文章编号：2095-2163(2023)05-0018-06 中图分类号：TP181 文献标志码：A

停车场车位自动化排布设计及优化

许哲阳，张易诚，沈 炜
（浙江理工大学 计算机科学与技术学院，杭州 310018）

摘 要：用智能的方式来设计停车场的车位排布一直都是工程上的难题。传统的启发式算法和基于深度强化学习的方法（DQN）均存在数据利用率低下的问题。本文提出了一种基于改进 DQN 算法的停车场自动车位排布方法，该方法结合前人的经验，设计了合理的奖励函数，并使用了一种多步提取算法（EnDQN）训练强化学习智能体。通过实验确定了一组较优的超参数，并验证了该方法在提高数据利用率上的有效性，为后续研究提供了参考。

关键词：停车场；车位排布；深度强化学习；多步提取算法；数据利用率

Design and optimization of automatic layout of parking spaces

XU Zheyang, ZHANG Yicheng
（School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China）

【Abstract】 Designing car parking space scheduling in an intelligent way has always been an engineering challenge. Both traditional heuristic algorithms (dynamic planning, evolutionary strategies) and deep reinforcement learning-based methods (DQN) suffer from low data utilization. To this end, this paper proposes a method for automatic parking lot scheduling based on an improved DQN (Deep Q-Network) algorithm. The method combines previous experience, designs a reasonable reward function, and uses a multi-step extraction algorithm (EnDQN) to train reinforcement learning intelligences. Through experiments, we identify a better set of hyperparameters and verify the effectiveness of the method in improving data utilization for subsequent research.

【Key words】 parking lot; parking space layout; deep reinforcement learning; multi step extraction method; data utilization

0 引 言

目前大多数关于车位排布问题的研究都是将其看成一个带有约束条件的单一尺寸矩形排样问题，大都采用启发式算法，如：模拟退火算法、动态规划、遗传算法、进化策略等进行优化[1]。Huang 等[2]设计了一套基于模拟退火算法的针对矩形区域内车位排布问题的通用算法；利润等[3]提出使用两段动态规划进行车道布局的算法；徐涵喆等人[4]提出了使用遗传算法来解决地下车库外圈车位排布的问题；但这些研究都仅考虑到了车位数最优的情况。为了加强便捷度的设计，余光鑫等人[5]采用进化策略，同时提出增加一个便捷度的评价体系来设计停车场。虽然得到了较好的实验效果，但是进化策略属于启发式算法，通常依靠猜测和搜索来解决问题，数据利用率不高。如进化策略仅仅将当前生成的数据用作一次策略更新，待策略更新结束后就将当前生成的数据丢弃，数据利用率十分低下[6]。采用深度强化学习方法，如 DQN（Deep Q-Network），使用经验回放，将每次从环境中采样得到的四元组数据（状态、动作、奖励、下一状态）存储到回放缓冲区中，训练网络的时候再从回放缓冲区中随机采样若干数据来进行训练，这样不仅可以使样本满足独立假设，还可以提高数据利用率。此外，随着大型综合体对客户体验越来越重视，停车场的设计目标越来越倾向于便捷度优先而兼顾停车位数量，这样的设计方式会更加依赖经验数据的使用，所以深度强化学习方法相较于启发式算法更适合去解决车位排布这种对数据利用率有要求的问题，但目前可供参考的基于深度强化学习方法来解决车位排布问题的文献较少。

使用深度强化学习方法仍然有其局限性，随着环境复杂性的增加，DQN 中智能体需要大量的时间和数据来学习[7]。为此，Yinlong Yuan 等[8]提出的一种新的多步提取方法，是针对深度强化学习做出

作者简介：许哲阳（1996-），男，硕士研究生，主要研究方向：强化学习；张易诚（2001-），男，本科生，主要研究方向：强化学习；沈 炜（1973-），男，博士，教授，硕士生导师，主要研究方向：智能控制、云计算。
通讯作者：沈 炜 Email：latitude@126.com

