一种基于视觉注意力机制的深度循环 Q 网络模型

刘 全[1),2),3)] 翟建伟[1)] 钟 珊[1)] 章宗长[1),2)] 周 倩[1)] 章 鹏[1)]

[1)](苏州大学计算机科学与技术学院 江苏 苏州 215006)

[2)](软件新技术与产业化协同创新中心 南京 210000)

[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)

摘 要 由现代强化学习和深度学习相结合形成的深度强化学习方法是目前人工智能领域一个新的研究热点,已经在各种需要感知高维度原始输入数据和决策控制的任务中取得了实质性的突破.尤其是一种被称为深度 Q 网络的模型在处理诸如 Atari 2600 游戏这类趋于真实环境的复杂问题时表现出了和人类玩家相媲美的水平.然而,当存在有延迟的奖赏而导致需要长时间步规划才能优化策略的情形中,深度 Q 网络的表现就会急剧下降.这说明深度 Q 网络并不擅长解决战略性深度强化学习任务.针对此问题,文中使用带视觉注意力机制的循环神经网络改进了传统的深度 Q 网络模型,提出了一种较为完善的深度强化学习模型.新模型的关键思想有两点:一是使用双层门限循环单元构成的循环神经网络模块来记忆较长时间步内的历史信息.这使得 Agent 能够及时使用有延迟的反馈奖赏来正确地指导下一步的动作选择;二是通过视觉注意力机制自适应地将注意力集中于面积较小但更具价值的图像区域,从而使得 Agent 能够更加高效地学习近似最优策略.该文通过选取一些经典的 Atari 2600 战略性游戏作为实验对象来评估新模型的有效性.实验结果表明,与传统的深度强化学习模型相比,新模型在一些战略性任务上具有很好的性能表现和较高的稳定性.

关键词 深度学习;强化学习;深度强化学习;深度 Q 学习;循环神经网络;视觉注意力机制;人工智能

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2017.01353

A Deep Recurrent Q-Network Based on Visual Attention Mechanism

LIU Quan[1),2),3)] ZHAI Jian-Wei[1)] ZHONG Shan[1)] ZHANG Zong-Zhang[1),2)]
ZHOU Qian[1)] ZHANG Peng[1)]

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

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

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

Abstract Reinforcement Learning, as a subject of study for over more than fifty years, investigates how an autonomous agent can learn what to do to maximize a numerical reward signal from interaction with the world by balancing exploration of the environment with exploitation of knowledge gained via evaluative feedback, without relying on exemplary supervision of an omniscient teacher or complete models of the environment. Deep learning is a cutting-edge approach to machine learning that concerns with using multi-layer artificial neural networks to learn the complicated representations that are expressed in terms of simpler ones. Currently, Deep Reinforcement Learning formed by combining modern reinforcement learning with deep learning is becoming a new research hotspot

收稿日期:2016-04-17;在线出版日期:2016-12-05.本课题得到国家自然科学基金项目(61272005,61303108,61373094,61472262,61502323,61502329)、江苏省自然科学基金(BK2012616)、江苏省高校自然科学研究项目(13KJB520020,16KJB520041)、吉林大学符号计算与知识工程教育部重点实验室基金项目(93K172014K04)、苏州市应用基础研究计划工业部分(SYG201422,SYG201308)资助.**刘 全**,男,1969 年生,博士,教授,博士生导师,中国计算机协会(CCF)高级会员,主要研究领域为智能信息助理、自动推理和机器学习.E-mail:quanliu@suda.edu.cn.**翟建伟**,男,1992 年生,硕士研究生,主要研究方向为强化学习、深度学习和深度强化学习.**钟 珊**,女,1983 年生,博士研究生,主要研究方向为机器学习和深度学习.**章宗长**,男,1985 年生,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为部分感知的马尔科夫决策过程、强化学习和多 Agent 系统.**周 倩**,女,1992 年生,硕士研究生,主要研究方向为强化学习.**章 鹏**,男,1992 年生,硕士研究生,主要研究方向为连续空间强化学习.

in the Artificial Intelligence community, and has made substantial breakthroughs in a variety of tasks—such as robot control, text recognition and games—requiring both rich perception of high-dimensional raw inputs and policy control. In particular, a state-of-the-art deep reinforcement learning model, termed Deep Q-Network, is able to perform human-level control using the same network architecture and hyper-parameters for handling problems approaching real-world complexity such as some Atari 2600 games. However, Deep Q-Network's performance falls far below human level in situations that exist delayed rewards and require planning under uncertainty within long-time horizon to optimize strategies. This implies that Deep Q-Network is not good at controlling agents in strategic deep reinforcement learning tasks. To alleviate the issue, this paper proposes a novel deep reinforcement learning model by improving Deep Q-Network with recurrent neural networks based on visual attention mechanism. Two key ideas are included in the new model: (1) it uses recurrent neural networks consisting of two-layer gated recurrent units in order to remember more historical information of multiple time steps. This can make an agent exploit delayed feedback in time to guide its next action selection online. By using recurrent neural networks, the scale of input state is reduced from four stacked images to one current raw image. This can substantially reduce the state space; (2) the visual attention mechanism is used to adaptively focus attention on smaller but more valuable regions of an input image, and make agents control the process of learning near optimal policies more effectively. As a result, the number of parameters to be learned by a stochastic gradient descent method during training can be decreased sharply by introducing the visual attention mechanism. This can speed up the process of learning near optimal policies. This new model is actually equivalent to an encoder-decoder architecture, where the convolutional neural networks play an encoder role for extracting useful features, and the recurrent neural networks based on visual attention mechanism play the other. We used five challenging strategic tasks from the set of classic Atari 2600 games, i.e., Seaquest, Alien, Gopher, Asteroids, and Gravitar, to verify the effectiveness of the new model. Experimental results show that artificial agents generated through our new model surpass DQN and its variant's performance in terms of the average reward per episode, training speed, and policy stability on them, especially on the Seaquest and Gopher games.

Keywords deep learning; reinforcement learning; deep reinforcement learning; deep Q-learning; recurrent neural network; visual attention mechanism; artificial intelligence

1 引 言

近年来,深度学习(Deep Learning,DL)作为机器学习领域中的一个研究热点[1],已在计算机视觉[2-4]和语音识别[5,6]等领域取得了令人瞩目的成功.其基本思想是结合使用多层的网络结构和非线性变换,从高维的原始特征中抽取出维度较低且高度可区分的特征.因此,深度学习方法忽视了控制过程,更侧重于对事物的认知和表达.随着人工智能的飞速发展,越来越多的实际问题需要利用 DL 来提取大规模输入数据的特征,并以此特征为依据进行自我激励的学习,优化解决问题的策略.与基于有监督训练的深度学习不同,强化学习(Reinforcement Learning,RL)强调的是一种从环境状态到动作映射的自我学习过程:Agent 在每个时刻与环境交互并选择动作,环境对此动作做出反应,并到达新的状态,然后通过值函数评价每个状态或状态动作对的好坏,最终确定到达目标状态的最优策略[7].目前,强化学习已经广泛应用到仿真模拟、工业控制和博弈游戏等领域[8-10].在输入状态维度较低的决策问题中,通过传统的强化学习算法,Agent 通常能学习到一个近似最优策略.然而当状态空间维度很高时,例如输入状态表示一幅图片或一段视频数据时,强化学习方法会因为无法感知到良好的、抽象的输入特征而导致算法性能急剧下降.由此可见,DL 和

RL 有各自的优势和局限. 因此，将深度学习的感知能力和强化学习的决策能力相结合，产生了人工智能领域新的研究热点，即深度强化学习(Deep Reinforcement Learning，DRL).

Lange 等人[11]最先结合深度学习模型和强化学习方法，提出了一种深度自动编码器(Deep Auto-Encoder，DAE)模型. 不过该模型只被证明适用于状态空间维度较小的控制问题，比如以视觉感知为基础的格子世界任务. Abtahi 等人[12]用深度信念网(Deep Belief Network，DBN)作为传统强化学习中的函数逼近器，极大地提高了 Agent 的学习效率. Lange 等人[13]又进一步提出了深度拟合 Q 学习算法(Deep Fitted Q-learning，DFQ)，并将该算法运用于车辆控制. Mnih 等人[14,15]结合深度学习中处理图像数据最通用的卷积神经网络(Convolutional Neural Networks，CNNs)和传统强化学习中求解最优动作值函数的 Q 学习算法，提出了一种深度 Q 网络模型(Deep Q-Network，DQN)来近似表示动作值函数，并在基于 Atari 2600 的游戏平台上做了大量的实验. 结果表明，在大部分游戏上，DQN 的游戏表现已经赶上甚至超过了人类玩家的水平. 这说明在很多基于视觉感知的控制任务中，可通过深度 Q 网络来指导强化学习 Agent 求解近似最优策略.

