文章编号:1671-6833(2020)04-0041-05

基于深度强化学习的自动驾驶车控制算法研究

王丙琛,司怀伟,谭国真

(大连理工大学 计算机科学与技术学院,辽宁 大连 116000)

摘 要:为了提高基于强化学习的自动驾驶车控制算法的学习效率,提出了一种结合专家经验的自动驾驶策略学习算法(deep deterministic policy gridient with expert,DDPGwE)。DDPGwE 采用基于 DDPG 的强化学习框架进行模型在线训练,使用真实的人类驾驶数据对 actor 网络进行预训练,同时在 actor 网络中加入 LSTM 预测机制,提升自动驾驶车对将要发生状况的预判。在仿真平台 TORCS 中的实验结果表明,所提算法相较于原始 DDPG 算法,训练时间大大缩短,收敛速度加快,提高了模型的稳定性和泛化能力。

关键词:神经网络;强化学习;自动驾驶;DDPG 算法;actor-critic 网络;LSTM

中图分类号:TB183 **文献标志码**:A **doi**:10.13705/j.issn.1671-6833.2020.04.002

0 引言

近年来,随着人工智能和无人驾驶技术的快速发展,无人驾驶车辆已经向实用化迈进,在未来将对提高道路安全、促进交通管理和改善城市环保等产生颠覆性影响,成为汽车交通领域的一项革命[1]。据统计,大约有 90%的交通事故是由于驾驶员的失误造成的,主要包括注意力不集中,判断失误和情境意识不足等因素[2]。但是,机器并不会出现劳累和注意力不集中等现象。因此,无人驾驶车的出现,将有可能极大程度上降低这部分交通事故的概率[3]。Google、特斯拉、百度、NVIDIA 等著名公司在无人驾驶上均投入了大量的人力和物力,并且已经有部分车辆在道路上进行了实测[4]。美国电气和电子工程师协会(IEEE)预测,至 2040 年自动驾驶车所占比例将达到 75%[5]。无人驾驶的决策和控制模块是决定无人驾驶车安全性、稳定性的关键技术[6]。然而,就当前的情况而言,开发出能够完全自主应对各种复杂多变的路况及充满不确定性的交通场景下的无人驾驶车仍然是一项巨大的挑战。

目前,有基于深度学习和基于强化学习的自动驾驶技术。基于深度学习的方法利用人工记录人类驾驶员的行为训练深度网络,利用训练好的深度网络作为“司机”完成自动驾驶。这种方法需要大量的人工标注信息,这是不现实的。而基于强化学习的自动驾驶算法则具有自己探索环境做出正确决策的能力,这符合人们对于自动驾驶汽车的期望。强化学习,其最通用的模型构造方法是构造一个列表存储所有的状态-动作对的评价值。但是,这种方法对于自动驾驶这种状态-动作空间较大的情况不能奏效。因此基于强化学习的自动驾驶算法一直未出现较大规模的应用。

Hinton 等[7]在 2006 年提出的深度置信网络(deep belief networks)开创了深度学习的一个新纪元。2013 年,Krizhevsky 等[8]在大规模视觉识别挑战赛(imagenet large scale visual recognition competition)使用卷积神经网络取得突出成绩之后,深度学习开始在计算机视觉等许多领域得到广泛应用。2016 年,Bojarski 等[9]提出使用卷积神经网络进行自动驾驶系统研究的方案。普林斯顿大学于 2016 年提出了改进的基于深度学习的自动驾驶系统方案[10]。随着深度学习在各大领域的广泛应用,研究人员开始尝试将深度学习和强化学习进行结合形成了较为成熟的深度强化学习框架。其中最具代表性的就是 Mnih 等[11]提出的 DQN(deep q-network)算法。由于 DQN 这种方法针对的是离散动作空间,所以这种方法并不适

收稿日期:2020-02-26;**修订日期**:2020-04-09

基金项目:国家自然科学基金委员会与辽宁省联合基金重点支持项目(U1808206)

