多智能体强化学习中条件协调行为的可解释性

本川佳成日本东京早稻田大学计算机科学系
y.motokawa@isl.cs.waseda.ac.jp

Toshiharu Sugawara日本早稻田大学计算机科学系，东京，日本sugawara@waseda.jp

arXiv:2304.10375v1 [cs.LG] 20 Apr 2023

摘要—**我们提出了一种名为** 条件注意力后的分布式注意力演员架构 **(DA6-X) 的无模型强化学习架构，以提高条件协调行为的可解释性。其基本原理是重用显著性向量，该向量表示环境的条件状态，例如智能体的全局位置。因此，将 DA6-X 灵活性融入其策略的智能体，能够在决策过程中考虑条件状态中的额外信息，从而表现出更优异的性能。我们通过在物体收集游戏中将所提出的方法与传统方法进行比较，对其有效性进行了实验评估。通过可视化 DA6-X 的注意力权重，我们证实，智能体能够通过正确识别各种条件状态，成功学习依赖于具体情境的协调行为，从而提高智能体的可解释性，并实现更优异的性能。**

索引词—**多智能体深度强化学习、可解释强化学习、分布式系统、注意力机制、协调、合作**

I. 引言

可解释强化学习（XRL）已在学术界和工业界引起广泛关注。针对智能体可解释性的 XRL 方法有多种类别 [21]。例如，可以通过分解奖励函数 [8]，或基于积分梯度方法可视化显著性图 [25]，来解释智能体的决策过程。除 Transformer 外， [26]，开发的注意力机制也在提升决策过程的透明度方面发挥着关键作用。因此，基于注意力机制和 Transformer 的各种神经网络模型，例如视觉 *Transformer* [4] 和决策 *Transformer* [1]，已成功实现对输入信息的可视化；这些信息对于在计算机视觉乃至深度强化学习（DRL）中实现最先进的性能至关重要。

然而，尽管阐明黑箱中的协调／合作机制对于提升整个系统的生产力和鲁棒性至关重要，鲜有研究关注多智能体系统（MAS）中的 XRL。多智能体注意力评论家（ MAAC） [6], 以 MADDPG [13], 的方式融入注意力机制，展示了智能体如何利用注意力机制，有选择地关注彼此之间的合作。 [17], [18] 通过分析单个智能体的注意力机制，提出了一种建立智能体协调行为可解释性的方法并通过实证阐明，智能体会有选择地识别值得与之协调的其他智能体。这些发现有望提高对个体所学习行为的可解释性，并提升整个系统的效率。同时，多智能体系统中的智能体必须根据包括其他智能体及其行为在内的条件和情境灵活地行动。智能体可以使用上述方法学习此类条件行为；然而，这些方法不足以确保学习结果的可解释性能够使智能体习得预期的条件行为。

为应对这一挑战，我们提出了分布式注意力演员架构（条件依赖注意力（DA6-X）。该模型架构通过重复利用显著性向量 [18] ，增强对所学习的条件协调行为的可解释性，适用于多智能体深度强化学习（MADRL）。DA6-X 是一种序贯框架，它将条件／情境识别模块与任意深度强化学习头（对应“X”）相结合，以输出所学习到的行为。与先前模型不同 [17], [18],，DA6-X 使用两层数据：条件状态（条件／情境）和局部观测。DA6-X 首先在前置条件模块（CM）中使用显著性向量表示条件状态。随后，显著性向量与局部观测一同输入后置局部变换器编码器，使 DA6-X 能够在各种情境下进行全面学习。此外，采用 DA6-X 的智能体（DA6-X）不仅能以更有条理的方式学习条件依赖动作，还能提供信息，帮助理解其动作的依据与理由。通过从 DA6-X 的局部变换器编码器中提取注意力权重，可以生成注意力热力图，以确定输入数据中描述条件因素的部分，以及这些因素如何影响对两层数据（尤其是局部观测）的关注强度。即使提供相同的局部观测，这些信息也有助于我们理解条件状态如何影响智能体的行为。

我们的主要贡献总结如下：1) DA6-X 是一种新颖的多智能体深度强化学习架构，通过复用显著性向量来解释依赖情境的/条件协调行为。2) 在提供透明性的同时，DA6-X 可以轻松集成到任何深度强化学习算法中（因此，*X* 包含在其中。

（该名称）在单智能体或多智能体深度强化学习系统中。3）具有各种条件状态的 DA6-X智能体在以物体收集游戏 为例的实验中，取得了优于若干现有算法的性能。4）通过层次化分析输入信息，利用注意力机制对改进后的可解释协调行为进行了定性展示。

II. 相关工作

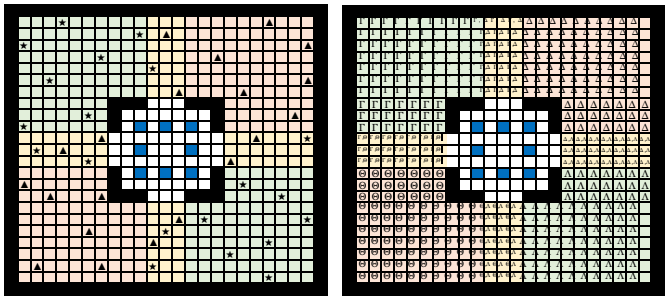
XRL 中基于注意力的方法： 除了开展视觉解释方面的研究，例如基于特征的 [7],基于嵌入的 [16], 基于扰动的 [29], 和基于梯度的方法 [25], 在模型中引入注意力机制也是 XRL 中最受欢迎的方法之一 [27]。最近，*DA3-X* [18] 被提出，作为 *MAT-DQN* [17] 的扩展，用于展示去中心化智能体如何基于注意力机制，通过突出相关任务、其他智能体以及局部观测中的噪声所产生的影响来建立协调。

然而，这些先前的研究仅通过关注性能提升，以聚合方式表示智能体输入中具有影响力的片段，而忽略了输入内部的依赖关系；换言之，它们并未确定输入中引导智能体将注意力转向其他片段部分的片段。相比之下，我们提出的方法明确地将输入信息划分为条件片段（例如观测智能体和其他智能体的全局位置）以及基线片段（例如局部可见区域），以识别条件片段中影响智能体在其局部观测内关注重点的部分，并促成条件化、情境相关且协调的行为。

多智能体深度强化学习中的注意力机制： 各种多智能体深度强化学习模型都采用了注意力机制。多项研究 [2], [6] 将注意力机制用作集中式通信处理器，以高效处理多智能体深度强化学习中智能体之间的编码消息。在多智能体系统中引入注意力机制，也有助于解决受限问题 [20], ，例如近似智能体的潜在行为 [11]以及轨迹预测 [10]。具体而言， [2] 提出了多焦点注意力网络，该网络通过多个并行注意力机制，帮助智能体在网格环境中关注重要的感官输入信息。 [28]研究了增强智能体高效适应复杂环境的能力.....