深度 Q 网络在接近于真实场景的各类任务上表现出的强大适用性使得更多的人工智能研究者投入到深度强化学习算法的研究当中. Nair 等人[16]针对训练 DQN 耗时大的问题，开发出一种大型的并发式架构(Gorila)，从而缩短网络的训练时间. Van Hasselt 等人[17]提出一种深度双 Q 学习网络(Deep Double Q-Network，DDQN)，成功解决了过度乐观评估值函数的问题. Schaul 等人[18]在训练 DQN 时，使用基于优先级的经验回放机制替代等概率的抽样方式，提高有价值样本的利用率，使得 Agent 在一些 Atari 游戏上取得了更高的得分. Narasimhan 等人[19]第一次将循环神经网络(Recurrent Neural Networks，RNNs)结构中的长短期记忆单元(Long-Short Term Memory，LSTM)引入到 DQN 中，提出一种带有 LSTM 单元的深度 Q 网络模型(Deep Q-Network with Long-Short Term Memory，LSTM-DQN)，并且在一种文本类游戏上取得了不错的效果. 然而，LSTM-DQN 在一种文本游戏上取得的成功并不能说明该模型广泛适用于各类基于视觉感知的 DRL 任务. Hausknecht 等人[20]首次在 Atari 2600 游戏平台上验证了 LSTM-DQN 模型的通用性，实验结果表明 LSTM-DQN 在大多数游戏中的表现都达到了 DQN 的水平. 不过由于单层 LSTM 单元构成的 RNNs 能够记忆的历史信息相对有限，LSTM-DQN 在指导 Agent 玩一些难度较大的游戏时，性能并没有显著得到的提升.

在面对需要讲究战略性的 DRL 任务时，DQN 及上述这些变体的表现与人类的水平相差甚远[14,15]，例如深海探险(Seaquest)游戏. 这是因为这类模型的输入状态是由堆叠离当前时刻最近的 4 幅连续视频帧(DQN)或单层 LSTM 单元记忆的相关历史信息(LSTM-DQN)而构成，Agent 能够感知的历史信息相对有限. 然而，对于战略性任务一个动作所带来的反馈可能在数十步甚至上百步之后才能在值函数中体现出来，所以需要跨较长的时间步去规划 Agent 的策略. 简单来讲，就是某个时刻选择 Agent 的动作需要综合考虑之前很多步的状态信息. 因此，对于此类战略性任务，通过有限时间步的历史画面来判断当前形势并做出决策，会存在着部分有价值信息不可感知的问题. 比如在国际象棋游戏中，仅根据 Agent 最近连续几次走子时的盘面构成的状态信息来判断当前局势并控制下一步走棋是远远不够的. 为了提高强化学习 Agent 在战略性任务上的性能，可以增加输入中连续视频帧的数目以缓和部分状态的不可观测性问题，但这会增加状态空间的维度，带来严重的计算负担. 本文通过在 DQN 中引入由双层门限循环单元(Gated Recurrent Units，GRU)构成的 RNNs 来记忆较长时间段内的历史信息，缓解了部分有价值状态的不可观测性问题，使得 Agent 面对战略性任务时能够及时得到延迟奖赏的反馈.

另一方面，受到人类视觉机理的启发，近年来基于注意力机制(Attention Mechanism，AM)的循环神经网络越来越多地被应用到机器语言翻译、图像识别和主题生成等领域[21-23]. Bahdanau 等人[21]基于传统的编码器-解码器(Encoder-Decoder)框架，利用 AM 使得模型可以自动搜索与目标输出相关联的源输入语句，在英语到法语的翻译任务上取得了令人满意的效果. Mnih 等人[22]提出了一种新颖的 RNNs 模型，该模型通过自我强化的 AM 自适应选择并提取图像中的重点区域或位置点的相关特征，提升了对图像的识别率. Xu 等人[23]首次将 AM 运用到视频图像数据的处理上，提出了视觉注意力

机制(Visual Attention Mechanism,VAM).在每个时刻,通过 VAM 将注意力集中于对识别主题有促进作用的图像区域,提高了模型识别图片主题的正确率.总的来说,通过各类任务的目标导向作用,VAM 以"高分辨率"的形式将 Agent 聚焦于图片中具有丰富信息的特定区域或像素位置,并通过训练不断调整其聚焦的区域,最终以较少的训练数据、较快的训练速度取得了更好的性能.本文在模型中加入 VAM,使得 Agent 在每个时刻能够自适应地将注意力集中于对提升累计奖赏有促进作用的图像区域,从而在较短时间内完成对 RNNs 记忆的多时间步内关键信息的感知.这大大减轻了网络训练时的运算代价,并加速了 Agent 学习近似最优策略的进程.

本文提出了一种基于视觉注意力机制的深度循环 Q 网络模型(Deep Recurrent Q-Network with Visual Attention Mechanism,VAM-DRQN).该模型主要在原有的 DQN 基础上做了两方面的改进:(1)引入由双层 GRU 构成的 RNNs 来记忆时间轴上的多步序列信息,缓解了 Agent 在解决战略性任务时存在的部分有价值状态不可观测的问题;(2)通过在模型中加入 VAM,Agent 能够在训练过程中自适应地将注意力集中于当前画面中对学习更具价值的区域,从而直观地、在线地做出正确的决策.并将 VAM-DRQN 应用于 Atari 2600 中的一些战略性游戏上,实验结果表明,通过 VAM-DRQN 模型的确能够提升强化学习 Agent 在战略性任务上的表现.

2 背景知识

2.1 强化学习

强化学习是一种从环境状态映射到动作的学习,目标是使 Agent 在与环境的交互中获得最大累积奖赏.马尔可夫决策过程(Markov Decision Process,MDP)可以用来对强化学习问题进行建模.通常将 MDP 定义为一个四元组(X,U,ρ,f),其中:

(1)X 为所有环境状态的集合,且 $x_t\in X$ 表示 Agent 在 t 时刻所处的状态;

(2)U 为 Agent 所有可执行动作的集合,且 $u_t\in U$ 表示 Agent 在 t 时刻所采取的动作;

(3)$\rho:X\times U\to R$ 为立即奖赏函数,表示 Agent 在 t 时刻位于状态 x_t 下执行动作 u_t 获得的立即回报值 r_t,可以表示为 $r_t\sim\rho(x_t,u_t)$;

(4)$f:X\times U\times X\to[0,1]$ 为状态转移函数,表示 Agent 在 t 时刻位于状态 x_t 下执行动作 u_t 转移到下一状态 x_{t+1} 的概率,可以表示为 $x_{t+1}\sim f(x_t,u_t)$.

在强化学习中,策略 $h:X\to U$ 是状态空间到动作空间的一个映射,表示为 $u_t\sim h(x_t)$,指 Agent 在时刻 t 位于状态 x_t 下采取动作 u_t.比如 Agent 在控制 Atari 2600 游戏过程中,状态 x_t 是由距时刻 t 最近的 M 幅视频帧组成:$x_t=(s_{t-M+1},\cdots,s_t)\in X$.然后,Agent 从离散的动作集中采取一个动作 $u_t=\{1,\cdots,K\}\in U$,产生新的游戏画面则代表 Agent 转移到下一状态 x_{t+1},并根据两幅画面中得分的差值确定奖赏信号 r_t.假设未来奖赏在每一个时间步都要乘以一个折扣因子 γ,所以从 t 时刻开始到 T 时刻情节结束时的奖赏之和定义为

$$R_t=\sum_{t'=t}^{T}\gamma^{t'-t}r_{t'} \tag{1}$$

其中 $\gamma\in[0,1]$,用来权衡未来奖赏对累积奖赏的影响力.

状态动作值函数 $Q^h(x,u)$ 指的是在当前状态 x_t 下执行动作 u_t,并一直遵循策略 h 到情节结束这一过程中 Agent 获得的累积回报,可以表示为

$$Q^h(x,u)=E[R_t\mid x_t=x,u_t=u,h] \tag{2}$$

对于所有的状态,如果一个策略 h^* 的期望回报大于或等于其它所有策略的期望回报值,那么该策略 h^* 就被称为最优策略.最优策略可能不止一个,但是它们共享相同的状态动作值函数:

$$Q^*(x,u)=\max_h E[R_t\mid x_t=x,u_t=u,h] \tag{3}$$

这被称为最优状态动作值函数,而最优状态动作值函数已经被证明是遵循最优贝尔曼方程的[7].通俗点讲,也就是如果下一状态 x' 在下个时间步上的关于所有动作的 Q 函数都是已知的,那么最优策略就是选择可以最大化期望回报 $r+\gamma Q^*(x',u')$ 的动作:

$$Q^*(x,u)=E_{x'\sim X}[r+\gamma\max_{u'}Q(x',u')\mid x,u] \tag{4}$$

在传统的 RL 中,贝尔曼方程发挥着举足轻重的作用.一般通过迭代贝尔曼方程,可得到

$$Q_{i+1}(x,u)=E_{x'\sim X}[r+\gamma\max_{u'}Q_i(x',u')\mid x,u] \tag{5}$$

其中,当 $i\to\infty$ 时,$Q_i\to Q^*$.即通过不断地迭代会使状态动作值函数最终收敛,从而得到最优策略.然而,对于解决实际问题,通过迭代上式求解最优策略显然很不可取.因为在大规模状态空间下,以表格式方法通过迭代贝尔曼方程求解各个状态动作对的值

函数的计算代价太大.所以在求解最优策略之前,需要先对状态空间进行泛化以降低其维度.

