



基于循环卷积神经网络的POMDP值迭代算法

于丹宁,倪 坤,刘云龙

(厦门大学 航空航天学院,福建 厦门 361102)

摘 要: 基于卷积神经网络的部分可观测马尔科夫决策过程(POMDP)值迭代算法 QMDP-net 在无先验知识的情况下具有较好的性能表现,但其存在训练效果不稳定、参数敏感等优化难题。提出基于循环卷积神经网络的 POMDP 值迭代算法 RQMDP-net,使用门控循环单元网络实现值迭代更新,在保留输入和递归权重矩阵卷积特性的同时增强网络时序处理能力。实验结果表明,RQMDP-net 在 10×10 网格地图规划任务中导航准确率高达 98.5%,且在 36×36 网格地图规划任务中相比 QMDP-net 最多提升 5.8 个百分点,具有更快的网络收敛速度和更强的导航任务规划能力。

关键词: 部分可观测马尔科夫决策过程;值迭代;卷积神经网络;循环卷积神经网络;智能体规划

开放科学(资源服务)标志码(OSID):



中文引用格式: 于丹宁,倪坤,刘云龙.基于循环卷积神经网络的 POMDP 值迭代算法[J].计算机工程,2021,47(2): 90-94,102.

英文引用格式: YU Danning, NI Kun, LIU Yunlong. Value iteration algorithm for POMDP based on recurrent convolutional neural network[J]. Computer Engineering, 2021, 47(2): 90-94, 102.

Value Iteration Algorithm for POMDP Based on Recurrent Convolutional Neural Network

YU Danning, NI Kun, LIU Yunlong

(School of Aerospace Engineering, Xiamen University, Xiamen, Fujian 361102, China)

[Abstract] The value iteration algorithm, QMDP-net, for Partially Observable Markov Decision Process (POMDP) based on Convolutional Neural Network (CNN) performs well in cases of no prior knowledge. However, it often suffers from instable training results, sensitive parameter and other optimization problems. For these problems, this paper proposes a value iteration algorithm called RQMDP-net for POMDP based on Recurrent Convolutional Neural Network (RCNN). The update of value iteration is realized by using Gated Recurrent Unit (GRU), which keeps the input and convolution features of the recursive weight matrix, and enhances the sequential processing ability of the network. Experimental results show that the navigation accuracy of RQMDP-net for 10×10 planning tasks in the grid map reaches 98.5%, and is up to 5.8 percentage points higher than that of QMDP-net for 36×36 planning tasks in the grid map, which demonstrates that RQMDP-net has a higher network convergence speed and better planning ability in navigation tasks.

[Key words] Partially Observable Markov Decision Process (POMDP); value iteration; Convolutional Neural Network (CNN); Recurrent Convolutional Neural Network (RCNN); agent planning

DOI: 10. 19678/j. issn. 1000-3428. 0057027

0 概述

随着人工智能技术的快速发展,智能体规划被广泛应用于组合调度、游戏博弈等任务^[1-2]中,然而现实世界中的动态系统多数面向部分可观测环境,针

对部分可观测环境下的智能体规划问题,部分可观测马尔科夫决策过程(Partially Observable Markov Decision Process, POMDP)模型应用而生^[3-5]。POMDP 模型的核心思想是将动态系统中的不确定性规划问题转化为最优化问题进行求解,但由于其基于系统

基金项目: 国家自然科学基金(61772438, 61375077)。

作者简介: 于丹宁(1994—),女,硕士研究生,主研方向为深度强化学习、智能体决策;倪 坤,硕士研究生;刘云龙(通信作者),副教授、博士。

收稿日期: 2019-12-25 **修回日期:** 2020-02-04 **E-mail:** ylliu@xmu.edu.cn

隐含状态空间进行建立,因此人为建立模型需要大量先验知识并且存在容易陷入局部极小值的问题^[6-7]。深度神经网络作为一种多层次特征学习网络,能够自动从训练数据中学习抽象特征^[8]。KARKUS等人在深度神经网络与QMDP模型的基础上,提出基于卷积神经网络(Convolutional Neural Network, CNN)的POMDP值迭代算法QMDP-net^[9],其是QMDP模型的网络化表示,能使POMDP模型所需的参数以网络中权值的形式通过训练数据进行自动学习,无需提供大量的先验知识或假设POMDP模型已知。此外,QMDP-net已被证明在未预先给定环境模型的情况下可有效解决2D网格地图上的导航规划问题^[10-12]。

