文章编号:1671-4598(2022)11-0226-07 DOI:10.16526/j.cnki.11-4762/tp.2022.11.033 中图分类号:TP242.6 文献标识码:A

基于深度Q网络和人工势场的移动机器人路径规划研究

王冰晨，连晓峰，颜 湘，白天昕，董兆阳
（北京工商大学 人工智能学院，北京 100048）

摘要：随着移动机器人在各个领域的研究与发展，人们对移动机器人路径规划的能力提出了更高的要求；为了解决传统的深度Q网络算法在未知环境下，应用于自主移动机器人路径规划时存在的收敛速度慢、训练前期产生较大迭代空间、迭代的次数多等问题，在传统DQN算法初始化Q值时，加入人工势场法的引力势场来协助初始化环境先验信息，进而可以引导移动机器人向目标点运动，来减少算法在最初几轮探索中形成的大批无效迭代，进而减少迭代次数，加快收敛速度；在栅格地图环境中应用pytorch框架验证加入初始引力势场的改进DQN算法路径规划效果；仿真实验结果表明，改进算法能在产生较小的迭代空间且较少的迭代次数后，快速有效地规划出一条从起点到目标点的最优路径。

关键词：路径规划；DQN；人工势场；栅格地图；pytorch

Research on Path Planning of Mobile Robot Based on Deep Q-Network and Artificial Potential Field

WANG Bingchen, LIAN Xiaofeng, YAN Xiang, BAI Tianxin, DONG Zhaoyang
(School of Artifical Intelligence, Beijing Technology and Business University, Beijing 100048, China)

Abstract: With the research and development of mobile robots in various fields, people put forward higher requirements for the ability of mobile robot path planning. In order to solve the problems of slow convergence speed, many iterations, and large iteration space in the early stage of training when the traditional deep reinforcement learning algorithm is applied to the path planning of mobile robots in an unknown environment, an artificial potential field is added when the traditional deep q-learning network (DQN) algorithm initializes the Q value. The attractive field of the algorithm is used as the prior information of the initial environment, and then the mobile robot is guided to move towards the target position, the algorithm reduces many invalid iterations caused by the environmental exploration in the initial stage of the algorithm, thereby reduces the number of iterations and speeds up the convergence speed. The pytorch framework in the grid map environment is used to verify the path planning effect of the improved DQN algorithm in the initial gravitational potential field. The simulation results show that the improved algorithm can quickly and effectively plan an optimal path from starting point to target point after generating smaller iteration space and fewer iterations.

Keywords: path planning; DQN; artificial potential field; grid map; pytorch

0 引言

伴随人工智能技术的兴起，移动机器人正在朝着自探索、自学习、自适应的方向发展[1]。行成路径的策略被称为路径规划。作为自主移动机器人运动规划的主要探究内容之一，路径规划的目的是通过环境感知与主动避障功能寻找一个合理的最优路线，该道路由起点到目标点在既定的目标距离内不和其他障碍物交叉，并且要调整机器人移动方向，使之尽量满足更短、更平缓的条件。依照自主移动机器人在它运行环境内信息的剖析水平，路径规划可分两种类别，即局部路径规划和全局路径规划。路径规划效果的优劣会立刻裁夺自主移动机器人完成任务的时效性和质量好坏，而机器人路径规划的核心内容是算法的设计[2]。常用的移动机器人路径规划算法有人工势场法[3]、粒子群优化算法[4]、模糊逻辑算法[5]、遗传算法[6]等。而这类常用的方式多数需依据环境建模，往往需求预先构造出地图相关信息，然后对环境路径做好控制规划。一旦建立出了不准确模型，必将恶化过程与结果的实时性和准确性，乃至将会影响移动机器人及其周边环境的安全性。按照训练方法的不同，路径规划可分成监督学习、无监督学习和强化学习几类方式，在实施路径规划时监督学习需要提前建立出

收稿日期:2022-09-14； **修回日期**:2022-09-16。
项目基金:国家级大创项目(G014)。
作者简介:王冰晨(2000-),男,河南新安人,大学本科,主要从事路径规划等方向的研究。
通讯作者:连晓峰(1977-),男,山西长治人,博士,副教授,主要从事智能控制与模式识别等方向的研究。
引用格式:王冰晨,连晓峰,颜 湘,等.基于深度Q网络和人工势场的移动机器人路径规划研究[J].计算机测量与控制,2022,30(11):226-232,239.