收稿日期：2023-02-03

的一个改进算法，与传统的多步 Q-learning 不同，多步提取方法使用了一种新的回报函数，即 n 个标准返回函数的平均值。在选择当前状态动作时，新的回报函数改变了未来奖励的折扣，同时减少了即时奖励的影响，多步方法更适合结合经验重放来提高经典 DRL(deep reinforcement learning)算法的性能。

本文尝试基于优化后的深度强化学习方法，在车位排布的约束条件、初始化条件、奖励设计、网络结构、算法选择等方面做出改进，以此来解决便捷度优先的车位自动化排布问题。

1　问题描述

1.1　问题定义

本文的车位自动化排布问题定义为：智能体在一个虚拟的车位场景中计算出一条最优的道路，这条道路可以让停车场内道路便捷度最高，在此基础上，车位数也要求尽可能多。

1.2　约束条件

停车场车位自动化排布的约束条件可以概括为以下 3 点：

(1)智能体计算得出的道路不能超出虚拟环境所设定的边界；

(2)道路与道路之间必须保证连通性，确保每条道路能到达除本身外任意一条道路；

(3)形成车位的空地必须满足与道路相邻的条件。

2　模型设计

2.1　网络结构

网络架构如图 1 所示，是有两个隐藏层的全连接网络，第一个隐藏层有 128 个节点，第二个隐藏层有 64 个节点。

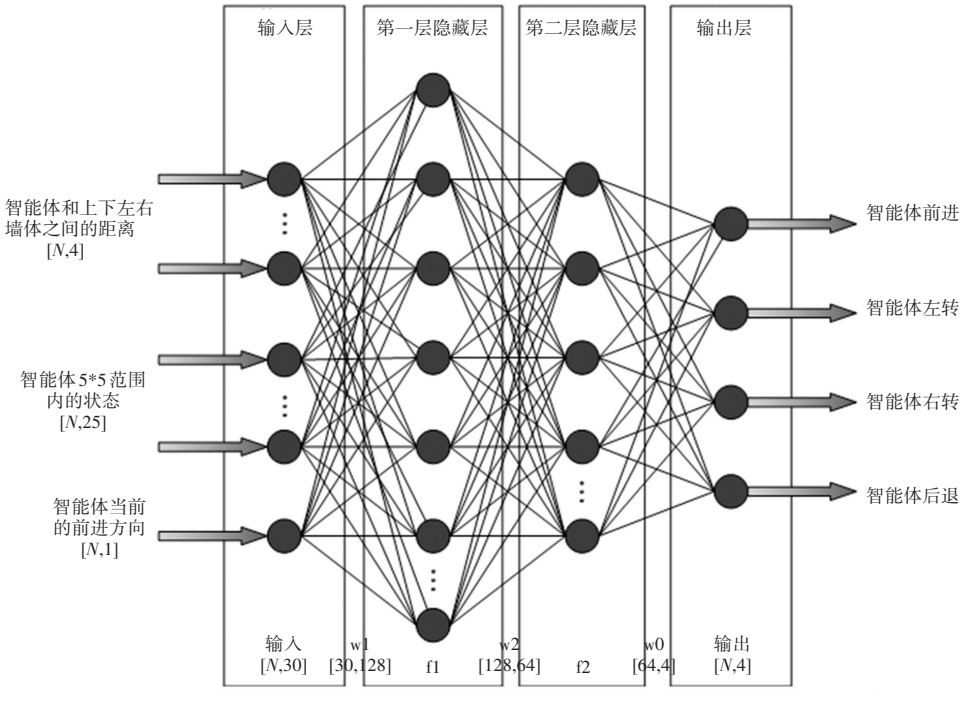

图 1　神经网络架构

Fig. 1　Neural network architecture

模型输入包括 3 部分：

(1)智能体和上下左右墙体之间的距离；

(2)以智能体为中心的 5×5 范围内的停车场状态信息；

(3)当前智能体所前进的方向信息，初始时设置为向上。

输出层则包含了智能体在当前位置可以选择的 4 个动作：

①当输出为 0 时，智能体策略为朝当前方向继续前进；

②当输出为 1 时，智能体策略为在当前方向的基础上左转；

③当输出为 2 时，智能体策略为在当前方向的基础上右转；

④当输出为 3 时，智能体策略为后退。

2.2　深度强化学习算法选择

本文使用深度强化学习中的 DQN 算法来取代进化策略结合神经网络的方法。采用进化策略和神

经网络结合的方法通常是不稳定的,甚至是发散的,而 DQN 相比于进化策略结合神经网络的方法优势在于 DQN 具有经验回放的功能,通过将经验数据 $e_t=(s_t,a_t,r_t,s_{t+1})$ 存储在 DQN 的经验重放内存 D 中,深度强化学习智能体可以记忆和重用过去的经验,其中 t 是时间节点,在该时间节点下,s_t 是状态,a_t 是智能体选择的动作,s_{t+1} 是 $t+1$ 时间节点的状态,r_t 是从 s_t 过渡到 s_{t+1} 获得的奖励。在训练过程中,任意两个元组 $e_{t1}\in D$ 和 $e_{t2}\in D$ 之间的训练数据弱相关。神经网络的参数是通过随机从 D 中均匀采样小批量经验来学习的,这有助于打破强相关更新,保证学习系统的稳定性[9]。同时用 DQN 中损失函数对训练过程进行改善,从而持续减少输出值的较大偏差,公式(1):

