强化学习如何使用内在动机?

本文在回顾内在动机的生理学知识的基础上，探讨了内在动机在强化学习中的应用。

机器之心分析师网络，作者：仵冀颖，编辑：Joni Zhong。

「**内在动机**」(Intrinsic Motivation) 这一概念最初是在心理学中提出并发展起来的。由于其在制造开放式学习机器和机器人方面的潜力，这一概念正日益受到认知科学的关注。

所谓动机（Motivation）是指生物体的行为受到三个因素影响：（1）不可抗拒的外部影响；（2）内在的需求、动力、计划等；（3）充当目标或动机的外部对象或情况。

第一个因素很大程度上独立于生物体的内部状态，例如，从痛苦刺激中反射性退出，这叫做**外在动机**（Extrinsic Motivation）。后两个因素涉及假设的内部状态，这些内部状态被认为是解释行为的必要条件，称之为**内在动机**。

从心理学的角度分析，研究内在动机的主要目的是解释克服行为主义学习和驱动理论的困难，例如：解释为什么动物会对一些中性刺激（突然的光照、喂食等）产生特定的条件反射等反应。另一方面，研究内在动机的目的是探讨行动在内在动机中的重要性，例如解释一个人设法通过其行为来影响环境或可以自主地设定自己的事实有关的重要性。

Baldassarre 在文献 [1] 中从生物学的角度探讨内在动机。特别地，他对于内在动机和外在动机的区别进行了详细的分析。外在动机是指因某些外部提供的奖励而做某事，而内在动机则是指「因为某件事本身具有趣味性或令人愉悦而做某事」。从进化的角度分析，外在动机指导人们学习直接提高适应度的行为，而内在动机推动人或者智能体本身知识和技能的获得，这些知识和技能有助于智能体产生只在后期才能够显现作用的行为。基于这一差异，外在动机根据涉及身体自我平衡调节的事件生成学习信号，而内在动机则根据发生在大脑内部的事件生成学习信号。

近年来，内在动机问题引起了计算建模和机器学习等领域研究人员的关注。在了解了内在动机的生物学原理的基础上，研究人员提出了一些计算理论和模型，这些模型和模型有助于以与生物学或理论观点相关的方式阐明内在动机的概念。

这类模型主要有两类：

一是**直接定义内在动机的生物学机制的模型**。针对具有学习预测因子（即世界模型）的系统所捕获的内在动机因素，通过构建自适应的神经网络模型生成基于预测误差或预测误差减少的内在学习信号，从而找到能够学习到最多知识的环境因素 [7]。

二是**与内在动机的进化理论直接相关的模型**，例如引入内在动机的强化学习（Reinforcement Learning，RL）框架 [8]，利用一个进化神经网络（强化器，Reinforcers）产生的内在强化信号学习基本的通用技能，进一步，能够将学习到的通用技能结合起来解决不同的机器人任务。

本文主要在回顾内在动机的生理学知识的基础上，探讨**内在动机在强化学习 RL 中的应用**。

在机器学习领域，一般认为强化学习 RL 框架只能处理外在动机，因为 RL 智能体（Agent）具有独特的输入通道，可以从其外部环境传递奖励信号。然而，研究人员证明，

RL 框架同样适合结合内在动机的原理。

RL 被称为「启发式动态规划」和「神经动力学规划」。RL 算法解决了行为智能体如何在与环境直接交互的同时学习最佳行为策略（通常称为策略 Policy）的问题。强化学习之父 Barto 在文献 [2] 中阐述了在 RL 框架中引入内在动机的可能性和重要性。由于环境学习的局限性，RL 存在智能体无法学习可靠策略的问题，通过引入内在动机可以帮助智能体解决环境局限性所带来的问题：**内在动机可使智能体能够学习有用的环境模型，从而帮助其更有效地学习其最终任务。**Tejas D. Kulkarni 在文献 [3] 中提出了一种 RL 框架，该框架将在不同时间范围内运行的层次化行动价值功能与目标驱动的内在动机的深度强化学习相结合。

另外，本文将介绍利用内在动机改进机器人强化学习的两项研究成果。

一、内在动机的生理学背景介绍 [1]

从生理学的角度讨论，因为进化（Evolution）是理解任何生物学现象的关键原理，所以内在动机和外在动机都可以看成进化的关键因素：生物所有行为的存在都是由有机体的生存和生殖的进化需要引起的：为了补偿有机体的生理缺陷（「组织需求」）、减少缺陷，生物激发了各种行为。因此所有行为都是通过其与原始驱动力的关联而被激发和引导的，成为直接学习或作为通过二次继续（加强）学习的结果。例如，生物体进化出可以增加其生存和特定生境中的存活率和生殖机会（健康度）的身体结构；肌肉和骨骼系统的进化使生物能够在环境中更好的移动；传感器（各种传感系统）的进化使生物能够更好的感知外部环境；大脑（神经系统）可以存储技能（即感觉运动图）和知识（即抽象感官和预测能力）等等。

虽然外在动机和内在动机都倾向于涉及共同的生物学进化的学习机制，并改变相同的大脑结构，但它们涉及不同的指导机制：内在动机是基于驱动技能和知识学习的机制的动机，以及基于在大脑内直接检测到的这些技能和知识的水平和变化而对行为和行为进行剥削和激励的动力。内在动机使得生物体能够学习技能和知识，而无需在获取技能时直接影响体内的稳态需求和健康状况。这些技能和知识通常有助于提高适应性，从而使得生物体可以用来相对较快地学习调节行为的稳态行为和复杂的行为。生物体感知世界，并在此基础上产生行为。行为会产生其他感知，并且如果生物体（即大脑本身）正在获得新的技能和/或知识，则大脑的内在动机机制会产生学习信号。

关于内在动机背后的机制，生物学家开展了关于多巴胺（dopamine，DA）、海马体（hippocampus，Hip）、神经调节剂乙酰胆碱（neuromodulators acetylcholine，ACh）和去甲肾上腺素（noradrenaline，NE）等的研究，分析了相关生物体对内在动机的反应情况。

总体来说，内在动机有两个区别于外在动机的特征：一是，内在动机可能在外在动机影响的情况下产生并发展，例如，家禽需要短期的父母照料，之后利用内在动机满足体内平衡的需要；二是，内在动机产生的学习信号在获得产生它们的技能或知识后趋于减少或消失，例如，当孩子学会可靠地产生一种新发现的作用或认识一个新颖的物体时，与它们有关的内在动机就会趋于停止，并开始将其活动引导到其他地方。

二、内在动机与强化学习 RL [2][3]

在本节中，我们展开讨论如何基于强化学习（Reinforcement Learning，RL）的计算理论引入与内在动机相关的概念。RL 解决的是行为智能体如何在与环境直接交互的同时学会学习最佳行为策略的问题。RL 由一些方法组成，这些方法可在控制器与被控制系统进行交互时，针对最佳控制问题逼近闭环解决方案。RL 强调如何学习预测值并用于指导行为，经典

的 RL 主要强调从外部通道（外部动机）获取奖励 (reward) 信号。因此，**RL 与动机研究自然相关。**

心理学家的研究表明，动机因素可以通过控制奖励的有效性以及控制学习结果在行为中的表达方式来影响学习。人工智能研究人员考虑，**是否可以向以从外部输入通道（外在动机）获取奖励信号的经典 RL 框架中引入内在动机来进一步改进 RL?** 具体而言，这种内在动机包括学习（Learning）、触发固定行为模式的先天机制等。

对于 RL 来说，学习和行为生成过程可以「不在乎」使用内在动机还是外在动机，即奖励信号是内在的还是外在的。这就决定了**RL 只需关注用于生成奖励信号的特定机制即可，并不需要专门区分该机制是内在还是外在动机。**

经典 RL 框架中 RL 智能体与其环境交互的一般视图如图 1 所示。智能体在该环境的感知状态的上下文中生成动作，它的行为（Action）会随着时间的推移影响环境的状态（State）。该环境（Environment）包含一个「批判函数（Critic）」，该批判函数在每个时间步骤为智能体提供对其正在进行的行为的评估（数字评分）。智能体通过在一段时间内学习如何从环境中传递更大幅度的奖励信号来提高其控制环境的技能，其中，从状态到批判函数实施的奖励信号的映射称为奖励函数（reward function）。

智能体的目标是通过行为以最大程度地衡量其期望在未来获得的奖励信号（Reward Signals）。该度量可以是预期在将来收到的奖励信号的简单总和，也可以是一种加权计算的结果，即较晚的奖励信号的权重小于较早的奖励信号的权重。由于智能体的行为会影响环境状态随时间的变化，因此，必须最大化预期奖励，即需要智能体对其环境状态的演变施加控制。智能体通过调整政策（Policy）来实现这一目标，该政策是将行动与观察到的环境状态相关联的规则。

注意图 1 中标有「环境（Environment）」的框不仅表示动物或机器人外部世界中的内容，还表示相对于奖励学习而言外部的内容，这部分内容可能仍位于动物或机器人内部。在 RL 框架中包含动机因素的起点是要弄清楚我们所说的「内部状态」是什么。

