第 58 卷　第 5 期
2022 年 9 月

南京大学学报(自然科学)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCE)

Vol. 58, No. 5
Sept., 2022

DOI:10.13232/j.cnki.jnju.2022.05.006

局部可观测环境下未来信息辅助的无模型深度强化学习

常芳芳,陈祺航,刘云龙*
(厦门大学自动化系,厦门,361102)

摘　要:深度强化学习结合了深度学习的特征提取能力和强化学习的决策能力,近年来在众多领域得到了广泛应用,但现有的针对深度强化学习的研究通常假定系统状态完全可观测,而在实际应用中,由于受到感知能力的限制,智能体往往不能完全确定所处状态,即所处环境为局部可观测环境.同时,现有的无模型强化学习算法往往仅依赖以往历史数据来确定决策策略,不能利用可辅助智能体决策的未来有关信息.以局部可观测问题为应用背景,通过利用对比预测编码(Contrastive Prediction Code,CPC)对未来信息的预测能力实现局部可观测环境下未来信息辅助的无模型决策学习,提出的算法既保留了无模型强化学习算法端对端的训练、性能优势,又能充分利用预测的信息来辅助智能体的决策.在不同的局部可观测环境任务上对提出的算法进行了验证和对比,实验结果验证了该算法的有效性.

关键词:深度强化学习,局部可观测环境,对比预测编码,未来信息,表征学习
中图分类号:TP391　　**文献标志码:**A

Model-free deep reinforcement learning with future information in partially observable domains

Chang Fangfang, Chen Qihang, Liu Yunlong*
(Department of Automation, Xiamen University, Xiamen, 361102, China)

Abstract: By combining the abilities of feature extraction of deep learning and decision-making of reinforcement learning, deep reinforcement learning algorithms have been widely applied in various domains in recent years. While current algorithms mainly focus on planning in fully observable environments, in reality, the states of many applications can only be partially observed due to the limitation of the agents′ perception, i.e., the environments are partially observable. Furthermore, for model-free reinforcement learning algorithms, the decision usually relies on historical data, and no future information that may help the decision making is utilized. In this paper, aims to address the planning problem in partially observable domains, we propose a model-free reinforcement learning algorithm where future information can be incorporated as in the model-based reinforcement learning framework, and the future information is predicted by Contrastive Prediction Code (CPC). Our proposed algorithm can not only retain the end-to-end training and performance advantages of the model-free reinforcement learning algorithm, but also utilize future information for the decision of the agent. The proposed algorithm has been verified and compared on different locally observable environmental tasks. Experimental results demonstrate the effectiveness of the proposed algorithm.

Key words: deep reinforcement learning (DRL), partially observable environment, contrastive prediction code (CPC), future information, representation learning

---

基金项目:国家自然科学基金(61772438,61375077)
收稿日期:2022－05－04
* 通讯联系人,E-mail:ylliu@xmu.edu.cn

深度强化学习结合了深度神经网络强大的特征抽取能力和强化学习的决策能力,以AlphoGo为代表之一,近年来在电脑游戏、医疗健康、机器人控制等领域得到了广泛应用[1-3]. 目前,针对深度强化学习的研究通常假定智能体所处环境完全可观测,即通过一帧或几帧输入图像即可完全确定智能体所处状态. 而在现实应用中,智能体由于受到其感知能力的限制,有可能感知不到所处环境的某些重要特征,同时,智能体采取的动作也往往达不到预期效果,此类问题被称为局部可观测问题或不确定性环境下的智能体的决策规划问题. 日常生活中,局部可观测问题并不鲜见,例如汽车无人驾驶、机器人导航、口语对话系统、用户兴趣获取等,都是典型的局部可观测问题.

对于局部可观测领域,一般的做法是先对环境进行建模,进而利用得到的环境模型确定智能体的规划策略. 其中,局部可观测马尔可夫决策过程(Partially Observable Markov Decision Process,POMDPs)模型[4]以系统隐含状态的概率分布表示信念状态,提供对局部可观测环境建模的方案. 此外,预测状态表示(Predictive State Representations,PSRs)[5]以未来事件发生概率组成的预测向量来表示信念状态,并对局部可观测环境建模. 此类方案主要应用于状态离散的较小规模系统,而现有的深度强化学习的研究通常直接以观测图像作为智能体输入,并使用深度神经网络抽取输入的特征作为状态表示,更符合实际系统的应用需求.