环境信息，需要提供大量的先验知识，不然将不能完成满足需求的路径规划。无监督学习同监督学习一样。与此相反，强化学习不必提前了解环境信息，换句话说即无须先验知识，所以这种学习方式被普遍地使用于智能移动机器人路径规划中。强化学习的智能体将会和环境持续不断试错和交互，并经由累积奖励于改良策略，这是类于周围状况映照至行为的学习方法[7]，它把学习当成是一个“试探—评价”的进程。Q 学习（Q-learning）是强化学习中的一种常用的基础模型，并且不需要了解具体模型就能够确保最终收敛，它也为当前运用到智能自主移动机器人路径规划的特别见效的模型之一，当于状态空间相对小的情境下可以轻松地取得满意的路径规划结果[8]，得到满意的相关参数，该算法是经由搭建 Q 值表来挑选最佳策略的，当维度比较高的时候，这将引起维数灾难[9]。

深度强化学习（DRL，deep reinforcement learning）算法将强化学习的决策能力与深度学习强大的感知能力融合到一起，在应对不同情境的工作中表现优秀，这非常利于移动机器人的自主路径规划或导航避障。深度 Q 网络（DQN，deep q-learning network）是众多深度强化学习算法中的非常典型的常用算法，它是一个贴近人们思考逻辑的人工智能算法，核心就是把 Q－table 的更新转化为函数问题，通过拟合一个 function 来代替 Q-table 产生 Q 值。Mnih 等人[10]提出了 DQN 技术，并将它使用到 Atari2600 中，在游戏内达到人类玩家甚至超越人类玩家的水准。Schaul 等人[11]提出训练 DQN 模型的方法为根据优先级的经验回放方式取代原本的同概率选择方式，优先回放的核心思路是给经验池里的经验分别规定优先级，之后在选择经验时会偏向于挑选优先级很高的经验。Xin 等人[12]在进行移动机器人路径规划时，第一次使用到了 DQN 算法。

为解决 DQN 算法在路径规划上存在的收敛速度慢，训练前期产生巨大迭代空间的问题，本文在现有的路径规划算法和深度 Q 网络基础上，提出一种基于深度强化学习和人工势场法融合的路径规划算法。

1　人工势场

人工势场法（APF，artificial potential field）路径规划是由 Khatib[13]提出的一种虚拟力算法，是局部路径规划里面频繁被用到的算法，物理学里面的势，也被称为“位”，为一种能量的概念。把经典力学里的“场”的思想加到此算法里面，假设使移动机器人于此类虚拟力场里实施运动动作。究竟怎么规划势场会影响此方法的实用性和性能。它的核心概念为把自主移动机器人在周围现实情境中的运动设想成一个在宽泛的人工力场中的运动，利用目标物对自主移动机器人形成“引力”影响，障碍物对自主移动机器人形成“斥力”影响，最终再利用求出的合力来限制自主移动机器人的运动，如图 1 所示，是人工势场法中对环境下自主移动机器人的受力分析。而应用势场法设计出来的运动路线，通常是是相对平滑而且安全的[14]。

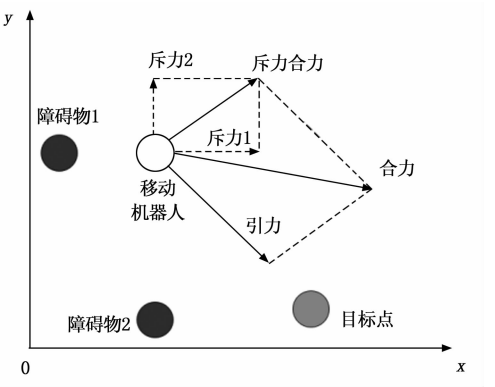

图 1　人工势场法中机器人的受力分析

最常见的引力场函数（Attractive Field）如公式（1）所示：

$$U_{att}(q)=\frac{1}{2}\zeta d^2(q,q_{goal}) \tag{1}$$

其中：ζ 为引力增益，d（q，q_{goal}）是当前点 q 与目标点 q_{goal} 之间的欧几里得距离。引力场有了，那么引力就是引力势场对距离的负导数：

$$F_{att}(q)=-\nabla U_{att}(q)=\zeta(q_{goal}-q) \tag{2}$$

最常见的斥力场函数（Repulsive Potential）如公式（3）所示：

$$U_{rep}(q)=\begin{cases}\frac{1}{2}\eta\left(\frac{1}{D(q)}-\frac{1}{Q^*}\right)^2 & ,D(q)\leqslant Q^*\\ 0 & ,D(q)>Q^*\end{cases} \tag{3}$$

其中：η 是斥力增益，D（q）是目前点 q 与相离最近的障碍物点之间的欧几里得距离，Q^* 是障碍物出现斥力影响的距离阈值，大于此阈值距离的障碍物不会产生排斥力作用来影响移动机器人。同理，斥力为：

$$F_{rep}(q)=-\nabla U_{rep}(q)=$$
$$\begin{cases}\eta\left(\frac{1}{D(q)}-\frac{1}{Q^*}\right)\frac{1}{D^2(q)}\nabla D(q) & ,D(q)\leqslant Q^*\\ 0 & ,D(q)>Q^*\end{cases} \tag{4}$$

将斥力场与引力场叠加，就形成了人工势力场：

$$U(q)=U_{att}(q)+U_{rep}(q) \tag{5}$$

$$F(q)=-\nabla U(q) \tag{6}$$

在排斥力势场和引力势场合力的驱动下，将移动机器人由高势能位置移动到低势能位置，同时找到一条能够到达目标点位置的无碰撞路径。地图上目标位置的引力（即重力）涵盖了整个环境地图，2D 空间的引力场示意图如图 2 所示，因此自主移动机器人可以从地图上的任何位置向目标点位置进行移动。