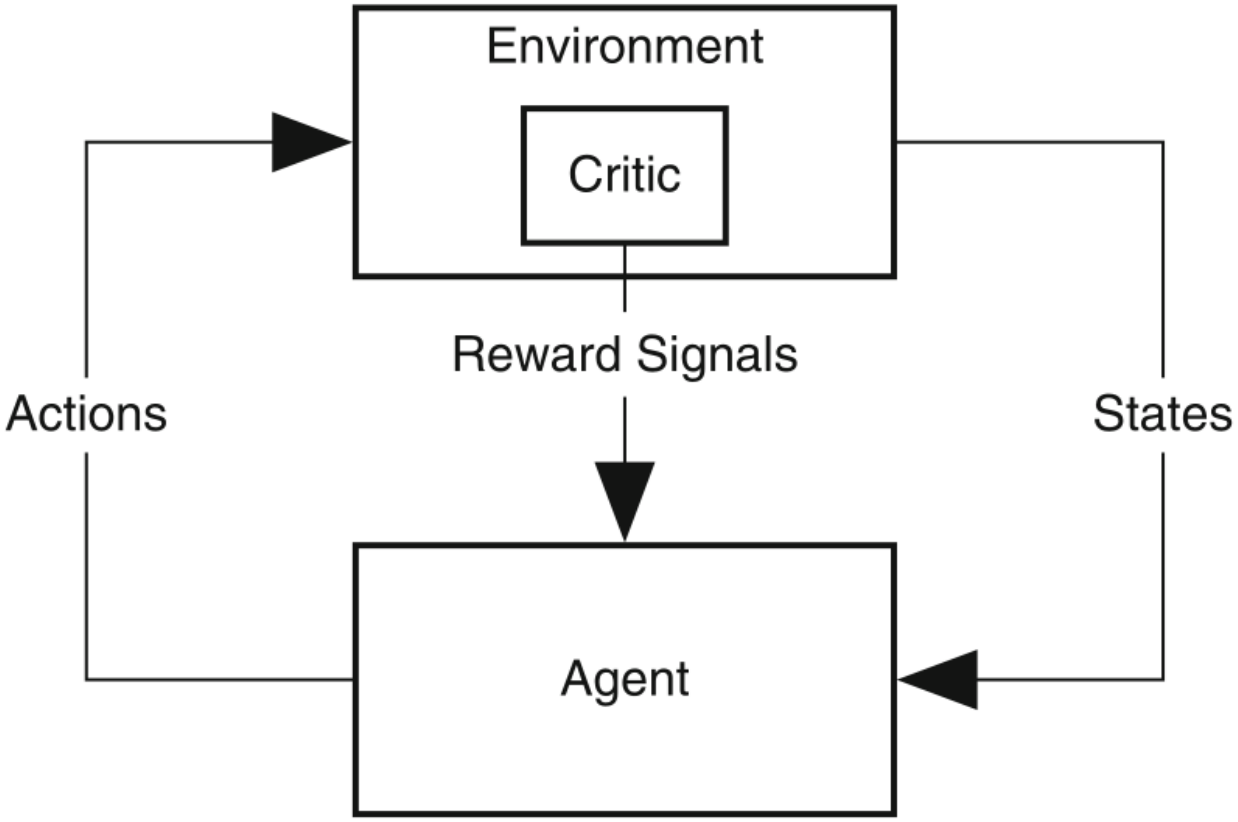


图 1. RL 中的主体与环境互动。主要奖励信号是从其环境中的「批评者」提供给智能体的。

图 2 是对图 1 的改进，将环境分为外部环境和内部环境。外部环境表示动物或机器人外部的内容，而内部环境则包含生物体内的组成部分。这两个组成部分共同构成了生物或机器人的环境。某些奖励信号可能是由外部环境中的物体或事件产生的感觉触发到身体产生的，例如拍打头部或赞美之词；其他因素可能是外部刺激和内部环境条件（如口渴的饮用水）共同引发的。但总体来说，所有奖励信号归根结底都是在生物体内生成的。

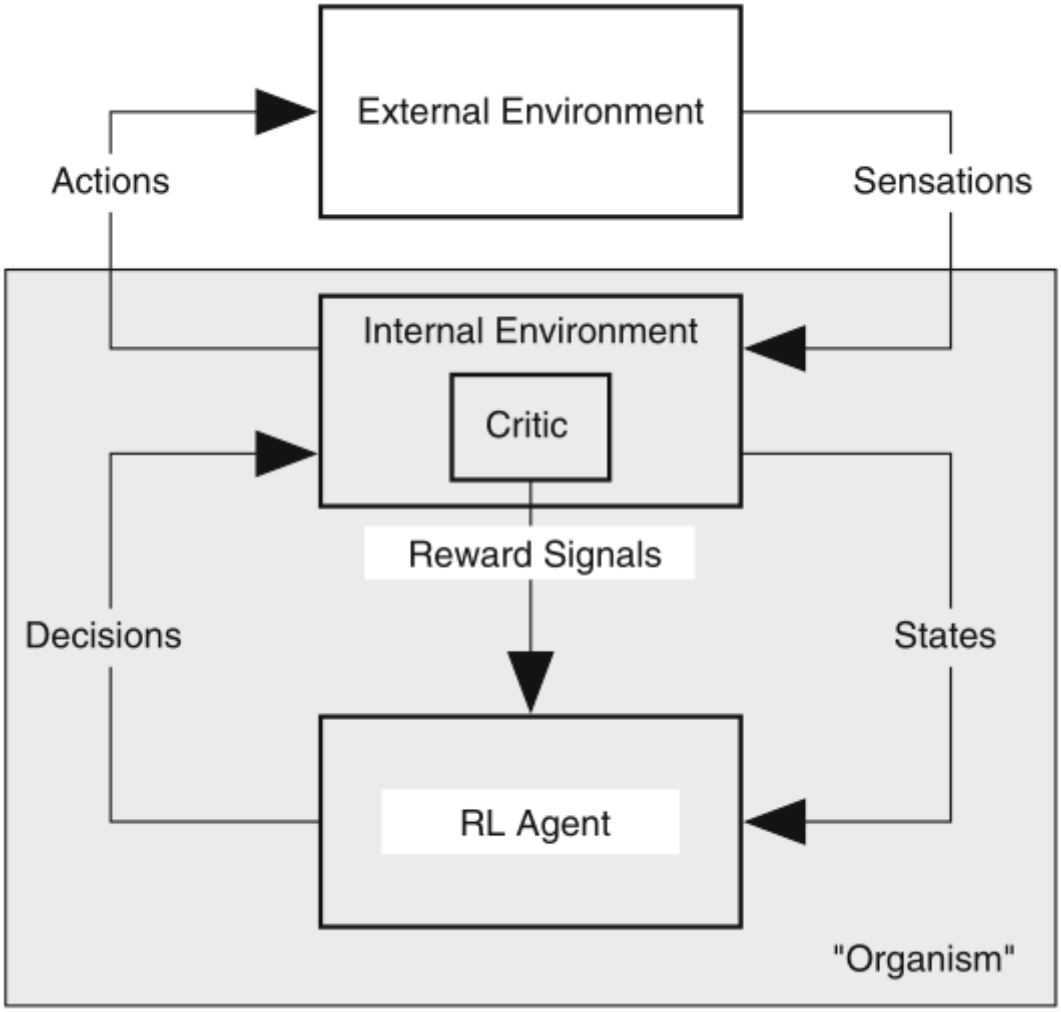


图 2. RL 中的智能体与环境互动，其中环境分为内部和外部环境，所有奖励信号均来自前者。阴影框对应于生物有机体。

然而，在构建 RL 框架的时候很难清晰地区分外部和内部奖励信号。在文献 [2] 中，从进化的角度提出从明显的外部奖励信号到明显的内部奖励信号其实是连续性的，内部奖励信号随着进化而改变，但这种改变相对于外部奖励信号来讲相对缓慢。因此，内部奖励信号与演化具有比较明显的恒定关系。因此，希望 RL 中的主要激励能够鼓励涉及学习系统这一部分环境特征的多种行为，包括涉及好奇心、新颖性、惊奇以及通常与内在奖励相关的其他内部介导特征的行为。这将给改进 RL 提供新的思路和方向。

为了解决在稀疏反馈的环境中学习目标导向的行为，文献 [3] 中提出了分层深度 Q 网络强化学习（hierarchical-DQN，h-DQN）框架。如图 3 所示，该框架集成了分层的行动价值函数，在不同的时间尺度上运行，并具有目标驱动的内在动机的深度强化学习。图 3 给出了 h-DQN 的整体框架，可以看出图 3 的框架是对图 2 框架的扩展。其中顶层 q 函数学习有关内在目标的策略，而下层 q 函数学习关于满足给定目标的原子动作的策略。

在该框架中，强化学习（RL）将控制问题的形式转化为寻找使预期的未来回报最大化的政策 π。价值函数 V(s) 是 RL 的总体目标，也可以概括为 V(s, g)，以表示状态 s 用于实现给定目标 g∈G 的情况。当环境提供延迟的奖励时，首先学习实现内在生成的目标的方法，然后学习将它们连接在一起的最佳策略。每个价值函数 V(s, g) 可用于生成一个策略，该策略在智能体到达目标状态 g 时终止。这些策略的集合可以在半马尔可夫决策过程的框架内按时间动态地分层安排，以进行学习或完成计划。在高维问题中，这些价值函数可以通过神经网络近似为 V(s, g; θ)。

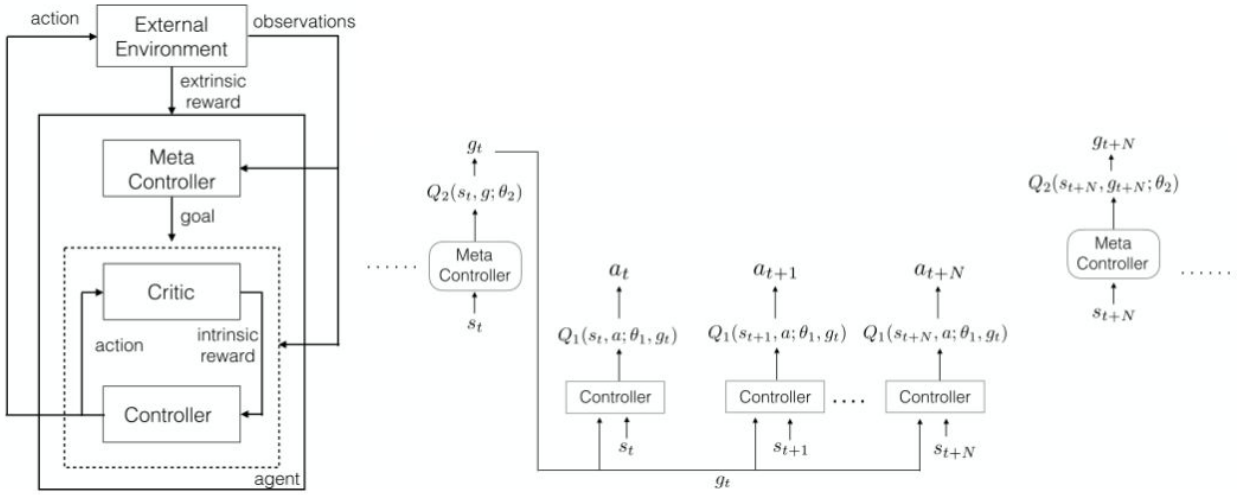


图 3. h-DQN 整体框架。

该框架具有在不同时间范围内工作的分层组织的深度强化学习模块，因此，允许灵活的目标定义。该模型在两个层次上进行决策：（a）顶层模块（元控制器）接受状态并选择新的目标；（b）下层模块（控制器）同时使用状态和所选目标来选择操作，直到达到目标或事件终止为止。

之后，定义元控制器选择另一个目标，并重复步骤（a-b）。在不同的时间尺度上使用随机梯度下降训练模型，以优化预期的未来内在（控制器）和外在奖励（元控制器）。其中，元控制器 (Meta-controller) 和控制器（Controller）使用单独的 DQN。元控制器查看原始状态，并通过估计动作值函数 Q2(s_t, g_t; θ_2) 来制定针对目标的策略（以最大化预期的未来外部奖励）。控制器接受状态和当前目标，并通过估计行动值函数 Q1(s_t, a_t; θ_1, g_t) 来估计行动（通过最大化预期的未来内在报酬）来产生行动策略。当且仅当达到目标时，内部批判函数才能为控制者提供积极的奖励。当一个阶段的工作结束或达到 g 的目标时，控制器终止。然后，元控制器选择一个新的 g，重复该过程。具体的算法学习流程如下：