针对深度强化学习的研究,从有无模型角度看可分为有模型和无模型的方法. 基于模型的方法需要首先对环境建模,再利用模型产生的预测/未来信息进行相关规划,但对于大部分环境,建立系统的准确模型非常困难,而模型不准确往往会导致预测轨迹出现偏差,影响决策过程,降低算法效果. 局部可观测环境下的无模型深度强化学习,有关算法往往仅能依赖以往的历史信息来进行决策,例如深度循环Q学习(Deep Recurrent Q-Network,DRQN)[6]首先利用循环神经网络整合历史信息,充分统计历史信息以表示信念状态,进而建立状态到动作的映射以找到最佳策略. 但现有的无模型强化学习的研究通常只能通过以往的历史数据直接学习观测动作的映射,缺少利用未来信息辅助决策的能力.

对比预测编码(Contrastive Prediction Code,CPC)[7]是一种将高维数据进行编码以获取全局特征的方法. 与一般的通过重构误差构建环境模型的方法不同,CPC使用自监督学习的方式最大化历史信息与未来之间的相关性来提取全局特征,得到建立于所学隐状态上的一步预测模型. 基于动作的对比预测编码算法(CPC|Action)[8]则通过对预测特征分类的方式对局部可观测环境的信念状态进行表征学习,提高了RL(Reinforcement Learning)智能体的性能.

本文以深度强化学习典型算法——深度Q网络(Deep Q-Network,DQN)为基础,借助对比预测编码算法,基于当前的信念状态与动作来预测未来的状态特征,并将预测的未来特征用于辅助无模型决策学习. 具体地,本文首先使用CPC|Action方法对当前统计得到的局部可观测环境信念状态进行表示,使信念状态包含未来信息;进而将预测特征作为决策的一部分,使决策过程不仅考虑当前状态,同时也将预测特征考虑进去. 提出的算法使用端到端的训练方式,在不同的局部可观测环境任务中进行实验,验证了算法的效果.

1　相关工作

目前深度强化学习的主流为无模型方法,无模型方法又可以分为基于值的深度强化学习方法和基于策略梯度的深度强化学习方法. DQN[9]通过神经网络拟合值函数,避免用表格方式计算Q,通过状态直接计算对应的Q以得到最优的动作. 演员评论家算法(Actor-Critic,A2C)[10]使用策略梯度更新的方式,引入优势函数,进一步提升算法效果. 近期,Nguyen et al[11]利用预测信息进行对比学习,根据前一状态和动作预测下一状态,与真实的状态进行对比学习,同时,预测误差还作为内在奖励被用于驱动探索机制. 此类工作在很多领域取得了成功,但相关算法主要针对全局可观测的环境.

局部可观测环境下,智能体不能得到完整的

状态信息,通常需要对环境建模. 值预测网络(Value Prediction Network, VPN)[12]使用神经网络对环境建模,而后利用蒙特卡罗算法查找最优动作. Karkus et al[13]使用贝叶斯过滤器对环境建立模型,而后通过Bellman方程计算最大的Q以得到最优策略. 然而,模型的准确性是有模型RL算法的关键,不准确的模型会严重损害策略的性能. 因此,局部可观测环境下信念状态的表示尤为重要.

局部可观测的信念状态表示中,Dosovitskiy and Koltun[14]使用监督学习的方式捕获状态信息以表示状态,并将其应用于不同的任务中. Higgins et al[15]学习状态表示并应用于3D任务的实验中. Van Den Oord et al[7]提出CPC,通过无监督学习的方式学习图像的有效表示,并将CPC应用于强化学习A2C的训练过程,在全局可观测的部分Atari环境中验证了有关算法的性能,实验效果和利用传统状态表示的A2C相比,有所提升. Guo et al[8]在此基础上用非监督学习的方式学习局部可观测环境的信念表示,提取状态的关键信息,但仅用于信念状态表示,没有将其用于解决强化学习问题.

2 研究背景

2.1 对比预测编码 CPC[7]是一种使用无监督学习方式训练、能够从高维数据中提取对预测未来有用的表示信息的一种方法. 在预测高维数据时,如果直接学习重构数据中每个细节的条件生成模型,需要很大的计算量,并且容易忽略数据集中的上下文关系,即,为了提取数据x和上下文关系c之间的共享信息,直接对$p(x|c)$建模比较困难. 因此,CPC通过最大化x和c之间的互信息实现有关信息的提取,如式(1)所示:

$$I(x;c)=\sum_{x,c}p(x;c)\lg\frac{p(x|c)}{p(x)} \tag{1}$$

由于对$p(x|c)$建模比较困难,提出密度比的概念以衡量x_{t+k}和c_t之间的互信息,关系表示如式(2)所示:

$$f_k(x_{t+k},c_t)\propto\frac{p(x_{t+k}|c_t)}{p(x_{t+k})} \tag{2}$$

其中,$\propto$表示正比于,$f_k(x_{t+k},c_t)$表示上下文c_t的预测和未来真实值x_{t+k}的相似程度,其正比于未来真实数据与随机采样数据的概率之比.

