基于双重注意力机制的异步优势行动者评论家算法

凌兴宏[1),2)] 李 杰[1),2)] 朱 斐[1),2)] 刘 全[1),2),3),4)] 伏玉琛[5)]

[1)](苏州大学计算机科学与技术学院 江苏 苏州 215006)

[2)](苏州大学江苏省计算机信息处理技术重点实验室 江苏 苏州 215006)

[3)](吉林大学符号计算与知识工程教育部重点实验室 长春 130012)

[4)](软件新技术与产业化协同创新中心 南京 210000)

[5)](常熟理工学院计算机科学与工程学院 江苏 常熟 215500)

摘 要 深度强化学习是目前机器学习领域发展最快的技术之一.传统的深度强化学习方法在处理高维度大状态的空间任务时,庞大的计算量导致其训练时间过长.虽然异步深度强化学习利用异步方法极大缩短了训练时间,但会忽略某些更具价值的图像区域和图像特征.针对上述问题,本文提出了一种基于双重注意力机制的异步优势行动者评论家算法.新算法利用特征注意力机制和视觉注意力机制来改进传统的异步深度强化学习模型.其中,特征注意力机制为卷积神经网络卷积后的所有特征图设置不同的权重,使得智能体聚焦于重要的图像特征;同时,视觉注意力机制为图像不同区域设置权重参数,权重高的区域表示该区域信息对智能体后续的策略学习有重要价值,帮助智能体更高效地学习到最优策略.新算法引入双重注意力机制,从表层和深层两个角度对图像进行编码表征,帮助智能体将聚焦点集中在重要的图像区域和图像特征上.最后,通过 Atari 2600 部分经典实验验证了基于双重注意力机制的异步优势行动者评论家算法的有效性.

关键词 注意力机制;双重注意力机制;行动者评论家;异步优势行动者评论家;异步深度强化学习

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2020.00093

Asynchronous Advantage Actor-Critic with Double Attention Mechanisms

LING Xing-Hong[1),2)] LI Jie[1),2)] ZHU Fei[1),2)] LIU Quan[1),2),3),4)] FU Yu-Chen[5)]

[1)](*School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu* 215006)

[2)](*Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou, Jiangsu* 215006)

[3)](*Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun* 130012)

[4)](*Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing* 210000)

[5)](*School of Computer Science and Engineering, Changshu Institute of Technology, Changshu, Jiangsu* 215500)

Abstract In recent years, deep reinforcement learning (DRL), which combines deep learning and reinforcement learning together, is a new research hotspot in artificial intelligence. As DRL takes advantage of deep learning, it is able to take raw images as input, which extends applications of reinforcement learning. At the mean while time, DRL retains the advantages of reinforcement learning in application such as intelligent policy decision or robotic control. However, traditional DRL such as deep Q-network (DQN) or double deep Q-network (DDQN), could hardly deal with complex tasks with high-dimensional state in a short time. Researchers have proposed many methods to solve this problem, and asynchronous advantage actor-critic (A3C) is one of the most used

收稿日期:2018-05-21;在线出版日期:2019-03-14.本课题得到国家自然科学基金(61772355,61303108,61373094)、江苏省高等学校自然科学研究重大项目(17KJA520004)、吉林大学符号计算与知识工程教育部重点实验室资助项目(93K172014K04)、苏州市应用基础研究计划工业部分(SYG201422)、苏州市民生科技项目(SS201736)和江苏高校优势学科建设工程资助项目资助.**凌兴宏**,博士,副教授,主要研究方向为机器学习、强化学习研究.E-mail:lingxinghong@suda.edu.cn.**李 杰**,硕士研究生,主要研究方向为深度学习、深度强化学习.**朱 斐**(通信作者),博士,副教授,中国计算机学会(CCF)专业会员,主要研究方向为机器学习、生物医学信息.E-mail:zhufei@suda.edu.cn.**刘 全**,男,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为机器学习、智能信息处理.**伏玉琛**,男,1968 年生,博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为强化学习、人工智能.

algorithm. As we know, traditional asynchronous deep reinforcement learning can use multi-threading techniques to reduce large amounts of training time. However, when it comes to high-dimensional large-state space tasks, some valuable and important image areas and features are often ignored, such as Atari 2600 games. The reason is that Agent's attention is focused on the entire input image and all features of the image, without any emphases on some important features. To handle this problem, we employ the attention mechanism to ameliorate the performance of traditional asynchronous deep reinforcement learning models. In recent years, inspired by human vision, the attention mechanism has been extensively used in machine translation, image recognition and speech recognition, becoming one of the most noteworthy and in-depth research techniques in the area of deep learning technologies. Based on this, we put forward an asynchronous advantage actor-critic with double attention mechanisms (DAM-A3C). In DAM-A3C, there are two main characteristics: visual attention mechanism (VAM) and feature attention mechanism (FAM). First, the application of visual attention mechanism can enable Agent to adaptively engage in the image region, especially in those more important areas which can enhance the cumulative reward at each moment, reducing the computational cost of the network's training and finally accelerating the process of learning the approximate optimal strategy. Second, via the exertion of FAM, an asynchronous advantage actor-critic is expected to pay more attention to those features with more value. What we know is that different convolution kernels can generate different feature maps by operating convolution on the image in convolutional neural network. And feature maps completely describe the image from different features. The traditional training of convolutional neural network treats each extracted feature equally, which means all features have the same proportion, instead of different levels of focus according to their value. However, some image features have a crucial role in the description of images, such as color features, shape features and spatial relationship features, etc. In order to alleviate this problem, FAM can assist Agent to converge on feature maps with rich values, which will facilitate Agent to make correct decisions. To sum up, we introduce FAM in VAM-A3C model and propose DAM-A3C model. DAM-A3C utilizes visual attention mechanism and feature attention mechanism to enable Agent to concentrate on the important areas and important features of the image, which advances the network model to recognize important information and key features of the image in a short time. We select some classic Atari 2600 games as experimental objects to evaluate the performance of the new model. The experimental result shows that the new model has better performance than the traditional asynchronous advantage actor-critic algorithm in experimental tasks.

Keywords attention mechanism; double attention mechanisms; actor-critic; asynchronous advantage actor-critic; asynchronous deep reinforcement learning

1 引 言

深度学习[1](Deep Learning)是机器学习领域的一种监督学习方法,在智能语音[2-3]、计算机视觉[4-6]和自然语言处理[7-8]等领域已取得了显著的应用.深度学习方法以多层感知机为整体架构,以激活函数和梯度反向传播等为训练算法,不仅能够提供端到端(end-to-end)的解决方案,而且能够在无需人工参与的前提下提取出有效的状态特征.强化学习[9](Reinforcement Learning)不同于传统的监督学习方法,其强调的是在环境中自主学习目标策略,主要应用在工业控制、仿真模拟和游戏博弈等领域[10-12].深度强化学习(Deep Reinforcement Learning,DRL)利用人工神经网络的特征表示能力和强化学习的策略学习能力,在复杂的高维状态空间任务中能够有效提取数据特征并作出最优策略.

Mnih 等人[13-14]在传统 Q 学习[15]算法中引入卷

积神经网络(Convolutional Neural Network,CNN)来拟合值函数,提出了深度Q网络(Deep Q-Network,DQN).DQN 模型用于处理基于视觉感知的控制任务,在 Atari 2600 平台上的大部分游戏中均表现出超出人类玩家的水平,是 DRL 领域的开创性工作.DQN 模型在训练过程中,每次选取动作都会以状态动作 Q 值作为衡量指标,这会导致学习模型出现过度拟合的问题.基于上述问题,双重深度 Q 网络模型(Double Deep Q-Network,DDQN)[16] 利用两种不同的网络参数完美解决了模型过拟合的问题,两套网络参数分别用作选择动作和评估策略.DDQN 使得动作和策略相互独立,使用两套不同的参数来表示它们,降低了过度乐观估计 Q 值的风险,在某些基于视觉感知的游戏任务中获得了更稳定有效的学习性能.

虽然 DQN 和 DDQN 算法在 Atari 2600 平台的大部分游戏上表现效果惊人,但是这两种 DRL 算法采用的都是等概率采样,无法充分发挥某些重要训练样本的价值.因此,Schaul 等人[17] 提出基于优先级的经验回放机制,该机制为所有的训练样本设置不同的优先级,以此代替等概率采样方式,帮助模型充分利用有价值的样本数据.

不同类型的深度神经网络为 RL 算法提供了高效的表征能力,同时传统的 DRL 算法均采用了基于优先级经验的回放机制来满足所有训练样本的独立性.然而经验回放机制具有其固有的欠缺性:

(1)经验回放机制导致训练模型的计算量非常大,对计算设备要求过高,需要如 GPU 等专门加速计算的硬件.

(2)经验回放机制需要将大量的训练样本存储在经验池中,因此需要较大的存储空间.

