脉冲强化学习算法研究综述

陈 鼎[1),3)] 黄杨茹[2)] 彭佩玺[2),3)] 黄铁军[2)] 田永鸿[2),3)]

[1)](上海交通大学计算机科学与工程系 上海 200240)

[2)](北京大学计算机学院 北京 100871)

[3)](鹏城实验室 广东 深圳 518055)

摘 要 借助人工神经网络(Artificial Neural Network, ANN),深度强化学习在游戏、机器人等复杂控制任务中取得了巨大的成功.然而,在认知能力与计算效率等方面,深度强化学习与大脑中的奖励学习机制相比仍存在着巨大的差距.受大脑中基于脉冲的通信方式启发,脉冲神经网络(Spiking Neural Network, SNN)使用拟合生物神经元机制的脉冲神经元模型进行计算,具有处理复杂时序数据的能力、极低的能耗以及较强的鲁棒性,并展现出了持续学习的潜力.在神经形态工程以及类脑计算领域中,SNN受到了广泛的关注,被誉为是新一代的神经网络.通过将SNN与强化学习相结合,脉冲强化学习算法被认为是发展人工大脑的一个可行途径,并能够有效解释生物大脑中的发现.作为神经科学与人工智能的交叉学科,脉冲强化学习算法涵盖了一大批杰出的研究工作.根据对不同领域的侧重,这些研究工作主要可以分为两大类:一类是以更好地理解大脑中的奖励学习机制为目的,用于解释动物实验中的发现,并对大脑学习进行仿真,例如R-STDP学习规则;另一类则是以实际控制任务中的性能、功耗等具体指标为导向,用作人工智能的一种鲁棒且低能耗的解决方案,在机器人、自主控制等领域具有巨大的应用潜力.本文首先介绍了脉冲强化学习算法的基础(即脉冲神经网络以及强化学习),然后对当前这两大类脉冲强化学习算法的研究特点与研究进展等进行分析.对于第一类算法,本文重点分析了利用三因素学习规则实现的强化学习算法,并回顾了其生理学背景以及具体实现方式.根据在训练过程中是否使用ANN,本文将第二类算法分为依托ANN实现的脉冲强化学习算法与基于脉冲的直接强化学习算法,并率先对这一脉冲强化学习算法的最新进展进行了系统性的梳理与分析,同时全面展示了在深度强化学习算法中应用SNN的不同方式.最后,本文对该领域的研究挑战以及后续研究方向进行了深入地探讨,总结了当前研究的优势与不足,并对其未来对神经科学以及人工智能领域可能产生的影响进行展望,以吸引更多研究人士参与这个新兴方向的交流与合作.

关键词 脉冲神经网络;强化学习;类脑智能;人工智能;神经形态工程

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2023.02132

Research on Spiking Reinforcement Learning Algorithms: A Survey

CHEN Ding[1),3)] HUANG Yang-Ru[2)] PENG Pei-Xi[2),3)] HUANG Tie-Jun[2)]
TIAN Yong-Hong[2),3)]

[1)](*Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai* 200240)

[2)](*School of Computer Science, Peking University, Beijing* 100871)

[3)](*Peng Cheng Laboratory, Shenzhen, Guangdong* 518055)

Abstract With the help of Artificial Neural Networks (ANNs), deep reinforcement learning algorithms have achieved great success in complex tasks, such as playing games and robotic control, etc. However, compared with the mechanism of reward-modulated learning in the brain, deep reinforcement learning still has a huge gap in cognitive ability and computational efficiency.

收稿日期:2022-08-25;在线发布日期:2023-04-12. 本课题得到广东省重点领域研发计划项目(2020B0101380001)资助. **陈 鼎**,博士研究生,主要研究领域为脉冲神经网络、强化学习、类脑计算. E-mail: lucifer1997@sjtu.edu.cn. **黄杨茹**,博士研究生,主要研究领域为类脑强化学习、计算机视觉. **彭佩玺**,博士,副研究员,主要研究领域为计算机视觉、强化学习等. **黄铁军**,博士,教授,中国计算机学会(CCF)会士,主要研究领域为视觉信息处理、类脑视觉. **田永鸿**(通信作者),博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为视频大数据分析处理、机器学习、类脑计算. E-mail: yhtian@pku.edu.cn.

Inspired by the spike-driven communication in the brain, Spiking Neural Networks (SNNs) adopt the spiking neuronal models for calculation and exchange information through discrete action potentials, i. e. , spikes, which greatly fit the mechanism of biological neurons. It is demonstrated by much research that SNNs have distinctive properties, such as complex time-series information processing capability, extremely low energy consumption, and strong robustness. In addition, SNNs also show the potential for continual learning. In the field of neuromorphic engineering and brain-inspired computing, SNNs are widely concerned and known as a new generation of neural networks. By combining SNNs with reinforcement learning, spiking reinforcement learning algorithms are considered as a feasible way to develop the artificial brain, and can effectively explain the discovery in the biological brain. As a cross-discipline research area of neuroscience and artificial intelligence, spiking reinforcement learning algorithms cover a large number of outstanding research works. According to the emphasis on different fields, these research works can be divided into two categories: one is to better understand the mechanism of reward-modulated learning in the brain, which is used to explain the findings in animal experiments and simulate the learning process of the brain, such as R-STDP learning rules; The other is to improve the performance, energy efficiency and other specific indicators of various control tasks requiring reinforcement learning algorithms to solve. With the unique advantages of SNNs, this kind of algorithm acts as a robust and energy-efficient solution for artificial intelligence, and shows great application potential in the fields of robotics and autonomous control. In this paper, the first part presents the cornerstone of spiking reinforcement learning algorithms, that is, spiking neural networks and reinforcement learning, and then analyzes the research characteristics and progress of these two kinds of spiking reinforcement learning algorithms. For the first kind of algorithms, this paper focuses on analyzing the reinforcement learning algorithms using the three-factor learning rules, and introduces their physiological background and specific implementation methods. Based on whether to use ANNs during training, this paper further divides the second kind of algorithms into spiking reinforcement learning algorithms using ANNs and spike-based direct reinforcement learning algorithms. As far as we know, this paper takes the lead in systematically sorting out and analyzing the latest progress of spiking reinforcement learning algorithms, and comprehensively shows the different ways of applying SNNs in deep reinforcement learning algorithms. Finally, this paper makes an in-depth discussion of the current challenges and follow-up research directions in this field. We systematically summarize the advantages and disadvantages of the current research, and look forward to its future impact on the field of neuroscience and artificial intelligence, so as to attract more researchers to participate in communication and cooperation in this new direction.

Keywords spiking neural network; reinforcement learning; brain-inspired intelligence; artificial intelligence; neuromorphic engineering

1 引 言

神经科学在人工智能(Artificial Intelligence, AI)发展史上扮演了重要的角色,许多经典神经网络结构的出发点都是为了理解大脑的工作机制[1-3]. 此外,神经科学不仅可以为已存在的AI技术提供生物学解释[4-7],还可以为构建人工大脑时所需的新算法与新架构提供丰富的灵感来源[8-10]. 近些年来,随着计算机算力的增强以及大数据的积累,以深度学习[11]为代表的人工智能领域得到了蓬勃的发展. 然而,现有的计算系统执行相同的任务所需要的能耗往往要比人脑高出至少一个数量级[12]. 因此,AI研究人员将目光转回大脑,对神经元之间脉冲驱动的

通信方式产生了极大的兴趣.在人脑的指引下,通过脉冲驱动通信实现的神经元-突触硬件计算系统有望解决当前深度学习算法面临的高能耗问题[13].这种神经形态计算技术[14]始于20世纪80年代,并在21世纪初期促成了大规模神经形态芯片的出现,例如IBM的TrueNorth芯片[15]、Intel的Loihi芯片[16]以及英国曼彻斯特大学的SpiNNaker芯片[17].通过采用存算一体的架构,神经形态芯片解决了传统冯•诺依曼计算架构中处理单元与存储单元物理分离(存算分离)的固有缺陷,从而减轻"内存墙瓶颈"对计算吞吐量和能源效率的影响,将硬件功耗降低到毫瓦级[13].

在硬件不断发展的同时,相关的算法也在不断协同演化.通过将生物神经元之间通信的稀疏脉冲信号和事件驱动的性质抽象为神经单元,生物学合理的脉冲神经元模型[18]被应用到神经网络之中,由此诞生了脉冲神经网络(Spiking Neural Network,SNN).SNN是为了弥合神经科学与机器学习之间的差异而设计的新一代神经网络[19],被认为是人工智能硬件实现的一种极具前景的解决方案[13].SNN与目前流行的神经网络和机器学习方法有着根本上的不同,即其使用脉冲,而非常见的浮点值进行学习.脉冲是一种发生在时间点上的离散事件,一般可以由0和1进行表示,与生物神经元中的动作电位(Action Potential)相对应.通常来说,SNN的输入和输出均为脉冲序列,神经元之间通过突触进行连接.理论分析表明,SNN在计算性能上与常规神经元模型相当[19].由于其处理复杂时序数据的能力、极低的能耗[13]以及深厚的生理学基础[20],SNN受到了广大学者的关注,在图像分类[21-23]、目标识别[24-25]、语音识别[26-27]以及其他领域[28-30]上取得了飞速的发展,展现出了极强的上升势头.最近的研究表明,SNN在许多领域接近或达到了与经典人工神经网络(Artificial Neural Network,ANN)相当的性能[21,27].相比ANN,SNN还表现出了较强的鲁棒性.首先,脉冲神经元动态中的随机性可以提高网络对外部噪声的鲁棒性[31].其次,近期有研究表明脉冲神经元的发放机制使得SNN之于对抗攻击存在内在的鲁棒性[32].此外,生物体的一生都在从与环境之间的交互中学习,而人工系统若要在现实世界中行动和适应,同样需要能够实现持续学习(Continual Learning)[33].为了解决这一难题,许多生物学启发的模型以及机制被应用到人工系统中,并取得了不错的效果[34].由于额外的时间维度,SNN被认为具有实现持续学习的潜力[35-36].

尽管深度学习在很多领域都取得了突破性的成就,达到甚至超过了人类水平,为SNN设下了很高的竞争门槛.研究表明,相比于目前已经较为成熟的计算机视觉任务,SNN能够在机器人、自主控制等领域取得优于深度学习的表现[13,37].在这些领域中,传统深度学习算法需要的大量计算资源在处理实际问题时往往难以满足,而借助专用的神经形态硬件,SNN能够极大地降低任务所需的能耗,这与移动设备上有限的主板能量资源之间具有天然适配性.

强化学习(Reinforcement Learning,RL)作为AI研究的一个重要分支,用于解决在智能体与环境交互过程中的序列决策问题,通过学习策略以实现期望未来奖励最大化,并且已经在广泛的控制任务上证明了其有效性[10,38-40].因此,通过将SNN与强化学习相结合,脉冲强化学习算法[41-43]为连续控制任务提供了一种低能耗的解决方案,已经被广泛应用在车辆、机器人等移动设备的控制任务中[29,44],受到了不少学者的关注.同时,借助神经形态传感器[45-46],脉冲强化学习算法能够充分利用多模态的脉冲序列数据,令智能体像人脑一样进行感知与决策,为仿生机器人的研究提供了一个可行的解决方案[44].更令人惊喜的是,脉冲强化学习算法能够有效解决强化学习中的鲁棒性问题[43,47],这是决定策略是否实用的关键因素.

此外,强化学习在诞生初期就与动物学习中心理学中的试错法以及神经科学中大脑的奖励学习机制密切相关,其中最显著的联系就是时序差分(Temporal Difference,TD)误差与多巴胺之间的相似关系,这被归纳为多巴胺的奖励预测误差假说[48].多巴胺的奖励预测误差假说认为,多巴胺的功能之一就是将未来期望奖励的新旧估计值之间的误差传递给大脑中的所有目标区域.这一假说利用强化学习中的TD误差概念,成功解释了哺乳动物中多巴胺神经元的相位活动特征.在计算神经科学领域里,大量的研究工作利用强化学习算法对大脑的奖励学习机制进行建模[49-51],这些都属于脉冲强化学习算法的研究范畴.

综上所述,脉冲强化学习算法不仅是脉冲神经网络与强化学习算法的有机结合,还是连通神经科学与AI两个领域的桥梁.根据时间顺序,脉冲强化学习算法的发展历程可以被分为三个时期:SNN与强化学习的基础研究时期、基于突触可塑性的脉冲

强化学习算法时期以及深度强化学习算法与SNN的结合时期，如图1所示．在基础研究时期，图1列举了SNN与强化学习各自的一些早期代表性工作，这些工作为后续脉冲强化学习算法的诞生与发展奠定了基础．以深度学习的兴起为时间节点，脉冲强化学习算法有着明显的不同．早期的算法注重突触可塑性与强化学习理论的结合，而后期的算法侧重于将SNN应用到深度强化学习算法中．由此，图1进一步地划分出了两个时期，分别列举了脉冲强化学习算法在早期与晚期的代表性工作．尽管近些年也出现了一些优秀的突触可塑性算法（例如e-prop[52]），但这已经不是计算机科学领域的主流，所以未被列入图1中．由于图的大小有限，部分SNN与强化学习的经典工作并未在图中展现，这将在本文的后续章节中进行更为系统的梳理．此外，出于美观考虑，图1根据事件线分为上下两侧，不存在事件类型的区别．

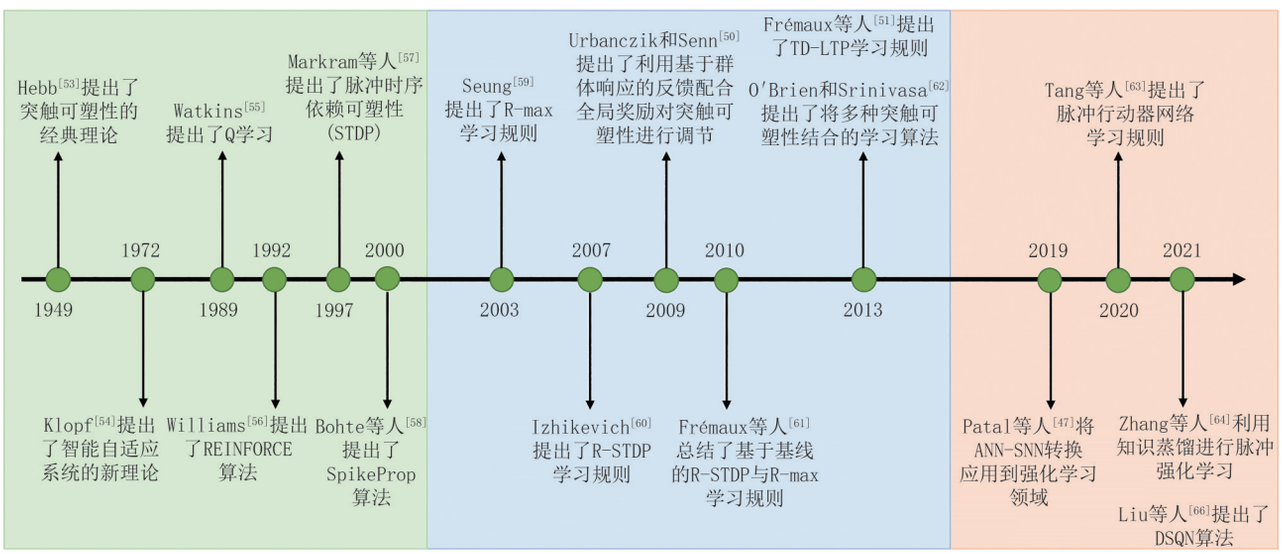

图1　脉冲强化学习算法的发展历程

关于相关工作在脉冲强化学习算法的发展历程中的地位以及作用，其概述如下：1949年Hebb[53]提出了突触可塑性的经典理论，对突触可塑性的基本原理进行了描述．1972年Klopf[54]提出了智能自适应系统的新理论，其中的一系列思想促成了资格迹（Eligibility Trace）在强化学习中的应用[48]．1989年Watkins[55]提出了Q学习，这是强化学习早期的一个重要突破，实现了异策（Off-policy）[48]下的时序差分控制．1992年Williams[56]提出了REINFORCE算法，这是一个经典的策略梯度算法，动作选择不再直接依赖于价值函数，而是可以直接学习参数化的策略．1997年Markram等人[57]提出了一个较为通用的SNN学习规则，即脉冲时序依赖可塑性（Spike-timing-dependent Plasticity，STDP），这是无监督学习的重要生物学基础．2000年Bohte等人[58]提出了SpikeProp算法，首次使用误差反向传播对SNN进行训练．2003年Seung[59]基于策略梯度算法提出了R-max学习规则．2007年Izhikevich[60]受到动物学实验的发现启发，提出了R-STDP学习规则．2009年Urbanczik和Senn[50]提出了利用基于群体响应的反馈配合全局奖励对突触可塑性进行调节，丰富了三因素学习规则中全局信号的选择范围．2010年Frémaux等人[61]总结了基于基线的R-STDP与R-max学习规则，利用基线函数对原本的学习规则进行改进，使其能够同时学习多个任务．2013年Frémaux等人[51]提出了TD-LTP学习规则，成功解决了如何在非离散框架下实现强化学习以及如何在神经元中计算奖励预测误差的问题．O′Brien和Srinivasa[62]提出了将多种突触可塑性结合的学习算法以解决同时学习多个远端奖励的问题．Patal等人[47]首次将ANN-SNN转换应用到强化学习领域，避免了强化学习直接训练SNN的困难，并证明了SNN能够提高模型对于遮挡的鲁棒性．Tang等人[63]提出了一个混合的行动器-评判器（Actor-Critic）网络，对脉冲行动器网络与深度评判器网络进行联合训练，证明了脉冲行动器网络可以作为原本深度行动器网络的一个低能耗替代方案．Zhang等人[64]受到知识蒸馏[65]启发，提出了一种间接训练SNN的方法，利用强化学习训练得到的ANN教师网络指导SNN学生网络的学习．Liu等人[66]提出了

DSQN算法，摆脱了原本的深度脉冲强化学习算法在训练过程中对ANN的依赖.