在 RL 算法中,通常是构造线性函数逼近器来近似表示状态动作值函数,$Q(x,u|\boldsymbol{\theta})\approx Q^*(x,u)$.当然也可以用非线性函数逼近器去近似表示值函数,例如神经网络,但这使得算法的性能很不稳定[24],该问题一直阻碍着深度强化学习的发展.

2.2　深度 Q 网络

最近,Mnih 等人[14,15]结合 CNNs 和传统 RL 中的 Q 学习算法,提出了一种名为深度 Q 网络的 DRL 算法,一定程度上解决了用非线性函数逼近器近似表示值函数时算法的不稳定性问题.如图 1 所示,DQN 主要在传统的 Q 学习算法上作了两处修改.

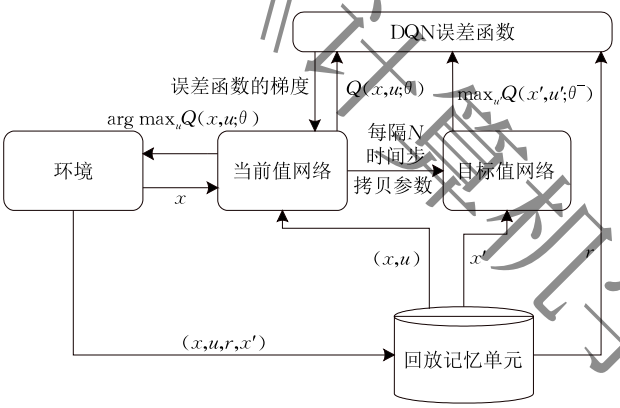

图 1　深度 Q 网络的原理图

首先,DQN 使用两个不同的 CNNs 模型分别近似表示状态动作对的当前值函数和目标值函数:$Q(x,u|\boldsymbol{\theta}_i)$代表当前值网络的输出,用来评估当前状态动作对的值函数;$Q(x,u|\boldsymbol{\theta}_i^-)$代表目标值网络的输出,一般用 $Y_i=r+\gamma\max_{u'}Q(x',u'|\boldsymbol{\theta}_i^-)$近似表示值函数的优化目标即最优值函数.当前网络的权值 $\boldsymbol{\theta}$ 是实时更新的,每经过 N 轮迭代,将当前网络的权值复制给目标值网络.初始时,$\boldsymbol{\theta}^-=\boldsymbol{\theta}$.通过优化目标值函数和当前值函数之间的均方误差函数来更新网络的权值.其中,误差函数如下:

$$L_i(\boldsymbol{\theta}_i)=E[(Y_i-Q(x,u|\boldsymbol{\theta}_i))^2] \tag{6}$$

其次,DQN 使用了经验回放机制[25](Experience Replay)在线处理训练过程中得到的转移样本.在每个时间步 t,将 Agent 与环境交互得到的转移样本 $e_t=(x_t,u_t,r_t,x_{t+1})$存放到回放记忆单元 $D_t=\{e_1,\cdots,e_t\}$.训练值网络时,每次从 D 中随机采样数量固定的转移样本(mini-batch),并且使用随机梯度下降(Stochastic Gradient Descent,SGD)算法更新网络的参数 $\boldsymbol{\theta}$.对式(6)关于参数 $\boldsymbol{\theta}$ 求偏导得到以下梯度:

$$\nabla_{\boldsymbol{\theta}_i}L(\boldsymbol{\theta}_i)=E_{x,u,r,x'}[(Y_i-Q(x,u|\boldsymbol{\theta}_i))\nabla_{\boldsymbol{\theta}_i}Q(x,u|\boldsymbol{\theta}_i)] \tag{7}$$

2.3　门限循环单元

循环神经网络是一种可以对时间维度的序列数据进行显示建模的深度学习模型.该模型通过添加跨越时间点的自连接隐藏层,使得网络中的处理单元之间既有内部的反馈连接又有前馈连接,从而记录动态的时序行为.即在 RNNs 中隐藏层的反馈不仅仅进入输出端,还进入到下一时间步的隐藏层中.传统的 RNNs 首先将输入序列$(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_t)$映射为隐藏层状态$(\boldsymbol{h}_1,\cdots,\boldsymbol{h}_t)$,然后再根据以下的转换关系得到输出$(\boldsymbol{z}_1,\cdots,\boldsymbol{z}_t)$:

$$\boldsymbol{h}_t=f(\boldsymbol{W}_{xh}\boldsymbol{x}_t+\boldsymbol{W}_{hh}\boldsymbol{h}_{t-1}) \tag{8}$$

$$\boldsymbol{z}_t=g(\boldsymbol{W}_{zh}\boldsymbol{h}_t) \tag{9}$$

其中 f 和 g 是非线性激活函数,常见的有 S 形生长函数(Sigmoid)和双曲正切函数(tanh),$\boldsymbol{W}_{ij}$ 代表连接着不同层神经单元之间的权值.

然而,当开始记忆时序信息的时间步与当前时刻的间距较大时,会产生关于时间轴的“梯度弥散”问题.即对于 t 时刻产生的误差信号在沿着时间轴向前传播几个时间步之后趋近于零,从而阻碍了网络参数的更新.

为了解决这种长期依赖问题,可以运用由 Hochreiter 等人[26]提出的长短期记忆模型来替代传统的循环神经网络模型.LSTM 在模型内部增加 4 种类型的门(Gate)结构,通过控制它们的开关实现记忆较长时间步状态的功能.经过若干年的发展,LSTM 衍生出了不少变体,并在结构上作了进一步的简化与改进.本文提出的 VAM-DRQN 模型中 RNNs 部分运用了其中一个较流行的变体,被称作门限循环单元[27],其结构如图 2 所示.

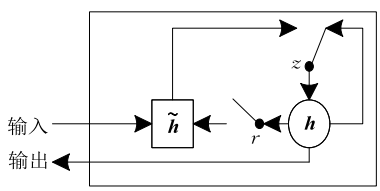

图 2　门限循环单元结构图

GRU 相对于 LSTM 来说参数更少,所以更不容易出现过拟合问题.GRU 通过代表不同功能门结构的组合,使得数据可以选择式地通过网络传播到下一阶段,从而输出更加抽象、紧凑的多时间步内的状态信息.

GRU 在 t 时刻的激活值 $\boldsymbol{h}_t$ 是上一时间步激活值 $\boldsymbol{h}_{t-1}$ 和候选激活值 $\tilde{\boldsymbol{h}}_t$ 的线性组合：

$$\boldsymbol{h}_t=(1-\boldsymbol{z}_t)*\boldsymbol{h}_{t-1}+\boldsymbol{z}_t*\tilde{\boldsymbol{h}}_t \tag{10}$$

其中，$\boldsymbol{z}_t$ 表示更新门的输出. 更新门用 S 形生长函数 $\sigma(\cdot)$ 激活输入信息 $\boldsymbol{x}_t$ 和上一时间步的激活值 $\boldsymbol{h}_{t-1}$ 的线性组合值. 得到的输出决定更新过去信息的程度：

$$\boldsymbol{z}_t=\sigma(\boldsymbol{W}_{xz}\boldsymbol{x}_t+\boldsymbol{W}_{hz}\boldsymbol{h}_{t-1}+\boldsymbol{b}_z) \tag{11}$$

候选激活值 $\tilde{\boldsymbol{h}}_t$ 的计算方式类似于传统的循环神经网络单元：

$$\tilde{\boldsymbol{h}}_t=\tanh(\boldsymbol{W}_{xh}\boldsymbol{x}_t+\boldsymbol{W}_{hh}(\boldsymbol{r}_t*\boldsymbol{h}_{t-1})+\boldsymbol{b}_h) \tag{12}$$

其中，$\boldsymbol{r}_t$ 表示重置门. 当 $\boldsymbol{r}_t$ 接近于 $\boldsymbol{0}$ 时，GRU 选择“忘记”之前的激活值并用当前的输入重置状态. 反之当 $\boldsymbol{r}_t$ 接近于 1 时，表示记忆之前全部的激活信息. 重置门的计算方式 $\boldsymbol{r}_t$ 如下：

$$\boldsymbol{r}_t=\sigma(\boldsymbol{W}_{xr}\boldsymbol{x}_t+\boldsymbol{W}_{hr}\boldsymbol{h}_{t-1}+\boldsymbol{b}_r) \tag{13}$$

其中，式(10)～式(13)中的 σ 表示非线性的 S 形生长函数，符号“$*$”表示矩阵间的内积操作符，$\boldsymbol{W}_{ij}$ 表示不同结构之间的连接权值矩阵.

3 一种基于视觉注意力机制的深度循环 Q 网络模型

本节将主要阐述 VAM-DRQN 模型的具体架构及训练方法. 其中，3.1 节具体分析 VAM-DRQN 各模块的构成及处理数据的流程，3.2 节详细介绍模型更新参数的方法.