CPC使用线性矩阵$W_1, W_2, \cdots, W_k$乘以c_t作为预测值,z_{t+k}^T为真实值,用向量内积来衡量相似度. 计算过程中使用式(3)表示相似度:

$$f_k(x_{t+k},c_t)=\exp(z_{t+k}^T W_k c_t) \tag{3}$$

CPC使用噪声对比估计[16,18]以及重要性采样[19]的方式估计概率之比. 使用基于噪声对比估计的损失又称InfoNCE,用来训练整个网络. 损失的计算如式(4)所示:

$$\mathcal{L}_N=-\mathbb{E}_X\left[\lg\frac{f_k(x_{t+k},c_t)}{\sum_{x_j\in X}f_k(x_j,c_t)}\right] \tag{4}$$

其中,$X=\{x_1,x_2,\cdots,x_N\}$是一组样本,(x_{t+k},c_t)可以视为正样本对,(x_j,c_t)可以视为负样本对. 优化该损失,也就是希望分子尽可能大,分母尽可能小,这满足对x和c之间互信息的要求,即正样本之间互信息更大,负样本之间互信息更小. 所以,优化该损失实际上就是最大化x_{t+k}和c_t之间的互信息.

2.2 基于动作的对比预测编码 基于动作的对比预测编码(CPC|Action)[8]对CPC进行拓展,通过引入动作对局部可观测环境的信念状态进行表示,使信念状态包含未来信息.

CPC|Action[8]将噪声对比估计作为学习分布的统计方法. 噪声对比估计通过区分来自数据分布的样本(正例)和来自另一分布的样本(负例)而学习,因此又被称为"从比较中学习". 实现此原理的一种简单方法是二进制分类,来自数据分布的样本被标记为正例,而来自另一个分布的样本被标记为负例,二进制分类的目标是训练区别样本的二进制分类器以学习数据分布上的信息,从而进行特征编码. 如式(5)、式(6)和式(7)所示,假设从概率密度为ρ^+的分布中采样N^+个样本$(o_i^+)_{i=1}^{N^+}$,从概率密度为ρ^-的分布中采样N^-个样本$(o_i^-)_{i=1}^{N^-}$. 训练一个二进制分类器f以最大化$\hat{J}(f)$,$\hat{J}(f)$为$J(f)$的经验估计,因此使用Jensen-Shannon散度[20]计算.

$$\hat{J}(f)=\frac{1}{N^+}\sum_{i=1}^{N^+}\lg\left(f\left(o^+\right)\right)+\frac{1}{N^-}\sum_{i=1}^{N^-}\lg\left(1-f\left(o^-\right)\right) \tag{5}$$

$$J(f)=\mathbb{E}_{o^+\sim\rho^+}\left[\lg\left(f\left(o^+\right)\right)\right]+\mathbb{E}_{o^-\sim\rho^-}\left[\lg\left(1-f\left(o^-\right)\right)\right] \tag{6}$$

$$\max_f J(f)=2D_{JS}\left(\rho^+\|\rho^-\right)-\lg 4 \tag{7}$$

CPC|Action通过循环神经网络表示的信念状态b_t,使用GRU(Gate Recurrent Unit)结合过去的动作预测特征$b^a=g\left(b_t,a_{t:t+f-1}\right)$,然后对预测特征进行分类,其损失计算如式(8)、式(9)和式(10)所示,其中,z_{t+1}表示正样本即真正的观测,z^-表示负样本即随机采样的观测.CPC|Action最终通过实验表明信念状态能够捕捉环境中的一些关键信息.

$$l^+=sigmoid\_cross\_entrop\left(h\left(b^a,z_{t+1}\right),1\right) \tag{8}$$

$$l^-=sigmoid\_cross\_entrop\left(h\left(b^a,z^-\right),0\right) \tag{9}$$

$$l=l^++l^- \tag{10}$$

3　基于对比预测编码的深度Q学习算法

主要面向全局可观测问题,深度强化学习的典型算法——深度Q学习算法[9]通过引入经验池和双神经网络结构训练有关策略,经验池的引入主要解决数据相关性以及非静态分布的问题,双神经网络的训练提高了算法的稳定性.对于局部可观测问题,深度循环Q学习算法[6]提出在深度Q学习算法的基础上增加循环神经网络,对历史信息进行整合以表示信念状态.但无模型的深度强化学习有关算法,例如深度循环Q学习,其信念状态通常不包含未来信息,决策过程中也没有考虑预测信息对策略的影响.

本文使用基于动作的对比预测编码算法对局部可观测环境信念状态进行表示,同时将预测特征作为一部分决策依据,基于深度Q学习算法,提出基于对比预测编码的深度强化学习算法.算法的模型结构分为记忆模块、特征预测模块、策略模块三部分,如图1所示.