(3)经验回放机制必须使用如 Q 学习等异策略学习算法,例如 Sarsa 算法的同策略强化学习算法无法利用该机制.

针对经验回放机制的上述三种问题,利用了 DRL 算法和强化学习中异步思想的异步深度强化学习(Asynchronous Deep Reinforcement Learning,ADRL)[18],使得传统的 DRL 模型不需要存储大量的训练样本,也不需要重放一定批量的样本来计算损失并且更新模型参数,因此极大减少了存储和计算的开销.与传统的 DQN、DDQN 等算法相比,ADRL 可以利用多线程技术加速 DRL 的训练,在较短的时间内获得更好的实验效果.

近年来,注意力机制(Attention Machanism,AM)被广泛应用于机器翻译、图像识别和语音识别等领域[19-21],是深度学习技术中最值得关注与深入研究的技术之一.AM 通常是基于编码器-解码器(Encoder-Decoder)框架应用于深度学习各个领域中,实现端到端的学习.Bahdanau 等人[19] 基于编码器-解码器框架,利用 AM 在英-法双语的翻译任务取得令人满意的结果;Xu 等人[20] 借鉴注意力机制在机器翻译中的应用,提出一种用于计算机视觉任务中的视觉注意力机制(Visual Attention Mechanism,VAM),VAM 使得算法模型将关注点聚焦于具有重要价值的图像区域,有效描述图片主题;Chorowski 等人[21] 首次利用 Attention 机制用于对输出序列的每个音素和输入语音序列中一些特定帧进行关联;Sorokin 等人[22] 首次将注意力机制和 DRL 算法结合,在深度循环 Q 网络中引入 Attention 机制,通过高亮显示智能体(Agent)正在关注的游戏屏幕区域,实现在线监测训练过程.因此,本文在异步深度强化学习模型中加入了 VAM,提出一种基于视觉注意力机制的异步优势行动者评论家算法(Asynchronous Advantage Actor-Critic with Visual Attention Mechanism,VAM-A3C),VAM-A3C 算法帮助学习模型在后续训练中充分利用重要的图像区域信息,从而 Agent 能够根据这些状态信息高效地学习策略.

另一方面,传统的深度强化学习算法在编码阶段利用 CNN 来提取图像的特征信息.如图 1 所示,CNN 利用多个卷积核(filter)对图像进行卷积运算,不同的卷积核会提取出不一样的特征信息,即提取出多种不同的特征图(Feature Map).本文基于 CNN 的特征图,提出一种特征注意力机制(Feature Attention Mechanism,FAM),该机制通过给所有的特征图初始化相应的权重并在训练过程中学习权重参数.FAM 能够帮助网络模型将注意力聚焦在有价值的特征图上,从而关注图像的某些重要特征.

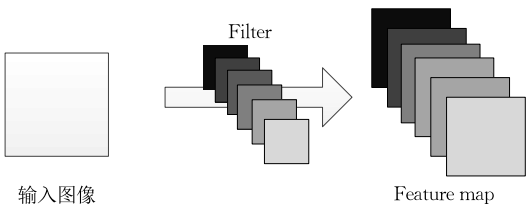

图 1　CNN 卷积过程

ADRL 算法利用异步方法不仅消除了训练样本的关联性,还加速了学习算法的训练过程,但是传统的 ADRL 无法将注意力集中在更有价值的图像

区域和图像特征上，原因在于 Agent 的注意力集中于整幅输入图像以及图像的所有特征. 本文提出一种基于双重注意力机制的异步优势行动者评论家算法（Asynchronous Advantage Actor-Critic with Double Attention Mechanisms，DAM-A3C），该算法在传统的基于循环神经网络的 A3C 算法的基础上做了以下改进：(1) 在传统异步深度强化学习模型中引入 VAM，Agent 能够根据不同图像区域设置的权重参数来不同程度地利用其区域信息；(2) 引入 FAM 到 A3C 模型中，使得 Agent 重点关注图像中有价值的特征，从而直观、有效地作出正确的决策. 实验表明，在 Atari 2600 游戏中，基于双重注意力机制的 A3C 算法能够提升传统 A3C 算法的性能.

2 背景知识

2.1 强化学习

强化学习是一种处理序贯决策任务的学习方法，其通过获得最大累积奖赏以解决决策优化的问题. 智能体根据观察到的环境状态来进行自主学习，因此满足马尔科夫决策过程[23-24]（Markov Decision Process，MDP）的学习条件，MDP 可以由元组 (S,A,P,R,γ) 来描述，其中：

(1) S 指状态集合，$\boldsymbol{s}_t \in S$ 表示 t 时刻的状态；

(2) A 指动作集合，$\boldsymbol{a}_t \in A$ 表示 t 时刻执行的动作；

(3) P 为当前任务的状态转移概率，$P(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t,\boldsymbol{a}_t)$ 表示在状态 $\boldsymbol{s}_t$ 下采用动作 $\boldsymbol{a}_t$ 转移到状态 $\boldsymbol{s}_{t+1}$ 的概率值；

(4) R 为当前任务的奖赏函数，r_t 表示 Agent 在状态 $\boldsymbol{s}_t$ 下执行动作 $\boldsymbol{a}_t$ 获得的立即奖赏；

(5) γ 为折扣因子，用来计算累计回报.

在强化学习中，智能体根据已学习到的策略 π 来执行动作 $\boldsymbol{a}_t$，从状态 $\boldsymbol{s}_t$ 开始所有时刻的累积奖赏值，称之为期望回报. Agent 所获得期望回报为

$$R_t=\sum_{t'=t}^{T}\gamma^{t'-t}r_{t'},$$

其中 $\gamma\in[0,1]$ 用来表示未来时刻的奖赏值对累计回报的影响程度.

强化学习的最终目标是最大化 Agent 在每个情节的累计回报值，以此学习到最优策略. 状态动作值 $Q^{\pi}(s,a)$ 表示智能体在当前状态 $\boldsymbol{s}_t$ 下根据已知的学习策略来优先采取动作 $\boldsymbol{a}_t$，最终得到的期望回报：

$$Q^{\pi}(s,a)=E[R_t|\boldsymbol{s}_t=s,\boldsymbol{a}_t=a].$$

同时，最优的 Q 值是指 Agent 在给定状态 s 和动作 a 时，策略 π 能够获取的最大奖赏值：

$$Q^{*}(s,a)=\max_{\pi}Q^{\pi}(s,a).$$

强化学习方法包括动作值拟合方法和基于动作概率的学习方法，动作值拟合的强化学习包括 Q 学习算法、Sarsa 算法等；基于动作概率的强化学习包括策略梯度方法[25-27]. 行动者评论家算法[28-29]（Actor-Critic，AC）结合了值函数学习方法和策略梯度学习方法，以策略梯度方法作为行动者算法，用于动作选择；以值函数方法作为评论家算法，用于评论动作的好坏.

2.2 行动者评论家算法

AC 算法可以将策略的获取和值函数的计算进行分离，策略结构被看作行动者，值函数计算的部分被看作评论家. AC 算法结构如图 2 所示，行动者表示 Agent 在当前状态下根据策略 π 采取一个动作，使环境迁移到下一个状态；评论家得到动作时，利用时间差分（Temporal Difference，TD）误差项来评论当前状态所采取动作的优劣性. TD 误差的计算公式如下所示：

$$\delta_t=\sum_{i=0}^{n=1}\gamma^i r_{t+i}+\gamma^n V(\boldsymbol{s}_{t+n})-V(\boldsymbol{s}_t),$$

其中 r_{t+i} 表示 Agent 在状态 $\boldsymbol{s}_{t+i}$ 根据策略 π 采取 a_{t+i} 所获得的立即奖赏，$V(\boldsymbol{s}_t)$ 表示在状态 $\boldsymbol{s}_t$ 的期望回报. TD 误差若大于 0，则学习算法在后续状态中应积极采用动作 $\boldsymbol{a}_t$；TD 误差若小于 0，则学习算法在后续状态中应降低采用动作 $\boldsymbol{a}_t$ 的概率.

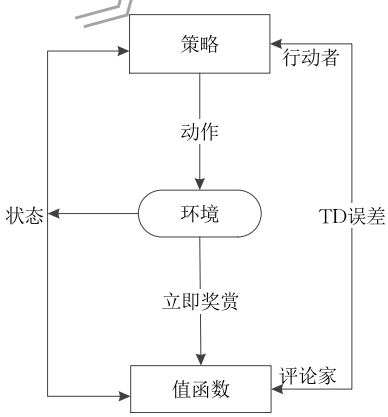

图 2 AC 算法结构图

AC 算法将行动者部分和评论家部分独立出来，能够对值函数和策略函数的训练同步进行，从而减少模型的训练时间；同时，AC 算法是一种策略梯度算法，当动作空间是连续的时候，动作选择时不需要为无穷的动作进行大量的计算.