由于脉冲强化学习算法的领域交叉性，脉冲强化学习算法的研究在脉冲神经网络与强化学习算法的文献综述中少有提及. 例如Taherkhani等人[67]简单提及了基于奖励的突触可塑性学习. Hu等人[68]介绍了三因素学习规则. Sutton和Barto[48]阐述了神经科学中大脑奖励系统与强化学习理论之间的对应关系，并对神经科学与强化学习如何相互影响进行了讨论. 此外，之前关于脉冲强化学习算法的综述由于时间较早，其关注的都是与突触可塑性相关的内容. 例如Frémaux和Gerstner[69]系统总结了利用三因素学习规则实现的强化学习算法. Bing等人[37]介绍了利用三因素学习规则实现的强化学习算法及其相应的机器人应用.

本文首先介绍了SNN与强化学习的基本原理，然后以SNN学习算法的分类为基础，从更长的时间线对传统的利用三因素学习规则实现的强化学习算法与近些年来出现的新型脉冲强化学习算法进行了系统性回顾与综述. 不同于已有的综述，本文率先对依托ANN实现的脉冲强化学习算法与基于脉冲的直接强化学习算法进行了系统梳理与全面总结，并介绍了最新的研究挑战与未来研究方向. 最后，本文对脉冲强化学习算法的优点与不足进行了总结，并展望了其对未来人工智能和神经科学领域的潜在影响，希望通过跨学科的交流与合作，推动该领域的快速发展.

2 脉冲神经网络

脉冲神经网络作为第三代神经网络，使用生物学合理的脉冲神经元模型和相应的学习算法对大脑中的神经元结构以及突触可塑性进行建模. 本章将按照脉冲神经元模型、脉冲编码方式以及学习算法这三大基本要素对脉冲神经网络进行简要地介绍[67-68].

2.1 脉冲神经元模型

在生物体中，神经元的典型结构主要包括树突、胞体以及轴突三个部分. 树突收集来自其他神经元的输入信号并将其传递给胞体；胞体作为中央处理器，当接受的传入电流积累导致神经元膜电位超过一定阈值时就会产生脉冲(即动作电位)；轴突传播脉冲并通过末端的突触结构将信号传递给下一个神经元.

1952年，Hodgkin和Huxley对鱿鱼巨型轴突的电位数据进行了研究，提出了一个详细的基于电导的神经元模型(H-H模型)[70]，成功重现了电生理测量结果，由此获得了1963年的诺贝尔生理学奖. 然而，H-H模型极高的计算复杂度限制了其在实际任务中的使用. 因此，后续的低阶神经元模型通常对H-H模型进行简化，以降低计算成本. 目前常见的几种一阶神经元模型包括IF模型[71]、LIF模型及其变体[21,72-75]以及脉冲响应模型(Spike Response Model，SRM)[20]. 这些神经元模型计算成本较低，但与H-H模型相比，它们的生物合理性较差. Izhikevich比较了各种脉冲神经元模型的生物合理性和计算成本[76]，并提出了一个二阶神经元模型(Izhikevich模型)[77]，在生物合理性与计算成本之间提供了一个很好的折中.

在计算机仿真过程中，脉冲神经元模型中各个变量的连续微分方程需要在时间上先进行离散化处理，然后按照离散的时间步骤对相应的变量进行迭代更新. 对于离散时间步骤，当选择较小的时间间隔时，模型的变化会更加接近原本的连续微分方程，精度的损失较小. 但是，这也提高了相同条件下离散时间步骤的数目，从而增大了计算复杂度. 因此，离散脉冲神经元模型存在一个精度与计算复杂度的权衡. 通常，离散脉冲神经元模型可以被描述为充电、放电以及重置这三个离散方程(如图2所示)：

$$H_t = f(V_{t-1}, X_t) \tag{1}$$

$$S_t = \Theta(H_t - V_{th}) \tag{2}$$

$$V_t = H_t(1 - S_t) + V_{reset}S_t \tag{3}$$

其中H_t，V_t分别表示在t时刻充电后的膜电位以及重置后的膜电位. $f(\cdot)$定义了具体的神经元充电方程，这在不同的脉冲神经元模型之间存在较大差异. $\Theta(\cdot)$为阶跃函数，一般被定义为

$$\Theta(x) = \begin{cases} 1, & if\ x \geqslant 0 \\ 0, & otherwise \end{cases} \tag{4}$$

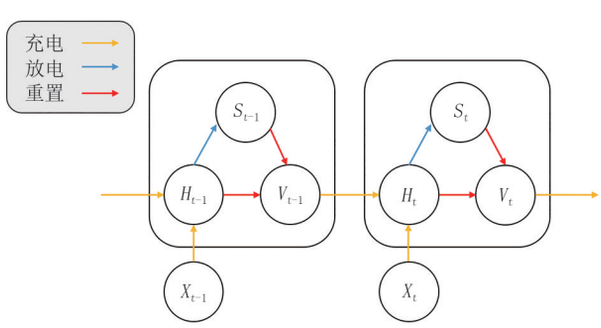

图2 通用的离散脉冲神经元模型[21]

充电过程中，神经元接收外源输入X_t，并对膜电位进行更新；放电过程中，当膜电位超过阈值V_{th}，神经元输出脉冲信号S_t；重置过程中，神经元根据脉冲信号S_t对膜电位进行更新. 若是采用软重置，则需要将公式(2)改为$V_t=H_t-V_{th}S_t$. 需要强调的是，这种方法描述的对象主要是深度SNN中常用的脉冲神经元类型，并不能涵盖所有的脉冲神经元动态.

2.2　脉冲编码方式

生物体对周围世界的感知、对内稳态的调节以及对感官刺激的行为反应，这些都由神经元脉冲序列实现信息传递. 神经元利用何种编码方式将信息存储到脉冲序列之中，这是神经科学领域中讨论的重要问题之一.

根据从大脑中得到的生理学发现，研究界提出了许多不同的脉冲编码方式. 目前常见的脉冲编码方式主要包括频率编码(Rate Coding)、时间编码(Temporal Coding)、爆发型编码(Bursting Coding)和群体编码(Population Coding)等.

频率编码主要考察神经元的脉冲发放率，即神经元发放的脉冲发放次数在仿真时长中的均值. 神经元的脉冲发放率可以反映出输入刺激的强弱程度[78]. 相比频率编码，时间编码更关注脉冲序列在时间结构上的差异，除了完整的脉冲发放时间信息之外，从接受刺激到发放首个脉冲的时间以及脉冲之间的时序逻辑都被认为可以编码脉冲序列的重要信息，其中前者被称为首次脉冲时间编码(Time-to-first-spike Coding)[79]，后者被称为排序编码(Rank Order Coding)[80]. 除了这两种常用的时间编码方案外，时间编码方案还包括延迟编码(Latency Coding)[81]、相位编码(Phase Coding)、同步编码(Synchrony Coding)[82]和多时化编码(Polychronisation Coding)[83]等. 爆发型编码考虑的是利用神经元爆发型放电这一行为对信息进行编码[84]. 在爆发型放电时，神经元在一段时间内快速地发放大量脉冲，随后在较长时间内保持静息. 与更多关注单个神经元活动的频率编码、时间编码等不同，群体编码认为刺激产生的信息由多个神经元的联合活动进行表征[85].

在脉冲神经网络的具体实现过程中，需要灵活地针对不同任务对脉冲编码方式进行选择. 不同的脉冲编码方式在任务性能、时延以及能耗等方面各有优劣，并且会对脉冲神经网络学习算法的选择产生巨大的影响.

2.3　学习算法

目前，脉冲神经网络领域并不存在统一的学习算法. 出于对生物合理性与任务性能的不同侧重，以及网络中脉冲神经元模型以及脉冲编码方式的不同，脉冲神经网络的学习算法存在较大的差异. 在脉冲神经网络发展初期，学习算法更注重生物合理性，主要可以分为三大类：无监督学习算法、三因素学习规则以及基于时序的监督学习算法. 随着深度学习的兴起，出于处理复杂任务的需求以及性能的考虑，如何将脉冲神经网络的层数做深成为了一大难题. 为了解决网络深度的问题，SNN研究界涌现出了两类新的学习算法：ANN-SNN转换以及基于激活值的监督学习算法. 通过将前后两个阶段的SNN学习算法进行对比，可以发现一个明显的差异，即前期的SNN学习算法主要选择的脉冲编码方式是时间编码，而后期的SNN学习算法则主要是频率编码. 发生如此改变的一个主要原因是频率编码能够较好地与现有的深度学习框架相兼容. 这不仅便于学习算法的实现，还大大提升了学习算法的计算效率.

2.3.1　无监督学习算法

脉冲神经网络的无监督学习算法利用突触可塑性规则进行实现. 对于突触可塑性，加拿大著名生理心理学家Donald Olding Hebb于1949年提出了一个经典的理论，其主要内容为“一起发放的神经元连在一起”(“Neurons fire together，wire together”)，即如果相连的两个神经元同时被激活，那么就增加这两个神经元之间的权重[53]. 该理论描述了突触可塑性的基本原理，也成为了无监督学习算法的生物学基础.

STDP[57]可以被视为Hebb规则的变种，根据突触前和突触后脉冲的相对时间，对突触权重变化进行了描述. 对于常见的STDP规则，其描述公式如下所示：

$$\Delta w_{ij}=\begin{cases}A_+e^{-\frac{\Delta t}{\tau_+}}, & if\ \Delta t\geqslant 0\\ -A_-e^{\frac{\Delta t}{\tau_-}}, & otherwise\end{cases}\tag{5}$$

其中，Δw_{ij}表示权重改变量，突触前和突触后脉冲之间的时间间隔为$\Delta t=t_i-t_j$，t_j与t_i分别表示突触前和突触后神经元的脉冲发放时间. A_+与A_-表示权重的最大改变幅度，而τ_+与τ_-表示影响STDP时间窗口的常数.

这些常数的取值与海马体中发现的两种可塑性现象密切相关[57]. 这两种现象分别被称为长时程增

强(Long-Term Potentiation，LTP)以及长时程抑制(Long-Term Depression，LTD).

2.3.2 三因素学习规则

为了弥合行为时间尺度和神经元动作电位之间的差距，现代突触可塑性理论假设突触前和突触后神经元的共同激活在突触处设置了一个标志，称为资格迹. 只有在标志设置时存在第三个因素，即奖励、惩罚、惊喜或新奇信号，突触权重才会改变. 因此，该突触可塑性理论被称为三因素学习规则. 具体而言，三因素学习规则对应的突触权重更新公式被定义如下：$\Delta w_{ij}=e_{ij}M_{3rd}(t)$，其中，$e_{ij}$表示突触的资格迹. $M_{3rd}(t)$表示第三个因素，这可能是由注意过程、意外事件或奖励触发的. 多巴胺、5-羟色胺、乙酰胆碱或去甲肾上腺素等神经递质的相位信号是第三个因素的首要候选[86-87]. 其中，相位信号被定义为当前值与运行平均值的偏差，以便$M_{3rd}(t)$可以选取正值与负值. 尽管该理论框架[69,88-90]是在过去几十年中发展起来的，但实验人员在最近几年中才收集到支持秒级时间尺度上资格迹的实验证据[91].

与无监督学习算法不同，三因素学习规则通过引入第三个因素作为监督信号，在生物体行为和局部突触可塑性之间建立了桥梁，为突触权重的变化指定了目标与方向，大大提高了学习效率.

2.3.3 基于时序的监督学习算法

在基于时序的监督学习算法中，标签被编码为具有时间结构的目标脉冲序列. 此类学习算法根据SNN输出的脉冲序列与目标脉冲序列之间的差异，对网络中的权重进行学习.

Bohte等人[58]提出了SpikeProp算法，这是首个使用误差反向传播进行训练的SNN学习算法. 通过利用SRM对脉冲发放时间与膜电位之间的函数关系进行解析，SpikeProp成功实现了对脉冲发放时间的求导，避开了脉冲发放函数在阈值处不连续导致的不可微难题. ReSuMe算法[92]采用精确时间编码，结合STDP和反向STDP过程，摆脱了对脉冲神经元模型的依赖. 该算法基于Widrow-Hoff规则[93]，在不需要显式梯度计算的情况下，最小化输出与目标脉冲序列之间的差异，其突触权重的更新公式定义如下：$\frac{\mathrm{d}w_{oi}}{\mathrm{d}t}=\left[S_d(t)-S_o(t)\right]\left[a_d+\int_0^{\infty}a_{di}(s)S_i(t-s)\mathrm{d}s\right]$，其中，$S_i(t)$，$S_o(t)$，$S_d(t)$分别表示输入(突触前)、输出(突触后)以及目标脉冲序列. 每个脉冲序列都是时间尺度上单个脉冲的叠加$S(t)=\sum_f\delta(t_f)$，其中$\delta(\cdot)$是狄拉克函数，t_f表示第f个脉冲的发放时间. a_d是用于对包含$a_{di}(s)$的Hebb项进行调节的非Hebb项. $a_{di}(s)$是类似STDP的学习窗口，其公式为

$$a_{di}(s)=\begin{cases}Ae^{-\frac{s}{\tau}}, & if\ s\geqslant 0\\ 0, & otherwise\end{cases}\tag{6}$$

其中A与τ的定义与STDP一致，都是用于描述学习窗口的常数. SPAN算法[94]与ReSuMe算法相似，它同样结合了STDP和反向STDP过程，并且源自Widrow-Hoff规则. 此算法的创新之处在于其将脉冲序列转换为模拟信号，以便于执行常见的数学运算.

2.3.4 ANN-SNN转换

近些年来，ANN的训练技术日趋成熟. 与ANN相比，目前训练SNN不仅需要大量的计算资源，还要克服脉冲发放函数不可微的难题. 为了避免训练SNN的困难，同时发挥SNN处理数据的优势，ANN-SNN转换成为了一个可行的方法. ANN-SNN转换通过使用每个脉冲神经元的发放率来近似模拟人工神经元中相应的ReLU激活值，将训练好的ANN转换为SNN[95-97]. 值得注意的是，此处脉冲神经元的发放率指的是在离散的仿真周期内脉冲发放总数与仿真时长的商. 研究证明软重置(Soft Reset)设置下的IF神经元是ReLU激活函数在时间上的无偏估计[97].

为了快速训练和收敛，研究人员在ANN中引入了批归一化(Batch Normalization). 由于经过批归一化，ANN的层输出的均值为0，这不符合转换后SNN的需求. 因此，需要对批归一化层进行改进，将其参数并入前面的参数层(卷积层、全连接层)[97]. 此外，根据转换理论中脉冲神经元发放率的定义，ANN的每层输入输出需要被限制在[0,1]范围内，这就需要对参数进行缩放(模型归一化)，其中包括阈值平衡以及权重归一化[97]. 利用这些技术手段，ANN-SNN转换可以保证较低的模型转换损失. 尽管如此，ANN-SNN转换在推理阶段会存在一个精度与延迟之间的权衡. 为了使得转换后的SNN能够达到与ANN相比接近无损的推理精度，在转换过程中需要将SNN的仿真时长设置为一个较大的数值(一般为数十或者上百). 因此，ANN-SNN转换的推理延迟与直接训练的SNN学习算法相比存在较大的差距.

2.3.5　基于激活值的监督学习算法

为了利用标准的基于反向传播的优化方法，基于激活值的监督学习算法采用替代梯度法[98-99]进行训练，即SNN使用阈值函数$\Theta(x)$进行前向传播，但使用替代函数$\sigma(x)$的导数进行反向传播，即$\Theta'(x)=\sigma'(x)$. 常用的替代函数包括Sigmoid函数、反正切函数等.

借助这种方式，SNN中的脉冲信号可以与ANN中的激活值一样进行前向与反向传播. 具体而言，在一定的仿真时长内，基于激活值的学习算法将脉冲序列数据按离散时间步骤输入到SNN中得到输出脉冲序列. 而当仿真时长结束时，算法对输出脉冲序列进行处理，并将其与目标值进行计算以得到连续的损失值. 对于图像分类任务，首先将输出层中每个神经元的脉冲发放次数除以整个仿真时长得到脉冲发放率，然后利用脉冲发放率与标签信号计算交叉熵损失或均方误差损失，实现完整的前向传播. 最后，SNN利用损失值进行反向传播，对网络参数进行更新. 总体而言，基于激活值的监督学习算法[21-22,100-101]成功实现了深度SNN的训练，在图像分类任务中表现出较强的竞争力，成为了直接训练的SNN学习算法中的主流.

3　强化学习

强化学习解决的是能够与环境交互的智能体序列决策问题，其过程被描述为如下循环：

（1）智能体从环境中接收状态s；

（2）基于状态s，智能体采取动作a；

（3）环境转换到一个新的状态s'；

（4）环境给予智能体奖励r.

这个循环可以被描述为马尔可夫决策过程（Markov Decision Process，MDP）[48]. MDP通常被定义为一个元组$M=(\mathcal{S},\mathcal{A},\mathcal{P},\mathcal{R},\gamma)$. 在这个元组中，$\mathcal{S}$表示状态空间，$\mathcal{A}$表示动作空间，$\mathcal{P}$表示状态转移矩阵，$\mathcal{R}$表示奖励函数. 智能体在与环境交互的过程中会输出一个状态、动作与奖励的序列$(S_0,A_0,R_1,S_1,A_1,R_2,\cdots)$，如图3所示. 基于奖励假设（Reward Hypothesis），强化学习的所有目标都可以被描述为期望累积奖励的最大化. 另一方面，为了避免持续性任务导致累积奖励趋于无穷，折扣率γ被引入，使得强化学习的目标变为最大化期望折后累积奖励，这也可以被称为期望未来奖励. t时刻后的期望未来奖励G_t被定义为如下形式：$G_t=R_{t+1}+\gamma R_{t+2}+\gamma^2 R_{t+3}+\cdots$，其中$0\leqslant\gamma\leqslant1$，用于表示智能体在短期奖励与长期奖励之间的偏好.

