基于数字孪生的多自动驾驶车辆分布式协同路径规划算法

唐 伦 戴 军* 成章超 张鸿鹏 陈前斌
(重庆邮电大学通信与信息工程学院 重庆 400065)
(移动通信技术重庆市重点实验室 重庆 400065)

摘 要：针对多辆自动驾驶车辆(AVs)在进行路径规划过程中存在的车辆之间协作难、协作训练出来的模型质量低以及所求结果直接应用到物理车辆的效果较差的问题，该文提出一种基于数字孪生(DT)的多AVs分布式协同路径规划算法，基于可信度加权去中心化的联邦强化学习方法(CWDFRL)来实现多AVs的路径规划。首先将单个AVs的路径规划问题建模成在驾驶行为约束下的最小化平均任务完成时间问题，并将其转化成马尔科夫决策过程(MDP),使用深度确定性策略梯度算法(DDPG)进行求解；然后使用联邦学习(FL)保证车辆之间的协同合作，针对集中式的FL中存在的全局模型更新质量低的问题，使用基于可信度的动态节点选择的去中心化FL训练方法改善了全局模型聚合质量低的问题；最后使用DT辅助去中心化联邦强化学习(DFRL)模型的训练，利用孪生体可以从DT环境中学习的优点，快速将训练好的模型直接部署到现实世界的AVs上。仿真结果表明，与现有的方法相比，所提训练框架可以得到一个较高的奖励，有效地提高了车辆对其本身速度的利用率，与此同时还降低了车辆群体的平均任务完成时间和碰撞概率。

关键词：数字孪生；自动驾驶；去中心化联邦强化学习；路径规划
中图分类号：TN929.5 **文献标识码**：A **文章编号**：1009-5896(2024)00-0001-08
DOI: 10.11999/JEIT230678

Distributed Collaborative Path Planning Algorithm for Multiple Autonomous vehicles Based on Digital Twin

TANG Lun DAI Jun CHENG Zhangchao ZHANG Hongpeng CHEN Qianbin
(*School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing* 400065, *China*)
(*Chongqing Key Laboratory of Mobile Communications Technology, Chongqing* 400065, *China*)

Abstract: Focusing on the problems of difficult cooperation between vehicles, low quality of the model trained by cooperation and poor effect of direct application of the obtained results to physical vehicles in the process of path planning for multiple Autonomous Vehicles (AVs), a distributed collaborative path planning algorithm is proposed for multiple AVs based on Digital Twin (DT). The algorithm is based on the Credibility-Weighted Decentralized Federated Reinforcement Learning (CWDFRL) to realize the path planning of multiple AVs. In this paper, the path planning problem of a single AVs is first modeled as the problem of minimizing the average task completion time under the constraints of driving behavior, which is transformed into Markov Decision Process (MDP) and solved by Deep Deterministic Policy Gradient algorithm (DDPG). Then Federated Learning (FL) is used to ensure the cooperation between vehicles. Aiming at the problem of low quality of global model update in centralized FL, this paper uses a decentralized FL training method based on dynamic node selection of reliability to improve the low quality. Finally, the DT is used to assist the training of the Decentralized Federated Reinforcement Learning (DFRL) model, and the trained model can be quickly deployed directly to the real-world AVs by taking advantage of the twin's ability of learning from DT environment. The simulation results show that compared with the existing methods, the proposed training framework can obtain a higher reward, effectively improve the utilization of the vehicle's own speed, and at the

收稿日期：2023-07-07；改回日期：2024-01-04；网络出版：2024-01-29
*通信作者： 戴军 878607471@qq.com
基金项目：国家自然科学基金(62071078)，川渝联合实施重点研发项目(2021YFQ0053)
Foundation Items: The National Natural Science Foundation of China (62071078), Sichuan and Chongqing Key R&D Projects (2021YFQ0053)

same time reduce the average task completion time and collision probability of the vehicle swarm.

Key words: Digital Twin (DT); Autonomous driving; Decentralized Federated Reinforcement Learning (DFRL); Path planning

1 引言

自动驾驶被认为是减少驾驶员工作量、提高驾驶安全性和舒适性的一种驾驶解决方案[1]。但是由于车辆通常处在高动态的复杂环境中，车辆需要实时的调整路线，所以自动驾驶车辆(Autonomous Vehicles, AVs)如何获得周围多个车辆的运动轨迹并成功的规划一条免碰撞、舒适并且可执行的轨迹一直备受关注。