人工势场法如同搭建了类似吸铁石的场景，里面容纳了引力场与斥力场。深色物体产生斥力，是障碍物，箭标指向是移动机器人接下来运行的方向。智能移动机器人依据箭标的指向抵至目标物，目标物有类似于“引力”似的招至移动机器人与它的距离减小。但于障碍物周围，移动机器人反着箭标的方向，类似于对机器人形成“斥力”。移

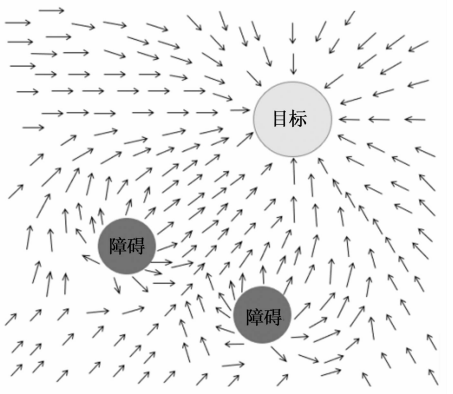

图 2 2D 空间引力场

动机器人移动的方向即斥力和引力的合力的指向。

人工势场法也存在诸多问题，比如在移动机器人距离目标位置相对远的时候，引力会变的非常大，且比较小的斥力在乃至能够被无视的情境下，移动机器人在路径上大概率将碰到障碍物；在目标点位置旁边有障碍物时侯，斥力会变得特别大，而引力特别小，移动机器人抵达目标点位置将会变得很麻烦；当环境中的一点，引力和斥力正好完全相等，方向相反，那么移动机器人将很轻易地深陷震荡或局部最优解。

对于上述存在的问题来说，接触到障碍物的问题，能够经由修改引力函数来处理，阻止遇到距离目标点位置过于远从而出现引力太大的情况；目标点位置旁边有障碍物从而引起目标不可达的问题，能够经由引进其它斥力函数来处理，此函数增添目标点和移动机器人距离的作用，一定程度而言，当移动机器人在目标点附近时，即使斥力作用变大，但与此同时距离在变小，所以此函数的增添能够产生对斥力场的拖拽影响；陷入局部最优解和震荡是人工势场法的一大难题，能够经由增添一个随机扰动，从而使得移动机器人脱离局部最优的情况。

2 深度强化学习

深度学习和强化学习的联结是人工智能领域的一个必然发展的趋势。深度强化学习可以经由端对端的学习方式完成从原始输入到输出的直接控制，即能够运用到强化学习的试错算法和积累奖励函数来加速神经网络设计，又能够使用到深度学习的高维数据处理本领与快捷特征提取本领来成功搞定强化学习中的值函数逼近问题，可以进行“从零开始”“无师自通”的学习方式。

2.1 卷积神经网络

人工神经网络（ANN，artificial neural networks）是一种模拟生物神经系统的结构和行为，从信息处理的视角对生物神经元网络实行抽象，搭建一种容易模型，依据不一样的连接方式构建不一样的网络，进行分布式并行信息处理的算法模型。人工神经网络借助改变结构中神经元与神经元之间的权重关联，进而完成处理相关信息的目标。卷积神经网络（CNN，convolutional neural network）是一种前馈型神经网络，是由一些卷积层和池化层构成的，对于图像处理技术方面卷积神经网络的运用特别广泛，并且表现结果格外优秀。

卷积神经网络主要由：输入层、卷积层、ReLU 层、池化（Pooling）层和全连接层（全连接层和常规神经网络中的一样）构成。把几个类型层相互叠加到一起，能够搭建成为一个完整的卷积神经网络。在现实情况的使用过程里，经常把卷积层与 ReLU 层一起叫作卷积层，因此卷积层在通过卷积步骤之后还需要通过激活函数。如图 3 是一种卷积神经网络结构。

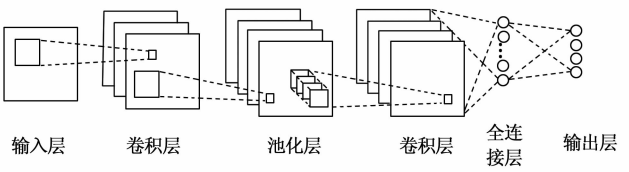

图 3 一种卷积神经网络结构图

将卷积神经网络与典型人工神经网络相比，成功处理的一类问题是“将复杂问题简化”，将诸多参数降低维度到些许参数，之后再进一步处理。最主要的原因为：在绝大多数情况下，降低维度的处理并不改变最终结果。举例像 2 000像素的图片被降维至 400 像素，此行为最终不改变人眼看出来图片中原来的小动物具体是什么种类，机器也和人眼一样。成功处理的另一类问题是“保留视觉特征”，从视觉的角度来看，在图片的具体内容（本质）不发生改变，不过图片位置发生了改变的情况。在我们改变图像中物体的位置时，用经典人工神经网络的方法处理得到的参数将与之相差特别多，此现象极度与图像处理的要求相矛盾。但卷积神经网络成功处理了此问题，它应用几乎相同于视觉的方法维持了图像的特征，当图像做翻转、旋转或者变换位置得动作时，它仍然可以快速地识别得出这也是类似的图像。

