高技术通讯 2024 年 第34 卷 第6 期:555-566

doi:10.3772/j.issn.1002-0470.2024.06.001

基于语言类任务的概念化强化学习框架①

彭少辉②[*][**][***] 胡 杏③[*] 支 天[*]

([*]中国科学院计算技术研究所处理器芯片全国重点实验室 北京 100190)

([**]中国科学院大学 北京 100049)

([***]中科寒武纪科技股份有限公司 北京 100080)

摘 要 语言类强化学习任务可以促进强化学习策略的泛化性,其关键问题是自动化学习观测和语言描述的通用表示。现有方法往往隐式学习联合表示,不可避免地引入训练集中的虚假相关信息,进而损伤策略的泛化性和训练效率。针对这一问题,本文提出了概念化强化学习框架(CRL),其利用概念化这种从实体提取相似性生成抽象表示的认知方式,通过基于注意力机制的概念编码器和限制性损失函数显式地学习概括且抽象的概念化表示作为强化学习策略的输入。本文在常用的语言条件任务和文本游戏任务上验证了CRL 的有效性,结果显示概念化表示大幅提升了策略的训练效率(最多 70%)和泛化性能(最多 30%),并有效提升了策略的可解释性。

关键词 深度强化学习(DRL);语言类强化学习任务;文本游戏;表示学习;互信息优化

深度强化学习(deep reinforcement learning, DRL)已经成功应用在诸多领域,然而将强化学习策略迁移至相似的新环境仍然是巨大的挑战[1]。研究者们提出了语言类强化学习任务,希望借助语言作为结构化的中间渠道增强策略在不同环境间的泛化性[2]。

在语言类任务中强化学习策略的关键挑战是如何学习可捕获环境间不变因素的语言和观测的联合表示,以作为增强策略泛化性的基础。现有工作往往遵循从实例中学习表示的范式,即基于训练环境中的实例隐式学习联合表示[3-6]。例如在语言条件任务中利用特定网络结构学习以实体为中心的联合表示[3],或者在文本游戏中基于知识图谱学习游戏状态的嵌入表示[6]。这样基于实例学习的表示很难捕获环境间不变的因素将不可避免引入训练集中的虚假相关信息,导致策略的泛化性能和训练效率受到损害[7]。

概念化是人类常用的认知方法,即提取不同实例之间的相似性以形成“概念”并加以理解和利用,从而加深对问题的抽象认识,提高同类任务的效率。受到概念化认知方式启发,本文提出了概念化强化学习框架(conceptual reinforcement learning, CRL),包含基于注意力机制的概念编码器和保证概念化表示性质的限制性损失函数,以显式为强化学习策略学习概念化的表示。概念化表示具有 2 个重要性质,即概括性和抽象性,可以分别提升策略在新环境上的泛化性能和策略的训练效率。

本文在语言条件任务和文本游戏这 2 类常见的语言强化学习任务上应用了 CRL 框架,并基于任务的不同输入形式,设计了具有不同细节的 CRL 框架。实验结果显示,CRL 框架相比于现有先进算法显著提升了泛化性能(最高 30%)和训练效率(最高 70%)。

具体来说,本文有以下 3 点贡献。

① 国家自然科学基金(62002338,U20A20227,U22A2028)和中国科学院稳定支持基础研究领域青年团队计划(YSBR-029)资助项目。

② 男,1996 年生,博士生;研究方向:强化学习算法;E-mail:pengshaohui18z@ict.ac.cn。

③ 通信作者,E-mail:huxing@ict.ac.cn。

(收稿日期:2023-02-23)

(1)创造性地提出了新的强化学习框架——概念化强化学习(CRL)。相比于标准强化学习框架,CRL 引入了概括且抽象的概念化表示及其无监督的学习方法,有效提升了策略的语义抽象性和智能程度。

(2)在常见的语言类任务(语言条件任务和文本游戏)中应用了概念化强化学习框架,将概念化表示作为强化学习策略的输入。实验结果验证了 CRL 可显著提升策略的泛化性能和训练效率。

(3)通过可视化结果验证了 CRL 框架具有提升策略可解释性的强大潜力,其通过无监督方式学习到的概念化表示类别可以很好地匹配人类先验知识,这意味着 CRL 可作为策略解释方法的基本框架。

1 相关工作

语言强化学习任务主要包括语言条件任务和文本游戏,本节将详细介绍相关任务和方法。除此之外,本节还介绍了与概念化强化学习相关的表示学习领域中的不变表示学习。

1.1 语言类任务

近年来,将语言和强化学习结合的任务和方法有很多,总的来说可以大致分为 3 类:语言条件任务[3-5,8]、文本游戏[6,8-9]和指令遵循任务[10-14]。

语言条件任务通过额外的文本描述为智能体提供环境先验知识,最常用的是“阅读击败怪物”(read to fight monsters, RTFM)[3]和“送信者”(messenger)[4]。现有的在该类任务上的方法都是在训练策略时隐式学习观测和文本的联合表示[3-5]。这种方式学到的隐式表示会不可避免地包含噪音和虚假相关信息,而本文提出的 CRL 框架则通过显式学习概念化表示来解决该问题。

文本游戏起源于人类的交互式小说,智能体与环境通过文本观测和语言化动作进行交互以完成某种任务,常用的文本游戏框架包括文本世界(TextWorld)[15]和 Jericho[16]。现有的基于文本游戏的强化学习算法往往专注于增强探索[9]、动作空间剪枝[8,17]及知识表示与应用[10]等问题。本文提出的 CRL 框架就是通过学习概括且抽象的知识图谱解决文本游戏中的知识表示问题。

除了本文解决的 2 类语言任务,还有一类在机器人控制中常见的语言任务场景,即指令遵循任务[10]。智能体需使用环境支持的简单动作完成各种语言指令描述的复杂任务,不同的语言指令实际表示了同一环境下的不同奖励函数。在该类任务上,传统的方法需要专家知识建模指令结构和环境观测之间的关系以完成策略的规划[11,13]。而最新的方法[12]一般直接学习指令和环境观测的联合嵌入,并将其作为策略网络的输入进行训练。

1.2 不变表示学习

与本文方法相关的另一类工作是不变表示学习,其目的是为强化学习策略学习在多环境或多任务中通用表示[18]。有的不变表示学习方法利用归纳偏置作为结构先验,例如基于对象的结构[19];有的方法试图显式地消除表示中的误导信息,像背景信息[20];还有一些其他方法通过定义不变性来直接优化表示,例如策略相似度度量[21]。现有的不变表示学习方法均无法学习多模态的联合表示,而 CRL 则可通过多层注意力机制和限制性函数学习概括且抽象的联合表示。

2 问题描述

本文主要研究概念化强化学习框架在 2 类常见语言任务上的应用,分别是语言条件任务和文本游戏。本节将具体介绍语言条件任务和文本游戏的问题设置以及在语言任务上的强化学习设定。

