DOI:10.19557/j.cnki.1001-9944.2021.01.008

DDPG 算法在实现无人车快速控制的研究

朱　坚,宋晓茹,高　嵩,高泽鹏
(西安工业大学 电子信息工程学院,西安 710021)

摘要:大多传统的无人车控制算法需要人为调整参数,需要算法设计精确规则,无法快速适应多种情况。针对上述问题,该文采用深度强化学习对无人车的方向、速度和刹车三方面进行控制,让无人车自主学习,自主调参。该文重点通过改进OU噪声信号和设计网络结构,实现对无人车的快速控制。在TORCS无人车模拟器的仿真结果表明,改进后的方法误差曲线可以快速的收敛,有效解决了传统无人车控制耗时耗力的问题,对无人车的控制研究有重要的意义。

关键词:无人车;深度强化学习;TORCS;OU噪声;网络结构
中图分类号:TP391　　文献标志码:A　　文章编号:1001-9944(2021)01-0031-05

Research of DDPG Algorithm in Realizing Fast Control of Autonomous Vehicle

ZHU Jian,SONG Xiao-ru,GAO Song,GAO Ze-peng
(School of Electronic Information Engineering,Xi'an Technological University,Xi'an 710021,China)

Abstract:Most traditional autonomous vehicle control algorithms need to adjust parameters manually and design precise rules,so they cannot adapt to different situations quickly. In view of the above problems. This paper adopts deep reinforcement learning to control the direction,speed and brake of the autonomous vehicle,so that the Autonomous vehicle can learn and adjust parameters autonomously. This paper mainly through improving OU noise signal vector and design network structure to realize the rapid control of autonomous vehicle. The simulation results of TORCS autonomous vehicle simulator show that the improved method can fast convergence error curve,effectively solve the traditional time-consuming problem of autonomous vehicle control,the control of autonomous vehicle research will have an significant meaning.

Key words:autonomous vehicle;deep reinforcement learning;TORCS;OU noise;network structure

车辆的普及带来了很多的交通安全问题。根据国家统计局的统计数据,平均每年发生交通事故近20000起。随着科技的发展,无人驾驶[1-3]作为汽车新的研究热点,新的智能算法[4]在无人的控制方面有了更加广泛的应用。

在无人驾驶汽车的各个部分中,行为决策是最关键的一部分。文献[5-6]使用A*算法实现机器人的控制,A*算法虽然从理论上来说,可以得到最优时间路径,但是如果是在情况比较复杂的情况下,计算量是非常庞大的,A*算法是无法处理的。文献[7]采用Dijkstra算法实现最短路径的求解,该算法简单明了,能够得到最优解,但是效率比较低,特别是

收稿日期:2020-10-28;修订日期:2020-12-15
基金项目:陕西省重点研发计划项目(2019GY-072);陕西省教育厅专项科研计划项目(17JK0369)
作者简介:朱坚(1996—),男,在读硕士研究生,研究方向为深度学习、强化学习、无人车路径规划;宋晓茹(1978—),女,博士,副教授,研究方向为智能控制、目标识别;高嵩(1964—),男,博士,教授,研究方向为目标探测与识别、自主系统及机器人、工业控制与自动化;高泽鹏(1995—),男,硕士研究生,研究方向为路径跟踪。

在实际情况下，并不需要求得最优的解，并且运算占用空间大。文献[8-9]使用的是粒子群算法，该算法虽然搜索速度快，但是对离散的问题处理效果不佳。

上述文献中基本使用建立栅格地图的方式来进行算法仿真，并不能看出算法在实际道路下的情况，因此本文使用 TORCS[10]无人车模拟器实现对算法的仿真。TORCS 无人车仿真平台，内部集成了各种各样的精确的车辆动力学模型和赛道，与建立栅格地图的方法相比，一方面环境更加复杂，由二维平面变为三维地图，并且算法处理的数据更加庞大，计算量也大大提高，所以使用 TORCS 模拟器更能体现出算法在实际情况下的表现，并且还不用考虑安全性问题。

由于传统的路径规划算法，如 A* 算法，Dijkstra 算法等，其决策方式是一个典型的有限状态机，只能采取保守的驾驶策略，需要人为设计精确的规则来应对复杂多状态的各种情况，如果其算法模型的参数设置精度不够，那么在实际情况下将无法达到较好的效果，也就无法实验无人车的快速控制。基于此。本文使用 DDPG 算法实现无人车的控制，文献[11]提出了一种新的多目标车辆跟随决策算法，解决了已有算法泛化性和舒适性差的问题。文献[12]将模仿学习(IL)和 DDPG 相结合，加快了强化学习的训练过程。