伴随着深度学习相关技术的研究与发展，卷积神经网络在各种图像数据集上的准确率越来越高，其结构也向着更深的方向发展，这得益于大数据技术和 GPU 的出现。这些进步使得卷积神经网络在计算机视觉领域获得了广泛的应用，从简单的图像分类，到精确到每一像素的图像分割，卷积神经网络都有着特别出色的表现。

2.2 马尔可夫决策过程

强化学习可以看作状态、动作和奖励三者的时间序列，其中状态又可以分为三种：环境状态，智能体状态和信息状态。信息状态包含历史的所有有用信息，一般指马尔可夫。马尔可夫状态中，当前状态只与前一个状态有关，一旦当前状态已知，就会舍弃历史信息，只需要保留当前状态。

经过前人数多年的持续探索和研究，最终一种能够成功处理多数强化学习问题的框架被发明揭示，此框架即马尔可夫决策过程（MDP，markov decision processes）。接下

来本文将相对详细地介绍马尔可夫决策过程：首先介绍马尔可夫性，接着介绍马尔可夫过程，最后介绍马尔可夫决策过程。

马尔可夫性指下一个状态只与当前状态有关，且与先前的状态无关。马尔可夫性定义为：状态 s_t 是马尔可夫性的，当且仅当：

$$P[s_{t+1} \mid s_t] = P[s_{t+1} \mid s_1, \cdots, s_t] \tag{7}$$

从上面的定义可以看出，某个状态是马尔可夫的，即该状态从历史中捕获了所有信息。因此，一旦得到了该状态，就可以舍弃历史信息了。换句话说，当前状态是未来的充分统计量。在强化学习过程中，状态 s 包含了足够多的历史信息，来描述未来所有的回报。

马尔可夫过程的定义：随机变量序列中的每个状态都是马尔可夫的，是一个二元组（S，P），S 为有限状态集，P 是状态转移概率。

对于马尔可夫状态 s 和他的后继状态 s'，定义状态转移概率为：

$$P_{ss'} = [s_{t+1} = s' \mid s_t = s] \tag{8}$$

状态转移矩阵 $\boldsymbol{P}$ 定义了所有由状态 s 到后继状态 s' 的转移概率，即：

$$\boldsymbol{P} = \begin{bmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{n1} & \cdots & p_{nn} \end{bmatrix} \tag{9}$$

马尔可夫决策过程由五元组$<S$，A，P，R，$\gamma>$组成。其中 S 为有限的状态集，A 为有限的动作集，P 为状态转移概率，R 为回报函数，γ 为折扣因子，用于计算累积回报。

强化学习的目标是，给定一个 MDP，寻找最优策略。这里的策略指从状态到行为的映射：

$$\pi(a \mid s) = \boldsymbol{P}[A_t = a \mid S_t = s] \tag{10}$$

该式含义为：策略 π 在任何状态 s 下指定一个动作概率，假设这是一个确定的动作，那么这个策略为确定性策略。事实上，强化学习的策略通常是随机策略，移动机器人通过不断测验其他动作从而找到更优秀的策略，为此引入概率因素。既然策略是随机的策略，那么状态变化序列也可能不同，因此累计回报也是随机的。

2.3 Q-learning

在强化学习中，大部分情况下都会采用时间差分（TD，temporal-difference）算法族。TD-Learning 联结了动态规划和蒙特卡罗算法，是强化学习的核心思想。实际上，TD-Learning 利用了马尔可夫属性，通过含蓄地构建 MDP 结构来利用它，然后从 MDP 结构来求解问题。

TD-Learning 策略迭代包括策略评估和策略改善，若策略评估和改善的更新方式相同则为 On-Policy，否则为 Off-policy。Q-learning 算法就是基于 Off-policy 的 TD 优化，其在更新下一策略时使用了 max 操作，为挑选最优动作，但是当前动作并不见得必能挑选到最优动作，所以这里策略评价的策略和策略改进的策略不同。

Q-learning 是 Watkins 于 1989 年提出的一种无模型的强化学习方法[15]。它是一种 value-based 算法，即通过判断每一步动作的价值来进行下一步的动作，Q-learning 的核心是 Q-Table，它可以评定所有可用行为的预期效用，并且不使用环境模型，即无需先验知识。在同一时间，它也能够处理并解决随机过渡问题和奖励问题，并且不必做任何调整。因为目前已经得到了证实，就是从目前状态出发，每个连续步骤对于收益总回报能得到最大的期望值，针对随机一个局限的 MDP，调整 Q-learning 学习最终结果是会找到一个最优策略。在起初学习前，Q 将被初始化成为一种不定的固定值。接下来在下一个时间 t，智能体会进行一次动作选择 a_t，并获得一个奖励 r_t，得到一个全新的状态 S_{t+1} 和 Q 值更新。值函数的迭代过程是该算法的重点，即：

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \cdot [r_t + \gamma \max_a Q(s_{t+1}, a_t) - Q(s_t, a_t)] \tag{11}$$

