一种深度强化学习空间关系与记忆融合方法研究

刘卉玲[1)] 刘 鹏[1)] 白辰甲[2)]

[1)](哈尔滨工业大学计算机科学与技术学院 哈尔滨 150008)

[2)](上海人工智能实验室 上海 200232)

摘 要 深度强化学习结合了深度学习在视觉上强大的感知能力来解决复杂环境的序列决策问题,但是由于采样效率低,对于复杂高维数据输入,学习其重要特征较为困难.为了从序列样本中更有效地提取信息,本文提出在深度强化学习中融合空间关系推理和记忆推理(Spatial Relationship Reasoning and Memory Reasoning,SRRMR)的模型结构.模型分为空间关系推理和记忆推理两部分,空间关系推理使用注意力机制作为空间关系学习方法隐式地推理任意两个实体间的关系,注意力机制中的查询向量融合了记忆推理的内容;记忆推理将输入图像的特征和关系作为记忆的输入,利用自注意力与记忆组成部分进行推理和交互,并将交互的结果存储在记忆单元中,使得记忆存储单元融合了空间信息与记忆信息.SRRMR模型在不同种类的Atari游戏中进行了训练和验证,结果表明,空间关系推理与记忆推理的融合在7/15个游戏环境中以更少的交互次数收敛到更好的结果,记忆推理网络在12/15个游戏中获得提升,提升智能体学习效率,更高效地利用序列中的样本,提高了强化学习的样本利用率.

关键词 空间关系推理;记忆推理;深度强化学习;注意力机制;状态表示

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2023.00814

Research on a Fusion Method of Spatial Relationship and Memory in Deep Reinforcement Learning

LIU Hui-Ling[1)] LIU Peng[1)] Bai Chen-Jia[2)]

[1)](*School of Computer Science and Technology*, *Harbin Institute of Technology*, *Harbin* 150008)

[2)](*Shanghai AI Laboratory*, *Shanghai* 200232)

Abstract Deep reinforcement learning combines the powerful visual perception of deep learning to solve the sequential decision-making problem in complex environments. However, due to the low sampling efficiency, it is difficult to learn the important features of complex high-dimensional data input. In order to extract information from sequence samples more effectively, this paper proposes a model structure integrating Spatial Relationship Reasoning and Memory Reasoning (SRRMR) in deep reinforcement learning. The model is divided into two parts: spatial relation reasoning and memory reasoning. Spatial relation reasoning uses attention mechanism as a spatial relation learning method to implicitly infer the relationship between any two entities, and the query vector in attention mechanism integrates the content of memory reasoning; Memory reasoning takes the characteristics and relations of the input image as the input of memory, uses the self attention mechanism to reason and interact with the memory components, and stores the interactive results in the memory unit, so that the memory storage unit integrates spatial information and memory information. The SRRMR model has been trained and verified in different

收稿日期:2022-06-23,在线发布日期:2022-11-25.本课题得到国家自然科学基金重点项目(No.51935005)、基础科研项目(No.JCKY20200603C010)、黑龙江省自然科学基金(No.LH2021F023)资助、黑龙江省科技计划项目(No.GA21C031)资助.**刘卉玲**,博士研究生,主要研究领域为深度强化学习,E-mail:19b903039@stu.hit.edu.cn.**刘 鹏**(通信作者),博士,教授,中国计算机学会(CCF)会员,主要研究领域为机器学习、模式识别.E-mail:pengliu@hit.edu.cn.**白辰甲**,博士,青年研究员,主要研究领域为深度强化学习.

Atari games. The results show that the integration of spatial relationship reasoning and memory reasoning converges to better results with less interaction times in 7/15 game environments, and the memory reasoning network is improved in 12/15 games, improving the learning efficiency of agents, making more efficient use of samples in sequences, and improving the sample utilization rate of reinforcement learning.

Keywords spatial relational reasoning; memory reasoning; deep reinforcement learning; attention mechanism; state representation

1 引 言

深度强化学习[1-2](Deep Reinforcement Learning, DRL)结合了深度学习[3-4]的感知能力和强化学习[5]的决策能力,可直接根据输入的图像进行控制,是一种更接近人类思维方式的人工智能,在许多具有挑战性的问题领域取得了显著的成绩,如游戏[6-7]、机器人控制[8]、计算机视觉[9]、自然语言处理[10]和医疗[11]等.这很大程度上是因为深度强化学习能够灵活地学习和利用观测和奖励信号,但也面临许多挑战,如样本效率低,需要进行长时间的训练.为了克服这种局限性,许多方法尝试在深度强化学习中引入关系归纳偏置[12-13]或者一些简单的记忆结构[14-15],尽管这些方法在一定程度上提高了样本效率,但仅限于相对简单的任务和数据,难以学习到状态丰富的表示.本文尝试对深度强化学习中的空间关系方法和记忆结构进行融合,使用关系归纳偏置方法学习空间实体和关系,其查询向量由记忆网络产生,外部记忆增强神经网络存储空间关系与记忆推理的结果,高效的利用已知的信息来提高样本利用率.

深度强化学习通常使用卷积神经网络[16]对当前图像输入的状态进行表示,指导智能体决策,这种方法的关系归纳偏差较弱,需要堆叠多层来获取对全局的理解,层数过高策略往往对环境过拟合,数据效率低,无法获取问题高层次的理解.空间关系学习能够提取图像中的抽象概念,并学习这些概念之间的关系,加速学习进程,Mott等人[12]和Zambaldi等人[13]在深度强化学习中引入以实体和关系为中心的状态表示的关系归纳偏差方法,通过低层卷积网络处理原始视觉输入数据获得实体集,使用注意力机制计算全局实体重要性程度,隐式地在空间中进行关系推理,提升智能体学习抽象概念及关系能力,处理复杂任务.

深度强化学习中引入的记忆网络如长短期记忆网络[17],由于记忆能力受可训练参数的数量影响较大,导致存储的内容也较少.记忆增强神经网络设置外部存储结构和读写机制,记忆能力与可训练参数无关,相对记忆能力更强,更容易建模长距离依赖关系,形式也更加灵活和丰富,如神经图灵机[14]和记忆网络[15].Loynd等人[18]提出一种利用自注意力学习记忆组成部分之间关系的方法增加了记忆的交互,使得记忆增强神经网络既能记忆又能推理.

深度强化学习在与环境进行交互时,空间关系推理通常会提取一帧或多帧输入图像的信息指导智能体进行决策,如Zambaldi等人[13]只分析了单帧物体空间实体之间的关系.单次观测不足以描述潜在的环境状态,过去的观测同样也与决策相关,缺少输入序列之间的联系,即记忆问题[19],导致需要大量交互才能学习到较为完整的环境状态.深度强化学习记忆以隐藏状态或者记忆片段等形式进行存储,但是记忆存储的内容通常是卷积神经网络对历史观测提取的表示[14-15,19],或者是人为指定的块因素[18],缺乏对空间关系的理解和主动感知任务相关的需要,学习效率低.为了更高效地利用样本信息,提高样本效率,本文提出一种在深度强化学习中融合空间关系推理和记忆推理(Spatial Relation Reasoning and Memory Reasoning, SRRMR)的模型结构,能够从可获得的数据中更高效的提取信息,提高深度强化学习智能体学习效率和学习能力.空间关系推理利用注意力机制提取单个时间步上输入图像特征之间的关系,其中注意力机制的查询向量融合了记忆推理的内容,使得智能体根据记忆中的先验知识主动感知空间关系.记忆推理使用自注意力学习空间关系与记忆信息的关系,并将记忆推理结果存储在外部记忆单元中,记忆网络除了记忆功能外还添加了推理能力.二者充分利用已学到的知识,高效地提取相关信息,最后整合空间关系推理和记忆推理的结果作为输出,生成强化学习策略和值函数.为了验证模型的有效性,本文在不同类型的Atari 2600上进行实验,结果显示,SRRMR模型在7/15个游戏环境

中提升了学习效率，在其他游戏中也具有竞争力，记忆推理网络在12/15个游戏中获得提升，表明SRRMR模型提升了收敛速度和学习效果，能够在相对较少的交互次数内学习到更好的策略，提高了强化学习智能体的样本利用率.