由图1可见,t时刻的观测o_t经过记忆模块整合后得到当前的信念状态表示b_t,特征预测模块

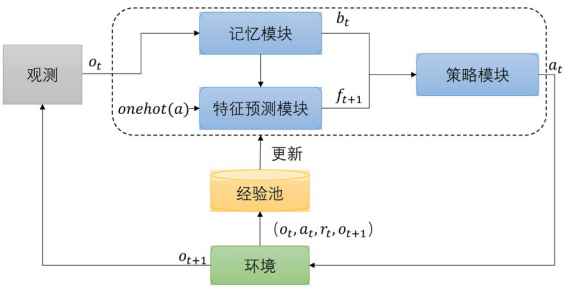

图1　算法的框架结构

Fig. 1　The framework of our method

根据b_t与一步动作预测未来特征f_{t+1}.b_t与f_{t+1}共同作为策略模块的输入,两者共同决定当前状态下策略模块的输出动作a_t.智能体采取策略输出的动作与环境进行交互,环境根据智能体的动作给出新的观测o_{t+1}与动作的奖励值r_t,得到新的经验$\left\{o_t,a_t,r_t,o_{t+1}\right\}$,将其保存到经验池中.网络训练过程中,不断从经验池中抽取训练数据更新网络参数.

3.1　记忆模块　记忆模块可对历史信息进行整合,记忆模块的引入可以将局部可观测环境马尔可夫决策过程近似看作马尔可夫决策过程.在局部可观测环境下,由于不满足马尔可夫性,信念状态表示就十分重要.算法在记忆模块使用了循环神经网络,通过对过去一定步数的经历进行总结来实现对历史信息的统计,以表示当前的信念状态.记忆模块首先对当前观测进行编码处理,编码器由卷积神经网络组成,记观测为o_t,编码器为ϕ,编码后的特征为z_t,如式(11)所示.循环神经网络使用具有门控单元的循环神经网络,记为f,其中包含的参数用θ_f表示,信念状态b_t由上一步信念状态b_{t-1}、特征z_t以及动作a_{t-1}经过网络f得到,如式(12)所示.

$$z_t=\phi\left(o_t\right) \tag{11}$$

$$b_t=f\left(b_{t-1},z_t,a_{t-1}\middle|\theta_f\right) \tag{12}$$

3.2　特征预测模块　特征预测模块使用信念状态预测后续特征,通过对特征进行分类的方式为整个网络添加噪声对比损失.此外,特征预测模块为策略模块提供了下一步可能的预测特征.

如上所述,记忆模块输出为信念状态表示b_t,特征预测模块根据b_t和动作的独热编码

$onehot(a)$预测后续状态的特征. 算法的特征预测模块结构采用循环神经网络构建. 训练过程中采用噪声对比估计损失,其表示如式(8)、式(9)和式(10)所示. 训练过程中,特征预测模块根据b_t与动作预测下一步观测的编码特征f_{t+1},通过特征分类的方式确定是否准确. 特征预测模块与决策过程同时训练. 记特征预测模块内的循环神经网络为g,参数为θ_g,其过程如式(13)所示:

$$f_{t+1}=g\left(b_t,onehot(a)\middle|\theta_g\right) \tag{13}$$

3.3 策略模块 策略模块的作用是根据过去的经历与预测特征进行动作选择,本算法中是计算当前状态下对应的最大的Q的动作.

策略模块根据信念状态b_t和预测特征f_{t+1}选择最大化奖励值对应的动作. 记忆模块使b_t包含当前状态的历史信息,通过特征预测模块网络更新,b_t同时包含了部分未来信息. 策略模块利用信念状态b_t和预测特征f_{t+1}两部分包含的信息,求出当前状态下的Q. 本文算法中策略模块采用双神经网络结构训练,由两层全连接层组成,与深度Q学习算法相似,求得当前局部观测对应的真实状态s下的每个动作a对应的$Q(s,a)$. 记两层全连接层为f_π,参数为θ_π,过程如式(14)所示:

$$Q(s,a)=f_\pi\left(b_t,f_{t+1}\middle|\theta_\pi\right) \tag{14}$$

智能体在选取动作时,为了保证对环境的探索率,选择ε-贪心算法. 智能体在选择动作时以概率ε随机选取动作,以概率1−ε采取算法给出的策略.