通信作者:谭国真(1964—),男,辽宁大连人,大连理工大学教授,博士,主要研究方向为强化学习、车联网,E-mail:youthbingchenw@163.com。

用于自动驾驶控制系统的开发。2016年,Google DeepMind又基于演员-评论家模型[12],将DQN算法改进为深度确定性策略梯度(deep deterministic policy gradient,DDPG)算法[13],实现了对于连续动作空间的控制。这种演员-评论家模型,分别使用一个价值网络对当前状况进行评估,一个策略网络做出下一步决策,两者的结合实现了更加符合人类决策过程的智能控制模型。

笔者根据上述研究提出一种结合了深度学习和强化学习的自动驾驶策略学习算法:采用基于DDPG的强化学习算法进行模型的在线训练,使用真实的人类驾驶数据对actor网络进行预训练,同时为了增强模型的泛化能力,在actor网络中加入LSTM预测机制。最后,笔者对原始DDPG算法和新的算法的实验效果进行了比较。

1 基于强化学习的方法

1.1 强化学习

强化学习思想来源于生物学中的动物行为训练,驯兽员通过奖励与惩罚的方式让动物学会一种行为和状态之间的某种联系规则[14]。强化学习框架如图1所示。智能体agent从环境感知初始状态s_t,采取动作a_t,此agent会得到来自环境的奖励r_t,如此产生一系列的"状态-动作-奖励",直到结束状态为止。agent的目的就是通过不断地探索环境得到反馈来最大化奖励值总和。下面将对一些比较成熟的强化学习算法进行介绍。

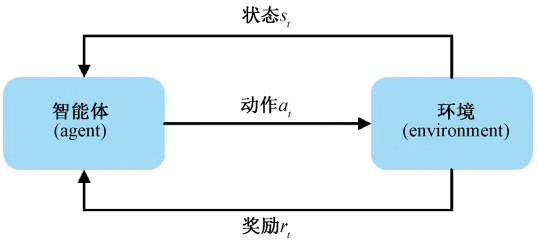

图1 强化学习框架

Figure 1 Reinforcement learning framework

1.2 DQN算法

DQN算法于2013年由Mnih等[11]提出,并于2015年进行了改进。DQN算法是传统强化学习的一个延伸,它将智能体agent和环境的交互与马尔可夫决策过程(MDP)形式化。马尔可夫决策过程就是智能体在初始状态s_0下,从动作空间A中挑选一个动作a_0执行,执行后,智能体按照一定概率转移到下一个状态s_1,然后再执行下一个动作,重复上述过程。尤其是在步骤t时,agent通过执行动作a_t,从状态s_t转移到新的状态s_{t+1},并且会得到环境反馈的奖励值r_t。在DQN之前,强化学习中的一个重要分支为Q学习算法。Q学习通过构造一个表来存储状态-动作对。Q学习的目的是根据Bellman方程将Q值函数最大化,其形式为:

$$Q(s_t,a_t)=r_t+\gamma\max_{a_{t+1}}Q(s_{t+1},a_{t+1}),\quad(1)$$

式中:γ是折扣因子。在DQN之前,当采用非线性结构(例如神经网络)来逼近Q学习的值函数时,不能保证探索的收敛性。DQN引入了经验回放和目标网络来解决这个问题[11]。首先,状态-动作对转移序列$\{s_t,a_t,r_t,s_{t+1}\}$保存在经验缓冲区中,然后训练一个深度神经网络,采用从经验缓冲区中随机抽取的转移序列来逼近Q函数,这种技术在很大程度上打破了连续转移的相关性,使得学习过程更加稳定。Q网络更新时,Q值的目标为:

$$y_t=r_t+\gamma\max_{a_{t+1}}\hat{Q}(s_{t+1},a_{t+1};\theta),\quad(2)$$

式中:θ是Q网络的参数集合;$\hat{Q}$是目标网络。学习的目标是通过最小化目标Q值与当前Q网络输出之间的误差来学习参数集θ的值。如式(3)所示:

$$J(Q)=\frac{1}{N}\sum_t(y_t-Q(s_t,a_t\mid\theta))^2。\quad(3)$$

DQN算法只考虑离散动作域,这种方法对于自动驾驶这种连续动作空间来说将不再适用。下面将介绍一种适用于连续动作空间的算法。

1.3 DDPG算法

DDPG算法基于确定性策略梯度DPG算法[15],同时采用了演员-评论家模型[12],而且保留了DQN的经验回放和目标网络技术[11]。

actor-critic算法[12]将策略梯度算法和值函数结合在一起。策略函数称为actor。值函数称为critic。基本上,演员actor会做出一个动作,评论家critic会评价这个动作。然后根据这些评价,演员将调整自己的动作,为了下次做得更好。

DDPG算法分别参数化评论家函数$Q(s,a)$和演员函数$\mu(s\mid\theta^\mu)$。评论家函数的定义类似于Q-learning中的值函数,并通过最小化来更新。如公式(4)和(5)所示:

$$J(Q)=\frac{1}{N}\sum_t L^2;\quad(4)$$

$$L=r_t+\gamma\hat{Q}(s_{t+1},\hat{\mu}(s_{t+1}\mid\theta^{\hat{\mu}})\mid\theta^{\hat{Q}})-Q(s_t,a_t\mid\theta^Q)。\quad(5)$$

演员函数将当前状态映射到当前最佳动作,

并通过以下方式更新,如式(6)所示:

$$\nabla_{\theta^\mu} J \approx \frac{1}{N}\sum_i \nabla_a Q(s,a \mid \theta^Q)\mid_{s=s_i,a=\mu_i} \cdot \nabla_{\theta^\mu}\mu(s \mid \theta^\mu)\mid_{s_i}, \tag{6}$$

式(4)~(6)中:t 为训练步数;N 为训练总步数;Q 和 μ 分别表示评论家和演员的函数;$\hat{Q}$ 和 $\hat{\mu}$ 分别表示目标网络的评论家和演员的函数;θ^Q 和 θ^μ 分别是评论家网络和演员网络的参数表示。最后,目标网络根据延迟因子 τ 复制原始网络参数来进行更新。如式(7)和式(8)所示:

$$\theta^{\hat{Q}} = \tau\theta^Q + (1-\tau)\theta^{\hat{Q}}; \tag{7}$$

$$\theta^{\hat{\mu}} = \tau\theta^\mu + (1-\tau)\theta^{\hat{\mu}}。 \tag{8}$$

2　DDPGwE 算法

基于深度学习的自动驾驶方法需要大量人工标注的数据,模型泛化能力不足。而基于强化学习的算法初始阶段学习速度缓慢,学习效率低。除此之外,希望智能体能够在驾驶过程中从专业司机的演示中学习一些驾驶风格,但是强化学习算法只能根据给定的奖励函数来进行优化,这很难描述专业司机的偏好,很难学习到类似人类的驾驶技巧。基于以上存在的问题,笔者在 DDPG 算法的基础上,提出融合了专家经验的 DDPGwE(determi-nistic policy gradient with expert)算法。DDPGwE 增加了采用由专业司机驾驶数据组成的专家经验对演员网络进行预训练的模块。同时,为了让智能体学习到类似人类在驾驶环境中对未来状况预判的能力,笔者在演员网络中加入 LSTM 预测模块,用来增强智能体在驾驶环境中对未来状况的预判能力,以便更好地做出决策,避免一些危险状况的发生。