Algorithm 1 Learning algorithm for h-DQN

Initialize experience replay memories $\{\mathcal{D}_1, \mathcal{D}_2\}$ and parameters $\{\theta_1, \theta_2\}$ for the controller and meta-controller respectively.
Initialize exploration probability $\epsilon_{1,g} = 1$ for the controller for all goals g and $\epsilon_2 = 1$ for the meta-controller.
for $i = 1, num\_episodes$ **do**
    Initialize game and get start state description s
    $g \leftarrow \text{EPSGREEDY}(s, \mathcal{G}, \epsilon_2, Q_2)$
    while s is **not** terminal **do**
        $F \leftarrow 0$
        $s_0 \leftarrow s$
        while not (s is terminal **or** goal g reached) **do**
            $a \leftarrow \text{EPSGREEDY}(\{s, g\}, \mathcal{A}, \epsilon_{1,g}, Q_1)$
            Execute a and obtain next state s' and extrinsic reward f from environment
            Obtain intrinsic reward $r(s, a, s')$ from internal critic
            Store transition $(\{s, g\}, a, r, \{s', g\})$ in $\mathcal{D}_1$
            $\text{UPDATEPARAMS}(\mathcal{L}_1(\theta_{1,i}), \mathcal{D}_1)$
            $\text{UPDATEPARAMS}(\mathcal{L}_2(\theta_{2,i}), \mathcal{D}_2)$
            $F \leftarrow F + f$
            $s \leftarrow s'$
        end while
        Store transition (s_0, g, F, s') in $\mathcal{D}_2$
        if s is **not** terminal **then**
            $g \leftarrow \text{EPSGREEDY}(s, \mathcal{G}, \epsilon_2, Q_2)$
        end if
    end while
    Anneal ϵ_2 and ϵ_1.
end for

h-DQN 使用 Deep Q-Learning 框架 [9] 来学习控制器和元控制器的策略。具体而言，控制器 Q1 估算以下 Q 值函数：

$$Q_1^*(s, a; g) = \max_{\pi_{ag}} \mathrm{E}\Big[\sum_{t'=t}^{\infty} \gamma^{t'-t} r_{t'} \mid s_t = s, a_t = a, g_t = g, \pi_{ag}\Big]$$
$$= \max_{\pi_{ag}} \mathrm{E}\big[r_t + \gamma \max_{a_{t+1}} Q_1^*(s_{t+1}, a_{t+1}; g) \mid s_t = s, a_t = a, g_t = g, \pi_{ag}\big]$$

其中 g 是状态 s 中代理的目标，而 π_ag 是操作策略。同样，对于元控制器 Q2，我们有：

$$Q_2^*(s, g) = \max_{\pi_g} \mathrm{E}\Big[\sum_{t'=t}^{t+N} f_{t'} + \gamma \max_{g'} Q_2^*(s_{t+N}, g') \mid s_t = s, g_t = g, \pi_g\Big]$$

其中 N 表示直到控制器停止给定当前目标为止的所用步长，g' 是状态 s_t+N 下智能体的目标，而 π_g 则是基于目标制定策略。本文使用参数为θ的非线性函数逼近器来表示 Q*(s, g) ≈ Q(s, g; θ)。每个 Q∈{Q1，Q2} 可以通过最小化相应的损耗函数 L1(θ1) 和 L2(θ2) 来训练。Q1 的损失函数可以表示为：

$$L_1(\theta_{1,i}) = \mathrm{E}_{(s,a,g,r,s') \sim D_1}\left[(y_{1,i} - Q_1(s, a; \theta_{1,i}, g))^2\right]$$

其中 i 表示训练迭代次数，y_1,i = r +γ max_a' Q1(s_0, a_0; θ_{1，i-1}, g)。可以使用梯度优化参数 θ1：

$$\nabla_{\theta_{1,i}} L_1(\theta_{1,i}) = \mathrm{E}_{(s,a,r,s' \sim D_1)} \left[\left(r + \gamma \max_{a'} Q_1(s', a'; \theta_{1,i-1}, g) - Q_1(s, a; \theta_{1,i}, g) \right) \nabla_{\theta_{1,i}} Q_1(s, a; \theta_{1,i}, g) \right]$$

损失函数 L2 及其梯度可以使用类似的过程计算得出。使用随机梯度下降法在不同的时间尺度上学习 h-DQN 的参数，其中在每个时间步都收集来自控制器的状态变化，但是仅在控制器终止时收集元控制器的状态变化。

三、内在动机在机器人学中的应用

本节中，我们选择了两篇论文具体探讨如何在构建 RL 框架的过程中引入内在动机，从而改进机器人的动作完成效果。

1. Intrinsically motivated model learning for developing curious robots

论文地址：http://www.cs.utexas.edu/users/pstone/Papers/bib2html-links/AIJ15-Hester.pdf

引入内在动机的 RL 框架能够使得智能体学习更有用的环境模型，从而帮助其更好的学习完成最终任务。这种 RL 框架非常适合机器人应用，因为机器人需要了解可以应用于不同任务的动态和能力。文献 [4] 提出了一种使用方差和内在创新奖励算法（Variance-And-Novelty-Intrinsic-Rewards algorithm，texplore-vanir）的基于内在动机的 RL 框架。该框架计算两种不同的内在动机：一种是探索模型的不确定内容，另一种是获得尚未对该模型进行过训练的创新经验。

Texplore-vanir 遵循了基于模型的 RL 智能体的典型方法。它使用其学习到的模型（包括内在奖励）来规划一个策略，按照该策略采取行动，获取新的经验来改进其模型，然后重复进行。为了适用于机器人任务，texplore-vanir 采用了基于实时模型的架构 [10]。该架构使用近似规划并将模型的学习、规划和行动并行化，使智能体可以在指定的频率下实时地采取行动。texplore-vanir 将模型学习任务作为一个监督学习问题，将当前状态和动作作为输入，下一个状态作为要预测的输出。以 (s，a) 作为输入，s'-s 和 r 作为待预测的输出。模型学习通过监督学习器对未访问或不经常访问的状态进行预测的能力来加速模型学习。

与基于动态贝叶斯网络（Dynamic Bayesian Network，DBN）的 RL 算法一样，本文提出的算法通过学习每个 n-状态特征和奖励的单独预测，学习一个事实域。马尔可夫决策过程（Markov Decision Process，MDP）模型由 n 个预测每个特征的模型 (featModel1 到 featModeln) 和一个预测奖励的模型 (rewardModel) 组成。每个模型都可以被查询到对特定状态行为的预测（featModel⇒query(s,a)），或者用新的训练经验更新（featModel⇒update(s,a,out)）。在 texplore-vanir 中，这些模型中的每一个都是一个随机森林。texplore-vanir 模型学习算法从计算状态的相对变化（s.^rel）开始，然后它用新的相对变化来更新每个特征的模型，并更新奖励模型。和基于 DBN 的算法一样，texplore-vanir 假设每个状态变量都是独立过渡的。因此，独立的特征预测可以组合起来，形成一个完整的状态向量的预测。智能体得到每个特征的变化值的预测值，然后将 s.^rel 加在一起，就可以得到 s' 的预测值。

本文的算法不是针对这个单一的模型进行规划，而是以随机森林的形式，对可能的树模型的分布进行规划来驱动探索。一个随机森林是一个决策树的集合，每个决策树都不同，因为它们是在一个随机的经验子集上训练的，并且在选择决策节点上的分叉时有一定的随机性。为了增加模型的随机性，在树中的每 m 个决策树只在智能体的经验子集

(<s,a,s',r>tuples) 上训练，因为它随着每一个新的经验以概率 w 更新。texplore-vanir 的动作值是：

$$Q(s,a) = \frac{1}{m}\sum_{i=1}^{m} R_i(s,a) + \gamma \frac{1}{m}\sum_{i=1}^{m}\sum_{s'} P_i(s'|s,a)\max_{a'} Q(s',a')$$

使用最佳的内在奖励（intrinsic reward）来提高模型学习效率，这在很大程度上取决于被学习模型的类型。对于本文提出的随机森林模型 texplore-vanirs，假设以下两种内在动机的表现最好：1）倾向于探索状态空间中模型存在较大不确定性的区域；2）倾向于探索状态空间中远离之前探索过的区域（不管模型的确定性如何）。对于 1），texplore-vanir 计算对给定的状态行为的每个状态特征的预测方差的度量：

$$D(s,a) = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{m} D_{KL}(P_j(x_i|s,a)||P_k(x_i|s,a))$$

其中，对于森林中的每一对模型 (j 和 k)，它是每个特征 i 的预测概率分布之间的 KL 差值之和。D(s, a) 衡量不同模型的预测结果的差异程度。将一个与这种变异度量成正比的内在报酬，即变异报酬，纳入智能体的计划模型中：

$$R(s,a) = vD(s,a)$$

其中 v 是一个系数，决定这个奖励应该有多大。这种奖励将驱使智能体达到一种状态反应中，而此时其模型尚未收敛到全局动态的单一假设。因此会出现智能体的所有模型都做出错误预测的情况。对于 texplore-vaniruses 的随机森林模型来说，当它的模型不得不将自己的预测一般化到离它所训练的经验更远的地方时，就更容易出错。引入 2）对应的内在动机，在特征空间中，从一个给定的状态动作和模型训练过的最近的动作之间的 L1 距离。这个距离是针对每个动作分别计算的。对于一个动作 a，X_a 是该动作被采取的所有状态的集合。那么，δ(s，a) 是从给定状态 s 到最近的行动 a 被采取的状态的最小 L1 距离：

$$\delta(s,a) = \min_{s_x \in X_a} ||s - s_x||_1$$

其中每个特征被归一化为 0 到 1 的范围。对于布尔特征，假定真和假之间的距离为 1。定义一个奖励函数（Novelty-reward），该奖励函数与这一距离成正比，它促使智能体去探索与之前所见过的状态反应相比最新颖、最具创新性的状态反应：

$$R(s,a) = n\delta(s,a)$$

其中 n 是一个系数，决定了这个奖励的大小。这个奖励的一个很好的属性是，给定足够的时间，它将驱动智能体探索该域中的所有状态行为，因为任何未被访问的状态行为在某些特征上都与被访问的状态行为不同。

本文提出的 texplore-vanir 是通过这两种内在奖励的组合完成的。它们可以与不同权重的系数 (v 和 n) 相结合，或者与定义任务的外部奖励相结合。两种内在奖励的结合能促使智能体更有效地学习一个模型，同时以发展和好奇的方式进行探索：寻找创新有趣的状态动作，同时探索领域中更加复杂的部分。其随机代码如下算法所示。