3.4 网络结构 算法训练过程的网络结构如图2所示. 黄色的conv表示编码操作,采用卷积神经网络实现. GRU(蓝色)为记忆模块中的门控单元循环神经网络f,整合了一定步数o_t之前的历史训练数据得到当前t时刻的信念状态表示b_t. GRU(橙色)是特征预测模块的门控单元循环神经网络g,输入为t时刻的信念表示b_t和t时刻要采取的动作a_t,输出为预测的$t+1$时刻预测特征表示f_{t+1}. MLP是多层感知机h,由多层全连接层组成,用于训练过程中判断预测特征f_{t+1}是否准确. 一个MLP的输入为下一时刻的预测特征表示f_{t+1}和正样本z_{t+1},另一个MLP的输入为下一

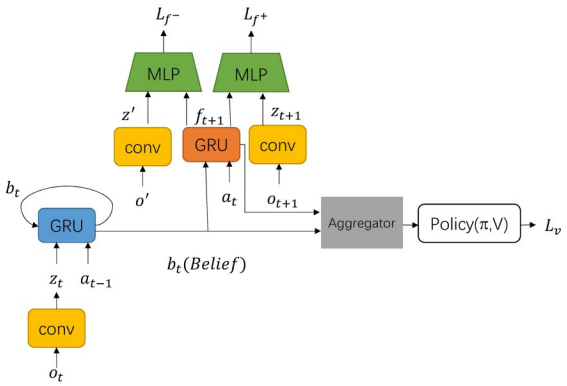

图2 算法的计算过程总览

Fig. 2 The pipeline of our method

时刻的预测特征表示f_{t+1}和负样本z'. 正样本z_{t+1}表示$t+1$时刻的观测o_{t+1}的特征表示,负样本z'则表示从经验池中随意选取的观测o'的特征表示,相应的损失函数$L_{f_t^+}$和$L_{f_t^-}$如式(15)和式(16)所示:

$$L_{f_t^+}=sigmoid\_cross\_entrop\left(h\left(f_{t+1},z_{t+1}\right),1\right) \tag{15}$$

$$L_{f_t^-}=sigmoid\_cross\_entrop\left(h\left(f_{t+1},z'\right),0\right) \tag{16}$$

$$L_f=L_{f_t^+}+L_{f_t^-} \tag{17}$$

优化噪声对比估计损失L_f,输入正样本z_{t+1}和预测特征f_{t+1}使MLP预测为1,输入负样本z'和预测特征f_{t+1}使MLP预测为0. 经过上述过程后,得到预测特征f_{t+1},将预测特征f_{t+1}与当前信念状态b_t聚集到一起作为策略网络的输入,共同作用于策略选择. 策略部分采用深度Q学习算法[9]的训练方式,即包含缓冲池与双神经网络(目标网络$\widehat{f_\pi}$参数为θ_π^-,训练网络f_π参数为θ_π,经过一定步数同步参数). 对于经验$\left(o_t,a_t,r_t,o_{t+1}\right)$,根据式(18)计算估计的Q,根据式(19)计算时序差分(Temporal Difference,TD)损失. 算法采用端到端的训练方式,网络结构损失如式(20)所示,其中,λ表示噪声对比估计损失占比. 算法训练过程中,计算损失函数梯度,不断更新各个模块的参数.

$$y_t=\begin{cases}r_t\\ r_t+\gamma\max_a\widehat{f_\pi}\left(b_t,f_{t+1}\middle|\theta_\pi^-\right)\end{cases} \tag{18}$$

$$L_v=\left(y_t-f_\pi\left(b_t,f_{t+1}\middle|\theta_\pi\right)\right)^2 \tag{19}$$

$$L=L_v+\lambda L_f \tag{20}$$

算法的具体过程如下所示.

算法　基于对比预测编码的深度Q学习算法

输入:Belief GRU f, Action GRU g, t时刻的信念状态 b_t, t时刻的预测特征 f_{t+1}

1. 初始化一个记忆池 D
2. 随机初始化训练网络 f_π,随机初始化目标网络 $\widehat{f_\pi}$
3. for $episode=1,2,\cdots,M$ do
4. 接收一个初始观测 o_1,并编码 $z_1=\phi(o_1)$
5. for $j=1,\cdots,T$ do
6. 以 ε 的概率随机选择一个动作 a_j
7. 以 $1-\varepsilon$ 的概率选择 $a_j=\max_a\left(f_\pi\left(f_{j+1},b_j\middle|\theta_\pi\right)\right)$, 其中,$f_{j+1}$ 根据式(13)计算,b_j 根据式(12)计算
8. 将 (o_j,a_j,r_j,o_{j+1}) 存入 D
9. D 中采样一个长度为 N 的样本
10. for $t=1,2,\cdots,N-1$ do
11. 随机选取观测 o'
12. $b_t=f(b_{t-1},z_t,a_{t-1})$
13. $f_{t+1}=g(b_t,onehot(a))$
14. 根据式(20)计算损失
15. end for
16. 最小化累积损失 L 更新网络参数
17. 每 C 步同步目标网络 f_π'
18. end for
19. end for
20. End

4　实验

在MiniPacman环境中四种不同的任务下进行实验,任务不同,需要学习的最优策略也不相同.实验中本文算法用FutureFeature表示.由于是基于无模型的方法,深度循环Q学习算法作为在高维局部可观测环境中的通用无模型RL方法,被作为首要的对比对象.DRQN仅使用 b_t 作为策略网络输入,且仅使用策略模块的TD损失 L_v 更新网络参数.