复杂环境（Box-World 和 星际争霸 II）需要通过注意力机制对结构化表示进行关系推理。 [9] 提出了联合注意力，该机制聚合其他所有智能体的注意力图，并证明了其在多智能体协作环境中的成本效益。

早期关于多智能体深度强化学习的研究主要着眼于在集中训练、分散执行 (CTDE)方法中利用集中式注意力提升智能体的性能；然而，由于策略更新的方差较小，完全分散式方法通常更为稳健，也更适用于现实领域 [14]。此外，利用注意力热图对协调进行定性分析



(a) 物体生成区域（绿色表示物体 ★，红色表示物体 ▲，米色表示物体 ★、▲）。(b) 任务执行区域，划分为四个不同的区域，分别标记为 Γ、 Δ、 Θ 和 Λ。

图 1：网格环境示例。

尚未得到深入研究。我们的目标是基于分散式注意力热力图，对协作智能体进行具体的行为分析，以阐明它们如何根据条件状态和局部观测采取不同策略与协作结构，从而提升可解释性。

III. 预备知识

分散式部分可观测马尔可夫决策过程（Dec-POMDP）：本研究聚焦于由 N 个智能体组成的去中心化部分可观测马尔可夫决策过程（Dec-POMDP） [22], 表示为元组 $\langle \mathcal{I}, \mathcal{S}, \{\mathcal{A}_i\}, p_T, \{r_i\}, \{\Omega_i\}, \mathcal{O}, H \rangle$。 $\mathcal{I} = \{1, \ldots, N\}$ 表示智能体集合； $\mathcal{S}$ 表示有限状态集合； $\mathcal{A}_i$ 表示 $i \in \mathcal{I}$的有限个体动作空间。假设 $a \in \mathcal{A}$ 和 $s, s' \in \mathcal{S}$； $p_T(s'|s,a)$表示转移概率； $r_i(s,a)$表示由 $i \in \mathcal{I}$获得的奖励（$\in \mathbb{R}$）； Ω_i 表示 $i \in \mathcal{I}$的有限观测集合； $\mathcal{O}(o|s,a)$表示 $o \in \Omega$的转移概率； H 表示该过程的时间范围。在 Dec-POMDP 中，智能体通过更新各自的策略π_i，旨在最大化折扣累积奖励$R_i = \sum_{t=0}^{H} \gamma^t r_i(s,a)$，其中 γ 表示折扣因子（$0 \leq \gamma < 1$）。

问题设定： 在物体收集游戏中，智能体在大小为 $G_X \times G_Y$的网格环境中尽可能多地收集物体，如图 1 所示。每个回合开始时，智能体被放置在图 1 中蓝色单元格所示的初始位置，并开始探索。在每一步中，智能体将观测到的局部环境编码为 $R_{\mathrm{X}} \times R_{\mathrm{Y}}$ 矩阵的N_{C} 个通道（其中 $R_{\mathrm{X}} \times R_{\mathrm{Y}}$ 是观测矩阵 $\in \mathbb{R}^{R_{\mathrm{X}} \times R_{\mathrm{Y}}}$的大小），并将形状为 $N_{\mathrm{C}} \times R_{\mathrm{X}} \times R_{\mathrm{Y}}$ 的局部观测输入各自的网络。然后，它们选择合适的动作$a_i \in \mathcal{A}_i = \{\text{上}, \text{下}, \text{右}, \text{左}\}$，其中每个元素都描述了智能体下一步的移动方向。

智能体 $\forall i$ 会根据下一步的状态获得奖励 r_i ，即，如果它移动到与物体相同的位置，就会获得正奖励 $r_{\mathrm{o}} > 0$ ；如果与其他智能体或墙壁发生碰撞，就会获得负奖励 $r_{\mathrm{c}} < 0$ ；否则，它会获得 $r_i = 0$。

多头注意力： [26] 提出了自注意力机制，用于计算序列中的相似度，如

$$\text{Attention}(Q,K,V)=\text{Softmax}(\frac{Q\cdot K^T}{\sqrt{d}})V,$$

其中，Q、K 和 V 分别表示查询、键和值 矩阵，d 表示查询/键的维度。多头注意力 （MHA）通过在 h 个并行注意力头中计算自注意力来确定，如下式所示：

$$\begin{aligned}\text{MHA}(Q,K,V)&=\text{Concat}(head_1,\ldots,head_h)W^O\\ head_l&=\text{Attention}(Q\cdot W_l^Q,K\cdot W_l^K,V\cdot W_l^V),\end{aligned}$$

其中，W_l^Q、W_l^K、W_l^V、W^O是注意力头 头$_l$ ($1\le l\le h$)的投影参数矩阵。更多详情请参阅原论文 [26]。

DA3-X： DA3-X [18] 是一种包含注意力机制的神经网络模型，如图 2a 所示。DA3-X 的重要意义在于能够提高分布式多智能体系统中智能体的可解释性。由于多智能体系统中的智能体需要与其他智能体建立协调关系，因此，阐明其协同行为如何通过黑箱决策过程产生至关重要。此前的一些研究致力于提高集中式多智能体系统中协调过程的透明度，但大多未考虑在现实应用中更具可行性的分布式系统。通过将 DA3-X 作为基线方法，我们分析了 DA3-X 中的注意力权重，验证了其可解释性，以及选择性地区分观测中重要片段的能力。其灵活的网络结构也值得关注，因为该结构能够应用任意强化学习方法（如 DA3-DQN、DA3-DDPG 等）。此外，由于 DA3-X 假设系统采用分布式架构，且不依赖其他智能体的模型，因此我们验证了其在智能体数量方面具有无尺度特性。