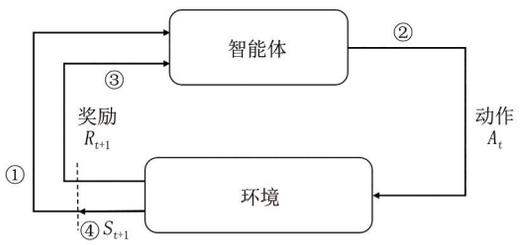

图3　马尔可夫决策过程中智能体与环境之间的交互

为了获取最大的长期奖励，最优策略或许会牺牲一些短期奖励. 这个问题被总结为探索与开发之间的权衡（Exploration/Exploitation Trade-off）[48]，是强化学习的一个重要主题. 探索是为了寻找有关环境的更多信息，开发是为了利用已知信息来最大化奖励. 具体而言，探索越多，获取的环境信息就越全面，便于后续做出更优的决策，但也可能会使得当前累积获得的奖励较小. 因此，智能体往往会陷入探索与开发的两难困境，这就需要算法设计者为策略制定良好的探索方案，在两者之间进行权衡. 对此，ϵ-贪心算法[48]是一种常用的策略. 采取该算法的智能体在进行决策时，会以概率ϵ非贪心地在动作空间内随机选择一个动作，剩下$1-\epsilon$的概率根据贪心策略选择当前最优的动作. 其中，$0\leqslant\epsilon\leqslant1$，通常初始设定或者随时间推移衰减为一个很小的固定正值.

除了奖励与环境这两个要素外，强化学习中还存在着影响智能体的两个关键要素，即策略和价值函数. 策略是从状态到每个动作的选择概率之间的映射，可以分为随机性策略与确定性策略. 前者输出的是不同动作的概率，而后者输出的是一个确定的动作. 随机性策略本身自带探索，可以通过探索产生各种数据，然后对当前的策略进行改进. 而在给定状态和策略参数时，确定性策略的动作是固定的，所以无法探索其他轨迹或者访问其他状态. 因此确定性策略无法探索环境，所以需要通过异策[48]方法进行学习，即策略评估时采用的策略与数据采集时的不同. 与随机性策略相比，确定性策略的优点在于需要采样的数据少，算法效率高. 但是，确定性策略可能会出现感知混叠（Perceptual Aliasing）[102]的问题. 这一问题通常发生在部分可观察马尔可夫决策过程（Partially Observable Markov Decision

Process，POMDP)中. 对于由POMDP建模的任务而言，环境动态由MDP决定，但是智能体无法直接得到状态，而只能得到基于状态的部分观察值. 当智能体在面对不同状态时，可能得到两个类似或者相同的观察值，即感知混叠，这通常需要采取不同的动作. 当面对这种问题时，采用确定性策略会使得智能体陷入循环，而采用随机性策略往往能够取得较好的结果.

为了学到最优策略，通常需要从长远的角度对策略进行评估. 因此，为了最大化长期奖励，强化学习中引入价值函数来表示智能体的期望未来奖励. 根据不同的定义，价值函数可以分为状态价值函数以及动作价值函数. 状态价值函数$V_\pi(s)$表示智能体在状态s时采取策略π将会得到的最大期望未来奖励，因此其定义公式可以表示为：$V_\pi(s)=\mathbb{E}_\pi\left[G_t|S_t=s\right], \forall s\in\mathcal{S}$. 动作价值函数$Q_\pi(s,a)$(Q函数)表示智能体在状态$s$时执行动作$a$后采取策略$\pi$将会得到的最大期望未来奖励，其被定义为：$Q_\pi(s,a)=\mathbb{E}_\pi\left[G_t|S_t=s,A_t=a\right]$. 状态价值函数与动作价值函数之间可以用以下两个公式互相表示：$V_\pi(s)=\sum_{a\in\mathcal{A}}\pi(a|s)Q_\pi(s,a)$以及$Q_\pi(s,a)=r(s,a)+\gamma\sum_{s'}p(s'|s,a)$.

由于智能体的期望未来奖励取决于智能体所选择的动作. 因此，价值函数与特定的策略相关联. 通过对价值函数的比较，可以对策略的优劣进行精确判断，即价值函数定义了策略上的一个偏序关系. 具体而言，若$\pi\geqslant\pi'$，那么应当对于$\forall s\in\mathcal{S}, V_\pi(s)\geqslant V_{\pi'}(s)$. 借助这个偏序关系，至少可以找到一个策略不劣于其他所有的策略，这就是最优策略π_*，其定义为：$\pi_*=\arg max_\pi V_\pi(s), \forall s\in\mathcal{S}$. 与最优策略对应的最优状态价值函数$V_*(s)$为：$V_*(s)=max_\pi V_\pi(s), \forall s\in\mathcal{S}$. 同理，最优动作价值函数$Q_*(s,a)$为：$Q_*(s,a)=\max_\pi Q_\pi(s,a), \forall s\in\mathcal{S}, a\in\mathcal{A}$. 最优状态价值函数和最优动作价值函数同样能够互相表示，并且由于其最优性，可以表示为不局限于任何特定策略的形式，即贝尔曼最优方程(Bellman Optimal Equation)[48]. 其中，贝尔曼最优方程具有以下两种形式：$V_*(s)=max_{a\in\mathcal{A}}\left[r(s,a)+\gamma\sum_{s'}p(s'|s,a)V_*(s')\right], \forall s\in\mathcal{S}$与$Q_*(s,a)=r(s,a)+\gamma\sum_{s'}p(s'|s,a)max_{a'}Q_*(s',a'), \forall s\in\mathcal{S}, a\in\mathcal{A}$.

最优价值函数可以根据贝尔曼最优方程进行求解. 但是由于利用贝尔曼最优方程求解的计算复杂度较高，受限于计算资源，在实际使用时通常会考虑需要使用迭代的数值方法来对其进行求解. 常用的迭代方法包括动态规划(Dynamic Programming，DP)算法、蒙特卡洛(Monte-Carlo，MC)方法以及TD学习[48]. 对于在给定环境模型的情况下求解MDP的最优策略，动态规划是一类很好的优化算法. 通过策略评估与策略改进这两个计算步骤，可以将最常用的动态规划算法分为两种：策略迭代与价值迭代. 然而，若是事先没有任何环境模型的信息，就需要借助动态规划以外的算法，即蒙特卡洛方法与TD学习. 前者在回合结束时收集奖励，然后计算最大期望未来奖励；而后者采用自举(Bootstrapping)[48]的思想，对每一步的奖励进行估计. 自举指的是根据对后继状态价值的估计来更新对当前状态价值的估计，这在动态规划算法中也有所体现.

以上介绍的三种迭代算法，在每次更新价值函数时都只能更新某个确定状态(或状态-动作对)下的价值估计，因此仅能用于解决一些相对简单问题. 比如，这些问题的状态空间与动作空间较小或者能够被离散化，可以用表格的形式表示价值函数并通过上述算法迭代求得精确的最优解. 但是对于实际的决策控制任务，其状态空间和动作空间往往非常大，甚至是无法遍历或者离散化的连续空间，这就使得原本的算法难以利用表格对价值函数进行表示并逐一对其进行更新. 因此，强化学习算法的目标从找到某些确定状态下的最优策略或最优价值函数的精确解转为使用有限的计算资源找到一个可以适用于所有状态的近似函数. 函数近似方法[48]用参数化的模型来近似价值函数，并在每次学习时通过改变参数对整个函数进行更新，成功地解决了此问题. 随着近些年深度学习的发展，绝大多数强化学习方法将深度神经网络作为策略函数、价值函数的近似函数，因此这些方法也被称为深度强化学习.

此外，按照学习过程中是否用到了环境的数学模型，强化学习可以被分为有模型(Model-based)学习算法和无模型(Model-free)学习算法[48]. 对于有模型学习算法，环境的模型可以是学习前已经明确的，也可以是通过学习得到的. 反之，无模型学习算法不需要任何有关环境的信息，也不需要对环境模型进行学习，训练所需的所有经验都是通过与真实环境交互得到的.

与有模型学习算法相比，无模型学习算法易于调整与实现，更受研究者的欢迎，并且在大量环境与任务中得到了更广泛的开发与测试. 因此，本章将

主要介绍无模型学习算法. 通常，根据智能体训练算法的差异，可以将无模型学习算法分为三大类：基于价值(Value-based)的方法、基于策略(Policy-based)的方法以及行动器-评判器方法.

3.1　基于价值的方法

基于价值的方法对价值函数进行参数化，借助价值函数对当前状态下的动作进行选择. 由于价值函数学到的是当前状态下策略的期望未来奖励，此类方法也被称为价值估计算法. 在价值估计算法中，最经典的算法即为Q学习[55]. 传统的Q学习算法会将状态和动作构建成一张Q表. Q表中的每一列对应于不同的动作，每一行对应于不同的状态，而每个表格的值对应于给定状态与动作下的最大期望未来奖励. Q学习使用时序差分方法进行异策学习，并利用贝尔曼方程对MDP求解得到最优策略，其定义为 $Q(S_t,A_t)\leftarrow Q(S_t,A_t)+\alpha\big(R_{t+1}+\gamma\max_a Q(S_{t+1},a)-Q(S_t,A_t)\big)$. 值得注意的是，Q学习是异策的，这意味着每次更新Q表都可以使用在训练期间任意时间点收集的数据，而无需顾及获取数据时智能体采取的策略.

当环境的状态空间较大时，生成和更新Q表将会消耗大量存储与计算资源. 因此，使用函数近似算法逐渐成为主流. DQN算法[10]将深度学习与强化学习相结合，用一个参数为θ的人工神经网络来拟合动作价值函数$Q(s,a)$. 该算法标志着深度强化学习算法的诞生，并对强化学习研究产生了深远的影响. 考虑到基于价值的方法在同时满足异策、自举以及函数近似这三个基本要素时，就一定会有不稳定和发散的危险[48]. DQN算法提出了经验回放(Experience Replay)以及目标网络(Target Network)对其进行改进. 其后，研究人员对DQN算法进行了一些独立的改进，这些方法包括双重Q学习(Double Q-learning)[103]、优先回放(Prioritized Replay)[104]、对偶网络(Dueling Networks)[105]、多步学习(Multi-step Learning)[106]、分布强化学习(Distributional RL)[107]、带噪网络(Noisy Networks)[108]等. Rainbow[109]对DQN算法的扩展进行了详细的研究，并通过经验研究了各种拓展的组合，在Atari 2600基准测试[110]上其数据效率和最终性能均取得了领先的水平.

3.2　基于策略的方法

基于策略的方法又称策略梯度算法，它不需要学习动作价值函数，而是将策略参数化为$\pi_\theta(a|s)$，并通过在性能目标$J(\pi_\theta)$或其局部近似上执行梯度上升算法对策略参数θ进行优化. 该策略梯度公式为 $\nabla_\theta J(\theta)=E_{\pi_\theta}\big[\nabla_\theta log\pi_\theta(s,a)Q_{\pi_\theta}(s,a)\big]$. 通过蒙特卡洛方法，可以将策略梯度公式简化为 $\theta_{t+1}=\theta_t+\alpha G_t\nabla_\theta ln\pi_{\theta_t}(A_t|S_t)$. 这是策略梯度的经典算法，被称为REINFORCE[56]. 通过引入基线函数，可以降低策略梯度算法学习过程中的方差.

3.3　行动器-评判器方法

行动器-评判器方法是基于价值的方法与基于策略的方法的混合体. 通常，行动器-评判器方法包含两个人工神经网络：一个参数为w的评判器，用于衡量所采取动作的期望未来奖励；一个参数为θ的行动器，控制智能体的行为. 在学习初期，行动器执行随机动作，评判器对行动器所执行的动作进行价值估计并提供反馈. 通过学习该反馈，行动器会对策略进行改进. 与此同时，评判器也将对价值函数进行更新，以提供更好的反馈. 两者并行运行，不断对两个网络的参数进行更新. 尽管有些文献[48]将行动器-评判器方法归入策略梯度算法之中，但由于其结合了价值估计算法，在此仍将其单独列为两者之外的第三种方法.

值得重视的是，尽管带基线函数的策略梯度算法同样既学习了一个策略函数，也学习了一个状态价值函数，但通常并不认为这是一种行动器-评判器方法. 因为算法中的状态价值函数仅被用作基线，而不是一个“评判器”，即其没有被用于自举. 只有采用自举法，才能保证出现依赖于函数逼近质量的偏差以及渐进性收敛. 在降低方差的同时加快学习速度[48]. 当在基本的动作价值行动器-评论器方法中将状态价值函数作为基线时，即可得到优势函数估计的单步时序差分形式 $R_{t+1}+\gamma V_w(S_{t+1})-V_w(S_t)$，由此可推导出经典的优势行动器-评判器(Advantage Actor-Critic，A2C)算法以及异步优势行动器-评判器(Asynchronous Advantage Actor-Critic，A3C)算法[111]. 此外，行动器-评判器方法还包括信任域策略优化(Trust Region Policy Optimization，TRPO)算法[112]、近端策略优化(Proximal Policy Optimization，PPO)算法[113]、柔性行动器-评判器(Soft Actor-Critic，SAC)算法[114-115]、深度确定性策略梯度(Deep Deterministic Policy Gradient，DDPG)算法[116]、双延迟深度确定性(Twin Delayed Deep Deterministic，TD3)策略梯度算法[117]等. 值得注意的是，除了DDPG与TD3算法采用确定性策略外，本节提到的其他算法采用的都是随机性策略. 除了

A3C、SAC、DDPG、TD3算法采用异策学习外，本节提到的其他算法采用的都是同策(On-policy)学习[48]，这意味着这些算法在每次更新过程中只能使用根据当前策略进行操作时收集的数据，因而样本效率较为低下.

4 脉冲强化学习算法研究进展与现状

脉冲强化学习算法在发展初期汲取了大量神经科学领域的发现，通过对神经递质与突触可塑性的相关性进行建模，推动建立起了以三因素学习规则为基础的现代突触可塑性理论. 近十年，随着以卷积神经网络(Convolutional Neural Network，CNN)为代表的深度学习热度不断飙升，脉冲强化学习的发展与脉冲神经网络一样，在网络结构、训练算法以及任务领域上与之前利用三因素学习规则实现的强化学习算法产生了较大的差异. 具体而言，后期的脉冲强化学习算法主要可以分为两大类：(1)依托ANN实现的脉冲强化学习算法：在训练过程中使用ANN用于辅助SNN进行强化学习，以避免直接训练SNN模型会遇到的问题.(2)基于脉冲的直接强化学习算法：在SNN的前向传播过程中，利用不同的脉冲编码方案将脉冲序列转换为强化学习所需的值，并且在反向传播过程中，借助替代梯度法对深度强化学习中的模型参数进行更新.

总体而言，脉冲强化学习算法的基本架构如图4所示. 尽管在训练阶段，脉冲强化学习算法可能会借助深度网络进行实现. 但是在实际的测试与部署时，脉冲强化学习算法的主体采用的都是脉冲神经网络.

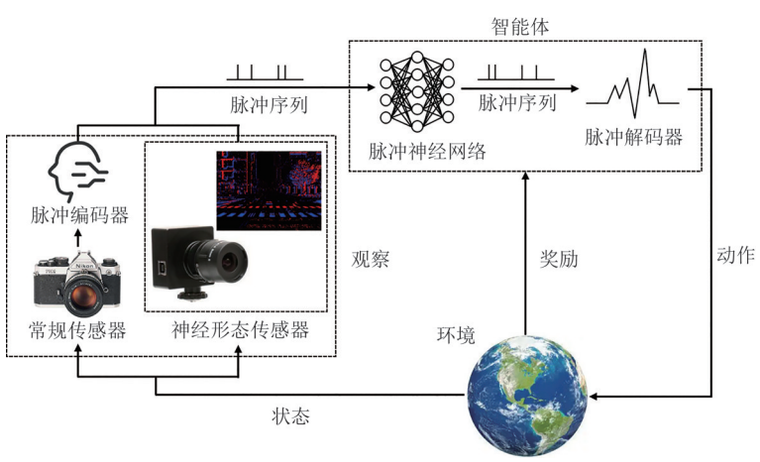

图4 脉冲强化学习算法的基本架构图

4.1 利用三因素学习规则实现的强化学习算法

利用三因素学习规则实现的强化学习算法利用全局的奖励相关信号以及局部的突触资格迹对突触权重进行更新，在脉冲神经网络中对强化学习算法进行实现[69].

4.1.1 策略梯度算法：R-max学习规则

对于奖励驱动学习的问题，对应的突触可塑性规则可以从奖励最大化原则的迭代中推导而出. 这种从奖励最大化导出的规则被称为R-max学习规则，该规则源于将策略梯度方法应用于随机脉冲神经元模型，旨在增加学习过程中的奖励.

Seung[59]假设大脑使用突触传递的随机性进行学习，设计了一个享乐主义突触网络[118]来实现强化学习，其中突触根据全局奖励之前的动作对囊泡释放或失败的概率进行调整，以完成对奖励的响应. 他将REINFORCE算法[56]在享乐主义突触模型上进行了实现，通过IF神经元模型对其进行数值仿真. 享乐主义突触模型被建模为两个状态，分别为可用状态和不应状态. 当突触可用时，突触前脉冲刺激囊泡释放的概率p被定义为关于释放参数q的函数：$p=\frac{1}{e^{-q-c}}$，其中c表示类似钙的变量. 当囊泡释放后，突触从可用状态变为不应状态，然后根据时间常数$\frac{1}{\tau_r}$进行恢复，即从不应状态变为可用状态. 对于从神经元j到神经元i的突触，其资格(Eligibility)e_{ij}被定义为如下形式：$e_{ij}(x_t,x_{t-1})=\frac{\partial}{\partial q_{ij}}logP_\theta(x_t|x_{t-1})$，其中状态矢量$x$表示网络的所有动态变量，$\theta$表示网络的参数，$t$表示数值仿真中的离散时间步骤，那么整个动态可以被描述为马尔可夫转移矩阵$P_\theta(x_t|x_{t-1})$. 当神经元j不发放脉冲的情况

下，$P_\theta(x_t|x_{t-1})$与q_{ij}之间相互独立，会导致资格消失. 此外，当突触处于不应状态时，资格也会消失. 只有当突触处于可用状态且神经元j发放脉冲时，下式成立：

$$logP_\theta(x_t|x_{t-1})=\begin{cases} logp_{ij}+IR, & if\ \ s=1 \\ log(1-p_{ij})+IR, & if\ \ s=0 \end{cases} \tag{7}$$

其中，s表示囊泡的释放行为，1为成功，反之则为失败.

