基于深度强化算法的机器人动态目标点跟随研究

徐继宁 曾 杰
(北方工业大学电气与控制工程学院 北京 100043)

摘 要 机器人的路径规划一直是机器人运动控制研究的热点。目前的路径规划需要耗费大量时间来构建地图,而基于不断"试错"机制的强化学习通过预先的训练可以实现无地图条件下的路径规划。通过对当前的多种深度强化学习算法进行研究和分析,利用低维度的雷达数据和少量位置信息,最终确定了在不同智能家居环境下的有效动态目标点跟踪策略,同时完成了避障功能。实验结果表明,基于优先采样的 DQN、Dueling Double DQN 和 DDPG 算法,在不同环境下呈现较强的泛化能力。

关键词 强化学习,路径规划,目标跟随
中图法分类号 TP181 **文献标识码** A

Dynamic Target Following Based on Reinforcement Learning of Robot-car

XU Ji-ning ZENG Jie
(School of Electrical and Control Engineering, North China University of Technology, Beijing 100043, China)

Abstract Robot path planning has always been a hot topic in robot motion control. The current path planning takes a lot of time to build the map, but the reinforcement learning based on continuous "trial and error" mechanism can realize the mapless navigation. Through the research and analysis of current various deep reinforcement learning algorithms, using low-dimensional radar data and a small amount of position information can follow a moving target and avoid collisions in indoor environments. The results show that DQN、Dueling Double DQN and DDPG algorithms based on priority sampling present strong generalization capabilities in different environment.

Keywords Reinforcement learning, Path planning, Target following

1 引言

传统的机器人路径规划方法有模拟退火算法、人工势场法、A * 算法和 Dijkstra 算法等[1]。这些算法有很强的路径搜索能力,但不能很好地适应复杂多变的环境。随着人工智能的热潮,深度神经网络的应用使得端对端具有泛化能力的模型变成可能。其中,深度增强学习最贴近人类思维模式,是实现通用人工智能较为重要的一环。目前,基于深度学习的路径规划大部分都是以摄像头原始图像信息作为输入使机器学习导航和避障等功能,小部分则是以激光雷达数据作为输入。

强化学习的基本原理是指智能体 Agent 的行为策略与环境 Env 交互,通过奖赏 Reward 对行为策略进行修正。如果得到正的奖赏 Reward,那么这个行为策略的趋势就会加强。

强化学习算法的分类如图 1 所示。根据策略的更新和学习方法,可以分为基于值函数的强化学习方法、基于直接策略的搜索强化算法[2];基于值函数的算法主要着眼于 Q 函数的更新和训练。基于策略搜索算法的策略梯度(Policy Gradient, PG)[3-4]可以学习随机策略,是一种直接将策略参数化的优化方法。通过不断计算总奖励的期望值关于策略参数的梯度来进行策略更新,最终收敛于最优策略。这种直接在策略空间中搜索最优策略,省去了繁琐的中间环节,不需要用状态值或者状态动作值来选择动作,与深度 Q 网络(DeepQ-network, DQN)[5]及其改进模型相比,策略梯度搜索算法的适用范围更广,效果更好。但纯策略梯度搜索易收敛到局部最小值,且方差较大。针对此问题,Konda[6]和 Lillicra[7]陆续提出了深度确定性策略梯度(Deep Deterministic Policy Gradient, DDPG)算法,借鉴了 DQN 对 Q-learning 的扩展,用神经网络的方法来模拟策略函数 μ 和 Q 函数,并在 Actor-Critic 算法框架上应用了确定性策略梯度(Deterministic Policy Gradient, DPG)算法,成功地解决了连续动作的控制问题。