IV. 提出的方法：DA6-X

A. 神经网络架构

DA6-X 的核心思想是，在 CM 和局部变换器编码器中始终使用相同的显著性向量，如图 2 所示。CM 和局部变换器编码器分别发挥识别环境条件以及通过注意力机制对局部信息进行加权的作用。DA6-X 是 DA3-X 的扩展；然而，两者的区别在于，DA6-X 引入了 CM，以解释情境依赖行为的依据，并便于管理员理解。我们在 CM 中引入了独立的 Transformer 编码器，以有效生成显著性向量。通过 CM 中 Transformer 编码器的注意力机制，DA6-X 能够灵活地识别条件状态，并将其表示在显著性向量中，最终该向量会在局部 Transformer 编码器中重复使用。由于还可以通过分析 CM 中 Transformer 编码器的注意力权重来解释其作用，因此我们能够明确 CM 的功能。与所展示的注意力分析类似，在 V-C节，还可以类似地进一步分析 DA6-X 如何处理条件模块中的条件状态。DA6-X 的数学过程如下所述：

$$\begin{aligned}g_{m,0}&=[\boldsymbol{u}_m^{\text{sal}};x_m^1E_m^{\text{cond}};x_m^2E_m^{\text{cond}};\ldots;x_m^{I_m}E_m^{\text{cond}}]+P_m^{\text{cond}}\\ g_{m,l}&=\text{TEL}_m^{\text{cond}}(g_{m,l-1}),\qquad l=1,\ldots,L_m\\ \boldsymbol{v}^{\text{sal}}&=\text{VectorIntegration}(g_{1,L_1}^0,\ldots,g_{m,L_m}^0,\ldots,g_{M,L_M}^0)\end{aligned}$$

上述方程表示了条件模块（CM）中条件状态的流动计算过程。假设总共有 M 个子模块，第 m 个（$1\le m\le M$）经过补丁处理的条件状态矩阵 $x_m\in\mathbb{R}^{I_m\times(P_m^2N_m)}$通过参数 $E_m^{\text{cond}}\in\mathbb{R}^{(P_m^2N_m)\times C_m}$进行嵌入，随后与显著性向量（可训练参数） $\boldsymbol{u}_m^{\text{sal}}\in\mathbb{R}^{C_m}$进行拼接，其中，$I_m$ 表示补丁操作后第 m 个条件状态的长度；P_m 表示第 m 个补丁大小；N_m 表示补丁操作前条件状态通道的数量；C_m 表示第 m 个显著性向量的长度。需要注意的是，该计算对应于图 2b 所示的状态嵌入器中的处理过程。在附加位置嵌入$P^{\text{cond}}\in\mathbb{R}^{(I_m+1)\times C_m}$ 后，第 m 个输入被送入第 m 个变换器编码器层（$\text{TEL}_m^{\text{cond}}$），如图 2b 中的绿色方框所示。经过 L_m 次循环后，所有子模块中的显著性向量 $g_{1,L_1},\ldots,g_{m,L_m},\ldots,g_{M,L_M}^0$通过向量整合 过程进行聚合，从而生成长度为 C_{L} 的最终显著性向量 $\boldsymbol{v}^{\text{sal}}\in\mathbb{R}^{C_{\text{L}}}$。

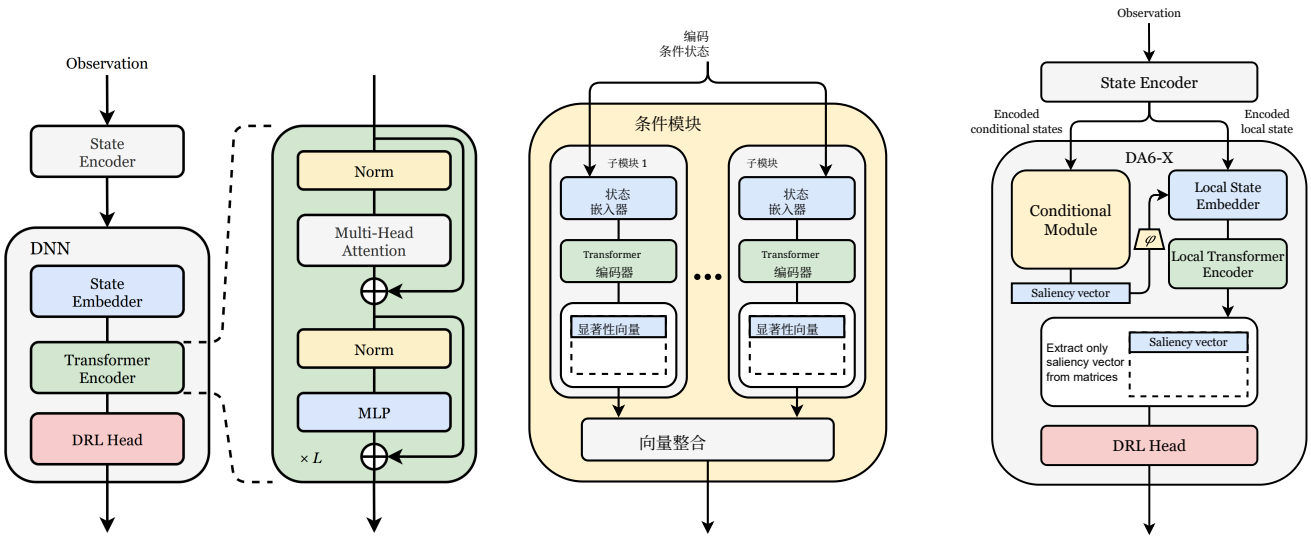
CM之后的计算过程如下：

$$\begin{aligned}h_0&=[\varphi(\boldsymbol{v}^{\text{sal}});y^1E^{\text{local}};y^2E^{\text{local}};\ldots;y^{I_{\text{L}}}E^{\text{local}}]+P^{\text{local}}\\ h_l&=\text{TEL}^{\text{local}}(h_{l-1}),\qquad l=1,\ldots,L_{\text{L}}\\ Q&=\text{DRLHead}(h_{L_{\text{L}}}^0)\end{aligned}$$

经过记为 $\varphi(\cdot)$ 的投影后，显著性向量 $\boldsymbol{v}^{\text{sal}}$会直接输入局部状态嵌入器，如图 2c 所示。与前一过程类似，经过补丁处理的局部观测 $y\in\mathbb{R}^{I_{\text{L}}\times(P^2{}_{\text{L}}N_{\text{L}})}$通过参数 $E_{\text{L}}\in\mathbb{R}^{(P^2{}_{\text{L}}N_{\text{L}})\times C_{\text{L}}}$进行嵌入，并与投影后的显著性向量 $\varphi(\boldsymbol{v}^{\text{sal}})$ 拼接，其中 I_{L} 是补丁处理后局部观测的长度；P_{L} 是局部补丁大小；N_{L} 是经过补丁处理的局部观测通道数。通过复用 CM 的显著性向量 $\boldsymbol{v}^{\text{sal}}$，DA6-X智能体在考虑条件状态后能够构建灵活的策略。类似地，在附加位置嵌入$P^{\text{local}}\in\mathbb{R}^{(I_{\text{L}}+1)\times C_{\text{L}}}$ 后，输入会被传入局部变换器编码器层（$\text{TEL}^{\text{local}}$） L_{L} 次。最后，只有显著性向量 $h_{L_{\text{L}}}^0$ 会输入*DRLHead*，如图 2c 所示。

