文章编号:1671-4598(2022)01-0258-08 DOI:10.16526/j.cnki.11—4762/tp.2022.01.040 中图分类号:TP181 文献标识码:A

基于一维卷积循环神经网络的深度强化学习算法

畅 鑫[1,2],李艳斌[1,2],田 淼[3],陈苏遂[3],杜宇峰[1,2],赵 研[1,2]
(1. 中国电子科技集团公司 第五十四研究所,石家庄 050081;
2. 河北省电磁频谱认知与管控重点实验室,石家庄 050081;
3. 电子科技大学 信息与通信工程学院,成都 611731)

摘要: 针对现有深度强化学习算法在状态空间维度大的环境中难以收敛的问题,提出了在时间维度上提取特征的基于一维卷积循环网络的强化学习算法;首先在深度Q网络(DQN,deep Q network)的基础上构建一个深度强化学习系统;然后在深度循环Q网络(DRQN,deep recurrent Q network)的神经网络结构基础上加入了一层一维卷积层,用于在长短时记忆(LSTM,long short-term memory)层之前提取时间维度上的特征;最后在与时序相关的环境下对该新型强化学习算法进行训练和测试;实验结果表明这一改动可以提高智能体的决策水平,并使得深度强化学习算法在非图像输入的时序相关环境中有更好的表现。

关键词: 强化学习;深度学习;长短时记忆网络;卷积神经网络;深度Q网络

Reinforcement Learning Algorithm Based on One-dimensional Convolutional Recurrent Network

CHANG Xin[1,2], LI Yanbin[1,2], TIAN Miao[3], CHEN Suyi[3], DU Yufeng[1,2], ZHAO Yan[1,2]
(1. The 54th Research Institute of China Electronics Technology Group Corporation (CETC54), Shijiazhuang 050081, China; 2. Hebei Key Laboratory of Electromagnetic Spectrum Cognition and Control, The 54th Research Institute of China Electronics Technology Group Corporation (CETC54), Shijiazhuang 050081, China; 3. School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China)

Abstract: Existing deep reinforcement learning algorithms have difficulty converging in environments with large state space dimensions. So a reinforcement learning algorithm based on one-dimensional convolutional recurrent networks that extracts features in the time dimension is proposed. Firstly, a deep reinforcement learning system based on DQN is built. Then a one-dimensional convolutional layer is added into the neural network architecture of DRQN for extracting the features in the time dimension before the LSTM layer. Finally, the new reinforcement learning algorithm is trained and tested in a timing-related environment. The experimental results show that this change can improve the decision-making level of the agent, making deep reinforcement learning algorithms have better performance in non-image input and timing-related environment

Keywords: reinforcement learning; deep learning; LSTM; convolutional neural network; DQN

0 引言

用数学方法寻找最优策略的研究既古老又新颖,最早可以追溯到20世纪50年代初,美国数学家贝尔曼(R. Bellman)等人在研究多阶段决策过程的优化问题时,提出了著名的最优化原理,从而创立了动态规划。然后随着时代发展,这个领域逐渐出现了蒙特卡罗法、时序差分法等优秀的算法,解决了许多动态规划所不能解决的问题。在传统强化学习时代,最为

收稿日期:2021-10-09; **修回日期:**2021-11-25。
基金项目:中国博士后科学基金(2021M693002)。
作者简介:畅 鑫(1990-),男,河北石家庄人,博士,工程师,主要从事智能博弈技术方向的研究。
引用格式:畅 鑫,李艳斌,田 淼,等.基于一维卷积循环神经网络的深度强化学习算法[J].计算机测量与控制,2022,30(1):258-265.

杰出和经典的就是 Q 学习（Q-learning）算法。Q-learning 采用表格记录状态－动作对价值，即 Q 值的方法探索最优策略，这也成为了后续深度强化学习算法中基于价值（value-based）分支的基石[1]。然而，在现实中的许多情况下，问题所包含的状态空间和动作空间都非常大，比如将一些连续状态离散化后形成的状态空间，这就使得借助表格存储 Q 值的方法难以为继。

幸运的是，随着计算机算力的飞速发展，在强化学习中引入深度学习来解决连续状态空间问题成为了可能。但人们很快就发现，使用神经网络这样的非线性函数逼近动作价值函数的强化学习算法都是不稳定甚至不收敛的。这就是所谓的“离线学习－函数逼近－自举检验”不可能三角（deadly triad issue），意思是强化学习无法同时使用这 3 种数学方法，否则将导致算法的不稳定甚至不收敛。造成这种情况的原因主要有 3 点：1）连续的状态之间的相关性；2）动作价值函数的微小变化可能导致策略的突变并显著地改变数据分布；3）动作价值函数与收敛目标之间的相关性。