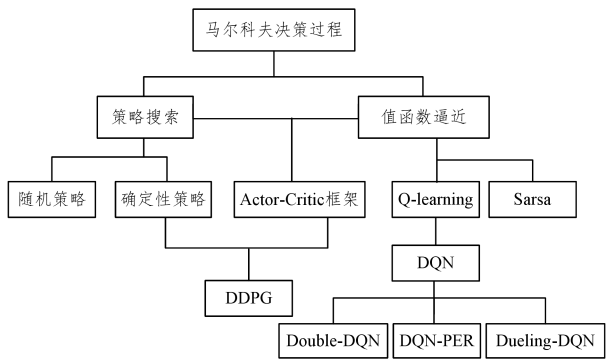


图 1 强化学习算法分类

本文在室内无地图环境下,利用激光雷达实现小车的自主运行。输入为 12 维的激光雷达数据和 2 维的目标点位置信息。通过建立的端对端模型对小车动作进行决策学习,对比离散动作输出的 DQN, Dueling-DQN[8]和连续动作输出的 DDPG 算法的实现效果,实现了小车对移动目标点的跟随和避障能力。

本文受北方工大科研专项项目(108051360018XN073)资助。

徐继宁(1970—),女,博士,副教授,主要研究方向为信号处理、工业控制和总线技术,E-mail:jxu0422@ncut.edu.cn(通信作者)。

2　算法对比

本文实现的算法主要包括基于值函数的 DQN，Double-DQN[9]，Dueling-DQN 和基于策略搜索的 DDPG。在设定环境中对效果进行了对比评价，并予以算法上的改进。

2.1　基于值函数的 DQN 算法

DQN 解决了当状态和动作空间是连续且维数较高时，Q-Table 难以储存每个状态、动作对的 Q 值问题。记忆回放单元(Experience Replay)将每个时间步 Agent 与环境交互得到的状态转移样本(S_t，a_t，r_t，S_{t+1})储存到经验池；用于学习阶段的样本训练。同时，构造两个结构相同的当前值网络(MainNet，θ)和目标值网络(TargetNet，θ')，解决了非线性网络表示值函数不稳定的情况，让目标 Q 值在一段时间内保持不变，一定程度上降低了当前 Q 值和目标 Q 值的相关性，提高了算法的稳定性。

Double-DQN 主要解决了 Q-learning 本身过估计的问题。分别用了不同的值函数来实现动作的选择和动作的评估。动作的选择是指将 S_t 作为网络的输入得到动作值函数 Q，然后选取值最大的 a^*，利用 a^* 处的动作值函数构造 TD 目标，一般情况下用同一个参数 θ 来选择和评估动作。而在 Double-DQN 中，选择动作值函数网络是基于 θ，而评估动作所用的动作值函数网络是基于 θ'。即在当前值网络中寻找动作值函数最大的动作 a^*，然后在目标值网络中选取 a^* 所对应的动作值函数。Double-DQN 的 TD 目标公式如下：

$$target^{Double\ Q}=R_{t+1}+\gamma Q(S_{t+1},\arg\max_a Q(S_{t+1},a;\theta_t);\theta_t') \tag{1}$$

Dueling-DQN 是由 Wang 等提出的一种竞争网络结构作为 DQN 的网络模型，主要是将 Q 值分解成两部分，并分别进行训练。Dueling-DQN 如图 2 虚线所示，Dueling-DQN 和 DQN 的输出都是 Q 函数的值，由全联接层得到。全联接层分成两部分：1)输出关于状态价值的标量 $\hat{V}$；2)输出关于动作的优势函数(Advantage Function)的矢量 $\hat{A}$。最后将两部分合起来作为 Q 价值函数，分解公式如下：

$$Q(s,a|\theta,\alpha,\beta)=\hat{A}(s|\theta,\beta)+\hat{A}(s,a|\theta,\alpha) \tag{2}$$