B. DA6-X 的优势

由于注意力机制具备选择能力，智能体在探索中效率更高，鲁棒协调能力也优于系统未引入注意力机制时的表现。因此，采用注意力机制的 DA6-X 智能体也有望超越传统的多智能体深度强化学习或深度强化学习算法，效果与先前方法类似 [17], [18]。此外，



(a) DA3-X [18]。 (b) 条件模块。 (c) DA6-X

图 2：基线方法与所提方法的架构。

DA6-X 方程中 DRLHead 的输出（记为 Q ）可以是 Q 值，也可以是随机策略，具体取决于所采用的深度强化学习算法，例如 DDPG [12],PPO [23],或 RAINBOW [5]。这一特性也使 DA6-X 在现实应用中具有更高的通用性。

与 DA3-X [18] 相比，在 DA6-X 中复用显著性向量不仅有助于提升性能，还有助于提高可解释性。如前所述，DA6-X智能体通过复用显著性向量，可以根据任意数量的可用条件状态，灵活改变其对局部观测中特定信息的加权策略。因此，可以利用局部观测中的注意力热图来解释DA6-X智能体的复杂行为。相比之下，DA3-X智能体将聚合后的条件状态直接输入黑盒 DRLHead，因此其局部观测的可解释性相对有限。

我们提出的方法 DA6-X 与 DA3-X 的主要区别在于条件模块（CM）。事实上，如图 2a 所示，DA3-X 可以视为不含 CM ($M = 0$) 的 DA6-X。通过在后续的局部变换器编码器中重复使用 CM 的显著性向量，可以看出，智能体在考虑 CM 中的条件状态（例如其全局位置（G位置））后，如何观察局部观测是可以解释的。因此，由于缺少 CM，只要局部观测相同，无论条件状态如何变化，DA3-X 的行为方式始终相同，其可解释性也保持一致。相比之下，DA6-X 在理解条件状态后，可以灵活调整获取局部观测的方式，因此即使获得相同的局部观测，也可能根据条件状态采取不同的行为。由于重复使用了显著性向量，DA6-X 的可解释性优于 DA3-X。

V. 实验与结果

A. 实验设置

我们的实验在环境 $\mathcal{E}$ 中使用物体收集游戏进行，如图 1 所示，其中（G_X，G_Y）=（25，25），每种颜色的格子代表一种实体：黑色表示墙壁，蓝色表示智能体的初始位置，白色表示空区域，绿色、红色和米色表示物体生成区域。在 $\mathcal{E}$ 中生成了两类物体：绿色区域中为 ★ 类物体，红色区域中为 ▲ 类物体，米色区域中则两类物体都会生成。每类物体的数量设为 20 个（总计 40 个物体）。如图 1b 所示，环境被划分为四个不同的区域（Γ、 Δ、 Θ 和 Λ）。智能体每收集一个物体，便会在生成区域内随机位置生成一个相同类型的物体。

8 个智能体（$N = 8$）被放置在蓝色格子中，并学习各自的策略 π_i，在 5，000 个回合内尽可能多地收集物体且避免碰撞，其中每个回合的长度为 $H = 200$ 步。这些智能体分为四种类型，如表I所示。表中列出了智能体类型、智能体数量、指定的物体类型以及收集区域。例如，有两个A类智能体，它们在对角区域收集 ★物体（图1b中的Γ 和 Λ ）；还有两个C类 智能体，它们只在上半区域收集 ★ 和 ▲ 物体（图1b中的Γ 和 Δ ）。换言之，协作智能体在另一个区域中可能并不协作。奖励机制设定为 $r_o = 1$ 和 $r_c = -1$，其中只有当智能体在表I指定的区域内收集到分配给它的物体类型时，才会获得 r_o 。

表I：智能体-任务规格。

智能体	Num	符号	区域
A 型	2	★	Γ, Λ
B 型	2	▲	Δ, Θ
C 型	2	★, ▲	Γ, Δ
D 型	2	★, ▲	Θ, Λ

本实验旨在分析局部Transformer编码器中的注意力权重，以探究DA6-X智能体在遇到各种条件状态时如何建立情境依赖协调，以及这些条件状态如何影响其行为决策。我们考察了八个*DA6-DQN* 智能体和*DA6-IQN* 智能体，它们的DRLHead中分别安装了多层感知机和隐式分位数网络 (IQN) [3] 。作为基线方法，我们训练了八个标准DQN智能体、隐式分位数网络智能体、DA3-DQN智能体和DA3-IQN智能体。除了输入局部状态编码器的局部观测之外，还向条件模块（CM）提供两类额外输入，以构成环境中的条件状态：观测智能体的全局位置（*G*位置）和所有物体的位置（*O*位置）。每个智能体通过局部视图 [15]获取其局部观测（$\{R_{\mathrm{X}}, R_{\mathrm{Y}}\} = \{7, 7\}$）。基线智能体通过相对视图 [15], 观测*G*位置 和*O*位置，而DA6-X智能体则通过合并视图 [24]获取条件状态。每种条件下的实验均重复三次，使用了约100 GPU小时的 *NVIDIAGeForceRTX3090*。

B. 性能比较

图3展示了每种强化学习方法在5,000个回合中的平均回合奖励。图3中各线条的颜色表示输入智能体的条件状态：蓝色表示同时提供*G*位置 和*O*位置，红色表示提供*G*位置，黑色表示未提供条件状态。表II列出了每种学习算法在各条件状态输入条件模块时，最后100个回合的平均回合奖励、收集的物体数量，以及智能体与墙壁之间发生的碰撞次数。总体而言，无论采用哪种深度强化学习方法，在决策过程中可用的输入信息越多，学习性能就越好；除深度 Q 网络和隐式分位数网络智能体外，其他方法也呈现出类似现象。