2 相关工作

本节主要介绍强化学习、空间关系推理方法以及记忆网络方法.

2.1 强化学习

强化学习[5]主要由智能体、环境、状态、动作、奖励组成，智能体和环境通过状态、动作和奖励进行交互.强化学习最为核心的是马尔可夫决策过程，组成元素有(S,A,P,R)，包含动作$a\in A$，状态$s\in S$，奖励函数$r=R(s,a)$，状态转移概率$P(s'|s,a)$.本文使用a_t，s_t和r_t分别表示时间步t的动作，状态和奖励.轨迹由一系列状态，动作和奖励序列组成，$\tau=(s_0,a_0,r_0,s_1,a_1,a_1,r_1,\cdots,s_T,a_T,r_T)$.强化学习通过最大化期望回报学习一种最优策略.本文研究的是无模型(model-free)的离策略(off-policy)强化学习算法，智能体通过与环境交互产生经验，并通过经验池采样的方式训练生成最优值函数，如DQN[20].还有一些方法使用两个网络分别学习策略"Actor"和值函数"Critic"，如DDPG[21]，A3C[22]等.本文应用分布式强化学习算法重要性加权Actor-Leraner架构[23](Importance Weighted Actor-Learner Architecture，IMPALA)训练智能体，其中参与者使用本地策略μ与环境交互n个时间步收集经验轨迹，本地策略μ来自学习者的最新策略π，隔一段时间更新一次本地策略，将收集n个时间步的经验轨迹$(s_0,a_0,r_0,\cdots,s_{n-1},a_{n-1},r_{n-1})$和策略分布$\mu(a_t|x_t)$发送给学习者；学习者根据多个参与者发送的轨迹连续地更新策略π.这样导致学习者策略π比参与者策略μ更新更快，通过引入V-trace修正策略间的差异估计值函数$v_s=V(s)+\sum_{t=0}^{n-1}\gamma^t\left(\prod_{i=0}^{t-1}c_i\right)\rho_t(r_t+\gamma V(s_{t+1})-V(s_t))$，其中$\rho_t=\min(\bar{\rho},\pi(a_t|s_t)/\mu(a_t|s_t))$，$c_i=\min(\bar{c},\pi(a_i|s_i)/\mu(a_i|s_i))$，$\bar{\rho}\geqslant\bar{c}$.由于可以并行独立执行计算，大大提高了系统的吞吐量和数据利用率，因此可以构建更深层网络结构.

2.2 空间关系推理方法

空间关系是指各实体空间之间的关系，包括拓扑、顺序和度量等空间关系.空间关系推理方法有很多，如图神经网络[24]、符号逻辑[25]、关系网络[13]等.图神经网络(Graph Neural Network，GNN)使用包含节点和边的图形结构，能够学习任意两个节点之间的关系.Kurin等人[26]提出使用Transformer学习图网络肢体节点关系的机器人连续控制方法.Jiang等人[27]提出使用网格到图的方法将离散二维观测的网格结构映射为关系图，再通过关系图卷积神经网络对关系图中节点之间关系进行推理生成动作分布.由于网格结构需要重新对环境进行设计并简化环境状态表示，将原始输入图像转换成10×10的网格结构，导致环境中一些信息丢失.符号逻辑通过专家人工指定一些符号参与逻辑推理过程，表示一组对象之间的关系.神经逻辑器[28](Neural Logic Machines，NLM)使用逻辑程序设计作为符号处理器，建立一阶逻辑应用于块世界和网格世界强化学习任务.Garnelo等人[25]尝试将符号逻辑融入深度学习框架.关系网络使用物体检测网络从图像中提取物体，经过关系模型得到关系似然度.Che[29]提出将视觉关系嵌入网络识别物体并检测之间的关系，用于生成图像的段落描述.关系深度强化学习[12-13]以原始图像作为输入，使用神经网络提取特征作为实体，采用注意力机制隐式地推理空间中实体之间成对的关系，结合结构感知和关系推理提高深度强化学习方法的效率.

图神经网络直接根据图进行节点之间的解释和推理，但其对于非结构化场景没有生成图的通用方法，节点的变化和不规则性导致计算困难无法自适应改变.符号逻辑可以自然地处理推理中的符号规则，然而通常做法是将归纳偏差硬编码在网络结构中，过于依赖人力不易进行修改和自主学习.关系学习通过端到端的方式训练深度网络学习图像中的实体，使用注意力网络推断它们之间的关系，解决以上两种方法面临的问题.但是这些关系网络通常只根据一帧或者多帧图像的关系进行决策，不能重复使用输入序列中预先计算的关系，不能对学习的关系进行存储，缺乏对关系的长期记忆.

2.3 记忆网络方法

人工智能对记忆模型的研究主要基于物体或事件的存储和检索，如带有循环神经网络的强化学习记忆模型[30-31]，使用隐藏状态向量作为记忆单元，整合时间上的观测更好地估计潜在状态，这些方法受可训练参数数量上的影响，存储的记忆较小.此外还有外部记忆神经网络，解决了可训练参数与记忆能

力之间的相关性，记忆能力更强更容易建模长距离依赖关系.Graves等人[14]提出一种可微的外部存储器称为神经图灵机(Neural Turing Machine，NTM)，使用注意力机制改进记忆单元的检索，可以学习复制和反转等算法.Zaremba等人[32]扩展了神经图灵机的寻址机制，模型不可微，使用梯度策略训练模型.Oh等人[33]提出的是神经图灵机在强化学习中的一个具体应用，利用额外记忆单元在过去观测中实现更复杂的寻址方案.Parisotto等人[34]开发了一个自适应写操作存储系统，使用具有空间结构的2D内存图像存储长时间内关于环境的任意信息.记忆网络[15]存储所有输入，根据问题检索相关记忆，引入具有独立可读写操作的记忆模块保存场景信息，可以灵活的保留输入信息，结构简单可塑性强.基于记忆神经网络的扩展有很多，如端到端记忆网络[35]、动态记忆网络[36]、键值记忆网络[37]、分层记忆网络[38]等.这些记忆网络支持长期检索，但是缺乏对记忆组成部分之间关系的表示.Loynd等人[18]提出的记忆推理模型，利用自注意力建模记忆之间的关系，通过记忆的交互形成记忆组成部分之间的推理，并根据推理结果更新记忆单元，使得记忆网络除存储检索外还具备记忆推理能力.

记忆增强神经网络的研究主要集中于记忆的存储、检索和交互的形式，而记忆网络存储的内容大多是输入图像的状态表示，由于记忆网络存储的是输入序列，为了减少计算量往往会对状态的表示进行压缩，导致存储的内容粗糙且无法控制.