本实验使用基于优先采样的 Double Dueling DQN 算法，其总体控制图如图 2 所示。

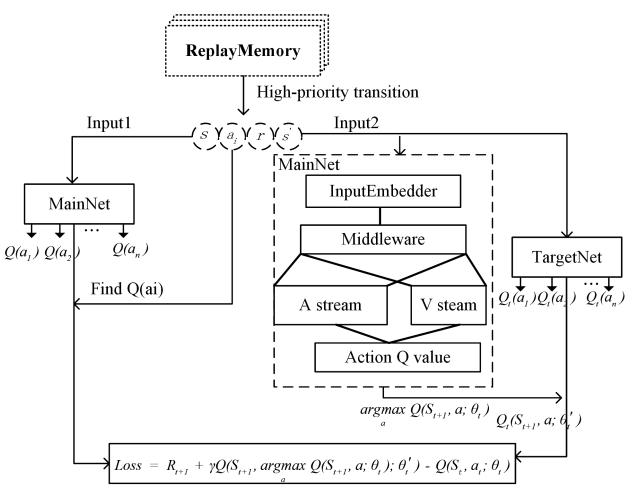


图 2　基于优先采样的 Double Dueling DQN

2.2　基于策略搜索的 DDPG 算法

DDPG 算法流程如图 3 所示，主要由 Actor-Critic 框架组成，包括 Q 网络和策略网络 μ；Q 网络是指用一个深度神经网络对 Q 函数进行拟合，记为 θ^Q，表示对 Actor 在每个时间步的评价；策略网络 μ 是指用神经网络对确定性行为策略 μ 函数进行拟合，每一时间步长的行为可以通过 $a_t=\mu(s_t)$ 获得，记为 θ^μ。根据以往研究者的实践经验，单个神经网络算法在路径规划学习过程中很不稳定，以至于学不到东西[10]，因此，DDPG 分别为 Q 网络和策略网络 μ 创建两个神经网络，分别记为目标网络 $\theta^{Q'}$ 和 $\theta^{\mu'}$，并每隔一定数量的步数对 Target 网络进行参数更新，即 soft-update；目标网络 TD error 损失函数如公式所示。总而言之，整个框架就是尽可能使 Actor 更新网络以获得更大的 Q 值。

$$y_i=r_i+\lambda Q(s_{i+1},\mu(s_{i+1}|\theta^{\mu'})|\theta^{Q'}) \tag{3}$$

$$L=\frac{1}{N}\sum_i(y_i-Q(s_i,a_i|\theta^Q))^2 \tag{4}$$

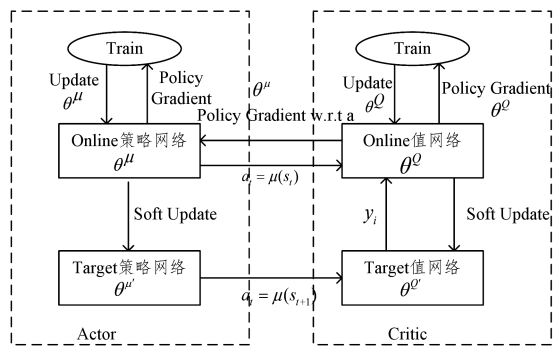


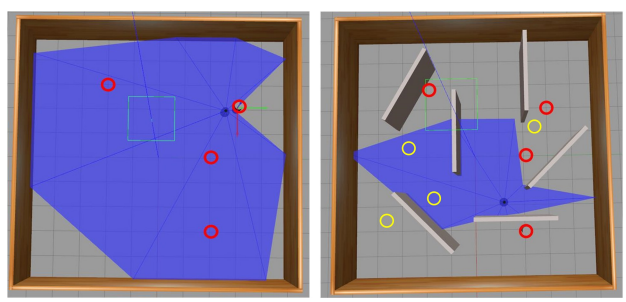
图 3　DDPG 算法流程图

2.3　优先采样

DQN 的成功归因于经验回放和独立的目标网络；无论是改变网络结构的 Dueling-DQN 或者是改进 Q-learning 中 max 操作的 Double-DQN，经验回放都采用均匀分布。然而，均匀分布采样并不会高效利用数据的分布。因此，有人提出了以 TD 偏差作为权重的优先采样方法，根据 TD 偏差来判断该状态的值函数与 TD 目标的差距，如果偏差越大，说明该处的学习效率就越高[11]。本文经过仿真实现的算法都是以优先采样为原则。