3.1 模型架构

如图 3 所示，基于视觉注意力机制的深度循环 Q 网络模型主要由 CNNs、VAM 和 RNNs 这 3 个模块组成. 该模型可以看成是对高维原始输入数据编码之后再解码成低维抽象特征的一个过程，因此可以通过编码器-解码器框架以 Atari 2600 游戏为例具体分析 VAM-DRQN 中各个模块的作用及处理输入数据时各模块之间的关联性.

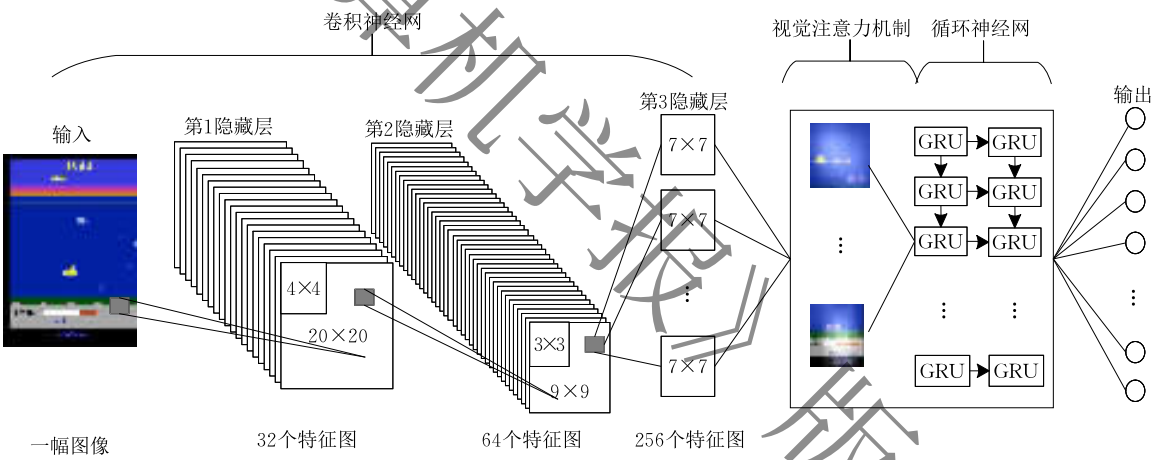

图 3 基于视觉注意力机制的深度循环 Q 网络模型

3.1.1 预处理

处理图像数据时，一般要通过各种预处理操作来消除图像中的无关信息以增强有价值特征的可检测性，从而最大限度地减小数据的复杂度. 由于 Atari 2600 游戏画面的尺寸大小为 210×160，如果直接将原始图像输入到网络中，处理数据所需的计算和存储代价过大. 这时就需要特定的预处理操作对原始图像进行处理. 首先，将原始的三原色(RGB)图像转换成灰度图；然后，通过降采样方法生成对应规模为 110×84 的缩略图；最后，去除边界一些无价值的像素点，将此缩略图裁剪为 84×84 大小. 该规模的图像能完整捕捉到当前的游戏画面，所以预处理并不会导致有价值特征的流失. 此外，该预处理操作仅需要对图像进行简单的灰度转换、裁剪和降采样，所耗费的计算资源远少于直接用 CNNs 处理原始图像的情形. 综上所述，该预处理操作是有必要进行的.

DQN 是将离当前时刻最近的 4 幅原始游戏画面经过预处理之后输入到网络模型中，因此输入状态的规模为 4×84×84. 本文提出的 VAM-DRQN，引入了 RNNs 来存储游戏过程中多时间步的历史状态信息，因此只需要将该预处理操作运用于当前的一幅游戏画面中. 也就是在每个训练时间步，输入状态的规模仅仅是 1×84×84，这使得状态空间的大小缩小至原来 DQN 的 1/4，大大减小了更新网络所需的计算量，加快了 Agent 的学习速度.

3.1.2 编码器：卷积神经网络

VAM-DRQN 以 CNNs 作为编码器，将高维的

输入数据编码成一系列低维的、抽象的特征表达. 从图 3 可直观地看出，CNNs 模块从输入到输出进行了 3 次卷积和 1 次映射操作. 具体地：

(1) 输入经过预处理之后大小为 1×84×84 的当前游戏画面. 通过 32 个规模为 1×8×8 的卷积核以步长 4 对输入进行卷积操作，得到 32 幅大小为 20×20 的特征图(Feature Maps)作为第一隐藏层. 然后用矫正函数 max(0，x)对第一隐藏层进行非线性变换.

(2) 通过 64 个规模为 32×4×4 的卷积核以步长 2 对第一隐藏层的激活值进行卷积操作，得到 64 幅大小为 9×9 的特征图作为第二隐藏层. 同样地，通过非线性变换来激活第二隐藏层.

(3) 利用 256 个规模为 64×3×3 的卷积核以步长 1 对上一隐藏层的输出进行卷积操作，得到 256 幅大小为 7×7 的特征图作为第三隐藏层. 同理，也对第三隐藏层进行非线性变换，输出 m 幅大小为 $n\times n$ 的特征图，其中，$m=256$，$n=7$.

在 DQN 中，将逐层卷积得到的特征图直接通过全连接层输入到包含 512 个神经单元的下一隐藏层. 而 VAM-DRQN 则是将第三隐藏层输出的 n 幅特征图映射为一个特征向量集合. 该集合的每个向量元素代表各个特征图中不同位置上像素点的组合，t 时刻该向量集合为

$$\boldsymbol{a}_t=\{\boldsymbol{a}_t^1,\cdots,\boldsymbol{a}_t^N\} \tag{14}$$

其中，$\boldsymbol{a}_t^i\in\Re^m$，$N=n\times n$. 综上，经过 3 次卷积操作和一层映射变换，将原始的输入状态编码成不同特征图中对应位置像素点的集合，从而有利于解码器中的视觉注意力机制区分各个像素点的视觉重要性.

3.1.3　解码器：基于视觉注意力机制的循环神经网络

在解码之前，需要进一步处理编码器输出的向量集合 $\boldsymbol{a}_t$，得到用于关联解码器和编码器的上下文向量 $\boldsymbol{C}_t$. 传统形式的 $\boldsymbol{C}_t$ 表示每个时刻关于输入图像各相关特征的一个动态表示. 然而，在某些复杂的基于视觉感知的战略性 DRL 任务中，Agent 需要在短时间内完成对输入状态中关键特征的感知并依据此特征做出动作的选择. 此时如果将 Agent 的注意力集中于整幅输入图像，会延缓网络模型的解码速度，使得 Agent 在短时间内无法及时感知到能对正确决策有促进作用的部分有价值信息. 为了缓解此问题，本文创新性地将 Xu 等人[23]提出的视觉注意力机制引入到网络中，以 $\boldsymbol{a}_t$ 为输入，通过该机制计算新的上下文向量 $\boldsymbol{C}_t$，使得 Agent 在每个时刻可以自适应地将注意力集中于当前画面中面积相对较小但具有丰富信息的图像区域，从而加快模型解码的速度. 下面具体介绍 VAM 在 VAM-DRQN 模型中的工作流程：

(1) 通过该模块中前两个全连接层计算向量集合 $\boldsymbol{a}_t$ 中各个像素点的视觉重要性：

$$vam(\boldsymbol{a}_t^i,\boldsymbol{h}_{t-1})=Linear(Tanh(Linear(\boldsymbol{a}_t^i)+\boldsymbol{W}\boldsymbol{h}_{t-1})) \tag{15}$$

其中：$Linear(x)=\boldsymbol{A}x+\boldsymbol{b}$ 是一种权值系数为 $\boldsymbol{A}$、偏移为 $\boldsymbol{b}$ 的线性仿射变换；$\boldsymbol{W}$ 为系数矩阵；$\boldsymbol{h}_{t-1}$ 为 RNNs 模块隐藏层的输出.

(2) 使用 Softmax 回归操作对第 1 步求得的结果进行归一化，得到每个像素点的相对视觉重要性：

$$\boldsymbol{\alpha}_t^i=\frac{\exp(vam(\boldsymbol{a}_t^i,\boldsymbol{h}_{t-1}))}{\sum_{k=1}^{N}\exp(vam(\boldsymbol{a}_t^k,\boldsymbol{h}_{t-1}))} \tag{16}$$

(3) 根据每个像素点的相对视觉重要性计算出最终要输入到 RNNs 模块的上下文特征向量 $\boldsymbol{C}_t$：

$$\boldsymbol{C}_t=\sum_{i=1}^{N}\boldsymbol{\alpha}_t^i\boldsymbol{a}_t^i \tag{17}$$

由上述过程可知，VAM 模块得到的上下文特征向量 $\boldsymbol{C}_t$ 代表 $\boldsymbol{a}_t$ 中所有像素点关于相对视觉重要性的一个线性加权. 根据 Donahue 等人[28]关于 RNNs 结构布局方面的研究，当使用双层记忆单元组件构成的 RNNs 作为模型中的解码器来解码上下文特征时，其记忆时序信息的效果比用单层或 4 层组件的网络更明显. 所以在 VAM-DRQN 中，首次将双层 GRU 构成的 RNNs 模块引入到深度 Q 网络中，用来对 VAM 生成的上下文特征向量 $\boldsymbol{C}_t$ 进行解码. 通过该方法，序列化处理一段时间内连续的部分有价值信息，从而使得 Agent 能够感知多个时间步上的关键信息. 从图 3 可以看出，RNNs 的每一层都是由 256 个的 GRU 组件构成.