为了验证本文方法与框架的有效性,还进行了部分消融实验.首先,与无预测特征的方法进行对比.该方法用NoFuture表示,在训练过程中加入 L_f 损失计算,但在决策过程中不加入预测特征 f_{t+1},即策略模块的输入只有 b_t,由于该方法不加入预测特征,因而可以验证额外使用预测特征进行值函数估计是否对决策有影响.最后,为了验证本文算法的性能是否是因为引入额外的模型而产生的,与无预测特征损失方法进行了对比.该方法用NoFutureLoss表示,在梯度更新时没有将噪声对比估计损失 L_f 加入计算,仅使用策略模块的TD损失 L_v 更新网络参数.

4.1　实验环境　实验环境为Deepmind团队设计开发的迷你网格游戏吃豆人(MiniPacman)[17],是迷宫的游戏画面.迷宫的全局状态是一个RGB的彩色图像,使用大小为15×19、通道数为3的图像表示.MiniPacman环境是一个迷宫游戏,其目标是智能体在迷宫中行走完成一些任务,以得到更高的奖励.如图3所示,绿色方块为智能体,红色方块为幽灵,浅蓝色方块为能量药片,蓝色区域为食物,白色区域为障碍,黑色表示食物已经吃掉.图4描述了智能体连续两步的观测变化.智能体(绿色)向右连走两步吃掉食物(蓝色),观测发生的变化.MiniPacman环境下可以定义不同的目标奖励,从而产生多种任务,本实验使用其中四种任务,分别为Regular,Avoid,Ambush和Rush.

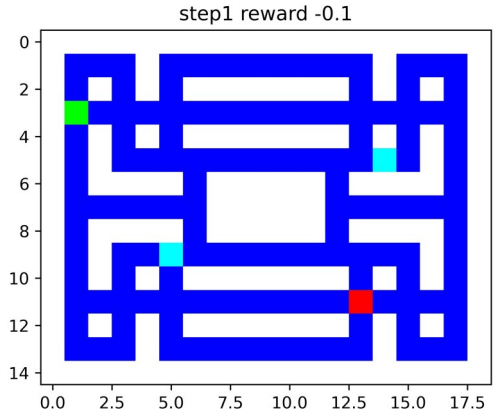

图3　MiniPacman游戏帧的图例

Fig. 3　Snapshot of MiniPacman

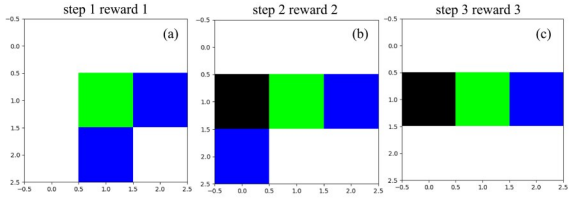

图4　MiniPacman中智能体连续观测变化的图例

Fig. 4　Illustrastion of consecutive observations of RL agent in MiniPacman

Regular下智能体要尽可能多地吃到食物,Avoid下智能体要尽可能避免被幽灵抓住,Ambush下智能体要尽可能捕食幽灵,Rush下智能体要尽可能多地吃能量药片.

4.2 实验设置 算法采用四帧连续时刻的观测图像作为算法的观测输入,即观测大小维度为$3\times3\times12$. 记忆模块和特征预测模块的conv编码操作均由卷积神经网络构成,将观测编码为长度为16的特征向量. 门控单元循环神经网络f的隐藏层个数为16,特征预测模块的门控单元循环神经网络g的隐藏层个数为512. 策略模块将b_t, f_{t+1}拼接后经过两个全连接层,第一层全连接层输出单元个数为256,第一层的输出作为第二层的输入,第二层的输出单元个数为环境动作空间对应的动作数,对应每一个动作下的Q. 参数设置,式(18)中$\gamma=0.99$,式(20)中$\lambda=0.1$,每隔10^4步对目标网络进行更新,经验池大小为10^6. 进行采样后,每一步都对网络进行训练,每500步进行一次评估,评估即测试当前的网络在环境中做决策一个完整的回合后能取得的奖励. 所有实验均在装备GeForce RTX 2080ti显卡、内存32 G、CPU 3.60 GHz的计算机上运行.

4.3 实验结果与分析 实验结果如图5至图8所示,采用任务完成奖励作为评价指标. 由总体实验结果可以看出,对于四种不同任务,本文提出的基于对比预测编码的深度Q学习算法,即图中的FutureFeature,在最终得分表现上均优于其他三种对比算法.