$$L(\boldsymbol{\theta})=\boldsymbol{E}_{(s,a,r,s')\sim U(D)}[(y^{DQN}-Q(s,a;\boldsymbol{\theta}))^2] \tag{1}$$

其中,$\boldsymbol{\theta}$ 是有限维权向量;$\boldsymbol{E}[\cdot]$ 是期望函数;s 是当前环境的状态;a 是智能体选择的动作;s' 是新的状态;r 是从当前状态过渡到新状态获得的奖励;$U(\cdot)$ 是随机抽样函数;$Q(\cdot)$ 是参数连续化函数。

用 DQN 代替进化策略的方法不仅可以进一步提高训练速度,而且可以显著提高模型性能。

本文还使用了 EnDQN 技术进一步优化 DQN 算法。与传统的强化学习方法不同,EnDQN 引入了一个新的回报函数,定义公式(2):

$$\hat{G}_t^{(n)}=\frac{1}{n}\sum_{k=1}^{k=n}G_{t+k-1} \tag{2}$$

即新的回报函数是 n 个标准返回函数 $G_t,G_{t+1},\cdots,G_{t+n-1}$ 的平均值。根据数学推导,新的回报函数可以分解为两部分,分解后表示为公式(3):

$$\hat{G}_t^{(n)}=\frac{1}{n(1-\gamma)}\sum_{k=1}^{k=n}(1-\gamma^k)r_{t+k-1}+\frac{1-\gamma^n}{n(1-\gamma)}\sum_{k>n}\gamma^{k-n}r_{t+k-1}=\hat{g}_1+\hat{g}_2 \tag{3}$$

第一部分是 $\hat{g}_1$,包含前 n 个未来奖励,$\hat{g}_1$ 这部分与传统回报函数有着本质区别,在选择当前状态行为时,该函数改变了未来奖励的折扣,同时减少了当前奖励的影响,即前 n 项未来奖励的重要性得到强化。在最后 n 步,回报函数 $\hat{G}_t^{(n)}$ 和 G_t 在参数 $\{(\gamma,k)\}$ 下的对比如图 2 所示[8],表明对于任意一组参数 $\{(\gamma,k)\}$,$\hat{G}_t^{(n)}$ 的值小于 G_t,说明新的回报函数可以增强负奖励($r_T=-1$)的影响,并有效地降低回报函数的值,所以智能体在每个训练步骤中都会加上一个负奖励($r_T=-1$),用于增强负奖励的影响。新回报函数的第二部分是 $\hat{g}_2$,其包括前 n 步之后的奖励。这些奖励的折扣逐渐减小,这与传统的强化学习算法一致,有利于系统的稳定性[10]。