本文在强化学习中将空间关系推理和记忆推理网络进行融合，前者提高神经网络的关系推理能力，后者增强系统长期记忆能力和记忆推理能力.空间关系推理提取图像中的抽象概念，并学习这些概念之间的关系.记忆推理学习空间关系与记忆之间的关系，并将记忆推理的结果进行存储.此外本文还在空间关系推理中保留输入状态的表示，丰富了记忆学习的内容，加快智能体学习速率.空间关系推理中的查询融合了前一时刻记忆网络的输出以及空间关系推理的结果，即每一次空间关系推理都包含了前一时刻的空间关系和记忆网络中预先计算的空间关系序列之间的联系.二者的融合使得智能体充分利用样本信息，提高样本效率.

3　空间关系推理与记忆推理模型

本节对模型的整体和各个模块的设计进行详细描述，3.1节介绍了模型的整体结构；3.2节介绍了空间关系推理模块的方法；3.3节介绍了记忆单元的存储、推理及更新方式；3.4节介绍了模型的输出；3.5节为模型算法的描述.

3.1　SRRMR模型的总体描述

为了提高深度强化学习的样本利用率，本文提出一种深度强化学习空间关系推理和记忆推理融合方法，模型如图1所示，主要由空间关系推理(Spatial Relation Reasoning，SRR)和记忆推理(Memory Reasoning，MR)两部分构成，其中黄色块表示神经网络.空间关系推理的输入来自输入的图像，其内部有两条通路分别是状态表示和关系提取.状态表示是卷积神经网络提取的输入图像特征后经过全连接网络压缩得到.关系提取是将卷积神经网络提取的图像特征作为输入，使用注意力机制“隐式”推理图像特征之间的关系，同样地，使用多层感知器将关系压缩成易于记忆推理存储的形式.两条通路分别提取了图像的特征以及特征之间的关系，丰富了空间

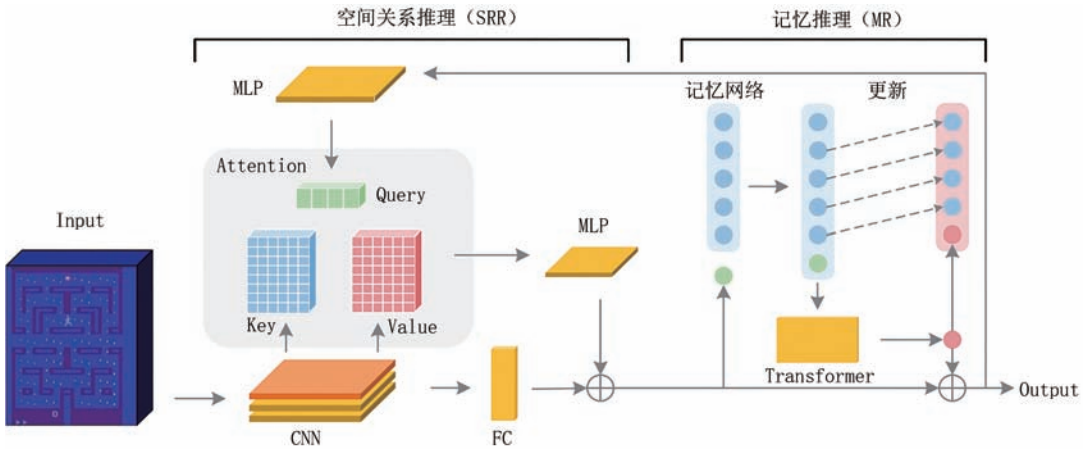

图1　SRRMR模型整体结构

信息的表示，并且将两个向量进行压缩合并后输送给记忆推理，这样做的好处是减小记忆推理中的存储空间与计算量.空间关系推理的输出包含图像的特征以及特征之间的关系.

记忆推理将输入序列中的图像特征以及特征之间的关系作为输入，利用自注意力[39]推理出当前时刻空间信息的输入(绿色球)与记忆片段之间(蓝色球)的关系，将其结果作为记忆推理的输出(红色球).记忆推理的结果还作为新的记忆存储在记忆单元中，记忆网络的更新在3.3节进行详细介绍.本文将新的记忆片段与空间关系推理的输出叠加作为模型的输出，此外还将当前时刻得到的空间向量和新的记忆片段作用于下一时刻空间关系提取中查询向量的生成.空间关系推理与记忆推理二者通过不断学习和重复使用预先计算关系，能够提高智能体逻辑推理能力和记忆推理能力，提高样本利用率.整个模型将空间关系推理和记忆推理分别提取的空间信息与记忆信息相融合生成强化学习的策略和值函数.以下对两种推理结构进行详细描述.

3.2 空间关系推理(SRR)

传统深度强化学习通常使用卷积神经网络或者循环神经网络提取输入的特征信息，缺乏对空间关系的理解[30].还有一部分研究考虑在深度强化学习中增强关系学习的能力[13]，往往忽视了低层图像特征中的信息.SRR在空间关系学习的基础上保留了低层图像特征的表示，由图1可以看出，空间关系推理中有两个分支，其中一个为卷积神经网络提取的低层图像特征的表示，另一个为利用注意力机制推理图像特征之间关系的学习.这样设置的目的在于丰富空间信息的表示以便后续记忆网络的学习，为了便于这些信息在记忆网络中进行存储和计算，SRR分别使用全连接网络和多层感知器网络对图像特征和特征之间的关系进行压缩，然后将两个向量叠加在一起共同组成空间关系推理的输出.本文在每一层卷积网络后使用批标准化(Batch Normalization)和非线性激活函数(ReLU)避免梯度爆炸和消失问题.

具体的，采用两层卷积神经网络和一层ConvLSTM提取当前时刻输入图像特征$x_t\in\mathbb{R}^{h\times w\times c}$，分别生成两个大小相同通道不同的向量，键向量$k_t\in\mathbb{R}^{h\times w\times c_k}$和值向量$v_t\in\mathbb{R}^{h\times w\times c_v}$，$c_k+c_v=c$.此外，还在这两个向量的通道上添加位置向量$s\in\mathbb{R}^{h\times w\times c_s}$，提供空间位置信息.查询向量$q_t\in\mathbb{R}^{n\times(c_k\times c_s)}$由上一时刻模型输出经过多层感知器(MLP)产生，上一时刻模型输出主要包含空间关系推理模块输出$\text{output}_{t-1}^{\text{SRR}}$以及记忆推理模块的输出$\text{output}_{t-1}^{\text{MR}}$，即空间关系推理的结果和记忆推理的结果共同决定下一时刻需要查询的目标.注意力机制通过关系计算函数点积计算查询向量与键向量的相似度，归一化得到图像特征的重要性程度，再与对应位置上的值向量相乘得到注意力向量$\text{Att}_t\in\mathbb{R}^{n\times(c_v\times c_s)}$，注意力的输出通过查询向量和注意力向量的按列叠加，展开后经过一层多层感知器压缩得到$\text{answer}_t\in\mathbb{R}^{1\times c_a}$.

$$q_t=\text{MLP}[\text{output}_{t-1}^{\text{SRR}};\ \text{output}_{t-1}^{\text{MR}};\ a_{t-1};\ r_{t-1}] \tag{1}$$

$$\text{Att}_t=\frac{\exp(k_i\cdot q_t)}{\sum_j\exp(k_j\cdot q_t)}v_i\quad i,j\in h\times w \tag{2}$$

$$\text{answer}_t=\text{MLP}[\text{Att}_t^n;\ q_t^n] \tag{3}$$

式(1)中a_{t-1}和r_{t-1}分别强化学习中上一时刻的动作和奖励.式(2)中i,j表示提取的图像特征的特征向量.