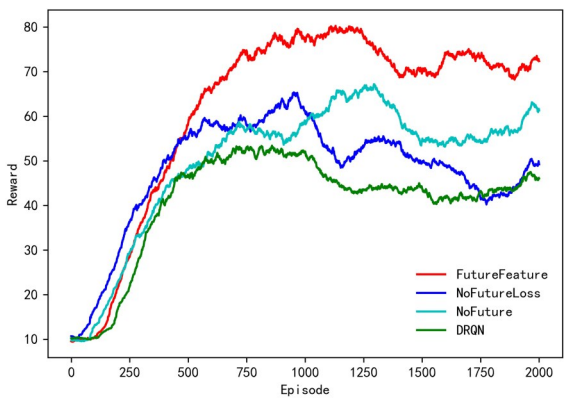

图5 Regular任务下的训练得分曲线

Fig. 5 Performance in Regular task

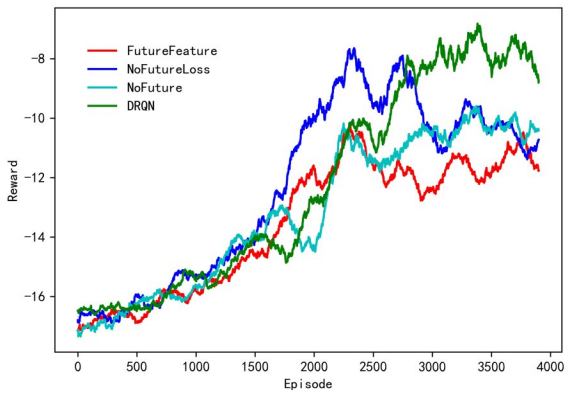

图6 Avoid任务下的训练得分曲线

Fig. 6 Performance in Avoid task

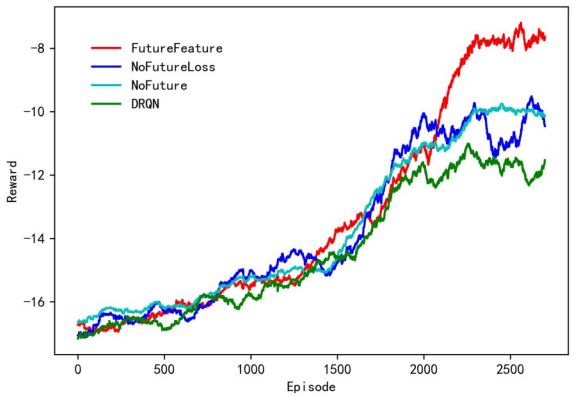

图7 Ambush任务下的训练得分曲线

Fig. 7 Performance in Ambush task

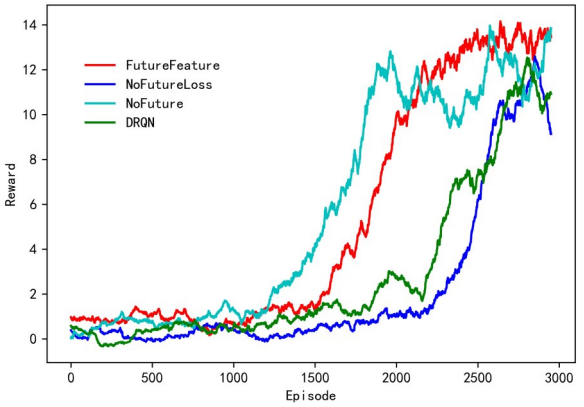

图8 Rush任务下的训练得分曲线

Fig. 8 Performance in Rush task

具体地,由于MiniPacman的任务特性,智能体需要尽可能存活更长的时间以达成各自任务的目标,如吃到更多的食物、抓到更多的幽灵等. 而幽灵这一不可控因素是决定游戏得分的关键之一,因为在大部分设定下,被幽灵抓住意味着回合

结束,这要求智能体需要及时预判游戏中幽灵的走位来调整自身的策略.

DRQN方法仅根据信念状态进行决策,既没有依据对比预测编码辅助表征学习,也没有使用任何的预测信息辅助决策,所以在几乎所有的任务中的表现都是最差的.

NoFutureLoss方法与本文提出的方法相比,不将任何噪声对比估计预测损失加入网络中训练,即完全使用TD损失更新策略与表征模块,这使NoFutureLoss方法在功能上无异于增加了模型参数量的DRQN,其用于决策和值估计的表征同样不具备任何的未来信息,因此在所有的任务上表现的最终性能仅优于DRQN. 这也从消融实验的角度验证了本文算法的性能提升并非来自模型参数量的增加.