由于IR项仅与除p_{ij}以外的其他突触释放概率相关，在对q_{ij}求导时可以忽略不计，以上的资格公式可以简化成如下形式：

$$e_{ij}(x_t,x_{t-1})=\begin{cases} 1-p_{ij}, & if\ \ s=1 \\ -p_{ij}, & if\ \ s=0 \end{cases} \tag{8}$$

而资格迹$\overline{e_t}$的定义为$\overline{e_t}=\beta\overline{e_{t-1}}+e(x_t,x_{t-1})$，其中$\beta$表示资格迹的衰减率. 由此推导出的REINFORCE学习规则可以表示为$\Delta\theta_t=\eta R(x_t)\overline{e_t}$，其中$\eta$表示学习率，$R(x_t)$表示$t$时刻下对应状态收到的奖励.

与突触相比，脉冲神经网络研究的重点更偏向于对神经元进行建模[67]. Xie和Seung[90]将不规则脉冲的波动与奖励信号相关联，学习规则在期望奖励上执行随机梯度上升. 在模型网络中，突触后神经元i接收的总突触电流I_i由突触前神经元j的贡献之和按下式给出：$I_i(t)=\sum_{j=1}^{n}w_{ij}h_{ij}(t)$，其中$w_{ij}$和$h_{ij}(t)$分别表示从神经元$j$到神经元$i$的突触权重以及$t$时刻的突触激活值. 在神经元$j$发放的脉冲，神经元$i$中突触电流的行为时间过程被定义为如下形式：$\tau_s\frac{\mathrm{d}h_{ij}}{\mathrm{d}t}+h_{ij}(t)=\sum_a\delta(t-T_j^a)\zeta_{ij}^a$，其中$\tau_s$表示时间常数，$\delta(\cdot)$为狄拉克函数，$T_j^a$表示神经元$j$发放第$a$个脉冲的时间. ζ_{ij}^a是一个二值随机变量，用于模拟神经递质响应突触前的随机释放. $\zeta_{ij}^a=1$表示神经递质从神经元j释放到神经元i以响应神经元j的第a个脉冲，而$\zeta_{ij}^a=0$则表示未进行神经递质的释放. 对于回合形式的学习规则，在每一个回合结束时，突触权重就会按下式进行更新：$\Delta w_{ij}=\eta Re_{ij}$，其中$\eta$是学习率，$R$是奖励信号，$e_{ij}$是从神经元$j$到神经元$i$的突触的资格迹，$e_{ij}=\int_0^T\Phi_i(I_i)\left[s_i(t)-f_i(I_i)\right]h_{ij}(t)\mathrm{d}t$，其中$s_i(t)=\sum_a\delta(t-T_j^a)$是神经元$i$的脉冲序列，$T$是一个回合的长度. $f_i(\cdot)$是神经元i的电流-放电关系，用于根据电流大小计算神经元对应的瞬时发放率. 假设$f_i(\cdot)$单调递增，函数$\Phi_i(I_i)=f_i'(I_i)/f_i(I_i)$是一个正因子. 该算法高度依赖于神经元的泊松特性，并且由于在使用时需要在神经元中注入噪声，因此难以与常用的神经模型（例如IF神经元）结合使用.

Florian[119-120]利用全局奖励信号对局部突触可塑性进行调节，提出了一种生物学合理的脉冲强化学习算法. 这种强化机制在生物学上是高度合理的，并在随后的动物实验研究[91]中得到了有力的证明. 该算法由OLPOMDP强化学习算法[121-122]针对随机IF神经元网络推导得出，这是一种在参数化随机策略控制的通用部分可观察的马尔可夫过程中生成平均奖励梯度的有偏估计的在线算法. 在模型网络中，对于每个时间步骤t，神经元i发放脉冲的概率为$\sigma_i(t)$，对应$f_i(t)=1$. 若不发放脉冲，$f_i(t)=0$. 在Florian的第一篇工作[119]中，突触按照如下可塑性规则进行更新：

$$w_{ij}(t+\Delta t)=w_{ij}(t)+\gamma r(t+\Delta t)e_{ij}(t+\Delta t) \tag{9}$$

$$e_{ij}(t+\Delta t)=\beta e_{ij}(t)+\zeta_{ij}(t) \tag{10}$$

$$\zeta_{ij}(t)=\begin{cases} \dfrac{1}{\sigma_i(t)}\dfrac{\partial\sigma_i(t)}{\partial w_{ij}}, & if\ f_i(t)=1 \\ -\dfrac{1}{1-\sigma_i(t)}\dfrac{\partial\sigma_i(t)}{\partial w_{ij}}, & if\ f_i(t)=0 \end{cases} \tag{11}$$

其中w_{ij}是从神经元j到神经元i的突触权重，Δt是一个时间步骤的持续时间，学习率γ是一个小常数参数，e_{ij}是资格迹，ζ表示上一个时间步骤中的活动导致资格迹的变化量. 折扣因子$\gamma=exp(-\Delta t/\tau_e)$是一个0到1之间的参数值，其中$\tau_e$是影响资格迹指数衰减速度的时间常数. Florian在后续工作[120]中对其进行了连续时间的拓展，并将突触前后神经元发放阈值作为可学习的适应性参数进行了推导分析.

与此同时，Baras和Meir[123]将策略梯度算法应用于SNN，推导出了一种STDP学习规则. 他们通过统计分析表明其与广泛使用的BCM规则[124]密切相关，具有良好的生物学依据. Kaiser等人[125]提出了一种基于突触采样的奖励学习规则，通过在线强化学习来评估突触可塑性，用于两个视觉运动任务：目标到达和车道保持. 借助开源软件，这两个任务实现了动态视觉传感器（Dynamic Vision Sensor，DVS）[126]视觉感知下的机器人模拟，并且避免了实际机器人技术中必要的人工监督所需要花费的大量时间.

4.1.2　现象学模型算法：R-STDP学习规则

上一节中讨论的R-max学习规则可以严格地从策略梯度算法中推导出来，但还存在更多启发式的学习规则. 这些学习规则受到动物行为实验中一

些发现的启发，能够在解释生物学现象的同时，实现对不同任务的学习，其中最突出的例子就是R-STDP学习规则[60].

在动物行为实验中，强化刺激会在它们强化的事件后很久才出现，此时动物会加强神经元活动以进行学习. 通常，动作与其导致的奖励或惩罚之间可能存在较大的时间间隔. 这种延迟强化的问题在行为学文献中被称为远端奖励问题(Distal Reward Problem)[127]；而在强化学习文献中，这被称为信度分配问题(Credit Assignment Problem)[48]. 为解决远端奖励问题，Izhikevich[60]提出了多巴胺调节的STDP学习规则，这在后续的研究[128-129]中被统称为奖励调节的STDP(Reward-modulated STDP，R-STDP)学习规则.

R-STDP学习规则依赖于两个关键假设：(1) Hebb可塑性受到奖励的调节，这可以从奖励与多巴胺的联系以及多巴胺对STDP进行调节的证据中得以证明[60]；(2)突触由资格迹标记，以弥合Hebb共同激活与奖励信号发生之间的时间差距. 这一个假设在纹状体、大脑皮层以及海马体中均取得了直接的生理证据[91]. 如图5所示，对于R-STDP学习规则，其突触权重的更新公式通常被定义为如下形式：

$$\frac{\mathrm{d}e_{ij}(t)}{\mathrm{d}t}=-\frac{e_{ij}(t)}{\tau_e}+STDP(\Delta t)\delta\left(t-t_{j/i}\right) \tag{12}$$

$$STDP(\Delta t)=\begin{cases}A_+e^{-\frac{\Delta t}{\tau_+}}, & if\ \ \Delta t\geqslant 0\\ -A_-e^{\frac{\Delta t}{\tau_-}}, & otherwise\end{cases} \tag{13}$$

$$\frac{\mathrm{d}w_{ij}(t)}{\mathrm{d}t}=R(t)e_{ij}(t) \tag{14}$$

其中t_j和t_i分别代表突触前和突触后神经元发放脉冲的时间，$\Delta t=t_i-t_j$. A_+和A_-是两个正常数，用于调节长时程增强和抑制的强度. τ_+和τ_-也是两个正常数，分别表示正负学习时间窗口的长度. $R(t)$是奖励信号，也可以用于描述诸如多巴胺这样的神经递质在细胞外的浓度.

图5对R-STDP学习规则中突触权重的更新流程进行了直观的描述，其中红色代表突触前神经元发放脉冲的时间早于突触后神经元发放脉冲的时间($\Delta t>0$)，蓝色代表突触前神经元发放脉冲的时间晚于突触后神经元发放脉冲的时间($\Delta t<0$). 图5中的各个流程与突触权重的更新公式存在对应关系，其中“STDP学习窗口”是公式(13)的函数曲线，“资格迹更新”对应于公式(12)，而“权重更新”描述了公式(14)中不同奖励条件下的权重更新. 对于公式(12)，第一项表示资格迹的衰减，其发生在任意时刻；而第二项表示突触前后的成对脉冲对资格迹的瞬时影响，其发生在成对脉冲发放结束的时刻，即当$\Delta t>0$时，$t_{j/i}$取t_i，反之则取t_j.

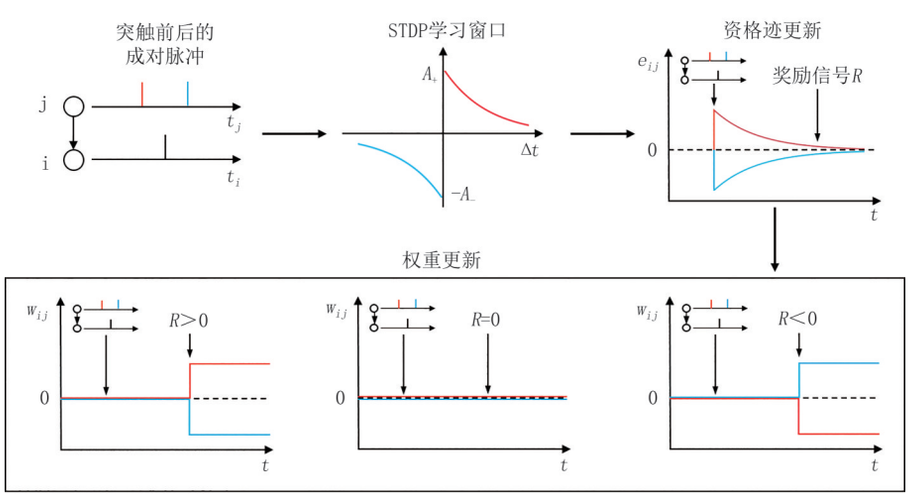

图5 R-STDP学习规则[61]，($\Delta t>0$为红色，$\Delta t<0$为蓝色)

作为奖励学习的重要神经基础，R-STDP学习规则被广泛应用于图像分类[130]、机器人控制[29,44]等任务中，并对三因素学习规则的发展起到了重要的推动作用[91]. Mahadevuni和Li[29]利用R-STDP学习规则在模拟环境中实现了车辆的绕墙避障导航任务. 而Bing等人[44]基于R-STDP学习规则实现了端到端的机器人控制，其中SNN使用DVS作为输入，并计算电机命令作为车道保持任务的输出，成功在仿真环境中取得了比静态控制器更好的性能.

4.1.3　类似评判器的带基线算法

上述的两类学习规则都可以分为代表神经元发放和奖励之间协方差的项和代表无监督学习影响的项. 如果全局的神经调节信号编码了奖励和期望奖励之差，则可以抑制通常对基于奖励的学习有害的无监督项，但前提是要为每个任务和刺激单独计算期望奖励. 如果要同时学习多个任务，则神经系统需要一个内部“评判器”，该“评判器”能够预测任意刺激的期望奖励[61]. 由此推导出的算法被称为类似评判器的带基线算法. 当必须同时学习多个任务时，类似评判器的带基线算法优于不带基线的R-max学习规则或R-STDP学习规则. 此外，对于任意数目的任务，带基线的R-max学习规则总是优于带基线的R-STDP学习规则. 这种优势会随着任务数目的增加而减小[61].

Vasilaki等人[131]在连续时空任务中实现了类似评判器的算法，并对策略梯度算法的局限性进行了深入分析. Frémaux等人[61]实现了一个能够预测期望奖励的内部“评判器”，成功对R-max学习规则与R-STDP学习规则进行了改进. 此外，Farries和Fairhall[128]利用奖励信号与最近收到的奖励的运行平均值之差对TD误差进行近似. Legenstein等人[129]同样通过在奖励调节信号中减去其均值对R-STDP学习规则进行改进，并为猴子生物反馈的基本实验结果提供了解释. 以上这些工作所实现的“评判器”与TD学习中用于实现奖励预测误差的评判器存在一个重大差异，即这种利用奖励的运行试验均值实现的内部“评判器”不需要进行自举，唯一的功能就是中和原本学习规则中的无监督倾向，降低算法的方差.

对于类似评判器的带基线算法，其权重规则具有如下形式：$\Delta w=(R-b)H(pre,post)$，其中，$\Delta w$表示突触权重的变化量，$R$表示全局奖励信号. b表示基线，由于其通常由奖励的运行试验均值实现，也常被写作$\bar{R}$. $H(\cdot)$是突触前脉冲序列与突触后神经元的状态组成的任意函数，用于表示赫布可塑性. 这是突触前和突触后神经元的联合激活引起的突触LTP，通常以资格迹的形式进行实现.

4.1.4　时序差分学习的脉冲时序实现

前几个小节所述的算法仅考虑到仅有行动器的策略梯度算法. 这些基于策略的方法依赖于足够频繁的奖励，但许多现实世界中的问题存在的奖励却是稀疏，延迟且不可靠的. 由于奖励稀疏的问题，智能体难以取得朝向目标的进步，可能会长时间漫无目的地选择动作进行执行. 即使利用基线对这些方法进行改进，它们仍存在着许多问题，例如学习比较缓慢，不便于在线实现，以及难以应用于持续性问题等. 这些问题可以通过时序差分学习进行解决，并结合脉冲时序进行生物学合理的实现.

Potjans等人[132]通过将局部可塑性规则与全局奖励信号相结合，首先实现了行动器-评判器方法. 该方法中的TD学习是TD(0)学习[48]中价值函数更新的生物学合理实现. 在后续的研究中，Potjans等人[49]将利用脉冲神经元网络实现的行动器-评判器中的TD学习与多巴胺能信号调节行为的实验证据[133]相结合，证明了具有真实发放率的多巴胺能信号能够作为第三个因素调节突触可塑性. Frémaux等人[51]根据Doya[134]提出的连续TD学习，利用简化的脉冲响应模型对行动器-评判器网络进行了实现，如图6所示. 他们所提出的TD-LTP学习规则摆脱了传统强化学习对离散框架的依赖，在更符合现实场景的连续时空任务中得到了验证. Zheng和Mazumder[135]提出了一种生物学合理且硬件友好的TD-STDP学习规则，对行动器-评判器结构进行了SNN硬件实现. 此外，Rasmussen和Eliasmith[136]也提出了一种用于脉冲神经元的TD学习算法，根据误差调节神经学习规则[137]对状态-动作价值(Q值)估计进行更新. Bellec等人[52]通过梯度下降实现生物学合理且硬件友好的在线突触可塑性规则，即基于奖励的e-prop学习方法. 他们通过在线行动器-评判器方法对循环连接的脉冲神经网络进行实现，成功解决了包括Atari游戏在内的大多数深度强化学习任务.