仿真结果表明，改进后的 DDPG 算法，可以解决传统算法解决不了的计算量大、占用运算空间多、对离散问题处理效果不佳等问题，与未改进前的算法相比，在无人车的自动控制方面有更好的表现。

1 算法模型设计

1.1 DDPG 算法

DDPG(Deep Deterministic Policy Gradient)算法，其算法原理在本质上是 Actor-Critic 算法和 DQN(Deep Q-Learning Network)算法的结合体。

DDPG 算法一方面使用了和 DQN 算法中相同的经验池和双网络结构来促进神经网络的学习；而算法中的“Deterministic”表示 Actor 网络不再输出两个动作的概率，而是一个具体的值。

如图 1 所示，算法原理的另一部分和 Actor-Critic 算法相同，DDPG 算法有一个 Actor 网络和 Critic 网络，Actor 网络和 Critic 网络都有目标值网络(Target-net)和估计值网络(Eval-net)。只需要训练两个 Eval-net 的网络参数，而 Target-net 网络的参数是由前面两个网络每隔一定的时间复制过去得到。

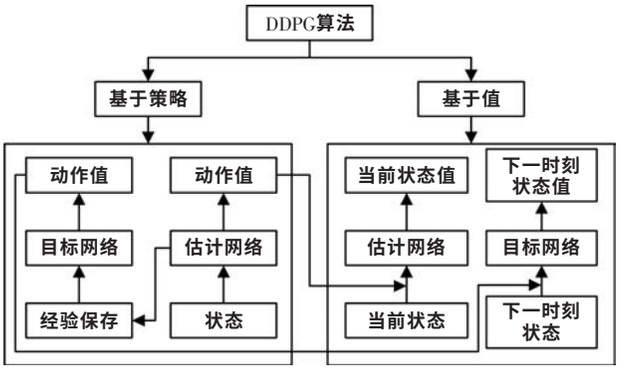

图 1 DDPG 算法原理结构

Fig.1 DDPG algorithm principle structure

Critic 网络的更新公式为

$$\begin{cases} y_i=r_i+\gamma Q'(s_{i+1},\mu'(s_{i+1}\mid\theta^{\mu'})\mid\theta^{Q'}) \\ L=\dfrac{1}{N}\sum\limits_{i=1}^{N}(y_i-Q(s_i,a_i\mid\theta^{Q}))^2 \end{cases} \tag{1}$$

Actor 网络的更新公式

$$\nabla_{\theta^{\mu}}J=\frac{1}{N}\sum_{i}\nabla_{a}Q(s,a\mid\theta^{Q})\mid_{s=s_i,a=\mu(s_i)}\nabla_{\theta^{\mu}}\mu=(s\mid\theta^{\mu})\mid_{s_i} \tag{2}$$

Actor 网络和 Critic 的网络参数，是通过网络的梯度进行更新的。Actor 网络的梯度用 Grand(μ)表示，Critic 网络的梯度由 Grand(Q)表示。关于 Actor 网络的更新，其参数的更新一方面是从 Critic 网络得到的，通过 Grand(Q)该梯度的正负和大小，得到 Actor 网络的更新方向；而 Grand(μ)来自 Actor 网络本身，这两个参数相结合，表示 Actor 网络要朝着获得最大 Q 值的方向来修正网络的参数。

1.2 DDPG 算法的改进

为了使模型可以快速学习和更新为网络参数，提高算法的探索能力，求取到最优解，因此采用不断衰减噪声信号的方法来改进 OU 过程。

Ornstein-Uhlenbeck 过程（也称为 OU 过程）是一种序贯相关的过程，在 DDPG 中用于实现 RL 的探索，OU 过程满足如下的随机微分方程：

$$\mathrm{d}x_t=\theta(\mu-x_t)\mathrm{d}t+\sigma\mathrm{d}W_t \tag{3}$$

式中：$\theta>0$；$\mu,\sigma>0$ 为参数；W_t 为维纳过程。

在代码实现中：

$$OU=\theta(\mu-X)+\sigma W \tag{4}$$

式中：W 为满足正态分布的一个随机数。

改进后的计算方法为在训练一开始，给 *noise*

设定一个较大的值，然后随着训练步骤衰减 *noise*，可以使得模型快速学习网络的参数，快速找到算法的最优解。

$$\begin{cases} noise=noise-\varepsilon \\ OU=\max(noise,0)*OU \end{cases} \tag{5}$$