最后,NoFuture方法和本文提出的方法相比,没有加入预测特征辅助决策,但仍然会使用对比预测损失辅助表征训练,因此在本质上NoFuture方法等价于CPC|Action算法. 虽然对比预测损失的加入可以有效地让算法学习的信念状态包含一定程度的未来信息,但本文方法能进一步结合预测的特征,提高智能体在决策时考虑的信息量. 通过实验结果不难发现,NoFuture方法在大部分任务中的最终性能得分均不如本文方法,证明本文算法使用预测特征辅助决策可以带来有效的性能提升. 此外,令人感兴趣的是,在Rush任务中可以看到NoFuture方法在收敛速度上具有一定优势,可能是因为NoFuture方法在决策时需要使用的信息较少,在Rush这样简单的任务中只需要使用由TD损失和对比预测损失训练的表征就足以快速达到一定的策略效果.

5　结论

本文提出一种基于对比预测编码的深度Q学习算法,利用对比预测编码对未来信息的预测能力,并通过结合无模型强化学习算法,避免了有模型强化学习算法建模困难、所建模型存在潜在不准确问题以及利用模型进行规划计算量大等缺点,同时实现了在无模型强化学习框架中利用未来信息的能力. 与多种不同算法进行比较,并在多个任务中进行验证,证明使用包含未来信息的信念状态和预测特征辅助决策能有效提高智能体的决策效果.

参考文献

[1] Jaderberg M, Czarnecki W M, Dunning I, et al. Human-level performance in 3D multiplayer games with population-based reinforcement learning. Science, 2019, 364(6443): 859−865.

[2] 范家伟,张如如,陆萌,等. 深度学习方法在糖尿病视网膜病变诊断中的应用. 自动化学报, 2021, 47(5): 985−1004. (Fan J W, Zhang R R, Lu M, et al. Applications of deep learning techniques for diabetic retinal diagnosis. Acta Automatica Sinica, 2021, 47(5): 985−1004.)

[3] 孙辉辉,胡春鹤,张军国. 移动机器人运动规划中的深度强化学习方法. 控制与决策, 2021, 36(6): 1281−1292. (Sun H H, Hu C H, Zhang J G. Deep Reinforcement Learning for motion planning of mobile robots. Control and Decision, 2021, 36(6): 1281−1292.)

[4] Kaelbling L P, Littman M L, Cassandra A R. Planning and acting in partially observable stochastic domains. Artificial Intelligence, 1998, 101(1−2): 99−134.

[5] James M R, Singh S. Learning and discovery of predictive state representations in dynamical systems with reset//Proceedings of the 21st International Conference on Machine Learning. Banff, Canada: ACM, 2004: 53.

[6] Hausknecht M, Stone P. Deep recurrent Q-learning for partially observable MDPs. 2015, arXiv: 1507.06527v4.

[7] Van Den Oord A, Li Y Z, Vinyals O. Representation learning with contrastive predictive coding. 2018, arXiv: 1807.03748.

[8] Guo Z D, Azar M G, Piot B, et al. Neural Predictive Belief Representations. 2018, arXiv: 1811.06407.

[9] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. Nature, 2015, 518(7540): 529−533.

[10] Konda V R, Tsitsiklis J N. Actor-critic algorithms. Advances in Neural Information Processing Systems, 2000(12): 1008−1014.

[11] Nguyen T, Luu T M, Vu T, et al. Sample-efficient reinforcement learning representation learning with curiosity contrastive forward dynamics model. 2021, arXiv:2103.08255.

[12] Oh J, Singh S, Lee H. Value prediction network. 2017, arXiv:1707.03497v2.

[13] Karkus P, Hsu D, Lee W S. QMDP-Net: Deep learning for planning under partial observability// Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, CA, USA: Curran Associates Inc., 2017: 4697—4707.

[14] Dosovitskiy A, Koltun V. Learning to act by predicting the future. 2016, arXiv:1611.01779.

[15] Higgins I, Pal A, Rusu A A, et al. DARLA: Improving zero-shot transfer in reinforcement learning//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: PMLR, 2017:1480—1490.

[16] Gutmann M, Hyvärinen A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models//Proceedings of the 13th International Conference on Artificial Intelligence and Statistics. Sardinia, Italy: JMLR.org, 2010:297—304.

[17] Racanière S, Weber T, Reichert D, et al. Imagination-augmented agents for deep reinforcement learning// Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, CA, USA: Curran Associates Inc., 2017: 5694—5705.

[18] Jozefowicz R, Vinyals O, Schuster M, et al. Exploring the limits of language modeling. 2016, arXiv:1602.02410.

[19] Bengio Y, Senecal J S. Adaptive importance sampling to accelerate training of a neural probabilistic language model. IEEE Transactions on Neural Networks, 2008, 19(4):713—722.

[20] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 2014:2672—2680.

(责任编辑 杨可盛)