基于双深度网络的安全深度强化学习方法

朱 斐[1),2),3),4)] 吴 文[1)] 伏玉琛[1),5)] 刘 全[1),2),3)]

[1)](苏州大学计算机科学与技术学院 江苏 苏州 215006)

[2)](软件新技术与产业化协同创新中心 南京 210000)

[3)](吉林大学符号计算与知识工程教育部重点实验室 长春 130012)

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

[5)](常熟理工学院计算机科学与工程学院 江苏 常熟 215500)

摘 要 深度强化学习利用深度学习感知环境信息,使用强化学习求解最优决策,是当前人工智能领域的主要研究热点之一.然而,大部分深度强化学习的工作未考虑安全问题,有些方法甚至特意加入带随机性质的探索来扩展采样的覆盖面,以期望获得更好的近似最优解.可是,不受安全控制的探索性学习很可能会带来重大风险.针对上述问题,提出了一种基于双深度网络的安全深度强化学习(Dual Deep Network Based Secure Deep Reinforcement Learning,DDN-SDRL)方法.DDN-SDRL 方法设计了危险样本经验池和安全样本经验池,其中危险样本经验池用于记录探索失败时的临界状态和危险状态的样本,而安全样本经验池用于记录剔除了临界状态和危险状态的样本.DDN-SDRL 方法在原始网络模型上增加了一个深度 Q 网络来训练危险样本,将高维输入编码为抽象表示后再解码为特征;同时提出了惩罚项描述临界状态,并使用原始网络目标函数和惩罚项计算目标函数.DDN-SDRL 方法以危险样本经验池中的样本为输入,使用深度 Q 网络训练得到惩罚项.由于 DDN-SDRL 方法利用了临界状态、危险状态及安全状态信息,因此 Agent 可以通过避开危险状态的样本、优先选取安全状态的样本来提高安全性.DDN-SDRL 方法具有通用性,能与多种深度网络模型结合.实验验证了方法的有效性.

关键词 强化学习;深度强化学习;深度 Q 网络;安全深度强化学习;安全人工智能;经验回放

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2019.01812

A Dual Deep Network Based Secure Deep Reinforcement Learning Method

ZHU Fei[1),2),3),4)] WU Wen[1)] FU Yu-Chen[1),5)] LIU Quan[1),2),3)]

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

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

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

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

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

Abstract Reinforcement learning is a widely studied class of machine learning method, where the agent of reinforcement learning keeps continuously interacting with the environment with the goal of getting maximal long term return. Reinforcement learning is particularly prominent in areas such as control and optimal scheduling. Deep reinforcement learning, which is able to take large-scale high-dimensional data, e.g. video and image, as original input data, takes advantage of deep learning methods to extract abstract representations of them, and then utilizes reinforcement

收稿日期:2018-05-22;在线出版日期:2019-03-19.本课题得到国家自然科学基金项目(61303108,61373094,61772355)、江苏省高校自然科学研究项目重大项目(17KJA520004)、符号计算与知识工程教育部重点实验室(吉林大学)项目(93K172014K04)、苏州市重点产业技术创新-前瞻性应用研究项目(SYG201804)、高校省级重点实验室(苏州大学)项目(KJS1524)、中国国家留学基金(201606920013)资助.

朱 斐,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为强化学习、深度强化学习和文本挖掘.E-mail:zhufei@suda.edu.cn.**吴 文**,硕士研究生,主要研究方向为深度强化学习.**伏玉琛**(通信作者),博士,教授,中国计算机学会(CCF)会员,主要研究领域为强化学习、智能信息处理.E-mail:yuchenfu@cslg.edu.cn.**刘 全**,博士,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为智能信息处理、自动推理和机器学习.

learning methods to attain optimal strategies, has recently become a research hotspot in artificial intelligence. There has emerged a large amount of work on deep reinforcement learning. For example, deep Q network (DQN), one of the most famous models in deep reinforcement learning, is based on convolutional neural networks (CNNs) and Q-learning algorithm, directly uses the unprocessed image as the input. DQN has been applied to learn strategy in complex environments with high-dimensional input. However, few deep reinforcement learning algorithms considers how to ensure security during the process of learning in the unknown environment. Even more, many reinforcement learning algorithms intentionally add random exploration approaches, e.g. ε-greedy, to guarantee the diversity of data sampling so that the algorithm could obtain a better approximate optimal solution. Nevertheless, exploration without any security constraint is very dangerous and likely to bring with high risk of leading to disastrous results. Aiming at solving this problem, an algorithm, named dual deep network based secure deep reinforcement learning (DDN-SDRL), is proposed. The DDN-SDRL algorithm sets up two experience pools. The first one is the experience pool of dangerous samples, including critical states and dangerous states that caused failure; and the second one is the experience pool of the secure sample, which excluded critical states and dangerous states. The DDN-SDRL algorithm takes advantage of an additional deep Q network to train dangerous samples and reconstructs a new objective function by introducing a penalty component. The new objective function is calculated by the penalty component and the original network objective function. The penalty component, which is trained by a deep Q network with samples in the critical state experience pool, is used to represent critical states before failure. As the DDN-SDRL algorithm fully uses information of critical state, dangerous state and secure state, the agent is able to improve security by avoiding most dangerous states during the training process. The DDN-SDRL is a general mechanism of enhancing security during the learning and can be combined with a variety of deep network models, such as DQN, dueling deep Q network (DuDQN), and deep recurrent Q network (DRQN). In the simulated experiments, DQN, DuDQN and DRQN were used as original deep network respectively, and at the same time DDN-SDRL was applied to ensure security. The results of six testing Atari 2600 games, CrazyClimber, Kangaroo, KungFuMaster, Pooyan, RoadRunner and Zaxxon, indicate that the proposed DDN-SDRL algorithm makes control safer, more stable and more effective. It can be concluded that the characteristics of the environment suitable for DDN-SDRL include: (1) there are many representable dangerous states that lead to failure in the environment; (2) the difference between dangerous states and secure states is discriminative; (3) there are not too many actions and the agent can attain improvement by self-training. In these cases, the DDN-SDRL improves original deep network much better.

Keywords　reinforcement learning; deep reinforcement learning; deep Q-network; safe reinforcement learning; safe artificial intelligence; experience replay

1　引　言

强化学习(Reinforcement Learning, RL)是一种受到动物心理学启发,并结合心理学、控制理论等相关学科而发展形成的机器学习方法[1-2],其通过智能体(Agent)的不断“试错式”学习,寻求累积奖赏最大的策略[3].强化学习获得了研究人员和产业界的关注,在优化、控制、博弈、生物医学等领域积累了大量的研究成果,也有不少的实际应用[4-6].深度学习(Deep Learning, DL)是机器学习领域一类重要的方法,其建立神经网络,模仿人脑的机制来解释数据[7].深度学习能从原始的高维数据中提取可区分的特征.近年来,深度学习已经在语音识别、图像处

理等领域涌现出众多良好的研究和应用[8-9].

随着机器视觉设备的推广,在很多系统中,原始的状态信息以视频、图像等方式呈现.因此,如何充分利用深度学习方法,以视频、图像等数据作为原始输入,使用深度学习来提取大规模数据的抽象表征(Abstract Representation),再利用强化学习方法学习获得最优策略,从而实现"类脑"控制,是当前重要的研究方向之一.谷歌人工智能团队 DeepMind 较早地将深度学习与强化学习相结合,形成了深度强化学习(Deep Reinforcement Learning,DRL)这一研究方向.深度强化学习包括了强化学习和深度学习两个部分,其中深度学习部分具有感知环境信息的功能;强化学习部分通过决策完成从状态到动作的映射,并获得奖赏.深度强化学习累积了一些研究工作基础.如由强化学习中的 Q 学习(Q-Learning)方法[10] 和卷积神经网络(Convolutional Neural Networks,CNNs)结合而成的深度 Q 网络(Deep Q-Network,DQN)[11-12] 是深度强化学习领域的一个重要方法,已被成功地用于高维度输入环境中的策略求解.Van Hasselt 等人[13] 在双 Q 学习(Double Q-Learning)[14] 方法的基础上提出了深度双 Q 网络(Double Deep Q-Network,DDQN),该方法在计算目标网络的 Q 值时使用两套不同的参数,有效地避免了 DQN 过高估计动作值的问题.Hausknecht 等人[15] 首次将长短时间记忆单元(Long-Short Term Memory,LSTM)引入 DQN 中,提出了一种带有 LSTM 单元的深度循环 Q 网络(Deep Recurrent Q-Network,DRQN),该网络对情节具有一定的记忆功能,在大多数 Atari 2600 游戏实验环境中取得了较佳的成绩.Wang 等人[16] 提出了竞争深度 Q 网络(Dueling Deep Q-Network,DuDQN),利用 CNN 从原始输入中提取特征,分为优势函数(Advantage Function)和与动作无关的状态值函数两个通道来生成动作值函数,该方法在不改变底层强化学习方法的情况下进一步提高了 DQN 在 Atari 2600 环境下的效果.