1.3 网络结构设计与参数设置

TORCS 无人车仿真平台，内部集成了各种各样的精确的车辆动力学模型和赛道，并且可以获得仿真环境下所有车辆的真实数据。

Actor 网络结构如图 2 所示，Critic 网络结构如图 3 所示。

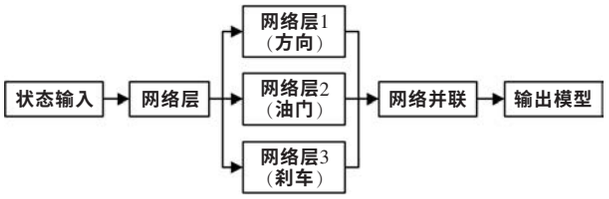

图 2 Actor 网络结构设计

Fig.2 Design of Actor network structure

图 3 Critic 网络结构设计

Fig.3 Design of Critic network structure

TORCS 模拟器中主要监测无人车的三部分：方向，油门，刹车，在网络结构中将此三部分做为网络结构中平行的三个部分，然后将这三部分的输出进行连接，作为整个神经网络的输出，从而形成 Actor 网络的网络结构模型（Actor-model）。

而 Critic 网络，将状态值和动作值分别输入到网络层中，将输出连接后输入另一个网络层中，形成 Critic 网络的网络结构模型（Output_Critic-model）。

网络的参数是通过梯度下降的方式进行训练调整的。学习率的设置会大大影响网络的学习速度。

梯度下降公式为

$$\theta=\theta-\alpha\nabla J \tag{6}$$

当 α 设置过大，梯度可能会在最小值附近震荡，甚至可能无法收敛；当 α 设置过小时，收敛速度会非常缓慢。通过同时调整 Actor 网络和 Critic 网络的学习率，使的无人车能更快的自主学习调整网络参数。

2 仿真实验

2.1 仿真条件

2.1.1 仿真平台

TORCS 是一个开源的赛车仿真模器，通过 UDP 协议进行通信。客户端（Client）可以向服务器（Server）发送数据请求，并根据得到的传感数据进行车辆的控制，控制效果由 TORCS 进行 3D 可视化。这一框架可以灵活地控制 TORCS 平台仿真的开始和终止，并可方便地获取车辆的状态（图像和其他传感）信息，根据所获信息实现车辆控制，基本满足了深度强化学习策略训练需要的所有条件。TORCS 模拟器框架如图 4 所示，模拟器道路如图 5 所示。

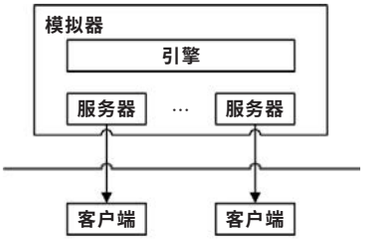

图 4 TORCS 模拟器框架

Fig.4 TORCS simulator framework

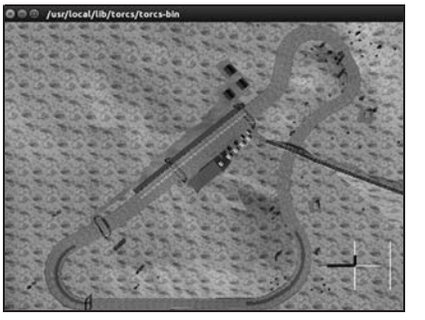

图 5 TORCS 模拟器道路

Fig.5 TORCS simulator road

2.1.2 软件版本

本文使用 python 版本是 3.5.2，tensorflow 版本是 1.4.0，keras 版本是 2.2.5，另外还有一些常见的 python 包。

2.2 实验结果

基于以上论述，本文将首先调整网络参数，然后在网络参数相同的情况下，对改进前后的算法进行仿真，观察算法的收敛速度。

为了提高网络的学习速度，本文中对主要的参数进行测试，观察参数的变化情况对实验结果的影响，从而选择最合适的参数进行网络的训练。表 1

中 ALR(Actor network learning rate)表示 Actor 网络的学习率,CLR(Critic network learning rate)表示 Critic 网络的学习率,TAU(Target Network HyperParameters)表示目标网络的超参数,GAMMA 为衰减因子。经过测试,本文网络主要参数设置如表 1 所示。