其中：α 是学习率，γ 为折扣因子。

Q-learning 算法通过一个 Q-tabel 来记录每个状态下的动作值，在动作空间或状态空间相对很大的时候，必要的存储空间便会更大。如果状态空间或动作空间连续，则该算法无法使用。所以，Q-learning 算法只能用于解决离散并且低维度状态空间和动作空间类问题。

表 1 为 Q-learning 算法的伪代码。

表 1　Q-learning：Off-policy TD Control 算法

1. 初始化 $Q(s_t, a_t)$，$\forall s \in S, a \in A(s)$，设终止状态下 $Q = 0$
2. 循环（对于每个实验）：
3. 初始化状态 s
4. 循环（对于实验中的每一步）：
5. 在状态 s 下根据 ε-Greedy 策略选择动作 a_t
6. 选择动作 a_t，得到回报 r_t 和下一个状态 s_{t+1}
7. 按照公式(11)更新迭代
8. $s \leftarrow s_{t+1}$
9. 直到 s_t 为终止状态
10. 输出最终策略 $\pi(s) = \mathrm{argmax}_a Q(s_t, a_t)$

2.4 深度 Q 网络

Mnih 等人把卷积神经网络和经典的 Q-learning 算法结合到一起，提出了 DQN 算法模型，该算法的提出开创了深度强化学习新的研究领域。DQN 算法的核心概念是以一个人工神经网络 q（s，a：ω），$s \in S$，$a \in A$ 来代替 Q-tabel，亦即动作价值函数，该算法将卷积神经网络作以媒介，将参数是 ω 的 f 网络约等取代为值函数，原理公式为：

$$f(s, a, \omega) \approx Q^*(s, a) \tag{12}$$

其中：f（s，a，ω）能够是任意类型函数，用函数取代，神经网络的输出能够拿来表征 Q 值，且不管状态空间的大小如何，s 为输入状态。网络的输入为状态信息，而输出则是每个动作的价值，因此 DQN 算法不仅可以用来解决连续状态空间而且可以解决离散动作空间问题。

DQN 相较于传统强化学习算法有两大非常重要的

改进：

1）引入深度学习中的神经网络，并不直接使用预更新的目前 Q 网络。使用的神经网络为双重网络结构，便可以一起使用 Q 估计网络和 Q 目标网络来完成模型的训练，并由此来降低目标值与当前值之间的关联性。在学习经历中，学习目标是应用目标网络进行自益从而获得回报的评估值。在更新的历程中，无需更新目标网络的权重，只需更新评估网络的权重。另外目标网络和估计网络的构成是一摸一样的。此方法极大地提高了网络训练的稳定性和收敛性。

卷积神经网络的训练是一种最优化问题，所谓的最优化就是最优化一个损失函数[16]，是标签与卷积神经网络之间输出的偏差值，其目标是使得损失函数值最小。所以，首先必须有一定的训练样本，其中含有许多的带标记数据，接着再经由以反向传播[17]方式的梯度下降来改变并更新卷积神经网络的参数。

此时使用 Q-learning 计算出来的正确 Q 值当做标记信息，为 Q-Network 提供需要的训练样本，不断地优化权重矩阵。因此，Q-Network 训练的损失函数为：

$$L(\omega)=E\left[\left(r+\gamma\max_{a_{t+1}}Q(s_{t+1},a_{t+1},\omega^{-})-Q(s_t,a_t,\omega)\right)^2\right] \tag{13}$$

其中：$Q(s_t,a_t,\omega)$ 是目前网络输出，用于评价与目前状态动作相应的值函数，$Q(s_{t+1},a_{t+1},\omega^{-})$ 是 Q-target 的输出，用于获得目标函数的 Q 值，ω^{-} 由参数 ω 延迟更新得到。

2）采用了经验回放机制，要求在每次进行网络更新时输入的信息数据符合独立同分布，从而打破数据间的关联性。记忆回放的基本思路是在每一回学习进程中随机地选择出记忆库中的部分样本数据，然后对它实行梯度下降学习。智能体会随机从经验池中抽取定量的 transition，以进行学习，既可以学习现在也可以学习过去的经验，同时随机的抽取也降低了样本之间的相关性导致的过拟合。要想能够把新的经验单元与旧经验单元任意地混杂并加以更换，进而打断相邻样本数据之间的关联性，需要利用到记忆矩阵中的经验单元 (s,a,r,s')。同时其中多数经验会被多次重复地利用或加以更新，对于数据取得相对麻烦的情境特别适用，以此提升样本的利用率。