2.1 语言条件任务设定

相比于传统强化学习任务,语言条件任务在环境观测之外额外提供描述环境先验知识的文本。该类任务通过在每幕(episode)重新生成环境文本以初始化相似但不同的任务。环境文本描述了环境中实体之间的关系、任务目标和一些环境动态。图 1 左侧显示了环境观测和文本的一般形式。语言条件任务上的强化学习是为了得到最优的语言条件策略 $\pi(\boldsymbol{a} \mid \boldsymbol{o}, \boldsymbol{t})$ 以最大化累积环境奖励,其中策略的输入包含环境观测 $\boldsymbol{o}$ 和文本 $\boldsymbol{t}$。

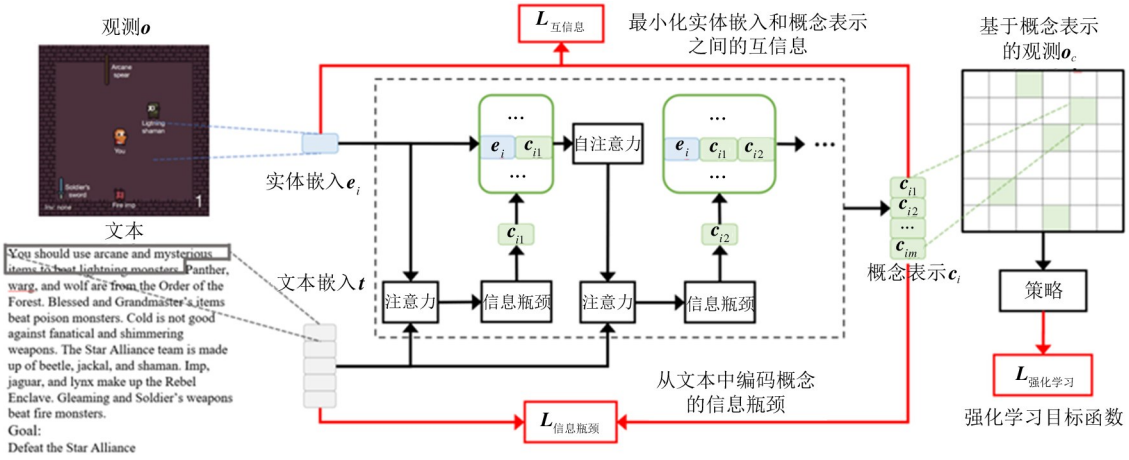


图 1　概念化强化学习在语言条件任务上的应用

下面将以常用语言条件任务 RTFM[3] 和 Messenger[4] 为例,具体描述该类任务的细节。在 RTFM 中,智能体通过阅读描述环境动态的文本,并进行多步推理以获知正确工具和目标怪物。RTFM 提供一组训练环境动态(包含实体组合和角色分配),以及预留的与训练集独立同分布的测试环境动态;其支持 4 种难度设置,包含最简单的 6×6 格子世界、多对一的角色分配(group)、可移动敌人(dyna)以及自然语言描述替代模板语言(nl)。Messenger 是基于格子世界的环境,智能体需从环境说明书中提取出不同实体的身份信息,以完成从送信人处取信并传递给收信人的任务。Messenger 中的关键挑战是分布外泛化问题(out-of-distribution, OOD),关键的实体身份分配在训练和测试环境之间完全不重复,且具有分布漂移(distribution shift)性,这样会导致较明显的虚假相关问题。Messenger 提供 3 个难度等级,包括获取或传递信件(S1)、先获取后传递信件(S2)以及增加诱饵实体和无关描述的送信任务(S3)。RTFM 和 Messenger 中的最优算法分别是 txt2π[3] 和基于多模态注意力的实体映射器(entity mapper with multi-modal attention, EMMA)[4]。txt2π 利用了文本描述的先验结构以构建特殊网络结构生成联合表示。相比于卷积神经网络和线性特征调制网络(feature-wise linear modulation, FiLM),txt2π 获得了显著的性能提升。EMMA 使用基于实体的注意力机制以在文本中提取任务相关信息,在 Messenger 中相比于其他方法(包括 txt2π)实现了性能提升。

2.2　文本游戏设定

文本游戏是完全由语言组成的环境,其环境观测和动作空间都由语言组成。智能体接收由语言描述的环境观测,同时输出由预定义词表中词汇组成的句子或者短语与环境交互。在文本游戏中策略的形式为 $\pi(\boldsymbol{a} \mid \boldsymbol{o}, \boldsymbol{g})$,其中输入包括环境文本观测 $\boldsymbol{o}$ 和可观测知识图谱 $\boldsymbol{g}$。

“做饭”(Cooking)是由文本游戏框架 TextWorld[16] 生成的基于选择性动作的文本游戏。智能体需要在具有多个相互连接地点的世界中寻找食谱,收集并处理对应食材,最终完成备餐。根据世界中地点数量以及食谱复杂程度,Cooking 提供多个游戏难度(1~5,难度 5 是前 4 个难度的均匀混合)。Cooking 同时提供了 2 组训练集,分别包含 100 和 20 个训练实例。Cooking 上的常用算法包括长短期记忆深度 Q 网络算法(long short-term memory-deep Q network, LSTM-DQN)[17]、图辅助的 Transformer 智能体算法(graph-aided Transformer agent, GATA)[6] 和真实图辅助的 Transformer 智能体算法(ground-truth graph aided Transformer agent, GATA-GTF)[6]。LSTM-DQN 是传统的学习语言条件策略的方法,核心是使用 LSTM 编码历史文本信息;GATA 则通过监督学习方式动态生成和更新知识图谱,由于数据集的限制,GATA 并不能很好地泛化到复杂的新环境上。GATA-GTF 算法则在 GATA 的基础上,借助环境提供的完美知识图谱获得了在复杂场景上的性

能和泛化性提升。相比之下,CRL 只使用环境提供的可观测知识图谱 $\boldsymbol{g}$,通过学习概括抽象的概念化知识图谱来提升策略的泛化性和训练效率。

2.3　语言条件强化学习设定

强化学习任务可以被建模为部分可见的马尔可夫决策过程(partial observed Markov decision process, POMDP),该过程可用六元组 (S, O, A, P, r, γ) 描述。其中 S 和 O 分别表示环境状态集合和观测集合,且存在一个不可见的观测函数将环境状态 s 映射为观测 $\boldsymbol{o}$;A 表示智能体可选择的动作集合;P 表示环境状态之间的转移概率函数;$r = R(\boldsymbol{o}, \boldsymbol{a}, \boldsymbol{o}')$ 是环境奖励函数,表示观测下 $\boldsymbol{o}$ 采用动作 $\boldsymbol{a}$ 时转移到 $\boldsymbol{o}'$ 获得的奖励;γ 是环境奖励折扣系数。在语言条件强化学习 $(S, O, A, P, T, r, \gamma)$ 中额外增加了文本描述的集合 T, 表示描述环境的文本的集合。语言条件强化学习的任务与标准强化学习类似(区别是前者的策略输入除了包含观测 $\boldsymbol{o}$ 外,还包含环境语言描述 $\boldsymbol{t}$),具体为学习最佳策略 $\pi(\boldsymbol{a} \mid \boldsymbol{o}, \boldsymbol{t})$ 以最大化与环境交互获得的累积奖励,目标函数如下。

$$L_{\text{强化学习}}(\theta) = \max_{\theta} E\left[\sum_{n=0}^{\infty} \gamma^n R(\boldsymbol{o}_n, \boldsymbol{\pi}_\theta(\boldsymbol{o}_n, \boldsymbol{t}_n), \boldsymbol{o}_{n+1})\right] \tag{1}$$