由于QMDP-net中的值迭代模块是通过卷积层与最大池化层相结合的网络结构进行表示,然而该网络结构使得QMDP-net存在训练结果不稳定、随机种子及超参数敏感等问题^[13-14]。为解决上述问题,本文提出一种基于循环卷积神经网络(Recurrent Convolutional Neural Network, RCNN)的POMDP值迭代算法RQMDP-net,使用门控循环单元(Gated Recurrent Unit, GRU)网络实现值迭代过程,并以经典游戏《格子世界》网格地图上的导航规划任务为例对RQMDP-net算法的有效性进行验证。

1 基于CNN的POMDP值迭代算法

1.1 POMDP模型

POMDP是一种对部分可观测环境规划问题进行系统建模的常用模型。POMDP模型由一个七元组构成: $M=(S, A, O, T, Z, R, b_0)$ ^[15],其中: S, A, O 分别表示动态系统的所有状态集合、动作集合和观测集合; $T(s, a, s') = \Pr(s'|s, a)$ 表示状态转移概率,即在状态 s 下执行动作 a 后,转移到其他状态 s' 的概率分布; $Z(s, a, o) = \Pr(o|s, a)$ 表示观测概率,即在状态 s 下执行动作 a 后,获得观测值 o 的概率分布; $R(s, a)$ 表示在状态 s 下执行动作 a 所获得的奖励; b_0 表示初始状态分布,即在初始时刻智能体在状态集合 S 上的分布。

在部分可观测环境下,智能体仅通过当前观测无法准确感知当前所处的状态,因此需要根据过去的历史序列 $\{a_1, o_1, a_2, o_2, \dots, a_t, o_t\}$ 对当前状态进行估计。POMDP引入信念状态 b 来表示智能体的当前状态,其中 b 是对过去所有历史信息的总体统计量,代表当前所有隐含状态的概率分布^[16]。在已知当前

信念状态 b 、执行动作 a 和获得观测值 o 的情况下,通过贝叶斯公式的更新来获得下一时刻的信念状态^[17]可表示为:

$$b'(s') = \Pr(s'|o, a, b) = \frac{\Pr(o|s', a, b)\Pr(s'|a, b)}{\Pr(o|a, b)} = \frac{Z(s', a, o) \sum_{s \in S} T(s, a, s')b(s)}{\Pr(o|a, b)} \quad (1)$$

1.2 值迭代对POMDP的求解

值迭代对POMDP的求解是在建立准确POMDP模型的基础上,使用值迭代算法进行动作选择以达到回报最大化的目的。值迭代作为求解马尔科夫决策过程(Markov Decision Process, MDP)的一种经典动态规划算法,其从任意初始状态值开始,使用贝尔曼方程组迭代求解状态的值函数。令 $V_k(s)$ 表示状态 s 在第 k 次迭代中的评估值,值迭代过程可表示为^[9]:

$$Q_{k+1}(s, a) = \sum_{s'} \Pr(s'|s, a)(R(s, a, s') + \gamma V_k(s')) \quad (2)$$

$$V_{k+1}(s) = \max_a Q_{k+1}(s, a) \quad (3)$$

POMDP模型使用信念状态 b 表示智能体当前所处状态,其向量中元素 $b(s)$ 表示智能体当前处于状态 s 的概率。当 k 趋于无穷大时,值函数 $V(s)$ 会收敛于最优值函数 $V^*(s)$,此时在 b 状态下执行动作 a 所得的最大回报 $Q(b, a) = \sum_{s \in S} b(s)Q_{\infty}(s, a)$,其对应的最优策略可表示为:

$$\pi^*(b) = \operatorname{argmax}_a Q(b, a) \quad (4)$$

1.3 QMDP-net算法

由于使用值迭代算法对POMDP问题进行求解的前提是建立准确的POMDP模型,然而学习动态系统的POMDP模型通常很困难,因此模型建立需要大量的先验知识。QMDP-net是一种用于解决部分可观测环境下动态规划问题的网络化值迭代算法。QMDP-net使用深度神经网络对POMDP算法的求解过程进行表示,使得所需POMDP模型的参数可以以网络中权值的形式通过训练数据进行自动学习^[9]。因此,QMDP-net可以在无先验知识的情况下对POMDP问题进行求解。