2.1　DDPGwE 整体框架

DDPGwE 算法的整体框架如图 2 所示。本文方法仍然采用原始 DDPG 算法的整体框架。但是,在原始网络中进行改进。本方法加入了采用基于专业司机的驾驶经验数据组成 Expert 模块来对演员网络进行预训练。因为人类在学习新知识的过程中,在初始阶段会有“老师”进行指导和传授经验,在后面的学习过程当中自己将会不断地探索,这一过程正是强化学习的过程。所以,在强化学习阶段之前使用专业司机经验对网络进行预训练,这一过程更加接近人类的学习过程。同时专家经验参与到强化学习过程当中,使用一个策略在演员和专家之间选择最佳动作。同时为了让智能体更好地做出决策,该算法在演员网络中加入 LSTM 预测模块,让智能体学习对未来状况进行预判。因为这也符合人类的驾驶习惯,有了对未来状况的预判,将很大程度地提高驾驶安全性。

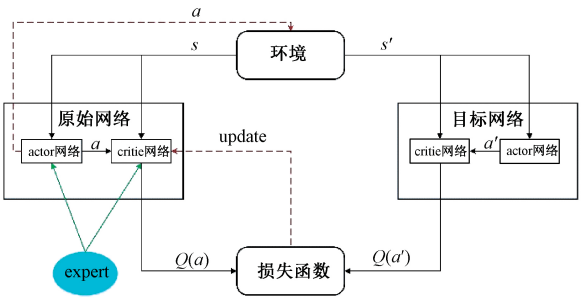

图 2　DDPGwE 框架

Figure 2　DDPGwE framework

2.2　LSTM 算法

长短期记忆网络(long short term memory network, LSTM)算法[16]是一种改进的循环神经网络算法,它使用一种被称为 LSTM 的记忆单元来判别哪些信息应该被保留,控制信息从前一时刻到下一时刻进行传输,是目前应用最为广泛的具有记忆功能的网络,其数学模型如式(9)~(13)所示:

$$i_t = \tanh(\boldsymbol{W}_{xi}x_t + \boldsymbol{W}_{hi}h_{t-1} + \boldsymbol{W}_{ci}c_{t-1} + b_i); \tag{9}$$

$$f_t = \tanh(\boldsymbol{W}_{xf}x_t + \boldsymbol{W}_{hf}h_{t-1} + \boldsymbol{W}_{cf}c_{t-1} + b_f); \tag{10}$$

$$c_t = f_t c_{t-1} + i_t\tanh(\boldsymbol{W}_{xc}x_t + \boldsymbol{W}_{hc}h_{t-1} + b_c); \tag{11}$$

$$o_t = \tanh(\boldsymbol{W}_{xo}x_t + \boldsymbol{W}_{ho}h_{t-1} + \boldsymbol{W}_{co}c_t + b_o); \tag{12}$$

$$h_t = o_t\tanh(c_t), \tag{13}$$

式中:$\boldsymbol{W}$ 是 LSTM 细胞单元的参数矩阵;b 是 LSTM 细胞单元的偏置;tanh 是激活函数,可以增强神经网络的非线性。

2.3　预训练演员网络

根据深度学习当中的预训练思想,对演员网络进行改进,如图 3 所示。通过人工采集的专业司机的驾驶数据组成的专家经验来对演员网络进行预训练,训练过程将状态-动作对作为神经网络的输入。同时专家经验也参与到动作的决策过程中,该方法确保了专家经验在初始阶段参与度高,而在接下来的强化学习阶段的参与度较低。最后,演员网络将掌握专家的驾驶经验,并且学习到专家经验之外的动作。同时为了让智能体学习类似人类真实驾驶行为中对未来状况预判的能力,在演员网络中加入 LSTM 预测模块,从而加强模型对未来状况的预判能力,从而更好地做出决策,避免危险状况的发生,使得算法具有更好的泛化