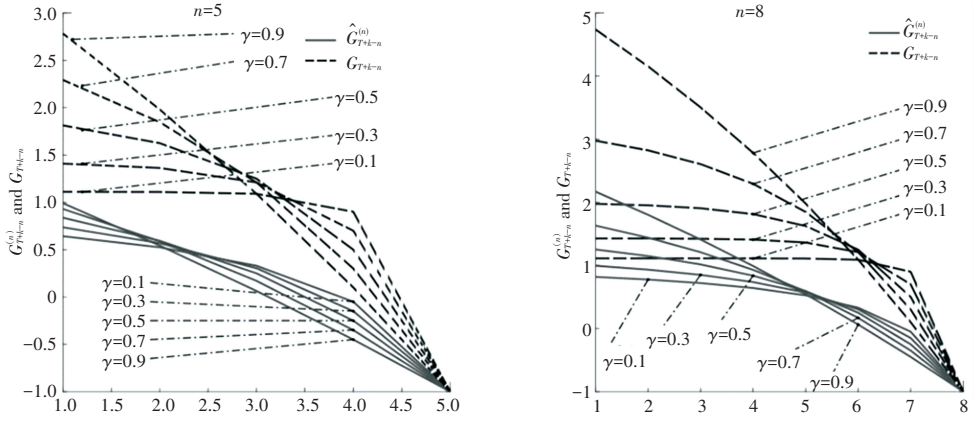

图 2 最后 n 步回报函数 $\hat{G}_t^{(n)}$ 和 G_t 在参数 $\{(\gamma,k)\}$ 下的对比图[8]

Fig. 2 In the last step, the comparison diagram of return function $\hat{G}_t^{(n)}$ and G_t under $\{(\gamma, k)\}$ parameter[8]

2.3 奖励设计

在奖励设计上,本文提出了一种高效的六车位算法,可以用极低的时间复杂度计算出虚拟地图中便捷度最高的道路,完全取代了复杂便捷度计算方法,在时间效率上得到了质的提升。该方法具体描述如下:

当智能体经过的道路周边有空地,且空地能满足形成一个最便捷的六车位组如图 3 所示,即智能体所经过道路所围成的空地刚好是一个六车位组,就给一个+6 的奖励,本文给此奖励乘上一个系数 ω 后加到总奖励中。由于+6 的奖励不方便对总奖励进行车位数换算,因此 ω 设置为 5/3,且在每一个训练模型步数结束时,当前所得到的总奖励会减去上一个训练模型步数中记录的总奖励来得到一个奖励

差，用来及时提供每一个训练模型步数的反馈，从而提高下一个训练模型步数的效率。

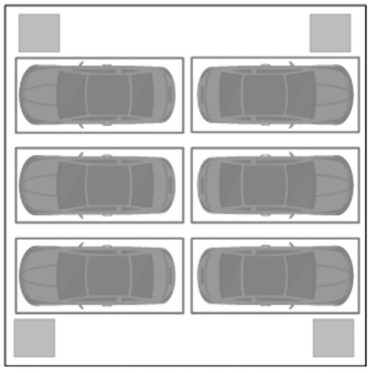

图 3　六车位组

Fig. 3　Six parking lot group

奖励差算法伪代码如下：

初始化训练模型步数 t

初始化期望训练模型的总步数 T

初始化存储上一次训练模型步数的总奖励 pre

初始化车位组奖励系数 ω

For $t = 1$ to T do

初始化奖励总数 $reward$

$reward = \omega \times \sum$ 满足要求的六车位组 $- pre$

If $t < T$ then

$pre = reward$

$reward - 1$

End

由于 EnDQN 算法更适合在负奖励中进行训练，所以每一个训练模型步数结束后会给一个-1 的奖励，尽可能让 800 个训练模型步数结束后奖励为负，可以增强负奖励的影响。