2.3　优势行动者评论家算法

行动者评论家网络(Actor-Critic Network，ACN)将深度学习算法引入到 A3C 算法中，ACN 可以在高维度状态空间内有效学习，使得传统的 AC 算法不需要进行复杂的人工预处理. 与 AC 算法相似，ACN 包括两个部分：

(1) 值网络，即 $V(s;\boldsymbol{\theta}_v)$，$\boldsymbol{\theta}_v$ 表示值网络的参数. 值网络用来评价 Agent 在当前状态 $\boldsymbol{s}_t$ 下所采取动作 $\boldsymbol{a}_t$ 的好坏；

(2) 策略网络，即 $\pi(\boldsymbol{a}_t|\boldsymbol{s}_t;\boldsymbol{\theta})$，$\boldsymbol{\theta}$ 表示策略网络的参数. 策略网络用来优化学习算法的策略.

行动者评论家网络在进行策略网络更新时，平等对待每一个状态动作对，对每一个状态动作值的计算均采用相同的权重. 然而在当前状态 $\boldsymbol{s}_t$ 下，采取的每一个动作所获得的奖赏是不同的，即有些状态动作对的回报值相对较高，有些动作获得的奖赏相对较低. 针对此问题，ACN 引入优势函数 $A(\boldsymbol{s}_t,\boldsymbol{a}_t;\boldsymbol{\theta},\boldsymbol{\theta}_v)$，用来评估执行动作的优劣性. 这种算法称为优势行动者评论家(Advantage Actor-Critic，A2C)，该算法中优势函数的计算公式如下所示：

$$A(\boldsymbol{s}_t,\boldsymbol{a}_t;\boldsymbol{\theta},\boldsymbol{\theta}_v)=\sum_{i=0}^{n-1}\gamma^i r_{t+i}+\gamma^n V(\boldsymbol{s}_{t+n};\boldsymbol{\theta}_v)-V(\boldsymbol{s}_t;\boldsymbol{\theta}_v),$$

其中 $\gamma\in[0,1]$ 表示折扣因子，是未来奖赏对累计奖赏的重要程度，r_{t+i} 表示立即奖赏. 当 $n=1$ 时，优势函数表示 1 步回报优势函数；当 $n=k$ 时，优势函数表示 n 步回报优势函数. A2C 算法中值网络参数和策略网络参数的梯度计算如下所示：

$$\mathrm{d}\boldsymbol{\theta}_v=\partial(R-V(\boldsymbol{s}_t;\boldsymbol{\theta}_v))^2/\partial\boldsymbol{\theta}_v,$$

$$\mathrm{d}\boldsymbol{\theta}=\nabla_{\boldsymbol{\theta}}\log\pi(\boldsymbol{a}_t|\boldsymbol{s}_t;\boldsymbol{\theta})A(\boldsymbol{s}_t,\boldsymbol{a}_t;\boldsymbol{\theta},\boldsymbol{\theta}_v),$$

其中 R 表示当前状态 $\boldsymbol{s}_t$ 下 Agent 根据已知策略 π 选择动作 $\boldsymbol{a}_t$ 所获得的奖赏值.

2.4　视觉注意力机制

在 Encoder-Decoder 模型结构中，Encoder 模块首先将原始输入状态信息进行编码，生成一个上下文向量，而后 Decoder 模块再对该向量进行解码输出信息. 传统 Encoder-Decoder 模型的上下文向量对所有的输入信息平等对待，即所有输入的权重均赋值为 1，导致了模型无法充分利用重要的区域信息. 因此，Bahdanau 等人[19]提出一种注意力机制，在神经机器翻译领域中引入 AM，不再使用统一的语义特征，使 Decoder 在输出语句时为输入序列赋值不同的权重，充分利用重要的语义信息；Xu 等人[20]首次将 AM 应用于图像处理，利用 AM 改进了 Encoder-Decoder 结构，提出一种视觉注意力机制. 下面具体分析 VAM 的计算过程：

(1) 计算 t 时刻图像各区域的视觉信息值：

$$e_{ti}=f_{att}(\boldsymbol{a}_{ti},h_{t-1}),$$

其中，$\boldsymbol{a}_t$ 表示 t 时刻图像各区域的输入向量集，$\boldsymbol{a}_{ti}$ 表示图像第 i 个区域位置的输入向量；h_{t-1} 表示 $t-1$ 时刻的隐层状态值；f_{att} 表示视觉信息值的计算函数.

(2) 使用 Softmax 回归函数对图像各区域的视觉信息值归一化，得到 t 时刻各区域的相对视觉重要性：

$$\alpha_{ti}=\frac{\exp(e_{ti})}{\sum_{k=1}^{N}\exp(e_{tk})},$$

其中 N 表示图像的区域总数.

(3) 根据图像的输入向量集和各区域的相对视觉重要性，计算 t 时刻 Encoder 模块的上下文向量：

$$\boldsymbol{C}_t=\sum_{k=1}^{N}\alpha_{ti}\boldsymbol{a}_{ti}.$$

3　算法描述

本节阐述了 DAM-A3C 算法的网络模型和算法的具体训练过程. 其中 3.1 节介绍学习模型对原始数据的预处理操作，3.2 节介绍学习模型用来提取输入状态特征信息的 CNN 层，3.3 节到 3.5 节介绍 VAM-A3C 和 DAM-A3C 网络模型，3.6 节介绍 DAM-A3C 模型的训练过程.

3.1　预处理

网络模型在训练 Atari 2600 游戏时，首先需要利用图像预处理来去除图像的边缘区域，以降低模型训练时处理图像信息的复杂度. Atari 2600 游戏中每一帧图像的尺寸为 210×160，模型需要对原始图像进行预处理操作来减少计算代价，预处理操作包括灰度转换、降采样以及裁剪. 通过这三种预处理操作，网络模型将原始的 RGB 三色图像转换成尺寸大小为 84×84 的灰度图像. 预处理操作使得网络模型将原始图像转换为尺寸更小的图像，在不导致重要信息流失的前提下，可以最大限度地提高 Encoder 对输入数据的编码速度.

3.2　CNNs

DAM-A3C 模型以四层卷积神经网络作为 Encoder，通过编码将输入数据变成一系列低维度的特征表示. 四层 CNN 的具体信息如下：

(1)第一层 CNN. 输入游戏画面的尺寸为 84×84×1,通过 64 个 3×3 的卷积核以步幅为 2×2 对图像进行卷积运算,得到 64 幅大小为 42×42 的特征图;

(2)第二层 CNN. 输入游戏画面的尺寸为 42×42×64,通过 64 个 3×3 的卷积核以步幅为 2×2 对图像进行卷积运算,得到 64 幅大小为 21×21 的特征图;

(3)第三层 CNN. 输入游戏画面的尺寸为 21×21×64,通过 128 个 3×3 的卷积核以步幅为 2×2 对图像进行卷积运算,得到 128 幅大小为 11×11 的特征图;

(4)第四层 CNN. 输入游戏画面的尺寸为 11×11×128,通过 128 个 3×3 的卷积核以步幅为 2×2 对图像进行卷积运算,得到 128 幅大小为 6×6 的特征图.

经过四层 CNN 的编码后,模型在每一帧都会输出 128 幅尺寸为 6×6 的特征图像.

3.3 VAM-A3C

A3C 模型使用 CNN 网络来提取游戏画面的特征信息,每一次卷积运算都会得到一个特征图,将这些特征图映射为特征向量,特征向量的元素则代表了图像中不同区域位置的信息. 传统的基于 CNN 的前馈 A3C 模型(FF-A3C)直接将特征向量集合作为上下文向量进行解码,t 时刻该特征向量集合为

$$\boldsymbol{a}_t=\{\boldsymbol{a}_t^1,\boldsymbol{a}_t^2,\cdots,\boldsymbol{a}_t^N\},$$

其中,N 表示在当前时刻特征图的个数,$\boldsymbol{a}_t^i$ 表示当前时刻第 i 个特征图的特征向量.

在 Atari 2600 游戏中,优异的网络模型需要在短时间内获取游戏画面的特征信息,若是将注意力聚焦在整幅图像上,会使 Agent 无法及时获得图像中的价值信息. 因此,本文在传统的 A3C 模型中引入了视觉注意力机制,提出一种基于视觉注意力机制的异步优势行动者评论家算法(VAM-A3C). VAM-A3C 算法在解码之前,以 $\boldsymbol{a}_t$ 作为输入,利用视觉注意力机制重新计算上下文向量 $\boldsymbol{C}_t$,帮助 Agent 聚焦于图像的重要区域,从而在较短时间内获得图像的有价值信息. Xu 等人[20]利用 VAM 模块处理图像描述的生成任务,通过长短时记忆网络(Long Short Term Memory Network)来对上下文向量进行解码,需要将上一时刻的隐藏状态考虑进视觉重要值的计算中. 与传统 VAM 不同,VAM-A3C 引入 VAM 模块是为了加强模型对每一帧图像的编码表征能力,而不是一直对一幅图像进行编码. 因此本文利用长短时记忆网络的每一个时间步来对每帧游戏画面进行编码,计算图像各区域点的视觉信息值,计算公式如下所示:

$$vam(\boldsymbol{a}_t^i,h_{t-1})=Linear(Tanh(Linear(\boldsymbol{a}_t^i)+Wh_{t-1})),$$

其中,$Linear$ 是一种线性函数,$Tanh$ 是一种非线性变换. 在计算视觉重要性值之后,再通过 Softmax 归一化和线性加权分别求出图像每个区域的相对视觉重要性以及上下文向量.