3　真环境和实验设计

算法仿真是基于 Gazebo 仿真平台。Gazebo 是一个强大的传感器模拟库，包括 camera，laser 和 IMU 等常见机器人传感器，可以通过添加这些传感器和自建地图，测试机器人小车 Turtlebot 在自建环境下的算法表现。本文设计了两类实验环境：无障碍物环境 Env1 和有障碍物环境 Env2，分别如图 4(a)，(b)所示。其中圆柱为小车要达到的目标点。



(a)Env1　　(b)Env2

图 4　仿真环境示意图(电子版为彩色)

设定规则为：小车从初始点出发，经过学习到达目标点。一旦小车达到目标点，目标点随机(此处的随机指的是多个固

定点以随机方式出现)更改位置,本文基于强化学习的小车路径规划主要是为了小车能够在智能家居环境下完成定点巡航,以实现回合内小车碰到障碍物或达到每回合所允许的最大步数为止。环境参数信息如表 1 所列。

表 1 环境参数信息

参数名	值	备注
GoalReward	500	到达目标点的奖励
CollisionReward	200	碰到障碍物的惩罚,为负值
DistanceReward	[0.1,12]	小车与目标点距离,为负值
AngleReward	[0,π]	小车与目标点的角度,为负值
LaserData	[0.1,12]	小车激光雷达测距数据
Action1	(v=0.5 m/s),(v=0.1 m/s),(w=0.2 rad/s),w=(−0.2 rad/s)	4 个离散动作线速度与角速度值
Action2	v=[0,1],w=[−1,1]	动作输出为区间内的任意值

每回合的奖励如式(5)所示,训练时随机障碍物出现位置如图 4 中用红圈所示,测试时的位置如图 4(b)黄圈所示。

$$reward_{total}=GoalReward+CollisonReward+DistanceReward+AngleReward \quad (5)$$

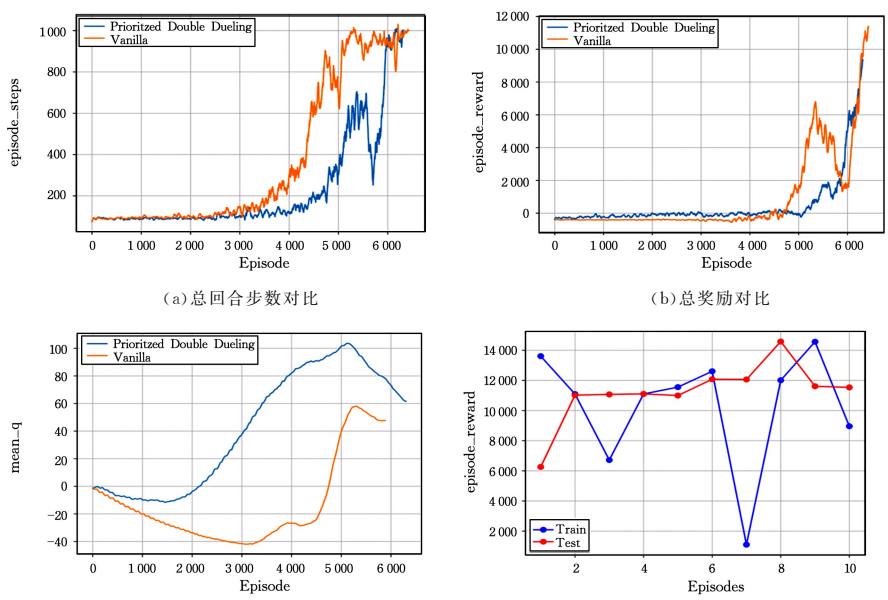
首先,在 Env1 下实现动态目标点跟随,使用的算法为 DQN 和 Prioritized Double Dueling DQN,待小车能够较好地实现动态目标点跟随;然后,在 Env2 下测试算法,发现 Prioritized Double Dueling DQN 在有障碍物环境下对动态目标点的跟随能力较差,仅能完成 1 个或 2 个目标点跟随,考虑到输出动作有限以及基于值的 Q-learning 算法效率较低,在 Env2 下采用了 DDPG 算法。