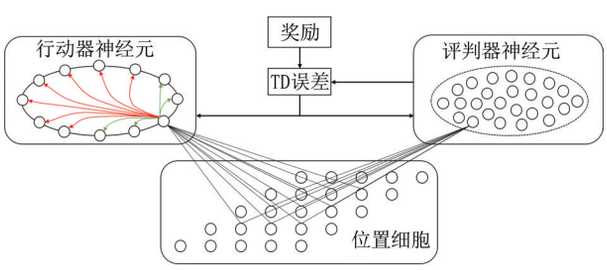

图6　行动器-评判器网络与TD-LTP学习规则[51]

对于三因素学习规则下的TD学习，其权重规则具有如下形式：$w=\delta^{TD}H(pre,post)$，其中，Δw表示突触权重的变化量，δ^{TD}表示TD误差. $H(\cdot)$项表示的含义与上一小节所述的赫布可塑性相同. 基于这类TD学习算法，Nichols等人[138]提出了一种基

于传感器融合的自组织SNN机器人控制系统，成功解决了沿墙绕行任务．在该控制系统中，机器人为SNN提供声呐和激光输入，而SNN为机器人提供电机控制．

4.1.5 多种突触可塑性的结合

大脑的可塑性十分复杂，由近千亿神经元通过百万亿个突触连接组成，这难以用单一的突触可塑性对其进行描述．考虑到突触可塑性的复杂性，通过协调多种突触可塑性，脉冲神经网络能够以快速且稳定的方式进行学习．

O′Brien和Srinivasa[62]提出了一种新颖的类似评判器的算法对R-STDP学习规则进行拓展，以解决同时学习多个远端奖励的问题．尽管其学习能力有限，他们增加了短期可塑性[139]以稳定学习动态．通过将这两种突触可塑性进行结合，整个算法的学习能力得到了极大的提高．

Yuan等人[140]提出了一种生物学合理的强化学习模型，该模型可以协调两种突触的可塑性：随机性和确定性．随机性突触的可塑性是通过享乐主义规则以调节突触神经递质的释放概率来实现的[59]，而确定性突触的可塑性是通过R-STDP学习规则的变体以调节突触权重来实现的．这种变体被命名为Semi-RSTDP学习规则，它仅在突触前脉冲先于突触后脉冲发放时修改突触权重．

4.1.6 超越奖励的学习算法

以多巴胺作为候选神经递质的基于奖励的学习算法，涵盖了大部分利用三因素学习规则实现的强化学习算法．然而，也有部分三因素学习规则用到了除奖励之外的其他神经调节信号，例如惊讶(Surprise)或好奇(Curiosity)信号[69]．

此外，三因素学习规则的一般框架还包含有与群体活动成比例的神经调节信号．例如，Urbanczik和Senn[50]在全局奖励的基础上，对有关群体响应的反馈与突触可塑性之间的调节进行了深入研究，提出了一种能够利用群体规模对学习进行加速的新算法．

4.1.7 最优学习规则搜索

由于三因素学习规则涉及较多的超参数，这些超参数共同决定了最终的学习规则．为了在最大程度上提升任务性能，可以通过超参数优化(Hyper-parameter Optimization)算法[141]找出最优学习规则．

Jordan等人[142]采用笛卡尔遗传规划(Cartesian Genetic Programming)[143]作为进化算法来发现脉冲神经网络中的可塑性规则．这是一种根据任务族定义、相关性能测量和生物学约束实现的自动化方法．对于奖励驱动的任务，他们的方法能够发现具有合适奖励基线的R-STDP学习规则，表现出较强的竞争力．

Lu等人[144]将该方法拓展到具有更高任务复杂性和扩展进化搜索空间的强化学习任务中，利用笛卡尔遗传规划发现的R-STDP学习规则成功解决了XOR和Cart-Pole任务．

4.2 依托ANN实现的脉冲强化学习算法

与ANN中常用的反向传播算法相比，三因素学习规则所使用的全局信号难以对突触权重进行精细地调节，这为训练多层SNN带来了极大的阻碍．通常而言，绝大多数利用三因素学习规则实现的强化学习算法仅能处理一些简单的低维控制任务．即使摆脱了任务维度的局限性，对于不同的任务，同一算法往往需要调整大量超参数，甚至是网络结构，这大大限制了此类算法的应用[52]．为了解决此问题，研究人员充分利用了目前较为成熟的深度强化学习算法，并依托ANN对脉冲强化学习算法进行实现．

4.2.1 ANN-SNN转换

在仿真过程中，脉冲神经元的发放率通常被定义为脉冲神经元的脉冲发放总数与仿真时长之间的商．因此，该发放率的取值范围介于0到1之间，而强化学习中涉及的连续值(连续动作、Q值等)往往不存在取值范围的限制，可以取到介于正负无穷之间的任意浮点值．即使这些连续值存在取值范围的限制，那通常也很难预先得知，无法用发放率的缩放结果对其进行表示．正是这种取值范围上的差异，极大地限制了基于频率编码的脉冲强化学习算法的应用范围．然而，利用训练好的强化学习ANN模型，SNN模型可以巧妙地避开强化学习训练过程中利用价值函数对TD误差的计算步骤．由于SNN的输出层中脉冲神经元的发放率与相应动作的Q值成正比，智能体可以根据发放率的相对大小来选择最优动作，而不需要得到准确的价值估计结果，成功避开了发放率的范围限制，将ANN网络的训练优势与SNN网络的低能耗推断相结合．无需对SNN进行训练，ANN-SNN转换就能够取得与所对应深度网络相当的性能．但是该方法较高的仿真时长，会带来较高的能耗与推理延迟．

Patel等人[47]首先将ANN-SNN转换方法从原先的图像分类领域扩展到使用强化学习算法训练的DQN网络．由于ANN-SNN转换过程需要对神经

网络的权重进行缩放，每一层权重的缩放比例都可以视为需要优化的独立超参数. 为了减少转换过程出现的误差，最大限度地提高SNN的性能，他们采用粒子群优化(Particle Swarm Optimization，PSO)算法[145]对最优缩放参数进行搜索. 转换得到的SNN在Atari Breakout游戏中取得了与原始ANN相当的性能，并通过实验证明了SNN对输入图像的遮挡攻击具有较强的鲁棒性. 然而，PSO算法需要探索的空间与神经网络层数密切相关，由此产生的计算消耗使其难以应对比DQN更深的网络结构，无法充分发挥ANN-SNN转换方法的优势.

Tan等人[42]回顾了ANN到SNN的转换理论，对SNN的发放率与ANN的激活值之间的关系进行了深入分析，其核心公式如下所述：$r_l(t)=\frac{W_l r_{l-1}(t)+b_l-\Delta V_l}{V_{th}}$，其中，$W_l$，$b_l$为具有$L$层的SNN中第$l$层的权重和偏差($l\in\{1,\cdots,L\}$)，$V_{th}$为脉冲神经元的阈值，$\Delta V_l=V_l(t)/t$，$V_l(t)$为$t$时刻第$l$层中脉冲神经元的膜电位. 该公式描述了相邻层之间发放率$r(t)$的关系，而输入层($l=0$)的发放率的公式如下：$r_0(t)=\frac{a_0-\Delta V_0}{V_{th}}$，其中$a_0$为ANN的输入值. 由于不同动作之间的Q值非常接近，为了区分具有相同发放率的脉冲神经元之间的实际最优动作，他们提出了一种发放率的可靠表示方法，以减少转换过程中的误差. 这种发放率的可靠表示方法被定义为如下形式：$f_L(t)=r_L(t)+\frac{V_L(t)}{tV_{th}}$，其中$f_L(t)$是$t$时刻第$L$层中脉冲神经元发放率更鲁棒的表征. 利用他们的方法转换得到的SNN在17个Atari游戏中均获得了与原始DQN相当的得分.

4.2.2　知识蒸馏

Zhang等人[64]从知识蒸馏[65]中得到启发，将原来的ANN教师网络训练ANN学生网络的方式转变为ANN教师网络训练SNN学生网络的方式，提出了带有脉冲蒸馏网络(Spike Distillation Network，SDN)的脉冲强化学习方法. 首先，该方法借助深度强化学习算法(DQN/DDQN[103])对ANN进行训练；然后，利用训练完的ANN搜索动作空间来得到后续SNN训练所需的经验池(Replay Buffer)[10]；随后，从经验池中将对应的环境状态分批输入到待训练的SNN与训练完的ANN中进行前向传播，并计算两者的交叉熵损失；最后，利用计算得到的交叉熵损失在SNN中进行反向传播，完成对其参数的更新.

该方法无需利用脉冲序列对价值函数进行表示，利用ANN教师网络的预测输出实现了对SNN学生网络的训练，为脉冲强化学习算法提供了一条全新的间接训练途径. 通过借鉴深度学习中知识蒸馏的最新技术[146]，这种训练方式在脉冲强化学习算法中将会有较强的竞争力.

4.2.3　脉冲行动器网络

脉冲行动器网络(Spiking Actor Network，SAN)算法基于Actor-Critic方法进行实现，其名字来源于该类方法的网络结构，由ANN实现的评判器网络与SNN实现的行动器网络组成，并使用梯度下降算法进行协同训练. 由于在推断过程中仅需要使用脉冲行动器网络，因此这类算法也被归入脉冲强化学习算法之中. 与ANN-SNN转换相比，基于SAN实现的混合框架协同训练算法可以大大降低仿真时长，减少推断过程中的时间延迟. 此外，ANN-SNN转换往往会由于转换过程中的误差带来性能上的下降，而SAN已经在强化学习任务中证明了其优势[41]. 此外，脉冲行动器网络与知识蒸馏存在一定的相似性，即两者都借助ANN对SNN进行训练. 脉冲行动器网络最大的不同在于其可以兼容于所有基于Actor-Critic方法实现的深度强化学习算法，在不改变原本训练流程的情况下实现对深度行动器网络的完美替代.

Tang等人[63]首先提出混合框架的Actor-Critic方法，并实现了由脉冲行动器网络与深度评判器网络组成的Spiking DDPG算法，其中两个网络使用梯度下降共同训练，其基本框架如图7所示. 由于脉冲发放阈值函数是不可微的，因此需要用替代梯度函数来近似原始函数的梯度. 他们选择矩形函数作为替代梯度函数，该函数被定义为如下形式：

$$z(v)=\begin{cases}a_1, & if\ |v-V_{th}|<a_2\\ 0, & otherwise\end{cases}\tag{15}$$

其中z是替代梯度，a_1是梯度放大器，a_2是传递梯度的阈值窗口. 通过将脉冲发放率fr根据动作空间范围$[V_{min},V_{max}]$进行缩放，脉冲行动器网络能够对连续动作值V进行表示，即$V=(V_{max}-V_{min})fr+V_{min}$. 在实际方法的评估过程中，他们将训练后的SAN部署在Intel的Loihi神经形态处理器上. 在仿真环境和真实的复杂环境中进行验证时，比起Jetson TX2上运行的DDPG算法，Loihi上运行的Spiking DDPG算法能够将每次推断所需要的能耗降低98.7%，并且导航到目标的成功率更高.

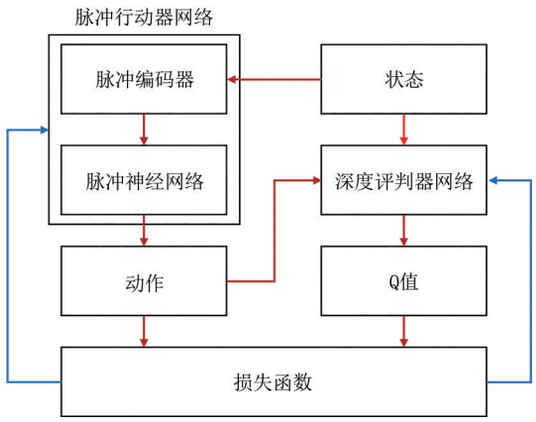

图7 脉冲行动器网络与深度评判器网络[63]

Tang等人[41]在之前SAN算法工作的基础上，提出了一种群体编码的脉冲行动器网络(Population-coded Spiking Actor Network，PopSAN)．通过遍布于大脑神经元网络的群体编码方案，PopSAN算法极大地增强了SNN的表征能力．在群体编码方案中，编码器模块负责将连续观察值转换为输入群体中的脉冲，而解码器模块负责将输出群体活动解码为实值动作．编码器模块分两个阶段计算输入群体的活动：首先将观察值转换为群体中每个神经元的刺激强度$A_E=e^{-\frac{1}{2}\left(\frac{s-\mu}{\sigma}\right)^2}$，其中$s$表示观察空间的连续状态值．$\mu$和$\sigma$分别表示高斯感受野的均值与标准差，这都是任务特定的可训练参数．然后，使用计算出的A_E来生成输入群体中每个神经元的脉冲．脉冲生成的方式可以分为两种：(1)概率编码：在每个时间步骤中以A_E作为概率生成所有神经元的脉冲；(2)确定性编码：在每个时间步骤中将A_E作为群体中每个神经元的突触前输入，其中神经元被建模为软重置的IF神经元．解码器模块分两个阶段将输出群体的活动解码为相应的浮点动作值：首先，将群体中每个神经元的脉冲计数除以仿真时长T得到发放率．然后，将发放率输入到全连接层中进行加权求和以得到浮点动作值．与Jetson TX2上运行的深度行动器网络相比，在Loihi上运行的PopSAN每次推断的能耗减少了99.29%，同时在任务性能上达到了相当的水平．

Zhang等人[147]在PopSAN算法的基础上，提出了多尺度动态编码改进的脉冲行为器网络(Multiscale Dynamic Coding improved Spiking Actor Network，MDC-SAN)．在对输入与输出的群体编码之外，该方法引入了动态神经元(Dynamic Neuron，DN)编码的机制，包含由四个动态参数影响的膜电位的二阶动态．在此之前，SAN算法与PopSAN算法使用的都是基于电流的LIF神经元，其稳定点可以用如下公式表示：$\tau\frac{\mathrm{d}V_t}{\mathrm{d}t}=-V_t+I$，其中$V_t$是$t$时刻的膜电压，$\tau$是膜时间常数，$I$是神经元的输入电流，突触前脉冲会不断整合到该项中．而对于具有高阶动态的神经元，其稳定点数量将不止一个．为简单起见，MAC-SAN算法中使用的是二阶动态神经元，其神经元动态被定义为如下形式：$\frac{\mathrm{d}V_{j,t}}{\mathrm{d}t}=V_{j,t}^2-V_{j,t}-U_{j,t}+I_{i,t}$，其中，$V_{j,t}^2$与$V_{j,t}$是$t$时刻突触后神经元$j$的不同阶动态膜电位，$U_{j,t}$是用于模拟超极化的电阻项，$I_{i,t}$是由来自突触前神经元$i$的脉冲整合得到的输入电流．除了膜电位之外，动态神经元中还使用四个超参数以及其他的一些隐变量来描述神经元的二阶动态，其公式如下所述：

$$\begin{cases}\dfrac{\mathrm{d}U_{j,t}}{\mathrm{d}t}=\theta_v V_{j,t}-\theta_u U_{j,t} & \\ V_{j,t}=\theta_r, & if\ V_{j,t}>V_{th} \\ U_{j,t}=U_{j,t}+\theta_s, & if\ spike\end{cases} \tag{16}$$

其中，$V_{j,t}$的稳定点由$U_{j,t}$与输入电流$I_{i,t}$共同决定．稳定点的数量和取值受到$\theta_{v,u,r,s}$四个参数的动态影响．这些参数分别表示V的电导率，U的电导率，重置膜电位以及U的脉冲改进．所有动态参数会在学习过程中连同突触权重共同针对一个任务进行调整．随后这些参数会用K-means算法[148]进行聚类，以获得$\theta_{v,u,r,s}$参数组的最优中心．最优中心所对应的四个关键参数将进一步作为所有任务中动态神经元的固定配置．在OpenAI Gym的四个连续控制任务[149]中，该二阶动态神经元都取得了比基于电流的LIF神经元更好的性能．

4.3 基于脉冲的直接强化学习算法

尽管脉冲神经元的发放率难以直接对强化学习中的连续值进行表示，例如价值函数与策略，但仍然有许多方法能够将脉冲序列解码成算法所需要的连续值．这些独特的脉冲编码方案促使一类新算法的产生，即基于脉冲的直接强化学习算法．与依托ANN实现的脉冲强化学习算法相比，该方法最大的特点在于无需借助ANN就能完成SNN的直接训练．

Liu等人[66]提出了一个深度脉冲Q网络(Deep Spiking Q-Network，DSQN)，并基于频率编码成功解决了令人困扰的Q值问题．对于输出层L的神经元，它们的输出O^L可以被描述为：$O^L=$

$W^L\left(\frac{1}{T}\sum_{t=1}^{T}S^{L-1,t}\right)$，其中$W^L$是输出层神经元的可学习权重，$T$是仿真时长，$S^{L-1,t}$表示第$t$步仿真时输出层的输入脉冲，$O^L$可以用于表示DSQN输出的Q值. 此外，他们采用反正切函数替代梯度函数以实现对DSQN的直接训练，该函数被定义为如下形式：$\sigma_{arctan}(\alpha x)=\frac{1}{\pi}arctan\left(\frac{\pi}{2}\alpha x\right)+\frac{1}{2}$，其中，$\alpha$是控制函数平滑度的参数. 反正切函数的梯度可以被表示为：$\sigma'_{arctan}(\alpha x)=\frac{\alpha}{2\left[1+\left(\frac{\pi}{2}\alpha x\right)^2\right]}$. 他们在17个人类玩家表现最好的Atari游戏进行了综合实验，证明了与最先进的ANN-SNN转换方法相比，此方法在性能、稳定性、鲁棒性和能源效率等方面的优越性. 同时，DSQN达到了与DQN大致相当的性能，并在稳定性方面超过了DQN.

与此同时，Chen等人[43]受到生物体内不发放脉冲的中间神经元的启发，利用不发放脉冲的神经元(Non-spiking Neurons)将输出的脉冲序列解码为连续值. 这种不发放脉冲的神经元在后续被简称为非脉冲神经元. 借助其膜电位整合输出的脉冲以表示Q值，他们提出了一种深度脉冲Q网络的实现方案，如图8所示. 在推断过程中，智能体从环境中接收状态，并通过脉冲神经元将它们编码为脉冲序列；随后脉冲序列被输入到SNN中，并在输出层的非脉冲神经元上累积膜电压，用于表示每个动作对应的Q值；最后，智能体选择具有最大Q值的动作. 对于该算法中使用的不发放脉冲的LIF神经元模型(这也被称为LI神经元模型)，其动态可以被描述为以下形式：$V_t=V_{t-1}+\frac{1}{\tau}\left(-(V_{t-1}-V_{reset})+X_t\right)$，其中$V_{reset}$是复位电压，$X_t$是神经元的输入，$\tau$表示神经元的膜时间常数. 在整个仿真时长$T$内，非脉冲神经元以脉冲序列作为输入，可以得到每个时间步骤t的膜电压V_t. 为了表示价值函数的输出，他们提出了三个统计量作为候选方案，分别是最后时刻膜电压($Q=V_T$)、最大膜电压($Q=\max_{1\leqslant t\leqslant T}V_t$)以及平均膜电压($Q=\frac{1}{T}\sum_{t=1}^{T}V_t$). 对17个Atari游戏进行的实验表明，他们所提出的DSQN是有效的，并且在大多数游戏中甚至优于相同网络规模的DQN. 此外，他们也对DSQN的低能耗以及对抗性攻击的鲁棒性进行了论证分析，全面展示了直接训练算法相比于ANN-SNN转换方法的优越性.