3.4 FAM

在图像处理中,学习模型会通过构造一组基来完整描述一张图像,这组基可以看作图像的多种描述角度. CNN 网络中的特征图可以理解为图像在同一层次上不同基的描述. Krizhevsky 等人[30]在 AlexNet 模型中通过 96 种 11×11×3 的卷积核对 227×227×3 的图像进行卷积运算,生成了 96 张 55×55 的特征图,并对这 96 张特征图进行了可视化. 如图 3 所示,这些特征图提取了与图像的频率、方向、颜色等特征的相关信息,从不同的角度全面描述了图像的细节.

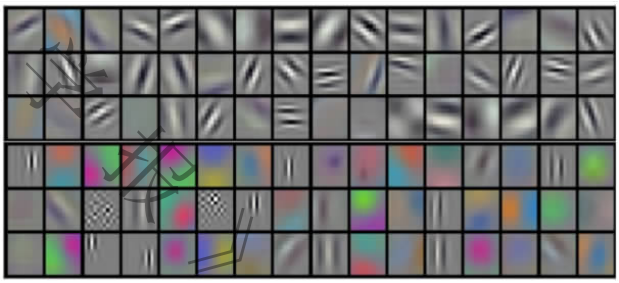

图 3 特征图的可视化

不同卷积核通过对图像进行卷积运算生成的特征图,从不同角度完整描述了图像信息. 然而,某些图像特征对图像的描述具有十分关键的作用. 传统 CNN 的训练平等对待每一个提取出的特征,即每张特征图的权重赋值相同,没有重点关注某些具有重要价值的特征. 为了缓解此问题,本文提出一种特征注意力机制,该注意力机制可以帮助 Agent 将注意力集中于具有丰富价值的特征图,这些特征图可以促进 Agent 进行正确决策. 下面具体介绍 FAM 的计算流程:

(1)计算 t 时刻各特征图的特征重要性值:

$$fam(\boldsymbol{s}_t^i,h_{t-1})=Linear(Tanh(Linear(\boldsymbol{s}_t^i)+Wh_{t-1})),$$

其中,$\boldsymbol{s}_t=\{\boldsymbol{s}_t^1,\boldsymbol{s}_t^2,\cdots,\boldsymbol{s}_t^N\}$ 表示所有特征图的输入向

量集合，$\boldsymbol{s}_t^i$表示第 i 个特征图的输入向量，N 是 t 时刻卷积层输出的特征图数量，h_{t-1}表示 $t-1$ 时刻的循环神经网络隐层输出值.

(2) 根据特征图的向量集合，计算 t 时刻每张特征图的相对特征重要性，即特征图的权重：

$$\partial_{ti}=\frac{\exp(fam(\boldsymbol{s}_t^i,h_{t-1}))}{\sum_{k=1}^{N}\exp(fam(\boldsymbol{s}_t^k,h_{t-1}))}.$$

(3) 对每一张特征图的输入向量和相应的权重进行线性加权，重新计算特征图的元素值：

$$s_{ti}=\partial_{ti}\,s_{ti}.$$

3.5　DAM-A3C

由上述 FAM 计算过程可知，FAM 模块使得图像的重要特征更加突出，权重的赋值更大，以此帮助 Agent 将注意力集中于更有价值的特征图. 因此在 VAM-A3C 的基础上，本文引入了 FAM 模块，提出一种基于双重注意力机制的异步优势行动者评论家算法(DAM-A3C). DAM-A3C 利用 VAM 和 FAM，使 Agent 将注意力聚焦在图像的重要区域和重要特征这两个方面，促进网络模型能够在短时间内感知图像的重要信息和特征. 如图 4 所示，DAM-A3C 网络模型包括 FAM、VAM 和 A3C 三个模块.

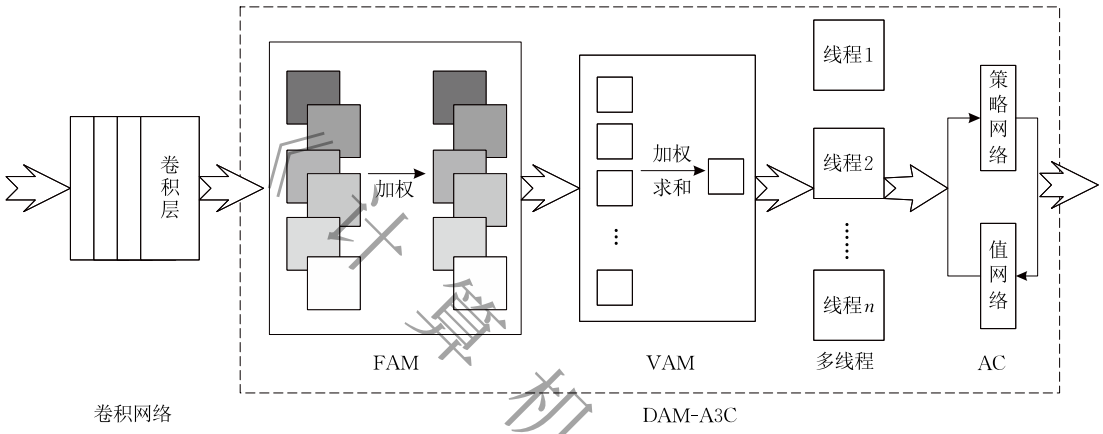

图 4　DAM-A3C 网络模型

DAM-A3C 模型以卷积神经网络层输入的特征图向量集合 $\boldsymbol{s}_t$ 作为输入，通过 FAM 模块重新计算所有特征图内的元素值，帮助 Agent 将注意力集中于重要的特征；其次，再通过 VAM 模块计算视觉上下文向量 $\boldsymbol{C}_t$，帮助 Agent 将注意力集中于重要的图像区域；最后，A3C 模块以最新的上下文向量 $\boldsymbol{C}_t$ 作为输入，通过多线程技术训练策略网络和值网络，所有的线程均有各自的网络模型和网络参数，且网络参数都是从共享网络中获取. A3C 中策略网络和值网络均使用了一个全连接层，其中策略网络的神经元个数是动作值的数量，用来选择最好的游戏动作；值网络的神经元个数只有一个，用来评估执行动作的好坏. DAM-A3C 模型的算法过程如下所示：

算法 1.　基于双重注意力机制的异步优势行动者-评论家.

输入：Atari 2600 游戏的图像画面

输出：共享网络参数

初始化共享网络中策略函数和值函数的参数向量 $\boldsymbol{\theta}$ 和 $\boldsymbol{\theta}_v$

初始化共享网络时间步数 $T=0$

初始化各线程中策略函数和值函数的参数向量 $\boldsymbol{\theta}'$ 和 $\boldsymbol{\theta}'_v$

初始化各线程时间步数 $t\leftarrow 1$

REPEAT

重置梯度 $\mathrm{d}\boldsymbol{\theta}\leftarrow 0$ 和 $\mathrm{d}\boldsymbol{\theta}_v\leftarrow 0$

各线程从共享网络中获取参数 $\boldsymbol{\theta}'=\boldsymbol{\theta}$ 和 $\boldsymbol{\theta}'_v=\boldsymbol{\theta}_v$

$t_{\mathrm{start}}=t$

获取特征图向量集合 $\boldsymbol{s}_t$

计算特征重要性值

$$fam(\boldsymbol{s}_t,h_{t-1})=Linear(Tanh(Linear(\boldsymbol{s}_t)+Wh_{t-1}))$$

使用 Softmax 计算特征权重

$$\partial_t^i=\exp(fam(\boldsymbol{s}_t^i,h_{t-1}))\Big/\sum_{k=1}^{N}\exp(fam(\boldsymbol{s}_t^k,h_{t-1}))$$

计算最新的特征图向量集合 $\boldsymbol{s}_t^i=\partial_t^i\boldsymbol{s}_t^i$

获取视觉向量集合 $\boldsymbol{a}_t=\boldsymbol{s}_t^{\mathrm{T}}$

计算视觉重要性值

$$vam(\boldsymbol{a}_t,h_{t-1})=Linear(Tanh(Linear(\boldsymbol{a}_t)+Wh_{t-1}))$$