2015 年，Mnih 及其同事提出的 DQN 通过采用经验回放（experience replay）和目标网络（target networks）技术解决了不稳定的问题，在 2 600 多个雅达利游戏上达到了人类玩家的水平，带来了深度强化学习的浪潮[2]。此后，对 DQN 的各种改进技术不断涌现。文献［3］提出了优先经验回放（prioritized experience replay），能让重要的经验被更频繁地利用，从而提升强化学习的效率。文献［4］于 2016 年提出的深度双 Q 网络（DDQN，double deep Q network），解决了过度估计的问题。同年，文献［5］向 DQN 加入了竞争结构（dueling architecture），提升了 DQN 的学习效率。这种带有竞争结构的 DQN 叫做竞争深度 Q 网络（Dueling DQN，dueling deep Q network）。除了上述提到的基于 DQN 的改进，深度强化学习领域还产生了更多的不同的技术路径[6-15]。

DQN 及其衍生的强化学习算法已经能算得上是非常强大的算法了，在许多领域，如简单的 2D 游戏的表现都超出常人。然而，这种优秀表现往往只停留在人为指定规则的环境中，如大多数棋牌和游戏等领域。DQN 在现实问题中仍然有着难以落地的问题。这是因为在过去的强化学习算法研究中，我们通常默认环境的状态我们是可以完全获取的。但是在现实世界中，我们显然没有棋牌和游戏中那样的上帝视角，我们对环境的状态的获取是通过观测（observation）得来的。而观测，或者说测量，必然会有信息误差甚至损失，从而使得无法通过观测获得完全的状态。这时，以马尔可夫决策过程为基本假设的 DQN 的性能自然就会受到较大的影响。

为了解决上述问题，文献［16］提出了 DRQN，在 DQN 的基础上将其第一个全连接层改为了相同大小的 LSTM 层，解决了现实环境部分观测的问题。为了解决强化学习与反馈神经网络参数更新之间的矛盾，Matthew Hausknecht 和 Peter Stone 又提出了序列自举更新和随机自举更新 2 种与之配套的参数更新方式。在部分观测的马尔科夫环境，DRQN 相比 DQN 有着明显的提升。

然而，深度强化学习在状态空间维度大的环境中仍然面临着难以收敛的问题。考虑到大多数环境中的状态在时间上都具有一定的相关性，若能让神经网络学会提取时间维度上的特征，则有可能改善强化学习在时间相关场景的学习效率。区别于以上研究，本文在 DRQN 的基础上展开研究，探究在时间维度上引入一维卷积对强化学习性能的影响，并设计了仿真实验与 DQN 的性能进行对比。

1　基于一维卷积循环网络的深度强化学习算法

1.1　深度强化学习基础

现实中许多决策问题都可以通过建模成由 5 个参数 (S,A,P,R,γ) 描述的马尔可夫决策过程（MDP，markov decision process）来进行研究[1,17]。这 5 个参数分别为状态空间 S、动作空间 A、状态转移概率函数 P、奖赏函数 R 和衰减因子 γ，在马尔可夫决策过程中的每一个时刻 t，智能体都会观察一个状态 $s_t \in S$ 然后选择一个动作 $a_t \in A$，这个过程将决定下一个时刻的状态 $s_t \sim P(s_t,a_t)$ 并收到一个奖赏 $r_t \sim R$。

1992 年由 Watkins 和 Dayan 提出的 Q-learning 通过在给定状态 s 下对动作 a 的长期回报进行预测来解决马尔可夫决策问题[1]。这样的长期动作回报叫做 Q 值。某个动作 a 的 Q 值越高，意味着在当前状态下选择该动作所获得的长期收益的期望越大。在 Q-learning 中，Q 值通过下式迭代更新：

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)] \tag{1}$$

Q-learning 伪代码。

输出:动作价值函数 Q

对所有状态 $s\in S$，$a\in A(s)$，随机初始化 Q，其中终止状态的动作价值为 0

对每个回合:

初始化状态 s

对回合中的每个时间步长:

使用基于 Q 的策略，如 ε－贪心算法，选择状态 s 对应的动作 a

执行动作 a，观察到 r，s

$Q(s,a)\leftarrow Q(s,a)+\alpha[r+\gamma\max\limits_{a'}Q(s',a')-Q(s,a)]$

$s\leftarrow s'$

直到状态 s 是终止状态

直到所有回合结束

许多现实问题的状态空间显然都是连续的，为了让强化学习在这些问题上得以运用，需要借助一种强大的函数拟合器代替 Q-learning 中的表格。而神经网络显然就是这样的强大函数拟合器。