当*G*位置 可用时：根据图3和表II，提供观测智能体的位置（*G*位置）后，DA3-X和DA6-X智能体收集到的物体更多。输入*G*位置 的DA3-深度 Q 网络和DA3-IQN智能体比未向条件模块提供输入（None）的DA3-深度 Q 网络和DA3-IQN智能体分别多收集8.52个（2.12%）和8.98个（2.19%）物体；而输入*G*位置的DA6-DQN和DA6-IQN智能体分别多收集54.65个（13.58%）和50.24个（12.24%）物体，回合奖励因此分别提高54.93和50.48分。此外，DA6-DQN和DA6-IQN智能体总共仅发生7.39次和8.01次碰撞，比输入*G*位置的DA3-DQN和DA3-IQN智能体分别少43.15%和33.80%。对于输入*G*位置 的标准深度 Q 网络和隐式分位数网络智能体，由于状态复杂度过载，其性能有所下降。

当*G*位置和*O*位置可用时：我们确认，当智能体和物体的位置（*G*位置 + *O*位置）可用时，DA3-X和DA6-X智能体取得了更高的学习性能。如表II所示，与不具备条件状态的DA3-DQN和DA3-IQN智能体相比，具备*G*位置和*O*位置的DA3-DQN和DA3-IQN智能体的回合奖励分别高出37.44分和38.99分；DA6-DQN和DA6-IQN智能体的回合奖励则分别提高61.29分和77.45分。此外，DA3-DQN与DA6-DQN智能体以及DA3-IQN与DA6-IQN智能体在收集的物体方面的差异分别为17.69（3.97%）和33.09（7.26%），这证实了DA6-X智能体借助条件状态成功构建了高效策略。与前一种情况类似，标准深度 Q 网络和隐式分位数网络智能体在*G*位置和*O*位置可用时，学习性能会下降。

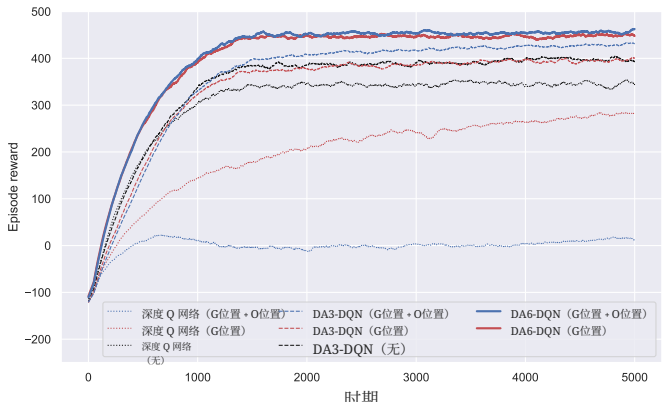
C. 注意力与协调

已确认，提供条件状态能够使DA6-X智能体学习到更高效的策略。为提升整体性能并实现成功协调，通过分析特定情境下注意力权重的强度，进一步研究了条件状态影响决策过程的方式。协调分析针对两种情况展开：仅有*G*位置可用，以及*G*位置 和*O*位置 均可用。

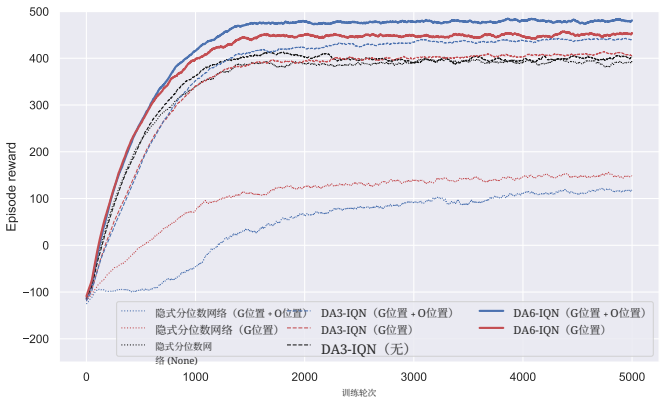
当 *G*位置 可用时： DA6-DQN智能体通过显著性向量理解其全局位置，从而提升了性能。因此，通过在两个不同位置进行测试，分析了坐标的变化方式：右上区域和左下区域（Δ 和 Θ）。首先，如图 4 所示，观察了*B*型 智能体在有两个距离不同的物体时，在*C*型 智能体位于附近与否两种情况下的行为。每幅图分别展示了*B*型 智能体的全局位置（左上）、局部观测（右上），以及 DA3-DQN（左下）和 DA6-DQN（右下）对应的注意力热图，以便比较可解释性。值得注意的是，DA3-DQN 的注意力热力图由图 2a 中不含条件状态的 Transformer 编码器的注意力权重生成，而 DA6-DQN 的注意力热力图则由图 2c 中的局部变换器编码器生成。全局位置图（图 4）中的绿色节点表示观测智能体本身，绿色方框表示可视区域（$\{R_{\mathrm{X}}, R_{\mathrm{Y}}\} = \{7, 7\}$）。观测图中的颜色表示环境中的实体：白色表示墙壁，黑色表示空区域，绿色表示观测到的*B*型 智能体，黄色表示 ▲ 物体，蓝色表示*C*型 智能体。局部观测中的绿色箭头表示 DA6-DQN 决定的下一步移动方向。

根据图 4a 和 4c，*B* 型 智能体总是向右移动，靠近距离更近的物体，因为无论全局位置如何，距离更近的物体都会比距离更远的物体吸引更多注意力（智能体位于右上区域时，DA6-DQN注意力分别为 0.177 和 0.089 Δ ；位于左下区域时，DA6-DQN注意力分别为 0.229 和 0.197 Θ）。DA3-DQN 的注意力热力图也对距离更近的物体（0.238）赋予了比距离更远的物体（0.076）更高的注意力。然而，图 4a 和 4c 中的注意力值完全相同，因为 DA3-X 不考虑条件状态。

有趣的是，当*C* 型智能体位于较近的物体附近（图 4b 和图4d）时，*B* 型智能体会根据自身的全局位置表现出不同的行为。在右上区域 Δ （图 4b），*B* 型智能体向左移动