DDPG 隐藏层层数为 3 层,每层有 300 个神经元,激活函数为 Relu;Actor 模型中对于线速度的输出激活函数为 sigmoid,线速度范围限制在(0,1)之间,不考虑小车后退的动作;对于角速度的输出激活函数为 tanh,速度范围限制在(−1,1)之间;而 DQN 的隐藏层层数为 3 层,每层有 100 个神经元,激活函数为 Relu。

4 效果评价和分析

结果评估由两种方式得出:1)根据得分情况,即在算法收敛后,观察小车每回合的总奖励值;2)观察小车每回合内平均 Q 值。据此来评价不同算法对于 Q 的估计程度。多次更改随机目标出现的位置来检验泛化能力。如图 5(a)为小车在环境 1 下跟随动态目标时回合数与每回合步数的曲线图,小车最后都收敛于最大步数 1000,表明小车在环境 1 下不触碰边界的情况下一直保持着对目标点的跟随;接着,由图 5(b)观察小车所获得的总奖励值可知,Vanilla DQN(后记为 Vanilla)前期上升较快,相比于 Prioritized Double Dueling DQN(后记为 PDD)更快学到了一个较优策略,但得分并不是最大值。这主要是因为 Vanilla 对于 Q 值存在过估计问题,过估计量分布不均匀,导致 Q-learning 收敛于次最优策略;Double-DQN 和 Dueling DQN 对与 Q-learning 的改进则是为了解决过估计的问题;图 5(c)为小车在环境 1 下神经网络输出的平均 Q 值与回合数的曲线图,PDD 算法比 Vanilla 的 Q 值增加更快,采样效率更高。

图 5(d)为环境 1 下小车使用 Prioritized Double Dueling DQN 在随机目标位置训练集和测试集 10 回合内的总奖励曲线。由图可知,在环境 1 下小车对新的随机位置有较好的跟随能力,不错的泛化能力。



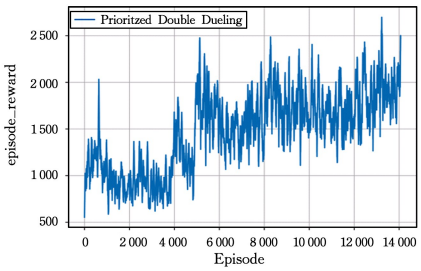
(a)总回合步数对比 (b)总奖励对比

(c)平均 Q 值对比 (d)训练集和测试集下总奖励对比

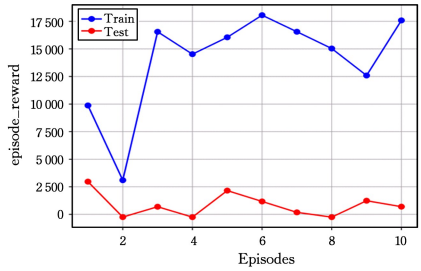
图 5 Env1 下实验数据图

图 6(a)为环境 2 下小车使用 DDPG 算法总奖励值和回合数的曲线,小车最终能跟踪 4～5 个目标点,且小车在复杂的环境 2 下算法前期能够迅速学到一个策略,使得分达到 1000 左右,但后期策略不太稳定。观察可知,对于某些随机

目标点，如处于障碍物正后方的目标点对于小车的跟随存在一定的难度。图 6(b)为环境 2 下训练集与测试集总奖励曲线，训练集中小车能达到较高的总奖励值，但个别点的奖励值过低，对于测试集则平均达到跟随 3～4 个随机目标点的结果。