使用 Softmax 计算视觉权重

$$\alpha_t^i=\exp(vam(\boldsymbol{a}_t^i,h_{t-1}))\Big/\sum_{k=1}^{N}\exp(vam(\boldsymbol{a}_t^k,h_{t-1}))$$

计算最新上下文向量 $\boldsymbol{C}_t=\sum_{i=1}^{N}\alpha_t^i\boldsymbol{a}_t^i$

获取状态 $\boldsymbol{s}_t\leftarrow\boldsymbol{C}_t$

REPEAT

根据策略 $\pi(\boldsymbol{a}_t\mid\boldsymbol{s}_t;\boldsymbol{\theta}')$ 选择动作 $\boldsymbol{a}_t$

获取奖赏 r_t 和新状态 $\boldsymbol{s}_{t+1}$

$t \leftarrow t+1$

$T \leftarrow T+1$

UNTIL 终止状态 s_t OR $t-t_{start}==t_{max}$

$$R=\begin{cases}0, & \text{终止状态}\\ V(s_t,\theta_v'), & \text{非终止状态 } s_t\end{cases}$$

FOR $i\in\{t-1,\cdots,t_{start}\}$ DO

$R \leftarrow r_i+\gamma R$

$d\theta \leftarrow d\theta+\nabla_{\theta'}\log\pi(a_i|s_i;\theta')(R-V(s_i;\theta_v'))$

$d\theta_v \leftarrow d\theta_v+\partial(R-V(s_i;\theta_v'))^2/\partial\theta_v'$

END FOR

通过 $d\theta$ 和 $d\theta_v$ 更新参数 θ 和 θ_v

UNTIL $T>T_{max}$

RETURN θ, θ_v

3.6 模型的训练过程

DAM-A3C 模型包括卷积神经网络层、FAM、VAM 和 A3C 四个模块，每个模块均是平滑、可微的，因此本文利用随机梯度下降算法来对网络模型的参数进行更新.

卷积神经网络模块包括四层，CNN、FAM 和 VAM 均包括两层全连接，该三个模块根据网络模型的输出值与目标值之间的差值平方来构造误差函数，采用 Adam 梯度下降方法来更新模块内的参数. 网络输出值通过状态 s 下 Agent 选择可能动作 a 的期望奖赏来表示，记作 $Q(s,a|\theta)$，其中 θ 为当前函数参数；目标值通过 Agent 执行动作的最大奖赏值来表示，计算公式如下：

$$y_i=r_i+\gamma\max_{a'}Q(s',a',\theta'),$$

其中，$\gamma\in[0,1]$为折扣因子，θ'为目标函数参数. 因此误差函数表示为

$$L_i(\theta_i)=E_{s,a,r,s'}[(y_i-Q(s,a|\theta_i))^2].$$

A3C 模块包含策略函数网络 $\pi(a_t|s_t;\theta)$和值函数网络 $V(s;\theta_v)$，在策略函数网络中，Agent 根据策略 π 在状态 s_t下选择回报值最高的游戏动作，同时 Agent 通过值函数评估采取动作的优势. DAM-A3C 利用异步方法更新网络参数，所有线程都有各自的网络参数，线程的网络参数均从共享网络中获取. 当 Agent 的情节结束或时间步达到最大值 t_{max}时，模型使用随机梯度下降方法更新共享网络的参数，而不是线程各自的网络参数. 策略函数和值函数的参数更新方法如下：

$$d\theta=\nabla_{\theta'}\log\pi(a_i|s_i;\theta')(R-V(s_i;\theta_v')),$$

$$d\theta_v=\partial(R-V(s_i;\theta_v'))^2/\partial\theta_v',$$

其中，R 表示状态 s_i下 Agent 采取动作 a_i的期望回报；θ 和 θ_v分别表示共享网络的策略函数参数和值函数参数；θ' 和 θ_v'分别表示当前线程的策略函数参数和值函数参数.

4 实验与分析

本节首先介绍实验使用的平台设置以及实验参数设置，其次比较了 DQN 与 A3C 系列算法的训练时间，最后在部分 Atari 2600 游戏中评估了 DQN、FF-A3C、VAM-A3C、FAM-A3C 以及 DAM-A3C 等模型的训练效果. 其中 DQN 表示深度 Q 网络模型，FF-A3C 表示传统的基于前馈网络的 A3C 算法，VAM-A3C 表示基于视觉注意力机制的 A3C 算法，FAM-A3C 表示基于特征注意力机制的 A3C 算法，DAM-A3C 表示基于双重注意力机制的 A3C 算法. 为了客观比较这 5 种模型的实验性能，本次实验均采用相同的实验参数.

4.1 实验平台描述

本文使用了 Intel Core i7-6800k CPU 作为实验的硬件环境，以 Atari 2600 游戏作为实验对象. Atari 2600 游戏是 OpenAI Gym 开源平台中一个环境，该游戏环境包括了策略类、竞技类、桌游类等游戏.

Mnih 等人[18]在 Atari 2600 大部分游戏中验证了 FF-A3C 算法比 DQN、DDQN 等基于经验重放机制的传统 DRL 算法具有更优的实验效果. 因此本文以 Gravitar、StarGunner、TimePilot、Seaquest、Centipede、Breakout、NameThisGame、Amidar、Assault 和 Boxing 等 Atari 2600 游戏为实验对象，着重比较了 FF-A3C、VAM-A3C、FAM-A3C 和 DAM-A3C 这 4 种 A3C 算法的性能差异. 同时，为了进一步验证模型的有效性，本文以训练好的模型参数又对这 10 种游戏进行了测试，分析其测试结果. 在 Agent 训练过程中，网络模型以游戏的原始画面和游戏得分作为输入，帮助 Agent 学习到近似最优策略，实现端到端的学习过程.

4.2 实验参数设置

为了有效比较 DQN、FF-A3C、VAM-A3C、FAM-A3C 和 DAM-A3C 这 5 种模型的性能，本文所有模型的实验参数均相同. 5 种模型对原始数据进行了相同的预处理，均采用 4 层卷积神经网络作为编码器且网络参数都相同.

本文采用 Adam 梯度下降来更新网络参数，参数设置如下：学习率 $\eta=0.001$、一阶矩估计衰减率

$\beta_1=0.9$、二阶矩估计衰减率 $\beta_2=0.99$、超参数 $\varepsilon=0.001$，折扣因子采用 $\gamma=0.99$. FF-A3C、VAM-A3C、FAM-A3C 以及 DAM-A3C 的异步更新方式如下：实验采用 12 个线程加速训练，每 20 步或者当情节结束时更新一次网络参数. 实验共训练 1000 个训练阶段，每个阶段共设 80000 步，共计 8000 万步.

4.3　实验评估和结果分析

基于深度学习的网络训练，需要大量的数据集和训练周期，以阶段结果为指标进行模型评估；而强化学习不需要大量数据和训练周期来逼近最优值，采用一个情节从开始到结束所获得的累计奖赏作为评判标准. DRL 算法结合深度学习和强化学习，利用每个阶段的平均情节奖赏数作为度量指标. 传统 DRL 算法利用了深度神经网络的非线性函数来学习值函数或动作值函数，从而逼近最优决策值，然而这种方法无法保证算法的收敛性. 针对此问题，Mnih 等人[18]利用了经验重放机制和目标网络，使 DRL 算法的收敛性得到保证. ADRL 算法中的异步方法替代了传统的经验重放机制，在加速模型训练的前提下，同样能提高算法的稳定性.

本文比较了 DQN 与 A3C 系列算法在训练 Agent 玩部分 Atari 2600 游戏的每步训练时间，涉及的游戏包括 Gravitar、StarGunner、TimePilot、Seaquest、Centipede、Breakout、NameThisGame、Amidar、Assault 和 Boxing，10 种游戏的简要介绍如表 1 所示. 在 Intel Core i7-6800k CPU 上 DQN 和 A3C 系列算法的训练时间如表 2 所示. 表 2 数据展示了通过使用异步方法的 4 种 A3C 算法的每步训练时间相差不大，而传统 DQN 算法比 A3C 系列算法多耗费几倍的训练时间. 由此可见，A3C 系列算法在保证性能的情况下，利用异步方法极大缩短了模型的训练时间.