QMDP-net共分为POMDP模型和值迭代过程两部分。QMDP-net将POMDP模型中的状态转移概率、观测概率和奖励函数参数化为:

$$T(s, a, s') = f_T(s, a, s'|W_T) \quad (5)$$

$$\mathbf{Z}(s, a, o) = f_z(s, a, o | W_z) \quad (6)$$

$$\mathbf{R}(s, a) = f_r(s, a | W_r) \quad (7)$$

其中,函数 f_z 、 f_r 和 f_r 分别使用卷积神经网络进行表示,其对应的内核权重 W_z 、 W_z 和 W_r 通过端到端的训练方式从训练数据中获得。

在使用卷积层来参数化规划所需模型的基础上,利用卷积层和最大池化层构造值更新过程,并通过循环更新操作达到价值迭代的目的。第 k 次状态值的更新过程可表示为:

$$\mathbf{Q}_{k+1}(s, \bar{a}) = W_R^{\bar{a}} \mathbf{R}(s, a) + W_V^{\bar{a}} \mathbf{V}_k(s) \quad (8)$$

$$\mathbf{V}_{k+1}(s) = \max_a \mathbf{Q}_k(s, a) \quad (9)$$

其中, $\bar{a} \in A$ 表示系统中可执行的动作, \mathbf{Q} 、 \mathbf{V} 、 \mathbf{R} 分别表示动作-状态值函数、状态值函数和奖励函数, $W_R^{\bar{a}}$ 、 $W_V^{\bar{a}}$ 是用于构造奖励函数和状态值函数的卷积网络中 \bar{a} 对应的卷积核。

2 基于RCNN的POMDP值迭代算法

2.1 算法思想

虽然QMDP-net在无先验知识的情况下具有较好的性能表现,但其存在训练效果不稳定、参数敏感等优化难题。QMDP-net使用卷积层与最大池化层相结合的网络结构表示状态值的更新过程,由于卷积神经网络不具备记忆功能,因此需要通过不断循环运行该网络模块来达到值迭代的效果。循环神经网络(Recurrent Neural Network, RNN)具有记忆功能,更适用于循环处理时序问题^[18],因此,将值迭代过程编码为循环神经网络可有效缓解QMDP-net的优化难题。

由于RNN无法解决长期依赖问题,当循环次数较多时容易出现梯度消失现象^[19],因此本文使用门控循环单元网络来模拟值迭代过程,提出基于循环卷积神经网络的POMDP值迭代算法RQMDP-net。GRU通过门控机制有效缓解了RNN的梯度消失问题,而且相比LSTM具有更简单的网络结构^[20]。将值迭代过程使用由GRU和CNN结合构造的循环卷积神经网络进行表示,具体为:

$$\mathbf{V}_k(s) = \text{GRU}(W_R^a \mathbf{R}(s, a) + W_V^a \mathbf{V}_{k-1}(s)) \quad (10)$$

其中: W_R^a 和 W_V^a 表示转移函数的卷积神经网络中的卷积核,卷积核数量为 $|A|$;GRU的隐含状态 $\mathbf{V}_{k-1}(s)$ 和奖励 $\mathbf{R}(s, a)$ 在进行CNN状态转移计算后作为输入GRU进行新一轮迭代,在实现循环计算的同时保留QMDP-net的输入和递归权重矩阵的卷积特性。