Algorithm 1 MODEL.

procedure INIT-MODEL(n) ▷ n is the number of state variables
  for $i = 1 \rightarrow n$ **do**
    $featModel_i \Rightarrow$ INIT() ▷ Init random forest to predict feature i
  end for
  for $i = 1 \rightarrow nactions$ **do**
    $X_i \leftarrow \emptyset$ ▷ Init visited state set for action i
  end for
end procedure

procedure UPDATE-MODEL($list$) ▷ Update model with $list$ of experiences
  for all $\langle s, a, s' \rangle \in list$ **do**
    $s^{rel} \leftarrow s' - s$ ▷ Calculate relative effect
    for all $s_i^{rel} \in s^{rel}$ **do**
      $featModel_i \Rightarrow$ UPDATE($\langle s, a \rangle, s_i^{rel}$) ▷ Train feature model
    end for
    $X_a \leftarrow X_a \cup s$ ▷ Add s to visited set for action a
  end for
end procedure

procedure QUERY-MODEL(s, a, v, n) ▷ Get prediction of $\langle s', r \rangle$ for s, a
  for $i = 1 \rightarrow$ LENGTH(s) **do**
    $s_i^{rel} \leftarrow featModel_i \Rightarrow$ QUERY($\langle s, a \rangle$) ▷ Sample prediction for feat i
  end for
  $s' \leftarrow s + \left\langle s_1^{rel}, \ldots, s_n^{rel} \right\rangle$ ▷ Get absolute next state
  $D(s, a) \leftarrow \sum_{i=1}^{n} featModel_i \Rightarrow$ VARIANCE($\langle s, a \rangle$) ▷ Calculate variance.
  ▷ Each forest returns $\sum_{j=1}^{m} \sum_{k=1}^{m} D_{KL}(P_j(x_i|s,a)||P_k(x_i|s,a))$
  $\delta(s, a) \leftarrow \min_{s_x \in X_a} ||s - s_x||_1$ ▷ Calculate model novelty
  $r_{var} \leftarrow v D(s, a)$ ▷ Calculate variance reward
  $r_{nov} \leftarrow n\delta(s, a)$ ▷ Calculate novelty reward
  $r \leftarrow r_{var} + r_{nov}$
  return $\langle s', r \rangle$ ▷ Return sampled next state and reward
end procedure

作者给出了模拟域和真实机器人的实验。图 4 为模拟域任务—光明世界。在每一个房间里，智能体必须导航到钥匙，拿起钥匙，导航到锁上，按一下，然后导航到下一个房间的门，再通过门出去，进入下一个房间。图 5 显示了这些分布之间的变异距离，在 5000 个采样的状态反应中的平均数。该图显示，与其他方法相比，texplore-vanirlearns 模型的准确度显著高于其他方法（p <0.025）。

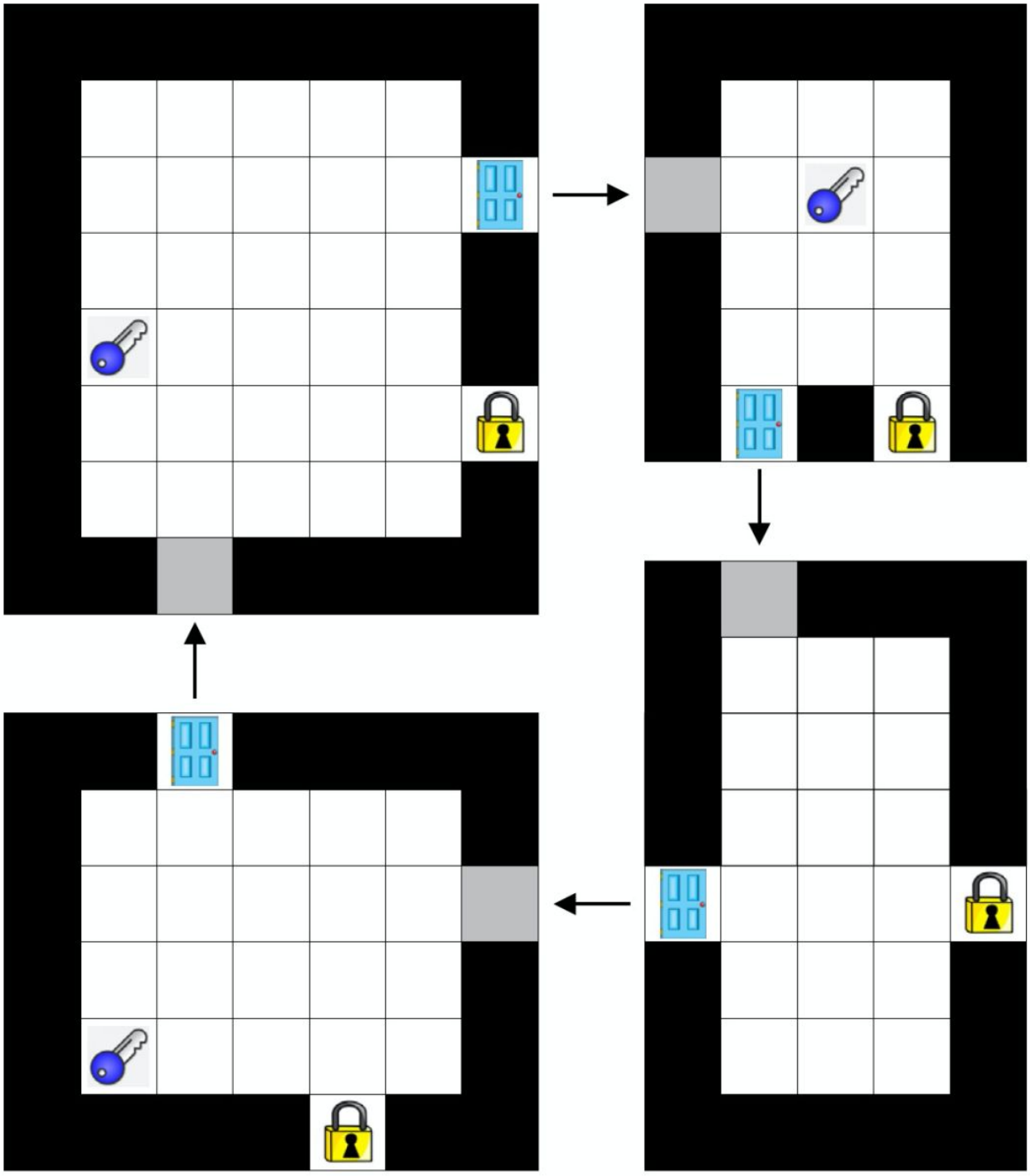


图 4. 光明世界任务。

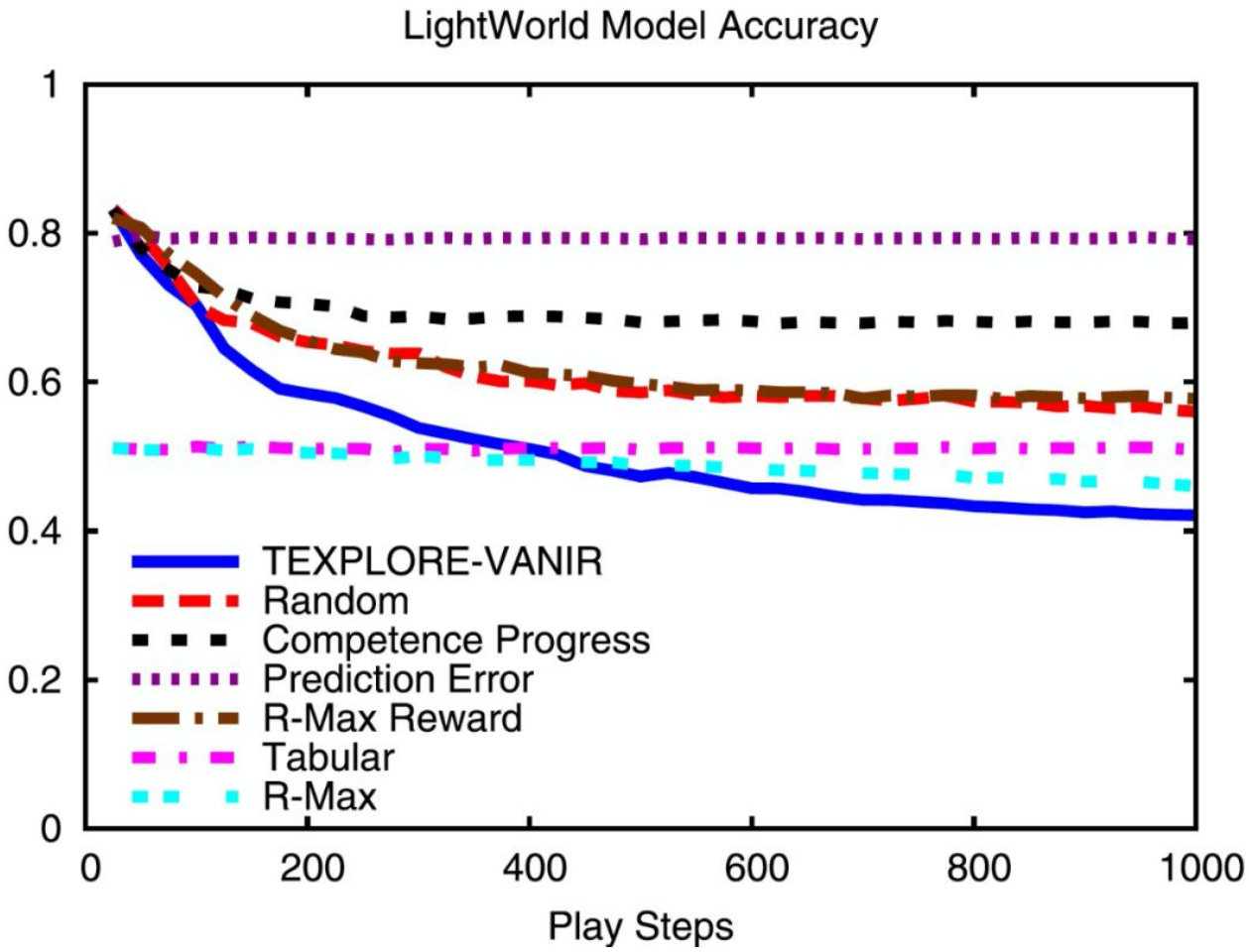


图 5. 不同算法模型的准确度与智能体所采取的步数相比，在 30 次实验和 5000 个随机抽样的状态作用下的平均值。

真实机器人实验是在一个特定的场景中，控制 Aldebaran Nao 机器人的手臂。机器人学习视频取自 https://www.cs.utexas.edu/~AustinVilla/?p=research/vanir。模型可以控制机器人的两个右肩关节。