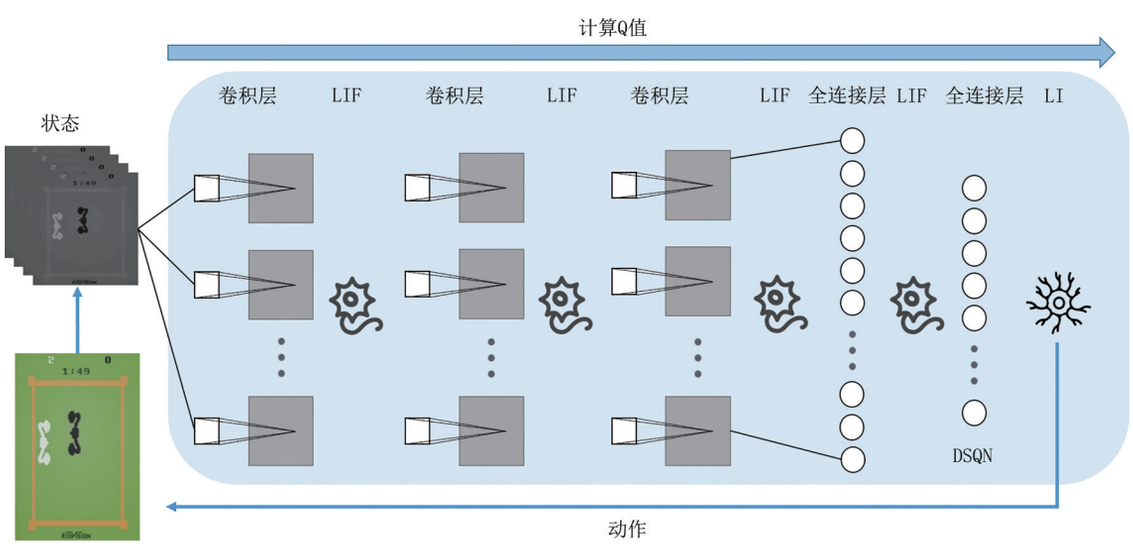

图8　深度脉冲Q网络[43]

如表1所示，我们选取了各大类脉冲强化学习算法中的代表性算法，对其在不同强化学习任务上的表现进行了总结，其中性能指任务训练过程中智能体评估获得的最大奖励均值. 对于表1中定性的性能对比结果，表2和表3提供了目前脉冲强化学习算法在Atari游戏与OpenAI Gym上的连续控制任务(使用强化学习环境MuJoCo)中更详细的定量性能对比. 由于某些脉冲强化学习算法仅提供了在指定游戏上的性能或者学习曲线，如表1的任务列所述. 表2中包含的算法都在较多的Atari游戏上进行了实验. 此外，在Atari游戏中，各个算法由于没有统一设置随机种子与超参数，在性能上存在较大

的差异.因此,表2提供的性能被表示为采用此算法与同样设置下DQN算法的智能体游戏得分之间的百分比,其中空出了部分无法用该指标准确比较的算法性能.表3提供的是智能体的平均得分,并且包含部分用于对比的深度强化学习算法,分别是深度行动器网络(Deep Actor Network, DAN)与群体编码的深度行动器网络(Population-coded Deep Actor Network, Pop-DAN)[147].

表1 各大类中代表性的脉冲强化学习算法总结

类型	脉冲强化学习算法	对应深度算法	任务	性能对比
利用三因素学习规则实现的强化学习算法	e-prop[52]	A3C-LSTM[111]	Atari Pong/ Fishing Derby	两者相当
ANN-SNN转换	转换DQN[42]	DQN[10]	Atari	两者相当
知识蒸馏	SDN[64]	DQN/DDQN[103]	Atari Pong	略优于对应的深度算法
脉冲行动器网络	MDC-SAN[147]	TD3[117]	OpenAI Gym连续控制	优于对应的深度算法
基于脉冲的直接强化学习算法	DSQN[43]	DQN	Atari	略优于对应的深度算法

表2 算法性能对比(Atari游戏)

游戏	转换DQN[42] (%DQN)	DSQN[66] (%DQN)	DSQN[43] (%DQN)
Atlantis	100.6%	98.8%	84.2%
Beam Rider	105.6%	97.5%	99.6%
Boxing	94.4%	99.2%	298.2%
Breakout	91.9%	90.9%	144.4%
Crazy Climber	96.8%	102.8%	109.8%
Gopher	94.6%	95.8%	149.0%
Jamesbond	103.3%	127.6%	113.9%
Kangaroo	98.8%	214.5%	94.6%
Krull	121.8%	106.8%	28.7%
Name This Game	103.5%	98.9%	152.4%
Pong	99.7%	100.5%	
Road Runner	101.3%	89.7%	917.3%
Space Invaders	102.9%	80.5%	106.9%
Star Gunner	94.7%	113.0%	153.7%
Tennis	95.2%		
Tutankham	106.8%	103.9%	280.5%
Video Pinball	106.4%	87.0%	160.0%

表3 算法性能对比(OpenAI Gym上的连续控制任务)

行动器网络	Ant	HalfCheetah	Hopper	Walker2d
DAN[147]	4671±777	10020±1696	3396±77	4139±429
Pop-DAN[147]	5146±669	10206±1089	3201±957	4361±621
PopSAN[41]	5416±673	10729±644	3599±68	4995±682
MDC-SAN[147]	5628±155	11657±400	3636±146	5566±332

5 研究挑战与未来研究方向

尽管深度强化学习算法在广泛的控制任务中展现出了极强的能力,但是在实际场景中,有限的主板能量资源无法满足深度强化学习算法极高的能耗需求,大大限制了其在机器人、自主控制等领域的应用.随着神经形态计算领域的发展,脉冲强化学习算法作为SNN的重要方向之一,成为了机器人控制以及脑科学等领域的热点.由于脉冲编码以及事件驱动特性带来的低功耗优势,脉冲强化学习算法可以作为许多实际控制任务的低功耗解决方案,具有广阔的市场前景.此外,基于事件的传感和信号处理的神经形态传感器已应用于许多领域,具有可喜的结果和若干概念优势[46].尤其是对于神经形态视觉传感器,因其具有低功耗、低数据冗余、高时间分辨率、高动态范围等优点,在自动驾驶、无人机导航、视频监控以及工业检测等传统机器视觉领域中都有着巨大的应用价值.依托于这些神经形态传感器,脉冲强化学习算法可以充分发挥事件驱动的特性,实现感知处理一体化的多模态网络,在复杂交互的真实环境中对不同形式的任务进行学习.同时,通过对大脑学习过程中神经元动态的建模,脉冲强化学习算法可以对实验生物学的发现进行解释与验证,并指导后续神经科学研究的进展,这也是探索大脑智能的有效途径之一.

若要充分发挥脉冲强化学习算法的理论优势,并将其实际落地,目前仍存在许多问题与挑战.除了全面剖析脉冲强化学习算法所面对的研究挑战外,本章还进一步讨论了这些研究挑战的解决方案以及未来研究方向.

5.1 研究挑战

然而,对于脉冲强化学习算法的研究与应用目前仍处于起步阶段,在解决实际应用问题或者解释大脑学习机制时存在许多问题与局限.具体而言,脉冲强化学习算法当前主要面临的是以下五个方面的挑战:

(1)脉冲编码方案的局限性:目前主流的脉冲编码方案是频率编码,但是其无法充分表达神经元脉

冲序列中的时序关系. 为了更好地将观察到的数据转换为脉冲信号，需要一套计算高效且生物学合理的脉冲编码方案. 有关大脑中信息处理的详细知识，能够在人工和生物脉冲神经网络之间建立映射关系，从中汲取灵感将有助于脉冲编码方案的设计.

(2)脉冲时序与强化学习时序的割裂：对于脉冲神经网络，脉冲发放的时间序列存储着重要信息. 而在强化学习场景下，智能体不断与环境进行交互，产生数据序列，这种决策顺序就是强化学习的时序. 除去利用三因素学习规则实现的强化学习算法，余下的两大类脉冲强化学习算法表现出两者时序之间明显的割裂. 对于这些算法，仿真时长内的脉冲时序发生在强化学习算法的单步决策之中，这与大脑学习中的真实时序大相径庭. 因此，将这两种时序关系巧妙地结合在一起，充分发挥SNN时序数据处理的能力，这是脉冲强化学习算法的一大关键点.

(3)生物学合理性与性能表现之间的矛盾：尽管目前的脉冲强化学习算法已经能够在Atari游戏以及OpenAI Gym上的连续控制任务中取得优于深度强化学习基线算法的性能，但比起Agent57[150]、MuZero[40]等最先进的深度强化学习算法仍有不小的差距. 此外，深度强化学习算法已经在广泛的控制任务上证明了其有效性，而脉冲强化学习算法的应用范围在目前仍然较为受限. 因此，相比于深度强化学习算法，脉冲强化学习算法能更好地解决哪些任务是一个重要的问题. 传统的脉冲神经网络模型非常类似于大脑中观察到的机制，但是事实证明很难建立能够学习与生物电路具有类似复杂性与性能的行为模型. 相反，深度神经网络与生物大脑完全不同，但它们有史以来第一次能够以与真实大脑的能力相媲美的水平解决复杂的问题. 这表明生物学合理性与性能表现之间似乎存在不可调和的矛盾. 如何深入挖掘脉冲架构的优势，利用脉冲强化学习算法解决传统算法难以完成的任务，实现生物学合理性与性能表现的双赢，这将是一个长期的研究课题.

(4)脉冲强化学习仿真平台的欠缺：对于强化学习算法，一个合适的仿真平台至关重要. 而现有的仿真平台都仅针对传统的强化学习算法，与脉冲强化学习算法并不契合. 对于脉冲强化学习仿真平台的搭建，目前主要有三种解决途径：将传统仿真平台经过脉冲编码映射为脉冲仿真平台、基于神经形态视觉传感器仿真构造脉冲仿真平台，以及借助神经形态视觉传感器采集的真实数据构造脉冲仿真平台. 然而，现在并没有一个主流的脉冲强化学习仿真平台可以用于算法验证，研究人员通常只能在传统的强化学习仿真平台(例如OpenAI Gym)上对经典控制任务进行仿真. 如何开发出一个计算高效、任务丰富且具有现实价值的脉冲强化学习仿真平台，这是一个亟待解决的问题.

(5)软件仿真低下的计算效率：尽管SNN最终部署在专门的神经形态硬件上，但研究人员仍然需要有效的工具来开发标准硬件(例如GPU)上的转换和训练算法. 目前大部分的SNN仿真工具都是为计算神经科学应用程序而设计的，因此缺乏对卷积层和批处理等常见深度学习元素的本地支持. 因此，许多研究人员选择在PyTorch、TensorFlow等深度学习平台上开发替代的SNN学习框架，以便他们可以继续利用这些平台提供的机器学习基础操作. 然而，这些深度学习平台的工作原理都是通过对张量数据类型进行诸如矩阵乘法和归约操作以完成计算图的组装. 这些操作被实现为高度优化的GPU核，利用CUDA对计算进行深度加速. 这种底层设计对SNN的稀疏性支持不佳，这使得它们在大型SNN模型仿真时计算效率低下. 考虑到强化学习算法本身大量的训练步骤，以及仿真时长与复杂神经元模型的影响，软件仿真低下的计算效率已经成为制约脉冲强化学习算法研究的重大问题.

5.2　未来研究方向

伴随着深度学习技术的发展，脉冲强化学习算法研究重新引起了广大学者的关注，出现了欣欣向荣的景象. 基于脉冲强化学习算法的研究现状与研究挑战，本章对该方向的未来进行展望，并讨论了七个可能的研究方向以及一些简单的研究思路.

5.2.1　脉冲编码方式

如何将连续的状态输入转化为脉冲序列，这是脉冲编码考虑的主要问题. 此外，由于目前采用的脉冲神经元发放率通常在0到1之间，无法准确表示强化学习算法中的连续值. 因此，需要借助特殊的解码方案将脉冲序列转换为强化学习算法所需的连续值. 这个方向的具体研究思路如下：

(1)群体编码方案：利用神经元群体对输入状态空间以及网络输出进行编码，对现有的脉冲发放率编码进行改进，以便更好地将脉冲序列与传统算法中输入/输出的连续值进行转换.

(2)神经元编码方案：参考H-H模型与Izhikevich

模型，采用复杂性与自适应性更强的脉冲神经元模型，涵盖更详细的神经元动态，以增强SNN模型的表征能力.

(3)基于非脉冲神经元膜电压的解码方案：通过仿真时间内非脉冲神经元膜电压的某个统计量对价值函数进行表示，对现有的强化学习算法进行实现，例如深度Q学习以及基于Actor-Critic的深度学习方法.

5.2.2 时序信息处理

在目前的脉冲强化学习算法中，仿真时长内的脉冲时序发生在强化学习算法的单步决策之中，这与大脑学习中的真实时序大相径庭，也无法充分发挥SNN处理时序数据的能力. 因此，需要进一步挖掘脉冲时序数据之中蕴含的信息. 这个方向的具体研究思路如下：

(1)脉冲强化学习算法处理神经形态传感器输出的脉冲时序数据：由于神经形态传感器的输出信号可以通过脉冲时序数据的形式输入SNN，因此可以直接利用脉冲强化学习算法对依托其实现的具体任务进行研究.

(2)脉冲循环神经网络处理强化学习任务：通过在深度强化学习算法中引入RNN，可以克服长期依赖给学习策略带来的困难，解决部分可观察的马尔可夫决策过程. 因此，与单纯利用神经元内部状态体现时序性的SNN不同，利用循环连接的脉冲神经元，可以充分发挥循环神经网络以及脉冲神经网络处理时序信息的能力，并且大大降低完成强化学习任务所需的能耗.

5.2.3 利用神经形态传感器的仿真平台开发与应用

脉冲强化学习仿真平台可以通过对神经形态传感器的仿真实现，将原有的强化学习仿真环境转变成脉冲版本，或者对传统强化学习无法完成的任务进行仿真平台的搭建. 通过搭载神经形态传感器的可控制设备，可以得到控制信号与真实脉冲数据构成的序列，利用这些真实的数据实现脉冲强化学习算法的应用. 这个方向的具体研究思路如下：

(1)传统室内场景以及室外交通平台脉冲版本的仿真实现：在传统的室内场景以及室外交通强化学习仿真平台中加入神经形态视觉传感器接口，实现脉冲版本的仿真实现.

(2)利用采集到的神经形态数据集进行算法研究：借助采集到的神经形态数据集，对算法进行研究. 这对数据集的采集有着较高的要求，因此需要针对应用进行大规模的数据采集. 目前，公开的大规模数据集较少，尤其是在复杂的控制任务，这是一个亟待解决的问题.

(3)利用神经形态视觉传感器的脉冲强化学习算法应用：在车辆、机器人、无人机等移动设备上搭载神经形态视觉传感器与神经形态芯片，通过脉冲强化学习算法实现类脑感知与决策. 对于自动驾驶任务，脉冲强化学习算法应用如图9所示. 在高速驾驶过程中，驾驶员在遇到紧急情况时往往因为心理因素需要较长的时间对其进行反应. 传统相机在遇到高速物体时，通常会出现运动模糊，难以敏锐地感知到具体情况，并且容易受到光照条件的限制. 与此相比，神经形态视觉传感器具有高时间分辨率，能够准确感知高速目标并在微秒级的时间内对紧急情况进行响应. 因此，神经形态芯片与脉冲强化学习算法的结合有望为自动驾驶提供一套可行的低能耗解决方案，用于提升自动驾驶的续航时间以及针对紧急事件的处理能力[151-153].

(4)不同神经形态传感器融合的多模态控制系统：将视觉、触觉、听觉等不同模态的神经形态传感器进行融合，在感知与决策层面对大脑进行模拟，应用于复杂交互环境下的智能控制系统.

5.2.4 基于开源脉冲神经网络学习框架的优化

基于现有的深度学习框架，对脉冲神经网络学习框架进行开发，并对其进行计算优化与CUDA加速. 同时，通过在开源脉冲神经网络学习框架中拓展强化学习算法与仿真应用示例，可以为脉冲强化学习研究提供一套完整的基准测试方案，便于后续研究人员部署与测试. 此外，针对脉冲强化学习算法，还需要在框架中支持神经形态芯片的测试，并且添加各类仿真平台的接口，这包括常见的强化学习仿真平台(例如OpenAI Gym)以及利用神经形态传感器的仿真平台. 通过在框架中集成现有的脉冲编码库、强化学习算法库以及仿真环境库，这能够极大地方便后续的研究人员加入到脉冲强化学习算法的研究之中.

5.2.5 神经科学启发的新型脉冲强化学习算法设计

作为脉冲强化学习算法的一个重要灵感来源，神经科学中的许多研究成果对设计新型脉冲强化学习算法都有着借鉴意义，具体以两个神经科学中的发现为例：

(1)横向相互作用：生物神经元中存在广泛的横向相互作用(Lateral Interactions)，而神经科学家发

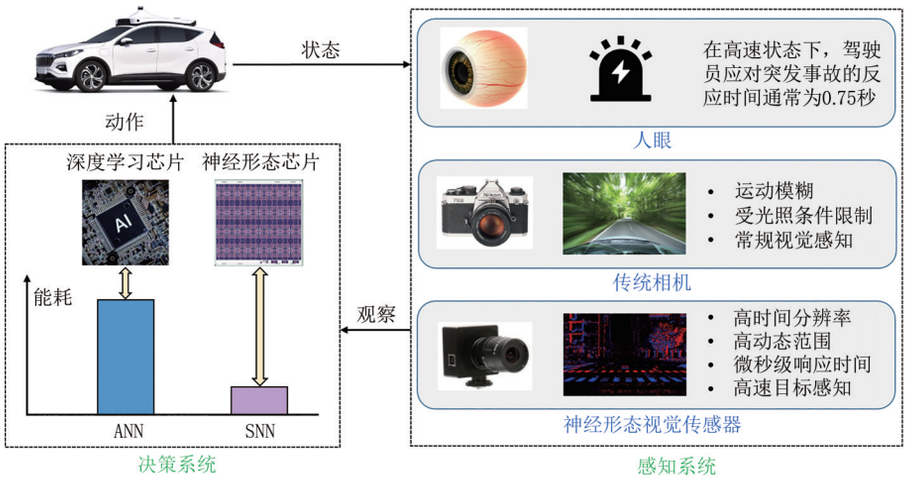

图9　脉冲强化学习算法在自动驾驶中的应用示意图

现视网膜神经元中的横向相互作用可以增强视觉对象边缘[154]. 在图像分类任务上，横向相互作用已被证明有助于提高SNN的性能与鲁棒性[155]. 而目前脉冲强化学习算法中仅在动作选择中实现了简单的横向相互作用[51]. 因此，通过对状态感知以及动作决策过程的改进，横向相互作用机制有望提高脉冲强化学习算法对任务相关目标的感知能力以及决策的鲁棒性，存在较大的研究空间与应用价值.

