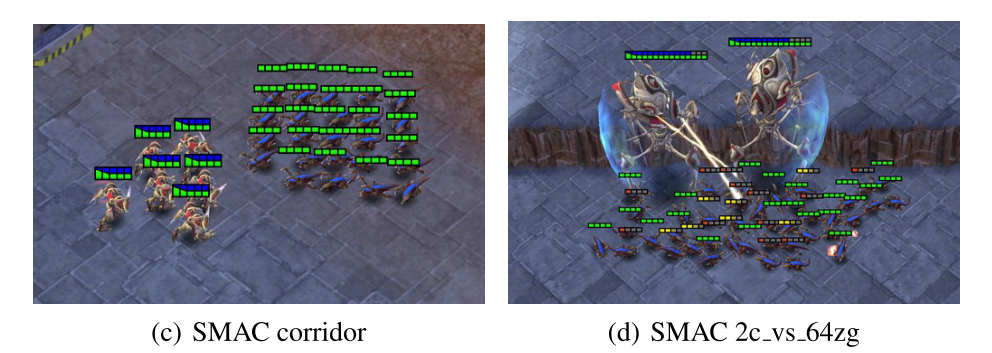
Introduction of a Popular Multi-Agent Reinforcement Learning Environment —StarCraft II

--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）



**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 & 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: 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