DQN 的算法运行结构图如图 4 所示。

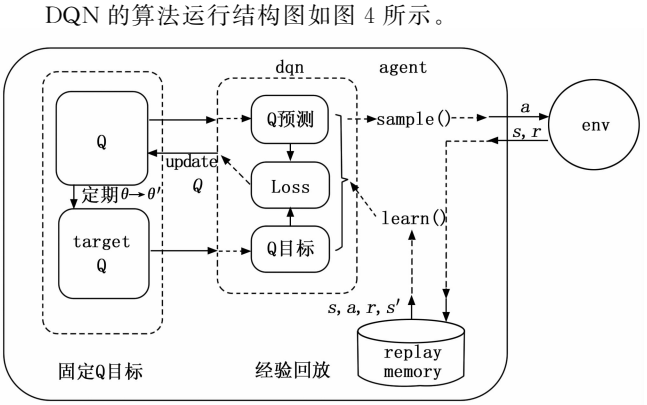

图 4 DQN 算法运行结构图

智能体不断与周围环境交互，并获取交互数据 (s,a,r,s') 存入 replay memory，当经验池中有足够多的数据之后，从经验池中随机抽取出一个 batch _ size 大小的数据，然后使用当前网络计算 Q 的预测值，再利用 Q-Target 网络计算出 Q 目标值，进而计算两者之间的损失函数，使用梯度下降来改变或更新当前网络参数，重复若干次后，把当前网络的参数复制给 Q-Target 网络。

3 深度 Q 网络和人工势场融合算法

深度强化学习是深度学习与强化学习的整合体，实际来说是融合了强化学习和深度学习的各自优点，但它仍然偏重于强化学习，处理的关键依旧是有关决策的，仅仅依靠神经网络强劲的表达本领来拟合 Q 表或径直拟合策略用来处理状态－动作空间问题[18]。

传统的 DQN 算法训练时间长，收敛速度慢。DQN 在 Q 表的初始化进程中通常被设计为绝对随机抽取，对环境并无先验信息，此设计导致自主移动机器人在各个初始状态对动作的挑选是完全任意的。通常使用 ε-贪婪策略进行动作选择。当自主移动机器人挑选出动作并进入到接下来的状态时，它通常会按照目前状态下动作的即时奖励与接下来状态的最大行为值函数来更新和调整目前状态的行为价值函数。DQN 算法是一种 Off-policy 算法。在最开始几轮训练规划中，尤其是在规模相对大的未知的环境下，会轻松地出现巨大的无效迭代空间[19]，之后伴着路径规划训练过程的慢慢增多，智能体对环境信息有了越来越多的了解，Q 值将逐步趋向收敛状态，进而路径规划回合的速率也会随着逐步变快[20]。

为了克服 DQN 算法收敛速度慢且在计算初始阶段容易陷入巨大的无效迭代空间问题，我们采用了 DQN 与人工势场算法结合的方式初始化 Q 值，提供给算法先验知识，鉴于地图上障碍物散播相对比较聚集，以便增强算法的实时性能，刨除障碍物出现的斥力场对移动机器人的排斥力作用，本文只考虑目标点对自主移动机器人的引力作用。本文的引力势场函数的引力如式（14）所示。

$$F'_{att}=-\nabla U'_{att}=\begin{cases}\zeta[L-\zeta d(q,q_{goal})] & ,d(q,q_{goal})\geqslant d^{*}\\ \zeta[L-\zeta d^{*}] & ,d(q,q_{goal})\leqslant d^{*}\end{cases} \tag{14}$$

其中：ζ 是引力因子，$d(q,q_{goal})$ 是当前点到目标点的欧几里得距离，d^{*} 是距离阈值，L 是提供的栅格地图的对角线距离。在 Q 值初始化的进程中，使用栅格 s_i 处的引力势能 $U'_{att}(s_i)$ 来初始化此状态情境下的任何价值函数 $V(s_i)$，最后经过式（15）来实现 Q 值的初始化。

$$Q(s,a,\omega)=r+\gamma V(s') \tag{15}$$

该算法的流程图如图 5 所示。

4 仿真实验

4.1 实验环境

为了验证本文提出的人工势场与 DQN 结合算法在路径规划中的收敛速度和较小迭代空间上的出众性能，对本文

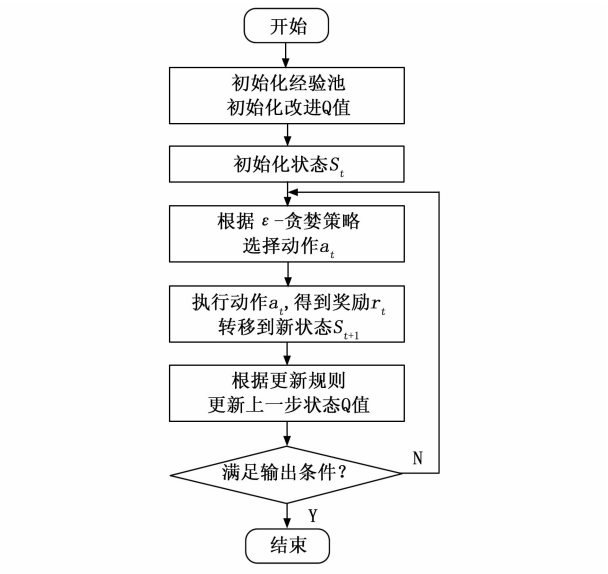

图 5　DQN 改进算法流程图

提出的改进 DQN 方法与传统 DQN 算法进行了效果对比。由于栅格法建模的广泛性和实用性[21]，实验环境使用栅格法进行建模。本文模拟软件使用的实验环境如表 2 所示。