在深度强化学习方法被成功应用于谷歌 AlphaGo[17] 并在围棋比赛中战胜世界冠军李世乭后,人们开始关注深度强化学习如何从研究迈向实际应用.然而,要完成这一阶段性的跨越还有很多工作需要完成.其中,保证决策的安全性是最重要的一项内容.有专家指出,人工智能的安全风险可能会带来严重的问题.而安全的人工智能必须保证所控制物体的安全性,降低导致危险的操作概率.2018 年以来,在美国发生了一系列无人驾驶车辆安全事故,更凸显出安全在自动控制任务中的重要性.但是,目前很多人工智能的方法没有充分地做到控制风险、增加安全性,甚至有些方法在求解过程中特意加入了带有随机性质的探索性学习.而不受安全限制的探索性学习很可能会带来重大风险.如果在现实世界的任务中直接应用强化学习的方法,让 Agent 进行"试错式"探索学习,所作出的决策就有可能使系统陷入危险状态.强化学习的安全性问题引起了越来越多的关注.有些研究人员开展了安全强化学习的研究工作.Berkenkamp 等人[18] 将控制理论中经典的李雅普诺夫函数(Lyapunov Function)引入强化学习,有效地缓解了 Agent 无限制探索带来的副作用.Garcia 等人[19] 提出了基于策略改进的安全强化学习方法,对动作和状态空间进行探索限制.然而,这两种方法的表现难以令人满意.

目前,安全深度强化学习是一个较新的研究内容.传统深度强化学习中,经典的 DQN 采用函数逼近的方法进行训练,为了去除训练过程中样本的相关性,DQN 使用经验重放(Experience Replay)的机制,以等概率的方式抽取样本更新网络.然而无差别地从经验池中采样并不能体现样本的区分度,很难辨别"好"的样本和"差"的样本,很有可能出现导致 Agent 陷入危险状态的样本再次被用于网络训练的情况.绝大部分深度强化学习方法没有采取预防机制来避免 Agent 陷入危险状态.因此,必须限制 Agent 无约束的探索以保障安全性.

针对上述问题,本文提出基于双深度网络的安全深度强化学习(Dual Deep Network Based Secure Deep Reinforcement Learning,DDN-SDRL)方法.DDN-SDRL 方法通过分离历史经验训练导致 Agent 陷入危险状态的经验样本,使其重新组成新的经验池,有效地甄别了导致 Agent 陷入危险状态的样本经验;在原有网络的基础上再增加一个深度网络,采用新增的深度网络对新经验池样本进行训练,充分利用了提取出的样本;将训练结果作为惩罚项改进目标函数.DDN-SDRL 方法具有较好的通用性,能与 DQN、DuDQN、DRQN 等网络相结合,并在 Atari 2600 游戏环境中表现良好.

2 背景知识

2.1 强化学习

在强化学习中,Agent 通过与环境交互以最大

化期望奖赏，是一种从环境状态映射到动作的过程. 强化学习问题可以用马尔可夫决策过程(Markov Decision Process，MDP)进行建模，使用四元组$\langle X,U,P,R\rangle$描述模型，其中：

X是状态集合，$x_t\in X$表示智能体在t时刻所处状态；

U是Agent可选动作集合，$u_t\in U$表示Agent在t时刻所采取的动作；

P是状态转移函数，表示在t时刻Agent采取动作u_t转至下一状态x_{t+1}的概率，通常形式化表示为$P(x_{t+1}\mid x_t,u_t)$；

R是奖赏函数，表示Agent在t时刻的状态x_t执行动作u_t转移到下一状态x_{t+1}后环境给出的立即奖赏r_{t+1}，通常形式化表示为$r_{t+1}=R(x_t,u_t,x_{t+1})$.

强化学习的累积奖赏是智能体从t时刻到T时刻的累积折扣奖赏之和

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

其中，γ为折扣因子，控制着未来奖赏的影响力.

强化学习的策略$h:X\to U$是从状态到动作的映射，即$u_t\sim h(x_t)$. 强化学习使用状态动作值函数来评估策略. 状态动作值函数$Q_h(x_t,u_t)$是在策略h下在当前状态x_t下执行动作u_t获得的期望奖赏，表示为

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

其中，E表示期望.

贝尔曼方程在传统强化学习中有着重要作用，状态动作值函数$Q_h(x_t,u_t)$遵循贝尔曼方程. 根据贝尔曼方程获得第$t+1$时刻的状态动作值$Q_{t+1}(x,u)$为

$$Q_{t+1}(x,u)=E_{x'\sim X}[r+\gamma\max_{u'}Q_t(x',u')\mid x,u] \tag{3}$$

当$t\to\infty$时，状态动作值函数会趋向最优. 即通过不断迭代，状态动作值函数最终会收敛.

Q学习是强化学习中被广泛应用的方法之一. 在Q学习中，Q值的计算方式为

$$\delta=r_{t+1}+\gamma\max_{u}Q(x_{t+1},u)-Q(x_t,u_t) \tag{4}$$

$$Q(x_t,u_t)=Q(x_t,u_t)+\alpha\delta \tag{5}$$

其中，γ为折扣因子，α为学习率，δ为TD误差. 通过式(4)、式(5)不断迭代更新Q值函数，从而收敛到最优状态动作值.

在实际应用中，通常状态空间和动作空间较为庞大，采用迭代算法寻找最优策略所产生的计算量较大，可行性较低. 因此，需要将大规模状态空间进行泛化处理，通常采用函数逼近的泛化方法近似得到Q值[20]. 例如，采用了资格迹的泛化方法为

$$\delta=r_{t+1}+\gamma\max_{u}Q(x_{t+1},u;\boldsymbol{\theta}_t)-Q(x_t,u_t;\boldsymbol{\theta}_t) \tag{6}$$

$$\boldsymbol{\theta}_{t+1}=\boldsymbol{\theta}_t+\alpha\delta_t\boldsymbol{e}_t \tag{7}$$

$$\boldsymbol{e}_t=\gamma\lambda\boldsymbol{e}_{t-1}+\nabla Q(x_t,u_t;\boldsymbol{\theta}_t) \tag{8}$$

其中，$\boldsymbol{e}$为资格迹.

2.2　深度Q网络(DQN)

DQN结合了传统强化学习的Q学习方法和卷积神经网络，采用经验回放机制和目标网络技术缓解了使用非线性神经网络作为函数逼近器出现的学习不稳定问题. DQN被成功应用于Atari 2600游戏环境中，其中部分游戏的成绩超过了人类玩家水平. DQN结构图如图1所示.

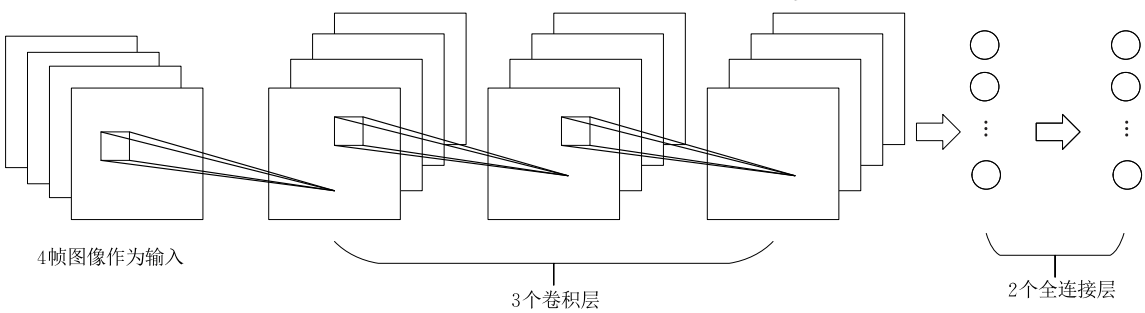

图1　深度Q网络(DQN)结构示意图

DQN的输入为4帧经过处理的图像，大小为84×84像素，之后由3个卷积层提取特征，每层卷积操作后利用非线性激活函数ReLU加入非线性因素，最后经过2个全连接层处理得到输出.

在深度学习中，训练数据需要满足相互独立的条件，而另一方面，强化学习的样本通过与环境交互产生，很多样本之间存在着较大的相关性. DQN利用经验回放机制打破样本之间相关性，从而使学习稳定，Agent在每个时间步t将状态(x_t)、动作(u_t)、奖赏(r_t)和执行动作后到达的下一个状态(x_{t+1})作为一个序列$\boldsymbol{e}_t=(x_t,u_t,r_t,x_{t+1})$保存在经验池$D=\{e_1,e_2,\cdots,e_N\}$中，并在训练阶段从中取出固定数量样本作为网络输入.