(a) 基于深度 Q 网络的算法所获得的回合奖励。



(b) 基于隐式分位数网络的算法的回合奖励。

图3：不同条件状态下的学习性能比较。

表II：定量性能比较。

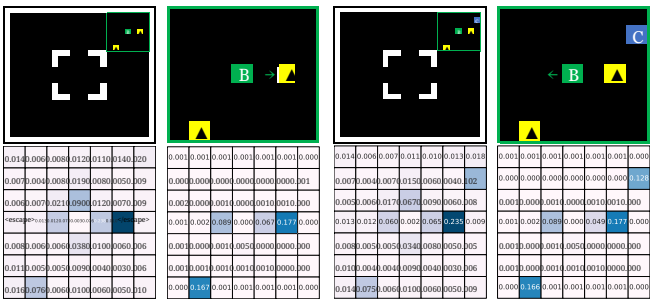
CM 的输入	模型	回合奖励	物体数量	智能体碰撞	墙壁碰撞
None	DQN	348.64 ± 35.83	360.22 ± 35.02	6.35 ± 2.50	5.23 ± 2.73
	DA3-DQN	394.76 ± 34.92	402.43 ± 35.07	3.94 ± 1.40	3.72 ± 1.29
	IQN	389.94 ± 28.43	398.85 ± 28.48	4.30 ± 1.57	4.61 ± 1.71
	DA3-IQN	402.16 ± 30.44	410.41 ± 30.43	4.10 ± 1.33	4.15 ± 1.69
*G*位置	DQN	282.21 ± 22.75	305.59 ± 21.26	14.24 ± 6.71	9.14 ± 4.32
	DA3-DQN	397.94 ± 25.43	410.95 ± 25.82	7.62 ± 3.68	5.38 ± 3.32
	DA6-DQN	**449.69 ± 24.96**	**457.08 ± 25.13**	**3.84 ± 1.21**	**3.55 ± 1.37**
	IQN	147.77 ± 27.12	323.51 ± 19.75	13.78 ± 7.45	161.97 ± 13.42
	DA3-IQN	407.29 ± 25.69	419.39 ± 24.18	6.55 ± 2.85	5.55 ± 5.23
	DA6-IQN	**452.64 ± 25.91**	**460.65 ± 25.73**	**4.09 ± 1.40**	**3.92 ± 1.68**
*G*位置 +*O*位置	DQN	14.60 ± 21.32	73.50 ± 10.90	39.36 ± 11.93	19.54 ± 6.54
	DA3-DQN	432.20 ± 22.96	446.11 ± 22.33	8.46 ± 3.36	5.45 ± 2.72
	DA6-DQN	**456.05 ± 24.28**	**463.81 ± 24.37**	**4.04 ± 1.39**	**3.72 ± 1.10**
	IQN	117.73 ± 29.48	342.13 ± 15.20	19.01 ± 10.28	205.40 ± 19.35
	DA3-IQN	441.15 ± 19.98	455.87 ± 18.72	8.37 ± 3.73	6.34 ± 4.40
	DA6-IQN	**479.61 ± 20.30**	**488.96 ± 19.76**	**5.30 ± 1.48**	**4.05 ± 1.47**

，让出位置给*C* 型智能体，从而接近较远处的物体；此时 DA6-DQN 的注意力权重为 0.128。*B* 型智能体理解到*C* 型智能体正瞄准同一目标，并认为值得通过协调来避免竞争。相比之下，在左下区域 Θ（图4d），*B* 型智能体分配给*C* 型智能体的 DA6-DQN 注意力权重仅为 0.106，并向右移动以接近较近的物体。*C* 型智能体无法收集 Θ 中的物体，因为根据表I，该物体位于其分配的任务区域之外。通过训练经验，*B* 型智能体成功学会了在左下区域 Θ 与*C* 型智能体建立协调并不值得。不同位置下，*B* 型智能体的这一策略可从其对*C* 型智能体的注意力权重降低得到验证（DA6-DQN 的注意力权重从 0.128 降至 0.106）。此外，DA3-X 智能体的条件状态如何影响协同行为，其行为分析变得非常复杂，因为 DA3-DQN 的注意力热力图不会随全局位置而变化。

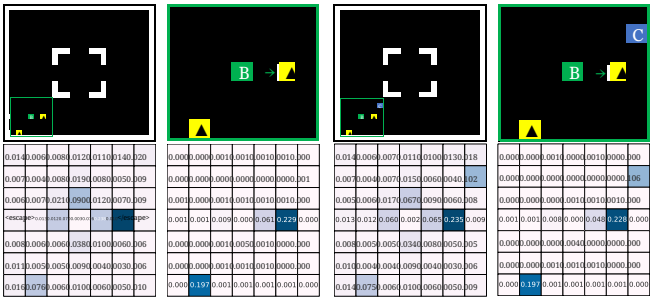
当*G*位置和*O*位置可用时：根据图3根据表II，当提供全局位置和物体的位置时，DA6-X智能体取得了更好的性能。与前述情况类似，我们分析了 DA3-DQN 和 DA6-DQN 的注意力热图，以比较并理解性能提升。在本研究中，在*B*型智能体周围相同距离处放置了两个 ▲ 物体，并将*C*型智能体放置在附近，同时将七个物体置于*B*型智能体的可见范围之外，以此分析其行为，如图5所示。图5中的颜色表示与图4相同的角色。

当*B* 型智能体在 Δ中观察到两个距离相同的物体时，它会朝左侧物体移动。左侧物体的注意力权重（DA3-DQN 为 0.177，DA6-DQN注意力为 0.169）高于右侧物体（DA3-DQN 为 0.169，DA6-DQN注意力为 0.130），如图5a所示。

当*C*型智能体位于可见范围内时，也测试了相同的情况（图5b）。在这种情况下，*B*型智能体对*C*型智能体的关注度相对较低（DA3-DQN为0.100，DA6-DQN注意力为0.062）