其中,θ 为待优化的策略参数,E 表示期望值,n 表示与环境交互的步数序号。

3　概念化强化学习框架

本节将介绍概念化强化学习框架(CRL)的具体细节。相比于标准强化学习框架,CRL 在输入和策略之间显式地学习具有概括性和抽象性的概念化表示,以去除表示中的噪音和虚假相关信息,达到策略可快速泛化到新环境的效果。CRL 包含 2 个主要设计,基于注意力的概念化表示编码器和保证表示的概括性和抽象性的无监督损失函数。下面的内容将从 CRL 与现有方法区别出发,并以语言条件任务和文本游戏为例详细介绍 CRL 的设计细节。

3.1　CRL 与现有方法区别

图 2 展示了现有的语言任务中的策略建模方式。在语言条件任务中,策略直接依赖于所有实体和文本实例。然而,实体和文本实例在每一幕任务中都会发生变化,且其采样分布在训练集和测试集之间也可能存在偏移。在文本游戏中也有类似现象,即策略直接依赖于文本描述实例和知识图谱实例。直接依赖实例学习到的联合表示可能包含训练集实例中的噪音和虚假相关信息,这将大大损害策略的训练效率和泛化性能。

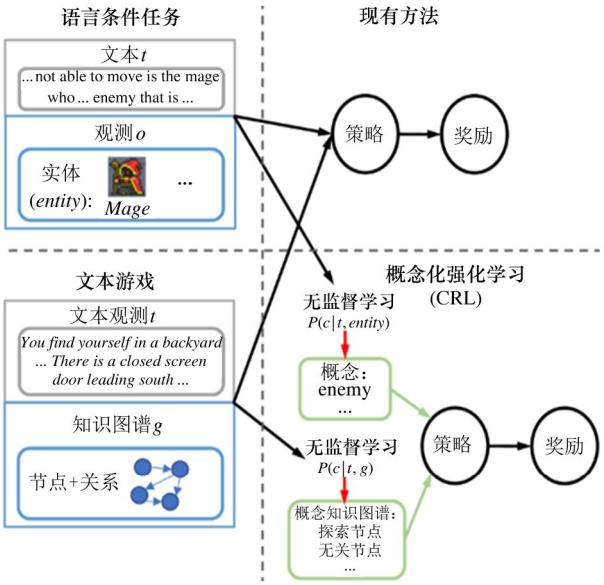


图 2　CRL 和现有方法策略建模方式的区别

如图 2 所示,概念化强化学习在输入和策略之间引入了一个显式学习的中间表示,即概念化表示。具体来说,在语言条件任务上,概念 c 是指从文本信息和实体中推导出的实体间共享性质。例如在 Messenger 任务中,存在角色的概念。任意一个实体,无论其名称或者行为模式,都是送信者、收信者或者是诱饵 3 种角色之一。CRL 通过无监督方式学习概念编码器建模概念表示 $P(\boldsymbol{c} \mid \boldsymbol{t}, \boldsymbol{entity})$,然后将观测中的实体替换为学到的概念进而强化学习策略可以依赖于实体间共享的概念去解决任务。类似地,在文本游戏中,CRL 在输入(文本观测 $\boldsymbol{t}$ 和知识图谱 $\boldsymbol{g}$)与策略之间引入概念化知识图谱 $\boldsymbol{c}_g$。$\boldsymbol{c}_g$ 描述了不同组成元素、不同复杂度文本游戏之间共用的节点性质和关系。例如在 Cooking 游戏中,环境中房间数和可交互物体个数都随难度变化,然而与完成任务相关的节点性质和抽象关系是不变的。CRL 通过建模概念化知识图谱 $P(\boldsymbol{c}_g \mid \boldsymbol{t}, \boldsymbol{g})$ 可以去

除无关节点以保留必要节点和关系。因此,在不同组成元素、不同复杂度的文本游戏中,策略都依赖于相似规模、相同抽象关系的概念化知识图谱,从而具有更好的泛化性能和训练效率。

3.2 语言条件任务中的 CRL 设计

本节首先介绍 CRL 在语言条件任务上的设计。

定义 1 每一幕游戏中,环境提供观测 $\boldsymbol{o} \in R^{h\times w\times d}$,包含 n 个实体 $\boldsymbol{e}=\{\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_n\}\in R^{n\times d}$,以及文本描述 $\boldsymbol{t}$。基于语言条件任务的性质,可以合理假设存在 m 种不可见的概念变量 $\boldsymbol{C}=\{\boldsymbol{C}_1,\boldsymbol{C}_2,\cdots,\boldsymbol{C}_m\}$ 来表征实体间的共享性质。每个实体 $\boldsymbol{e}_i$ 可以由其所对应的概念值的联合来表示,即实体的概念表示 $\boldsymbol{c}_i=\{\boldsymbol{c}_{i1},\boldsymbol{c}_{i2},\cdots,\boldsymbol{c}_{im}\}$。概念化强化学习中的概念编码器将从实体表示和文本描述中推理建模实体的概念表示 $P(\boldsymbol{c}_i\mid\boldsymbol{e}_i,\boldsymbol{t})$。CRL 中概念编码器所编码的概念化表示可能与实际概念值不完全相同,这是因为概念编码器将通过无监督的方式进行训练。然而,CRL 通过限制性损失函数可确保概念化表示具有与真实概念相似的概括性和抽象性。

3.2.1 结构描述

图 1 展示了 CRL 在语言条件任务上的应用过程。在语言条件任务中,概念化强化学习的关键是将观测 $\boldsymbol{o}$ 中的实体 $\boldsymbol{e}_i$ 替换为学习到的概念表示 $\boldsymbol{c}_i$ 以组成基于概念的观测 $\boldsymbol{o}_c$,并提供给策略进行训练。因此,作为关键部分的概念编码器 $f_\theta(\boldsymbol{e}_i,\boldsymbol{t})$ 以实体 $\boldsymbol{e}_i$ 和文本描述 $\boldsymbol{t}$ 作为输入,并输出概念表示 $\boldsymbol{c}_i$。如图 1 中的黑框内所示,概念编码器基于多层注意力机制,其包含 m 个注意力模块,每层输出一个单独的概念值 $\boldsymbol{c}_{ij}(1\leqslant j\leqslant m)$。这些注意力模块逐步产生从文本中直接推导的基础概念以及从文本和已产生概念经过多层推理的复杂概念。

具体来说,第 1 层注意力模块可生成从文本中直接推导的简单概念。例如,在 Messenger 中,不同实体有不同的角色概念,而文本描述会直接给出相关信息,"the thing that is not able to move is the mage who possesses the enemy that is deadly"指出不能移动的"mage"是"enemy"。在这个注意力模块中,

$$\begin{cases}\boldsymbol{q}_1 = MLP(\boldsymbol{e}_i)\\ \boldsymbol{k}_1 = GRU_{\text{key}_1}(\boldsymbol{t})\\ \boldsymbol{v}_1 = GRU_{\text{val}_1}(\boldsymbol{t})\end{cases}\tag{2}$$