表 1 主要参数表

Tab.1 List of main parameters

编号	参数	参数值
1	ALR	0.0001
2	CLR	0.001
3	TAU	0.001
4	GAMMA	0.99

未改进算法的误差曲线如图 6 所示,奖励曲线如图 7 所示。由测试结果来看,未改进的 DDPG 算法在 TORCS 模拟器上表现并没有较好的表现,而经过改进算法之后,代价曲线可以实现快速的收敛,快速实现赛车的控制。

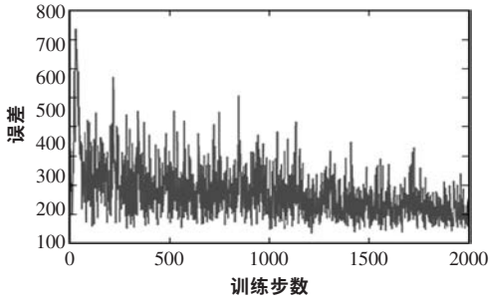

图 6 DDPG 误差曲线图

Fig.6 DDPG error graph

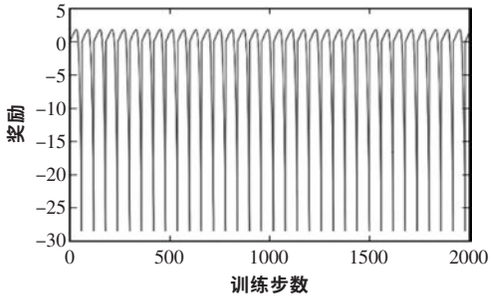

图 7 DDPG 奖励曲线

Fig.7 DDPG reward graph

由以上仿真结果看出,控制效果并不理想。图 8 和图 9 为改进后算法的仿真结果,改进后的算法可以实现快速的收敛,达到较好的训练效果。算法与改进前的情况相比,代价曲线在较短的时间内快速收敛,同时奖励函数曲线达到一个较为稳定的值;而在改进前,在相同的参数设置和实验环境下,由仿真的代价曲线和奖励曲线来看,无法达到较为理想的实验效果。由仿真结果可以看出,改进后的算法在无人车控制的快速性上有了显著提高。

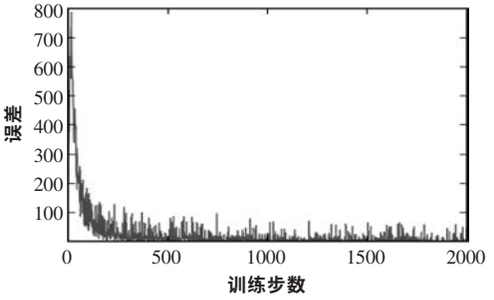

图 8 改进后 DDPG 代价曲线图

Fig.8 Improved DDPG cost curve

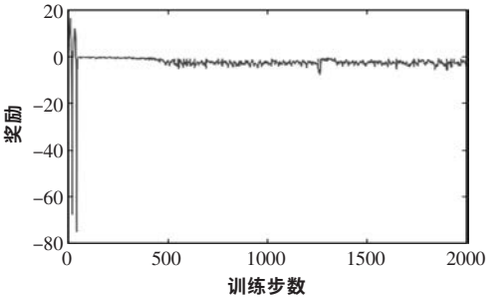

图 9 改进 DDPG 奖励曲线

Fig.9 Improved DDPG reward curve

实验仿真的对比结果如表 2 所示。

表 2 改进前后收敛步数的对比

Tab.2 Comparison of the number of convergent steps before and after improvement

算法模型	收敛时的步数
DDPG	曲线波动较大,收敛困难
改进后的 DDPG	300 步左右

3 结语

基于深度强化学习的无人车控制方法,避免了传统无人车控制方法手动调参,耗时费力的弊端。同时在 OU 噪声和网络结构两个方面的调整,并将其与该进前的算法进行比较,通过实验对算法模型进行仿真。在 TORCS 模拟器上实验结果表明,在实验环境、网络参数、训练步骤完全相同的情况下,与未改进的算法相比,改进后的算法在控制的快速性上有了明显的提高。

参考文献:

[1] 郭旭.人工智能视角下的无人驾驶技术分析与展望[J].电子世界,2017(20):64-65.

[2] 杨帆.无人驾驶汽车的发展现状和展望[J].上海汽车,2014(3):35-40.