2015 年，文献［2］提出的 DQN 使用经验回放和目标网络解决了神经网络在强化学习问题中不稳定的问题，开启了深度强化学习的时代。DQN 算法在对神经网络进行训练时，经验 (s_t,a_t,r_t,s_t) 会被存储在经验池（replay buffer）中，需要使用时再从经验池随机采样，这个过程就叫做经验回放。经验回放可以消除数据间相关性并平滑数据分布变化，为神经网络的收敛创造数学条件。在深度强化学习训练阶段实时更新的网络叫做在线网络（online networks）。而大多数时候都保持原有参数，仅周期性地从在线网络复制参数的网络叫做目标网络（target networks）。目标网络的引入能够有效地削弱动作价值函数 $Q(s,a)$ 与收敛目标 $r+\gamma\max\limits_{a'}Q(s',a')$ 之间的相关性，从而使收敛过程变得稳定。用于训练在线网络的代价函数为:

$$(r+\gamma\max_{a'}Q(s',a';\theta_i^-)-Q(s,a;\theta_i))^2 \qquad (2)$$

其中：θ_i 为在线网络在第 i 次迭代时的参数，θ_i^- 为目标网络在第 i 次迭代时的参数。θ_i^- 周期性地复制 θ_i。下面为 DQN 的伪代码。

输出:关于动作价值函数 Q 的神经网络初始化经验池 D

初始化在线动作价值网络 Q 的参数 θ 为随机数

初始化目标动作价值函数 $\hat{Q}$ 的参数 $\theta^-=\theta$

对每个回合:

初始化状态 s_1

对回合中的每个时间步长 t:

根据 ε－贪心算法 选择动作

$$a_t=\begin{cases}\text{随机动作} & \text{概率为 }\varepsilon\\ \arg\max\limits_a Q(s_t,a;\theta) & \text{其他}\end{cases}$$

执行动作 a_t，观测奖赏 r_t 和下一个状态 s_{t+1}

将经验 (s_t,a_t,r_t,s_{t+1}) 存入经验池 D

//经验回放

D 随机采样一批次的经验 (s_j,a_j,r_j,s_{j+1})

设 $y_j=\begin{cases}r_j,\text{回合在第 }j+1\text{ 步结束}\\ r_j+\gamma\max\limits_{a'}Q(s_{j+1},a';\theta^-),\text{其他}\end{cases}$

反向传播 $[y_j-Q(s_j,a_j;\theta)]^2$，并用梯度下降法更新 θ

//周期性更新目标网络

每过 C 步，将在线网络 Q 复制给目标网络 $\hat{Q}$，即，设 $\theta^-=\theta$

直到状态 s_t 是终止状态

直到所有回合结束

在现实的环境中，智能体往往很难获得完整的状态。换句话说，现实世界的环境通常不严格符合马尔可夫性[16]。部分可观测马尔可夫决策过程（POMDP，partially observable markov decision process）对观测与真实状态之间的联系进行了数学建模，因而能更好地描述现实环境的动态性[18]。POMDP 在 MDP 的基础上引入了观测空间 Ω 与条件观测概率函数 O，并将智能体对环境的一次感知定义为观测 $o\in\Omega$。观测与真实状态之间有着某种联系，这种联系通过概率描述，即 $o\sim O(s)$。如此，POMDP 就可以被 6 个参数 (S,A,P,R,Ω,O) 描述，分别表示状态空间、动作空间、状态转移概率函数、奖赏函数，以及相对于 MDP 新增加的观测空间 Ω 与条件观测概率函数 O。显然，当观测 o 与状态 s 一一对应时，POMDP 就变为了 MDP。2017 年 Matthew Hausknecht 和 Peter Stone 提出的 DRQN 对 DQN 的网络结构进行了修改，将其第一个全连接层改为了相同大小的 LSTM 层。

因为引入了记忆能力，使得神经网络能更好地对抗由于观测带来的信息不完整。DRQN 的神经网络结构如图 1 所示。

1.2 算法结构

本文在 DQN 的基础上构建了一个深度强化学习系统，如图 2 所示。

与大多数强化学习系统一样，从宏观层面上看，

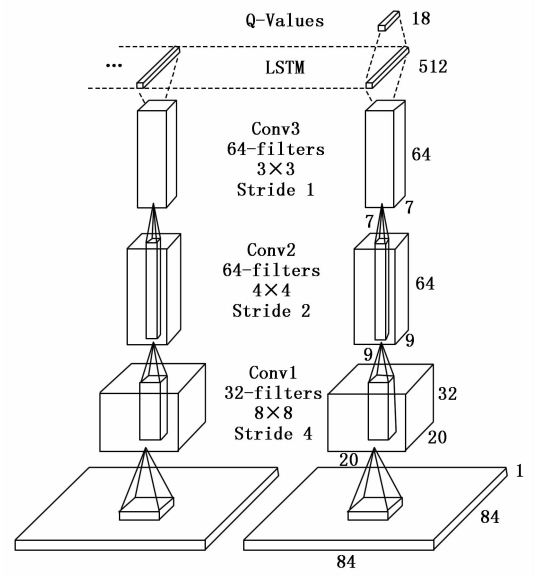

图 1　DRQN 结构示意图[16]