现有的路径规划方法一般分3种：基于启发式算法的[2]、基于模型预测控制器(Model Predictive Control, MPC)控制器的[3]以及基于机器学习[4]的方法。前两种方法在求解时对初始条件敏感性较高并且在每一时间步中求解都有可能是计算密集型的，这会导致实时系统中的延迟，这对于时间敏感的自动驾驶任务来说是无法容忍的。随着计算能力的增长，深度强化学习(Deep Reinforcement Learning, DRL)[5]极大地提高了智能体解决大规模复杂问题的能力。

真实世界中的交通情景复杂多变，单个车辆进行感知决策往往存在着激光雷达和摄像头等传感器存在视距感知以及其他交通参与者的意图不明问题。车辆之间通过相互协作可以很好的解决上述问题。联邦学习(Federated Learning, FL)利用了数据和计算资源通常分布在最终用户设备的特点进行计算和资源协同的机器学习模型训练[6]。基于FL的协同训练方法具有以下几个优点：保证用户数据的隐私[7] 、提升训练模型的性能[8]。但是传统的FL是中心化的，这就存在着一些模型更新的质量低下问题，如果中央服务器发生故障或者无法访问，整个系统的训练过程将会受到影响。去中心化联邦学习(Decentralized Federated Learning, DFL)不需要集中式的服务器来协调设备之间的模型更新，减少了对中央服务器的依赖并极大的降低了攻击风险，同时也使模型更加具有扩展性和健壮性[9]。

直接在车辆实体上使用机器学习的方法进行路径规划的效果往往不佳。原因如下：首先，很难为处于高速移动的车辆建立高保真的仿真环境；其次，从车辆自身模拟环境中学到的机器学习方法不能直接部署到现实世界的AVs上[10]；第三，由于数据采集效率和车辆计算能力低下，训练时延对于AVs是不可接受的。为了解决上述问题，本文将注意力转向数字孪生(Digital Twin, DT)技术。DT是网络空间中真实世界的高保真镜像，根据历史数据、传感器数据和物理对象及时反映真实世界的状态[11]。在DT的帮助下，机器学习方法可以很容易地获得真实世界的高保真状态信息并用于模型训练[12]。此外，通过监测现实世界中混合交通系统的变化，模型还可以随着时间的推移不断更新。如何将DT结合DFL+DRL体系结构用于多个AVs协作下的路径规划是一个挑战。

综上所述，本文的工作主要如下：(1)将单个AVs路径规划问题建模成马尔科夫决策过程(Markov Decision Process, MDP)问题，并针对路径规划中的驾驶行为设计了有效的奖励函数。(2)为多个AVs协作的路径规划问题提出了一种DT辅助的分布式联邦强化学习(Decentralized Federated Reinforcement Learning, DFRL)训练框架，在DT的帮助下，本文提出的模型可以快速的获得车辆群体驾驶场景的高保真状态信息并用于训练。训练好的模型可以直接部署到真实世界的AVs上。(3)提出了一种基于数字孪生的可信度加权的去中心化联邦强化学习算法(Digital Twin-based Credibility Weighted Decentralized Federated Reinforcement Learning, DT-CWDFRL)，在DT-CWDFRL中不需要中央服务器来将聚合模型参数，而是通过对车辆的可信度进行比较来选择模型聚合节点，减少了传统联邦学习对中央服务器的依赖的同时还提高系统的鲁棒性等，更好地适应现实世界中动态变化的驾驶场景。

2 系统建模和问题陈述

在本节中定义了自动驾驶车辆群体协作路径规划问题的系统模型，并将其路径规划问题转化为数学问题进行求解。

2.1 系统模型

本文考虑了一段具有多种车辆复杂驾驶场景的路段，如图1所示，真实世界中有N辆AVs，M辆人类驾驶车辆(Human Driving Vehicles, HDV)。每辆AVs都配有若干个传感器，通过传感器收集自身及周围车辆的运行状态等信息，并将采集到的车辆数据通过网络传输到基站(Base Station, BS)，然后在BS内使用物理建模和仿真技术，将车辆的物理特性、驾驶行为等转换为DT。通过将运行中的AVs映射到DT中，可以实现对AVs的实时监测、仿真预测以及决策优化等。在DT环境中，AVs_i在t时刻的物理位置和速度可以表示为$v^t_{\mathrm{A}_i}$和