其中,$\boldsymbol{q}$、$\boldsymbol{k}$、$\boldsymbol{v}$ 分别表示注意力模块中的查询向量、键向量和值向量,多层感知机(multi-layer perceptron, MLP)是常用的神经网络结构,循环门控单元(gated recurrent unit, GRU)代指使用了 GRU 的循环神经网络。通过标准的注意力模块计算得到的关于概念 $\boldsymbol{c}_{i1}$ 的信息被输入到信息瓶颈(information bottleneck, IB)模块以生成概念 $\boldsymbol{c}_{i1}$,这一过程将在下一节中详细描述。

考虑到某些复杂概念不仅需要文本中的直接信息,而且需要前置已产生的概念,第 1 层之后的注意力模块均增加了自注意力机制以模拟多层推理过程并生成复杂概念。例如在 RTFM 的文本中,"fire monsters are defeated by fanatical and shimmering weapons"解释了"fire"相关的怪物和"fanatical"属性工具的关系。但是为了推导"fanatical sword"实体是否具有概念"useful tool",则需要知道"fire"属性实体是否具有"enemy"的概念。因此 CRL 通过具有多层注意力模块的编码器结构,将浅层概念和实体 $\boldsymbol{e}_i$ 连接起来以生成后续的查询变量。如图 1 所示,查询向量 $\boldsymbol{q}_j(j>1)$ 由环境中所有实体和已生成概念通过自注意力机制生成。

$$\begin{cases}\boldsymbol{q}_j = SelfAttn(Concat(\boldsymbol{e}_i,\boldsymbol{c}_{i1},\cdots,\boldsymbol{c}_{i,j-1}))\\ \boldsymbol{k}_j = GRU_{\text{key}_j}(\boldsymbol{t})\\ \boldsymbol{v}_j = GRU_{\text{val}_j}(\boldsymbol{t})\quad 1<j\leqslant m\end{cases}\tag{3}$$

最终所有注意力模块生成的单个概念被连接在一起生成概念表示 $\boldsymbol{c}_i$,并替换 $\boldsymbol{o}$ 中的 $\boldsymbol{e}_i$ 生成新的基于概念的观测 $\boldsymbol{o}_c$。换言之,CRL 使用 $\boldsymbol{o}_c$ 作为替代原始观测 $\boldsymbol{o}$ 和文本 $\boldsymbol{t}$ 作为环境状态的表示来训练策略 $\pi(\boldsymbol{a}\mid\boldsymbol{o}_c)$。

3.2.2 限制性损失函数

除了基于多层次注意力模块的编码器结构,为了保证生成的概念具有概括性和抽象性,CRL 在训练过程中为编码器增加了额外的无监督损失函数——概括性限制函数和抽象性限制函数。具体来说,本框架从信息论的角度考虑概念的概括性和抽象性。

首先,阐述概括性限制。在语言条件任务中,由于实体和文本描述每幕都会发生变化,同样的实体在不同幕中性质可能不同。因此,概念化表示应该

只反映在不同幕间共享的性质而与每一幕特定的实体信息无关,也就是概念化表示的概括性。基于概括的概念化表示训练的策略只依赖于抽象性质而不是每幕特定的实体,因此可以消除无关信息以防止虚假相关。为了达到这样的目标,需要概念表示的生成分布和实体的采样分布相互独立。互信息(mutual information, MI)是衡量变量间独立性的基本度量,因此 CRL 通过最小化学习到的概念化表示和实体之间的互信息来使得它们互相独立。变量 X 和 Y 之间的互信息计算公式如下所示。

$$I(X;Y) = E_{p(x,y)}\left[\log \frac{p(x,y)}{p(x)p(y)}\right] \tag{4}$$

如图 1 中显示,CRL 使用 CLUB(contrastive log-ratio upper bound)[22] 方法最小化某个实体 $\boldsymbol{e}$ 及其概念 $\boldsymbol{c}$ 之间的互信息 $I(\boldsymbol{e};\boldsymbol{c})$。当有 N 个样本数据 $\{\boldsymbol{e}_i, \boldsymbol{c}_i\}$ $(1 \leqslant i \leqslant N)$ 时,CLUB 估计的互信息上界如式(5)所示。

$$I_{\text{CLUB}}(\boldsymbol{e};\boldsymbol{c}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[\log q(\boldsymbol{c}_i \mid \boldsymbol{e}_i) - \log q(\boldsymbol{c}_j \mid \boldsymbol{e}_i)\right] \tag{5}$$

其中,$q(\boldsymbol{c} \mid \boldsymbol{e})$ 是 CLUB 方法中额外学习的以实体 $\boldsymbol{e}$ 为输入预测概念表达 $\boldsymbol{c}$ 的估计器。为了最小化上述互信息上界,CRL 引入了基于 CLUB 的损失函数 L 互信息最小化每幕游戏 n 个实体及其概念化表示之间的互信息:

$$L_{\text{互信息}}(\theta) = \sum_{i=1}^{n} I_{\text{CLUB}}(\boldsymbol{c}_i;\boldsymbol{e}_i) = \sum_{i=1}^{n} I_{\text{CLUB}}(f_\theta(\boldsymbol{e}_i,\boldsymbol{t});\boldsymbol{e}_i) \tag{6}$$

其次,概念化表示还需要抽象性限制。由于在 CRL 框架中策略的输入仅有概念化表示,因此强化学习最大化累积奖励的目标函数间接保证了学习到的概念化表示的有效性,但这样正向的目标函数不能保证表示的抽象性。为了使得概念化表示只包含任务相关的必要信息,CRL 引入了抽象性限制函数来过滤掉无关信息。信息瓶颈(IB)技术可以压缩输入变量 X 中的噪音信息,并将对输出变量 Y 有表达力的信息保留在中间变量 Z 中,目标函数如下:

$$L_{\text{IB}} = \min_{p(z\mid x)} I(X;Z) - \beta I(Z;Y) \tag{7}$$

其中 β 表示拉格朗日乘数。

如上所述,强化学习目标函数可以隐式地保证在概念化表示中保留有效信息,因此仅需使用式(7)中的前一项来压缩从文本 $\boldsymbol{t}$ 编码概念表示 $\boldsymbol{c}$ 时的无关信息。具体来说,如图 1 所示,CRL 使用深度变分信息瓶颈(deep variational information bottleneck, Deep VIB)[23] 技术来最小化文本 $\boldsymbol{t}$ 和概念表示 $\boldsymbol{c}$ 的互信息,即将概念化表示的编码分布与正态分布 $\boldsymbol{N}(0,1)$ 之间的 KL 散度作为损失函数 $L_{\text{信息瓶颈}}(\theta)$:

$$L_{\text{信息瓶颈}}(\theta) = \frac{1}{N}\sum_{k=1}^{N}\sum_{i=1}^{n}\sum_{j=1}^{m} E_{\varepsilon\sim p(\varepsilon)}\left[KL(p_\theta(c_{ij} \mid t_{ij}), \boldsymbol{N}(0,1))\right] \tag{8}$$