机器人可以选择五种动作之一：将左（右）关节的角度增加 8 度、将左（右）关节的角度减少 8 度、或者什么都不做。智能体的状态由 8 个状态特征组成：两个肩关节的角度、机器人的手相对于胸前的三维位置（以毫米为单位）、机器人的摄像头图像中能看到多少个粉红色的像素、机器人的右脚按钮是否被按下、机器人的麦克风上听到的能量大小。机器人以坐着的姿势放置，它的右脚按钮可以被右手触摸，在它的右手边有一个钹。另外它的手臂上还挂着一个粉红色的方块，可以在摄像头前移动。

机器人以 3 赫兹的频率做动作，让手臂在每次动作后有时间到达新的位置（如图 6）。实验目的是计算在探索过程中机器人右手能够按下按钮的次数。

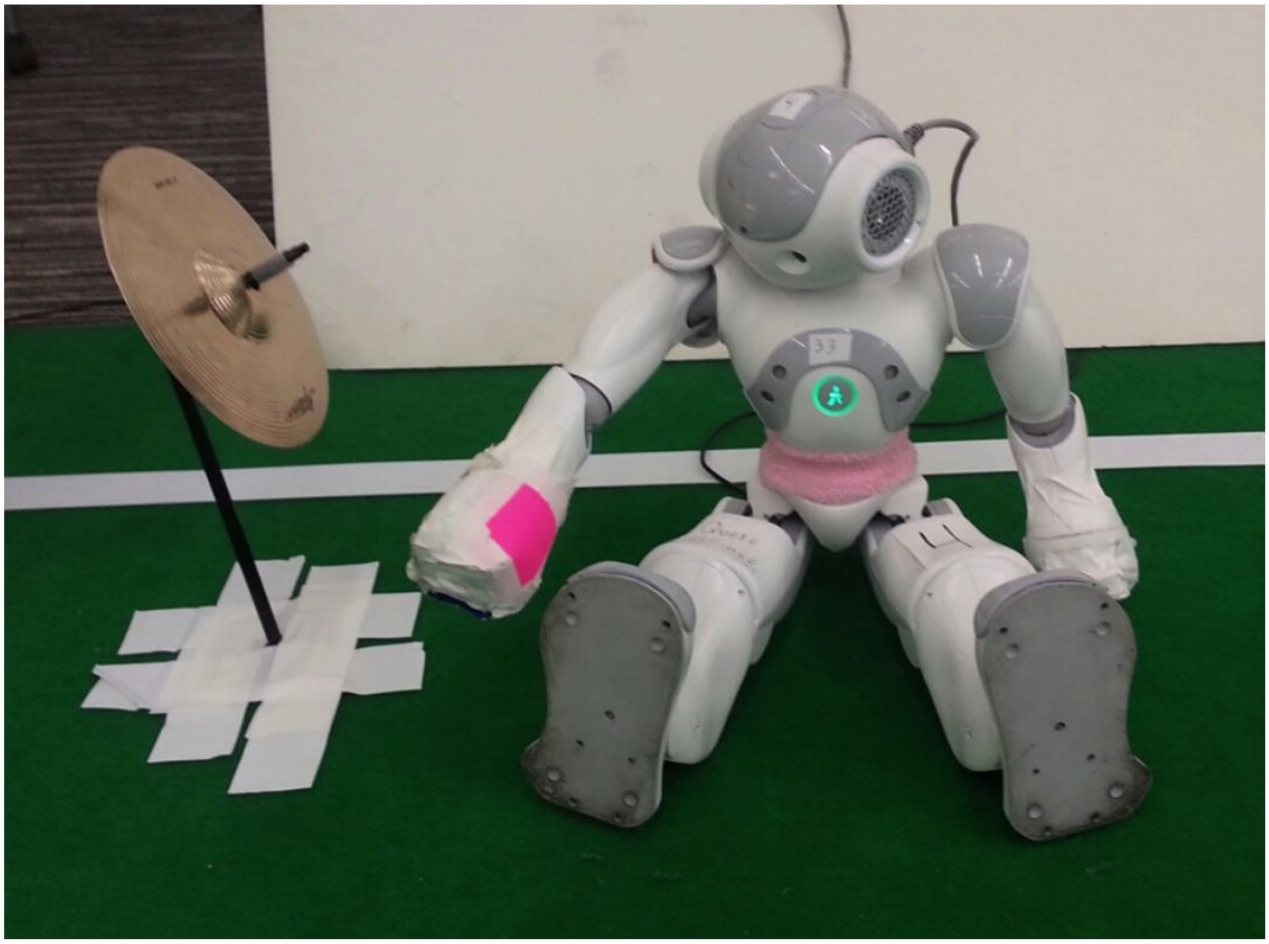


图 6. Aldebaran Nao 机器人。

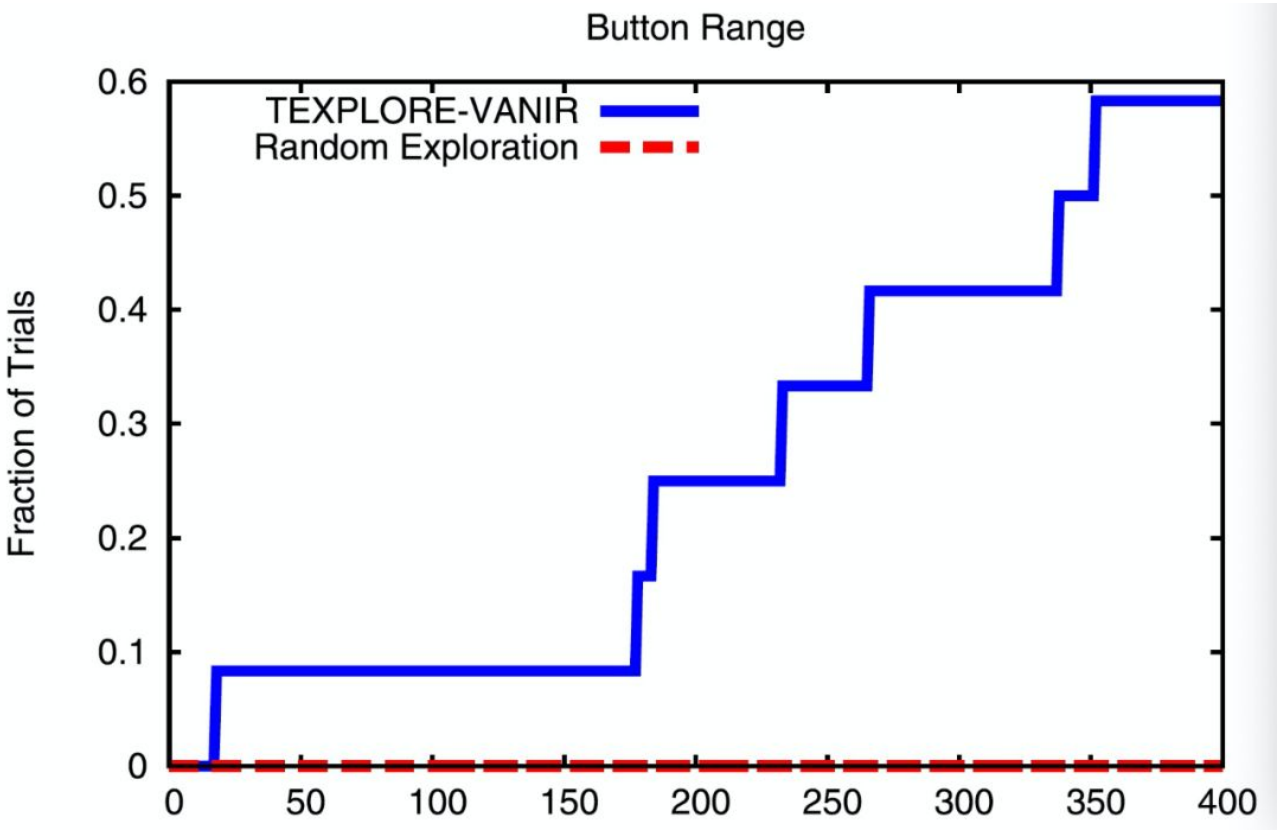


图 7. 在 13 次实验中，机器人在 400 个探索步骤中按下右脚按钮的比例。

图 7 给出在 400 步探索阶段中，机器人能够完成按下右脚旁边按钮任务的实验的比例。在 400 步探索阶段结束时，texplore-vanir 智能体在 13 个实验中的 7 个实验中按下了右脚旁边的按钮。相比之下，在进行随机探索时，没有一个实验按下了按钮。

2. Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning

论文地址：https://arxiv.org/pdf/1810.08647.pdf

这篇文章聚焦的是解决 RL 跨多个领域学习的问题，即多主体强化学习（multi-agent reinforcement learning，MARL）。在这类问题中，如果智能体对其它智能体的行为具有因果影响，就获得奖励。本文关注一个重要的内在动机，也是人类学习的关键：**社会交往**。本文提出了一个针对多智能体 RL（MARL）设计的内在动机奖励函数，该函数可以奖励对其他智能体行为有因果影响的智能体。作者使用反事实推理法评估因果影响。在每个时间点上，智能体模拟自己本可以采取的另一种反事实行动，并评估这些行动对另一个智能体行为的影响。能够导致其他智能体发生相对较高变化的行为被认为是影响较大的行为，会得到奖励。作者假设：奖励性的影响可能因此鼓励智能体之间的合作。

本文所采用的实验环境是具有挑战性的多智能体环境，该环境具有类似于囚徒困境的游戏理论奖励结构。对于每个智能体个体来说，「叛逃」（不合作行为）的奖励最高。然而，如果所有的智能体都选择合作，那么集体奖励会更好。这些任务奖励规则会自相矛盾，但也模拟了合作社会的动态性，对于典型的 RL 智能体来说具有极大的挑战性。

考虑一个由元组 <S,T,A,r> 定义的 MARL 马尔可夫博弈。在这个博弈中，多个不共享权重的智能体被训练成独立的个体，目标是让奖励最大化。在每个时间点 t，每个智能体选择一个动作。根据状态转换函数 T，所有 N 个智能体的行动组合成一个联合行动，根据状态转换函数 T，在环境中产生一个转换。每个智能体都会得到自己的奖励，该奖励可能取决于其他智能体的行为。本文考虑一个部分可观察的设置，意思是每个智能体只能看到一部分真实状态。每个智能体都寻求最大限度地提高自己的预期未来总回报。