由于 RNNs 是通过在网络中循环传递状态的方式处理时序数据，所以该模块的输入应由 VAM 处理得到的上下文特征向量 $\boldsymbol{C}_t$ 和上一时刻的隐藏层激活值 $\boldsymbol{h}_{t-1}$ 两部分组成. 下面具体分析 RNNs 处理数据的过程：

(1) 组合上下文特征向量和上一时刻的 RNNs 隐藏层激活值 $\boldsymbol{I}_t=\{\boldsymbol{C}_t,\boldsymbol{h}_{t-1}\}$，并将其作为当前时刻 RNNs 模块的第 1 层的输入；

(2) 将第 1 层的输出 $\boldsymbol{h}_t=\{\boldsymbol{h}_t^1,\cdots,\boldsymbol{h}_t^{256}\}$ 作为第 2 层的输入，其中的 $\boldsymbol{h}_t^i$ 代表 t 时刻 RNNs 模块中第 1 层的第 i 个 GRU 组件的隐藏层激活值；

(3) RNNs 模块输出当前时刻 t 更新后的隐藏

层激活值 $\boldsymbol{h}_t$. 该输出值有三方面的用途：

① 作为 RNNs 模块下一时间步 $t+1$ 时刻的输入激活值；

② 作为 VAM 模块在下一时间步 $t+1$ 时刻产生上下文向量 $\boldsymbol{C}_{t+1}$ 的输入；

③ 用来近似表示网络在当前输入状态下，Agent 采取各种可能动作 u_t 的 Q 值.

综上所述，VAM-DRQN 处理图像数据的流程如下：首先将 CNNs 用作编码器，接受当前的画面 x_t 作为输入，通过逐层的卷积和非线性变换将输入转化成 m 幅规模为 $n\times n$ 的特征图. 然后通过一个映射操作将 m 幅特征图转换到一个向量集合中 $\boldsymbol{a}_t=\{\boldsymbol{a}_t^1,\cdots,\boldsymbol{a}_t^N\}$，其中 $\boldsymbol{a}_t^i\in\Re^m$，$N=n\times n$，$\boldsymbol{a}_t$ 中的每个元素分别对应卷积不同图像区域后得到的抽象特征. 接下来通过基于 VAM 的 RNNs 进行解码，将向量集合 $\boldsymbol{a}_t$ 输入到 VAM 模块中，经过一系列操作得到上下文特征向量 $\boldsymbol{C}_t\in\Re^m$. 最后将 $\boldsymbol{C}_t$ 和上一时刻 RNNs 模块隐藏层激活值 $\boldsymbol{h}_{t-1}$ 一起作为双层 GRU 模块的输入，通过处理单元内部的自反馈及前反馈产生新的隐藏层激活值 $\boldsymbol{h}_t$. 该激活值既作为下一时间步调整视觉注意力的根据输入到 VAM 中，又要作为评估当前状态下 $\phi(x_t)$ 所有可采取动作的 Q 值的依据输入到模型的输出层，其中 $\phi(\cdot)$ 表示对原始图像的预处理操作.

3.2 网络的训练过程

VAM-DRQN 是一个平滑的、相互连接的模型，并且模型中的三大部分 CNNs、VAM 和 RNNs 也都是可微的，即该网络每个模块可训练的参数都存在关于自身的梯度. 所以本文使用了自适应学习率的随机梯度下降算法来对网络参数进行端对端(End-to-End)的更新.

由于 RL 在与环境的交互过程中沿时间轴产生的转移样本之间是高度相关的，而 DL 要求训练数据之间必须相互独立，所以在每个更新步都使用经验回放机制从样本池 D 中随机采样 mini-batch 数量的转移样本 $(\phi(x_j),u_j,r_j,\phi(x_{j+1}))$ 作为训练数据来更新网络的权值，其中 mini-batch 一般设置为 32. 另一方面，根据网络输出值和目标值之间的差值的平方项来构造误差函数，这也是训练深度网络模型最为关键的一步. 由 VAM-DRQN 的结构可知，当前网络的输出值为 $Q(\phi(x_j),u_j|\boldsymbol{\theta})$，代表当前状态 $\phi(x_j)$ 下，采取各种可能动作 u_j 的预期累积奖赏. 这里的目标值设定为

$$Y_t=E_{x_{j+1}\sim\varepsilon}[r_j+\gamma\max_{u_{j+1}}Q(\phi(x_{j+1}),u_{j+1})|\phi(x_j),u_j,\boldsymbol{\theta}_t^-] \tag{18}$$

所以误差函数的形式如下：

$$L_t(\boldsymbol{\theta}_t)=E_{\phi(x_j),u_j\sim h(\cdot)}[(Y_t-Q(\phi(x_j),u_j|\boldsymbol{\theta}_t))^2] \tag{19}$$

其中：ε 代表环境的先验分布，r_j 表示在 $\phi(x_j)$ 状态下采取 u_j 动作得到的立即奖赏，$\gamma\in[0,1]$ 是一个折扣因子，$h(\phi(x_j),u_j)$ 表示当前状态 $\phi(x_j)$ 下可采取动作 u_j 的分布，即 Agent 的行为策略. 这里采取的是 ε-greedy 策略，表示 Agent 以 $(1-\varepsilon)$ 的概率选择对应 Q 值最大的动作，以 ε 的概率随机选取一个动作来鼓励探索. 评估目标值 Y_t 时，采取对应 Q 值最大的贪心动作. 由于 Agent 行为策略与评估策略的不同，所以该方法是离策略(Off-Policy)的学习.

由于训练 VAM-DRQN 参数时采取的是端对端的形式，所以 $\boldsymbol{\theta}_t$ 代表当前 VAM-DRQN 中所有可训练的参数，$\boldsymbol{\theta}_t^-$ 代表阶段性固定的目标值网络的参数. 通过对误差函数关于参数 $\boldsymbol{\theta}_t$ 求偏导得到梯度，再使用标准的 Q 学习更新规则来更新网络的参数

$$\boldsymbol{\theta}_{t+1}=\boldsymbol{\theta}_t+\alpha(Y_t-Q(\phi(x_j),u_j|\boldsymbol{\theta}_t))\nabla_{\boldsymbol{\theta}_t}Q(\phi(x_j),u_j|\boldsymbol{\theta}_t) \tag{20}$$

以上就是 VAM-DRQN 参数学习的过程，训练过程中的一些细节设置将在下一节实验部分具体阐述.

4 实验及结果分析

本节首先介绍了实验依据的平台和实验过程中需要设置的参数，随后分别在训练期间和训练完成后，评估了 DQN、LSTM-DQN 和 VAM-DRQN 在一些 Atari 2600 战略性游戏中的表现，并结合实验结果分析说明 VAM-DRQN 的优势和适用范围.

4.1 实验平台描述

Bellemare 等人[29]开发了基于 Atari 2600 游戏环境的 Arcade Learning Environment(ALE)实验平台. ALE 提供了各式各样的 Atari 2600 游戏接口，游戏类型包括体育竞技类、桌游类和训练思维能力的战略性游戏等. 在 Agent 的学习过程中，仅仅将游戏过程中原始视频帧和由两个时刻游戏得分之差所构成的奖赏作为输入. 这为人工智能研究者提供了丰富且具有挑战性的 RL 问题.

DQN 等模型运用深度 Q 学习算法[14,15](Deep Q-learning，DQL)来训练 Agent，使其在 Atari 2600 上大部分游戏的得分赶上甚至超过了人类玩家. 但是对于需要经过长时间步的规划才能做出决策的游戏而言，这些模型的表现远远不能与专业的人类玩

家相比.为了提升 Agent 在战略性游戏上的性能，本文提出了 VAM-DRQN 模型，并选取 5 个单 Agent 的 Atari 2600 战略性游戏：深海探险(Seaquest)、星球大战(Alien)、萝卜保卫战(Gopher)、爆破彗星(Asteroids)和邪恶进攻(Gravitar)来设计实验，根据实验结果评估 VAM-DRQN 在战略性游戏上的表现，并与 DQN 和 LSTM-DQN 模型进行比较.

4.2　实验参数设置

首先强调一点，不同模型在训练过程中使用了相同的参数集合.另外，由于各个模型中包含的 CNNs 结构能够自动地学习到良好的特征表达，因此实验之前不需要针对特定任务去人工设计一些特征作为输入数据，而只需输入经过预处理后的游戏画面和游戏奖赏.这充分说明此类深度网络模型在解决基于视觉控制的 DRL 问题时是任务无关的，具有很强的泛化能力.