其中,N 为训练数据的个数,n 为实体个数,m 为概念个数。

因此在语言条件任务中,为了生成概括且抽象的概念化表示并训练基于概念化表示的策略,CRL 的目标函数如下:

$$L_{\text{CRL}}(\theta) = L_{\text{强化学习}}(\theta) + \alpha_1 L_{\text{互信息}}(\theta) + \alpha_2 L_{\text{信息瓶颈}}(\theta) \tag{9}$$

其中,$L_{\text{强化学习}}(\theta)$ 是强化学习的目标函数,α_1 和 α_2 系数是可调超参数。

3.3 文本游戏中的 CRL 设计

定义 2 在文本游戏中,环境提供文本观测 $\boldsymbol{t}$ 以及当前可观测知识图谱 $\boldsymbol{g}$。根据每一幕任务的不同,无论节点名称,知识图谱的节点均可以抽象地划分为有用节点(与任务目标相关的物体,待探索的空间等)和噪音节点(无用的物体或者已经探索完全的空间),可以用概念掩码 $\boldsymbol{m}_i$ 来表示。概念编码器从文本信息和节点信息中建模知识图谱的概念掩码 $P(\boldsymbol{m}_i \mid \boldsymbol{t}, \boldsymbol{n}_i)$,再利用概念掩码对原始知识图谱进行遮掩处理以得到概念化知识图谱 $\boldsymbol{g}_c$。通过所有掩码的统一值域(0 ~ 1)以及对掩码的稀疏化,CRL 可以保证概念化知识图谱的概括性和抽象性,进而提升语言条件策略的泛化性和训练效率。

在文本游戏中,概念化强化学习的关键是将复杂且多变的知识图谱 $\boldsymbol{g}$ 通过概念化的掩码 $\boldsymbol{m}$ 替换为概括且抽象的概念化知识图谱,并提供给策略进行训练。因此如图 3 中黑色虚线框内所示,概念化知识图谱编码器 $f_\theta(\boldsymbol{n}_i, \boldsymbol{t})$ 以文本信息嵌入 $\boldsymbol{t}$ 和原始知识图谱中的节点嵌入 $\boldsymbol{n}_i$ 作为输入,并输出概念化

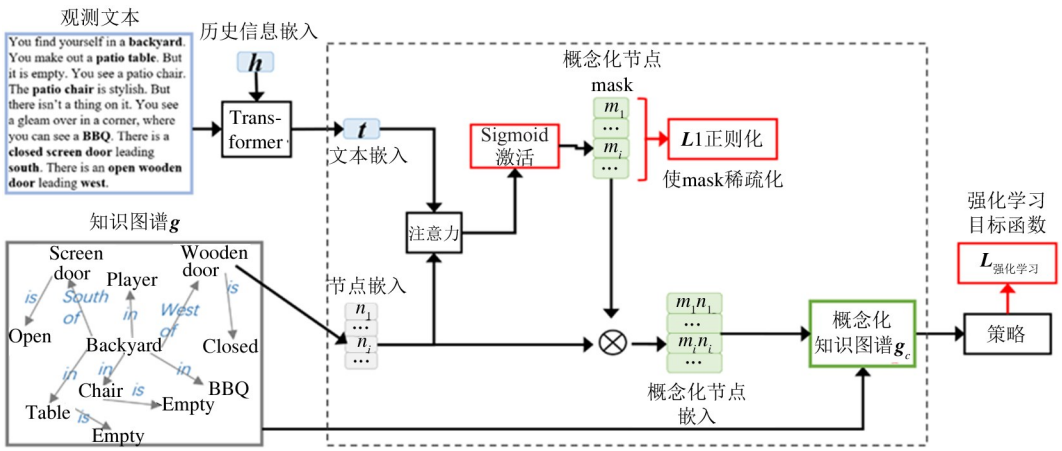


图 3　概念化强化学习在文本游戏上的应用

知识图谱 $\boldsymbol{g}_c$。首先是一个注意力模块,查询向量$\boldsymbol{q}$由单个节点嵌入$\boldsymbol{n}_i$产生,键值向量$\boldsymbol{k}$和值向量$\boldsymbol{v}$则由文本信息产生:

$$\begin{cases} \boldsymbol{q} = MLP(\boldsymbol{n}_i) \\ \boldsymbol{k} = GRU(\boldsymbol{t}) \\ \boldsymbol{v} = GRU(\boldsymbol{t}) \end{cases} \tag{10}$$

接下来 CRL 通过 Sigmoid 激活函数将生成掩码的值控制在 0 和 1 区间,以保证概念化掩码的统一性。为了将原始知识图谱中的节点中的有用节点保留而无用节点去除,CRL 将概念化掩码 $\boldsymbol{m}_i$ 与节点嵌入 $\boldsymbol{n}_i$ 相乘得到了概念化的节点嵌入 $\boldsymbol{m}_i\boldsymbol{n}_i$,并最终用概念化节点嵌入替代原始知识图谱中节点得到概念化知识图谱 $\boldsymbol{g}_c$ 作为策略的输入。

概念化知识图谱相比于原始知识图谱具有概括性以及抽象性,但相比于语言条件任务的概念化表示 $\boldsymbol{o}_c$ 与当前幕的实体 $\boldsymbol{e}_i$ 无关,概念化知识图谱 $\boldsymbol{g}_c$ 通过掩码机制保留当前幕知识图谱中的有用节点信息,这是由文本游戏的动作生成需要有用节点信息的性质决定的。

在基于注意力的概念化掩码生成机制中,Sigmoid 激活函数保证了掩码权重的统一性,即任意幕或任意环境的掩码的取值范围都是 0 ~ 1,这样统一的概念化掩码事实上保证了概念化知识图谱节点的概括性。同时,为了使得掩码尽可能稀疏,保证概念化知识图谱的抽象性,如图 3 中浅色框所示,CRL 对生成的掩码引入了 $L1$ 正则化进行稀疏化:

$$L1 = \sum_{i=1}^{n} | m_i | \tag{11}$$

其中 n 是节点个数。

与语言条件任务中概念化表示的机制类似,最终 CRL 只将概念化知识图谱作为策略的输入,因此强化学习的目标函数隐式地保证了概念化掩码所保留信息的有效性。同时,通过 Sigmoid 激活函数和 $L1$ 正则化稀疏的掩码也保证了概念化知识图谱的概括性和抽象性。

因此在文本游戏中,CRL 的目标函数如下:

$$L_{\mathrm{CRL}}(\theta) = L_{强化学习}(\theta) + \sum_{i=1}^{n} | m_i | \tag{12}$$

4　实　验

本节首先展示了概念化强化学习框架在 3 种语言条件任务中与对应任务的先进算法的性能比较,包括 RTFM 上的 txt2π 算法,Messenger 上的基于多模态注意力的实体映射器(EMMA)算法和 Cooking 上的长短期记忆深度 Q 网络算法(LSTM-DQN)、图辅助的 Transformer 智能体算法(GATA)和真实图辅助的 Transformer 智能体算法(GATA-GTF),以证明 CRL 框架的有效性。衡量标准包括了迁移性能,零样本(zero-shot)泛化性能和训练效率,以分别证明概念化表示的概括性和抽象性对强化学习策略的作用。然后,本文进行了消融实验以展示概念化强化学习框架的各个关键部件对性能的影响。最后,本文通过概念表示的可视化结果直观地展示了 CRL 学到的概念表示的物理含义,其结果清晰地显示了概念化表示与人类先验概念的匹配性,证明了 CRL