DQN使用两个独立的Q网络，即当前值网络

和目标值网络，分别用 $\boldsymbol{\theta}$ 和 $\boldsymbol{\theta}^-$ 表示两种网络的参数. 其中，$\boldsymbol{\theta}$ 是实时更新的，$\boldsymbol{\theta}^-$ 在每过 L 步后由 $\boldsymbol{\theta}$ 复制得到. 当前值网络和目标值网络的输出分别表示为 $Q(x,u\mid\boldsymbol{\theta})$ 和 $Q(x,u\mid\boldsymbol{\theta}^-)$. 损失函数由目标值函数和当前值函数的均方误差决定

$$L(\boldsymbol{\theta})=E[(Y-Q(x,u;\boldsymbol{\theta}))^2] \tag{9}$$

$$Y=r+\gamma\max_{u'}Q(x',u';\boldsymbol{\theta}^-) \tag{10}$$

其中，式(10)为监督学习的标签，用于近似表示值函数的优化目标. 为了求解最小化损失函数，在式(9)中对参数 $\boldsymbol{\theta}$ 求导

$$\nabla_{\boldsymbol{\theta}}L(\boldsymbol{\theta})=(r+\gamma\max_{u'}Q(x',u';\boldsymbol{\theta}^-)-Q(x,u;\boldsymbol{\theta}))\nabla_{\boldsymbol{\theta}}Q(x,u;\boldsymbol{\theta}) \tag{11}$$

可以采用随机梯度下降(Stochastic Gradient Descent，SGD)方法更新式(11)中的参数 $\boldsymbol{\theta}$.

DQN 通过贪心策略更新动作值函数，用 ε-贪心策略(ε-greedy)方法选择动作，即以 $1-\varepsilon$ 的概率选择贪心动作，以 ε 的概率随机选择动作.

2.3 安全强化学习

近来，有研究人员开展了安全强化学习的相关研究工作. Garcıa 等人[21]从安全的角度将 Agent 在训练过程中遇到的状态进行分类并系统地阐述了强化学习中的安全性概念，认为现有安全强化学习方法主要包括改变目标函数和改进探索两类方法.

第一类方法通过改变目标函数对 Agent 的动作进行限制. 强化学习的目标是最大化 Agent 在情节中的期望奖赏，目标函数通常只是最大化奖赏

$$Q(x,u)=r+\gamma\max_{u'}Q(x',u') \tag{12}$$

然而，式(12)忽视了危险状态对 Agent 的损害. 因此，第一类安全强化学习方法在目标函数中加入最小化 Agent 陷入危险状态的风险项

$$Q(x,u)=\min(Q(x,u),r+\gamma\max_{u'}Q(x',u')) \tag{13}$$

有人以改善奖赏方差的方式改进第一类安全强化学习方法. Chow 等人[22]将风险敏感性条件引入 MDP 模型，建立了风险控制和决策控制的统一框架. Hans 等人[23]提出一种基于级别的探索方案，一定程度上提高了 Agent 探索过程的安全性.

第二类方法通过外部环境信息改进探索过程. 由于大多数强化学习方法在开始学习阶段有外部知识，因此通常采用 ε-贪心策略作为探索策略获得状态动作空间先验知识. 然而，这种方法将耗费大量的探索时间和资源，ε-贪心策略的无限制探索也有可能使 Agent 陷入危险. 这在大规模连续空间任务中显得尤为突出. 针对这一问题，Song 等人[24]通过初始化适当的 Q 值改善算法效率. Pradyot 等人[25]利用 Agent 探索产生的经验改进探索算法，在格子世界等问题中取得了良好的效果.

在安全深度强化学习方面也有一些研究. Agarwal 等人[26]发现，随着样本的增加，经验也在同时积累，有效地利用这些经验可以控制决策的风险、提高决策的安全性从而降低错误决策的代价. 在多智能体的安全深度强化学习方面，Mhamdi 等人[27]提出了一种动态安全中断机制，完成了在离散状态空间下安全中断的任务. 然而，该方法在中断后缺乏学习能力，无法通过自主学习进行改进.

3 基于双深度网络的安全深度强化学习

本文提出一种基于双深度网络的安全深度强化学习方法. 整体模型由两个网络构成，即在原始网络模型(如 DQN、DRQN、DuDQN 等)中增加一个深度网络，用以训练临界于危险状态的临界样本和处于危险状态的危险样本，并采用双经验池方法，增加临界样本和危险样本经验池，充分利用其经验. 该模型将高维输入编码为抽象表示然后再解码为特征. 本节以实验环境 Atari 2600 为例分析各个网络功能及网络间关系.

3.1 安全深度强化学习

为了将安全强化学习概念引入深度强化学习，定义相关名词.

定义 1. 当 Agent 进入某个状态导致任务失败时，将该状态定义为危险状态，记为 x_d.

定义 2. 在危险状态 x_d 的前 Δ 个状态范围定义为安全距离，记为 d.

定义 3. 在安全距离 d 范围之内的状态定义为临界状态.

在本文的实验中，安全距离定义为 m，临界状态是 Agent 在游戏失败的前 m 帧状态，危险状态是 Agent 在游戏失败时的状态.

3.2 图像预处理

本文实验环境 Atari 2600 游戏产生 210×160 像素的 RGB 图像. 考虑到灰度图像和 RGB 图像作为输入对网络训练结果影响较小，而使用灰度图像

作为输入的计算量将会大幅减少，因此，本文将原始RGB图像处理为210×160像素的灰度图像；然后，使用降采样方法将灰度图像缩减为110×84×1的缩略图；最后，通过图像边界裁剪形成84×84×1的缩略图. 经过预处理操作，在有效信息不丢失的情况下减少了计算和存储量. 原始网络的输入是当前时刻最近的4帧经过预处理的图像；另一个深度网络的输入是处于临界状态和危险状态时经过预处理后的图像.

3.3　模型结构及分析

基于双深度网络的安全深度强化学习方法采用双网络模型架构，其中，增加的深度网络模型用于训练Agent历史遭遇的临界状态和危险状态. 训练结果用于惩罚Agent的动作值函数. 新加入的DQN对临界状态进行有针对性的训练，从而使得临界状态样本得到充分利用，减少Agent再次陷入危险状态的次数.

使用双网络模型有两方面优点：(1) 双网络模型充分利用了历史数据，使Agent有效地避免再次陷入危险状态，增加了安全性，加快了训练速度；(2) 通过额外训练临界状态与危险状态并改进动作值函数，使得目标函数避免朝危险状态方向收敛. 目标函数的改进提升了Agent探索的安全性. 基于双深度网络的安全深度强化学习方法如算法1所示.

算法1.　基于双深度网络的安全深度强化学习(Dual Deep Network Based Secure Deep Reinforcement Learning，DDN-SDRL)方法.

输入：原始网络模型、深度网络模型、安全距离 m

输出：训练完毕后的原始网络参数 $\boldsymbol{\theta}_1$、深度网络参数 $\boldsymbol{\theta}_2$

1. 初始化：经验回放单元 D、危险经验回放单元 D_d，原始网络情节结束终止奖赏 r_{T1}，深度网络情节结束终止奖赏 r_{T2}，抽样样本数量 minibatch 的大小 32，$\boldsymbol{\theta}_1$、$\boldsymbol{\theta}_2$ 初始化为随机较小数值
2. Repeat(对每一个情节)：
3. 使用预处理操作 $\phi()$ 对原始图像 x_1 进行处理，获得预处理后的图像 ϕ_1，$\phi_1=\phi(x_1)$
4. Repeat(对于情节中的每个时间步)：
5. 用 ε-贪心策略随机选择动作 u_t
6. 执行动作 u_t，得到观察的图像 o_{t+1} 和立即奖赏 r_t
7. 设置 $x_{t+1}=x_t$，形成元组 x_{t+1}，u_t，o_{t+1}，并预处理图像 $\phi_{t+1}=\phi(x_{t+1})$
8. If 到达当前情节终止状态
9. $y_1=r_{T1}$
10. $y_2=r_{T2}$
11. 将 $(\phi_{t-m},u_{t-m},r_{t-m},\phi_{t-m+1})$ 至 $(\phi_t,u_t,r_t,\phi_{t+1})$ 之间的样本存储到 D_d 中
12. Else
13. $y_1=r_1+\gamma\max\limits_{u'}Q(\phi_{i+1},u';\boldsymbol{\theta}_1)-\eta\text{avg}(Q(\phi_{i+1},u';\boldsymbol{\theta}_2))$，其中，使用式(12)计算 Q 值
14. $y_2=r_2+\gamma\max\limits_{u'}Q(\phi_{i+1},u';\boldsymbol{\theta}_2)$，其中，使用式(12)计算 Q 值
15. 将 $(\phi_t,u_t,r_t,\phi_{t+1})$ 作为一个样本存储到 D 中
16. End If
17. 从 D 中抽取 minibatch 个样本送入原始网络训练
18. 从 D_d 中抽取 minibatch 个样本送入深度网络训练
19. 分别对损失函数 $L(\boldsymbol{\theta}_1)=(y_1-Q(\phi,u;\boldsymbol{\theta}_1))^2$ 和 $L(\boldsymbol{\theta}_2)=(y_2-Q(\phi,u;\boldsymbol{\theta}_2))^2$ 的 $\boldsymbol{\theta}_1$ 和 $\boldsymbol{\theta}_2$ 进行梯度下降操作以更新权重
20. Until 智能体达到当前情节终止状态
21. Until 达到预期训练次数
22. Return $\boldsymbol{\theta}_1$，$\boldsymbol{\theta}_2$

