Counterfactual Multi-Agent Policy Gradients——hoho

论文试图解决什么问题？

在多智能体环境：

1. 过去常用independent actor-critic模型，每个智能体单独训练自己的策略actor-critic模型，导致全局信息共享不足，难以做到coordination（协调）
2. 在一个合作的场景，联合动作只能生成全局的奖励，导致如何衡量每个智能体对团队的贡献比较困难（multi-agent credit assignmane problem）
3. 模型往往需要求不同agent采取action的联合概率分布，联合动作（joint action space）空间十分巨大，导致计算复杂度巨大

这是否是一个新的问题？

对于多智能体环境的credit assignment问题，不是一个新问题。

这篇文章要验证一个什么科学假设？

- 进行中心化的全局价值计算（centralised critic），以使获取全局的价值信息
- counterfactual baseline：基于奖励差分的思想（difference reward），即对每个智能体进行奖励塑形，用一个默认的动作（default action）替换后，计算其奖励，并与全局的奖励进行对比，得出这个智能体对团队的贡献，以此作为奖励分配。
- efficient critic representation，缓解计算复杂度问题

有哪些相关研究？如何归类？谁是这一课题在领域内值得关注的研究员？

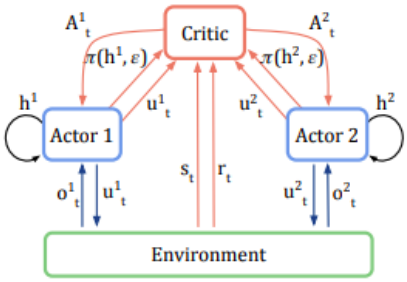
todo

论文中提到的解决方案之关键是什么？

借鉴优势函数的思想，本文提出方法COMA，基于中心化的价值函数（critic）,单独计算每个智能体的奖励基线——（此时其他智能体的动作保持固定），然后跟全局的价值函数作对比，进行反事实的推理：假如当时这个智能体采取另一个动作，会（不会）怎样！

- 中心化价值函数 centralised critic

整体架构如下：



critic学习一个全局的Q函数$Q(s,\mathbf{u})$，其中$\mathbf{u}$为所有智能体的联合动作。critic只会在训练时使用。

输入：环境的全局状态s_t，立即奖励r_t，还有各个智能体的动作输入u_t^a（表示第a个智能体的动作）和策略$\pi(h^a,\epsilon)$

当全局观测不可获得，则将当前所有agent的"action-observation"的历史记录τ代替全局状态 s

输出：每个智能体的动作优势A_t^a，以此来衡量每个智能体的贡献

2. 反事实的奖励基线 counterfactual baseline

通常的策略梯度可以建模为

$$\begin{aligned} g &= \partial_{\theta^\pi} \log \pi(u|\tau_t^a)(Q(u_t^a,s_t)-V(s_t)) \\ &\approx \partial_{\theta^\pi} \log \pi(u|\tau_t^a)(r+\gamma V(s_{t+1})-V(s_t)) \end{aligned}$$

但是如此一来就无法解决credit assignment问题：TD error考虑的是全局reward的影响，对于每个智能体来说（actor）无法显式确认它对于全局reward的贡献。

受difference reward思想的启发，本文使用了反事实推理：如果智能体当时不采取xxx动作，它的奖励将如何？

$$D^a = r(s,\mathbf{u}) - r(s,(\mathbf{u}^{-a},c^a))$$

将其他agent的action固定，只研究当前agent的action变化所产生的影响：

c^a为agent的默认动作，$(\mathbf{u}^{-a},c^a)$表示当前agent a 采取"默认行为" c^a后所有 agent 的联合动作空间，从而得到相对于默认动作，agent a实际所采取动作的优势。

但是要想计算出每一个动作的D^a值，就需要将每个动作都替换成默认行为 c^a去与环境互动一次得到最终结果，这样采样次数会非常多；另外，到底选择哪一个行为当作默认行为才是最合适的也是比较难决定的。因此，文中提出使用"函数拟合"的方式来计算D^a

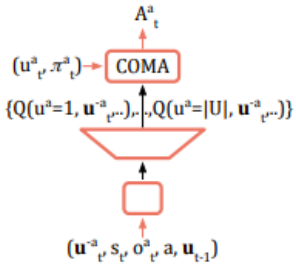
参考优势函数的思想，D^a可以拟合为：

$$A^a(s,\mathbf{u}) = Q(s,\mathbf{u}) - \sum_{u'^a}\pi^a(u'^a|\tau^a)Q(s,(\mathbf{u}^{-a},u'^a))$$

其中$(\mathbf{u}^{-a},u'^a)$表示在联合动作中用智能体a的动作u'^a替换其在联合动作$\mathbf{u}$中的部分。

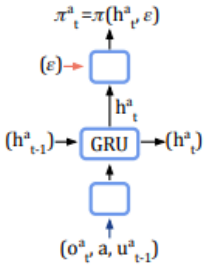
3. efficient critic representation

critic网络输出所有agent的所有动作即联合动作空间复杂度是$O(|U|^n)$，计算量太大。于是本文设计只输出当前agent的各个动作的Q值，网络架构如下：



将其他agents的action $\mathbf{u}^{-a}$，并结合历史信息作为输入。输出则是基于输入的其他agent的action下的当前agent的所有action的Q值（即$Q(s,(\mathbf{u}^{-a},u^{'a})$的部分），然后结合当前agent的$u_t^a$和$\pi_t^a$，相乘求和得到baseline，并与全局价值函数作差得到Advantage funciton，最终用于梯度下降。