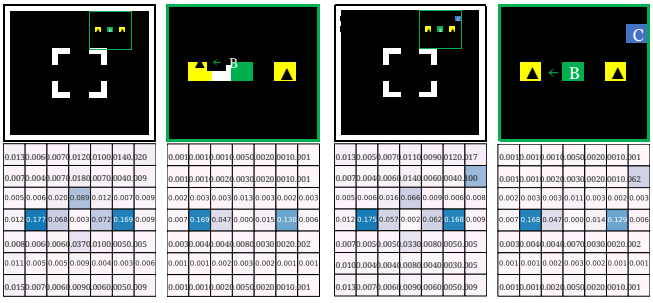


(a) 在 Δ 中单独。 (b) 在 Δ中与C 型智能体协作。

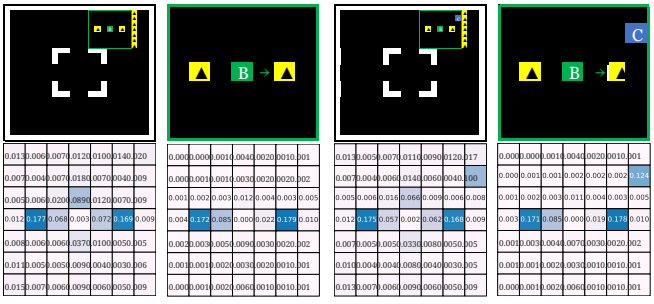


(c) 在 Θ中单独。 (d) 在 Θ中与C 型智能体协作。

图 4：使用G位置的协调研究。



(a) 单独。 (b)具有类型 E 智能体。



(c) 存在许多物体时单独行动。(d) 有对手和物体时。

图5：采用G位置 ＋ O位置的协调研究。

并移动到左侧物体。该行为是合理的，因为B 型智能体始终打算收集左侧物体，而右侧C 型智能体的存在并未影响其策略，如注意力热力图所示。

在 B型智能体右侧四个单位处添加了七个物体，以考察 DA6-IQN 智能体如何利用通过重复利用显著性向量（该向量表示G位置和O位置条件），可以实现这一点。有趣的是，如图5c所示，在考虑了可见范围之外的七个物体的位置后，B型智能体对右侧物体赋予了较高的注意力权重（DA6-DQN注意力权重为0.179），而对左侧物体赋予了较低的注意力权重（DA6-DQN注意力权重为0.172）。随后，该智能体向右移动，以接近更多物体。这验证了DA6-IQN智能体能够通过显著性向量成功理解视野之外其他物体的大致位置，并据此合理地制定高效策略。

最后，测试了B 型智能体在C 型智能体在场且右侧有七个物体时的行为。根据图5d中的注意力热图，该智能体分别将较高的注意力权重0.178和0.124赋予右侧物体和 C 型智能体。考虑到可见范围之外存在七个物体，B 型智能体分配给C 型智能体的 DA6-DQN注意力几乎增加了一倍，因为C 型智能体可能是争夺相同物体的竞争者。实际上，B 型智能体与C 型智能体一同朝右侧的物体移动，并持续锁定C 型智能体。

由于 DA6-X 在条件模块和局部Transformer中复用了显著性向量，因此我们可以分析条件状态如何影响获取局部观测的方式，并促成协调。因此，DA6-X 在提升学习性能的同时也提高了可解释性，而且还能处理多个条件状态，这一点不同于基线算法。

VI. 结论与讨论

本研究提出了 DA6-X，通过复用显著性向量来提高智能体的可解释性。我们在物体收集游戏中提供了若干条件状态并开展实验，以验证所提方法的有效性。定量结果表明，DA6-X智能体通过复用代表环境条件状态的显著性向量，成功构建了比基线方法更高效的策略。对局部变换器编码器中注意力权重生成的注意力热图进行分析后发现，提供条件状态能够改善智能体的局部观测、协调策略和决策过程，而这些此前因黑箱问题而不为人所知。

这项工作可以扩展到物体收集游戏之外的连续环境。针对不同环境中的 DA6-X 智能体开展协调研究，或许有助于更好地理解多智能体深度强化学习和 XRL，因此需求可能很高。此外，将 DA6-X 中的变换器编码器替换为其他变换器编码器（例如 [30] 提出的*TAFA*或 [19]提出的*GTrXL* ），可能会提升学习性能。

我们的方法在定量评估智能体的可解释性方面仍可能存在局限。在本研究中，我们通过注意力热图展示了DA6-X智能体观察环境的机制，并在若干情境下定性地证明了其可解释性。缺乏对可解释性的定量评估，在

自动驾驶系统等关键应用中可能是致命的，因为这些系统必须确保高安全性和算法问责制。因此，我们下一步的研究将通过注意力热图，对智能体决策过程的透明度进行定量考察。

致谢

本研究部分得到了 JSPS KAKENHI 项目资助，项目编号为 20H04245。

参考文献

[1] Lili Chen、Kevin Lu、Aravind Rajeswaran、Kimin Lee、Aditya Grover、Misha Laskin、Pieter Abbeel、Aravind Srinivas 和 Igor Mordatch。Decision Transformer：通过序列建模实现强化学习。神经信息处理系统进展，34，2021。[2] Jinyoung Choi、Beom-Jin Lee 和 Byoung-Tak Zhang。用于高效深度强化学习的多焦点注意力网络，2017。[3] Will Dabney、Georg Ostrovski、David Silver 和 Remi Munos。用于分布式强化学习的隐式分位数网络。载于 Jennifer Dy 和 Andreas Krause 编，第 *35* 届国际机器学习会议论文集，第 80 卷，机器学习研究论文集，第 1096–1105 页。PMLR，2018 年 7 月 10–15 日。[4] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly、Jakob Uszkoreit 和 Neil Houlsby。一张图像胜过 16×16 个词：用于大规模图像识别的 Transformer。载于国际学习表征会议，2021 年。[5] Matteo Hessel、Joseph Modayil、Hado van Hasselt、Tom Schaul、Georg Ostrovski、Will Dabney、Dan Horgan、Bilal Piot、Mohammad Azar 和 David Silver。Rainbow：融合深度强化学习中的改进方法。*AAAI* 人工智能会议论文集，32(1)，2018 年 4 月。[6] Shariq Iqbal 和 Fei Sha。用于多智能体强化学习的 Actor-Attention-Critic。载于 Kamalika Chaudhuri 和 Ruslan Salakhutdinov 编，第 *36* 届国际机器学习会议论文集，第 97 卷，机器学习研究论文集，第 2961–2970 页。PMLR，2019 年 6 月 9–15 日。[7] Rahul Radhakrishnan Iyer、Yuezhang Li、Huao Li、Michael Lewis、Ramitha Sundar 和 Katia P. Sycara。深度强化学习神经网络的透明度与解释。*2018* 年 *AAAI/ACM* 人工智能、伦理与社会会议论文集，2018 年。[8] Z. Juozapaitis、A. Koul、A. Fern、M. Erwig 和 F. Doshi-Velez。通过奖励分解实现可解释强化学习。载于国际人工智能联合会议论文集：可解释人工智能研讨会，2019 年。[9] Dennis Lee、Natasha Jaques、Chase Kew、Jiaxing Wu、Douglas Eck、Dale Schuurmans 和 Aleksandra Faust。用于多智能体协作与社会学习的联合注意力，2021 年。[10] Jiachen Li、Fan Yang、Masayoshi Tomizuka 和 Chiho Choi。Evolve-Graph：通过动态关系推理进行多智能体轨迹预测。载于神经信息处理系统会议（*NeurIPS*）论文集，2020 年。[11] Max Guangyu Li、Bo Jiang、Hao Zhu、Zhengping Che 和 Yan Liu。用于多智能体行为建模的生成式注意力网络。载于*AAAI*，2020 年。[12] Timothy P. Lillicrap、Jonathan J. Hunt、Alexander Pritzel、Nicolas Manfred Otto Heess、Tom Erez、Yuval Tassa、David Silver 和 Daan Wierstra。深度强化学习下的连续控制。*CoRR*，abs/1509.02971，2016 年。[13] Ryan Lowe、Yi Wu、Aviv Tamar、Jean Harb、Pieter Abbeel 和 Igor Mordatch。面向混合协作竞争环境的多智能体 Actor-Critic。载于第 *31* 届神经信息处理系统国际会议论文集，NIPS'17，第 6382–6393 页，美国纽约州雷德胡克，2017 年。Curran Associates Inc.[14] Xueguang Lyu、Yuchen Xiao、Brett Daley 和 Chris Amato。比较多智能体强化学习中的集中式与分散式评论家。*ArXiv*，abs/2102.04402，2021 年。