另外，实验中还运用了一些常用的技巧来提高模型在训练过程中的稳定性.Agent 是根据网络输出的 Q 值来选择下一步的动作，这一过程需要的时间远远大于网络一次前向传播的时间.所以实验中采用跳帧的技巧来缓和两者计算速度之间的差异.具体设置为：每隔 4 帧 Agent 才根据 ε-greedy 策略来选择下一步的动作，而在这之前只是重复执行当前的动作.由于不同游戏的得分区间有很大的差异，为了缩小 Q 值的范围，实验时将所有的正奖赏设置为 1，负奖赏设置为 −1，零奖赏保持不变.另外，实验中将梯度裁剪到[−5，5]区间，将误差项 $r+\gamma\max_{u'}Q(x',u'|\boldsymbol{\theta}_i^-)-Q(x,u|\boldsymbol{\theta}_i)$ 裁剪到[−1，1]区间.这是因为裁剪 Q 值、梯度和误差项到特定的区间有利于在不同游戏之间使用相同量级的学习率，并有效防止由于策略出现大的波动而导致结果发散或者陷入局部最优解，提高了训练时的稳定性.

本文使用基于均方根的随机梯度下降法(RMSProp)来更新参数，其中动量系数设置为 0.95.每次更新用的样本集都是通过经验回放机制从样本池 D 中随机抽取 mini-batch 个样本得到的，其中 mini-batch 的规模为 32.同时，折扣因子 γ 设置为 0.99，学习率 α 和行为策略 ε-greedy 的参数 ε 都设置为从游戏开始到 100 万幅视频帧区间内线性递减的形式，即学习率 α 从 0.005 降到 0.000 25，探索因子 ε 从 1.0 下降到 0.1.样本池 D 最大容量为 100 万个转移样本.值得注意的是，在训练的开始阶段，样本池 D 中并没有足够的转移样本来训练网络.实验中，在 25 000 更新步之前，Agent 先采取随机策略存储足够多的转移样本到样本池 D 中，以防止在学习的初期由于训练数据太少，而导致学习的偏向性.另外，LSTM-DQN 中 LSTM 单元和 VAM-DRQN 中 GRU 的隐藏层初始状态值都赋值为 **0** 向量.

4.3　实验评估与结果分析

深度网络模型的训练一般需要大量的训练数据和长时间的训练周期.在传统的 DL 方法中可以通过分阶段(Epoch)来评估网络模型的训练过程.而对于 RL，通常采用一个情节从开始到结束获得的累积奖赏作为评价标准.本文提出的 VAM-DRQN 是一种 DRL 模型，因此结合以上两种评估形式定义了一种新的度量学习过程的标准：模型训练过程中各阶段的平均每情节奖赏数.

Mnih 等人[14,15]训练 DQN 时采用了 200 个训练阶段，每个阶段包含 250 000 时间步的参数更新和 125 000 时间步的评估过程.借助于图形处理器(Graphic Processing Unit，GPU)，DQN 需要大约两周的训练周期.为了保证不同方法的参数一致性，各个模型都采用 100 个阶段作为训练周期.其中，每个阶段的规模设置为 50 000 时间步的更新参数过程和 25 000 时间步的评估过程.这样，借助 GPU 只需不到 48 h 就能训练出一个模型.训练周期缩短的原因有两个：(1) VAM-DRQN 模型引入了双层 GRU 构成的 RNNs 来记忆多时间步的历史信息，使得网络的输入状态仅是当前的一幅游戏画面，减小了状态空间的维度，并降低了问题的复杂度；(2) 通过视觉注意力机制使得 Agent 有选择性地将注意力集中于一幅图像中面积较小但具有丰富信息的区域，大大减小了网络可训练参数的数目，从而加速了网络的训练速度.在评估不同阶段的训练模型时采取的是 ε-greedy 行为策略，其中 ε 设置为 0.05 并一直保持不变.

实验首先比较了 3 种模型在训练 Agent 玩 Seaquest 游戏过程时各阶段平均每情节获得的奖赏数.从图 4 可以看出，VAM-DRQN 的训练效果最优，并且与另外两种模型之间的性能差距会随着训练时间的推移而变大.而 LSTM-DQN 模型的训练效果是稍优于 DQN 的，这是因为相比于 DQN 在每个时刻只能感知离当前时刻最近的连续 4 幅图像，LSTM-DQN 中由单层 LSTM 单元构成的 RNNs 能记忆的历史信息相对较多，所以延迟奖赏反馈到

Agent 的可能性也大一点. 另外从图中还可以看出，VAM-DRQN 模型的训练效果是远远优于另外两种模型的. 下面结合 Seaquest 游戏，对不同模型在训练时存在较大性能差异的原因展开具体分析.

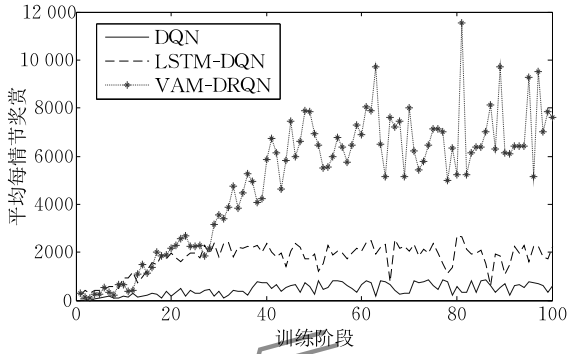

图 4 Seaquest 实验中采取不同模型训练时，各阶段中平均每情节奖赏对比

在 Seaquest 的游戏过程中，有些动作带来的效益在很多时间步之后才体现在游戏画面中并被 Agent 所感知. 比如，潜水艇在氧气不足时浮上水面储备氧气的动作. 所以当输入状态信息仅仅是由离当前时刻最近的 4 幅连续游戏画面(DQN)或单层 LSTM 单元记忆的有限步长内的历史信息(LSTM-DQN)所构成时，Agent 很可能就无法及时地得到有延迟的奖赏，从而阻碍了学习的过程. 本文提出的 VAM-DRQN 通过双层 GRU 构成的 RNNs 模块来记忆较长时间步内的历史信息，有效缓解了动作回报信号的延迟问题. 不仅如此，VAM-DRQN 中还引入了视觉注意力机制，这使得 Agent 在每个更新步都能有针对性地选择将注意力集中于对决策具有直接引导作用的一块面积较小的图像区域，从而减小了网络可训练的参数个数，进一步促进了 Agent 学习近似最优策略的过程.

从图 4 还可以看出，在 VAM-DRQN 的训练过程中，各阶段平均每情节得到的奖赏是有所波动的. 这主要是因为模型在训练时，网络的参数在不断地更新. 那么即使参数产生很小的变动，输出的动作 Q 值也会发生改变，从而可能会导致下一阶段策略的分布发生很大的变化[14]. 不过总体而言，VAM-DRQN 在训练过程中，平均每情节所获得的奖赏是会随着 Agent 的训练持续增加的.

为了说明 VAM-DRQN 模型在训练过程中的稳定性，图 5 对比了 3 种模型在训练 Agent 玩 Seaquest 游戏时各阶段平均每情节的最大动作 Q 值. 这里的 Q 值表示在任何给定的状态下，Agent 采取当前策略到情节结束能够获得的折扣累积奖赏. 经过裁剪之后，Q 值的量级缩小了数倍，充分保证了学习过程的稳定性.

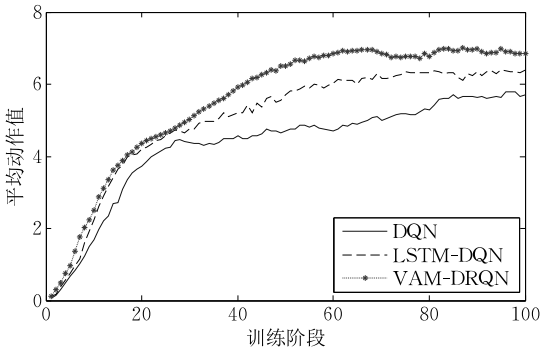

图 5 Seaquest 实验中采取不同模型训练时，各阶段中平均每情节最大状态动作值对比

从图 5 可以看出，VAM-DRQN 在训练过程中各阶段平均每情节的最大 Q 值曲线是优于 DQN 和 LSTM-DQN 的，这也是因为当用 VAM-DRQN 训练时，由双层 GRU 构成的 RNNs 记忆的较长时间步内的历史信息可以将有延迟的奖赏及时反馈给 Agent，促进了 Q 值函数的增长. 而对于 DQN 和 LSTM-DQN，由于存在部分有价值信息不可感知的问题，Agent 始终学习不到能够大幅提升游戏性能的关键策略. 比如在 Seaquest 游戏中当潜水艇的氧气不足时，DQN 和 LSTM-DQN 始终没有学会浮上水面去补充氧气这一关键策略，导致其游戏性能停滞不前. 同时，Q 值曲线保持平稳上升并趋向于收敛，这充分说明了 VAM-DRQN 模型在 DRL 任务中的有效性和稳定性.

另外，为了直观地说明在模型中引入 VAM 的作用，实验中可视化地分析了 Seaquest 游戏中视觉注意力机制工作的一段场景，如图 6 所示.