表 1　游戏的简要介绍

游戏名	游戏类别	游戏动作	任务目标
Gravitar	射击类	18	Agent 通过射击来摧毁红色掩体
StarGunner	射击类	18	Agent 要避障并射击敌人
TimePilot	射击类	10	Agent 驾驶飞机避障并射击敌人
Seaquest	射击类	18	Agent 水下避障并射击敌人
Centipede	策略类	18	Agent 避开敌人子弹并击打敌人
Breakout	策略类	4	Agent 使用墙壁或桨敲击板砖
NameThisGame	策略类	6	Agent 从屏幕顶部的章鱼中保护宝藏并试图用触角捕捉宝藏
Amidar	策略类	10	Agent 访问棋盘的每个位置，同时避开敌人
Assault	射击类	7	Agent 控制坦克攻击外星环境的表面力量
Boxing	竞技类	18	Agent 控制拳击手攻击对手

表 2　DQN 和 A3C 系列模型的每步训练时间　　（单位：ms/步）

游戏	模型				
	DQN	FF-A3C	VAM-A3C	FAM-A3C	DAM-A3C
Gravitar	8.4	1.8	2.2	1.2	2.3
StarGunner	6.8	1.6	1.1	1.6	2.1
TimePilot	7.5	1.9	2.9	2.3	1.2
Seaquest	6.9	2.9	1.5	2.3	1.7
Centipede	7.2	1.7	1.7	1.7	1.8
Breakout	5.8	1.2	1.7	1.1	1.3
NameThisGame	6.0	1.7	1.1	1.1	1.5
Amidar	13.4	3.0	3.1	3.1	3.1
Assault	10.8	2.3	1.7	3.4	2.1
Boxing	8.3	1.8	1.2	1.8	1.2

本文首先以 Gravitar 游戏为例，评估了 4 种 A3C 模型在各阶段平均每情节的奖赏值. 在 Gravitar 游戏中，Agent 控制一个蓝色航天器在星域中飞行，并通过射击按钮不断摧毁红色掩体来获得奖赏值，直到 Agent 撞到壁垒或者遭到敌人的射击才情节结束. 由于该游戏必须通过摧毁红色掩体来获得奖赏，Agent 在飞行过程中奖赏为 0，因此该游戏的奖赏是稀疏的，必须在多个时间步后才能获得有效奖赏. 4 种 A3C 模型在 Gravitar 游戏上的训练如图 5 所示，横坐标表示模型的训练阶段，纵坐标表示该阶段平均每情节的奖赏值.

4 种 A3C 模型随机初始化网络参数，其参数初始值可能会对算法模型的最终结果产生影响. 因此本文对 Gravitar 游戏进行了 3 次独立实验，各模型

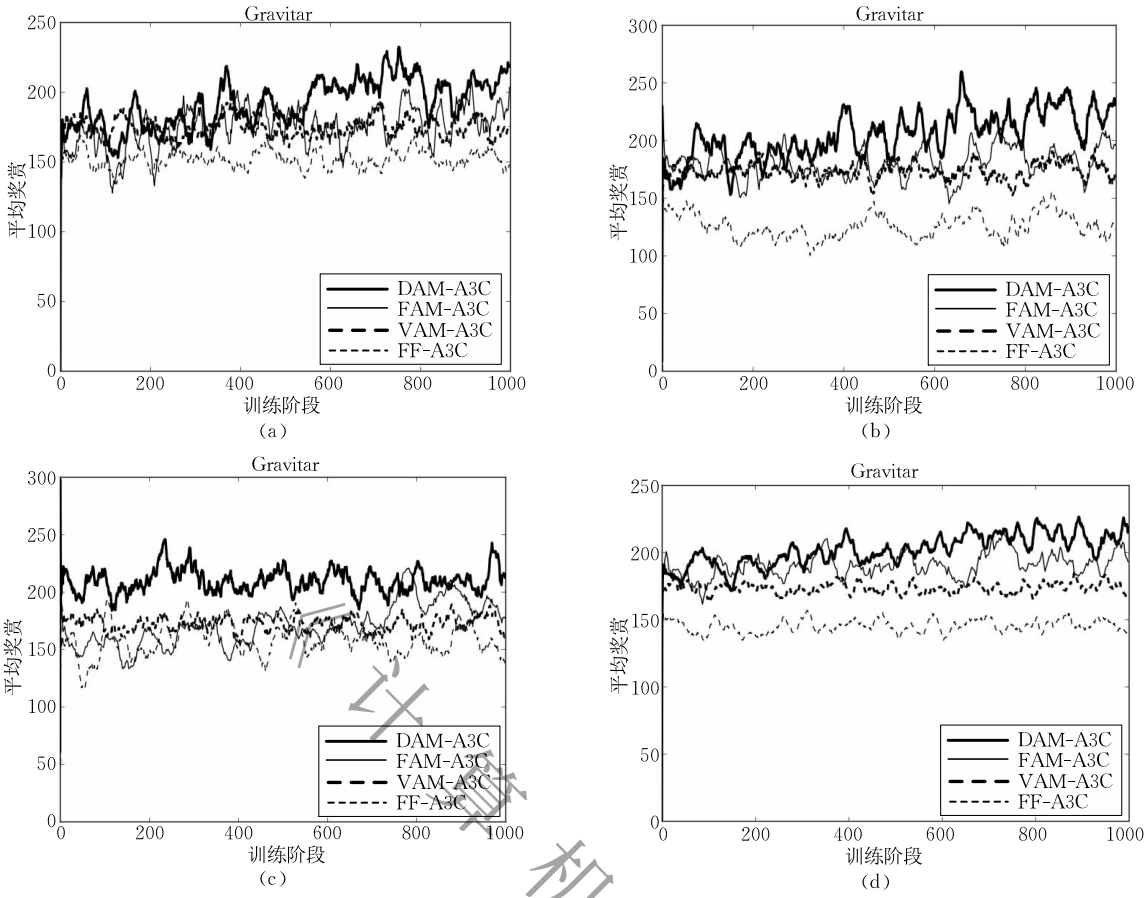

图 5 Gravitar 游戏的训练过程

共训练了 1000 个阶段，图 5 中的(a)、(b)和(c)分别是 3 次训练结果. 由 3 种训练结果可知，FF-A3C 模型在训练过程中的奖赏一直在 150 分上下波动；VAM-A3C 在传统异步优势行动者评论家模型上引入视觉注意力机制，使得智能体将关注点聚焦于游戏画面的关键区域，VAM-A3C 的训练结果比 FF-A3C 模型更好，奖赏值处于 150 到 200 的区间内；FAM-A3C 利用特征注意力机制帮助 Agent 将注意力集中在重要的图像特征上，训练结果与 VAM-A3C 模型相差不大；DAM-A3C 结合了 VAM 模块和 FAM 模块，Agent 在关注游戏画面的关键区域的前提下，又能将注意力聚焦在重要的特征图上，因此 DAM-A3C 比 VAM-A3C 和 FAM-A3C 具有更好的实验性能，游戏得分在训练后期可以高于 200 分. 同时，图 5 中的(d)表示的是 3 次独立实验各模型训练结果的平均曲线图.

另一方面，图 5 展示了 4 种 A3C 算法在训练过程中，各阶段平均每情节的奖赏值会出现不同程度的波动. 分析波动原因，学习模型在训练过程中，网络参数会不断更新，即使生成非常小的变化，都会导致下一阶段策略的分布产生很大的变动. 不过总体而言，DAM-A3C 模型的训练趋势是不变的，各阶段平均每情节的奖赏值都会随着 Agent 的不断学习而持续增加.

此外，本文还选择了 Gym 平台的另外 9 种 Atari 2600 游戏进行模型训练，包括 StarGunner、TimePilot、Seaquest、Centipede、Breakout、NameThisGame、Amidar、Assault 和 Boxing. 在这 9 种游戏中，分别使用了 4 种 A3C 模型进行训练，训练结果如图 6 所示.

从图 6 可知，4 种 A3C 模型在训练 StarGunner、TimePilot、Seaquest、Centipede、Breakout、NameThisGame、Amidar、Assault 和 Boxing 等 Atari 2600 游戏时，利用视觉注意力机制的 VAM-A3C 模型和利用特征注意力机制的 FAM-A3C 模型要比传统基于前馈神经网络的 A3C 模型表现更好；同时，引入 VAM 和 FAM 两种注意力机制的 DAM-A3C 模型比 VAM-A3C 和 FAM-A3C 整体性能更好. 因此，双重注意力机制的引入能够帮助传统深度强化学习算法提升学习性能.