另外，actor的架构如下：



论文中的实验是如何设计的？

使用星际争霸作为实验环境。设计多种不同的多智能体协作方式：

3个海军单位（3m）

5个海军单位（5m）

5个幽灵单位（5w）

2条龙和3个狂战士单位（2d_3z）

让算法控制的作战小队和游戏AI控制的小队进行对战，并计算胜率。

实验还加入了 "部分可观测" 条件的限制，视野范围等于攻击范围。这意味着当敌人没有进入攻击范围内时，作战单位是不知道敌人位置信息的，因此agent不仅要学会如何去探索敌方目标，还需要与队友共享敌方目标的位置信息。

- 动作空间与奖励设计

每个agent都有着相同的动作空间：{ move[direction]，attack[enemy_id]，stop，noop }。一个回合的全局reward：

$$global\_reward = \text{对敌人造成的伤害} - \frac{1}{2}\text{我方受到的伤害} + 10 * \text{被消灭的对方的作战单位数量}$$

当赢得游戏时，则获得的回报为整个队伍剩余血量另外加200。

- 环境状态设计

actor接收agent的局部观测信息；

critic接收全局状态信息。

1. 局部观测信息

由于作战单位的视野范围等于攻击范围，因此观测到的视野是以该单位为中心的一个圆。局部观测信息是指在视野圆圈内，**每一个单位**（包括敌方和友方）的以下几个信息：**distance，relative x，relative y，unit type，shield**。（shield指护盾，有些兵种攻击后有冷却期，该护盾用来吸收短时间对方的攻击）

2. 全局观测信息

全局观测信息包含了所有单位的**relative x，relative y，unit type，shield，healthy point，cooldown**信息，其中 relative 的坐标信息是相对据地图中心的相对坐标，不再是针对于某一个特定目标的坐标。

用于定量评估的数据集是什么？代码有没有开源？

- 没数据集
- 没官方代码

可参考第三方代码：https://github.com/opendilab/DI-engine/blob/main/ding/policy/coma.py

论文中的实验及结果有没有很好地支持需要验证的科学假设？

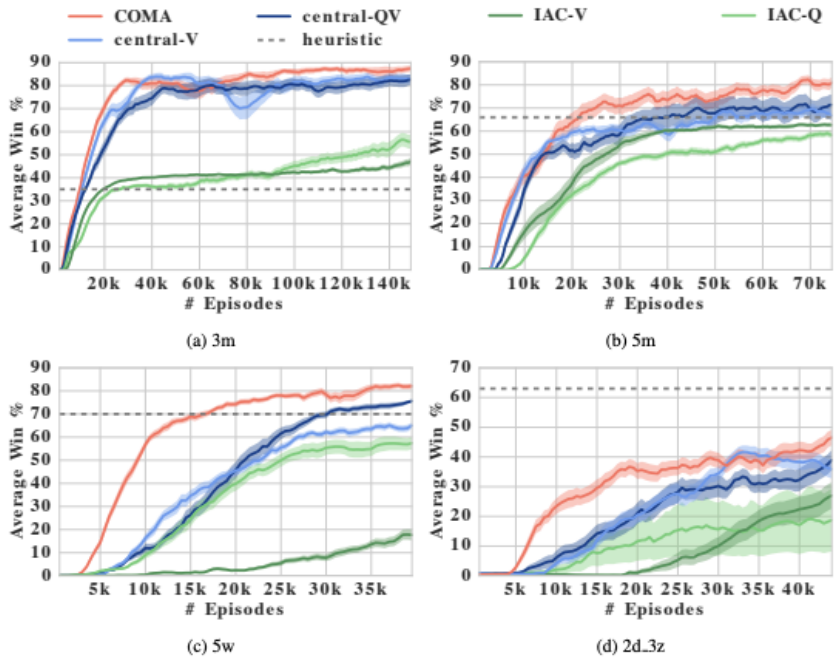
实验总体结果如下：

	Local Field of View (FoV)							Full FoV, Central Control		
map	heur.	IAC-V	IAC-Q	cnt-V	cnt-QV	COMA mean	COMA best	heur.	DQN	GMEZO
3m	35	47 (3)	56 (6)	83 (3)	83 (5)	**87** (3)	98	74	-	-
5m	66	63 (2)	58 (3)	67 (5)	71 (9)	**81** (5)	95	98	99	100
5w	70	18 (5)	57 (5)	65 (3)	76 (1)	**82** (3)	98	82	70	74[3]
2d_3z	**63**	27 (9)	19 (21)	36 (6)	39 (5)	47 (5)	65	68	61	90

- IAC-V：基于传统的多智能体独立actor-critic模型，只输出单个agent的V值
- IAC-Q：基于传统的多智能体独立actor-critic模型，只输出单个agent每个动作的Q值

- cnt-V：即central-V，学习中心化的critic，但只学习V函数，并使用TD error（$r + \gamma V(s_{t+1}) - V(s_t)$）进行优势函数的计算
- cnt-QV，即central-QV，同时学习Q值和V值，将本文COMA的countefactual baseline用V值替换来计算优势函数
- Local Field of View: 加入"部分可观测" 条件的限制，视野范围等于攻击范围
- Full FoV: 当不加入“部分可观测”的条件限制

各种兵种的实验对比：



这篇论文到底有什么贡献？

todo

下一步呢？有什么工作可以继续深入？

todo