(2)奖惩分离机制：大脑的背侧纹状体具有多种行为功能，包括运动控制，强化学习等. 它主要由表达D1型或D2型多巴胺受体的中型多棘神经元(Medium Spiny Neurons, MSNs)组成，分别引起“直接通路”(Direct Pathway)和“间接通路”(Indirect Pathway). 一种流行的模型认为，直接和间接通路具有相反的功能，前者促成运动并促进奖励或正强化，后者抑制运动并促进惩罚或负强化[156]. 然而，近期的研究表明，在运动过程中直接和间接通路中的神经元共同激活，而不会相互拮抗[157]. 因此，通过兴奋性和抑制性的神经元，可以在脉冲强化学习算法中用不同网络对这两个通路进行实现. 这种奖惩分离机制有助于人类更好地理解大脑中的学习机制，从而得到更加具有生物学合理性的新型脉冲强化学习算法.

5.2.6　基于具身图灵测试的模型设计

近期，包括Yoshua Bengio和Yann LeCun两位图灵奖获得者在内的一大批机器学习与神经科学领域的研究学者发布了神经人工智能(NeuroAI)白皮书[158]. 他们提出，为了加速人工智能的进步，必须推动神经人工智能的基础研究，而其中的核心组成部分即为具身图灵测试(Embodied Turing Test). 具身图灵测试要求人工智能模型能够以与动物相似的技能水平与世界互动，从而将研究重点从游戏和语言等人类特别发达或独特的能力转移到了那些经过5亿多年的进化之后所有动物继承下来的共同能力. 白皮书指出，一个合格的人工智能系统需要具有三个物种之间的共同特征：能够以有目标的方式四处移动并与环境互动、行为灵活以及计算效率高，而这正是脉冲强化学习算法的一个主要研究内容，即为机器人、自主控制等领域提供一套鲁棒且低能耗的解决方案. 因此，借助脉冲强化学习算法，我们有希望构建出能够通过具身图灵测试的模型，从而为下一代人工智能的研究提供路线图.

5.2.7　针对强化学习痛点问题的脉冲解决方案

尽管深度强化学习已经在广泛的控制任务上证明了其有效性，但是仍存在一些痛点问题亟待解决：

(1)较差的鲁棒性：以像素游戏为例，与人类玩家不同，深度强化学习智能体很脆弱，对小扰动非常敏感——稍微改变游戏规则，甚至改变输入上的几个像素，都可能导致灾难性的低性能[159]. 研究表明，脉冲强化学习算法由于其内在的脉冲发放机制，对于图像遮挡以及对抗攻击表现出了较强的鲁棒性[43, 47]. 如何从理论角度解释说明这种鲁棒性并将其充分利用，将会是一个很好的研究课题. 此外，脉冲强化学习算法是否对带噪奖励同样存在鲁棒性，这也值得深入研究.

(2)较大的性能方差：深度强化学习算法的性能与超参数以及随机种子的选取有着密切的关系，往往表现出较大的方差，即缺乏稳定性. 目前已有研究表明脉冲强化学习算法在稳定性上超过了对应的深度算法[66]. 然而，这个结果是否具有普遍性，仍

有待后续研究的验证.或许从脉冲强化学习算法对状态与奖励中各种扰动的鲁棒性切入,我们将能够得出一个极具价值的结论.

(3)较大的回放缓存(Replay Buffer):深度强化学习算法往往采用经验回放来提高样本效率,但随之也带来了较大的回放缓存.由于脉冲的二值特性,若将其作为状态值存入回放缓存中,就可以大大降低算法对设备内存的需求.因此,研究重点转为了如何设计出一种脉冲编码方案,能够最大程度地保留原始数据的特征信息,并在强化学习训练之前对原始数据进行脉冲化的预处理.

6 结 论

脉冲强化学习算法属于神经科学与人工智能方向的交叉研究,其终极目标之一就是模拟大脑结构与学习机理,建立一个人工大脑.同时,人工大脑的搭建也能反过来促进我们对生物大脑的理解.在发展初期,脉冲强化学习算法侧重于利用强化学习算法作为理论依据,解释动物学实验的结果.通过对实验采集的脉冲序列数据进行建模,研究人员总结出了一套完整的三因素学习规则理论.同时,伴随着三因素学习规则理论的发展,脉冲强化学习算法也为动物学实验的设计提供了灵感来源,推动着人类不断加深对大脑工作机制的理解.两者紧密相连,形成了一个良性的循环.随着深度学习的兴起以及随后的快速发展,SNN从中汲取智慧,在许多领域接近甚至达到了ANN的性能.由于其独特的信息编码模式,SNN具有低功耗、高鲁棒性等特点,并在处理复杂时序数据、持续学习等方面展现出了不错的潜力,而这些也是当前强化学习领域亟待解决的关键问题.因此,后期的脉冲强化学习算法注重于在强化学习领域发挥SNN的优势,解决将SNN应用于深度强化学习过程中遇到的训练问题.

与ANN相比,SNN还处于研究的初级阶段,其与强化学习的结合目前也才刚刚起步.尽管如此,脉冲强化学习算法已经能够在某些基准任务上达到甚至超过深度强化学习基线算法的性能,并在能耗以及鲁棒性等方面取得了明显的优势.然而不可否认的是,伴随着海量人力与物力的投入,深度强化学习算法取得了巨大的进展,比起脉冲强化学习算法在适用任务范围以及典型场景上的性能仍有着一定的优势.因此本文的目的之一就是希望引起人们对脉冲强化学习算法这一重要领域的关注,推动该领域的快速发展.

今后,脉冲强化学习算法将立足于神经科学与人工智能领域.在理论层面,将SNN与强化学习进行深度融合,实现生物学合理性与性能表现的双赢.通过突触可塑性理论与优化理论的突破,脉冲强化学习算法将在计算能效、任务性能、模型鲁棒性、生物可解释性、持续学习以及现实应用等方面取得重大进展,有望成为延续数十年来神经科学与人工智能之间良性循环的关键,为下一代人工智能的研究提供路线图.在应用层面,通过与神经形态芯片以及神经形态传感器的结合,脉冲强化学习算法极其有望为机器人、自主控制等领域提供一个鲁棒且低能耗的解决方案,用于应对资源受限情况下复杂环境的现实任务,并对紧急事件进行快速处理,具有巨大的应用潜力.

参 考 文 献

[1] Hubel D H, Wiesel T N. Receptive fields and functional architecture of monkey striate cortex. The Journal of Physiology, 1968, 195(1): 215-243

[2] Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, 1980, 36(4): 193-202

[3] Riesenhuber M, Poggio T. Hierarchical models of object recognition in cortex. Nature Neuroscience, 1999, 2(11): 1019

[4] Lindsay G W. Attention in psychology, neuroscience, and machine learning. Frontiers in Computational Neuroscience, 2020, 14: 29

[5] Banino A, Barry C, Uria B, et al. Vector-based navigation using grid-like representations in artificial agents. Nature, 2018, 557(7705): 429-433

[6] Mattar M G, Daw N D. Prioritized memory access explains planning and hippocampal replay. Nature Neuroscience, 2018, 21(11): 1609-1617

[7] Dabney W, Kurth-Nelson Z, Uchida N, et al. A distributional code for value in dopamine-based reinforcement learning. Nature, 2020, 577(7792): 671-675

[8] Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 1958, 65(6): 386

[9] Hinton G E, Dayan P, Frey B J, et al. The "wake-sleep" algorithm for unsupervised neural networks. Science, 1995, 268(5214): 1158-1161

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

[11] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436-444

[12] Cox D D, Dean T. Neural networks and neuroscience-inspired

computer vision.Current Biology，2014，24(18)：921-929

[13] Roy K，Jaiswal A，Panda P. Towards spike-based machine intelligence with neuromorphic computing. Nature，2019，575(7784)：607-617

[14] Mead C. Neuromorphic electronic systems. Proceedings of the IEEE，1990，78(10)：1629-1636

[15] Merolla P A，Arthur J V，Alvarez-Icaza R，et al. A million spiking-neuron integrated circuit with a scalable communication network and interface.Science，2014，345(6197)：668-673

[16] Davies M，Srinivasa N，Lin T H，et al.Loihi：A neuromorphic manycore processor with on-chip learning. IEEE Micro，2018，38(1)：82-99

[17] Furber S B，Galluppi F，Temple S，et al.The spinnaker project. Proceedings of the IEEE，2014，102(5)：652-665

[18] Gerstner W，Kistler W M. Spiking neuron models：Single neurons，populations，plasticity. Cambridge，UK：Cambridge University Press，2002

[19] Maass W.Networks of spiking neurons：The third generation of neural network models. Neural Networks，1997，10(9)：1659-1671

[20] Gerstner W，Kistler W M，Naud R，et al. Neuronal dynamics：From single neurons to networks and models of cognition. New York，USA：Cambridge University Press，2014

[21] Fang W，Yu Z，Chen Y，et al. Incorporating learnable membrane time constant to enhance learning of spiking neural networks//Proceedings of the IEEE/CVF International Conference on Computer Vision.Montreal，Canada，2021：2661-2671

[22] Fang W，Yu Z，Chen Y，et al.Deep residual learning in spiking neural networks//Proceedings of the Advances in Neural Information Processing Systems. Montreal，Canada，2021：21056-21069

[23] Guo S，Yu Z，Deng F，et al. Hierarchical Bayesian inference and learning in spiking neural networks. IEEE Transactions on Cybernetics，2017，49(1)：133-145

[24] Wang Y，Xu Y，Yan R，et al. Deep spiking neural networks with binary weights for object recognition.IEEE Transactions on Cognitive and Developmental Systems，2020，13(3)：514-523

[25] Kim S，Park S，Na B，et al. Spiking-yolo：Spiking neural network for energy-efficient object detection//Proceedings of the AAAI Conference on Artificial Intelligence. New York，USA，2020：11270-11277

[26] Dominguez-Morales J P，Liu Q，James R，et al. Deep spiking neural network model for time-variant signals classification：A real-time speech recognition approach//Proceedings of the International Joint Conference on Neural Networks. Rio de Janeiro，Brazil，2018：1-8

[27] Yin B，Corradi F，Bohté S M. Accurate and efficient time-domain classification with adaptive spiking recurrent neural networks.Nature Machine Intelligence，2021，3(10)：905-913

[28] Yu Q，Li S，Tang H，et al. Toward efficient processing and learning with spikes：New approaches for multispike learning. IEEE Transactions on Cybernetics，2020，52(3)：1364-1376

[29] Mahadevuni A，Li P.Navigating mobile robots to target in near shortest time using reinforcement learning with spiking neural networks//Proceedings of the International Joint Conference on Neural Networks.Anchorage，USA，2017：2243-2250

[30] Liu Q，Xing D，Tang H，et al. Event-based action recognition using motion information and spiking neural networks//Proceedings of the International Joint Conference on Artificial Intelligence.Montreal，Canada，2021：1743-1749

[31] Stromatias E，Neil D，Pfeiffer M，et al. Robustness of spiking deep belief networks to noise and reduced bit precision of neuro-inspired hardware platforms. Frontiers in Neuroscience，2015，9：222

[32] Sharmin S，Rathi N，Panda P，et al. Inherent adversarial robustness of deep spiking neural networks：Effects of discrete input encoding and non-linear activations//Proceedings of the European Conference on Computer Vision. Glasgow，US，2020：399-414

[33] De Lange M，Aljundi R，Masana M，et al.A continual learning survey：Defying forgetting in classification tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence，2021，44(7)：3366-3385

[34] Kudithipudi D，Aguilar-Simon M，Babb J，et al. Biological underpinnings for lifelong learning machines. Nature Machine Intelligence，2022，4(3)：196-210

[35] Zuo F，Panda P，Kotiuga M，et al. Habituation based synaptic plasticity and organismic learning in a quantum perovskite. Nature Communications，2017，8(1)：1-7

[36] Skatchkovsky N，Jang H，Simeone O. Bayesian continual learning via spiking neural networks. arXiv preprint arXiv：2208.13723，2022

[37] Bing Z，Meschede C，Röhrbein F，et al. A survey of robotics control based on learning-inspired spiking neural networks. Frontiers in Neurorobotics，2018，12：35

[38] 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

[39] Vinyals O，Babuschkin I，Czarnecki W M，et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature，2019，575(7782)：350-354

[40] Schrittwieser J，Antonoglou I，Hubert T，et al.Mastering atari，go，chess and shogi by planning with a learned model. Nature，2020，588(7839)：604-609

[41] Tang G，Kumar N，Yoo R，et al. Deep reinforcement learning with population-coded spiking neural network for continuous control//Proceedings of the Conference on Robot Learning. Cambridge，USA，2021：2016-2029

[42] Tan W，Patel D，Kozma R. Strategy and benchmark for converting deep Q-networks to event-driven spiking neural networks//Proceedings of the AAAI Conference on Artificial Intelligence.Online，2021：9816-9824

[43] Chen D，Peng P，Huang T，et al. Deep reinforcement learning with spiking q-learning.arXiv preprint arXiv：2201.09754，2022

[44] Bing Z，Meschede C，Huang K，et al. End to end learning of spiking neural network based on r-stdp for a lane keeping vehicle//Proceedings of the IEEE International Conference on Robotics and Automation. Brisbane，Australia，2018：4725-

4732

[45] Li Jia-Ning, Tian Yong-Hong. Recent advances in neuromorphic vision sensors: A survey. Chinese Journal of Computers, 2021, 44(6): 1258-1286 (in Chinese)
(李家宁, 田永鸿. 神经形态视觉传感器的研究进展及应用综述. 计算机学报, 2021, 44(6): 1258-1286)

[46] Tayarani-Najaran M H, Schmuker M. Event-based sensing and signal processing in the visual, auditory, and olfactory domain: A review. Frontiers in Neural Circuits, 2021, 15

[47] Patel D, Hazan H, Saunders D J, et al. Improved robustness of reinforcement learning policies upon conversion to spiking neuronal network platforms applied to Atari Breakout game. Neural Networks, 2019, 120: 108-115

[48] Sutton R S, Barto A G. Reinforcement learning: An introduction. Second edition. Cambridge, UK: MIT Press, 2018

[49] Potjans W, Diesmann M, Morrison A. An imperfect dopaminergic error signal can drive temporal-difference learning. PLoS Computational Biology, 2011, 7(5)

[50] Urbanczik R, Senn W. Reinforcement learning in populations of spiking neurons. Nature Neuroscience, 2009, 12(3): 250-252

[51] Frémaux N, Sprekeler H, Gerstner W. Reinforcement learning using a continuous time actor-critic framework with spiking neurons. PLoS Computational Biology, 2013, 9(4): e1003024

[52] Bellec G, Scherr F, Subramoney A, et al. A solution to the learning dilemma for recurrent networks of spiking neurons. Nature Communications, 2020, 11(1): 1-15

[53] Hebb D O. The organization of behavior: A neuropsychological theory. New York, USA: John Wiley and Sons, 1949: 62

[54] Klopf A H. Brain function and adaptive systems: A heterostatic theory. Air Force Cambridge Research Laboratories Special Report, 1972

[55] Watkins C J C H. Learning from delayed rewards [Ph.D. thesis]. University of Cambridge, Cambridge, UK, 1989

[56] Williams R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning, 1992, 8(3): 229-256

[57] Markram H, Lübke J, Frotscher M, et al. Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. Science, 1997, 275(5297): 213-215

[58] Bohte S M, La Poutre J A, Kok J N. Error-backpropagation in temporally encoded networks of spiking neurons. Neurocomputing, 2002, 48(1-4): 17-37

[59] Seung H S. Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. Neuron, 2003, 40(6): 1063-1073

[60] Izhikevich E M. Solving the distal reward problem through linkage of STDP and dopamine signaling. Cerebral Cortex, 2007, 17(10): 2443-2452

[61] Frémaux N, Sprekeler H, Gerstner W. Functional requirements for reward-modulated spike-timing-dependent plasticity. Journal of Neuroscience, 2010, 30(40): 13326-13337

[62] O'Brien M J, Srinivasa N. A spiking neural model for stable reinforcement of synapses based on multiple distal rewards. Neural Computation, 2013, 25(1): 123-156

[63] Tang G, Kumar N, Michmizos K P. Reinforcement co-learning of deep and spiking neural networks for energy-efficient mapless navigation with neuromorphic hardware//Proceedings of the International Conference on Intelligent Robots and Systems. Las Vegas, USA, 2020: 6090-6097

[64] Zhang L, Cao J, Zhang Y, et al. Distilling neuron spike with high temperature in reinforcement learning agents. arXiv preprint arXiv: 2108.10078, 2021

[65] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv: 1503.02531, 2015

[66] Liu G, Deng W, Xie X, et al. Human-level control through directly-trained deep spiking q-networks. IEEE Transactions on Cybernetics, 2022, to appear