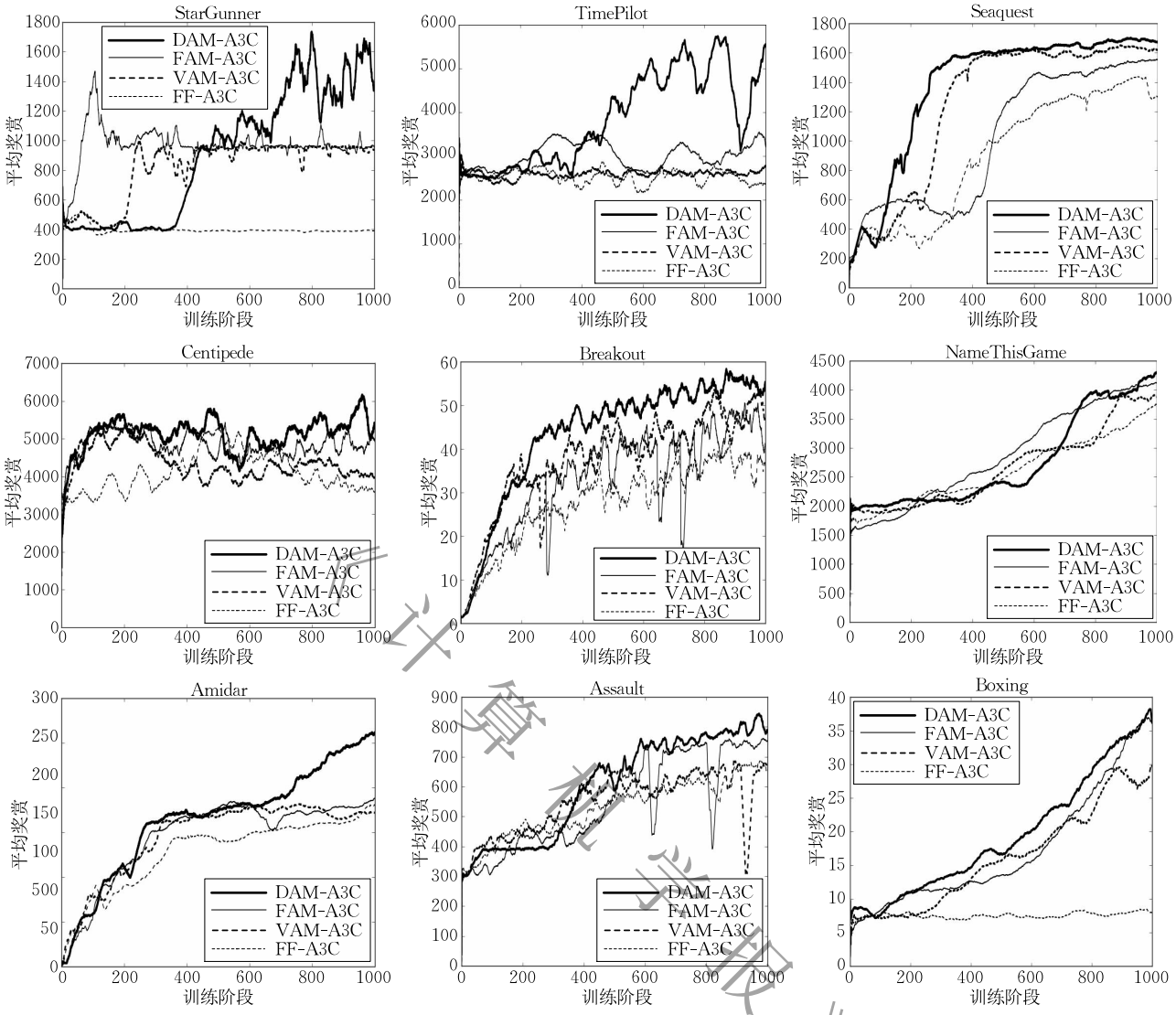

图 6　A3C 系列模型在 Atari 2600 九种游戏中的训练过程

本文通过 DQN、FF-A3C、VAM-A3C、FAM-A3C 和 DAM-A3C 这 5 种模型对 Gravitar 等 Atari 2600 游戏进行了学习训练，每种模型训练共 1000 个阶段，每个阶段包括 8 万个时间步，共计 8000 万步. 同时，一个性能优异的算法还需要能够通过已训练好的网络模型和网络参数，在每一次的决策任务中都能指导 Agent 高效完成任务. 因此本文使用 5 种模型训练之后的网络参数来测试 Agent 在 Atari 2600 游戏中的决策能力.

本文使用 5 种模型在 Gravitar、StarGunner、TimePilot、Seaquest、Centipede、Breakout、NameThisGame、Amidar、Assault 和 Boxing 这 10 种 Atari 2600 游戏进行实验测试. Agent 使用已有参数直接在这 10 种游戏上进行测试，每种游戏的测试共有 3 次试玩阶段，每次试玩共计 80 000 个时间步，通过计算 3 次试玩阶段所有情节的平均奖赏值以及情节最大奖赏值来进行游戏评估. 同时，考虑到测试阶段不同情节的奖赏波动，本实验计算了 5 种模型最后 50 次情节的奖赏标准差以评估测试奖赏间的差异性. 5 种模型评估结果如表 3 所示.

由评估表可知，与深度 Q 网络、基于前馈网络的 A3C 算法、基于视觉注意力机制的 A3C 算法和基于特征注意力机制的 A3C 算法相比，本文提出的基于双重注意力机制的 A3C 算法在指导智能体试玩 10 种 Atari 2600 游戏时都取得了不错的提升. 从平均奖赏一列中可以看出，在 10 种测试游戏中，VAM-A3C、FAM-A3C 和 DAM-A3C 这 3 种 A3C 改进模型相较于传统的 DQN 和基于前馈网络的 A3C 模型，测试性能提升较大. 从评估表的最大奖赏一列中可以看出，Boxing 游戏中，DAM-A3C 的

表 3 各模型在 10 种 Atari 2600 游戏中的测试结果

游戏	模型	平均奖赏	标准差	最大奖赏
Gravitar	DQN	58.06	17.50	254.52
	FF-A3C	123.38	10.52	920.83
	VAM-A3C	174.08	7.48	964.26
	FAM-A3C	204.29	11.71	1000.48
	DAM-A3C	**230.08**	13.50	**1025.00**
StarGunner	DQN	376.27	3.82	832.50
	FF-A3C	388.83	2.78	766.67
	VAM-A3C	936.75	19.22	1441.67
	FAM-A3C	954.33	20.68	1547.00
	DAM-A3C	**1561.83**	134.91	**5133.33**
TimePilot	DQN	1036.59	89.56	4188.92
	FF-A3C	2317.25	109.02	6183.33
	VAM-A3C	2782.17	58.94	7241.67
	FAM-A3C	3012.17	220.37	8275.00
	DAM-A3C	**5547.67**	800.36	**10341.67**
Seaquest	DQN	1108.92	52.88	1548.72
	FF-A3C	1303.48	55.00	1766.67
	VAM-A3C	1618.28	7.68	1828.33
	FAM-A3C	1555.15	6.79	1803.33
	DAM-A3C	**1672.82**	7.37	**1863.20**
Centipede	DQN	1844.98	6.52	1879.05
	FF-A3C	3626.43	166.39	12259.92
	VAM-A3C	3972.36	151.53	15396.58
	FAM-A3C	4929.26	216.84	18541.08
	DAM-A3C	**5398.44**	318.14	**19584.00**
Breakout	DQN	36.52	3.67	102.74
	FF-A3C	36.22	2.14	108.00
	VAM-A3C	46.89	3.12	126.67
	FAM-A3C	41.04	16.57	115.58
	DAM-A3C	**55.34**	1.17	**238.00**
NameThisGame	DQN	3764.47	86.61	6480.83
	FF-A3C	3750.60	163.22	6607.50
	VAM-A3C	3875.46	94.07	6850.83
	FAM-A3C	4120.38	77.93	7148.67
	DAM-A3C	**4289.35**	141.62	**7294.36**
Amidar	DQN	71.90	19.75	102.10
	FF-A3C	180.98	21.74	241.08
	VAM-A3C	172.05	29.17	268.83
	FAM-A3C	188.50	26.88	323.08
	DAM-A3C	**261.62**	25.89	**367.33**
Assault	DQN	465.01	75.21	816.56
	FF-A3C	654.88	121.93	1038.42
	VAM-A3C	674.13	118.55	1016.75
	FAM-A3C	753.81	197.34	1419.25
	DAM-A3C	**788.69**	199.38	**1514.42**
Boxing	DQN	16.92	8.17	35.84
	FF-A3C	8.02	9.86	19.75
	VAM-A3C	29.86	11.88	51.58
	FAM-A3C	36.16	10.65	**60.00**
	DAM-A3C	**36.75**	9.44	56.92

最优表现稍劣于 FAM-A3C，但两者相差不大；在其他 9 种游戏中，DAM-A3C 模型在指导 Agent 试玩的最优表现则优于其他模型，尤其在 TimePilot 和 Centipede 中的游戏表现甚至超过了部分专业玩家的水平.