空间关系推理模块的输出还包含了输入图像的低层特征信息$x_t'\in\mathbb{R}^{1\times c'}$，将ConvLSTM替换普通卷积神经网络后，经全连接层压缩得到.空间关系推理模块的输出(绿色球)为：

$$\text{output}_t^{\text{SRR}}=[\text{answer}_t;\ x_t'] \tag{4}$$

3.3 记忆推理(MR)

记忆推理使用具有外部存储模块以及相应的读写机制的记忆增强神经网络[18]，也称为外部记忆.利用自注意力学习记忆组成部分之间的关系，分析当前时刻的空间信息与记忆之间的联系，增强记忆网络的推理能力.为了与空间关系推理进行融合，做了两点改进，第一点以空间关系推理信息作为输入，而不仅仅只是图像特征[33]，丰富了空间信息的表示；第二点将空间关系推理与记忆的关系作为新的记忆片段进行存储，能够重复使用预先计算的关系，高效地利用历史信息.记忆增强神经网络的输出为当前时刻空间信息与记忆推理的结果，本文将其与空间关系推理的结果叠加在一起作为模型的输出，同时用于生成下一时刻空间关系推理中关系查询向量的生成.总的来说，空间关系推理融合了记忆推理的输出，记忆推理又融合了空间关系推理的输出，二者信息的融合提高了样本利用率.

记忆单元只保存M个记忆信息，记忆的更新为先进后出，删除最后一个记忆信息后，将$M-1$个记忆信息向后移动一个位置，再将新生成的记忆推理信息保存在记忆单元中.记忆单元初始化为零.本文

不直接将得到的空间信息存储在记忆当中，而是先推理其与记忆的关系，再将关系进行存储，这样做的好处是重复使用了预先计算的关系，提高样本利用率.为了尽可能多的获取历史记忆，进行先读取后写入的顺序，其中读操作包含了记忆的交互.

3.3.1　读操作

时刻t时，将空间关系推理输出作为记忆单元的输入$\text{Output}_t^{\text{SRR}}\in\mathbb{R}^{1\times d_M}$，$d_M=c_a+c'$.上一时刻记忆网络表示为$\text{Memo}_{t-1}\in\mathbb{R}^{n_M\times d_M}$，其中$n_M$表示记忆片段个数，$d_M$表示记忆片段大小.输入向量与记忆单元按行叠加组成新的向量作为控制器Transformer的输入.控制器计算输入向量与记忆片段之间的关系得到新的记忆向量，通过交互新的记忆向量包含了当前输入与历史记忆信息推理的结果.

$$T^{in}=\begin{bmatrix}\text{output}_t^{\text{SRR}}W_{ans}+b_{ans}\\ \text{Memo}_{t-1}W_{mem}+b_{mem}\end{bmatrix}\tag{5}$$

其中$T^{in}\in\mathbb{R}^{n_T\times d_T}$表示控制器Transformer的输入，$W_{ans}\in\mathbb{R}^{d_M\times d_T}$、$W_{mem}\in\mathbb{R}^{d_M\times d_T}$、$b\in\mathbb{R}^{d_T}$为控制器网络的参数，$T^{out}\in\mathbb{R}^{n_T\times d_T}$为控制器输出，其中$n_T=1+n_M$表示输入或者输出的数量，$d_T$表示向量大小.其中第一行表示输入向量与历史记忆片段的推理，即记忆推理结果$\text{output}_t^{\text{MR}}=T_{0:}^{out}\in\mathbb{R}^{1\times d_T}$.在每个周期开始前，初始化记忆单元为零.

3.3.2　写操作

在外部记忆网络更新中，为了减少计算量，一次只改变记忆单元一行记忆片段.记忆推理模块的输出$\text{output}_t^{\text{MR}}$为当前时刻输入向量与记忆单元交互的结果，将其经过线性变换存储在新记忆单元的第一行，删除最后一行记忆片段，其余记忆片段只移动位置不改变内容.由于保存的记忆片段都是当前时刻输入向量与历史记忆信息交互的结果，因此进行一次交互计算都是多步推理，而且当前时刻产生的记忆向量能够保持M个时间步，很大程度增加记忆的长期依赖.记忆单元更新方式如下，其中$W_M\in\mathbb{R}^{d_T\times d_M}$，$b_M\in\mathbb{R}^{d_M}$：

$$\text{Memo}_t=\begin{bmatrix}\tanh(\text{output}_t^{\text{MR}}W_M+b_M)\\ \text{Memo}_{t-1[0:-1]}\end{bmatrix}\tag{6}$$

3.4　策略和值函数生成

本文整体结构如图1所示，模型的输出为空间关系推理和记忆推理两部分推理结果叠加而成，接下来对强化学习策略和值函数的生成进行详细地描述.

SRRMR模型的输出包含了当前时刻空间关系推理模块输出$\text{output}_t^{\text{SRR}}$和记忆推理模块的输出$\text{output}_t^{\text{MR}}$，在强化学习策略和值函数生成中，还特别添加了上一时刻强化学习智能体决策a_{t-1}与奖励r_{t-1}.根据模型的输出，本文分别使用两个不同的线性网络生成智能体策略π和值函数V^{π}.

$$\text{output}_t^{\text{SRRMR}}=[\text{output}_t^{\text{SRR}};\ \text{output}_t^{\text{MR}};\ a_{t-1};\ r_{t-1}]\tag{7}$$

$$\pi=\text{output}_t^{\text{SRRMR}}W_{\pi}\tag{8}$$

$$V^{\pi}=\text{output}_t^{\text{SRRMR}}W_V\tag{9}$$

其中$W_{\pi}\in\mathbb{R}^{(c_a+d_T+2)\times d_{\pi}}$表示生成策略的线性网络，$W_V\in\mathbb{R}^{(c_a+d_T+2)}$表示生成值函数的线性网络.表1总结了模型所有网络类型和大小.模型训练时间随着网络的增大而增加.

表1　网络类型和大小

模型	类型	大小
输入	CNN	kernel size:8×8，stride: 4，channels: 32
		kernel size:4×4，stride: 2，channels: 64
		kernel size:3×3，stride: 2，channels: 64
	$\text{MLP}_{\text{input}}$	hidden units: 512
空间关系推理	Conv-LSTM	kernel size:3×3，channels: 128
	$\text{MLP}_{\text{query}}$	hidden units: 256
		hidden units: 128
		hidden units: 288
	$\text{MLP}_{\text{answer}}$	hidden units: 512
		hidden units: 256
外部记忆	$\text{Linear}_{\text{ans}}$	hidden units: 256
	$\text{Linear}_{\text{mem}}$	hidden units: 256
	$\text{Linear}_{\text{memory}}$	hidden units: 768
	n_M	8
	d_M	64
输出	$\text{Linear}_{\text{policy}}$	Actions
	$\text{Linear}_{\text{value}}$	1

3.5　模型算法描述

算法1.空间关系推理和记忆推理.