[67] Taherkhani A, Belatreche A, Li Y, et al. A review of learning in biologically plausible spiking neural networks. Neural Networks, 2020, 122: 253-272

[68] Hu Yi-Fan, Li Guo-Qi, Wu Yu-Jie, et al. Spiking neural networks: A survey on recent advances and new directions. Journal of Control and Decision, 2021, 36(1): 1-26 (in Chinese)
(胡一凡, 李国齐, 吴郁杰等. 脉冲神经网络研究进展综述. 控制与决策, 2021, 36(1): 1-26)

[69] Frémaux N, Gerstner W. Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules. Frontiers in Neural Circuits, 2016, 9: 85

[70] Hodgkin A L, Huxley A F. A quantitative description of membrane current and its application to conduction and excitation in nerve. The Journal of Physiology, 1952, 117(4): 500-544

[71] N Brunel, Mark C. W. van Rossum. Lapicque's 1907 paper: From frogs to integrate-and-fire. Biological Cybernetics, 2007, 97(5-6): 337-339

[72] Koch C, Segev I. Methods in neuronal modeling: From ions to networks. Cambridge, UK: MIT Press, 1998

[73] Dayan P, Abbott L. Theoretical neuroscience: Computational and mathematical modeling of neural systems. Philosophical Psychology, 2001, 15(1): 154-155

[74] Brunel N, Latham P E. Firing rate of the noisy quadratic integrate-and-fire neuron. Neural Computation, 2003, 15(10): 2281-2306

[75] Brette R, Gerstner W. Adaptive exponential integrate-and-fire model as an effective description of neuronal activity. Journal of Neurophysiology, 2005, 94(5): 3637-3642

[76] Izhikevich E M. Which model to use for cortical spiking neurons? IEEE Transactions on Neural Networks, 2004, 15(5): 1063-1070

[77] Izhikevich E M. Simple model of spiking neurons. IEEE Transactions on Neural Networks, 2003, 14(6): 1569-1572

[78] Adrian E D, Zotterman Y. The impulses produced by sensory nerve-endings, Part II: The response of a single end-organ. The Journal of Physiology, 1926, 61(2): 151-171

[79] VanRullen R, Guyonneau R, Thorpe S J. Spike times make sense. Trends in Neurosciences, 2005, 28(1): 1-4

[80] Thorpe S, Gautrais J. Rank order coding. Computational

Neuroscience.Boston，USA：Springer，1998：113-118
[81] Hu J，Tang H，Tan K C，et al.A spike-timing-based integrated model for pattern recognition. Neural Computation，2014，25(2)：450-472
[82] Ponulak F，Kasinski A.Introduction to spiking neural networks：Information processing，learning and applications. Acta Neurobiologiae Experimentalis，2011，71(4)：409-433
[83] Izhikevich E M. Polychronization：Computation with spikes. Neural Computation，2014，18(2)：245-282
[84] Izhikevich E M，Desai N S，Walcott E C，et al.Bursts as a unit of neural information：Selective communication via resonance. Trends in Neurosciences，2003，26(3)：161-167
[85] Georgopoulos A P，Schwartz A B，Kettner R E. Neuronal population coding of movement direction. Science，1986，233(4771)：1416-1419
[86] Schultz W.Predictive reward signal of dopamine neurons.Journal of Neurophysiology，1998，80(1)：1-27
[87] Ranganath C，Rainer G. Neural mechanisms for detecting and remembering novel events. Nature Reviews Neuroscience，2003，4(3)：193-202
[88] Crow T J.Cortical synapses and reinforcement：A hypothesis. Nature，1968，219(5155)：736-737
[89] Barto A G.Learning by statistical cooperation of self-interested neuron-like computing elements.Human Neurobiology，1985，4(4)：229-256
[90] Xie X，Seung H S.Learning in neural networks by reinforcement of irregular spiking.Physical Review E，2004，69(4)：041909
[91] Gerstner W，Lehmann M，Liakoni V，et al. Eligibility traces and plasticity on behavioral time scales：Experimental support of neohebbian three-factor learning rules. Frontiers in Neural Circuits，2018，12：53
[92] Ponulak F，Kasiski A. Supervised learning in spiking neural networks with ReSuMe：Sequence learning，classification，and spike shifting.Neural Computation，2010，22(2)：467-510
[93] Widrow B，Hoff M E. Adaptive switching circuits. Stanford，USA：Electronics Laboratory，Stanford University，1960：7-22
[94] Mohemmed A，Schliebs S，Matsuda S，et al. Span：Spike pattern association neuron for learning spatio-temporal spike patterns.International Journal of Neural Systems，2012，22(4)：1250012
[95] Cao Y，Chen Y，Khosla D. Spiking deep convolutional neural networks for energy-efficient object recognition. International Journal of Computer Vision，2015，113(1)：54-66
[96] Diehl P U，Neil D，Binas J，et al. Fast-classifying，high-accuracy spiking deep networks through weight and threshold balancing//Proceedings of the International Joint Conference on Neural Networks.Killarney，Ireland，2015：1-8
[97] Rueckauer B，Lungu I A，Hu Y H，et al. Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. Frontiers in Neuroscience，2017，11：682
[98] Lee J H，Delbruck T，Pfeiffer M. Training deep spiking neural networks using backpropagation. Frontiers in Neuroscience，2016，10：508
[99] Neftci E O，Mostafa H，Zenke F.Surrogate gradient learning in spiking neural networks. IEEE Signal Processing Magazine，2019，36(6)：51-63
[100] Lee C，Sarwar S S，Panda P，et al. Enabling spike-based backpropagation for training deep neural network architectures. Frontiers in Neuroscience，2020，14：119
[101] Sengupta A，Ye Y，Wang R，et al. Going deeper in spiking neural networks：VGG and residual architectures. Frontiers in Neuroscience，2019，13：95
[102] Chrisman L. Reinforcement learning with perceptual aliasing：The perceptual distinctions approach//Proceedings of the AAAI Conference on Artificial Intelligence. San Jose，USA，1992：183-188
[103] Van Hasselt H，Guez A，Silver D.Deep reinforcement learning with double q-learning//Proceedings of the AAAI Conference on Artificial Intelligence.Phoenix，USA，2016：2094-2100
[104] Schaul T，Quan J，Antonoglou I，et al. Prioritized experience replay//Proceedings of the International Conference on Learning Representations.San Juan，Puerto Rico，2016：1-21
[105] Freitas N D，Lanctot M，Hasselt H V，et al. Dueling network architectures for deep reinforcement learning//Proceedings of the International Conference on Machine Learning. New York，USA，2016：1995-2003
[106] Sutton R S. Learning to predict by the methods of temporal differences.Machine Learning，1988，3(1)：9-44
[107] Bellemare M G，Dabney W，Munos R. A distributional perspective on reinforcement learning//Proceedings of the International Conference on Machine Learning.Sydney，Australia，2017：449-458
[108] Fortunato M，Azar M G，Piot B，et al. Noisy networks for exploration//Proceedings of the International Conference on Learning Representations.Vancouver，Canada，2018：1-21
[109] Hessel M，Modayil J，Hasselt H V，et al.Rainbow：Combining improvements in deep reinforcement learning//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans，USA，2018：3215-3222
[110] 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
[111] 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
[112] Schulman J，Levine S，Abbeel P，et al. Trust region policy optimization//Proceedings of the International Conference on Machine Learning.Lille，France，2015：1889-1897
[113] Schulman J，Wolski F，Dhariwal P，et al. Proximal policy optimization algorithms.arXiv preprint arXiv：1707.06347，2017
[114] Haarnoja T，Zhou A，Abbeel P，et al. Soft actor-critic：Off-policy maximum entropy deep reinforcement learning with a stochastic actor//Proceedings of the International Conference on Machine Learning.Stockholm，Sweden，2018：1856-1865
[115] Haarnoja T，Zhou A，Hartikainen K，et al. Soft actor-critic algorithms and applications. arXiv preprint arXiv：1812.05905，

2018

[116] Lillicrap T P, Hunt J J, Pritzel A, et al.Continuous control with deep reinforcement learning//Proceedings of the International Conference on Learning Representations. San Juan, Puerto Rico, 2016: 1-14

[117] Fujimoto S, Hoof H, Meger D. Addressing function approximation error in actor-critic methods//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018: 1582-1591

[118] Klopf A H. The hedonistic neuron: A theory of memory, learning, and intelligence. Washington, USA: Hemisphere Publishing, 1982

[119] Florian R V. A reinforcement learning algorithm for spiking neural networks//Proceedings of the International Symposium on Symbolic and Numeric Algorithms for Scientific Computing. Timisoara, Romania, 2005: 299-306

[120] Florian R V. Reinforcement learning through modulation of spike-timing-dependent synaptic plasticity. Neural Computation, 2007, 19(6): 1468-1502

[121] Baxter J, Bartlett P L. Direct gradient-based reinforcement learning//Proceedings of the International Symposium on Circuits and Systems.Geneva, Switzerland, 2000: 271-274

[122] Baxter J, Bartlett P L. Infinite-horizon policy-gradient estimation. Journal of Artificial Intelligence Research, 2001, 15: 319-350

[123] Baras D, Meir R L. Reinforcement learning, spike-time-dependent plasticity, and the BCM rule. Neural Computation, 2007, 19(8): 2245-2279

[124] Bienenstock E L, Cooper L N, Munro P W. Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. Journal of Neuroscience, 1982, 2(1): 32-48

[125] Kaiser J, Hoff M, Konle A, et al.Embodied synaptic plasticity with online reinforcement learning. Frontiers in Neurorobotics, 2019, 13: 81

[126] Lichtsteiner P, Posch C, Delbruck T.A 128×128 120 dB 15 μs latency asynchronous temporal contrast vision sensor. IEEE Journal of Solid-State Circuits, 2008, 43(2): 566-576

[127] Hull C L. Principles of behavior: An introduction to behavior theory.New York: Appleton-Century-Crofts, 1943

[128] Farries M A, Fairhall A L. Reinforcement learning with modulated spike timing-dependent synaptic plasticity. Journal of Neurophysiology, 2007, 98(6): 3648-3665

[129] Legenstein R, Pecevski D, Maass W. A learning theory for reward-modulated spike-timing-dependent plasticity with application to biofeedback. PLoS Computational Biology, 2008, 4(10): e1000180

[130] Mozafari M, Kheradpisheh S R, Masquelier T, et al. First-spike-based visual categorization using reward-modulated STDP. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(12): 6178-6190

[131] Vasilaki E, Frémaux N, Urbanczik R, et al. Spike-based reinforcement learning in continuous state and action space: When policy gradient methods fail. PLoS Computational Biology, 2009, 5(12): e1000586

[132] Potjans W, Morrison A, Diesmann M.A spiking neural network model of an actor-critic learning agent. Neural Computation, 2009, 21(2): 301-339

[133] Schultz W, Dayan P, Montague P R. A neural substrate of prediction and reward.Science, 1997, 275(5306): 1593-1599

[134] Doya K. Reinforcement learning in continuous time and space. Neural Computation, 2000, 12(1): 219-245

[135] Zheng N, Mazumder P. Hardware-friendly actor-critic reinforcement learning through modulation of spike-timing-dependent plasticity. IEEE Transactions on Computers, 2016, 66(2): 299-311

[136] Rasmussen D, Eliasmith C. A neural reinforcement learning model for tasks with unknown time delays//Proceedings of the Cognitive Science Society.Berlin, Germany, 2013: 35

[137] MacNeil D, Eliasmith C. Fine-tuning and the stability of recurrent neural networks.PloS One, 2011, 6(9): e22885

[138] Nichols E, McDaid L J, Siddique N.Biologically inspired SNN for robot control. IEEE Transactions on Cybernetics, 2012, 43(1): 115-128

[139] Maass W, Markram H. Synapses as dynamic memory buffers. Neural Networks, 2002, 15(2): 155-161

[140] Yuan M, Wu X, Yan R, et al.Reinforcement learning in spiking neural networks with stochastic and deterministic synapses. Neural Computation, 2019, 31(12): 2368-2389

[141] Bergstra J, Bardenet R, Bengio Y, et al.Algorithms for hyper-parameter optimization//Proceedings of the Advances in Neural Information Processing Systems.Granada, Spain, 2011: 2546-2554

[142] Jordan J, Schmidt M, Senn W, et al. Evolving to learn: Discovering interpretable plasticity rules for spiking networks. arXiv preprint arXiv:2005.14149, 2020

[143] Miller J F, Harding S L. Cartesian genetic programming//Proceedings of the Genetic and Evolutionary Computation Conference.Montreal, Canada, 2009: 3489-3512

[144] Lu J, Hagenaars J J, de Croon G C H E. Evolving-to-learn reinforcement learning tasks with spiking neural networks.arXiv preprint arXiv:2202.12322, 2022

[145] Bonyadi M R, Michalewicz Z. Particle swarm optimization for single objective continuous space problems: A review. Evolutionary Computation, 2017, 25(1): 1-54

[146] Wang L, Yoon K J.Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(6): 3048-3068

[147] Zhang D, Zhang T, Jia S, et al. Multiscale dynamic coding improved spiking actor network for reinforcement learning//Proceedings of the AAAI Conference on Artificial Intelligence. Online, 2022

[148] Forgy E W. Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. Biometrics, 1965, 21(3): 768-769

[149] Brockman G, Cheung V, Pettersson L, et al.Openai gym.arXiv preprint arXiv:1606.01540, 2016

[150] Badia A P, Piot B, Kapturowski S, et al. Agent57: Outperforming the atari human benchmark//Proceedings of the International Conference on Machine Learning. Online, 2020: 507-517

[151] Yurtsever E, Lambert J, Carballo A, et al. A survey of autonomous driving: Common practices and emerging technologies. IEEE access, 2020, 8: 58443-58469

[152] Li J, Dong S, Yu Z, et al.Event-based vision enhanced: A joint detection framework in autonomous driving//Proceedings of the IEEE International Conference on Multimedia and Expo. Shanghai, China, 2019: 1396-1401

[153] Vemprala S, Mian S, Kapoor A. Representation learning for event-based visuomotor policies//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2021: 4712-4724

[154] Ratliff F, Hartline H K, Lange D. The dynamics of lateral inhibition in the compound eye of Limulus. I. Studies on Excitation and Inhibition in the Retina: A Collection of Papers from the Laboratories of H.Keffer Hartline, 1974: 463

[155] Cheng X, Hao Y, Xu J, et al.LISNN: Improving spiking neural networks with lateral interactions for robust object recognition// Proceedings of the International Joint Conference on Neural Networks.Online, 2020: 1519-1525

[156] Dudman J T, Krakauer J W. The basal ganglia: From motor commands to the control of vigor. Current Opinion in Neurobiology, 2016, 37: 158-166

[157] Barbera G, Liang B, Zhang L, et al. Spatially compact neural clusters in the dorsal striatum encode locomotion relevant information.Neuron, 2016, 92(1): 202-213

[158] Zador A, Escola S, Richards B, et al. Catalyzing next-generation Artificial Intelligence through NeuroAI. Nature Communications, 2023, 14(1): 1597

[159] Huang S, Papernot N, Goodfellow I, et al.Adversarial attacks on neural network policies.arXiv preprint arXiv:1702.02284, 2017

CHEN Ding, Ph. D. candidate. His main research interests include spiking neural networks, reinforcement learning and brain-inspired computation.

HUANG Yang-Ru, Ph. D. candidate. Her research interests include brain-inspired reinforcement learning and computer vision.

PENG Pei-Xi, Ph. D., associate research fellow. His main research interests include computer vision and reinforcement learning.

HUANG Tie-Jun, Ph. D., professor, CCF Fellow. His current research interests include visual information processing and brain-inspired vision.

TIAN Yong-Hong, Ph. D., professor, CCF Fellow. His current research interests include visual information processing and brain-inspired vision.

Background

In recent years, artificial intelligence, represented by deep learning, has developed rapidly. However, compared with the human brain, the energy consumption of existing hardware systems to perform the same tasks is often at least one order of magnitude higher. Under the guidance of the human brain, the hardware systems that implement neuronal and synaptic computations through spike-driven communication is expected to solve the energy consumption problem faced by current deep learning algorithms. With the continuous development of hardware systems, related methods are also evolving. Especially, Spiking Neural Network (SNN) is a new generation of neural networks corresponding to hardware systems. Although deep learning has made breakthrough achievements in many fields, evidence shows that SNN can perform better in robotics, autonomous control and other fields. In these fields, deep learning algorithms often need a large number of computing resources that are difficult to meet when dealing with practical problems. With the help of neuromorphic hardware, SNN can greatly reduce energy consumption. Reinforcement learning, as an important branch of AI research, has proved its effectiveness in a wide range of control tasks. Therefore, by combining SNN with reinforcement learning, spiking reinforcement learning algorithm provides a low energy consumption solution for continuous control tasks. Using neuromorphic sensors, spiking reinforcement learning algorithms can make the agent perceive and make decisions like the human brain. In addition, in the field of computational neuroscience, a large number of researchers have used reinforcement learning algorithms to model the brain's reward learning mechanism, which also belongs to the research category of spiking reinforcement learning algorithms.

In this paper, we present the research content of spiking reinforcement learning algorithms, and systematically introduce the related work of spiking reinforcement learning algorithms. Then we provide a comprehensive overview of the

research progress of spiking reinforcement learning algorithms. After that, we highlight several problems and limitations of spiking reinforcement learning algorithms consisting of coding schemes, the separation between the spike-train and the sequence of reinforcement learning, the trade-off between biological plausibility and performance, simulation platform and open-source framework. Finally, we discuss some meaningful future research directions on spiking reinforcement learning algorithms. We hope that this work can pave the way for spiking reinforcement learning algorithms and more researchers can join this field.

This work is partially supported by grants from the Key-Area Research and Development Program of Guangdong Province under contact No. 2020B0101380001, and the National Natural Science Foundation of China under contract No. 62027804, No. 62372010, No. 61825101 and No. 62088102.