表 2　实验环境与配置

实验环境	环境配置
操作系统	Windows10
CPU	Inter@Core™i7-9750H@2.60 GHz
GPU	NVIDIA GeForce GTX 1050 3G
内存	8 GB
编程工具	PyCharm
深度学习工具	PyTorch

4.2　参数配置

机器人的动作 $A=$｛向上，向下，向左，向右｝。

奖励函数将会对强化学习挑选动作的优劣给给出评价，在强化学习过程中具有积极的任务完成指引效果，用 r 表示。本文算法奖励函数设置为：

$$r=\begin{cases}1, & s_{t+1}\text{ 为目标位置}\\ 0, & s_{t+1}\text{ 为其他位置}\\ -1, & s_{t+1}\text{ 为障碍位置或出界}\end{cases}\tag{16}$$

仿真实验的各个参数为如表 3 所示。

表 3　实验参数

参数名称	参数值	参数名称	参数值
批处理数目	200	ε	1
迭代次数	1 000	探索衰减值	0.99
ζ	0.9	γ	0.9
τ	0.3	d^*	3

4.3　实验结果与分析

运动环境大小设置为 16×16 的栅格环境，其中每个栅格的尺寸为 1×1 像素。将栅格地图环境左上角设置为坐标的原点，横向方向设为 x 轴，纵向方向设为 y 轴，定义起点位置为（1，1），目标点位置为（15，15），黑色的栅格象征障碍区域，白色的栅格象征自由移动区域，仿真环境如图 6 所示。将自主移动机器人近似作为一个点，忽略其体积对仿真实验的影响。通过实验仿真，在全局静态环境下，利用改进 DQN 算法得到的从起点到终点路径规划和传统 DQN 算法结果相同，如图 7 所示。利用传统的 DQN 算法与改进的 DQN 得到的迭代收敛对比如图 8 和图 9 所示。

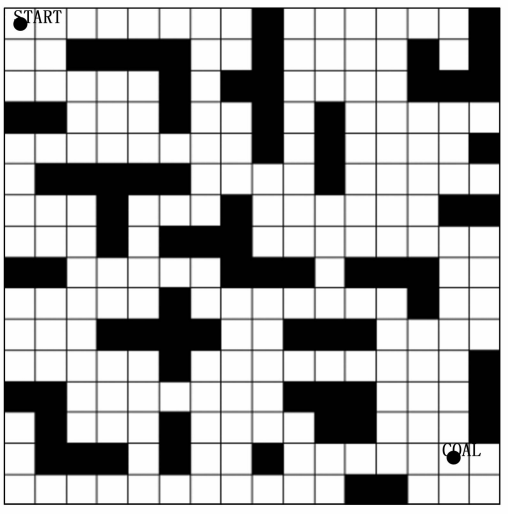

图 6　仿真环境

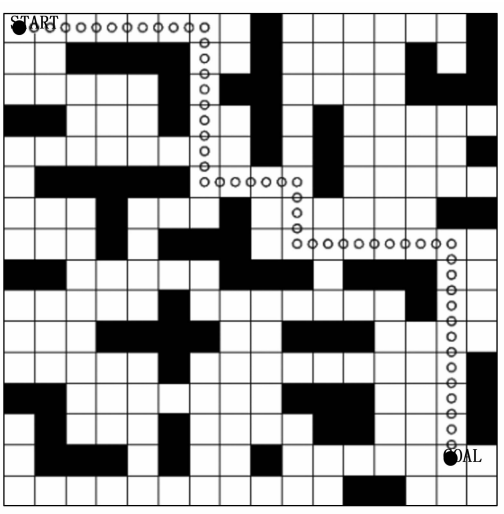

图 7　最短路径图

图 7 中左上角点为任务起点，右下角点则为目标点，其轨迹是移动机器人从起点到目标点的正确且最优路径。说明由于机器人与环境做出的不断交互，本文提出的改进深度强化学习算法可以使得机器人在该环境中进行正确的路径规划。