5 结束语

异步深度强化学习在传统 DRL 算法中引入多线程技术，利用异步方法加速学习模型的训练. 作为异步深度强化学习的经典算法，异步优势行动者-评论家虽然能够利用异步方法替代经验重放机制来加快模型训练，但是无法将重点集中在图像的重要区域和重要特征上. 因此，为了提升 Agent 在基于视觉感知 DRL 任务上的表现性能，本文提出了一种基于双重注意力机制的异步优势行动者评论家算法(DAM-A3C). DAM-A3C 在传统 A3C 算法中引入了视觉注意力机制和特征注意力机制，视觉注意力机制能够帮助 Agent 将关注点聚焦于图像中有重要价值的区域，特征注意力机制则可以将 Agent 的焦点集中于图像的重要特征，帮助卷积神经网络有效提取出能够促进 Agent 做出近似最优决策的特征. 本文选取 10 种 Atari 2600 游戏作为训练对象，并以训练后的网络参数对该 10 种游戏进行评估，验证 DAM-A3C 在处理基于视觉感知的决策任务时的有效性. 实验结果表明 DAM-A3C 在训练过程中所获得的奖赏值都高于其它的 A3C 模型，尤其是在 StarGunner 游戏和 TimePilot 游戏中性能提升最为明显. 此外，本文以训练完成后的网络模型参数对 10 种游戏直接进行测试，结果表明新模型的平均每情节奖赏和最优奖赏均优于其他模型. 另一方面，所有模型在 10 种不同游戏环境中使用了同样的网络参数进行训练，因此 DAM-A3C 模型具有较强的泛化能力.

然而，基于双重注意力机制的 A3C 模型无法有效解决强化学习算法中稀疏奖赏的问题，导致模型在上述 Atari 2600 游戏中收敛速度较慢，无法保持良好的策略稳定性. 因此下一步的研究重点是考虑将优先级扫描算法引入到基于双重注意力机制的 A3C 算法中，在保证 Agent 将注意力聚焦于重要的图像区域和特征的前提下，可以利用优先级扫描来避免 Agent 过多探索无意义的环境状态，使得基于双重注意力机制的 A3C 算法在基于视觉感知的 DRL 任务中具有更好的决策性能，帮助深度强化学习方法在未来可以更好地应用到真实场景中.

参 考 文 献

[1] Yu Kai, Jia Lei, Chen Yu-Qiang, et al. Deep learning: Yesterday,

today, and tomorrow. Journal of computer Research and Development, 2013, 50(9): 1799-1804(in Chinese)
(余凯, 贾磊, 陈雨强等. 深度学习的昨天、今天和明天. 计算机研究与发展, 2013, 50(9): 1799-1804)

[2] Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada, 2013: 6645-6649

[3] Li Ya-Xiong, Zhang Jian-Qiang, Pan Deng, et al. A study of speech recognition based on RNN-RBM language model. Journal of Computer Research and Development, 2014, 51(9): 1936-1944(in Chinese)
(黎亚雄, 张坚强, 潘登等. 基于 RNN-RBM 语言模型的语音识别研究. 计算机研究与发展, 2014, 51(9): 1936-1944)

[4] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks//Proceedings of the International Conference on Neural Information Processing Systems. Nevada, USA, 2012: 1097-1105

[5] Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 2014, 115(3): 211-252

[6] Liang Shu-Fen, Liu Yin-Hua, Li Li-Chen. Face recognition under unconstrained based on LBP and deep learning. Journal on Communications, 2014, 35(6): 154-160(in Chinese)
(梁淑芬, 刘银华, 李立琛. 基于 LBP 和深度学习的非限制条件下人脸识别算法. 通信学报, 2014, 35(6): 154-160)

[7] Tai K S, Socher R, Manning C D. Improved semantic representations from tree-structured long short-term memory networks. Computer Science, 2015, 5(1): 36

[8] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1724-1734

[9] Sutton R S, Barto A G. Reinforcement Learning: An Introduction. Cambridge, USA: MIT Press, 1998

[10] Gao Yang, Zhou Ru-Yi, Wang Hao, et al. Study on an average reward reinforcement learning algorithm. Chinese Journal of Computers, 2007, 30(8): 1372-1378(in Chinese)
(高阳, 周如益, 王皓等. 平均奖赏强化学习算法研究. 计算机学报, 2007, 30(8): 1372-1378)

[11] Fu Qi-Ming, Liu Quan, Wang Hui, et al. A novel off policy $Q(\lambda)$ algorithm based on linear function approximation. Chinese Journal of Computers, 2014, 37(3): 677-686(in Chinese)
(傅启明, 刘全, 王辉等. 一种基于线性函数逼近的离策略 $Q(\lambda)$ 算法. 计算机学报, 2014, 37(3): 677-686)

[12] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of go without human knowledge. Nature, 2017, 550(7676): 354

[13] Mnih V, Kavukcuoglu K, Silver D, et al. Playing Atari with deep reinforcement learning//Proceedings of the Workshops at the 26th Neural Information Processing Systems 2013. Lake Tahoe, USA, 2013: 201-220

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

[15] Watkins C J C H. Learning from delayed rewards. Robotics & Autonomous Systems, 1989, 15(4): 233-235

[16] Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double Q-learning//Proceedings of the Workshops at the 30th AAAI Conference on Artificial Intelligence. Phoenix, USA, 2015: 2094-2100

[17] Schaul T, Quan J, Antonoglou I, et al. Prioritized experience replay//Proceedings of the Workshops at the 4th International Conference on Learning Representations. San Juan, Puerto Rico, 2016: 322-355

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

[19] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate//Proceedings of the International Conference on Learning Representations. San Diego, USA, 2015: 1-15

[20] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention//Proceedings of the International Conference on Machine Learning. Lille, France, 2015: 2048-2057

[21] Chorowski J K, Bahdanau D, Serdyuk D, et al. Attention-based models for speech recognition//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2015: 577-585

[22] Sorokin I, Seleznev A, Pavlov M, et al. Deep attention recurrent Q-network//Proceedings of the Workshops at the 26th Neural Information Processing Systems 2013. Lake Tahoe, USA, 2013

[23] Busoniu L, Babuska R, De Schutter B, et al. Reinforcement Learning and Dynamic Programming Using Function Approximators. Florida, USA: CRC Press, 2010

[24] Wiering M, Otterlo M V. Reinforcement Learning: State-of-the-Art. West Berlin, Germany: Springer Publishing Company, 2012

[25] Sutton R S, McAllester D A, Singh S P, et al. Policy gradient methods for reinforcement learning with function approximation//Proceedings of the Advances in Neural Information Processing Systems. Denver, USA, 2000: 1057-1063

[26] Kakade S. A natural policy gradient//Proceedings of the International Conference on Neural Information Processing Systems: Natural and Synthetic. Vancouver, Canada, 2001: 1531-1538

[27] Silver D, Lever G, Heess N, et al. Deterministic policy gradient algorithms//Proceedings of the International Conference on International Conference on Machine Learning. Beijing, China, 2014: 387-395

[28] Konda V R, Tsitsiklis J N. Actor-critic algorithms//Proceedings of the Advances in Neural Information Processing Systems. Denver, USA, 2000: 1008-1014

[29] Bhatnagar S, Ghavamzadeh M, Lee M, et al. Incremental natural actor-critic algorithms//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2008: 105-112

[30] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks//Proceedings of the Advances in Neural Information Processing Systems. Lake Tahoe, USA, 2012: 1097-1105

LING Xing-Hong, Ph. D., associate professor. His main research interests include machine learning and reinforcement learning research.

LI Jie, M. S. candidate. His main research interests include deep learning and deep reinforcement learning.

ZHU Fei, Ph. D., associate professor. His main research interests include machine learning and biomedical research.

LIU Quan, Ph.D., professor, Ph.D. supervisor. His main research interests include machine learning and intelligence information processing.

FU Yu-Chen, Ph. D., professor. His main research interests include reinforcement learning and artificial intelligence.

Background

Deep reinforcement learning (DRL) combines deep learning (DL) and reinforcement learning (RL). Using DL's feature representation capability and RL's decision ability, DRL can effectively extract data features and get optimal strategies in complex high-dimensional state space tasks. Such as deep Q-network and double deep Q-network, traditional DRL methods have the problem of excessive training time. In order to overcome the problem, asynchronous deep reinforcement learning (ADRL) combines the asynchronous method and the DRL algorithm, and makes traditional DRL models not need to store a large number of training samples and replay a certain batch of samples, thus greatly reducing the storage space and calculation cost. However, while dealing with high-dimensional state space tasks such as Atari 2600 games, some of valuable image areas and image features are often ignored. And the reason is that Agent's attention is focused on the entire input image and all features of the image. To solve the problem, we propose an asynchronous advantage actor-critic with double attention mechanisms (DAM-A3C). We select some classic Atari 2600 games as experimental objects to evaluate the performance of the new model. The experimental result shows that the new model has better performance than the traditional asynchronous advantage actor-critic algorithms in experimental tasks.

This paper is supported by the National Natural Science Foundation of China (61772355, 61303108, 61373094), the Jiangsu Province Natural Science Research University Major Projects (17KJA520004), the Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University (93K172014K04), the Suzhou Industrial Application of Basic Research Program Part (SYG201422), the Suzhou Livelihood Science and Technology Projects (SS201736). These projects aim to enrich the reinforcement learning and deep reinforcement learning theories and help these theories be better applied to real-life scenarios.