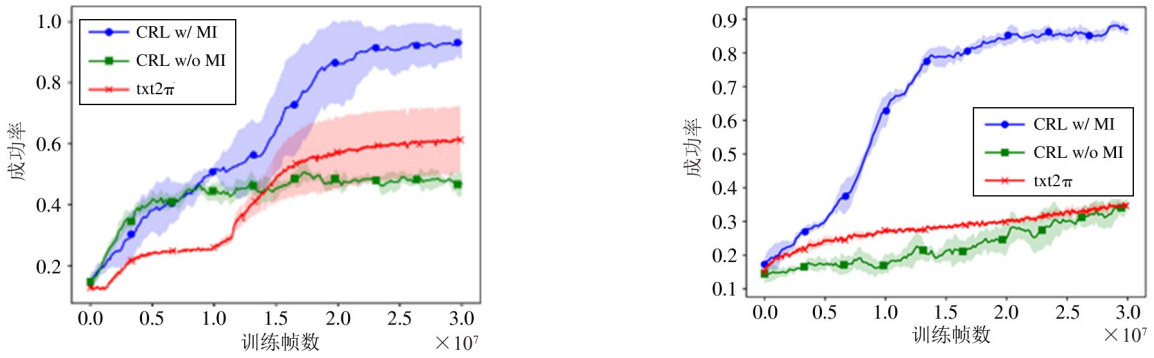
对强化学习策略可解释性的提升。

4.1　迁移性能和效率比较

表 1 比较了 CRL 和 txt2π 在 RTFM 上的迁移性能和效率。迁移性能是指将源任务中训练过的策略迁移至新的目标任务上训练最终达到的性能。表 1 展示了 CRL 和 txt2π 分别从随机初始化的策略(没有经过训练的策略)和 6×6 环境上训练的策略到 8 个更难的目标任务上的迁移性能。下划线表示在同样设置中的最佳性能。在迁移性能和迁移中的训练效率 2 个方面,CRL 都显著超越了 txt2π。在迁移性能方面,当从基础的 6×6 环境向所有其他复杂任务迁移时,CRL 的性能明显超过了 txt2π,最多达到了 46% 的成功率提升。这是因为 CRL 在基础环境上学习了概括的概念,这些概念在复杂环境上得到了有效的重用。在训练效率方面,CRL 也获得了较大提升,从随机初始化策略训练时节省了 70% 的训练时间,在迁移训练时也节省了 40% 的训练时间。训练效率的提升受益于 CRL 中概念化表示的抽象性,从图 4(a)中学习曲线也可以直观地看到训练效率的提升。图中 CRL w/MI 和 CRL w/o MI 分别为有互信息损失函数限制的概念强化学习和缺少互信息损失函数限制的概念强化学习。

表 1　CRL 和基线方法在 RTFM 环境上的任务成功率(%)

源任务	方法	迁移目标任务								
		6×6	6×6 dyna	6×6 groups	6×6 nl	6×6 dyna groups	6×6 groups nl	6×6 dyna nl	6×6 dyna groups nl	训练步数
随机	txt2π	84±20	26±7	25±3	45±6	23±2	25±3	23±2	23±2	100×10^6
	CRL	93±3	36±10	33±10	38±4	17±2	20±2	17±3	16±3	30×10^6(↓70%)
6×6	txt2π	——	85±9	82±19	78±24	64±12	52±13	53±18	40±8	50×10^6
	CRL	——	89±9	96±3	97±2	85±1	87±2	85±4	86±3	30×10^6(↓40%)



(a) RTFM-简单环境　　(b) RTFM-困难环境

图 4　CRL 在 RTFM 上关于互信息(MI)损失函数的消融实验结果

4.2　零样本泛化性能和效率比较

本文在 Messenger 环境和 Cooking 环境上比较了 CRL 与基线算法的训练性能和分布外测试环境的零样本泛化性能,其中零样本泛化性能指的是训练场景上的策略在未见过的测试场景上不经过再次训练的性能。

Messenger环境:表2展示了CRL在更困难、更具区分度的 S2 和 S3 环境上零样本泛化性能均超过了之前的最优算法 EMMA,具体表现为更高的任务成功率(最多 20%)和更低的泛化差距(最多 30%)。结果证明了概念化表示的概括性可以较好地缓解分布外泛化场景中的虚假相关问题。同时,CRL 也显示了一定的训练效率的提升,如图 4(b)中学习曲线所示。

表 2　CRL 和基线方法在 Messenger 环境上的训练及测试成功率

	方法	训练场景/%	测试场景/%	训练步数/ $\times 10^6$
S1	EMMA	98 ± 2.1	85 ± 1.4	30
	CRL	98 ± 1.0	88 ± 2.5	30
S2	EMMA	96 ± 2.0	45 ± 12.0	30
	CRL	98 ± 2.0	76 ± 5.0	30
S3	EMMA	19 ± 2.9	10 ± 0.8	30
	CRL	43 ± 5.0	32 ± 1.9	30

Cooking 环境:在不同训练和难度等级设置下的结果比较如表 3 和表 4 所示。CRL 相比于基线方法在零样本泛化性能上有显著提升,其中提升最明显的是小训练集(20 个训练游戏)和复杂难度(难度等级 4)。这说明了概括且抽象的概念化知识图谱有效提升了策略的训练效率和泛化性。同时结果还显示了在混合难度设置上(难度等级 5),CRL 相对于传统的知识图谱方法提升有限,这可能是因为不同难度环境中知识图谱的概念化掩码无法用统一的推导方式得到,从而影响了学习概念化知识图谱的效果。

表 3　CRL 和基线方法在 Cooking 环境上的训练任务成功率(%)

测试性能 难度等级	20 个训练游戏						100 个训练游戏					
	1	2	3	4	5	提升↑	1	2	3	4	5	提升↑
LSTM-DQN	90.8	36.9	69.3	31.5	61.2	——	63.4	33.2	66.1	31.9	55.2	——
GATA	90.9	41.4	66.1	31.3	58.8	+0.6	67.1	33.1	73.6	29.9	51.2	+0.7
GATA-GTF	100.0	96.9	98.3	44.9	76.6	+56.5	98.6	90.5	99.0	38.9	73.4	+61.5
CRL	100.0	96.2	92.5	73.5	60.0	+67.1	100.0	92.5	90.0	83.5	63.3	+89.8
CRL-ablation	100.0	78.8	82.5	53.5	78.3	+48.1	100.0	88.5	76.3	45.0	71.7	+62.1

注:CRL-ablation 为去掉 L1 正则化的消融变体

表 4　CRL 和基线方法在 Cooking 环境上的测试任务成功率(%)