以迭代步数（episode）为横坐标，到达目标点所需步

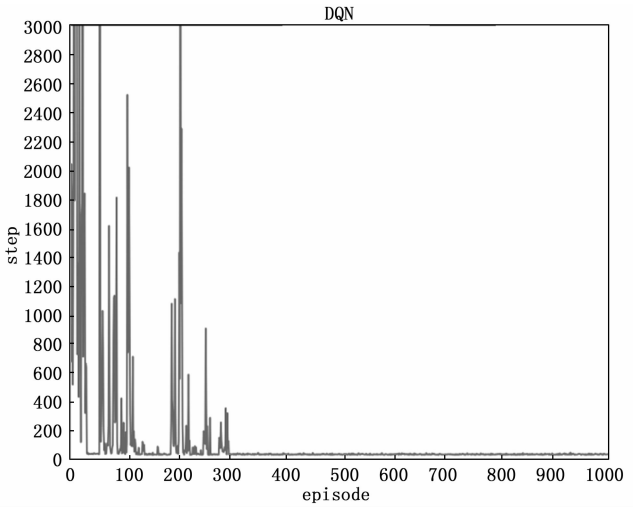

图 8 DQN 算法迭代变化曲线图

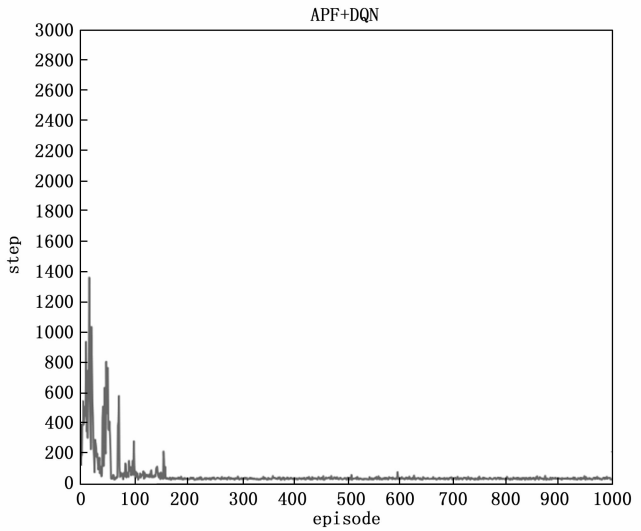

图 9 APF＋DQN 改进算法迭代变化曲线图

长（step）为纵坐标做出曲线图，从而进行两种算法的对比。当迭代步数的慢慢扩展增多，从起点至目标点的规划步长数目也逐渐趋向变小，最后会收敛至最优步长。

通过迭代图对比，可以得出改进的 DQN 算法得出的路径规划前期产生的迭代空间更小，收敛更早。传统的 DQN 算法在前约 220 次迭代产生较大的迭代空间，需要迭代约 300 次才可以收敛。改进的 DQN 算法在前约 90 次迭代中产生了较小的迭代空间，需要约 160 次就达到了收敛。

传统 DQN 算法在训练初始阶段由于缺乏样本池，所以动作选择的随机性相对较大，且只有得到大批的样本后便能训练出正确的模型。通过结合了深度 Q 网络算法与人工势场法，使得在训练初始阶段能够提供给模型适量的导向，进而减小了训练过程的盲目性和随机性，也因此进一步减少了训练时间，可以较快地得到优化模型。表 4 概括地对比了两种算法的性能，数据是对两种算法分别运行了 10 次取得的数据平均数。

表 4 两种算法性能对比

算法	收敛时间/s	收敛前迭代次数	1000 次迭代最优路径长度(栅格数)
DQN	299.4	353.4	29
APF＋DQN	116.8	192.7	29

上述仿真结果表明，本文所提出的算法能够实现智能移动机器人行走过程中的全局路径规划，对环境的路径规划有着良好的表现。在收敛时间方面，改进算法相较于传统算法减少了 60.99%，同时在收敛前的迭代次数方面减少了 45.47%。所以对于传统 DQN 算法而言，本文所采用的融合人工势场法的 DQN 算法效率更高，收敛速度更快，前期训练产生的迭代空间更小。

5 结束语

改进 DQN 算法针对传统 DQN 算法训练时间长、收敛速度慢、训练前期产生巨大迭代空间等问题，在原始 DQN 算法的基础上引入基于人工势场法的引力场来辅助初始化 Q 值进行改进，应用二维栅格环境进行试验验证。实验结果表明，改进 DQN 算法减少了算法训练时间，加快了收敛速度，减小了训练前期产生的迭代空间，对解决实际问题有一定的应用价值。本文主要面对静态环境下的全局路径规划研究，至于深度 Q 网络算法在繁杂的动态环境下的相关问题，尚有待继续的研究和讨论。

参考文献：

[1] 吴运雄，曾 碧. 基于深度强化学习的移动机器人轨迹跟踪和动态避障 [J]. 广东工业大学学报，2019，36 (1)：42－50.

[2] 李 辉，祁宇明. 一种复杂环境下基于深度强化学习的机器人路径规划方法 [J]. 计算机应用研究，2020，37 (S1)：129－131.

[3] KOREN Y，BORENSTEIN J. Potential field methods and their inherent limi－tations for mobile robot navigation [C] //Proc of IEEE International Conference on Robotics and Automation. Piscataway，NJ：IEEE Press1991：1398－1404.

[4] CLERE M，KENNEDY J. The particle swarm explosion，stability，and convergence in a multidimensional complex space [J]. IEEE Trans on Evolutionary Computation，2002，6 (1)：58－73.

[5] HEE R B，KYUNG S C. A sensor-based navigation for a mobile robot using fuzzy logic and reinforcement learning [J]. Systems Man & Cy bernetics IEEE Transactions，1995，25 (3)：464－477.

[6] CASTILLO O，LEONARDO T，PATRICIA M. Multiple objective genetic algorithms for path-planning optimization in autonomous mobile robots [J]. Soft Computing，2007，11 (3)：269－279.

[7] 童 亮，王 准. 强化学习在机器人路径规划中的应用研究 [J]. 计算机仿真，2013，30 (12)：351－355.

（下转第 239 页）