3　实验

3.1　仿真环境

本文使用用于自动化车位排布的环境 carEnv，将地图的网格数从 10×10 和 20×20 改成 21×21，可以让智能体计算出的车道完美贴合虚拟地图边缘，使最终效果更美观，暂时忽略环境中的障碍物，完全把停车场当成一个空地，且用 21×21 的网格展示出来。在这个环境中，存在一个智能体，为了提高样本的多样性，让智能体在初始化时随机选择虚拟地图边缘的一个网格作为起始点，提升样本随机性，从而大幅节约随机生成虚拟地图的时间成本。开始训练后，在每个时间步骤中，智能体有 4 个可能的动作：0，1，2，3，分别表示为智能体以当前方向前进、向左转、向右转以及向后退，当时间步骤大于等于 800 时退出。

本文的模型均在 11th Gen Intel（R）Core（TM）i7-1165G7 @2.80 GHz 1.69 GHz 上进行训练。

3.2　实验验证

在进化策略结合神经网络实验中，种群大小设置为 50，σ 为 0.1，学习率设置为 0.05，总代数为 800 代。平均奖励基线图如图 4 所示。

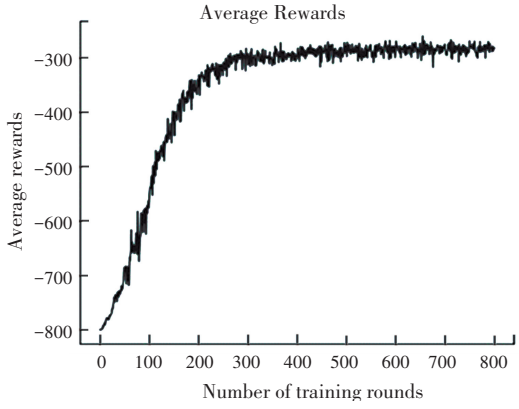

图 4　进化策略（ES）+神经网络（NN）的平均奖励基线图

Fig. 4　Average reward baseline of evolutionary strategy（ES）+ neural network（NN）

在 DQN 算法和 EnDQN 算法实验中，学习率设置为0.000 1，经验重放内存的大小是40 000元组。对内存进行采样，以每一步更新网络，批处理大小为 64。每次目标网络更新之间的步数是 800。在多步学习中，n 的值是一个敏感的超参数，本文只给出了 $n \in [1,2]$ 的结果，$n = 1$ 表示 DQN 算法，$n = 2$ 表示 EnDQN 算法。

DQN、EnDQN 和进化策略结合神经网络算法的实验中得到的平均奖励基线比较如图 5 所示。

EnDQN 算法的平均奖励会明显优于进化策略和神经网络的结合，同时在训练超过 500 轮后其基线显著优于 DQN 算法，这主要是因为进化策略本身就是一个数据利用率较差的方法，而作为经典深度强化学习算法之一的 DQN 算法又过度依赖于经验重放，在每个数据元组中，只考虑单个奖励信号，这同样导致其学习过程中存在数据利用率较低的问题，特别是当奖励信号稀疏时。而 EnDQN 方法使用了一个新的回报函数，改变了未来奖励的折扣，在选择当前状态行为时不再强调当前奖励是主要影响因素。因此在每次更新中，可以使用多步奖励信号，使得所选择的奖励信号更加有效。但当参数 n 较大时，每个元组 $\hat{e} \in D$ 内的数据相关性将变得更弱，

EnDQN 算法的性能将被削弱,因此选择一个合适的参数值是很重要的。

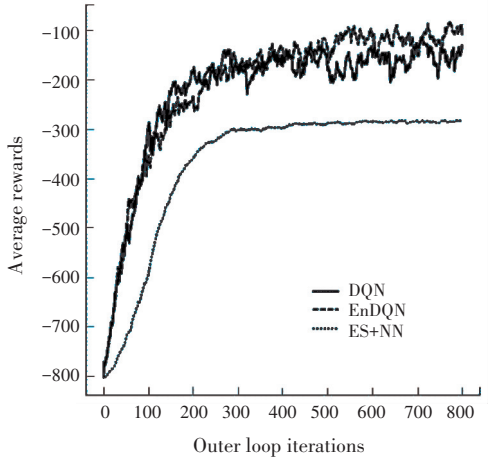

图 5 DQN、EnDQN 和进化策略(ES)+神经网络(NN)的平均奖励基线比较图

Fig. 5 Comparison chart of average reward baseline of DQN, EnDQN and evolutionary strategy (ES)+neural network (NN)