能力和预测能力。

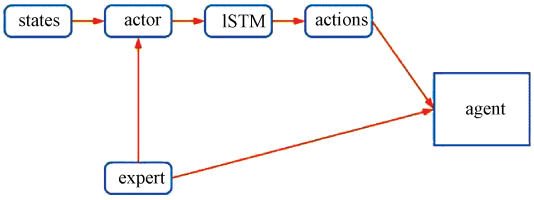

图 3 加入预训练的 actor 网络

Figure 3 Adding pre-trained actor network

2.4 算法流程

DDPGwE 算法的参数更新过程和 DDPG 算法的过程相似。使用状态-动作对作为专家经验对初始演员网络进行预训练。在强化学习阶段,在同样的输入状态下,希望演员网络生成的动作跟专家经验相似。同时,也希望 agent 所产生的这一动作在与环境交互的过程当中能够得到更高的奖励值。算法流程如下。

1: 随机初始化原始演员网络 θ^{μ}、原始评论家网络 θ^{Q}。

2: 随机初始化目标演员网络 $\theta^{\mu'}$、目标评论家网络 $\theta^{Q'}$。

3: 初始化经验回放池 D。

4: 首先采用专业司机数据预训练原始演员网络并保存权重。

5: for $episode = 1, M$ do

6: 加载预训练权重。

7: 获得初始状态 $s1$。

8: for $step = 1, T$ do

9: 利用策略在专家经验和演员网络产生的动作中选择一个执行动作。

10: 将执行完动作之后得到的序列数据 (s_t, a_t, r_t, s_{t+1}) 存储到 D 中。

11: 利用固定的比例从 D 中采样一批训练数据。

12: 利用式(4)和式(5)更新评论家网络。

13: 利用式(6)更新演员网络。

14: 利用式(7)和式(8)更新目标网络。

15: end for

16: end for

3 实验环境及设置

Torcs 是一款开放式的、跨平台的赛车模拟器[17]。它通过模拟真实车辆的发动机、离合器、变速箱等车辆物理模型来实现车辆与环境的交互,其高度的模块化和可移植性使其成为人工智能研究领域众多研究工作者的理想选择。本文方法也是采用了基于 Torcs 平台的仿真环境来验证算法的可行性。Torcs 有多个可用的地图,笔者采用如图 4 所示的地图进行算法的测试。

采用一个包含 29 个传感器值的向量(速度、角度、测距仪等)作为状态输入,3 个连续值(转向、加速和制动)作为动作输出。对于转向,范围为[-1,1],其中-1 表示最大右转,1 代表最大左转。对于加速度,它在[0,1],其中 0 表示加速度为 0,1 表示加速度最大。对于制动,在[0,1],其中 0 表示无制动,1 表示全制动。

在本文中,采用一种相对简单的方式来定义奖励函数,如式(14)所示:

$$R = v_x \cos\theta - v_x \sin\theta - v_x \mid trackPos \mid 。 \quad (14)$$

为了防止 agent 经常偏离轨道中心,也将 $trackPos$ 作为奖励函数计算的一部分,$trackPos$ 表示车辆中心线与轨道边缘的距离。希望最大化 agent 纵向速度 $v_x \cos\theta$,最小化 agent 横向速度 $v_x \sin\theta$。

图 4 模拟赛道

Figure 4 Simulated racetrack

4 结果与讨论

笔者分别记录了两种方法在训练过程中所获得的平均奖励值的统计分布。如图 5 所示,图 5 中黑色虚线是原始 DDPG 算法在训练过程中所得到的奖励值的分布,黑色实线是改进的 DDPGwE 算法在训练过程中的平均奖励值的分布。从图中可以看出,随着训练的不断进行,智能体所得到的奖励值在不断地增加,说明了智能体很好地学习了驾驶技能。

表 1 记录了原始 DDPG 算法和改进的算法的训练过程所消耗的时间和碰撞次数,以及达到收敛的迭代次数。根据仿真时间记录并且结合图 5,原始 DDPG 算法需要 240 min 的学习才能在 Torcs 中