输入：奖励折扣因子γ、训练终止时间步T、样本批量大小B、批量样本长度T'、actor数量N、经验池容量N'；

输出：目标策略网络参数

初始化：经验池为空，记忆单元为零，初始化行为策略μ网络；

① 建立N个actor，使用行为策略μ网络模型收集N'条轨迹信息，每条轨迹包含T'时间步的样本；

② 建立目标策略π网络模型；

③ FOR $t=1$ to Tdo

④　在经验池N'中按照先进先出原则获取B大小的轨迹信息；

⑤ 将B个轨迹中的观测输入目标策略网络；

⑥ 空间特征提取，得到x'_t； /*3.2节*/

⑥ 空间关系推理提取，得到answer_t； /*3.2节*/

⑦ 空间特征与空间关系推理进行残差连接得到空间关系推理的输出$\text{output}_t^{\text{SRR}}$，也是记忆推理的输入；

⑧ 记忆推理提取$\text{output}_t^{\text{MR}}$，更新记忆单元；

/*3.3节*/

⑨ 空间关系推理与记忆推理的结果进行残差连接，分别使用两个不同网络生成策略π和值函数V^{π}；

/*3.4节*/

⑩ 计算策略梯度损失、策略π熵损失、带有V-trace的值函数估计损失之和；

/*2.1节*/

⑪ 根据损失反向传播更新目标策略网络参数并保存；

⑫ 使用目标策略网络参数更新行为策略网络参数，生成轨迹信息补充经验池；

⑬ $t=B+T'$；

⑭ END FOR；

其中⑥～⑨步骤为本文网络模型主要结构，按照先空间关系推理后记忆推理的形式，但是在内容上进行融合的方法.空间关系推理中查询向量由上一时刻记忆推理生成，记忆推理存储空间中的关系特征.二者通过不断学习和重复使用预先计算关系，提高智能体空间推理能力和记忆推理能力，进而提高样本利用率.

4 实验结果分析

本节首先介绍了实验平台、环境和SRRMR模型参数的选择，其次阐述了两个基线强化学习算法IMPALA和空间关系推理方法以及本文提出的模型结构和变体，最后通过对比实验分析提出的模型有效提升了智能体学习效率和学习能力.

4.1 实验环境与平台

4.1.1 实验环境

本文使用的计算平台为浪潮英信服务器NF5280M5，配置2颗Intel Xeon Gold 5218 CPU、2块NVIDIA Tesla V100显卡(GPU)和128 GB内存.深度神经网络基于Pytorch开源库实现，使用分布式强化学习算法重要性加权Actor-Leraner架构(IMPALA)训练智能体，同时使用一种V-trace离策略校正方法，其带有RMSProp优化器的VTRACE损失.

模型在所有游戏环境中使用相同的超参数，以验证模型的通用性.RMSProp优化器中学习率为2e-4，梯度平方的移动均值衰减率0.99，模糊因子0.01.VTRACE损失中交叉熵损失0.01，折扣因子0.99.在训练过程中，每次使用3个随机种子数，在有些训练曲线波动较大有较高的随机性的游戏环境中采用10个随机种子数，训练5e7帧，取两次结果的平均值作为最终结果，Actor数量设置为36，Learner一次学习8条轨迹的80个时间步Actor交互的数据，记忆推理模块中记忆单元个数设置为16.

4.1.2 实验平台

实验环境为基于Atari 2600游戏平台Arcade Learning Environment(ALE)[40]的OpenAI集成环境.ALE是一个软件框架，允许用户通过操纵杆信息和屏幕信息模拟平台与Atari 2600交互.ALE包含超过50个2600游戏，并且提供可视化工具.游戏的观测包含像素尺寸为160×210的单个游戏屏幕，18个离散动作空间(8个方向、8个同时移动和射击、射击、无操作)通过操纵杆控制器定义.每个时间步还会获得一个即时奖励值，通常通过在帧之间的得分的差异指定.每个episode会在游戏终止时结束，也可以在预定义的时间步结束后终止，避免智能体进入死循环.由于Atari游戏的帧率很高，每秒可以生成60帧，本文选择将所选择的动作发送给环境4次，而不是将单独一帧都传递给智能体，以加快学习速度.本文选择ALE平台中的15个游戏进行实验，游戏风格和类型多种多样，包括探索、规划、反应性游戏等.

4.2 实验对比方法

本文选择两个基线算法进行对比，分别是IMPALA强化学习算法[23]和空间关系推理方法Attention[12].IMPALA是一个大规模强化学习训练的框架，负责采样的actor与策略学习learner有一定的滞后，通过V-trace技术对off-policy样本进行修正训练.IMPALA的网络结构是三层卷积神经网络、一层全连接网络和一层LSTM，值函数和策略分别使用一层全连接网络生成；空间关系推理方法Attention在IMPALA基础上对网络结构进行了修改，只改变空间特征提取而不改变记忆网络LSTM.Attention方法将卷积神经网络最后一层替换为Conv-LSTM，使用注意力机制中的Multi-Head Attention进行关系推理，后接多层感知器MLP，IMPALA中的值函数和策略的生成保持不变.SRRMR同样修改的是IMPALA中的网络结构，其中方法MR是对IMPALA的记忆网络进行修改，空

间特征提取保持不变，使用外部记忆网络代替LSTM，将IMPALA前馈网络的输出与记忆网络的输出进行残差连接.方法SRRMR在IMPALA中将空间关系推理Attention方法与MR方法进行融合，外部记忆网络保存空间关系特征，空间关系利用记忆信息进行查询与提取，值函数和策略的生成不变.

本文提出的在深度强化学习上融合了空间关系推理和记忆推理SRRMR模型结构如第3节所述，此外还训练一个单独的外部记忆网络(MR).MR网络结构只将强化学习前馈神经网络提取的图像特征作为输入，即SRRMR网络中去掉空间关系推理部分，与Loynd等人[18]不同的是记忆网络的输入不是目标的特征而是当前整个输入图像的特征.这样设置的目的有两个：可以看出在不同的环境中各个部分所起的作用；空间关系推理本身学习过程较慢需要长时间训练，Attention[12]在3e9帧上进行训练才获得一个较好结果，为了加快学习速率，在部分空间关系推理作用较小环境中只使用外部记忆推理结构.

4.3 实验分析

为了实验的一致性，在两种基线方法和本文方法上使用相同的超参数.不同的方法在游戏上训练5e7帧，每次使用3个随机种子，训练结果取均值和方差，如图2所示，横轴代表帧数，纵轴表示一个episode内获得的奖励分数.随着智能体与环境交互次数的增加，不断根据奖励进行学习，在训练过程中可能产生过拟合，因此奖励值曲线会产生较大波动，但四种方法的学习能力是不断提高的.本文方法整体上相对两种基线方法来说学习速率更快.

与强化学习算法IMPALA[23]相比，SRRMR模型在10/15个游戏中具有竞争力，特别是在Alien、Asterix、Berzerk、Boxing、Centipede、Pong和Riverraid这7个游戏中学习速率甚至超过强化学习算法.图2中的训练曲线表明本文提出的SRRMR模型能够以更少的交互次数达到最大奖励值，提升强化学习智能体的学习效率，提高样本利用率.分析SRRMR方法在有些环境中不如强化学习算法IMPALA的原因，可能是游戏机制和环境信息过于复杂，空间关系推理可能分配给任务相关信息较低的权重造成干扰，而强化学习提取的图像特征权重相同不容易丢失信息.