各方法每轮训练的耗时、训练过程中的最高奖励值、利用训练 800 轮的模型进行推理所得到的空地格子数和车位数量见表 1。

表 1 算法对比表

Tab. 1 Comparison table of algorithm

算法	每轮耗时/s	总用时/h	最高奖励值	满足奖励空地格子数	车位数
进化策略+神经网络	1 分 24	13.5	−221	58	348
DQN	24.43	5.43	−21	78	468
EnDQN	**23.81**	**5.29**	**59**	**86**	**516**

从表 1 中可以看出,在时间方面,同样的奖励设计下,EnDQN 算法的运行时间快于进化策略和神经网络的结合,很大程度上减少工程上的时间消耗,尤其是在环境非常复杂的现实工程问题中;同时根据 EnDQN 算法经验来看,实验效果理论上要比 DQN 的效果高出较多,但是在本文实验的结果来看,效果只是有提升,并没有特别大的改善。

EnDQN 在 800 轮训练中得到的道路空地效果如图 6 所示,预期可达到的效果如图 7 所示,不难看出图 6 与图 7 十分接近,可见本文使用的 EnDQN 算法在地下车库车位规划问题上取得了较好的效果。

本文将 EnDQN 在 800 轮训练中预期得到的效果图转换成真实场景如图 8 所示。

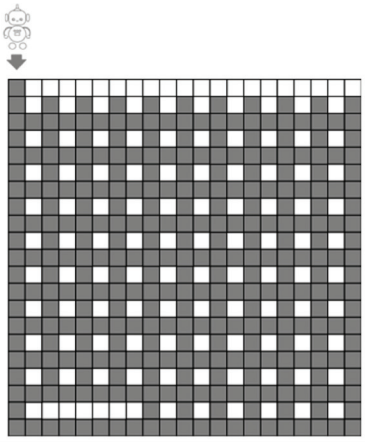

图 6 EnDQN 在 800 轮训练中得到的道路空地效果图

Fig. 6 EnDQN′s effect picture of road open space obtained in 800 rounds of training

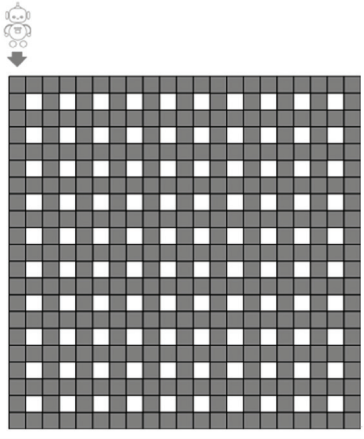

图 7 EnDQN 在 800 轮训练中预期可达到的效果图

Fig. 7 EnDQN′s expected effect in 800 rounds of training

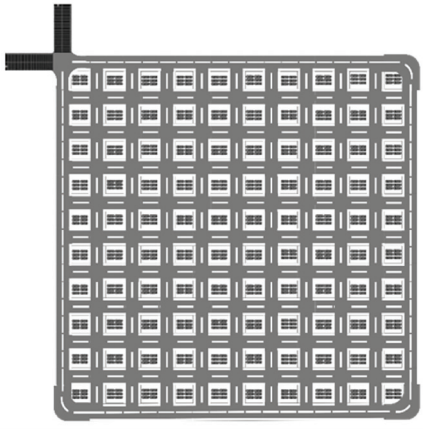

图 8 EnDQN 在 800 轮训练中预期得到的真实场景图

Fig. 8 EnDQN′s expected real scene renderings in 800 rounds of training

4 结束语

本文设计了一种新的停车场车位自动化排布模

(下转第 31 页)