引入内在动机因素的奖励函数为：

$$R^k = \alpha E^k + \beta I^k$$

其中，E 为外部或环境奖励，I 为因果奖励。作者通过生成智能体在每个时间点可能采取的反事实行动来计算 I，并评估采取这些行动会对其他智能体的行为产生怎样的影响。假设有两个智能体 A 和 B，智能体 B 在时间 t 时接收到 A 的行为作为输入。然后，智能体 B 使用它来计算其自身行为的分布。因为已经为 B 建立了 LSTM 模型，所以我们知道它的所有输入以及它自己内部的 LSTM 状态，如图 8 所示。这可以让我们通过限制在这个时间步长上观察到的其他输入的值来精确隔离 A 的行为对 B 的因果关系。

给定 A 的行为，B 的边际策略（marginal policy）与 B 的条件策略（conditional policy）之间的差异是 A 对 B 的因果影响的度量。它给出了 B 由于 A 的行为而改变其计划的行动分配的程度：

$$I_t^A = D_{KL}\left[p(a_t^B \mid a_t^A, z_t) \middle\| \sum_{\tilde{a}_t^A} p(a_t^B \mid z_t, \tilde{a}_t^A) p(\tilde{a}_t^A \mid z_t)\right] = D_{KL}\left[p(a_t^B \mid a_t^A, z_t) \middle\| p(a_t^B \mid z_t)\right]$$

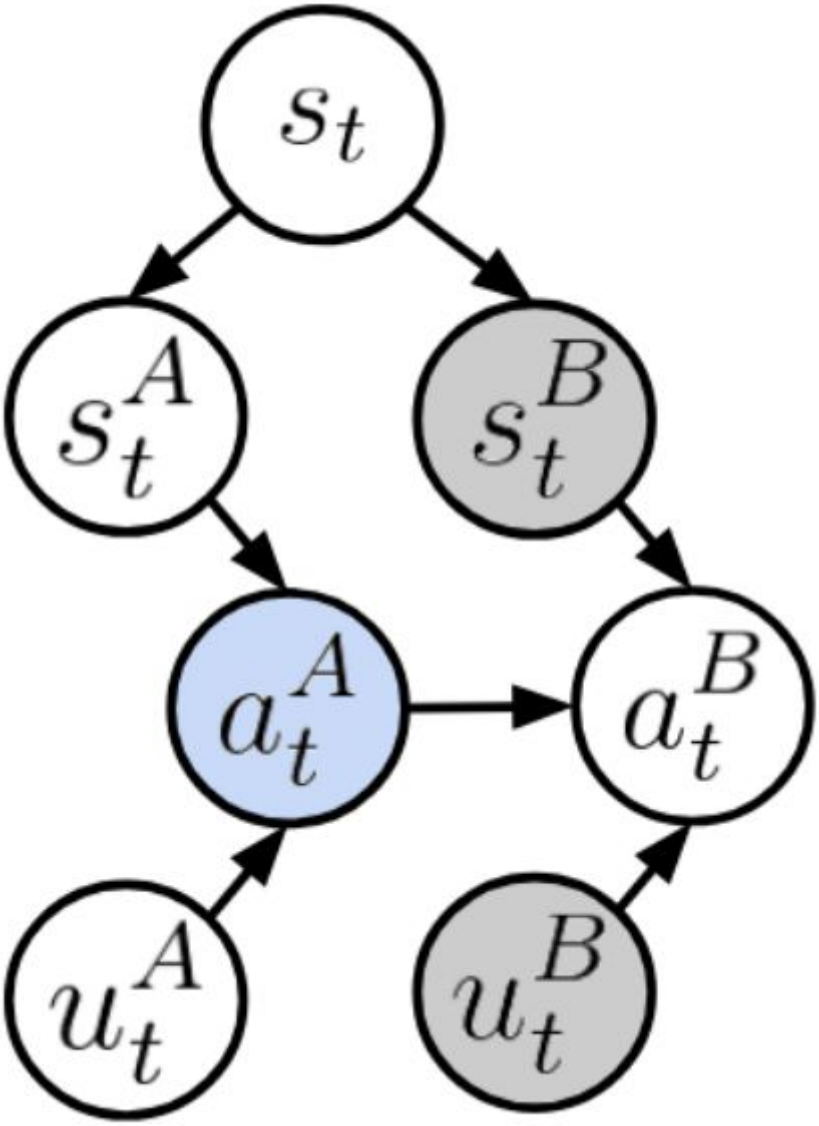


图 8. A 对 B 动作影响的因果图。以每个智能体对环境和 LSTM 状态（阴影节点）为条件，并对蓝色点）进行干预。

上式中的奖励函数与 A 和 B 之间的互信息有关：

$$I(A^B;A^A|z)=\sum_{a^A,a^B}p(a^B,a^A|z)\log\frac{p(a^B,a^A|z)}{p(a^B|z)p(a^A|z)}=\sum_{a^A}p(a^A|z)D_{\mathrm{KL}}\Big[p(a^B|a^A,z)\Big\|p(a^B|z)\Big]$$

与互信息的联系很有意思，因为单智能体 RL 经常使用的一个内在动机是赋权，它奖励智能体在其行为与环境的未来状态之间拥有较高的互信息，从而使其行为与环境的未来状态之间的互信息得到奖励。给定互信息的蒙特卡洛近似：

$$I(A^A;A^B|z)=\mathbb{E}_\tau\Big[D_{\mathrm{KL}}\big[p(A^B|A^A,z)\big\|p(A^B|z)\big]\Big|z\Big]\approx\frac{1}{N}\sum_n D_{\mathrm{KL}}\big[p(A^B|a_n^A,z)\big\|p(A^B|z)\big]$$

由此，定义社会影响奖励是智能体行为之间的互信息。

图 9 给出了一个关于不同智能体之间产生高影响力的瞬间示例。其中，紫色的影响者选择了朝向在黄色智能体的视线范围之外的苹果移动。因为影响者只有在有苹果的时候才会移

动，这就向黄色智能体发出了一个信号，那就是它的上方一定有一个它看不到的苹果。这就改变了黄色智能体的行动计划分布，从而对紫色智能体产生了影响。

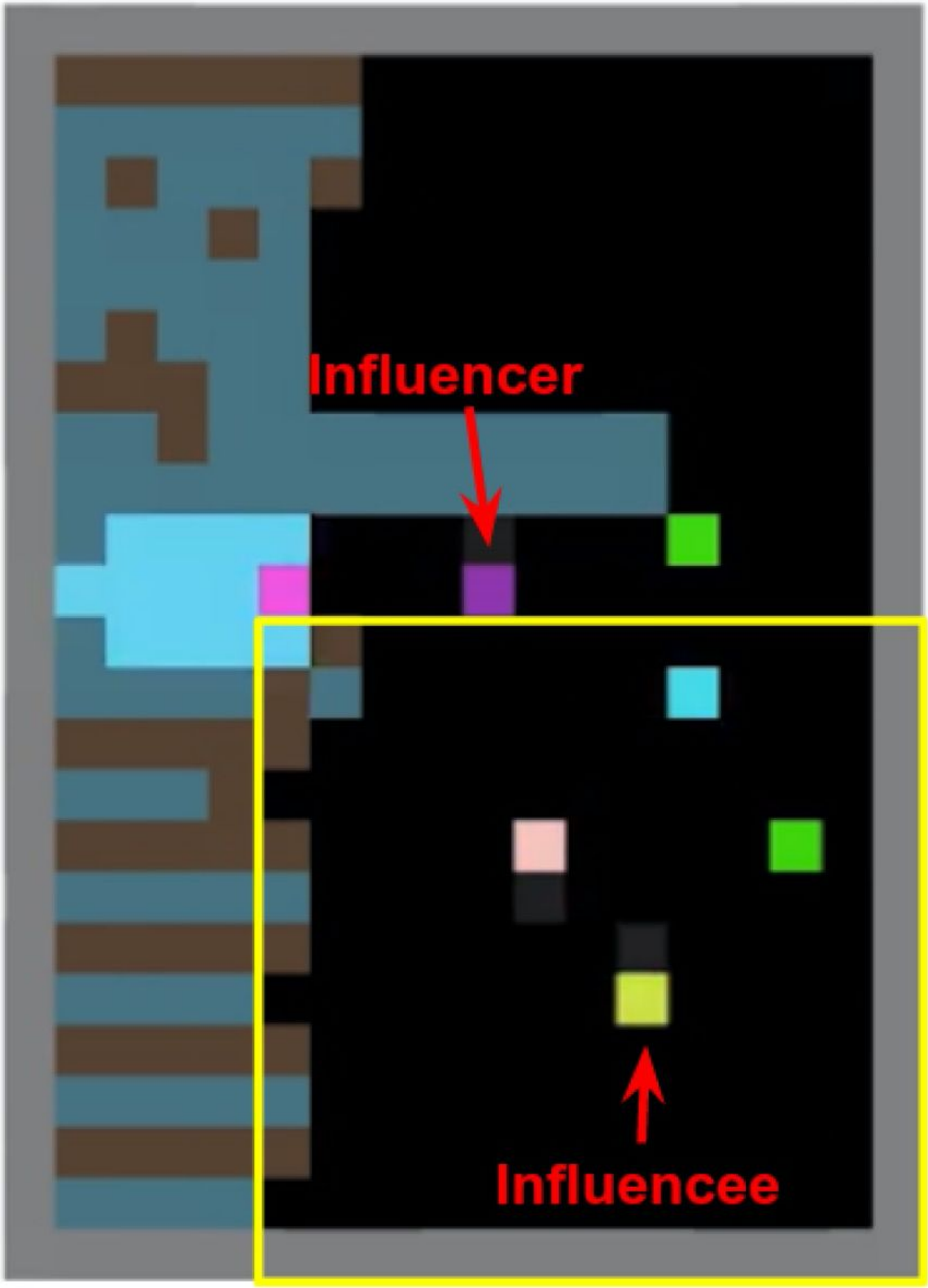


图 9. 当紫色的影响者发出信号，表示在黄色影响者的视野外 (黄色轮廓框内) 出现了一个苹果 (绿色方块)，会产生一个高影响的瞬间。

本文作者还讨论了社会影响力奖励的第二种用途：学习智能体之间的通信协议。为智能体提供了明确的沟通渠道。在每个时间点内，每个智能体选择一个离散的通信符号。这些符号被串联成 N 个智能体的组合消息向量。然后，此组合消息向量在下一个时间点中显示给所有其他智能体，如图 10 所示。社会影响力奖励包括环境奖励和沟通奖励。在图 9 示例的基础上，为了训练智能体的通信能力，在初始网络中增加了一个输出头，它可以学习一个通信策略和值函数，以确定要发出哪个符号，从而训练智能体的通信能力。