2.2 RQMDP-net算法

RQMDP-net在经典游戏《格子世界》网格地图上的导航规划任务中,系统状态空间为 $N \times N$ (其中 N 为网格数量),对应信念状态 \mathbf{b} 可由 $N \times N$ 矩阵表示,该模型已知包含地图和任务目标信息的环境参数 X 。

对于POMDP模型的建立,本文使用双卷积神经网络结构。对实现状态更新的贝叶斯公式进行分解并将其表示为神经网络,其模型网络结构表达式为:

$$\mathbf{b}'_i(s', a) = \text{CNN}(\mathbf{b}_i(s) | W_T) \quad (11)$$

$$\mathbf{b}'_i(s') = \omega_{a_i} \mathbf{b}'_i(s', a) \quad (12)$$

$$\mathbf{b}'_i(s', o) = \text{CNN}(\mathbf{b}'_i(s') | W_Z) \quad (13)$$

$$\mathbf{b}_{i+1}(s') = \omega_{o_i} \mathbf{b}'_i(s', o) \quad (14)$$

其中: ω_{a_i} 表示输入动作 a_i 经过one-hot编码形成的权重向量; ω_{o_i} 表示当前观测 o_i 经过全连接层后输出的权重向量; $\mathbf{b}'_i(s', a)$ 、 $\mathbf{b}'_i(s', o)$ 分别为具有 $|A|$ 个卷积核和 $|O|$ 个卷积核的卷积网络的输出,其大小为 $N \times N \times |A|$ 和 $N \times N \times |O|$ 。

本文使用循环卷积神经网络实现值迭代过程,RQMDP-net网络结构如图1所示。可以看出,表示网格地图和任务目标的图像信息 θ 通过表示奖励函数 f_r 的网络转换为大小为 $N \times N \times |A|$ 的奖励信息 $\mathbf{R}(s, a)$,此网络是由两个卷积层组成的卷积神经网络:第一层卷积包含150个大小为 3×3 的卷积核,并使用线性整流函数(ReLu)作为激活函数,其作用是对输入图像信息进行特征提取;第二层卷积包含 $|A|$ 个 1×1 的卷积,其作用是将前一层输出的特征转换为用于价值迭代计算的 $\mathbf{R}(s, a)$ 。在奖励信息计算完成后,通过GRU实现价值迭代的计算过程,此处循环神经网络的神经元个数设置为150。在每次迭代时,作为状态价值 $\mathbf{V}(s)$ 的GRU隐含状态 h_i 经过表示转移函数 f_r 的网络后转换为表示 $\mathbf{Q}(s, a)$ 的 h'_i ,其网络由一个包含 $|A|$ 个大小为 3×3 卷积核的卷积层组成,之后 h'_i 与 $\mathbf{R}(s, a)$ 分别作为GRU的隐含状态和输入参与下一次迭代的值计算。经过 K 次迭代后的 h'_{i+K} 与当前信念状态 $\mathbf{b}(s)$ 相乘并相加得到 $\mathbf{Q}(\mathbf{b}, a)$,即在当前信念状态 \mathbf{b} 下,执行动作可获得 \mathbf{Q} 值。最终经过全连接(Fully Connected, FC)层和softmax层计算得到表示关于所有可执行动作的概率分布 $\text{Pr}(a)$,并选择对应 $\text{Pr}(a)$ 最大的 a 作为最优动作。