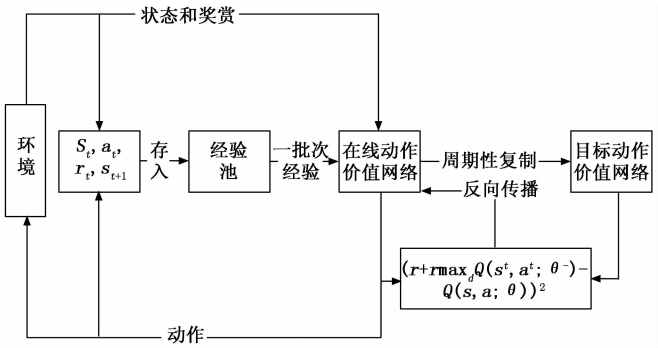

图 2　深度强化学习系统示意图

本文所构建的系统同样为环境与智能体进行交互的闭环系统。在每个步长里，智能体需要从环境获取当前步长的状态和奖赏，并选择一个动作反作用到环境中。

具体到内部结构，智能体主要由 4 个部分组成，分别为经验池、在线动作价值网络、目标动作价值网络和神经网络优化器。在每个步长里，经验池会将这一步长的状态、动作、奖赏以及下一步长的状态组合成一条经验储存起来，并随机选择一个批次的经验供神经网络训练使用；在线动作价值网络会根据当前步长的状态选择一个动作；神经网络优化器会计算代价函数，并将其计算结果反向传播给在线动作价值网络，优化神经网络的参数。在设定好的参数复制周期到来之时，目标动作价值网络会复制在线动作价值网络的参数并更新自身的参数。

1.3　伪代码

一维卷积循环网络的伪代码与 DQN 的伪代码形式基本一致，但因为包含了 LSTM 层，需要对经验回放部分进行修改，使其变为随机自举更新（bootstrapped random updates）[16]。下面为一维卷积循环网络的伪代码。

一维卷积循环网络伪代码。

输出：关于动作价值函数 Q 的神经网络初始化经验池 D

初始化在线动作价值网络 Q 的参数 θ 为随机数

初始化目标动作价值函数 $\hat{Q}$ 的参数 $\theta^- = \theta$

对每个回合：

初始化状态 s_1

对回合中的每个时间步长 t：

根据 ε－贪心算法 选择动作

$$a_t = \begin{cases} \text{随机动作}, & \text{概率为 } \varepsilon \\ \arg\max_a Q(s_t, a; \theta), & \text{其他} \end{cases}$$

执行动作 a_t，观测奖赏 r_t 和下一状态 s_{t+1}

将经验 (s_t, a_t, r_t, s_{t+1}) 存入经验池 D 中本回合的位置

//经验回放

随机选取一个序列长度 *seq_len*

从经验池 D 随机选取若干个回合的数据

从选取的回合数据中随机选取若干个时间点，并取出长度为 *seq_len* 的经验序列

$$\text{设 } y_j = \begin{cases} r_j, & \text{回合在第 } j+1 \text{ 步结束} \\ r_j + \gamma \max_{a'} Q(s_{j+1}, a'; \theta^-), & \text{其他} \end{cases}$$

反向传播 $[y_j - Q(s_j, a_j; \theta)]^2$，并用梯度下降法更新 θ

//周期性更新目标网络

每过 C 步，将在线网络 Q 复制给目标网络 $\hat{Q}$，即，设 $\theta^- = \theta$

直到状态 s_t 是终止状态

直到所有回合结束

2　一维卷积循环神经网络

为了在图像作为输入的 Atari 游戏环境上进行测试，DQN 与 DRQN 的神经网络都包含了二维卷积层。通常情况下，如果输入不为图像，而仅仅是特征向量，DQN 与 DRQN 所使用的神经网络将不会包含卷积层。然而，卷积层的特征提取能力不仅可以应用于提取图像特征，也可以应用于提取时间维度上的特征[19]。因此，本文探究了将卷积层的时间维度特征提取能力应用于深度强化学习的可能性。

图 2 系统中的在线动作价值网络与目标动作价值网络结构如图 3 所示，在 DRQN 所用神经网络的基础上加入了一维卷积层，称为一维卷积循环神经网

络。一维卷积层将在时间维度上对输入的数据进行卷积，并提取其在时间维度上的特征。实验表明这样做能提高神经网络的特征提取能力和拟合能力，从而提高智能体的决策水平，使得智能体在与时序相关的环境中有更好的表现。