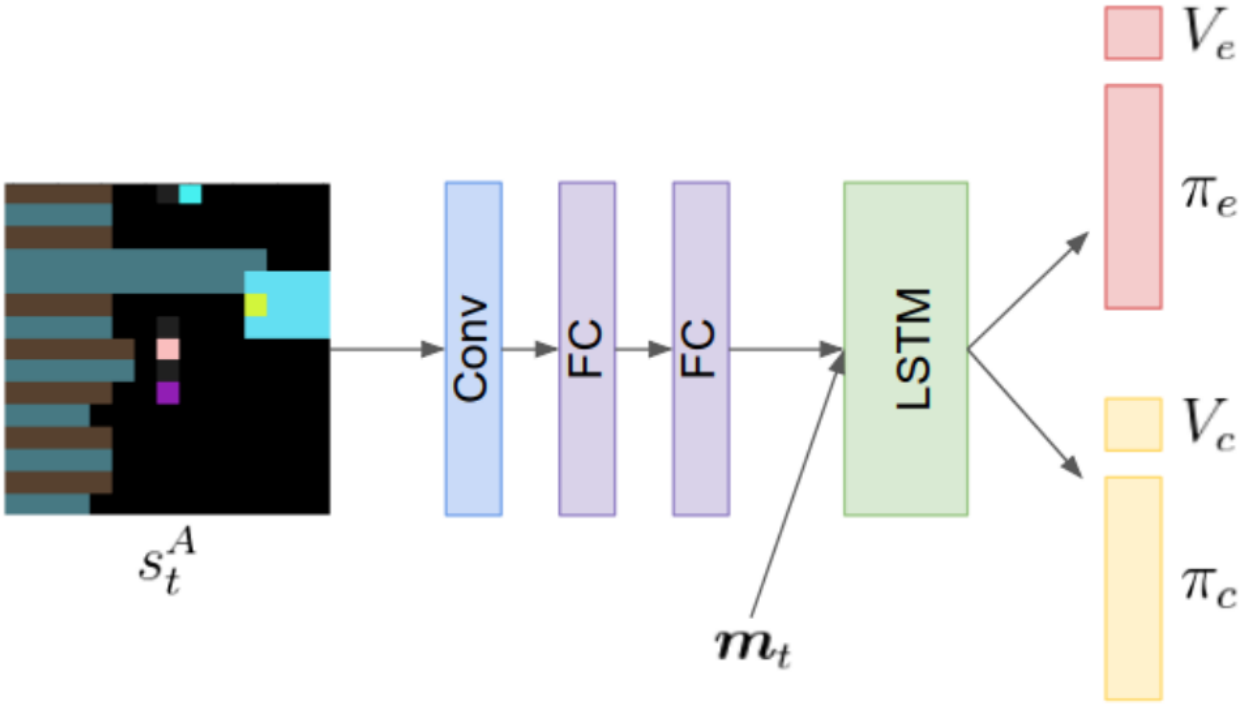


图 10. 双头通信模式，学习一个通信策略和值函数，将其他智能体的通讯信息 m_t 输入到 LSTM。

为满足 MARL 问题集中式训练框架的需求，本文通过为每个智能体训练自己的内部其它智能体模型（Model of Other Agents，MOA）来完成独立训练。MOA 与智能体的卷积层连接的第二组完全连接的 LSTM 层组成，同时训练 MOA 以预测所有其他智能体的下一步行动，并且是以智能体的自我为中心的状态（视线范围内）在给定之前行动的前提下进行预测，见图 11。使用观察到的动作轨迹和交叉熵损失来训练 MOA。MOA 的因果分析过程具体见图 12，只有当智能体试图影响的智能体在其视域范围内时，才会给予社会影响力奖励。

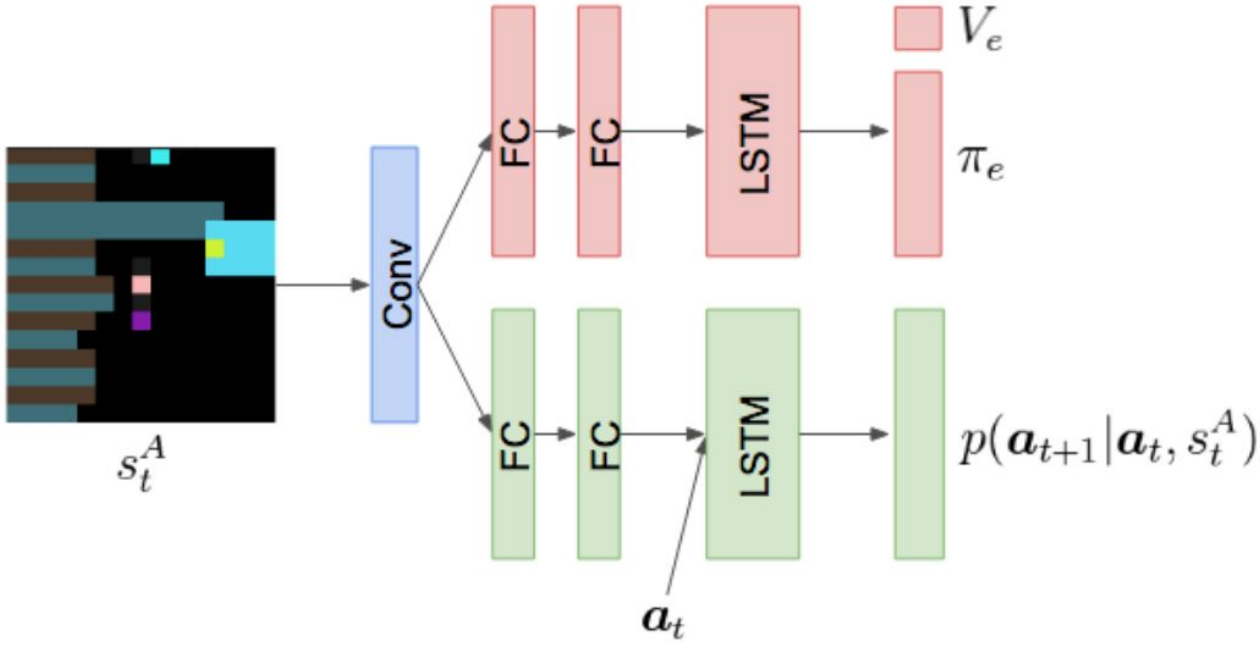

V_e
π_e
FC
FC
LSTM
Conv
FC
FC
LSTM
$p(\boldsymbol{a}_{t+1}|\boldsymbol{a}_t, s_t^A)$
s_t^A
$\boldsymbol{a}_t$