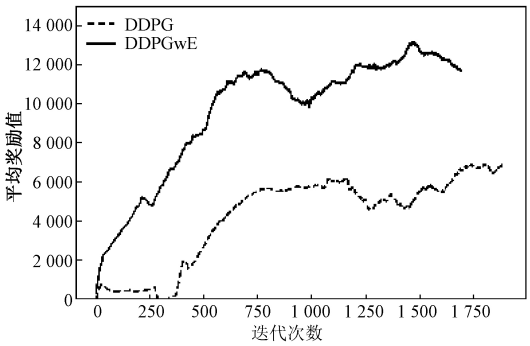

图 5　平均奖励值

Figure 5　Average reward

完整跑完一圈。而本文的算法仅用30 min即可在Torcs 中完整跑完一圈。在相同的迭代次数下,原始 DDPG 算法的学习过程不稳定,而且学习速度较慢,并且收敛到相对稳定的值所需要的时间较长。而改进过的算法则表现较好,根据图 5 和表 1 可以看出,DDPGwE 算法的学习速度较快,可以很快地收敛到一个较稳定的奖励值。并且,学习的过程非常稳定,所需要的时间也大大缩短。

表 1　训练信息

Table 1　Training information

方法	时间/min	碰撞次数	迭代次数
DDPG	240	100	1 000
DDPGwE	30	50	300

5　结论

提出了一个基于深度强化学习的自动驾驶策略学习算法(DDPGwE)。基于 DDPG 算法,首先使用专业司机驾驶数据对网络进行预训练,同时在强化学习过程中,专业司机经验也将参与到决策过程中。并且在 actor 网络中加入了 LSTM 模块来增强网络的稳定性和泛化能力。实验结果显示,笔者所提出的算法与原始 DDPG 算法相比,能够加快智能体的学习速度,并且能够快速地收敛到一个稳定的奖励值。同时,算法的泛化能力也得到了提高,具有很好的实际应用价值。

参考文献:

[1] 左思翔. 基于深度强化学习的无人驾驶智能决策控制研究[D].哈尔滨:哈尔滨工业大学,2018.

[2] TOURAN A, BRACKSTONE M A, MCDONALD M. A collision model for safety evaluation of autonomous intelligent cruise control[J]. Accident analysis & prevention, 1999, 31(5): 567-578.

[3] PADEN B, CAP M, YONG S Z, et al. A survey of motion planning and control techniques for self-driving urban vehicles[J]. IEEE transactions on intelligent vehicles, 2016, 1(1): 33-55.

[4] 夏伟,李慧云.基于深度强化学习的自动驾驶策略学习方法[J].集成技术,2017,6(3):29-34.

[5] 翁岳暄,多尼米克·希伦布兰德. 汽车智能化的道路: 智能汽车, 自动驾驶汽车安全监管研究[J]. 科技与法律, 2014 (4): 632-655.

[6] GONZALEZ D, PEREZ J, MILANES V, et al. A review of motion planning techniques for automated vehicles[J]. IEEE transactions. intelligent transportation systems, 2016, 17(4): 1135-1145.

[7] HINTON G E, OSINDDRO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural computation, 2006, 18(7): 1527-1554.

[8] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks [J]. Advances in neural information processing systems, 2012,25(2): 1097-1105.

[9] BOJARSKI M, DEL TESTA D, DWORAKOWSKI D, et al. End to end learning for self-driving cars[EB/OL]. (2016-3-25)[2019-09-31].https://arxiv.org/abs/1604.07316.

[10] CHEN C, SEFF A, KORNHAUSER A, et al. Deepdriving: Learning affordance for direct perception in autonomous driving[C]//Proceedings of the IEEE International Conference on Computer Vision. Santiago: IEEE, 2015: 2722-2730.

[11] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning[EB/OL]. (2013-10-19)[2019-10-22].https://arxiv.org/abs/1312.5602.

[12] KONDA V R, TSITSIKLIS J N. Actor-critic algorithms [C]//Advances in Neural Information Processing Systems. [S.l.]:The MIT Press, 2000: 1008-1014.