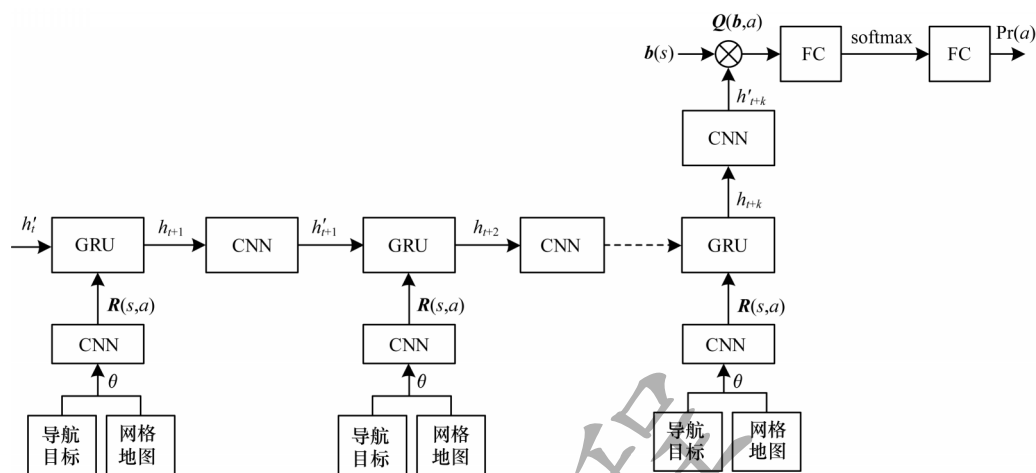


图 1 RQMDP-net 网络结构

Fig.1 RQMDP-net network structure

本文采用反向传播算法^[21]最小化交叉熵损失函数来优化深度神经网络模型,并将表示动作选择错误程度的损失函数定义为:

$$L = -\frac{1}{T} \sum_{t=1}^T \sum_{i=1}^{|A|} y'_{t,i} \log_a y_{t,i} \quad (15)$$

其中, $y'_{t,i}$ 为训练样本中 t 时刻标签动作经过 one-hot 编码后对应向量中第 i 个元素的值, $y_{t,i}$ 为 t 时刻网络输出的动作概率分布中第 i 个元素的值。

3 实验与结果分析

为验证基于循环卷积神经网络的值迭代算法 RQMDP-net 的有效性,实验在经典游戏《格子世界》网格地图上的导航规划任务中对 RQMDP-net 与 QMDP-net 的执行情况进行对比,并基于 TensorFlow 实现算法网络框架的搭建,同时使用 NVIDIA 1060 GPU 加速图像处理。

3.1 实验环境

实验任务是使智能体在 $N \times N$ 网格地图中进行导航。智能体已知的环境参数为标明障碍物和导航目标的 $N \times N$ 网格地图,其能观测四周是否有障碍物信息,而不同的位置周围障碍物的分布情况可能相同,因此智能体无法仅根据当前观测信息来获知自身在网格中的准确位置,即智能体状态。智能体可执行的动作包括向四周走动和原地不动 5 个。

3.2 实验设置与结果分析

在实验中,将来自 1 300 种随机环境下的 65 000 条专家轨迹(每个环境对应 50 条专家轨迹)作为数据集,其中,1 000 种随机环境的 50 000 条轨迹作为训练集,300 种环境的 15 000 条轨迹作为测试集。在网络训练过程中使用 ADAM 优化器更新网络参数,其初始学习率为 0.000 1。

本文实验将导航准确率和交叉熵损失值作为算法性能评价指标,其中,导航准确率为智能体导航至

目标位置的概率,交叉熵损失值为当前网络动作选择错误的概率。实验中有网格数量 N 和值迭代次数 K 2 个控制变量,其中, N 取值为 10、18、24、36, K 取值为 3、5、10、15。本文通过两组实验验证算法有效性及控制变量变化对算法性能的影响。

第 1 组实验通过设置不同的网格数量和值迭代次数来对比 RQMDP-net 和 QMDP-net 的导航准确率。由表 1 可以看出,在不同的网格数量下, RQMDP-net 的导航准确率高于 QMDP-net。在相同的网格数量下,随着值迭代次数的增加, RQMDP-net 的导航准确率在多数情况下相比 QMDP-net 增长更快。可见, RQMDP-net 在 10×10 网格地图中的导航准确率高达 98.5%,并且在 36×36 网格地图中相比 QMDP-net 最多提升 5.8 个百分点。

表 1 在 $N \times N$ 网格地图中 K 次值迭代的算法导航准确率对比

Table 1 Comparison of algorithm navigation accuracy of K iterations in the $N \times N$ grid map		%			
K	算法	导航准确率			
		$N=10$	$N=18$	$N=24$	$N=36$
3	RQMDP-net	94.0	75.4	66.7	42.3
	QMDP-net	91.9	76.2	64.3	40.6
5	RQMDP-net	95.2	78.0	68.4	44.7
	QMDP-net	93.4	75.7	65.0	43.5
10	RQMDP-net	98.2	83.0	71.3	49.3
	QMDP-net	96.9	78.6	67.8	44.2
15	RQMDP-net	98.5	89.4	75.0	52.8
	QMDP-net	95.6	80.7	69.8	47.0