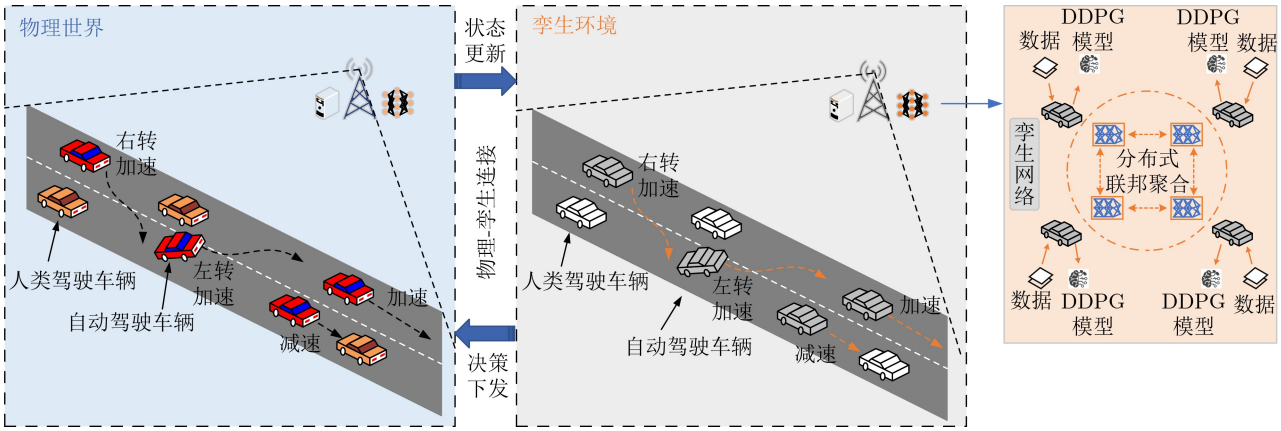

图 1 系统框图

$u_{\mathrm{A}_i}^t$，AVs的速度范围在$[v_{\min}, v_{\max}]$。每个AVs_i都有自己的目的地G_i，自动驾驶车辆群体的目的地集合可以表示为$\mathcal{G}=\{G_1, G_2, \cdots, G_N\}$。本文认为车辆$i$从起点安全行驶到终点为任务完成，记车辆$i$的任务完成时间为$T_i^c$，每辆车都安全到达终点才算车辆群体完成任务。

为了模仿HDV的运动，本文使用智能驾驶模型(Intelligent Driving Model, IDM)和改进的最大限度地减少变道引起的整体制动模型(Minimize Overall Braking Induced by Lane change, MOBIL)来控制HDV，HDV根据IDM模型和MOBIL模型分别输出加速命令和转向角命令，令受控车辆输出转向角以切换到目标车道上。

2.2 问题陈述

本文主要针对以下几个对路径规划影响较大的因素进行了分析。这样既可以在保证驾驶的安全可靠，还方便对算法进行优化。影响路径规划的因素可以总结为以下几个方面：

(1)尽快到达：保证自动驾驶车辆群体在前往目标的过程中整体速度尽可能的大，假设在T时刻所有车辆都抵达目标，上述目标等价于

$$\max \frac{\sum_{t=1}^{T}\sum_{i=1}^{N} v_i^t}{T} \tag{1}$$

(2)避免碰撞：本文采用多种碰撞指标来约束车辆之间的位置和速度。假设AVs之间的安全距离为d_1，AVs和HDV之间的安全距离为d_2，车辆i与其他车辆的碰撞时间记为$\mathrm{TTC}=d/\Delta v$，保证车辆之间不发生碰撞的最小安全时间是T_s，车辆之间的位置和速度关系可以描述为

$$\left.\begin{aligned} &||u_{\mathrm{A}_i}^t - u_{\mathrm{A}_j}^t||_2 \geq d_1 \quad \forall i,j \in \mathcal{N} \\ &||u_{\mathrm{A}_i}^t - u_{H_k}^t||_2 \geq d_2 \quad \forall i \in \mathcal{N}, k \in \mathcal{M} \\ &\mathrm{TTC} \geq T_s \end{aligned}\right\} \tag{2}$$

(3)车辆之间保持连接：在物理世界中，任意两辆AVs之间的通信距离要小于最大通信范围，以保持其连接性，假设每个AVs的最大通信范围为d_3，则AVs之间的位置应该满足

$$||u_{\mathrm{A}_i}^t - u_{\mathrm{A}_j}^t||_2 \leq d_3 \quad \forall i,j \in \mathcal{N} \tag{3}$$

(4)驾驶舒适性：为了使乘客有较好的乘坐体验，AVs要保证在驾驶途中尽量保持车辆的速度变化应较为平缓，并且要符合车辆的物理操作限制，假设AVs的最大加速度为$\mathrm{acc}_{\max}$，转向角δ的最大值是$\delta_{\max}$，则AVs的动作应满足

$$\left.\begin{aligned} &\mathrm{acc}_i^t \in [-\mathrm{acc}_{\max}, \mathrm{acc}_{\max}] \\ &\delta_i^t \in [-\delta_{\max}, \delta_{\max}] \end{aligned}\right\} \tag{4}$$