图 11. MOA 结构。

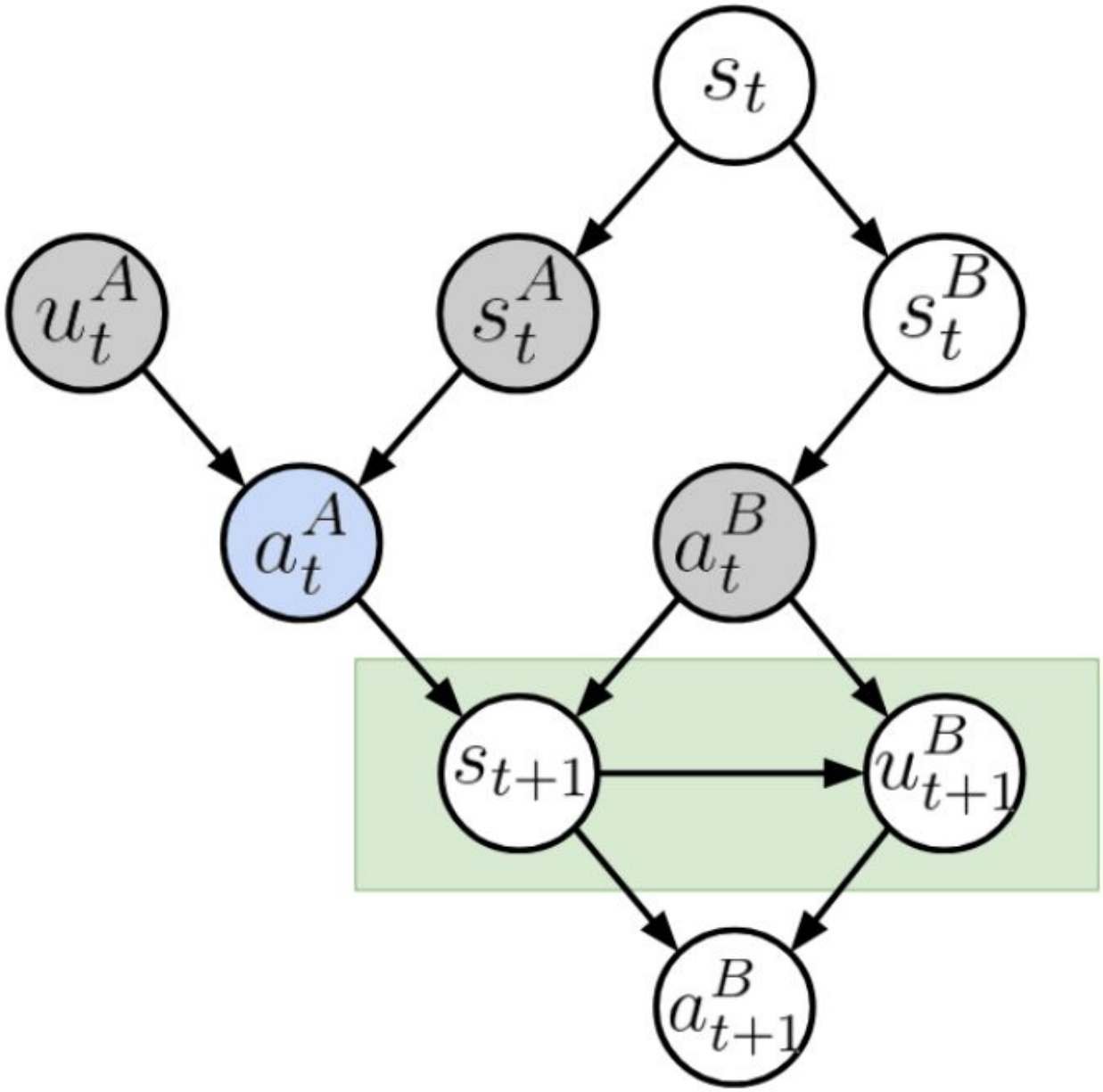

s_t
u_t^A
s_t^A
s_t^B
a_t^A
a_t^B
s_{t+1}
u_{t+1}^B
a_{t+1}^B

图 12. MOA 结构中的因果图。阴影的节点是有条件的，通过用反事实代替蓝色节点来干预蓝色节点。带绿色背景的节点必须使用 MOA 模块进行建模。

本文实验中考虑了两个具有挑战性的多智能体环境：一是公益游戏 Cleanup，另一个是悲剧性的大众游戏 Harvest，如图 13 所示。

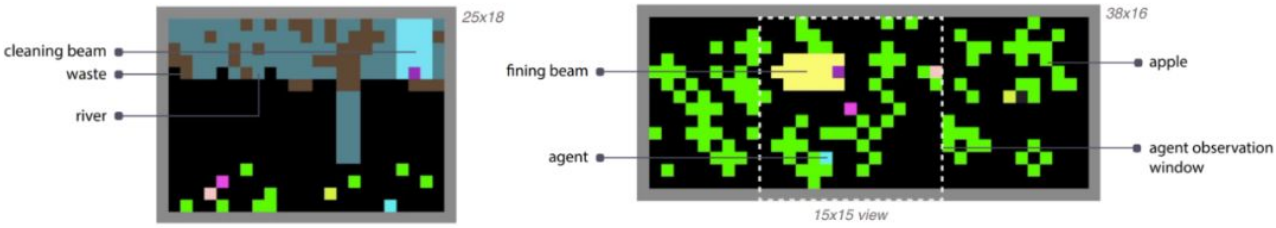
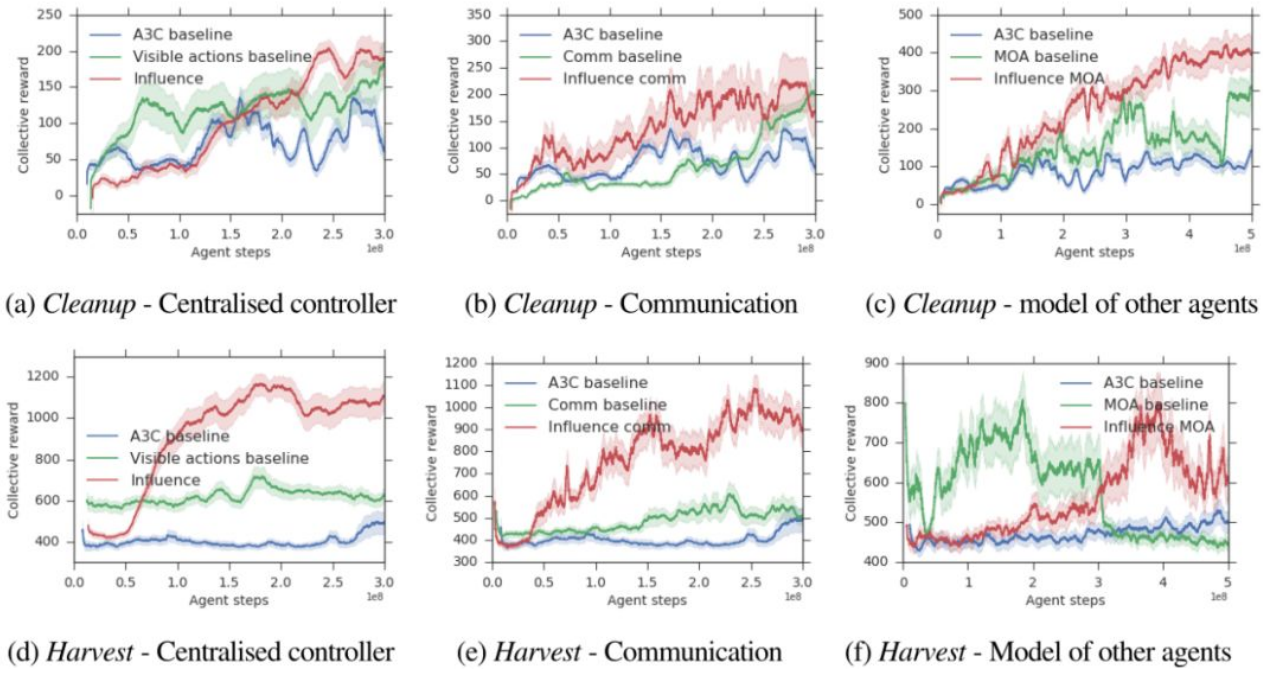


图 13. 多智能体环境图示，左：Cleanup，右：Harvest。

在不同实验条件下获得的总集体奖励如图 14。误差条显示了 5 个随机种子的 99.5% 的置信区间（CI），在 200 个智能体步骤的滑动窗口内计算。用影响奖励（红色）训练的模型明显优于基线模型和消融模型。



(a) *Cleanup* - Centralised controller (b) *Cleanup* - Communication (c) *Cleanup* - model of other agents

(d) *Harvest* - Centralised controller (e) *Harvest* - Communication (f) *Harvest* - Model of other agents

图 14. 总集体奖励。

为了分析智能体学习的交流行为，作者引入三个指标。发言者一致性：是一个归一化的分数，它评估的是一个归一化的分数，用来判断说话者智能体在采取特定动作时发出特定符号的一致性，反之亦然；以及两个瞬时协调 (IC) 的衡量标准，这两个标准衡量都是互信息。图 15 的实验结果表明，用社会影响力奖励训练出来的模型表现出更一致的沟通和更多的协调性，尤其是在影响力大的时候。

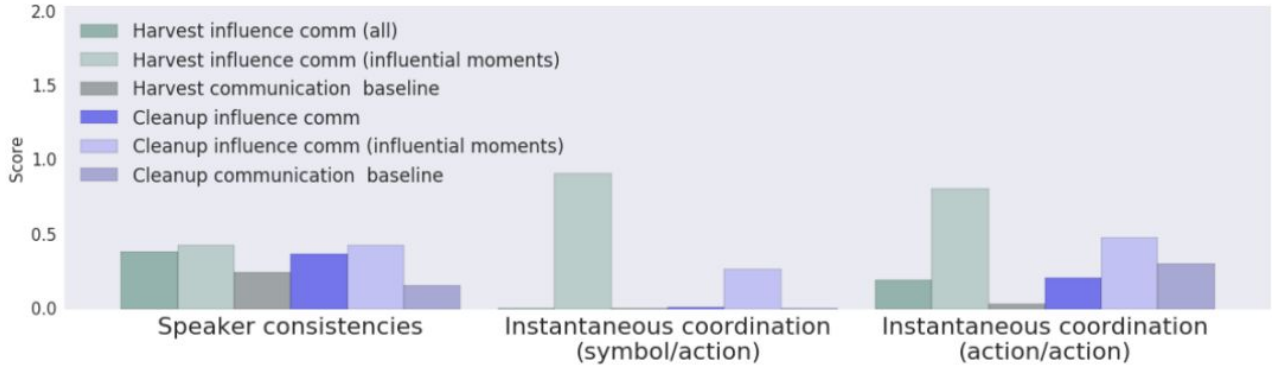


图 15. 描述学习通信协议质量的指标。

四、小结

在这篇文章中，我们讨论了内在动机及其在强化学习中的应用。在从心理学角度介绍内在动机基本概念的基础上，结合强化学习框架探讨了在机器学习方法中引入内在动机的方法。最后，详细介绍了两篇应用于机器人任务的强化学习框架，包括单智能体和多智能体两种应用场景。

正如开篇我们提到的，对于强化学习本身来说，并不严格区分内在动机、外在动机。强化学习强调的只是在与环境直接交互的同时学会学习最佳行为策略，而具体这种交互是出于内在动机还是外在动机并不影响学习的效果。只是人们普遍认为的是强化学习智能体奖励必须是外在的，因为这些奖励信号有明显的外部输入通道。研究人员改进经典的强化学习框架（如本文图 2），目的是更好的将内在动机原则融入其中。

将内在动机植入人工智能，可能会让人想起科幻小说中关于真正自主机器人的危险性的所有警告。但是，目前研究人员仍然是希望智能体拥有内在驱动的学习能力。自主性（Autonomy）正在成为自动化系统越来越常见的属性，因为它使自动化系统能够在动态的、复杂的、危险的、先验知识很少的环境中成功地长时间运行。为这些系统提供精心设计的内在动机，是一种使智能体有足够能力完成任务的研究方向。

参考文献

[1] Gianluca Baldassarre, What are Intrinsic Motivations? A Biological Perspective, http://core.ac.uk/download/pdf/37835544.pdf

[2] Andrew G. Barto, Intrinsic Motivation and Reinforcement Learning, https://eee.uci.edu/17s/68730/papers/barto-2013.pdf

[3] Tejas D. Kulkarni, Hierarchical Deep Reinforcement Learning: Integrating Temporal Abstraction and Intrinsic Motivation, http://papers.nips.cc/paper/6233-hierarchical-deep-reinforcement-learning-integrating-temporal-abstraction-and-intrinsic-motivation.pdf

[4] Hester, Todd，Stone, Peter, Intrinsically motivated model learning for developing curious robots, Artificial Intelligence, Volume 247, June 2017, Pages 170-186, http://www.cs.utexas.edu/users/pstone/Papers/bib2html-links/AIJ15-Hester.pdf

[5] Jane X Wang, Edward Hughes, et al., Evolving intrinsic motivations for altruistic behavior, ICLR 2019, https://arxiv.org/pdf/1811.05931v1.pdf

[6] Natasha Jaques, Angeliki Lazaridou, et al., Intrinsic Social Motivation via Causal Influence in Multi-Agent RL, https://arxiv.org/pdf/1810.08647v1.pdf

[7] H. V . Neto and U. Nehmzow,「Visual novelty detection with automatic scale selection,」Robotics and Autonomous Systems, vol. 55, no. 9, pp. 693–701, 2007

[8] S. Singh, R. Lewis, A. Barto, and J. Sorg,「Intrinsically motivated reinforcement learning: An evolutionary perspective,」Autonomous Mental Development, IEEE Transactions on, vol. 2, no. 2, pp. 70–82, 2010

[9] V . Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. V eness, M. G. Bellemare, A. Graves, M. Riedmiller, et al. Human-level control through deep reinforcement learning. Nature, 518(7540):529–533, 2015

[10] T. Hester, M. Quinlan, P. Stone, RTMBA: a real-time model-based reinforcement learning architecture for robot control, in: Proceedings of the 2012 IEEE International Conference on Robotics and Automation, ICRA, 2012

作者介绍：

仵冀颖，工学博士，毕业于北京交通大学，曾分别于香港中文大学和香港科技大学担任助理研究员和研究助理，现从事电子政务领域信息化新技术研究工作。主要研究方向为模式识别、计算机视觉，爱好科研，希望能保持学习、不断进步。

关于机器之心全球分析师网络 Synced Global Analyst Network

机器之心全球分析师网络是由机器之心发起的全球性人工智能专业知识共享网络。在过去的四年里，已有数百名来自全球各地的 AI 领域专业学生学者、工程专家、业务专家，利用自己的学业工作之余的闲暇时间，通过线上分享、专栏解读、知识库构建、报告发布、评测及项目咨询等形式与全球 AI 社区共享自己的研究思路、工程经验及行业洞察等专业知识，并从中获得了自身的能力成长、经验积累及职业发展。

提交申请：http://jiqizhixin.mikecrm.com/rg2RY52