[15] Yuki Miyashita 和 Toshiharu Sugawara。利用多智能体深度强化学习分析协调行为结构。应用智能，51(2)：1069–1085，2021 年 2 月。资助信息：本研究部分由 JSPS KAKENHI 资助，项目编号 17KT0044。出版方版权：© 2020，作者。[16] Volodymyr Mnih、Koray Kavukcuoglu、David Silver、Andrei A. Rusu、Joel Veness、Marc G. Bellemare、Alex Graves、Martin Riedmiller、Andreas K. Fidjeland、Georg Ostrovski、Stig Petersen、Charles Beattie、Amir Sadik、Ioannis Antonoglou、Helen King、Dharshan Kumaran、Daan Wierstra、Shane Legg 和 Demis Hassabis。通过深度强化学习实现类人控制。自然，518(7540)：529–533，2015 年。[17] Yoshinari Motokawa 和 Toshiharu Sugawara。MAT-DQN：面向协调活动、可解释的多智能体深度强化学习。载于人工神经网络与机器学习——*ICANN 2021*：第 *30* 届国际人工神经网络会议，斯洛伐克布拉迪斯拉发，*2021* 年 *9* 月 *14–17* 日，论文集，第 *IV* 部分，第 556–567 页，2021 年。[18] Yoshinari Motokawa 和 Toshiharu Sugawara。面向抗噪鲁棒协调的分布式多智能体深度强化学习。载于*2022* 年国际神经网络联合会议（*IJCNN*），第 1–8 页，2022 年。[19] Emilio Parisotto、Francis Song、Jack Rae、Razvan Pascanu、Caglar Gulcehre、Siddhant Jayakumar、Max Jaderberg、Raphael Lopez Kaufman、Aidan Clark、Seb Noury 等。稳定强化学习中的 Transformer。载于国际机器学习会议，第 7487–7498 页。PMLR，2020 年。[20] P. Parnika、Raghuram Bharadwaj Diddigi、Sai Koti Reddy Danda 和 Shalabh Bhatnagar。用于多智能体约束协作强化学习的注意力 Actor-Critic 算法。载于第 *20* 届自主智能体与多智能体系统国际会议论文集，AAMAS '21，第 1616–1618 页，美国南卡罗来纳州里奇兰，2021 年。自主智能体与多智能体系统国际基金会。[21] Erika Puiutta 和 Eric M. S. P. Veith。可解释强化学习：综述。载于 Andreas Holzinger、Peter Kieseberg、A Min Tjoa 和 Edgar Weippl 编，机器学习与知识提取，第 77–95 页，瑞士，2020 年。Springer International Publishing。[22] Martin L. Puterman。马尔可夫决策过程：离散随机动态规划。John Wiley & Sons, Inc.，美国，第 1 版，1994 年。[23] John Schulman、Filip Wolski、Prafulla Dhariwal、Alec Radford 和 Oleg Klimov。近端策略优化算法。*CoRR*，abs/1707.06347，2017 年。[24] Ken Smith、Yuki Miyashita 和 Toshiharu Sugawara。通过多智能体深度 Q 学习分析部分观测协作智能体的协作结构。载于*PRIMA 2020*：多智能体系统的原理与实践：第 *23* 届国际会议，名古屋，日本，*2020* 年 *11* 月 *18–20* 日，论文集，第 150–164 页，2020 年。[25] Mukund Sundararajan、Ankur Taly 和 Qiqi Yan。深度网络的公理化归因。载于第 *34* 届国际机器学习会议论文集——第 *70* 卷，ICML'17，第 3319–3328 页。JMLR.org，2017 年。[26] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Ł ukasz Kaiser 和 Illia Polosukhin。注意力就是你所需的一切。载于 I. Guyon、U. V. Luxburg、S. Bengio、H. Wallach、R. Fergus、S. Vishwanathan 和 R. Garnett 编，神经信息处理系统进展，2017 年。[27] Wenguan Wang、Jianbing Shen、Xiankai Lu、Steven C. H. Hoi 和 Haibin Ling。关注视频物体模式理解。*IEEE* 模式分析与机器智能汇刊，43(7)：2413–2428，2021 年。[28] Vinicius Zambaldi、David Raposo、Adam Santoro、Victor Bapst、Yujia Li、Igor Babuschkin、Karl Tuyls、David Reichert、Timothy Lillicrap、Edward Lockhart、Murray Shanahan、Victoria Langston、Razvan Pascanu、Matthew Botvinick、Oriol Vinyals 和 Peter Battaglia。通过关系归纳偏置实现深度强化学习。载于国际学习表征会议，2019 年。[29] Matthew D. Zeiler 和 Rob Fergus。卷积网络的可视化与理解。载于 David Fleet、Tomas Pajdla、Bernt Schiele 和 Tinne Tuytelaars 编，计算机视觉——*ECCV 2014*，第 818–833 页，瑞士，2014 年。Springer International Publishing。[30] Qiyuan Zhang、Xiaoteng Ma、Yiqin Yang、Chenghao Li、Jun Yang、Yu Liu 和 Bin Liang。学习发现与任务相关的特征，以实现可解释强化学习。*IEEE* 机器人与自动化快报，6(4)：6601–6607，2021 年。