[3] Shuhua Su,Gang Chen.Lateral robust iterative learning control for unmanned driving robot vehicle[J].Systems and Control Engineering,2020,234(7):792-808.
[4] 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.
[5] 王殿君.基于改进 A* 算法的室内移动机器人路径规划[J].清华大学学报:自然科学版,2012,52(8):1085-1089.
[6] 赵晓,王铮,黄程侃,等.基于改进 A* 算法的移动机器人路径规划[J].机器人,2018,40(6):903-910.
[7] 金婷,方欢,方贤文.改进型 Dijkstra 算法的最短路径求解[J].软件导刊,2016,15(2):129-131.
[8] 刘艳红,陈田田,张方方.基于改进粒子群算法的移动机器人路径规划[J].郑州大学学报:理学版,2020,52(1):114-119.
[9] 蒲兴成,李俊杰,吴慧超,等.基于改进粒子群算法的移动机器人多目标点路径规[J].智能系统学报,2017,12(3):301-309.
[10] 何宁,赵治国,朱阳.基于 TORCS 平台的虚拟车辆仿真系统开发[J].中国制造业信息化,2010,39(15):37-41.
[11] 邓小豪,侯进,谭光鸿,等.基于强化学习的多目标车辆跟随决策算法[J/OL].控制与决策:1-7[2020-09-15].https://doi.org/10.13195/j.kzyjc.2020.0426.
[12] Zhenzhong Chu,Bo Sun,Daqi Zhu,et al. Motion control of unmanned underwater vehicles via deep imitation reinforcement learning algorithm[J].Iet Intelligent Transport Systems,2020,14(7):764-774.

(上接第 30 页)远程监控数据可实时掌握农作物的生长情况,同时,1 位农技员可以同时控制多个大棚的种植,相对于普通温室可以节约大量的人力投入,为该地区的精准扶贫提供了参考范例。

在系统运行期间,环境因素尚存在以下问题:①由于大棚内种植菌菇,大棚内高温高湿的环境对电子器件存在影响,需要定期对其进行维护;②由于大棚建在山区,尽管使用了路由器增益天线,但偶尔会出现断电和信号弱等非技术性问题;③该系统的远程控制界面尚有待进一步开发,在实现监控的同时,可以根据农作物的生长情况实时改变参数。

尽管如此,该系统的实施运行仍为项目团队后续开发基于 5G 通讯的智能大棚远程监控系统,提供了大量的试验数据和技术积累。

5 结语

为了解决传统温室大棚种植需要对专业农技人员进行技能培训,人力成本高等问题,在此设计并实现了基于阿里云的智能大棚远程监控系统,详细分析了系统的组成和工作模式,给出了大棚主控系统架构以及大棚现场智能控制电气系统,并结合手动和自动 2 种控制模式给出了远程监控系统软件设计,分析了远程监控系统的工作过程。实际运行结果表明,智能大棚远程控制系统运行稳定,能准确显示现场实时数据。该系统对于普通大棚的升级改造具有很好的借鉴意义。

参考文献:

[1] 李道亮,杨昊.农业物联网技术研究进展与发展趋势分析[J].农业机械学报,2018,49(1):1-14.
[2] 江涛,麻洪欧.自动化技术在现代农业中的应用[J].南方农机,2015,64(7):24,28.
[3] 程力,郭晓金,谭洋.智能农业大棚环境远程监控系统的设计与实现[J].中国农机化学报,2019,40(6):174-177.
[4] 王嘉宁,牛新涛,徐子明,等.基于无线传感器网络的温室 CO_2 浓度监控系统[J].农业机械学报,2017,48(7):280-285.
[5] 孟令月.基于 ZigBee 的农业数字大棚系统设计与实现[D].大连:大连理工大学,2016.
[6] 裴晓辉.基于深度学习的智慧大棚监控系统的研究[D].沈阳:沈阳理工大学,2018.
[7] 孟子涵,龙振超,宋明智,等.基于 S7-300 型 PLC 的农业大棚智能养护系统设计[J].自动化控制理论与应用,2019,(10):12-14.
[8] 汪言康,周建平,许燕,等.基于物联网的温室大棚智能监控系统研究[J].机床与液压,2019,47(17):104-107.
[9] 张宝峰,杨雷,朱均超,等.温室大棚温湿度智能监控系统的设计与实现[J].自动化仪表,2017,38(10):83-85.
[10] 龚尚福,潘虹.智能温室大棚监控系统的研究与设计[J].现代电子技术,2017,40(19):119-122.