DDN-SDRL方法将Agent与环境交互过程中产生的样本分为临界样本、危险样本和一般样本，其中，临界样本为从Agent任务失败开始的倒数 m 帧，危险样本为处于危险状态的样本. 这两种样本存储在经验回放单元 D_d 中，一般样本存储在经验回放单元 D 中. DDN-SDRL方法在原有深度网络的基础上新增了一个网络训练临界样本和危险样本，其目的在于计算临界状态对Agent再次陷入危险状态的影响程度. 在训练过程中，原始深度神经网络和深度网络分别从 D 和 D_d 中抽取样本进行随机梯度下降训练. 损失函数为

$$L(\boldsymbol{\theta}_1)=(y_1-Q(x,u;\boldsymbol{\theta}_1))^2 \tag{14}$$

其中，y_1 的计算方式为

$$y_1=r_1+\gamma\max_{u'}Q(\phi_{i+1},u';\boldsymbol{\theta}_1)-\eta\,\text{avg}(Q(\phi_{i+1},u';\boldsymbol{\theta}_2)) \tag{15}$$

式(15)中 $\eta\text{avg}(Q(\phi_{i+1},u';\boldsymbol{\theta}_2))$ 代表临界样本模型对当前值网络的惩罚项，η 控制着深度网络模型对原始深度网络模型的影响力，$\text{avg}(Q(\phi_{i+1},u';\boldsymbol{\theta}_2)$ 为经过深度网络训练后危险样本的平均动作值函数. 无惩罚项的原目标函数根据损失函数进行梯度下降操作时易陷入局部最优，对加入惩罚项之后的式(15)进行平方操作，得到

$$y_1^2=[r_1+\gamma\max_{u'}Q(\phi_{i+1},u';\boldsymbol{\theta}_1)-\eta\text{avg}(Q(\phi_{i+1},u';\boldsymbol{\theta}_2))]^2 \tag{16}$$

由式(16)可以看出，加入惩罚项后的 y_1 距离未加入惩罚项的 y_1 相差了 $\eta\text{avg}(Q(\phi_{i+1},u';\boldsymbol{\theta}_2))$. 从二维平面来看，加入惩罚项后的 y_1 所能取得的值为以

$r_1+\gamma\max_{u'}Q(\phi_{i+1},u';\boldsymbol{\theta}_1)$为圆心，$\eta\text{avg}(Q(\phi_{i+1},u';\boldsymbol{\theta}_2))$为半径之外的区域，因此，式(16)可使 y_1 值在较为安全的范围之内，有效地避免进入危险状态的可能，从而保障 Agent 的安全性.

DDN-SDRL 方法改变了目标函数(增加了惩罚项)，相应地，最优函数也会随之改变. 然而，DDN-SDRL 方法针对性地训练了临界样本，在危险状态边缘的状态所训练得到的最优策略会受到较大影响，使得函数避免陷入局部最优，因此可以有效限制 Agent 向危险状态方向探索. 此外，由于安全样本距离危险状态边缘较远，目标函数惩罚项对其影响较小，保证了算法的稳定性.

3.4 网络训练过程

经过预处理的图像再由卷积神经网络提取特征，卷积过程为

$$C=f\left(\sum_{t=0}^{T-1}\boldsymbol{W}_t\cdot\boldsymbol{X}_t+\boldsymbol{b}\right) \tag{17}$$

其中，$\boldsymbol{X}$ 为经过处理后的图像矩阵，$\boldsymbol{W}$ 为卷积核，$\boldsymbol{b}$ 为偏置向量，f 为 ReLU 激活函数，T 为卷积核数量. 本文网络模型采用了 3 层卷积神经网络，在每层网络后又进行了非线性变换，提取图像特征.

由于池化(Pooling)操作会使得提取得到的特征忽视位置信息，而在实验环境中位置信息也会对奖赏产生较大影响，因此，DDN-SDRL 方法在通过卷积神经网络提取特征后未进行池化操作.

在网络的全连接层中，DDN-SDRL 方法通过两层全连接层以完成最终的从状态到动作的映射. 对于有 m 个输入节点 n 个输出节点的全连接层而言，其更新过程可以表示为

$$\boldsymbol{y}=\boldsymbol{w}\cdot\boldsymbol{x}+\boldsymbol{b} \tag{18}$$

其中，$\boldsymbol{x}\in\boldsymbol{R}^n$ 表示输入层信息，$\boldsymbol{w}\in\boldsymbol{R}^{n\times m}$ 表示全连接层的参数矩阵，$\boldsymbol{b}\in\boldsymbol{R}^n$ 表示偏置向量.

网络最终输出值为动作值函数，在 DDN-SDRL 方法中，以其均值作为惩罚项.

4 实验结果及分析

本节首先介绍实验平台和实验过程中的主要参数. 接着介绍了对比实验所使用的方法和实验环境. 为了评估 DDN-SDRL 方法在原始网络选择上的通用性，本文实验使用 DQN、DuDQN 和 DRQN 这 3 种不同的深度网络作为原始网络；分析结合这 3 种不同原始网络的情况下，DDN-SDRL 方法在 Atari 2600 实验平台部分游戏上的效果. 最后，根据实验结果分析 DDN-SDRL 方法的优缺点和适用范围.

4.1 实验平台及参数设置

本文采用的游戏环境是基于人工智能公司 OpenAI gym 工具包中的 Atari 2600 游戏实验平台. Atari 2600 游戏主要包括射击类、战略类、体育竞技类等方面的策略游戏，为研究人员提供了种类多样的实验环境.

实验采用的处理器为 Intel i7-7820X(8 核)，主频为 3.60 GHz，内存为 16 GB. 由于模型中大多用到了卷积运算和矩阵运算，因此使用了 GTX 1080Ti 图形处理器对模型进行辅助加速运算.

本文在 Atari 2600 游戏中选取了 6 个游戏进行评估，表 1 为相关游戏任务的简要介绍.

表 1 部分 Atari 游戏的简要介绍

英文名称	动作数	智能体任务简述
CrazyClimber (疯狂攀登者)	9	Agent 避开障碍物爬上高楼
Kangaroo (袋鼠)	18	Agent 通过楼梯向上爬并避开障碍物
KungFuMaster (功夫大师)	14	Agent 需要不断击败出现的敌人
Pooyan (猪小弟)	6	Agent 躲避子弹并击败每一个出现的敌人
RoadRunner (公路奔跑者)	18	Agent 在公路上躲避前后的敌人
Zaxxon (空间逃脱)	18	Agent 驾驶飞船左右躲避障碍物

为了更好地比较各方法的优劣，实验中每一种游戏都使用了相同的参数设置. 鉴于在将强化学习方法应用于深度学习问题时，会导致不稳定现象的发生. 本文在实验中采用了一些措施以保证模型的稳定性，主要包括：(1) 由于 Agent 根据 Q 值选择动作所消耗的时间远大于网络传播时间，为了平衡两者差异，采用跳帧技术[28]来缓和，即在每 4 帧状态下都采取相同的动作，并以累积奖赏作为总奖赏，每过 4 帧再根据 ε-贪心策略选择下一个动作，若在跳帧过程中出现终止状态则实验结束；(2) 在 Atari 2600 游戏环境中，不同游戏的奖赏范围波动较大，这会造成较大的得分差异，因此，实验中将正奖赏设为+1，负奖赏设为−1，其余不变. 这样不仅可以更方便快捷地在不同算法间进行比较，而且简化了奖赏的设置，可以明确动作带来的优劣，使得判断更加简明；(3) 为了防止策略陷入局部最优，提升算法稳定性，将损失函数进行裁剪：损失值在区间[−1,1]之外的取损失值的绝对值；损失值在[−1,1]范围

之内的则采用式(14)的损失函数进行随机梯度下降操作.

实验中所有模型均采用均方根随机梯度下降方法(Root Mean Square Propagation,RMSProp)来更新网络参数.为了降低动能,RMSProp 方法中的动量系数设置为 0.95.经验回放单元最大存放 100 万个样本.由于训练开始阶段没有充足的样本数量,不能够满足样本多样性的要求,因此,在训练的前 50000 步,Agent 根据随机策略生成足够的样本分别保存到经验回放单元 D 和 D_d.网络训练时利用的样本数量 mini-batch 设置为 32.网络更新的学习率设置为 0.005,奖赏折扣率 γ 设置为 0.99.学习率设置过大则会使训练容易陷入局部最优,奖赏折扣率较小则对奖赏信息的利用率会降低,最终导致更新缓慢.对于行为策略 ε-贪心策略而言,在训练后期所需要的探索要求比训练前期少,训练后期较大的训练步幅可能导致其不收敛.因此,实验中探索因子 ε 是动态变化的,在训练过程中每过 100000 步,ε 降低 0.1,使得 ε 从 1.0 逐步下降至 0.1;网络更新参数 α 为原来的 0.96,从 0.005 逐步降至 0.00025.