(5)车辆物理约束：由于AVs在物理世界中的硬件性能限制，其运动轨迹不能在两个连续的时间步长内任意切换，本文假设上一时间步长的速度为v_i^{t-1}，当前时间步长内的车辆速度为v_i^t，最大转向角的限制为$\delta_{\max}$，则车辆的速度必须满足以下条件

$$\arccos\left(\frac{v_i^{t-1}\cdot v_i^t}{|v_i^{t-1}|\times|v_i^t|}\right) \leq \delta_{\max} \tag{5}$$

本文的目标是为车辆群体找到一个在不发生碰撞的前提下，能够快速且舒适的完成任务的最优驾驶策略π^*，以此来最小化车辆群体的平均任务完成时间，即

$$\min_{\pi^*} \frac{1}{N}\sum_{i=1}^{N} T_i^c$$
$$\text{s.t.} \left\{\begin{aligned} &||u_{\mathrm{A}_i}^t - u_{\mathrm{A}_j}^t||_2 \geq d_1 \quad \forall i,j \in \mathcal{N} \\ &||u_{\mathrm{A}_i}^t - u_{H_k}^t||_2 \geq d_2 \quad \forall i \in \mathcal{N}, k \in \mathcal{M} \\ &\mathrm{TTC} \geq T_s \\ &||u_{\mathrm{A}_i}^t - u_{\mathrm{A}_j}^t||_2 \leq d_3 \quad \forall i,j \in \mathcal{N} \\ &\mathrm{acc}_i^t \in [-\mathrm{acc}_{\max}, \mathrm{acc}_{\max}] \\ &\delta_i^t \in [-\delta_{\max}, \delta_{\max}] \\ &\arccos\left(v_i^{t-1}\cdot v_i^t \ /|v_i^{t-1}|\times|v_i^t|\right) \leq \delta_{\max} \end{aligned}\right. \tag{6}$$

其中前3个约束保证AVs与其他AVs以及HDV之间

不发生碰撞，第4个约束保证物理车辆实体之间的连接以保持通信，第5个和第6个约束保证了物理世界中车辆驾驶的舒适度，最后1个约束是车辆实体的硬件约束，保证了车辆的正常驾驶。

3 基于DT-CWDFRL的路径规划算法

3.1 状态、动作以及奖励设置

一个MDP问题通常用一个五元组$(\boldsymbol{S},\boldsymbol{A},\boldsymbol{P},\boldsymbol{R},\gamma)$来表示，对于时间步长$t$处的AVs$_i$，其观测值由以下几部分组成：和目的地之间的相对位置$\boldsymbol{g}_i(t)=[G_i,u_{\mathrm{A}_i}^t]$，之所以采用相对位置关系，是为了减弱模型训练对绝对位置的敏感性；与HDV之间的相对位置$\boldsymbol{h}_i(t)=[u_{\mathrm{A}_i}^t-u_{H_1}^t,u_{\mathrm{A}_i}^t-u_{H_2}^t,\cdots,u_{\mathrm{A}_i}^t-u_{\mathrm{H}_k}^t]$，其中$k\in\mathcal{M}_k$；以及与其他AVs之间的相对位置$\boldsymbol{q}_i(t)=[u_{\mathrm{A}_i}^t-u_{A_1}^t,u_{\mathrm{A}_i}^t-u_{A_2}^t,\cdots,u_{\mathrm{A}_i}^t-u_{\mathrm{A}_j}^t]$，其中$j\in\mathcal{N}_i$。所以AVs$_i$在$t$时刻的观测状态$s_i^t$可以表示为$\boldsymbol{s}_i(t)=[\boldsymbol{g}_i(t),\boldsymbol{h}_i(t),\boldsymbol{q}_i(t)]$。

为了使AVs的轨迹更加平滑，本文的模型采用的是连续动作空间，每个车辆AVs_i的动作由加速度acc和前轮转向角σ组成，表示为$A=[\mathrm{acc},\sigma]$。为了让AVs有更好的驾驶体验，本文定义加速度acc的范围为$\mathrm{acc}\in[-5,5]\ \mathrm{m/s^2}$，转向角$\sigma$的范围为$\sigma\in[-0.25,0.25]$ rad。

对于多个AVs的路径规划任务来说，其主要包括以下几个方面：(1)在驾驶过程中不与周围的车辆发生碰撞，(2)驾驶过程中，AVs之间要保持适当的距离以保证其连接性，(3)驾驶过程中要保证乘客的体验感，即保证驾驶舒适性，(4)驾驶效率，尽快到达目的地。基于这些目的，本文定义奖励函数如下：