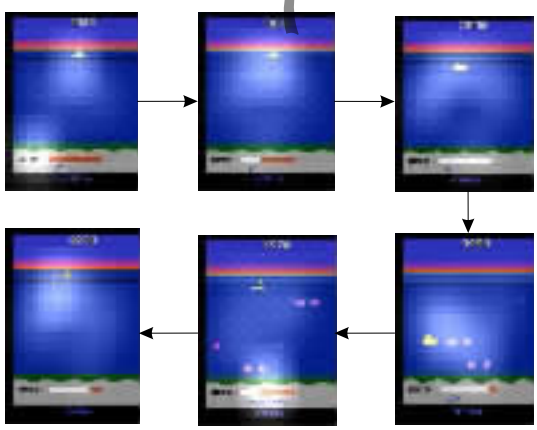

图 6 训练完成后 VAM-DRQN 中视觉注意力的可视化过程(其中白色阴影部分表示当前时刻 Agent 注意力集中的区域)

首先，从图 6 左上方中的一幅游戏画面可以看出，当潜艇的氧气刚耗光且已露出水面准备存储氧气时，此时的视觉注意力集中于画面下方的氧气指示器和潜艇自身. 随后，潜艇一直浮在水面储备氧气，该过程不需要再观测氧气探测仪，所以仅将注意力集中于潜艇上. 当氧气存储完毕之后，潜艇的注意力开始向两边及下方扩散，以保证再次潜水战斗的安全性. 从图 6 右下方画面中可以看出，当进入到战斗状态时，潜艇消耗氧气，Agent 将注意力转移到离自己最近的敌方所处的那部分区域中. 随着潜水时间的增加，氧气的消耗越来越多，此时注意力也更偏向于氧气指示器，当氧气不足时，Agent 开始浮上水面准备补充氧气. 在补充氧气的过程中，再次将注意力逐步转移到潜艇上. 总体上，视觉注意力机制使得 Agent 面对不同的处境时，能够正确地将注意力转移到有利于 Agent 快速做出决策的区域，直观地提升了 Agent 在此类战略性任务上的性能表现.

此外，为了进一步说明 VAM-DRQN 模型对于各类战略性游戏上的适用范围，本文也评估了 VAM-DRQN 训练 Agent 玩 Alien、Gopher、Asteroids 和 Gravitar 这 4 种战略性游戏时的表现. 图 7 针对上述各游戏，比较了不同模型训练时各阶段的平均每情节奖赏. 从图中可以看出，在 Alien 和 Gopher 游戏中，VAM-DRQN 的学习曲线和平均每情节奖赏都是明显优于另外两种模型的. 这充分说明在 DQN 的基础上引入带有 VAM 的 RNNs，能够提升 Agent 的学习性能，从而更好地解决一些基于视觉感知的战略性 DRL 任务. 然而从图 7 中也可以看出，在 Asteroids 和 Gravitar 游戏中，尽管通过训练 VAM-DRQN 模型在最后阶段取得了一些进展，但是总体上性能提升并不是很明显. 尤其是在 Gravitar 游戏上，平均每情节奖赏仍然存在着较大的波动. 这是由于 Asteroids 和 Gravitar 属于操作难度比较大的游戏，游戏场景的复杂性使得 Agent 在学习过程中得到的正向回报相对很少，从而阻碍了基于奖惩机制的学习. 另外，由于 Gravitar 游戏的操作难度相对更大，使得 Agent 在训练时需不断地重置游戏以进入新的情节. 而该游戏每次重启后的场景是随机变化的，这导致在新的游戏情节中，过去 VAM-DRQN 中通过 VAM 聚焦的那部分图像区域对 Agent 的学习意义不大. 因此 VAM-DRQN 模型还无法在该类游戏上取得突破. 总而言之，VAM-DRQN 模型更适用于有频繁的正向回报且游戏场景固定的一类战略性任务.

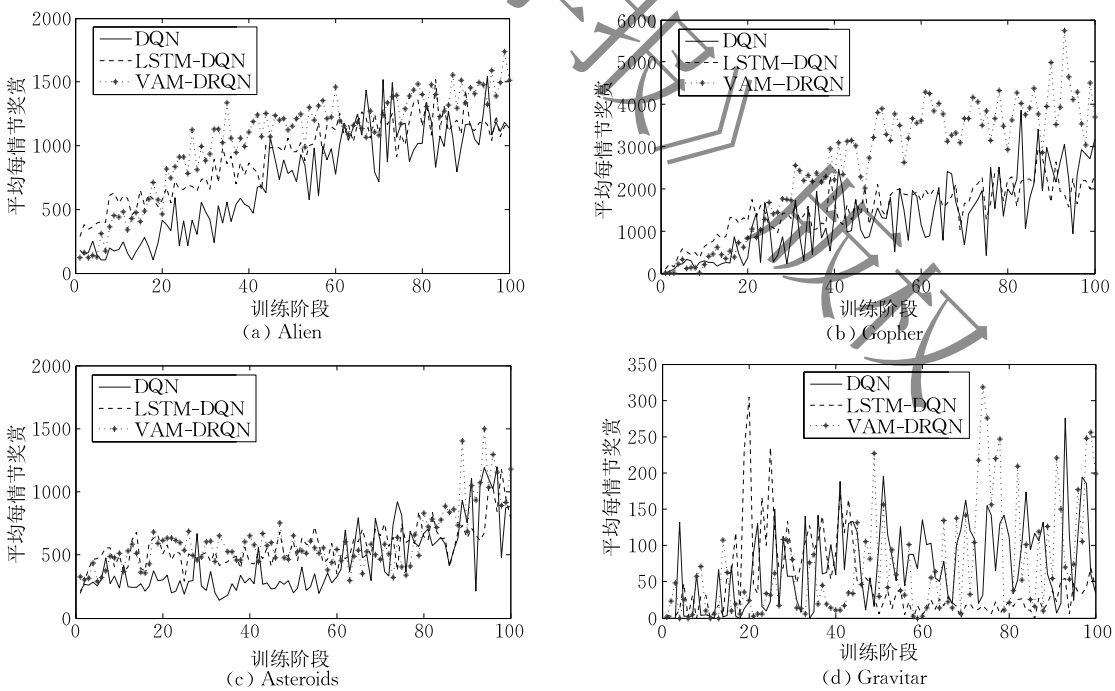

图 7　采用不同模型训练 Agent 玩 Alien、Gopher、Asteroids 和 Gravitar 时各阶段中平均每情节奖赏比较

另一方面，一个性能优异的深度网络模型不仅要在训练阶段表现出良好的性能，更重要的是能够通过训练得到一个完整的、可反复测试的策略模型，从而在每一次的 DRL 任务中，都可以依据该模型训练得到的策略来指导 Agent 取得优异的表现. 由不同模型在训练阶段的表现可以推断，训练完成后，VAM-DRQN 在指导 Agent 玩战略性游戏时的性能是优于其它模型的.

为了验证上述猜想，接下来将对比经过训练之后的不同模型在操作上述 5 种战略性游戏时的性能差异. 实验中，通过一个步长为 25 000 的游戏测试过程来评估训练完成后不同模型的性能好坏. 并且 Agent 所执行的策略是 ε-greedy 策略，采取这种策略是为了最小化过度拟合的可能性[15]. 其中，$\varepsilon=0.05$. 针对不同游戏，每个训练完成的模型都会被测试 50 次，且每次游戏时的初始状态都设置为不同，这充分保证了测试结果的多样性. 每次测试都获得一个得分，代表该次游戏的平均每情节所得奖赏. 一方面，实验中比较了不同游戏中各个模型 50 次测试的平均得分值. 另一方面，为了说明训练完成后模型的稳定性，实验中设置了 95%的置信区间来评估 50 次测试得分之间的差异性. 由于采用不同方法测试游戏时平均每情节得分在量级上有较大差距，比如 Seaquest 游戏中 VAM-DRQN 的得分均值，就远远大于 DQN 和 LSTM-DQN 的得分均值，因此可以通过置信等级这个评价标准来衡量模型性能的稳定性. 置信等级的定义如下

$$\text{置信等级}=\left(1-\frac{\text{置信上界}-\text{置信下界}}{\text{平均值}}\right)\times 100\% \quad (21)$$

由式(21)可知，置信等级越高，模型在游戏上的表现越稳定.

评估结果如表 1 所示. 从表中的平均值一列可以看出，与 DQN 和 LSTM-DQN 相比，每个训练完成后的 VAM-DRQN 模型在指导 Agent 玩战略性游戏时的表现均取得了一定幅度的提高. 从最大值一列中可以看出，训练完成后的 VAM-DRQN 在各游戏中的最优表现也基本优于其它模型. 尤其在 Seaquest 和 Gopher 游戏中，VAM-DRQN 的游戏表现接近甚至赶上了有经验的人类玩家. 不仅如此，由置信等级这一列的结果可以得出结论，训练完成后的 VAM-DRQN 模型不仅在游戏得分上已经超过了其它模型，而且能够在多次游戏中保持较好的策略稳定性.