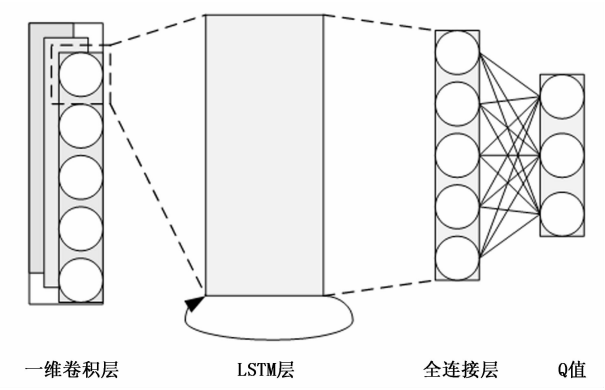

图 3 一维卷积循环神经网络示意图

2.1 一维卷积层

为了解决深度强化学习在状态空间维度大的环境中的快速收敛问题，本文用到了一维卷积层来提取状态在时间维度上的特征。设输入为 $\boldsymbol{X} \in R^{N\times C_{\text{in}}\times L_{\text{in}}}$，输出为 $\boldsymbol{Y} \in R^{N\times C_{\text{out}}\times L_{\text{out}}}$，则一维卷积层的数学表达式为：

$$\boldsymbol{Y}[i,j,:] = \boldsymbol{\beta}[j] + \sum_{k=0}^{C_{\text{in}}-1} \boldsymbol{\alpha}[j,k,:] \bigstar X[i,k,:] \tag{3}$$

式（3）中，符号 ★ 为互相关运算，N 为一个批次训练数据的大小，C_{in} 和 C_{out} 分别为输入和输出数据的通道数，L_{in} 和 L_{out} 分别为输入和输出数据的长度，$kernel\_size$ 表示一维卷积核大小。$\alpha \in R^{C_{\text{out}}\times C_{\text{in}}\times kernel\_size}$ 为该层的一维卷积核，$\beta \in R^{C_{\text{out}}}$ 为该层的偏置项。

2.2 LSTM 层

LSTM 层是一种循环神经网络，能给神经网络带来记忆能力。一般地，LSTM 层的输入为某一特征向量的时间序列 $\boldsymbol{x} \in R^{N\times L_{\text{in}}\times H_{\text{in}}}$。为简单起见，假设一个批次只包含 1 条数据且该特征向量只包含 1 个特征，即 $\boldsymbol{x} \in R^{L_{\text{in}}}$。由此可知 $\boldsymbol{x} = [x_1, x_2, \cdots, x_t, \cdots, x_{L_{\text{in}}}]^T$，则对于 x 中的任意一个时刻的元素 x_t，LSTM 层的数学表达式为：

$$\begin{cases} i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \\ f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}) \\ g_t = \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}) \\ o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \\ c_t = f_t \odot c_{t-1} + i_t \odot g_t \\ h_t = o_t \odot c_t \end{cases} \tag{4}$$

式（4）中，符号 ⊙ 表示哈达玛积，N 为一个批次训练数据的大小，L_{in} 为时间序列在时间维度上的长度，H_{in} 为时间数列包含的特征数。i_t、f_t、g_t 和 o_t 分别被称为 t 时刻的输入门（input gates）、遗忘门（forget gates）、元胞门（cell gates）和输出门（output gates）。c_t 和 h_t 分别被称为 t 时刻的元胞状态（cell states）和隐藏状态（hidden states）。

2.3 全连接层

全连接层是神经网络最经典的组成部件。按照经典的形式，设全连接层的输入为特征向量 $\boldsymbol{X} \in R^{N\times H_{\text{in}}}$，输出为 $\boldsymbol{Y} \in R^{N\times H_{\text{out}}}$，则全连接层的数学表达式为：

$$Y[i,:] = \sigma(X[i,:]A + b) \tag{5}$$

其中：σ 为某一非线性激活函数，常用的有 sigmoid 函数和 ReLU 函数等。N 为一个批次训练数据的大小，H_{in} 和 H_{out} 分别为输入和输出数据的特征数。$\boldsymbol{A} \in R^{H_{\text{in}}\times H_{\text{out}}}$ 为该层的权重，$\boldsymbol{b} \in R^{1\times H_{\text{out}}}$ 为该层的偏置项。

2.4 神经网络详细结构

具体地，以在 MountainCar-v0 环境中时为例。在训练阶段，深度强化学习训练器会在每个训练步长从经验池提取一个批次的经验用于训练神经网络，一个批次包含 512 条训练数据；每条训练数据皆为时间序列，序列长度在每个训练步长开始前随机选择；序列中每个时刻都包含小车当时的位置和速度信息。训练数据首先会被视为通道数为 2 的一维向量输入进一维卷积层，用于提取时间维度上的特征；然后被视为特征数为 2 的时间序列输入进 LSTM 层，增强对数据时间相关性的利用；最后将训练数据展开为一维向量输入到全连接层得到最终对每个动作价值的估计。为了加快收敛速度，在每一层后还加入了批归一化处理（batch normalization）。神经网络的详细结构如图 4 所示。

在测试阶段，神经网络的输入为由当前时刻小车的位置和速度组成的状态信息，为特征数为 2 的一维向量，输出每个动作价值的估计。

3 实验验证与分析

为了验证本文所提出的在时间维度上引入一维卷积层的有效性，设计仿真实验在 Open AI Gym 提供的 MountainCar-v0 环境下测试其性能，并在使用相同超参数的情况下与 DQN 的性能进行对比。