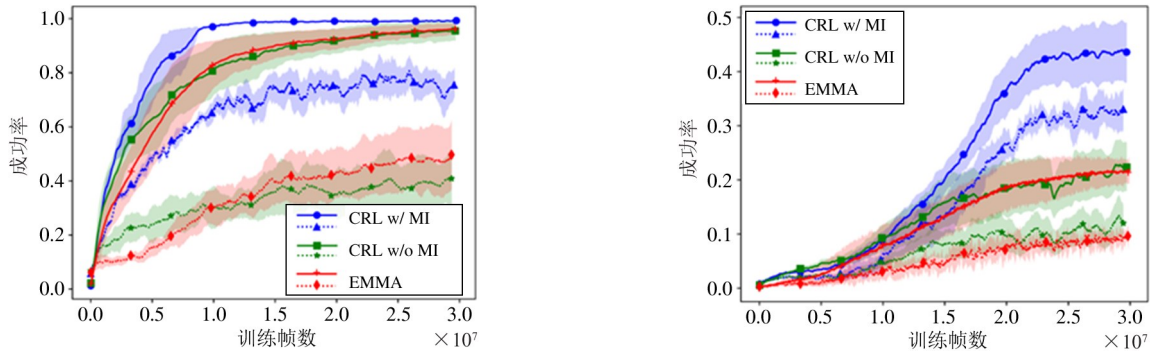
测试性能 难度等级	20 个训练游戏						100 个训练游戏					
	1	2	3	4	5	提升↑	1	2	3	4	5	提升↑
LSTM-DQN	66.2	26.0	16.7	18.2	27.9	——	62.5	32.0	38.3	17.7	34.6	——
GATA	68.8	33.0	41.7	17.7	27.0	+34.9	62.5	33.0	46.7	25.9	33.4	+24.2
GATA-GTF	48.7	61.0	46.7	23.6	28.9	+64.2	95.0	95.0	70.0	37.3	52.8	+81.6
CRL	100.0	61.3	66.3	55.0	28.3	+131.4	100.0	81.2	72.5	73.3	48.3	+137.5
CRL-ablation	95.0	53.7	43.8	16.8	26.3	+59.8	100.0	63.7	62.5	32.0	45.0	+66.6

注:CRL-ablation 为去掉 L1 正则化的消融变体

4.3　消融实验

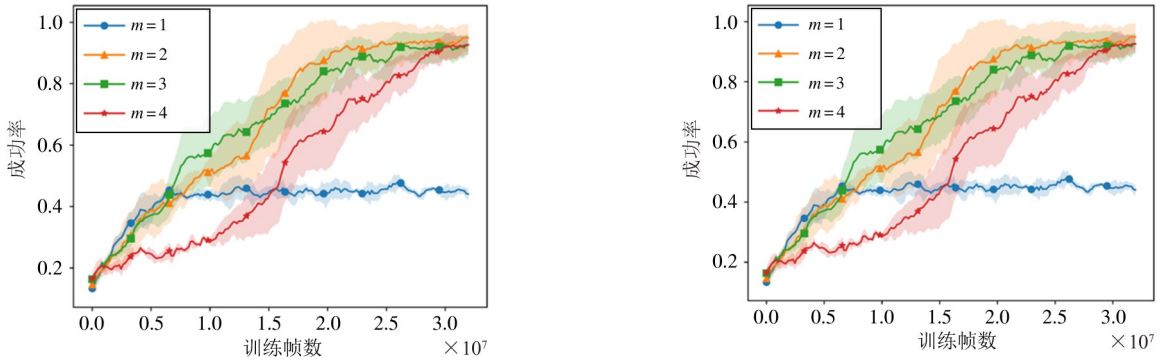
互信息(MI)损失函数:为了证明基于互信息的限制性损失函数在学习概念过程中的作用,本文验证了消融变体 CRL w/o MI 的性能。CRL w/o MI 与 CRL 具有同样的概念编码器结构,但是缺少了基于互信息的限制。图 4 和 5 显示了 CRL 及其消融变体的学习曲线,其中图 5 中实线表示训练场景性能,虚线表示测试场景性能。结果显示了具有互信息限制的 CRL 明显超过了其消融变体,尤其在泛化能力方面;而 CRL w/o MI 性能退化为类似隐式学习表示的基线方法。该消融实验证明了互信息限制对于学习概括且抽象的概念化表示的重要性。

概念数量 m:本文探究了概念数量 m 对于 CRL 的影响。考虑到在 RTFM 中存在 2 种概念,本文测试了 $m=1\sim4$ 的 CRL 在 RTFM 上的迁移性能。如图 6 所示,当 m 大于真实值时,CRL 的效率被轻微影响,但是渐近性能保持稳定。原因是概念编码器的待训练参数变多,导致了训练效率的轻微下降,但 CRL 仍然能保证学习到的概念具有概括性和抽象性,因此渐进性能保持稳定。当预定义 m 小于真实值时,概念编码器无法提取到足够的任务相关信息,因此性能下降。这样的消融结果启发用户调参时设置较大的概念数 m。



（a）Messenger-S2 环境　　（b）Messenger-S3 环境

图 5　CRL 在 Messenger 上关于互信息(MI)损失函数的消融实验结果



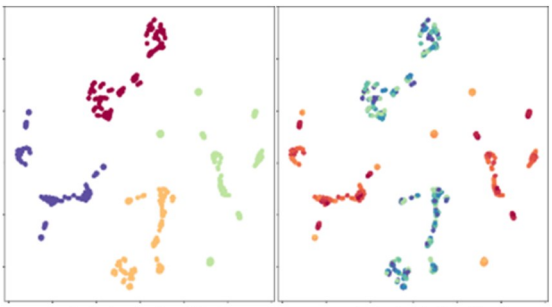
（a）RTFM-简单环境　　（b）RTFM-困难环境

图 6　CRL 在 RTFM 上概念数量 *m* 的消融实验结果

知识图谱:在文本游戏任务中,为了显示 *L*1 正则化损失函数对于概念化知识图谱生成的重要性,本文进一步比较了 CRL 和去掉 *L*1 正则化的消融变体(CRL-ablation)的性能,结果如表 3 和表 4 的最后一行所示。从结果可以看到,不做掩码稀疏化的概念化知识图谱性能退化为类似传统知识图谱方法(GATA-GTF),甚至在复杂难度设置上,性能有所下降。原因是 CRL ablation 只使用当前可见知识图谱且未做稀疏化,而 GATA-GTF 使用的是完美知识图谱。换言之,GATA-GTF 使用更多的环境信息降低了任务难度。

4.4　可视化实验

为了直观地展示 CRL 学到的概念,本文使用 T 分布和随机近邻嵌入(t-distributed stochastic neighbor embedding, TSNE)可视化技术对概念表示可视化,如图 7 和图 8 所示。具体来说,本文对环境进行随机初始化并保存 1 000 个实体生成的概念化表示变量进行可视化。从人类视角来看,RTFM 中存在 2 个变量,即目标/诱饵怪物和有用/无用工具。图 7 中展示了分别被人类理解的概念真实类别(概念学习过程不可见)和环境提供的实体类别染色的 TSNE 结果。结果显示学习到的概念向量分离为 4 个类别且与概念真实类别相匹配,而实体类别染色不能达到这样的效果。这样的现象直观地展示了 CRL 学习的概念表示具有概括性和抽象性,能够有效地消除实体分布带来的虚假相关;除此之外,可与人类先