(1)碰撞惩罚：这项奖励函数鼓励车辆与其周围的车辆保持安全距离，防止发生碰撞

$$r_1=-r_{\mathrm{HDV}}-r_{\mathrm{AD}}-r_{\mathrm{TTC}} \tag{7}$$

其中

$$\left.\begin{aligned} r_{\mathrm{HDV}}&=\begin{cases} w_c,\|u_{\mathrm{A}_i}^t-u_{H_j}^t\|_2\le d_2\forall i\in\mathcal{N},\forall j\in\mathcal{M}\\ 0,\text{其他}\end{cases}\\ r_{\mathrm{AD}}&=\begin{cases} w_c,\|u_{\mathrm{A}_i}^t-u_{\mathrm{A}_j}^t\|_2\le d_1\forall i,j\in\mathcal{N}\\ 0,\text{其他}\end{cases}\\ r_{\mathrm{TTC}}&=\frac{w_c}{\mathrm{TTC}-T_s+1}\end{aligned}\right\} \tag{8}$$

(2)连接维持奖励：这项奖励旨在引导AVs与其他AVs之间保持连接，以便在协作期间建立更好的合作

$$r_2=\begin{cases} w_{\mathrm{connec}}, & d_1\le\|u_{A_i}^t-u_{A_j}^t\|_2\le d_3\\ 0, & \text{其他}\end{cases} \tag{9}$$

(3)接近目标的奖励：此奖励为了鼓励AVs朝着目的地前进，其基本思想是在每一步中朝着自己的目标点的行进距离尽可能的大，因此在时刻t接近目标的奖励可以定义为$r_3=w_1\|u_{\mathrm{A}_i}^t-G_i\|_2$

(4)舒适度奖励：引导AVs在行驶过程中转向角的改变不要太大，以保证乘坐的舒适性，舒适度奖励定义为$r_4=-w_2(1-(\psi_i^t\times v_i^t)/4)$，其中$\psi_i^t$是$t$时刻AVs的转向角，$v_i^t$是$t$时刻AVs的速度

(5)效率奖励：鼓励AVs在驾驶过程中以较大的速度向着目的地前进

$$r_5=w_3\frac{v_i^t-v_{\min}}{v_{\max}-v_{\min}} \tag{10}$$

因此，自动驾驶车辆群体在时间步长t中的平均奖励函数可以推导为

$$R_t=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{5}r_{i,j} \tag{11}$$

在自动驾驶车辆群体的任务完成时，车辆群体的总累计奖励可以表示为

$$R=\sum_{t=1}^{T}R_t \tag{12}$$

3.2 DT-CWDFRL训练框架

MDP模型通常由动态规划(Dynamic Programming, DP)或者DRL的方法来求解[13]。由于常见的Q学习和深度Q网络算法(Deep Q Network, DQN)并不适用于连续性动作，所以本文为每辆AVs在本地运行深度确定性策略梯度(Deep Deterministic Policy Gradient, DDPG)。

在多个车辆协作的路径规划任务中，规划出的路径好坏取决于协作训练的模型质量。把所有车辆的数据集中起来进行集中式训练的传统方式往往存在着训练模型的更新质量低下、训练速度慢等缺点。而在一般的DFRL框架中，某个智能体可能会为了满足自己的利益而进行不可靠的局部模型更新，从而降低全局模型的聚合质量。

针对上述问题，本文提出了一个DT赋能的CWDFRL训练框架DT-CWDFRL，如图2所示。计算资源有限的AVs$_i$在时刻t的DT可以表示为$\mathrm{DT}_i(t)=(m_i(w_t),f_i(t),s_i(t),z_i(t),o_i(t))$其中$w_t$是车辆的当前训练参数，$m_i(t)$表示车辆的模型训练状态，$f_i(t)$表示车辆的计算能力，$s_i(t)$表示车辆的运行状态，$z_i(t)$表示车辆的任务进行时间，$o_i(t)$表示车辆是否完成任务。由于在对车辆进行DT建模时，很难精确的获取车辆的计算能力，因此CPU频率在真实值和映射值之间存在偏差，为了校准DT的偏