第 2 组实验通过设置不同的网格数量和值迭代次数来对比 RQMDP-net 和 QMDP-net 的交叉熵损失值下降情况。由图 2 可以看出,与 QMDP-net 相比, RQMDP-net 的交叉熵损失值下降更快,可经过更少的数据集迭代次数达到最低值,主要原因为 RQMDP-net

利用GRU网络使其时序处理能力更强,最终交叉熵损失值也更小,即相同条件下的RQMDP-net动作选择错误的概率小于QMDP-net。

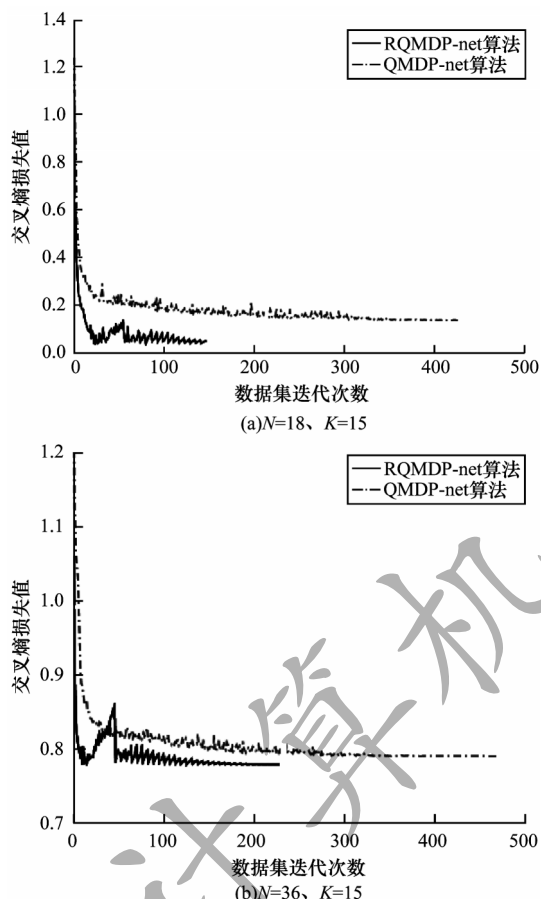


图2 交叉熵损失值与数据集迭代次数的关系

Fig.2 The relationship between cross entropy loss value and the number of iterations of the dataset

4 结束语

本文提出一种基于循环卷积神经网络的POMDP值迭代算法RQMDP-net。利用GRU网络与CNN实现值迭代过程,解决了仅由卷积层和最大池化层构成的QMDP-net训练不稳定、超参数设置敏感等问题,并且通过GRU网络的强时序处理能力,提升了RQMDP-net的算法运行速度。实验结果表明,与QMDP-net相比,RQMDP-net在训练过程中网络收敛速度更快,任务规划能力更强。后续可将RQMDP-net扩展至具有更复杂状态空间的导航规划任务中,进一步提高其适用性与通用性。

参考文献

[1] HU Bo, WANG Qiyao, FENG Hui, et al. Adaptive sensor scheduling algorithm for target tracking in wireless sensor networks[J]. Journal of Electronics and Information Technology, 2018, 40(9): 2033-2041. (in

Chinese)

胡波,王祺尧,冯辉,等. 一种无线传感器网络中目标跟踪的自适应节点调度算法[J]. 电子与信息学报, 2018, 40(9): 2033-2041.

[2] TESAURIO G. TD-gammon, a self-teaching backgammon program, achieves master-level play[J]. Neural Computation, 1994, 6(2): 215-219.