(a) 真实概念类别染色　　(b) 真实实体类别染色

图 7　RTFM 上产生的概念化表示可视化结果

验类别相匹配也说明了概念化表示可以有效地增强策略的可解释性。

5 结 论

语言类强化学习任务借助语言作为结构化的中间渠道来增强强化学习策略在不同环境间的泛化性。本文提出了概念化强化学习框架(CRL),利用基于注意力的概念编码器和限制性损失函数显式学习具有概括性和抽象性的概念化表示,以解决在语言类任务中自动学习环境间通用表示的挑战。本文在常见的2类语言强化学习任务上验证了CRL对提升策略训练效率和泛化性能的有效性。

对概念化表示可视化过程揭示了CRL通过无监督范式学习到的概念类别和人类先验知识的高度匹配性,意味着概念化表示不仅可以提升强化学习策略的性能,而且在提升策略可解释性方面具有很大潜力。除此之外,本文通过结合人类认知方法和机器学习范式,有效提升了人工智能算法的效率和鲁棒性,这可能启发未来更多的结合人类行为模式和机器学习以提升人工智能水平的工作。

参考文献

[1] KIRK R, ZHANG A, GREFENSTETTE E, et al. A survey of generalisation in deep reinforcement learning[EB/OL]. (2021-11-18)[2023-02-22]. https://arxiv.org/pdf/2111.09794.pdf.

[2] NARASIMHAN K, BARZILAY R, JAAKKOLA T. Grounding language for transfer in deep reinforcement learning [J]. Journal of Artificial Intelligence Research, 2018, 63:849-74.

[3] ZHONG V, ROCKTÄSCHEL T, GREFENSTETTE E. RTFM: generalising to novel environment dynamics via reading[C]//International Conference on Learning Representations. Virtual:ICLR, 2020: 1-17.

[4] HANJIE A W, ZHONG V Y, NARASIMHAN K. Grounding language to entities and dynamics for generalization in reinforcement learning[C]//International Conference on Machine Learning. Virtual: ICML, 2021:4051-4062.

[5] WU Y, FAN Y, LIANG P P, et al. Read and reap the rewards: learning to play Atari with the help of instruction manuals[EB/OL]. (2023-02-12)[2023-02-22]. https://arxiv.org/pdf/2302.04449.pdf.

[6] ADHIKARI A, YUAN X, CÔTÉ M A, et al. Learning dynamic belief graphs to generalize on text-based games [J]. Advances in Neural Information Processing Systems, 2020,33:3045-57.

[7] ZHONG V, HANJIE A W, WANG S, et al. SILG: the multi-domain symbolic interactive language grounding benchmark[J]. Advances in Neural Information Processing Systems, 2021,34:21505-21519.

[8] CHEN C, DAI Y, POON J, et al. Ananalysis of deep reinforcement learning agents for text-based games[EB/OL]. (2022-09-12)[2023-02-22]. https://arxiv.org/pdf/2209.04105.pdf.

[9] 宋浩楠,赵刚,王兴芬. 融合知识表示和深度强化学习的知识推理方法[J]. 计算机工程与应用, 2021,57(19):189-197.

[10] CHAPLOT D S, SATHYENDRA K M, PASUMARTHI R K, et al. Gated-attention architectures for task-oriented language grounding[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI Press, 2018:1-11.

[11] 陈稳舟. 面向精确指令与自然指令的移动机器人室内导航研究[D]. 杭州:浙江大学, 2022.

[12] 吕小晗. 基于长短期记忆网络的机器人视觉语言航行算法研究[D]. 秦皇岛:燕山大学, 2021.

[13] MEI H, BANSAL M, WALTER M R. Listen, attend, and walk: neural mapping of navigational instructions to action sequences[C]//The 30th AAAI Conference on Artificial Intelligence. Phoenix, USA: AAAI Press, 2016: 2772-2778.

[14] GOYAL P, NIEKUM S, MOONEY R J. Using natural language for reward shaping in reinforcement learning [EB/OL]. (2019-05-31)[2023-02-22]. https://arxiv.org/pdf/1911.11404.pdf.

[15] CÔTÉ M A, KÁDÁR A, YUAN X, et al. Textworld: a learning environment for text-based games[C]//The 7th Workshop in Conjunction with the 27th International Conference on Artificial Intelligence. Stockholm, Sweden:ICML, 2019:41-75

[16] HAUSKNECHT M, AMMANABROLU P, CÔTÉ M A, et al. Interactive fiction games: a colossal adventure[C]//Proceedings of the AAAI Conference on Artificial Intelli-

gence. New York, USA: AAAI Press, 2020:7903-7910.

[17] NARASIMHAN K, KULKARNI T, BARZILAY R. Language understanding for text-based games using deep reinforcement learning[EB/OL]. (2015-09-11)[2023-02-22]. https://arxiv.org/pdf/1506.08941.pdf.

[18] 王江晴,王雪言,孙翀,等. 用于多表连接优化的深度强化学习嵌入表示[J]. 计算机工程与设计, 2023,44(2):576-581.

[19] ZHANG X, SONG Y, UEHARA M, et al. Efficient reinforcement learning in block MDPs: a model-free representation learning approach[C]// International Conference on Machine Learning. Virtual: ICML, 2022: 26517-26547.

[20] WANG X, LIAN L, YU S X. Unsupervised visual attention and invariance for reinforcement learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual: IEEE, 2021: 6677-6687.

[21] AGARWAL R, MACHADO M C, CASTRO P S, et al. Contrastive behavioral similarity embeddings for generalization in reinforcement learning[EB/OL]. (2021-03-18)[2023-02-22]. https://arxiv.org/pdf/2101.05265.pdf.

[22] CHENG P, HAO W, DAI S, et al. Club: a contrastive log-ratio upper bound of mutual information[C]// International Conference on Machine Learning. Vienna, Austria: ICML, 2020:1779-1788.

[23] ALEMI A A, FISCHER I, DILLON J V, et al. Deep variational information bottleneck[EB/OL]. (2016-12-01)[2023-02-22]. https://arxiv.org/pdf/1612.00410.pdf.

Conceptual reinforcement learning for language-assisted tasks

PENG Shaohui[*,**,***], HU Xing[*], ZHI Tian[*]

([*]State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

([**]University of Chinese Academy of Sciences, Beijing 100049)

([***]Cambricon Technologies, Beijing 100080)

Abstract

Language-assisted tasks are proposed to facilitate the generalization ability of reinforcement learning policy. The key question is to learn the general representation across different scenarios. Existing studies often implicitly learn the joint representation, which may include spurious correlation information and consequently compromise policy's generalization performance and training efficiency. To address this issue, a conceptual reinforcement learning framework (CRL) is proposed, which exploits the motivation of human cognition that extracts similarits from numerous instances to generate conceptual abstraction, and incorporates a multi-level attention encoder and restricted loss functions to learn compact and invariant conceptual representation for the policy. Evaluated in challenging language-assisted tasks, the results demonstrate that CRL significantly improves the policy's training efficiency (up to 70%) and generalization ability (up to 30%). Additionally, the conceptual representation also shows better interpretability than other representations.

Key words: deep reinforcement learning (DRL), language-assisted reinforcement learning task, text game, representation learning, mutual information