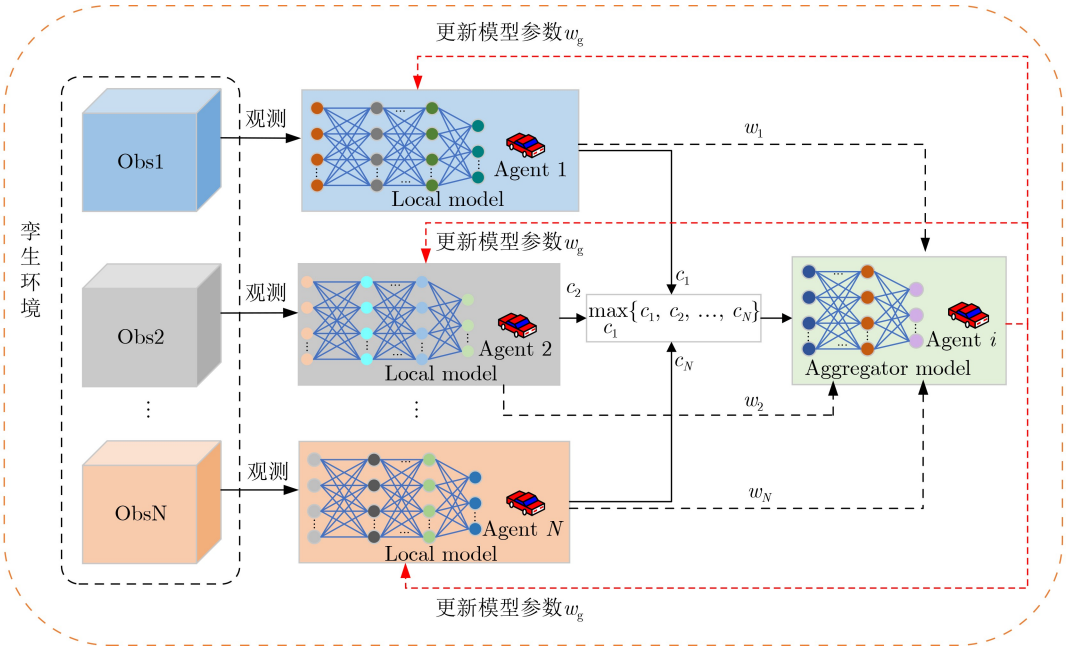

图 2 DT-CWDFRL训练框架

差，本文引入CPU频率偏差$\Delta f_i(t)$来测量DT映射计算能力上的误差，校正映射偏差后，车辆i的DT模型可表示为$\mathrm{DT}_i(t)=(m_i(t), f_i(t)+\triangle f_i(t), s_i(t), z_i(t), o_i(t))$。

每个车辆使用各自的采集到的数据进行基于DDPG的局部模型训练。每个AVs首先通过传感器收集到车辆的实时运行数据，将其同步到其DT中并在DT内训练各自的DDPG模型，其次，每个车辆将各自训练出来的DDPG模型参数进行可信度加权处理，第三，在车辆孪生网络中选择一个可信度最高的车辆作为中心节点，收集每个车辆的训练模型，最后在中心节点将所有车辆的模型参数合并，然后将聚合之后的全局模型参数转移到其他车辆的DT中，然后更新所有车辆的DT模型并进入下一轮训练，直到最后模型收敛。

本文在DT-CWDFRL中提出了一种可信度加权机制，车辆的可信度用$\boldsymbol{c}=[c_1, c_2, \cdots, c_N]$来表示，在每个聚合周期内，都通过对可信度进行比较，选出一个中心节点，并使用上一轮训练中车辆群体的全局模型来衡量当前时隙车辆孪生体DT_i训练出的局部模型的偏离程度，DT_i的模型偏离程度可以用$Q_i^{\mathrm{DT}}(t)$来表示，即$Q_i^{\mathrm{DT}}(t)=\sqrt{|w_i^{\mathrm{DT}}(t)-w_g(t-1)|}$。由于车辆的可信度不仅与局部模型的偏离程度相关，还和DT的映射精确度相关，所以本文在选择中心节点的时候需要同时考虑到这两个因素，因此车辆i的可信度可以建模为

$$c_i(t)=\frac{1}{Q_i^{\mathrm{DT}}(t)}\left(1-\frac{\Delta f_i(t)}{f_i(t)}\right) \tag{13}$$

由式可知，当模型偏离程度$Q_i^{\mathrm{DT}}(t)$越小，车辆作为中心节点的可信度越高，DT的映射偏差$\Delta f_i(t)/f_i(t)$越大时，车辆作为中心节点的可信度就越低。在计算出所有车辆的可信度之后，根据车辆的可信度选择出一个车辆作为模型聚合的节点，中心节点的选择过程可以表示为$\max\limits_{c_i}[c_1, c_2, \cdots, c_N]$，局部模型在中心节点上的聚合过程可以表示为$w^g(t)=\frac{1}{N}\sum_{i=1}^{N} c_i(t) w_i^{\mathrm{DT}}(t)$，最后将聚合之后的模型参数$w_g(t)$发送到各车辆，准备下一次训练，直到整个训练过程收敛。