[3] LIU Feng, WANG Chongjun, LUO Bin. A probability-based value iteration on optimal policy algorithm for POMDP [J]. Acta Electronica Sinica, 2016, 44(5): 1078-1084. (in Chinese)

刘峰,王崇俊,骆斌. 一种基于最优策略概率分布的POMDP值迭代算法[J]. 电子学报, 2016, 44(5): 1078-1084.

[4] SILVER D, VENESS J. Monte-Carlo planning in large POMDPs[M]. Cambridge, USA: MIT Press, 2010.

[5] LITTMAN M L, CASSANDRA A R, KAEHLBLING L P. Learning policies for partially observable environments: scaling up[C]//Proceedings of International Conference on Machine Learning. New York, USA: ACM Press, 1995: 362-370.

[6] HAN Bing. The design and implementation of point-based POMDP policy iteration algorithm[D]. Nanjing: Nanjing University, 2014. (in Chinese)

韩冰. 基于点的POMDP策略迭代算法设计与实现[D]. 南京: 南京大学, 2014.

[7] LIU Yunlong, LI Renhou, LIU Jianshu. Q-learning algorithm based on predictive state representations[J]. Journal of Xi'an Jiaotong University, 2008, 42(12): 1472-1475. (in Chinese)

刘云龙, 李人厚, 刘建书. 基于预测状态表示的Q学习算法[J]. 西安交通大学学报, 2008, 42(12): 1472-1475.

[8] LIU Quan, ZHAI Jianwei, ZHANG Zongzhang. A survey on deep reinforcement learning [J]. Chinese Journal of Computers, 2018, 41(1): 3-29. (in Chinese)

刘全, 翟建伟, 章宗长. 深度强化学习综述[J]. 计算机学报, 2018, 41(1): 3-29.

[9] KARKUS P, HSU D, LEE W S. QMDP-Net: deep learning for planning under partial observability[EB/OL]. [2019-11-04]. <https://arxiv.org/abs/1703.06692>.

[10] YU Kai, JIA Lei, CHEN Yuqiang, et al. Deep learning: yesterday, today, and tomorrow[J]. Journal of Computer Research and Development, 2013, 50(9): 1799-1804.

[11] HAARNOJA T, AJAY A, LEVINE S, et al. Backprop KF: learning discriminative deterministic state estimators [C]// Proceedings of the 30th International Conference on Neural Information Processing Systems. New York, USA: ACM Press, 2016: 4376-4384.

[12] KIM W, LEE H, KIM H J. Predictive modeling of time-varying environmental information for path planning [C]// Proceedings of IEEE International Conference on Systems. Washington D. C., USA: IEEE Press, 2013: 3639-3644.

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

(下转第102页)

(上接第 94 页)

- [14] TAMAR A, WU Y, THOMAS G, et al. Value iteration networks[C]//Proceedings of International Joint Conference on Artificial Intelligence. Washington D. C., USA: IEEE Press, 2016: 26-31.
- [15] SHANI G, PINEAU J, KAPLOW R. A survey of point-based POMDP solvers[J]. Autonomous Agents and Multi-Agent Systems, 2013, 27(1): 1-51.
- [16] SONDIK E J. The optimal control of partially observable Markov processes over the infinite horizon: discounted costs[J]. Operations Research, 1978, 26(2): 282-304.
- [17] MURPHY K P. A survey of POMDP solution techniques[EB/OL]. [2019-11-04]. https://www.researchgate.net/publication/2275247_A_survey_of_POMDP_solution_techniques.
- [18] KOUTNÍK J, GREFF K, GOMEZ F, et al. A clockwork RNN[EB/OL]. [2019-11-04]. <https://arxiv.org/abs/1402.3511>.
- [19] PASCANU R, MIKOLOV T, BENGIO Y. On the difficulty of training recurrent neural networks[C]//Proceedings of the 30th International Conference on Machine Learning. Washington D. C., USA: IEEE Press, 2013: 1310-1318.
- [20] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Pittsburgh, USA: Association for Computational Linguistics, 2014: 1724-1734.
- [21] WERBOS P J. Backpropagation through time: what it does and how to do it[J]. Proceedings of the IEEE, 1990, 78(10): 1550-1560.

编辑 陆燕菲