(a)DDPG 总奖励曲线



(b)训练集和测试集下总奖励对比

图 6　Env2 实验数据图

结束语　本文将采用优先采样的 Vanilla DQN，Double Dueling DQN 和 DDPG 3 种算法作用于移动小车，分析了不同环境下小车对移动随机目标点的自主跟随能力。结果表明，在简单的环境 1 下，输出为离散动作的小车，仅仅利用较少的输入信息就能够学习到对移动目标的跟随能力，对于新的随机目标点也具有较好的泛化能力。在复杂的环境 2 下，相对于 DQN 和 DDPG，输出连续动作的 DDPG 算法，能够较快地学习到最优策略，较好地实现在部分环境信息未知下小车对随机移动目标点的跟随。但是在有障碍物的环境下，算法的泛化能力仍需要进一步提升。针对这个问题，未来的研究可以考虑改进小车避障能力和目标跟随能力的分开训练，或者对奖励值部分进行细分等方面。

参 考 文 献

[1] 王春颖，刘平，秦洪政. 移动机器人的智能路径规划算法综述[J]. 传感器与微系统，2018，37(8)：5-8.

[2] 刘全，翟建伟，章宗长，等. 深度强化学习综述[J]. 计算机学报，2018，41(1)：1-27.

[3] HASSELT H V，GUEZ A，SILVER D. Deep Reinforcement Learning with Double Q-learning[J]. Computer Science，2015.

[4] SILVER D，LEVER G，HEESS N，et al. Deterministic policy gradient algorithms[C]// InternationalConference on International Conference on Machine Learning. JMLR. org，2014：387-395.

[5] MNIH V，KAVUKCUOGLU K，SILVER D，et al. Playing Atari with Deep Reinforcement Learning[J]. Computer Science，2013.

[6] KONDA V. Actorcritic algorithms[J]. Siam Journal on Control & Optimization，2003，42(4)：1143-1166.

[7] LILLICRAP T P，HUNT J J，PRITZEL A，et al. Continuous-control with deep reinforcement learning[J]. Computer Science，2015，8(6)：A187.

[8] WANG Z，SCHAUL T，HESSEL M，et al. Dueling network architectures for deep reinforcement learning[J]. arXiv：1511.06581，2015.

[9] HASSELT H V，GUEZ A，SILVER D. Deep Reinforcement Learning with Double Q-learning[J]. Computer Science，2015.

[10] 郭宪，方勇纯. 深入浅出强化学习原理入门[M]. 北京：电子工业出版社，2018：125-141.

[11] SCHAUL T，QUAN J，ANTONOGLOU I，et al. Prioritized Experience Replay[J]. Computer Science，2015.

(上接第 79 页)

[8] LI L L，ZHU W N，YU C，et al. Esports analysis data acquisition algorithm based on convolutional neural network[C]// MATEC Web of Conferences. EDP Sciences，2018，189：03003.

[9] 于诚，朱皖宁. 基于战场热点图的 MOBA 类游戏战术分析研究[J]. 计算机科学，2018，45(S2)：149-151，175.

[10] VINYALS O，BLUNDELL C，LILLICRAP T，et al. Matching networks for one shot learning[C]// Advances in Neural Information Processing Systems. 2016：3630-3638.

[11] SANTORO A，BARTUNOV S，BOTVINICK M，et al. Meta-learning with memory-augmented neural networks[C]// International Conference on Machine Learning. 2016：1842-1850.

[12] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017：5998-6008.

[13] 李海林，郭崇慧. 时间序列数据挖掘中特征表示与相似性度量研究综述[J]. 计算机应用研究，2013，30(5)：1285-1291.

[14] 马宏伟，张光卫，李鹏. 协同过滤推荐算法综述[J]. 小型微型计算机系统，2009，30(7)：1282-1288.

[15] SAHU S K，ANAND A. Drug-Drug Interaction Extraction from Biomedical Text Using Long Short Term Memory Network[J]. Journal of Biomedical Informatics，2017：S1532046418301606.