4.2　实验结果及分析

在强化学习中,通常使用累积奖赏作为评判策略优劣的标准.由于深度学习训练周期长、训练数据庞大,训练效果不稳定,通常利用一个情节所获得的累积奖赏作为评估标准,通过分阶段统计每个情节奖赏大小评估模型优劣.

实验中所有模型进行 2000 个阶段的训练,每个训练阶段 10000 步,在每个阶段统计模型训练的每情节平均奖赏值、模型损失值、游戏轮数等指标.由于在有些游戏中,gym 包会在失败后再给若干复活机会,所有机会消耗完后一个情节才彻底结束,而本文的实验主要验证在深度强化学习中加入安全性后的效果,临界样本出现于每一次游戏失败前的若干步骤.因此,重新定义强化学习的一个情节为从开始游戏到游戏第一次失败的过程.

在实验中,分别选择 DQN、DuDQN 和 DRQN 网络模型作为原始网络,将 DDN-SDRL 应用到上述 3 种网络模型中,形成带安全机制的 DQN、带安全机制的 DuDQN 和带安全机制的 DRQN,分别记为 DQN-SE、DuDQN-SE 和 DRQN-SE.

4.2.1　CrazyClimber 实验及分析

在 CrazyClimber 中,Agent 不断爬上高楼并躲避存在的各种障碍物以获取更多的得分.

图 2 对比了 3 种原始网络(DQN、DuDQN 和 DRQN)与 3 种带安全机制的网络(DQN-SE、DuDQN-SE 和 DRQN-SE)在 CrazyClimber 游戏中的训练结果.图中纵坐标为平均每情节奖赏值,横坐标为训练情节数.

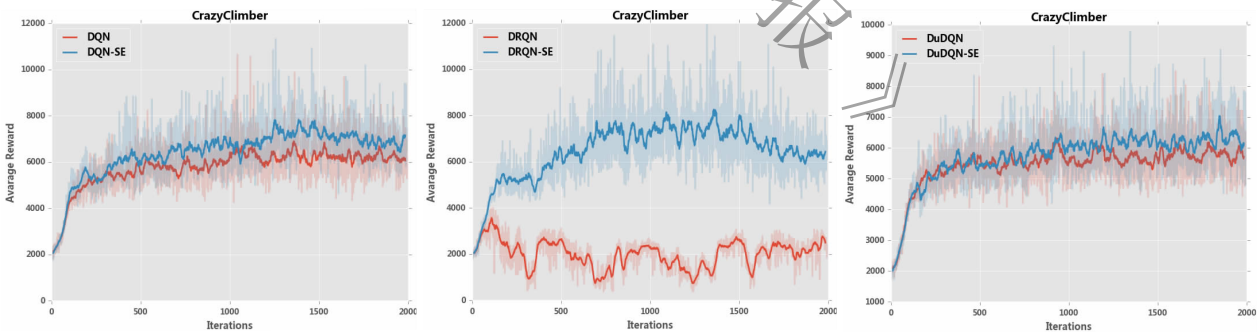

图 2　DDN-SDRL 在 CrazyClimber 实验中的效果对比图

由图 2 可以看出,在 CrazyClimber 环境中,基于 DDN-SDRL 方法的 DQN-SE、DRQN-SE 和 DuDQN-SE 较原始模型在平均每情节的奖赏有一定程度的提高.然而,除了 DRQN-SE,加入了安全机制的 DQN-SE 和 DuDQN-SE 方法的提升效果并不突出,其原因在于 CrazyClimber 游戏中,从临界状态调整至安全状态所需步骤较为复杂,需要经过一系列不同的步骤才能脱离临界状态,导致 DDN-SDRL 方法学习临界状态信息较慢,提升效果有限.此外,由于该任务较为复杂,所需训练时间步较多,使得应用了安全机制的模型比原始模型的优势不明显.

4.2.2　Kangaroo 实验及分析

Kangaroo 是一种较为复杂的策略类游戏,Agent 共有 18 个动作可供选择,游戏目的在于躲避或消灭出现的敌人并不断攀登楼梯以得到更多的分数.Kangaroo 游戏难度在于游戏中有较多陷阱使得游

戏失败，Agent 不仅需要躲避来自各个方向的子弹，还需避开脚下的陷阱. 此外，游戏过程中还有时间限制.

图 3 对比了 3 种原始网络(DQN、DuDQN 和 DRQN)与 3 种带安全机制的网络(DQN-SE、DuDQN-SE 和 DRQN-SE)在 Kangaroo 游戏中的训练结果. 图中纵坐标为平均每情节奖赏值，横坐标为训练情节数.

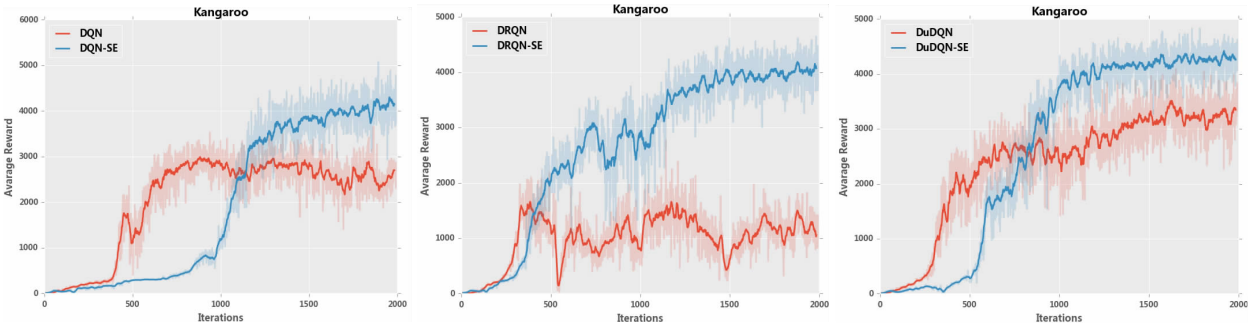

图 3 DDN-SDRL 在 Kangaroo 实验中的效果对比图

从图 3 中可以看出，在训练开始阶段，利用了 DDN-SDRL 的模型经历了少数若干次失败后即获得经验，对于相似的失败状态划定了一定数量的临界状态，当 Agent 再次遇到类似的状态后能够做出合理的动作选择. 这种新的探索机制增加了 Agent 获取更高分数的可能性，因此利用了 DDN-SDRL 方法的模型表现高于原始模型.

4.2.3 KungFuMaster 实验及分析

在 KungFuMaster 游戏中，Agent 通过击败从两边出现的敌人以获得更高的得分.

图 4 对比了 3 种原始网络(DQN、DuDQN 和 DRQN)与 3 种带安全机制的网络(DQN-SE、DuDQN-SE 和 DRQN-SE)在 KungFuMaster 游戏中的训练结果. 图中纵坐标为平均每情节奖赏值，横坐标为训练情节数.

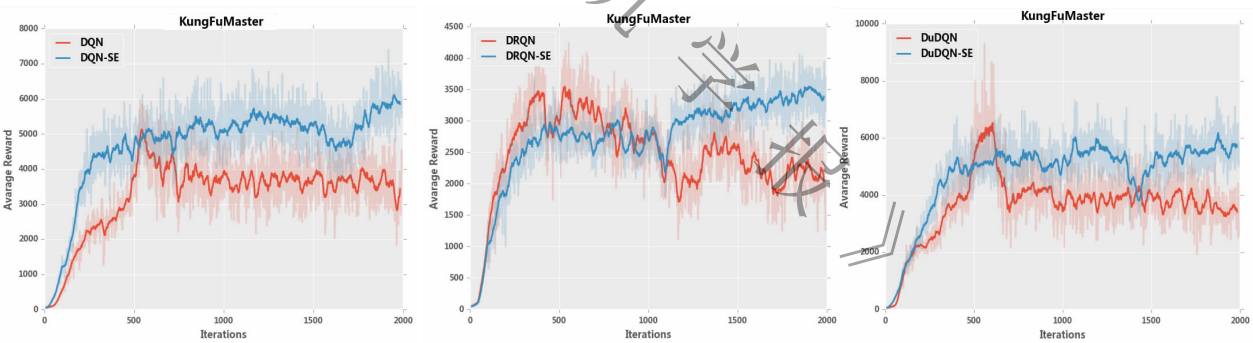

图 4 DDN-SDRL 在 KungFuMaster 实验中的效果对比图

在 KungFuMaster 游戏环境中，Agent 与近距离的敌人对抗，若击败敌人则获得奖赏，反之则游戏失败. 这个游戏的难点在于得分情况的图像和失败情况的图像相似，造成临界状态(濒临危险)和安全状态(安全)的区分度较低，难以通过图像信息加以区分，网络模型也很难从中提取有效信息选择近距离动作. 因此，DDN-SDRL 方法中危险经验回放单元 D_d 所收集的临界状态存在一定误差，导致终使得 DDN-SDRL 方法所提升的效果有限. 尽管如此，采用了 DDN-SDRL 方法的网络平均奖赏值依然有一定的提高.