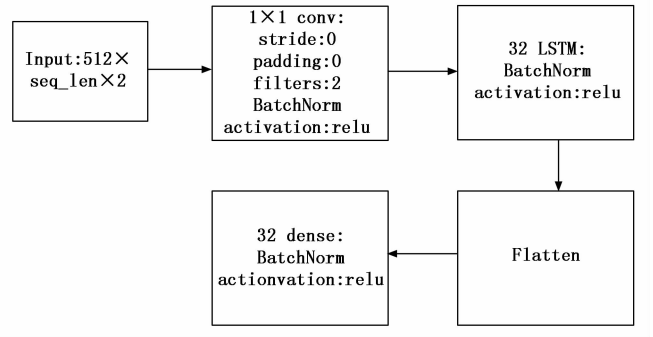

图 4　神经网络详细结构

在 MountainCar-v0 环境中，一辆小车处于两个山峰之间的一条一维轨道上，如图 5 所示。小车的目标是到达右边的山峰上，可是由于马力不足，小车必须学会积攒能量才能完成这一目标。

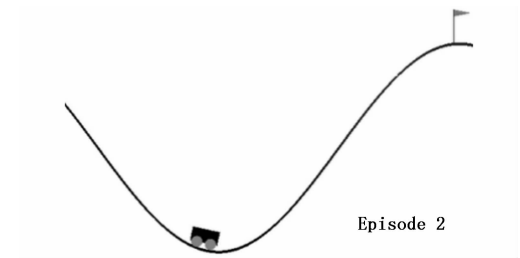

图 5　MountainCar-v0 环境示意图[20]

具体地，在 MountainCar-v0 环境中，神经网络输入的状态信息为小车的位置和速度，组成特征数为 2 的一维向量，输出的动作为小车的前进方向，共有向左、向右和空挡 3 种选择。

在测试中，学习率为 0.01，衰减因子为 0.9，探索度为 0.1；目标网络更新周期为 100，经验池大小为 4 096，一个批次包含 512 条训练数据，即 batch size=512，训练数据序列长度在 1～32 中随机选择。深度强化学习超参数总结如表 1 所示。

表 1　深度强化学习超参数表

超参数	取值
学习率	0.01
衰减因子	0.9
探索度	0.1
目标网络更新周期	100
经验池大小	4 096
批大小	512
训练数据序列长度	random(1～32)

下面首先给出一维卷积循环神经网络获取的总奖赏随训练轮次的变化曲线。在 MountainCar-v0 环境中，奖赏设定为当前时刻小车所具有的能量，即小车动能与势能之和。在具体代码实现中，设 p_t 和 v_t 为当前时刻小车的位置和速度，则奖赏 r_t 的定义如下：

$$r_t = abs(p_t + 0.6) + 10 \times abs(v_t) \tag{6}$$

DQN 与一维卷积循环神经网络在 MountainCar-v0 环境中获取总奖赏的表现如图 6 与图 7 所示。

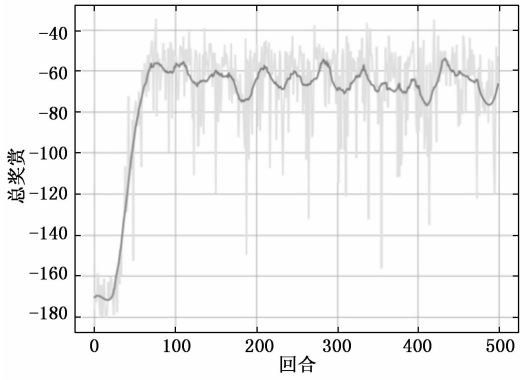

图 6　DQN 的总奖赏随训练轮次的变化

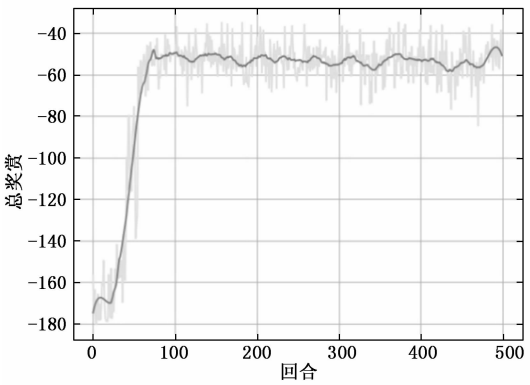

图 7　一维卷积循环神经网络的总奖赏随训练轮次的变化

图 6 与图 7 分别为 DQN 和一维卷积循环神经网络的总奖赏变化曲线。其中浅色部分表示原始数据，深色部分是平滑滤波后的结果。对比两者的总奖赏变化曲线，可以看出一维卷积循环神经网络相比 DQN 有着明显的提升。首先，在收敛过程中，一维卷积循环神经网络的总奖赏曲线斜率更大，上升速度更快，这说明一维卷积循环神经网络相比 DQN 有着更高的收敛效率；其次，更为突出的是，从最终达到的总奖赏来看，一维卷积循环神经网络学习到的策略所获取的总奖赏比 DQN 明显高出一部分，大约为 10 分。