与空间关系推理方法Attention[12]相比，本文提出的SRRMR模型的不同之处在于增强了记忆网络.Attention方法使用LSTM存储空间关系，由于隐藏状态少输出只与最近的状态相关，相当于只具备空间关系推理能力.SRRMR模型将空间关系与外部记忆网络相融合，同时增强了智能体空间关系推理能力和长期记忆推理能力.图2训练曲线结果显示，SRRMR模型在15个游戏中智能体学习速率和学习能力都获得了很大提升.表明在相同实验条件下，SRRMR模型智能体能够更快学习到最优策略.Attention方法在一些速度较快如Breakout、Pong游戏环境中，智能体基本没有学习到策略，分析其原因：小球目标小、速度快，注意力容易忽略小球的位置分配，较低的权重导致智能体很难捕捉到小球，这也是Attention方法远远不如强化学习本身学习能力的原因.SRRMR模型为了克服这一问题，在空间关系基础上还添加了输入图像的状态表示，并以记忆片段形式存储和推理，保留了图像的任务相关信息、全局信息与历史信息，明显提高了智能体学习效率和学习能力.

为了进一步分析模型不同部分的作用，将SRRMR中空间关系推理部分舍去，只保留外部记忆网络得到记忆推理MR.与基线算法IMPALA相比，MR在Asterix、Berzerk、Boxing、Centipede和SpaceInvaders这5个游戏中有明显提升，在多数环境中的训练曲线与基线算法IMPALA相似，表明在强化学习算法中添加外部记忆网络的MR方法提升了学习效率和学习能力.在一些学习效果不如IMPALA的和环境中，如Alien和Boxing中，通过与空间关系推理的融合学习效率和学习能力获得进一步提升.实验表明记忆推理MR能够提高智能体的学习能力，在相同实验条件下能够学习到更好的策略.

SRRMR虽然提升了收敛速率，但是在有些环境中最终收敛得分不如MR.空间关系使用低层卷积神经网络提取输入图像的局部特征，同时使用注意力机制推理这些局部特征之间的关系并自主学习分配权重，这就导致智能体对当前奖励影响不大，但是对长期获得奖励的重要信息产生忽略，例如Seaquest中底部的氧气条与后期奖励有很大关系.但是在目标较大、空间关系相对简单的环境中，如Boxing，智能体在不丢失重要信息的情况下能快速学到获得较高奖励的策略，大约需要1e7帧左右，基线强化学习算法在2.2e7帧数上才有所提升.表明SRRMR方法在强化学习算法中进行空间关系推理和记忆推理融合在能够正确识别目标物体

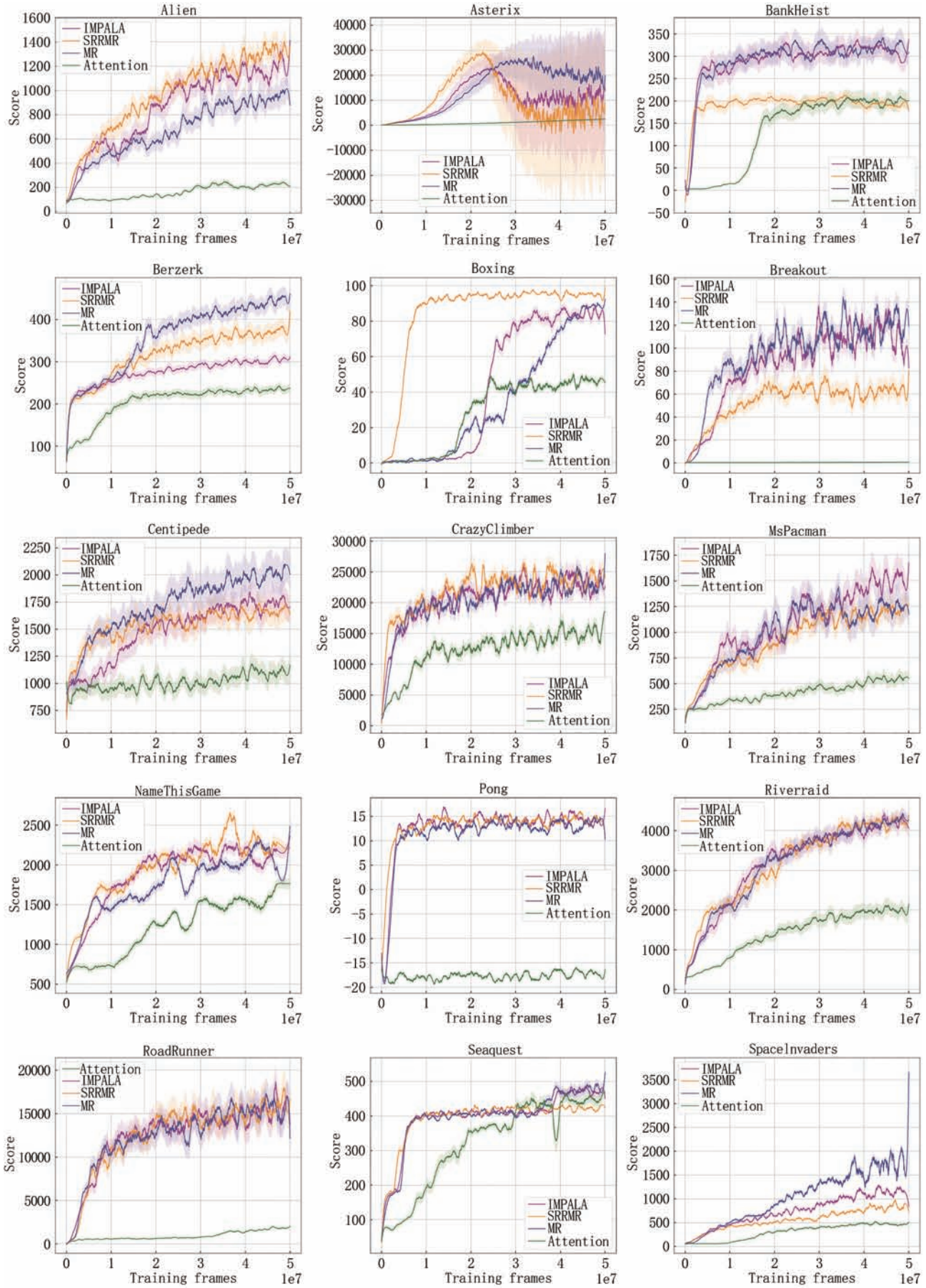

图2 15个Atari游戏训练曲线

以及与长期奖励关系较为明显的环境中能够快速提升学习速率，提高样本利用率，减少训练的样本数.本文提出的方法SRRMR和MR在其他强化学习算法上同样能够提升训练收敛速率，提高样本利用率，如图3所示.

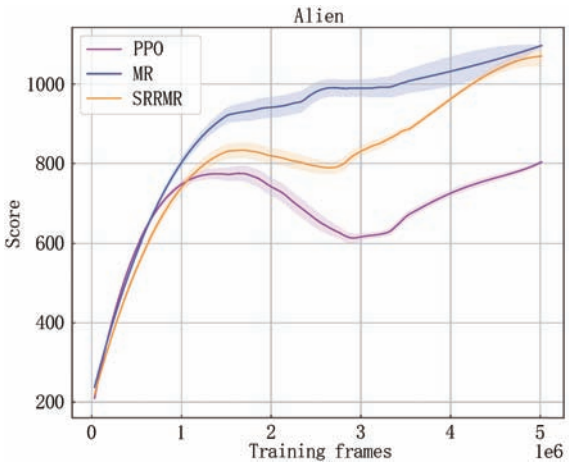

图3　SRRMR与MR适用性训练曲线

本文还进行了最优策略的测试，反映了智能体最终的学习结果.训练结束后，选择保存当前最优策略作为测试策略.通过比较这几种算法在这15个游戏上30个测试周期的得分平均值和方差，对其学习效果进行评价.在训练阶段通常设置失去一条命则游戏结束，而在测试阶段初始值设为3，失去所有生命则游戏结束.15个Atari游戏最优策略测试结果列于表2，显示了不同模型智能体在15个游戏中最优策略学习情况，其中随机智能体(Random)、人类专家(Human)的分数来源于[41].强化学习算法A3C的分数来源于[22].IMPALA在相同时间步2e8上最终收敛分数优于A3C达10/15个，表明本文选择的强化学习算法IMPALA具有优越性.