从图 4 中可以看出，3 个原始网络模型在 KungFuMaster 环境中的平均每情节奖赏值在训练阶段初期均上升，在训练阶段中期有较为明显的下降趋势. 这是因为在 KungFuMaster 游戏环境中，我方击败敌方的图像和敌方击败我方的图像显示信息较为相似，原始模型并未加以区分，导致了训练效果存在下降的趋势.

4.2.4 Pooyan 实验及分析

Pooyan 实验是一种益智类游戏，Agent 通过发射子弹击毁敌方，若未及时击败敌方，则当未击败的敌人累积到一定程度后游戏失败.

图 5 对比了 3 种原始网络(DQN、DuDQN 和 DRQN)与 3 种带安全机制的网络(DQN-SE、DuDQN-

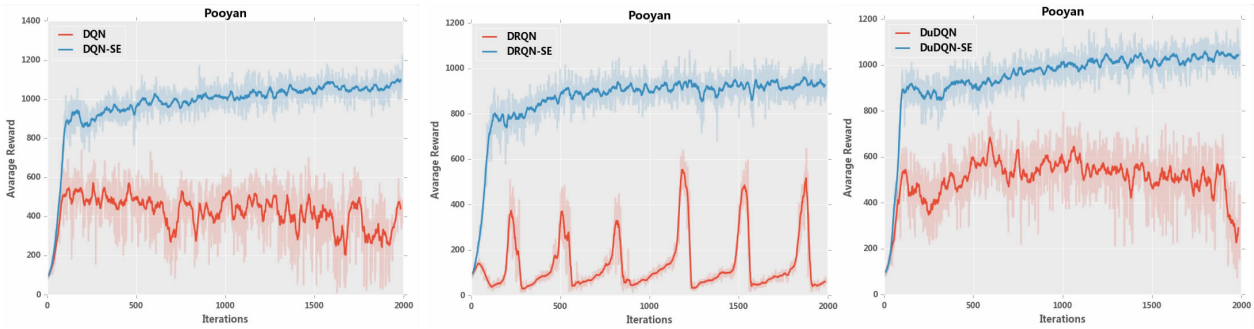

图 5　DDN-SDRL 在 Pooyan 实验中的效果对比图

SE 和 DRQN-SE)在 Pooyan 游戏中的训练结果. 图中纵坐标为平均每情节奖赏值,横坐标为训练情节数.

从图 5 中可以看到,利用了 DDN-SDRL 方法的网络在 Pooyan 中从训练开始到结束阶段平均每情节奖赏都较原始模型保持着较大的优势,其主要原因在于:(1)Pooyan 游戏环境仅有 6 个动作构成,训练较为简单,出现的状态情况较为相似;(2)Pooyan 游戏中安全状态与临界状态区别明显,较为容易通过卷积网络提取特征,而 DDN-SDRL 能有针对性地对临界样本进行训练. 这两点原因使得 DDN-SDRL 方法在 Pooyan 环境中训练效果较好.

4.2.5　RoadRunner 实验及分析

在 RoadRunner 中,Agent 向前运动,同时需要躲避往来的车辆. 在该游戏环境中,Agent 只能前进而不能后退,并且必须保持向左的动作以躲避后方出现的敌人.

图 6 对比了 3 种原始网络(DQN、DuDQN 和 DRQN)与 3 种带安全机制的网络(DQN-SE、DuDQN-SE 和 DRQN-SE)在 RoadRunner 游戏中的训练结果. 图中纵坐标为平均每情节奖赏值,横坐标为训练情节数.

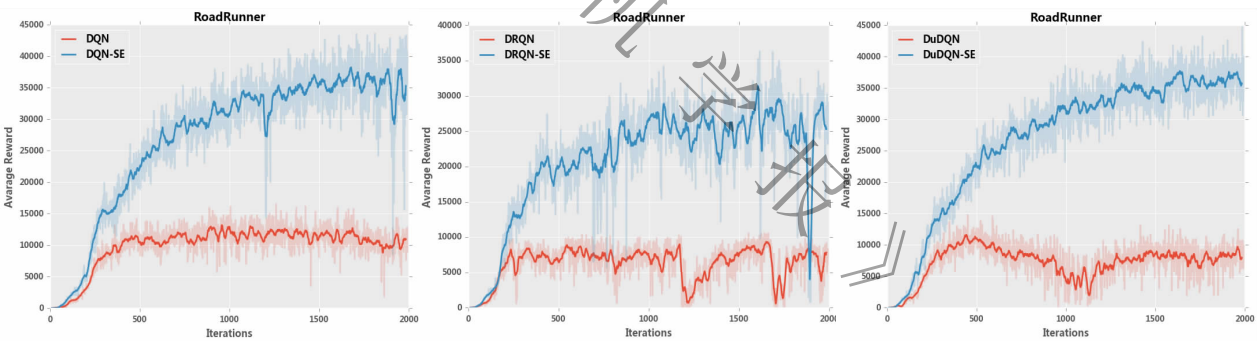

图 6　DDN-SDRL 在 RoadRunner 中的效果对比图

从图 6 中可以发现,DDN-SDRL 方法表现均超过了原始模型,以原始模型 DQN 为例,DQN 模型在 500 轮训练阶段之后平均每情节奖赏达到 12000 左右并保持稳定,而利用了 DDN-SDRL 方法的 DQN-SE 在 500 轮训练阶段已经平均每情节奖赏已达到 23000 左右且呈现上升趋势,在 2000 轮训练阶段结束后平均每情节奖赏更是达到了 35000. 虽然 RoadRunner 环境动作数量达到了 18 个,然而在较多状态下所采取的动作较为固定,因此训练难度较小. 另外,该游戏环境临界状态区分较为明显. 因此,同 Pooyan 环境类似,DDN-SDRL 方法在 RoadRunner 环境中训练效果较好.

4.2.6　Zaxxon 实验及分析

Zaxxon 需要 Agent 操纵飞机不断躲避障碍物并击毁敌方基地.

图 7 对比了 3 种原始网络(DQN、DuDQN 和 DRQN)与 3 种带安全机制的网络(DQN-SE、DuDQN-SE 和 DRQN-SE)在 Zaxxon 游戏中的训练结果. 图中纵坐标为平均每情节奖赏值,横坐标为训练情节数.

从图 7 中可以看出,由于 Zaxxon 游戏中躲避障碍物时状态与未躲避状态时动作极为相似,DDN-SDRL 方法较难精确识别临界状态与安全状态,从而导致训练提升效果有限. 在训练开始阶段甚至出现了原始模型训练效果好于 DDN-SDRL 方法的现象.

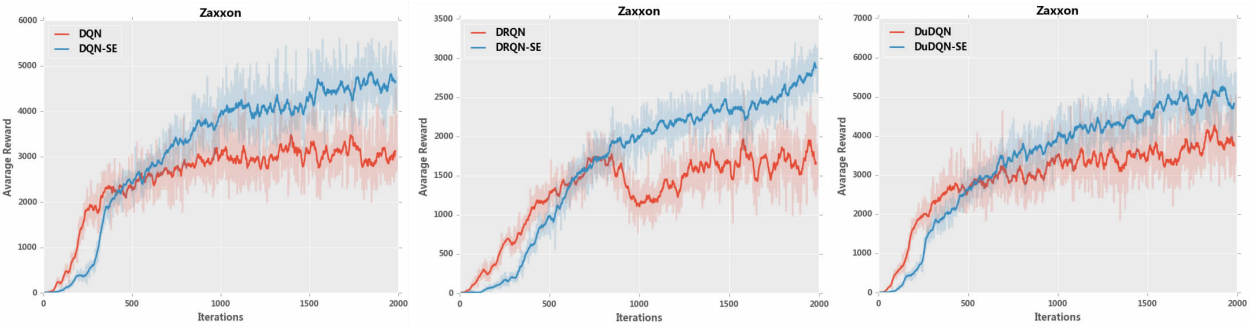

图 7 DDN-SDRL 在 Zaxxon 中的效果对比图

4.2.7 实验中其他问题的分析

在 Kangaroo、RoadRunner 和 Zaxxon 游戏的训练过程中，DQN-SE、DRQN-SE 和 DuDQN-SE 这 3 种模型在训练开始阶段都出现了平均每情节奖赏低于原始网络模型的现象，原因在于 DDN-SDRL 方法需要针对临界状态和危险状态进行训练，而这 3 种游戏的动作数量较多，在探索训练阶段需要耗费较大时间步以完成最佳动作的探索，从而产生足够的临界样本送入网络训练. 这导致使用了 DDN-SDRL 的网络模型在训练的开始阶段效果不如原始模型，而随着迭代步数的增加，DDN-SDRL 能有效避免重复陷入危险状态的效果逐渐突出，平均每情节奖赏在训练中后期超过原有模型并最终趋于稳定. 另一方面，由于游戏动作数量较多，在每一个情节下所需要的探索步骤也较为多样，因此出现训练过程中的波动现象，由于算法的自我调整，训练结果总体趋势向好. 值得注意的是，在训练阶段，由于 Agent 选择的动作是通过 ε-贪心策略获得，即选取动作时有一定概率采用随机动作而不是选择最优动作，采用该方法是为了保证强化学习中动作选择的探索性. 因此，增加了安全机制的 DDN-SDRL 算法训练效果是一个波动上升的过程.