图 8 与图 9 分别为 DQN 和一维卷积循环神经网络所作出动作选择的平均动作价值的变化曲线。其中

浅色部分表示原始数据，深色部分是平滑滤波后的结果。

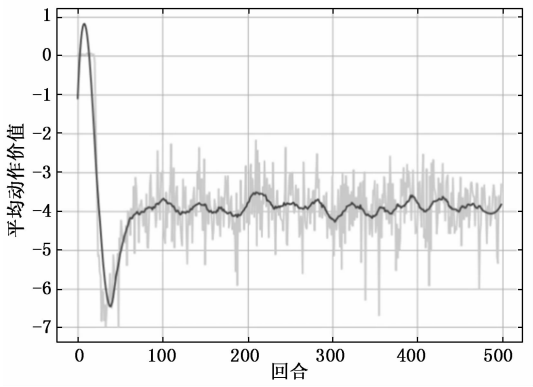

图 8 DQN 的平均动作价值随着训练轮次的变化

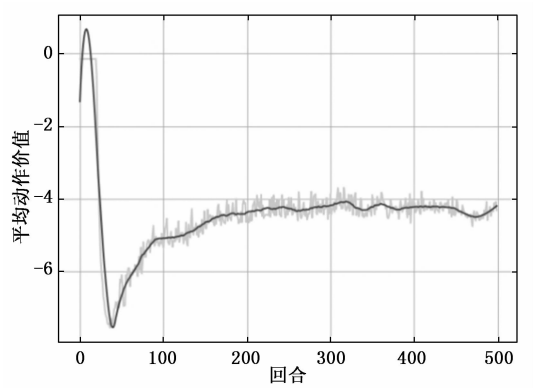

图 9 一维卷积循环神经网络的平均动作价值随着训练轮次的变化

可以看出在平均动作价值收敛的稳定性上，一维卷积循环神经网络相比 DQN 有着明显的提升。通过观察图 8 和图 9 中浅色部分的原始数据可以发现，DQN 的平均动作价值曲线波动较大，说明收敛过程不稳定；一维卷积循环神经网络的平均动作价值曲线波动较小，说明收敛过程相对稳定。

结合 DQN 和一维卷积循环神经网络的训练历史进行对比分析，不难发现一维卷积循环神经网络在最终结果还是收敛速度上都要优于 DQN。这是因为 LSTM 层赋予了一维卷积循环神经网络记忆性，使其可以利用更多的历史信息来辅助决策，并削弱 POMDP 的影响，从而让一维卷积循环神经网络在时间相关的环境中最终获得的总奖赏超过 DQN。同时，LSTM 层之前的一维卷积层在训练的过程中在时间维度上进行特征提取，使得整个一维卷积循环神经网络相比 DQN 有着更快的收敛速度以及稳定性。故相比于 DQN 简单的全连接结构，一维卷积循环神经网络在状态空间维度大且状态之间在时间上相关的环境中有着更好的表现。

4 结束语

在使用深度强化学习解决现实问题时，许多问题所构造的环境都存在着状态空间维度大且状态之间在时间上相关的特征。如果能够利用好状态在时间上的相关性就可以有效提升神经网络在大维度状态空间中的收敛效率。就本文所提出的一维卷积循环神经网络来说，LSTM 层的引入使得其拥有了一定的记忆能力，而一维卷积层的加入则让其在具备记忆能力的基础上有了更强的特征提取能力，进而可以更高效地处理时间维度上的信息。这使得改进后的算法能在 MountainCar-v0 这样与时序相关的环境中能够得到得到更高的总回报。同时，一维卷积层还增加了神经网络的拟合能力以及稳定性，使得深度强化学习的训练过程更加平稳。

参考文献：

[1] LI Y. Deep reinforcement learning: An overview [Z]. arXiv preprint arXiv: 1701.07274, 2018: 1-85.

[2] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning [J]. nature, 2015, 518 (7540): 529-533.

[3] SCHAUL T, QUAN J, ANTONOGLOU I, et al. Prioritized experience replay [C] // International Conference on Learning Representations. Puerto Rico, 2016.

[4] VAN HASSELT H, GUEZ A, SILVER D. Deep reinforcement learning with double q-learning [C] // Proceedings of the AAAI conference on artificial intelligence. 2016.

[5] WANG Z, SCHAUL T, HESSEL M, et al. Dueling network architectures for deep reinforcement learning [C] // International conference on machine learning. PMLR, 2016: 1995-2003.

[6] SILVER D, LEVER G, HEESS N, et al. Deterministic policy gradient algorithms [C] // International conference on machine learning. PMLR, 2014: 387-395.