3.3 CWDF-DDPG算法

4 仿真结果

为了评估本文所采用方案的相关性能，本文将进行大量实验仿真进行数值分析。本文采用的是基于Pytorch的深度学习框架，并使用GYM库中的highway-env环境进行模拟仿真。在每轮迭代开始的时候，AVs随机出现在起点附近，HDV随机分布在起点到终点的路径上。具体仿真参数总结于表1。

为了验证模型的有效性，本文首先把本文提出的算法与其他算法做了损失值和奖励值的对比，通过图3可以看出刚开始系统整体的奖励非常的低，大约在–30左右，这是因为这个时候车辆之间并没有学习到一个很好的策略来进行驾驶，所以车辆大概率都会发生碰撞，但是随着迭代次数的增加，车辆通过经验的积累学习到了一个良好的策略，奖励随之增加并最终稳定在一个区间。由图3、图4可以看出，AVs之间在没有进行协作时效果是最差的，

算法 1 本地车辆训练算法

输入： 车辆数N，噪声n，全局模型参数$w_g=(\pi_g,\theta_g)$

输出： 每个车辆训练模型的可信度c_i

(1) **for** vehicle $\in 1,2,\cdots,N$ **do**

(2) 将全局模型参数同步到本地运行的DDPG网络

(3) 根据当前环境做出动作，并增添随机噪声进行探索：$\boldsymbol{a}(t)=\pi(s(t)|\theta^{\pi})+n$

(4) 执行动作$\boldsymbol{a}(t)$，得到状态$s(t+1)$以及奖励$r(t)$

(5) **if** 经验回放池还没存满 **then**

(6) 将$(\boldsymbol{s}(t),\boldsymbol{a}(t),r(t),\boldsymbol{s}(t+1))$存入经验回放池中

(7) **else**

(8) 用$(\boldsymbol{s}(t),a(t),\boldsymbol{r}(t),\boldsymbol{s}(t+1))$代替经验池中的经验

(9) **end if**

(10) 从经验池中随机选择batch-size条经验构成样本

(11) 通过目标评论家网络得到$Q(\boldsymbol{s}(t+1),\boldsymbol{a}(t+1)|\theta^{Q'})$，计算损失函数$L(\theta^{Q})$

(12) 然后更新估计评论家网络参数θ^{Q}

(13) 根据估计评论家网络得到$Q\left(\boldsymbol{s}(t),\boldsymbol{a}(t)|\theta^{Q}\right)$，用策略梯度更新估计行动家网络参数$\theta^{\pi}$

(14) 软更新目标行动家网络和目标评论家网络的参数$\theta^{\pi'},\theta^{Q'}$

(15) 通过式(13)计算出车辆节点的可信度c_i

(16)**end for**

算法 2 车辆协同训练算法

输入： 回合数M，每回合训练次数step_per_episode，车辆数N，聚合周期A_g

输出： 最优策略π^*

(17) **for** episode $=1,2,\cdots,M$ **do**

(18) 为每辆车初始化全局模型参数$w_g=(\pi_g,\theta_g)$

(19) 初始化环境$s(0)$

(20) **for** step $\in$ step_per_episode **do**

(21) **for** vehicle $\in 1,2,\cdots,N$ **do**

(22) 调用算法1，得到每个车辆节点训练模型的可信度c_i

(23) **if** 处于聚合周期A_g，则需要进行全局模型的更新

(24) 选择聚合节点，聚合全局模型$w_g=(\pi_g,\theta_g)$

(25) 使用全局模型权重$w_g=(\pi_g,\theta_g)$来更新$\theta^{\pi},\theta^{Q},\theta^{\pi'},\theta^{Q'}$

(26) **end if**

(27) **end for**

(28) **end for**

(29) **end for**

系统整体的奖励很低，损失函数也收敛在较高的部分。相比之下，多智能体深度确定性策略梯度(Multi-Agent Deep Deterministic Policy Gradient, MADDPG)，多智能体演员评论家(Multi-Agent Actor Critic, MAAC)以及联邦平均深度确定性策略梯度(Federated average Deep Deterministic Policy Gradient, Fedavg-DDPG)[14]的累计平均奖励值相对较高，这是因为在分布式架构中AVs之间可以通过共享彼此之间的经验相互协作来降低单车智能中存在的安全问题，AVs之间通过协作可以采取比单车智能更优的动作，与MADDPG相比，MAAC解决了MADDPG中存在的信度分配问题，突出了每个智能体对系统整体的贡献，改善了车辆联合训练稳定性差的问题。在Fedavg-DDPG算法中，由于车辆之间只进行训练模型参数的交换，而不进行数据的交换，所以对于计算资源有限的车辆，收敛速度会略快于MADDPG算法。