4.3 算法适用性分析

通过对比实验，可以分析得到适用 DDN-SDRL 方法的环境所具备的特点，主要包括：(1) 在 Agent 探索的环境中存在较多可表示的导致游戏失败的状态，由于 DDN-SDRL 是针对临界和危险状态样本进行训练的模型，环境中导致危险状态的可能越多，其改进能力越强；(2) 在一个情节中的危险状态与完全安全状态之间存在着较大的间隔. 由于网络输入状态是 Agent 在学习过程中的灰度图像，若危险性状态与完全安全状态间隔较小，图像区分度也变得较低，这会使得 DDN-SDRL 对危险状态和安全状态的区分造成偏差，进而影响训练结果；(3) 在动作较少的环境中，DDN-SDRL 可以通过自我训练快速提高 Agent 的表现，反之，对于存在较多动作的环境，DDN-SDRL 在训练开始阶段可能存在着训练效果不明显的状况，随着训练的推进，训练效果会逐渐改善，最终效果优于原始模型. 为了直观表示 DDN-SDRL 的最终效果，实验统计了模型在不同训练阶段产生的危险状态数量. 图 8 显示了随着训练的不断进行，各个模型在不同训练阶段产生危险状态的数量. 纵坐标代表平均每阶段产生的危险状态数量.

从图 8 中可以看出，在大多数情况下，利用了 DDN-SDRL 网络模型的危险状态数量在训练开始阶段比未用 DDN-SDRL 方法的模型高；随着训练的进行，DDN-SDRL 方法在每阶段的危险状态数量逐渐减少并开始收敛，而原始网络模型不仅危险状态数量较 DDN-SDRL 方法多，并且每阶段危险状态数量上下波动范围大，说明其学习不稳定. 其原因在于 DDN-SDRL 方法可以对临界状态和危险状态进行针对性训练，因此有效地降低了危险状态再次发生的可能，进而提高训练结果和稳定性. 此外，可以看出，在部分游戏实验环境中，利用了 DDN-SDRL 方法的网络模型在接近训练结束时产生的危险状态数量仅比原始模型略少，而从图 2 至图 7 看出其奖赏却相差较大，说明 DDN-SDRL 方法在训练末期偏向于采用了奖赏值较大的动作，在有效避免危险状态的前提下保证了训练效果.

好的网络模型应该在测试阶段依然有良好的表现，Agent 会根据训练好的策略执行任务也能获得优秀的表现. 本文在 2000 个训练阶段结束后对训练所得的模型进行测试. 在测试中，步长设置为 10 个阶段，每个阶段有 10000 个时间步，Agent 的行为策略为 ε-贪心策略，ε 设为 0.05. 本文对每一个模型进

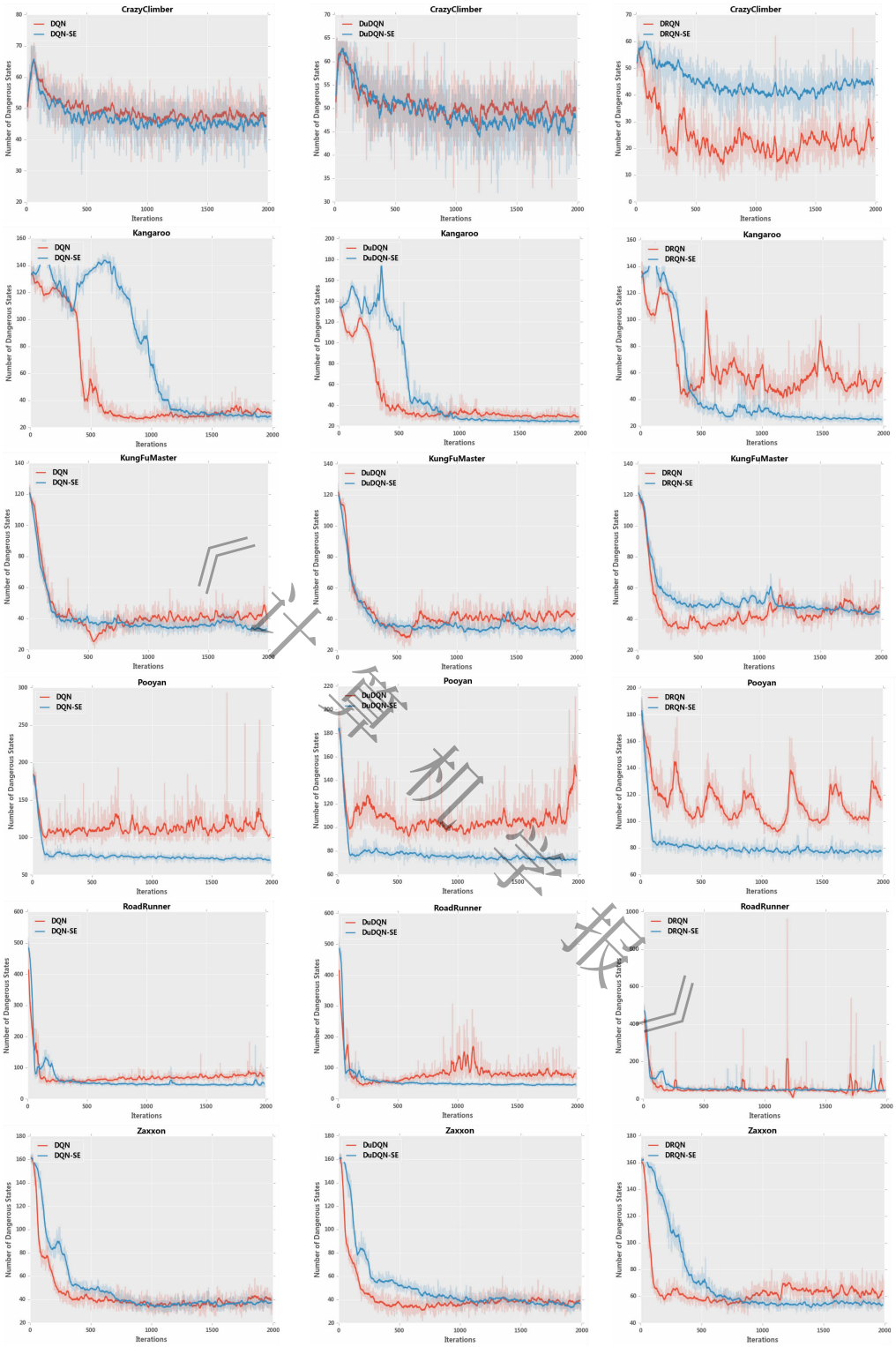

图 8　DDN-SDRL 在训练阶段产生危险状态数量对比图

行 20 轮测试. 表 2 给出了各模型在 6 个游戏中的测试结果.

从表 2 中可以看出，利用了 DDN-SDRL 方法的模型在测试后的最高奖赏和平均奖赏指标中都是最高的，即 DQN-SE、DuDQN-SE 和 DRQN-SE 这 3 种模型表现分别优于 DQN、DuDQN 和 DRQN. 采用了 DDN-SDRL 的网络模型在测试阶段遭遇的灾难性数量也是最少的，说明在模型中增加了安全性限制后，可以减少带来灾难性状态的选择，使得控制决策更为安全有效. 此外，DDN-SDRL 在 3 种不同的网络模型中均取得较好的效果，也体现出模型具有一定的通用性.