由表2数据可知，Attention在训练过程中获得的奖励值远远小于其他三种方法，甚至在有些环境中获得奖励均值接近于零，表明Attention方法没有学习到任何技能导致游戏失败.SRRMR、MR和IMPALA三种方法在有些环境中甚至超过了人类得分，表明三种方法在所有环境中都能获得胜利，并且学习到游戏的内在机制.记忆推理方法MR与强化学习方法IMPALA相比，在相同条件下取得最好成绩达12/15个，验证了MR方法能够提高强化学习算法IMPALA的性能，SRRMR方法与IMPALA相比，在相同条件下取得最好成绩达6/15个，SRRMR在Alien、Boxing、CrazyClimber、NameThisGame和RoadRunner环境中还能够进一步提升MR的奖励均值，两种方法在其他游戏上与强化学习算法IMPALA也具有竞争力.由于每个游戏得分差异较大，将测试结果的方差归一化求和后，IMPALA为4.7，MR为3.5，SRRMR为2.9.SRRMR训练方差之和最小，有较好的稳定性.实验结果表明，在深度强化学习基础上融合空间关系推理和记忆推理，大大提高智能体学习效率和样本利用率，使得智能体以更少的迭代次数收敛到更好的效果，稳定性也有所提升.

表2　15个游戏实际得分表

游戏	Random[41]	Human[41]	A3C[22](2e8)	IMPALA[23]	Attention[12]	MR	SRRMR
Alien	227.8	7127.7	945.3	1220.0(87.2)	39.3(13.6)	1375.0(563.9)	**1517.9(35.7)**
Asterix	210.0	742.0	17 244.5	9292.2(8905.4)	380.6(114.4)	**38 194.4**(19862.0)	14 911.1(**8760.0**)
BankHeist	14.2	734.4	932.8	318.9(19.2)	58.1(99.8)	**335.9(3.5)**	219.7(11.1)
Berzerk	123.7	2630.4	862.2	306.6(18.3)	213.7(30.9)	**510.6(16.7)**	442.3(103.6)
Boxing	0.1	4.3	37.3	66.8(56.9)	4.2(3.2)	97.1(**1.3**)	**97.5**(4.3)
Breakout	1.7	30.5	766.8	**355.8**(91.5)	0.8(0.4)	333.8(**2.7**)	131.8(78.0)
Centipede	2090.9	11 963.2	1997.0	1664.7(261.7)	1222.8(222.1)	**2104.9(161.2)**	1458.3(273.7)
CrazyClimber	10 780.5	35 410.5	13 8518.0	**29 064.4**(4551.0)	9881.1(2663.0)	26 191.1(**1662.9**)	26936.7(8677.5)
MsPacman	307.3	6951.6	850.7	**1761.1(258.2)**	163.8(83.4)	1518.7(530.2)	1030.2(427.5)
NameThisGame	2292.3	4076.2	12 093.7	3179.0(**278.5**)	2199.2(451.8)	3351.6(384.2)	**3549.6**(341.7)
Pong	−20.7	14.6	10.7	21.0(0.0)	−21.0(0.0)	21.0(0.0)	**21.0(0.0)**
Riverraid	1338.5	17 118.0	6591.9	4393.0(324.4)	1185.2(775.3)	**4525.7(171.3)**	3722.9(622.5)
RoadRunner	11.5	7845.0	73 949.0	4907.8(7426.2)	321.1(1.9)	10 216.7(7653.3)	**15 207.8(125.1)**
Seaquest	68.4	42 054.7	1326.1	518.0(111.5)	184(36.9)	**520.9**(105.5)	460.9(**2.1**)
SpaceInvaders	148.0	1668.7	23 846.0	2164.7(**159.0**)	368.1(116.1)	**4711.6**(4418.9)	1802.9(289.5)

5 结 论

深度强化学习中智能体与环境进行实时交互，由于采样效率低，对于高维复杂数据输入，学习特征较难而且需要很长时间才能得到较好的策略.为了提高样本利用率，本文提出在深度强化学习中融合空间关系推理和记忆推理的模型框架，增强智能体空间关系学习和长期记忆能力，加快理解环境的潜在状态.空间关系推理中包含利用注意力机制的关系学习和神经网络提取的状态表示，分别提取任务相关的重点信息和全局信息.记忆推理使用外部记忆网络存储空间信息，并利用自注意力进行记忆之间的交互，增强记忆推理能力，提取历史信息.记忆推理的结果用于生成空间关系中的查询向量，根据历史信息决定下一时刻智能体需要关注的部分.在Atari2600中的实验结果表明，二者的融合能够有效提升智能体学习效率和样本利用率，减少与环境交互次数，改善学习效果，提升最优策略的质量.

本文提出的在强化学习算法IMPALA中添加了外部记忆网络MR方法能够在大部分游戏环境中提升了学习效果，SRRMR在MR基础上与空间关系推理进行融合，能够加快训练收敛速率，提高样本利用率，但是空间关系推理本身具有限制性，以增加计算量为代价将局部关系转换成全局关系计算，此外注意力机制自主学习分配空间特征权重容易忽略与长期奖励有关的信息.针对空间关系推理中的不足之处，未来研究内容为深度强化学习时空特征，分析与任务相关信息在时序上的变化，减少空间背景信息的重复计算，同时利用注意力机制计算与任务相关信息在时序上的关系.

参 考 文 献

[1] Liu Q, Zhai J W, Zhang Z Z, et al. A survey on deep reinforcement learning. Chinese Journal of Computers, 2018, 41(1): 1-27

[2] Francois-Lavet V, Henderson P, Islam R, et al. An introduction to deep reinforcement learning. Foundations and Trends® in Machine Learning, 2018, 11(3-4): 219-354

[3] Kamilaris A, Prenafeta-Boldú F X. Deep learning in agriculture: A survey. Computers and Electronics in Agriculture, 2018, 147: 70-90

[4] Dargan S, Kumar M, Ayyagari M R, et al. A survey of deep learning and its applications: a new paradigm to machine learning. Archives of Computational Methods in Engineering, 2020, 27(4): 1071-1092

[5] Sutton R S, Barto A G. Reinforcement learning: An introduction. 2nd Edition. Cambridge, USA: The MIT Press, 2018

[6] Sieusahai A, Guzdial M. Explaining deep reinforcement learning agents in the Atari domain through a surrogate model// Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment. Vancouver, Canada, 2021, 17(1): 82-90

[7] Ye D, Liu Z, Sun M, et al. Mastering complex control in moba games with deep reinforcement learning//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020, 34(04): 6672-6679

[8] Ibarz J, Tan J, Finn C, et al. How to train your robot with deep reinforcement learning: lessons we have learned. The International Journal of Robotics Research, 2021, 40(4-5): 698-721

[9] Le N, Rathour V S, Yamazaki K, et al. Deep reinforcement learning in computer vision: a comprehensive survey. Artificial Intelligence Review, 2021: 1-87