[7] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning [Z]. arXiv preprint arXiv: 1509.02971, 2015: 1-14.

[8] SCHULMAN J, LEVINE S, ABBEEL P, et al. Trust

region policy optimization [C] // International conference on machine learning. PMLR, 2015: 1889 - 1897.
[9] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms [Z]. arXiv preprint arXiv: 1707. 06347, 2017: 1 - 12.
[10] WANG Z, BAPSt V, HEESS N, et al. Sample efficient actor-critic with experience replay [Z]. arXiv preprint arXiv: 1611. 01224, 2017: 1 - 20.
[11] MNIH V, BADIA A P, MIRZA M, et al. Asynchronous methods for deep reinforcement learning [C] // International conference on machine learning. PMLR, 2016: 1928 - 1937.
[12] SZITA I, LÖRINCZ A. Learning Tetris using the noisy cross - entropy method [J]. Neural computation, 2006, 18 (12): 2936 - 2941.
[13] MANIA H, GUY A, RECHT B. Simple random search provides a competitive approach to reinforcement learning [Z]. arXiv preprint arXiv: 1803. 07055, 2018: 1 - 22.
[14] SALIMANS T, HO J, CHEN X, et al. Evolution strategies as a scalable alternative to reinforcement learning [Z]. arXiv preprint arXiv: 1703. 03864, 2017: 1 - 13.
[15] HESSEL M, MODAYIL J, VAN HASSELT H, et al. Rainbow: Combining improvements in deep reinforcement learning [C] // Thirty-second AAAI conference on artificial intelligence. 2018.
[16] HAUSKNECHT M, STONE P. Deep recurrent q-learning for partially observable mdps [C] // 2015 aaai fall symposium series. 2015.
[17] SUTTON R S, BARTO A G. Reinforcement learning: An introduction [J]. Robotica, 1999, 17 (2): 229 - 235.
[18] SPAAN M T J. Partially observable Markov decision processes [M]. Reinforcement Learning. Springer, Berlin, Heidelberg, 2012: 387 - 414.
[19] ALBAWI S, MOHAMMED T A, Al-ZAWI S. Understanding of a convolutional neural network [C] // 2017 International Conference on Engineering and Technology (ICET). Ieee, 2017: 1 - 6.
[20] MOORE A. Efficient Memory-Based Learning for Robot Control [D]. University of Cambridge, 1990.

(上接第 233 页)

[7] 白英奎，孟宪江，丁　东，等. 利用神经网络提高偏最小二乘法的 NIR 多组分分析精度 [J]. 光谱学与光谱分析，2005 (3): 381 - 383.
[8] BIAGIONI D J, ASTLING D P, GRAF P, et al. Orthogonal projection to latent structures solution properties for chemometrics and systems biology data [J]. Journal of Chemometrics, 2011, 25 (9): 514 - 525.
[9] 李俊南. 基于正交偏最小二乘方法的代谢组学数据分析研究 [D]. 哈尔滨：哈尔滨医科大学，2014.
[10] 李俊南，侯　艳，李　康. 核正交偏最小二乘在代谢组学数据分析中的应用 [J]. 中国卫生统计，2015, 32 (1): 14 - 17.
[11] RANTALAINEN M, MAX BYLESJÖ, CLOAREC O, et al. Kernel－based orthogonal projections to latent structures (K - OPLS) [J]. Journal of Chemometrics, 2007, 21: 376 - 385.
[12] MULLER K, MIKA S, RATSCH G, et al. An introduction to kernel - based learning algorithms [J]. IEEE Transactions on Neural Networks, 2001, 12 (2): 181 - 201.
[13] BYLESJ M, RANTALAINEN M, NICHOLSON J K, et al. K - OPLS package: Kernel - based orthogonal projections to latent structures for prediction and interpretation in feature space [J]. Bmc Bioinformatics, 2008, 9 (1): 106 - 112.
[14] BLEKHERMAN G, LAUBENBACHER R, CORTES D F, et al. Bioinformatics tools for cancer metabolomics [J]. Metabolomics, 2011, 7 (3): 329 - 343.
[15] 刘根兰，倪永年. 荧光光谱法结合多元曲线分辨-交替最小二乘法研究伞形花内酯与牛血清白蛋白的相互作用 [J]. 高等学校化学学报，2008 (7): 1339 - 1343.
[16] ZHANG Y, MA C. Fault diagnosis of nonlinear processes using multiscale KPCA and multiscale KPLS [J]. Chemical Engineering Science, 2011, 66 (1): 64 - 72.
[17] 高恒振，万建伟，粘永健，等. 组合核函数支持向量机高光谱图像融合分类 [J]. 光学精密工程，2011, 19 (4): 878 - 883.
[18] 白　云，范川鹏，李素燕，等. 紫外分光光度法测定血液中的百草枯 [J]. 国际检验医学杂志，2013, 34 (11): 1421 - 1422.