表 1 训练完成后的不同模型在战略性游戏上的测试得分评估

游戏	Agent	平均值	标准差	最大值	置信下界	置信上界	置信等级/%
Seaquest	DQN	1106.2	470.6	3980.0	975.8	1236.6	76.4
	LSTM-DQN	2237.2	320.9	2640.0	2148.2	2326.2	92.1
	VAM-DRQN	8682.2	3076.6	17550.0	7829.4	9535.0	80.4
Alien	DQN	1461.4	379.4	2560.0	1356.2	1566.6	85.6
	LSTM-DQN	1501.8	458.5	2740.0	1374.7	1628.9	83.1
	VAM-DRQN	1704.2	606.0	4030.0	1536.2	1872.2	80.3
Gopher	DQN	2750.4	1024.7	4600.0	2466.4	3034.4	79.4
	LSTM-DQN	2932.0	1016.0	5820.0	2650.4	3213.6	80.8
	VAM-DRQN	5347.2	2196.2	13080.0	4738.4	5956.0	77.2
Asteroids	DQN	830.8	374.4	2250.0	727.0	934.6	75.0
	LSTM-DQN	1009.6	392.8	2350.0	900.7	1118.5	78.4
	VAM-DRQN	1185.2	473.2	3000.0	1054.0	1316.4	77.9
Gravitar	DQN	55.0	82.5	250.0	0.1	27.5	0.11
	LSTM-DQN	101.0	155.3	500.0	57.9	144.1	14.70
	VAM-DRQN	107.0	154.9	500.0	64.1	149.9	19.80

5 结束语

深度 Q 网络模型以及它的一些变体可以高效地求解各类基于视觉感知的 DRL 问题，其代表性的成功应用是 Atari 2600 游戏博弈. 然而，当面对的是存在延迟奖赏的战略性任务时，这类模型的表现就与人类相差甚远. 这是由于 Agent 需要在每个时刻考虑较长时间步的历史信息才能规划出较优的动作，而 DQN、LSTM-DQN 等模型能够记忆的历史状态信息是相对有限的. 为提高 Agent 应对战略性任务的能力，本文提出了一种基于视觉注意力机制的深度循环 Q 网络模型(VAM-DRQN). 一方面，通过在 DQN 的基础上引入由双层 GRU 构成的 RNNs 来记忆多时间步长的历史状态信息，这能在很大程度上缓解因延迟奖赏不能及时反馈给 Agent 的问题. 另一方面，Agent 通过视觉注意力机制自适应地将注意力集中于面积较小但更具价值的图像区域中，减少了网络可学习参数的总数，并提高了 Agent 学习的速率. 本文通过 5 个战略性 Atari 2600 游戏，验证了 VAM-DRQN 在面对此类战略性决策任务时的有效性. 实验结果表明该模型在训练过程中的平均每情节奖赏数和训练速度总体上优于 DQN 和 LSTM-DQN，尤其在 Seaquest 游戏中性能的提升最为明显. 进一步地，本文还评估了各模型训练完成之后的性能，结果表明训练好的 VAM-DRQN 模

型，不仅能提升 Agent 应对战略性任务的能力，并且还保持了相当稳定的性能表现. 值得注意的是，在不同的游戏中，Agent 只需要固定的一套参数、一个模型(VAM-DRQN)和一种算法(DQL)来学习近似最优策略，因此本文的方法具有较强的泛化能力.

然而，VAM-DRQN 在战略性游戏上的表现总体上还是达不到专业的人类玩家水平. 下一步的研究重点是考虑如何将 VAM-DRQN 和监督学习策略网络(Supervised Learning of Policy Networks)相结合，指导 Agent 更智能、快速地学会玩这类战略性游戏. 将监督学习和强化学习结合达到完美性能的典范是围棋机器人程序 AlphaGo[10]. 它先通过带标签的 3000 万幅专业棋手对局的棋谱来训练监督学习策略网络，使机器人学会了围棋高手的走棋方式，再通过深度强化学习方法进一步提升策略的质量，最终使得 AlphaGo 的棋力达到了世界冠军的水准. 类似地，为了进一步提高 Agent 在战略性任务时的性能，可以先使用监督学习训练出策略模型的初始参数，再通过基于奖惩机制的强化学习来更新策略.

参　考　文　献

[1] Yu Kai, Jia Lei, Chen Yu-Qiang, Xu Wei. 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，徐伟. 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] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks//Proceedings of the 26th Annual Conference on Neural Information Processing Systems. Nevada, USA, 2012: 1097-1105

[3] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 2015, 115(3): 211-252

[4] 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)

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

[6] Li Ya-Xiong, Zhang Jian-Qiang, Pan Deng, Hu Dan. 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，胡惮. 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)

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

[8] Gao Yang, Zhou Ru-Yi, Wang Hao, Cao Zhi-Xin. 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，曹志新. 平均奖赏强化学习算法研究. 计 算机学报，2007，30(8)：1372-1378)

[9] 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)

[10] Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search. Nature, 2016, 529(7587): 484-489

[11] Lange S, Riedmiller M. Deep auto-encoder neural networks in reinforcement learning//Proceedings of the 7th International Joint Conference on Neural Networks. Barcelona, Spain, 2010: 1-8

[12] Abtahi F, Fasel I. Deep belief nets as function approximators for reinforcement learning. Frontiers in Computational Neuroscience, 2011, 5(1): 112-131

[13] Lange S, Riedmiller M, Voigtlander A. Autonomous reinforcement learning on raw visual input data in a real world application//Proceedings of the 9th International Joint Conference on Neural Networks. Brisbane, Australia, 2012: 1-8

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

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

[16] Nair A, Srinivasan P, Blackwell S, et al. Massively parallel methods for deep reinforcement learning//Proceedings of Workshops at the 32nd International Conference on Machine Learning. Lille, France, 2015

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

[18] Schaul T, Quan J, Antonoglou I, Silver D. Prioritized experience replay//Proceedings of Workshops at the 4th International Conference on Learning Representations. San Juan, Puerto Rico, 2016

[19] Narasimhan K, Kulkarni T, Barzilay R. Language understanding for text-based games using deep reinforcement learning //Proceedings of Workshops at the 2015 Conference on Empirical Methods on Natural Language Processing. Lisbon, Portugal, 2015

[20] Hausknecht M, Stone P. Deep recurrent Q-learning for partially observable MDPs. Computer Science, 2015

[21] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. Eprint Arxiv: 1409.0473, 2014

[22] Mnih V, Heess N, Graves A. Recurrent models of visual attention//Proceedings of the 28th Annual Conference on Neural Information Processing Systems. Montreal, Canada, 2014: 2204-2212

[23] Xu K, Ba J, Kiros R, Courville A, et al. Show, attend and tell: Neural image caption generation with visual attention// Proceedings of Workshops at the 2015 Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015

[24] Tsitsiklis J N, Van R B. An analysis of temporal-difference learning with function approximation. IEEE Transactions on Automatic Control, 1997, 42(5): 674-690

[25] Lin L J. Reinforcement learning for robots using neural networks. USA: Defense Technical Information Center, DTIC Technical Report: ADA261434, 1993

[26] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780

[27] Cho K, Merrienboer B V, Bahdanau D, et al. On the properties of neural machine translation: Encoder-decoder approaches. Computer Science, 2014

[28] Donahue J, Anne H L, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 2625-2634

[29] Bellemare M G, Naddaf Y, Veness J, et al. The arcade learning environment: an evaluation platform for general agents. Journal of Artificial Intelligence Research, 2012, 47: 253-279

LIU Quan, born in 1969, Ph. D., professor, Ph. D. supervisor. His main research interests include intelligence information processing, automated reasoning and machine learning.

ZHAI Jian-Wei, born in 1992, M. S. candidate. His main research interests include reinforcement learning, deep learning and deep reinforcement learning.

ZHONG Shan, born in 1983, Ph. D. candidate, lecturer. Her research interests include machine learning and deep learning.

ZHANG Zong-Zhang, born in 1985, Ph. D., associate professor. His research interests include POMDPs, reinforcement learning and multi-agent systems.

ZHOU Qian, born in 1992, M. S. candidate. Her main research interest is reinforcement learning.

ZHANG Peng, born in 1992, M. S. candidate. His main research interest is continuous space reinforcement learning.

Background

Deep reinforcement learning, as a novel combination of reinforcement learning and deep learning in the artificial intelligence community, has achieved unprecedented progress in a variety of domains—involving both rich perception of high-dimensional raw inputs and action selection—such as robot control, text recognition and games. The Deep Q-network (DQN) is a state-of-the-art deep reinforcement learning method, and gains wide attention due to its excellent human-level performance over several Atari 2600 games. However, DQN's performance falls far below human level in some strategic games, especially those require long sequential decision making to find a near optimal solution. To be efficient in handling these kinds of strategic games, our paper proposes a novel deep reinforcement learning model, deep recurrent Q-network based on visual attention mechanism. Experimentally, our preliminary results demonstrate that training agents generated through our new model surpass DQN's performance on five strategic Atari 2600 games.

This paper is partially supported by National Natural Science Foundation of China (61272005, 61303108, 61373094, 61472262, 61502323, 61502329), the Natural Science Foundation of Jiangsu (BK2012616), the High School Natural Foundation of Jiangsu (13KJB520020, 16KJB520041), the Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, the Jilin University (93K172014K04), and the Suzhou Industrial Application of Basic Research Program Part (SYG201422, SYG201308). These projects aim to enrich the reinforcement-learning theory and develop efficient approximate algorithms to expand the power and applicability of reinforcement learning on large scale problems.