[10] Jiang Y, Gu S S, Murphy K P, et al. Language as an abstraction for hierarchical deep reinforcement learning// Proceedings of the Conference on Neural Information Processing Systems, Berlin, Germany 2020:8761-9553

[11] Leroy G, Rueckert D, Alansary A. Communicative reinforcement learning agents for landmark detection in brain images// Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, 2020:177-186

[12] Mott A, Zoran D, Chrzanowski M, et al. Towards interpretable reinforcement learning using attention augmented agents// Proceedings of the Conference on Neural Information Processing Systems, 2020:11941-12726

[13] Zambaldi V, Raposo D, Santoro A, et al. Relational deep reinforcement learning. arXiv preprint arXiv:1806.01830, 2018

[14] Graves A, Wayne G, Danihelka I. Neural turing machines. arXiv preprint arXiv:1410.5401, 2014

[15] Weston J, Chopra S, Bordes A. Memory networks. arXiv preprint arXiv:1410.3916, 2014

[16] Zhou Fei-Yan, Jin Lin-Peng, Dong Jun. A survey of convolutional neural networks. Chinese Journal of Computers, 2017, 40(6):1229-1251 (in Chinese)
周飞燕，金林鹏，董军.卷积神经网络研究综述.计算机学报，2017, 40(6):1229-1251

[17] Graves A. Long short-term memory. Supervised sequence labelling with recurrent neural networks, 2012: 37-45

[18] Loynd R, Fernandez R, Celikyilmaz A, et al. Working memory graphs// Proceedings of the International Conference on Machine Learning. PMLR, 2020: 6404-6414

[19] Jaunet T, Vuillemot R, Wolf C. DRLViz: Understanding decisions and memory in deep reinforcement learning. Computer Graphics Forum. 2020, 39(3): 49-61

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

[21] Silver D, Lever G, Heess N, et al. Deterministic policy gradient algorithms// Proceedings of the International conference on machine learning. PMLR, Detroit, USA, 2014: 387-395

[22] Mnih V, Badia A P, Mirza M, et al. Asynchronous methods for deep reinforcement learning// Proceedings of the International conference on machine learning. PMLR, New York, USA, 2016: 1928-1937

[23] Espeholt L, Soyer H, Munos R, et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures// Proceedings of the International Conference on Machine Learning. PMLR, Hanoi, Vietnam, 2018: 1407-1416

[24] Zhou J, Cui G, Hu S, et al. Graph neural networks: A review of methods and applications. AI Open, 2020, 1: 57-81

[25] Garnelo M, Shanahan M. Reconciling deep learning with symbolic artificial intelligence: representing objects and relations. Current Opinion in Behavioral Sciences, 2019, 29: 17-23

[26] Kurin V, Igl M, Rocktäschel T, et al. My body is a cage: the role of morphology in graph-based incompatible control. arXiv preprint arXiv:2010. 01856, 2020

[27] Jiang Z Y, Minervini P, Jiang M Q, et al. Grid-to-graph: Flexible spatial relational inductive biases for reinforcement learning// Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems. Richland, USA, 2021,674-682

[28] Jiang Z, Luo S. Neural logic reinforcement learning// Proceedings of the International Conference on Machine Learning. PMLR, Vancouver, Canada, 2019: 3110-3119

[29] Che W, Fan X, Xiong R, et al. Visual relationship embedding network for image paragraph generation. IEEE Transactions on Multimedia, 2019, 22(9): 2307-2320

[30] Hausknecht M, Stone P. Deep recurrent Q-learning for partially observable MDPs. arXiv preprint arXiv:1507. 06527, 2015

[31] Singla A, Padakandla S, Bhatnagar S. Memory-based deep reinforcement learning for obstacle avoidance in UAV with limited environment knowledge. IEEE Transactions on Intelligent Transportation Systems, 2019, 22(1): 107-118

[32] Zaremba, Wojciech and Sutskever, Ilya. Reinforcement learning neural turing machines. arXiv preprint arXiv: 1505. 00521, 2015

[33] Oh J, Chockalingam V, Lee H. Control of memory, active perception, and action in minecraft// Proceedings of the International Conference on Machine Learning. PMLR, New York,USA, 2016: 2790-2799

[34] Parisotto E, Salakhutdinov R. Neural map: Structured memory for deep reinforcement learning. arXiv preprint arXiv: 1702. 08360, 2017

[35] Sukhbaatar S, Weston J, Fergus R. End-to-end memory networks//Proceedings of the 28th International Conference on Neural Information Processing Systems. Bali, Indonesia, 2015, 2: 2440-2448

[36] Kumar A, Irsoy O, Ondruska P, et al. Ask me anything: Dynamic memory networks for natural language processing// Proceedings of the International conference on machine learning. PMLR, New York,USA, 2016: 1378-1387

[37] Miller A, Fisch A, Dodge J, et al. Key-value memory networks for directly reading documents//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. New York, USA, 2016: 1400-1409

[38] Chandar S, Ahn S, Larochelle H, et al. Hierarchical memory networks. arXiv preprint arXiv:1605. 07427, 2016

[39] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in Neural Information Processing Systems, 2017, 30: 6000-6010

[40] Bellemare M G, Naddaf Y, Veness J, et al. The arcade learning environment: An evaluation platform for general agents. Journal of Artificial Intelligence Research, 2013, 47: 253-279

[41] Wang Z, Schaul T, Hessel M, et al. Dueling network architectures for deep reinforcement learning// Proceedings of the International Conference on Machine Learning. PMLR, New York,USA, 2016: 1995-2003

LIU Hui-Ling, Ph. D. candidate. Her research interest is deep reinforcement learning.

LIU Peng, Ph.D., professor. His main research interests are machine learning and pattern recognition.

BAI Chen-Jia, Ph. D., young researcher. His research interest is deep reinforcement learning.

Background

Deep reinforcement learning combines the perception ability of deep learning and the decision-making ability of reinforcement learning, and can be controlled directly according to the input image. It is an artificial intelligence that is closer to the way of human thinking. This is largely because deep reinforcement learning can flexibly learn and utilize observations and reward signals, but it also faces many

challenges, such as low sample efficiency and long training times.

In order to make more efficient use of sample information and improve sample efficiency, this paper proposes a model structure that integrates spatial relationship reasoning and memory reasoning in deep reinforcement learning, which can extract information from available data more efficiently and improve the learning efficiency and learning ability of deep reinforcement learning agents Spatial relationship reasoning uses attention mechanism to extract the relationship between input image features in a single time step. The query vector of attention mechanism integrates the content of memory reasoning, so that the agent can actively perceive the spatial relationship according to the prior knowledge in memory memory reasoning uses self attention to learn the relationship between spatial relationship and memory information, and stores the memory reasoning results in external memory units. Besides memory function, the memory network also adds reasoning ability They make full use of the learned knowledge, extract relevant information efficiently, and finally integrate the results of spatial relationship reasoning and memory reasoning as outputs to generate reinforcement learning strategies and value functions. The model proposed in this paper improves the learning efficiency in 7/15 game environments, and is also competitive in other games. It shows that srrmr model has faster convergence speed, can learn better strategies in relatively few interaction times, and improves the sample utilization rate of reinforcement learning agent.

This paper is supported by Key projects of National Natural Science Foundation of China (No. 51935005), Basic scientific research projects (No. JCKY20200603C010), Natural Science Foundation of Heilongjiang Province (No. LH2021F023), and Heilongjiang Science and technology planning project (No. GA21C031).