表 2 不同模型在部分 Atari 游戏中的测试结果

游戏名	模型	平均奖赏	最大奖赏	标准差	危险状态数量
CrazyClimber	DQN	5568.22	7463.41	688.27	45.2
	DQN-SE	6126.26	7865.85	808.26	39.8
	DuDQN	5963.26	7227.27	543.98	48.4
	DuDQN-SE	6979.78	9405.58	984.80	42.5
	DRQN	2524.23	2817.39	232.17	45.4
	DRQN-SE	6578.79	7884.21	699.56	43.5
Kangaroo	DQN	2734.61	2920.00	173.74	31.2
	DQN-SE	4118.96	4784.00	294.45	28.5
	DuDQN	3313.79	3840.00	241.72	30.8
	DuDQN-SE	4288.00	4621.73	204.37	24.4
	DRQN	994.06	1423.91	206.49	56.5
	DRQN-SE	4064.00	4643.47	284.64	25.5
KungFuMaster	DQN	3441.07	4548.57	458.74	43.6
	DQN-SE	5848.43	6532.25	305.68	35.5
	DuDQN	3313.79	3840.00	241.72	28.5
	DuDQN-SE	4288.00	4621.73	204.37	25.2
	DRQN	2010.39	2840.54	376.21	53.4
	DRQN-SE	3423.86	3943.90	263.61	43.5
Pooyan	DQN	453.53	505.35	50.35	108.4
	DQN-SE	1096.44	1221.84	50.14	69.5
	DuDQN	305.11	436.44	110.32	122.5
	DuDQN-SE	1047.73	1157.94	47.23	72.4
	DRQN	62.21	79.95	9.17	119.0
	DRQN-SE	930.83	989.25	35.69	77.5
RoadRunner	DQN	11121.22	12800.00	1208.64	77.2
	DQN-SE	36556.52	43317.94	6399.38	48.8
	DuDQN	7778.31	10721.21	1507.39	84.6
	DuDQN-SE	35332.58	44720.51	3089.63	45.5
	DRQN	8095.74	9190.00	1077.93	62.4
	DRQN-SE	26333.74	30238.63	3803.14	43.2
Zaxxon	DQN	3597.96	4553.12	454.05	42.2
	DQN-SE	4358.66	4917.94	273.18	39.5
	DuDQN	3690.93	5038.70	496.37	41.2
	DuDQN-SE	4705.40	5625.00	460.11	36.5
	DRQN	1675.30	2139.58	226.54	58.5
	DRQN-SE	2895.19	3166.03	171.87	52.6

值得注意的是，测试结果中不同游戏环境表现出的标准差差别较大，而获得平均奖赏较大的模型其标准差也较高. 多种因素造成了该现象，其中一个重要的原因就是模型训练结果有一定的波动，强化学习的探索性导致其训练无法达到十分精确的结果，最终导致在测试阶段的波动. 而获得平均奖赏较高的模型，其相对波动更大. 因此，测试结果表现出平均奖赏较大的模型标准差较大.

5 结束语

以 DQN 模型为代表的深度强化学习方法已经在基于视觉的深度强化学习问题中取得了突破. 然而这些网络模型并没有考虑实际环境中对 Agent 及所控制对象的保护. 在实际环境中，由于成本问题，不能使 Agent 无限制地陷入危险状态. 为了避免 Agent 在训练过程中陷入危险状态，本文提出一种基于双深度网络的安全深度强化学习方法. DDN-SDRL 方法通过建立双经验池分离临界样本，并使用 DQN 网络对临界样本进行针对性训练；另一方面，目标函数加入惩罚项，对 Agent 的探索进行适当的限制.

本文通过 6 个 Atari 2600 游戏实验验证了 DDN-SDRL 方法的有效性，并在 DQN、DuDQN 和 DRQN 这 3 种网络模型中应用了 DDN-SDRL 方法. 实验表明，加入了 DDN-SDRL 方法的网络模型在游戏平均每情节奖赏中具有优势. 而成功应用于 3 种深度强化学习模型则说明 DDN-SDRL 方法具有较强的泛化能力. 此外，针对训练结束后的模型进行测试的结果也表明，利用了 DDN-SDRL 方法的网络模型在应用任务中具有较好的稳定性.

然而，加入了 DDN-SDRL 方法的网络模型在训练中依然存在着波动范围大、训练不稳定的现象. 有一些相关研究工作，如 Anschel 等人[29]提出的平均深度 Q 网络模型可以有效地降低模型方差并提升实验效果. 因此，下一步将针对模型的稳定性开展研究，通过进一步改进算法，达到减小方差并提升稳定性的效果.

参 考 文 献

[1] Sutton R S, Barto A G. Reinforcement Learning: An Introduction. Cambridge, USA: MIT Press, 2018

[2] Lee D, Seo H, Jung M W. Neural basis of reinforcement learning and decision making. Annual Review of Neuroscience, 2012, 35: 287-308

[3] Liu Quan, Zhai Jian-Wei, Zhang Zong-Zhang, et al. A survey on deep reinforcement learning. Chinese Journal of Computers, 2018, 41(1): 1-27(in Chinese)
(刘全，翟建伟，章宗长等. 深度强化学习综述. 计算机学报，2018，41(1)：1-27)

[4] Xu X, Zuo L, Huang Z. Reinforcement learning algorithms with function approximation: Recent advances and applications. Information Sciences, 2014, 261(5): 1-31

[5] Zhu F, Liu Q, Zhang X, Shen B. Protein-Protein Internation network constructing based on text mining and reinforcement learning with application to prostate cancer. Systems Biology Iet, 2015, 9(4): 106-112

[6] Silver D, Sutton R S, Müller M. Temporal-difference search in computer Go. Machine Learning, 2012, 87(2): 183-219

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

[8] Zhang Q, Yang L T, Chen Z, et al. A survey on deep learning for big data. Information Fusion, 2018, 42: 146-157

[9] Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 1725-1732

[10] Watkins C, Dayan P. Q-learning. Machine Learning, 1992, 8(3-4): 279-292

[11] Mnih V, Kavukcuoglu K, Silver D, et al. Playing Atari with deep reinforcement learning//Proceedings of the Neural Information Processing Systems. Lake Tahoe, USA, 2013: 211-219

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

[13] 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

[14] Van Hasselt H. Double Q-learning//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2010: 2613-2621

[15] Hausknecht M, Stone P. Deep recurrent Q-learning for partially observable MDPs//Proceedings of the AAAI Conference on Artificial Intelligence. Texas, USA. 2015: 1552-1560

[16] Wang Z, Freitas N D, Lanctot M. Dueling network architectures for deep reinforcement learning//Proceedings of the International Conference on Machine Learning. New York, USA, 2016: 1995-2003

[17] 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

[18] Berkenkamp F, Turchetta M, Schoellig A, et al. Safe model-based reinforcement learning with stability guarantees//Proceedings of the Advances in Neural Information Processing Systems. Long Beach, USA, 2017: 908-918

[19] Garcia J, Fernandez F. Safe exploration of state and action spaces in reinforcement learning. Journal of Artificial Intelligence Research, 2012, 45(1): 515-546

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

[21] Garcıa J, Fernandez F. A comprehensive survey on safe reinforcement learning. Journal of Machine Learning Research, 2015, 16(1): 1437-1480

[22] Chow Y, Tamar A, Mannor S, et al. Risk-sensitive and robust decision-making: A CVaR optimization approach//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2015: 1522-1530

[23] Hans A, Schneegaß D, Schäfer A M, et al. Safe exploration for reinforcement learning//Proceedings of the European Symposium on Artificial Neural Networks. Bruges, Belgium, 2012: 143-148

[24] Song Y, Li Y B, Li C H, et al. An efficient initialization approach of Q-learning for mobile robots. International Journal of Control Automation & Systems, 2012, 10(1): 166-172

[25] Pradyot Korupolu V N. Beyond Rewards: Learning From richer Supervision [Ph. D. dissertation]. Indian Institute of Technology, Madras, Indian, 2012

[26] Agarwal A, Kumar V, Dunovan K, et al. Better safe than sorry: Evidence accumulation allows for safe reinforcement learning//Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 1-8

[27] Mhamdi E M E, Guerraoui R, Hendrikx H, et al. Dynamic safe interruptibility for decentralized multi-agent reinforcement learning//Proceedings of the Advances in Neural Information Processing Systems. Long Beach, USA, 2017: 130-140

[28] Lakshminarayanan A S, Sharma S, Ravindran B. Dynamic frame skip deep Q network//Proceedings of the Workshop on Deep Reinforcement Learning of International Joint Conferences on Artificial Intelligence. New York, USA, 2016: 111-117

[29] Anschel O, Baram N, Shimkin N. Averaged-DQN: Variance reduction and stabilization for deep reinforcement learning//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia, 2017: 176-185

ZHU Fei, Ph. D., associate professor. His main research interests include reinforcement learning, deep reinforcement learning and text mining.

WU Wen, M. S. candidate. His main research interest is deep reinforcement learning.

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

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

Background

Deep reinforcement learning is currently a research focus in the field of artificial intelligence. However, during the process of the exploration and decision of the agent, it will cause serious damage to the agent after the failure of exploration. What's more, the failure will also cause significant economic losses in practical applications. Therefore, it is important for the agents to guarantee the security during the exploration and decision.

In order to guarantee the security of the agents, this paper proposes a novel secure exploration method based on deep Q-network, named Dual Deep Network Based Secure Deep Reinforcement Learning Method. The model divide experience into critical experience and general experience, and improve the objective function through targeted training critical experience. Experimentally, our preliminary results demonstrated that training agents generated through our new model gets less number of dangerous states on six Atari 2600 games. In addition, the new model is can be applied to many deep networks, such DQN, DuDQN and DRQN.

This paper is supported by the National Natural Science Foundation of China (61303108, 61373094, 61772355), the Jiangsu College Natural Science Research Key Program (17KJA520004), the Program of the Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education (Jilin University) (93K172014K04), the Suzhou Key Industries Technological Innovation-Prospective Applied Research Project (SYG201804) Program of the Provincial Key Laboratory for Computer Information Processing Technology (Soochow University) (KJS1524), the China Scholarship Council Project (201606920013). These projects aim to enrich the reinforcement learning theory and develop efficient approximate algorithms to expand the power and applicability of reinforcement learning on large scale problems.