[13] LILLICRAP T P,HUNT J J,PRITZEL A,et al.Continuous control with deep reinforcement learning[EB/OL]. (2015-03-19)[2019-12-20].https://arxiv.org/abs/1509.02971.

[14] 刘赫. 动物行为训练的理论基础[J]. 中国动物保健, 2014,16(2):23-25.

[15] SILVER D, LEVER G, HEESS N, et al. Deterministic policy gradient algorithms[C]// Proceedings of the 31st International Conference on International Conference on Machine Learning. [S.l.]:JMLR, 2014:387-395.

[16] GERS F A, SCHMIDHUBER J , CUMMINS F. Learning to forget:continual prediction with LSTM[J]. Neural computation,2000,12(10):2451-2471.

(下转第 80 页)

Research on Electromagnetic Interference of Lightning Impact Test of Primary and Secondary Distribution Switches

CHENG Xian[1,3], ZHU Jianpeng[1,3], ZHAO Haiyang[1,3], YUAN Xiaodong[1,3], HE Xiang[2], XU Mingming[2]

(1.School of Electrical Engineering, Zhengzhou University, Zhengzhou 450001, China; 2.Electric Power Research Institute, State Grid Henan Electric Power Company, Zhengzhou 450000, China; 3.Henan Province Transmission and Distribution Equipment and Electrical Insulation Engineering Technology Research Center, Zhengzhou 450000, China)

Abstract: With the standardization and integration of primary and secondary equipment of power distribution switches, higher requirements for the reliable operation of secondary equipment under long-term complex electromagnetic environment were put forward. A lightning impulse voltage test platform was used to collect the secondary side of the voltage transformer, the feeder terminal unit (FTU) input and output voltage signals, and to analyze the above-mentioned waveform to obtain the electromagnetic interference spectrum distribution of the secondary device port caused by the impulse voltage, which was effectively suppressed by the filtering method. The test results showed that the lightning impulse would produce transient high frequency electromagnetic interference to the voltage transformer and feeder terminal unit, the peak value of the interference voltage caused by radiation coupling was about 3 kV, the frequency band was 3.6-16.4 MHz; the peak value of the interference voltage caused by conduction coupling was about 4.5 kV, the frequency bands were mainly distributed between 1.2-6.7 MHz and 12.5-20 MHz. After adding the filter, the positive and negative interference voltage peaks were reduced by more than 23%, and no fault occursed. The test results could provide reference for the EMC research of secondary equipment for primary and secondary fusion distribution switches.

Key words: primary and secondary equipment; lightning impulse voltage; feeder terminal unit; spectrum distribution; anti-electromagnetic interference

(上接第45页)

[17] WYMANN B, ESPIE E, GUIONNEAU C, et al. TORCS: the open racing car simulator[EB/OL]. (2013-12-15)[2019-11-12]. http://www.cse.chalmers.se/~chrdimi/papers/torcs.pdf.

Research on Autopilot Control Algorithm Based on Deep Reinforcement Learning

WANG Bingchen, SI Huaiwei, TAN Guozhen

(School of Computer Science and Technology, Dalian University of Technology, Dalian 116000, China)

Abstract: In order to improve the learning efficiency of the autopilot car control algorithm based on reinforcement learning, this paper proposes an autopilot strategy learning algorithm DDPGwE (Deep Deterministic Policy Gridient with Expert, DDPGwE) combined with expert experience. DDPGwE used a DDPG-based reinforcement learning framework to conduct online training of the model; used real human driving data to pretrain the Actor network, and added an LSTM prediction mechanism to the Actor network to improve the prediction of the future status of autonomous vehicles. The experimental results in the simulation platform TORCS showed that Compared with the original DDPG algorithm, the algorithm proposed in this paper greatly reduced the training time and speedsed up the convergence speed, which improved the stability and generalization ability of the model.

Key words: neural network; reinforcement learning; autopilot; DDPG algorithm; actor-critic network; LSTM