可以看出，本文提出的算法优于其他对比算法，因为相比于Fedavg-DDPG而言，本文提出的

表 1 CWDF-DDPG仿真参数

仿真参数	值	仿真参数	值
经验池大小	100 000	AVs的速度范围	[0,15]
批尺寸	128	HDV的数量	10
折扣因子	0.99	碰撞系数(w_c)	–50
价值网络学习率	0.001	连接系数(w_{connec})	0.2
策略网络学习率	0.000 1	接近目标系数(w_1)	0.1
软更新系数	0.01	舒适度奖励系数(w_2)	1
最大回合数	10 000	效率奖励系数(w_3)	1
隐藏层单元数	256	AVs之间的安全距离(d_1)	2m
聚合周期	5	AVs与HDV的安全距离(d_2)	2m
免碰撞的最小安全时间	2.5s	AVs的最大通信距离(d_3)	50m

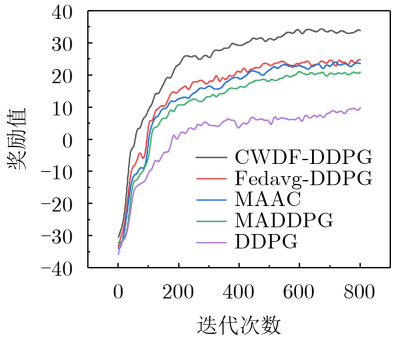

图 3 平均累计奖励

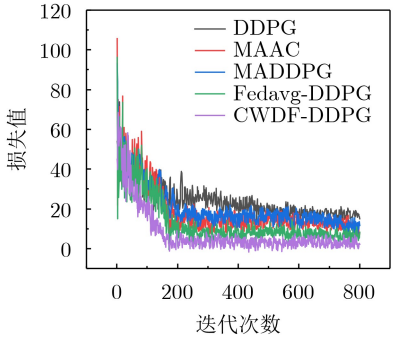

图 4 损失值对比

算法在进行聚合节点选择的时候是根据模型训练的质量来判断的，这就避免了由于拜占庭攻击[15]而导致的低质量全局模型更新从而使得全局模型不能很好的收敛的问题，最终使交通系统整体更快的达到更好的收敛状态。

图5、图6表示的是碰撞概率以及归一化平均转向角与AVs数量之间的关系。可以看出，当只有一辆AVs时，四种基于DDPG的算法的碰撞概率是一样的，这是因为在只有一个AVs的情况下，不管是哪种算法都等效于独立的运行DDPG算法，而单个车辆的AC算法在进行梯度更新时，行动家和评论家网络相互依赖，导致收敛效果并不好。随着AVs数量的增加，无协作的AVs群体的碰撞概率会因为车辆数的增加而增加，对于有协作的AVs群体，由于车辆之间可以通过共享自己学习到的知识从而进行协作，AVs的碰撞概率随之降低。由于highway-env是一个小型的自动驾驶仿真环境，当车辆的数目增加到一定数量时，碰撞概率会随着车辆数的增加而增加。相比于DDPG, MADDPG, MAAC, Fedavg-DDPG，本文提出的算法的平均碰撞概率分别降低约47.22%, 23.12%, 18.13%以及13.29%；在AVs较少的时候，道路上的车流量很小，AVs可以连续的进行变道，所以车辆的转向角很大，当AVs数增加时，尽管车辆之间没有协作，但是车辆自身会尽量避免碰撞而不进行连续的变道操作，可以看出本文提出的方案通过车辆之间的协作可以让车辆的转向角保持在很小的幅度,保证了车辆的驾驶舒适性。

图7、图8表示的是车辆群体的平均任务完成时间以及归一化车辆平均速度与AVs数量之间的关系。可以看出本文所提出的算法可以更充分的利用车辆本身的最大速度，当车流量的增加到一定程度时，车辆对自身速度的利用率明显降低，与之相反的是车辆群体从起点到达终点所消耗的时间会随着车辆速度的增加而降低，即当车辆的速度利用率降低时，车辆群体的平均任务完成时间会增加。

5　结束语

本文研究了AVs群体协作下的路径规划问题，针对车辆之间存在的协作问题，本文提出了一种新的分布式架构，并基于此架构提出了一种数字孪生辅助的可信度加权的去中心化联邦强化学习算法(DT-CWDFRL)，在路径规划问题中，本文把车辆群体的平均任务完成时间建立为优化目标，并使用碰撞概率、归一化平均转向角、归一化平均速度以及平均任务完成时间作为性能指标来衡量。实验结果表明，与DDPG, MADDPG, MAAC以及Fedavg-DDPG算法相比，本文所提算法把碰撞概率分别降低了约47.22%, 23.12%, 18.13%以及13.29%，有效提升了车辆的驾驶安全性和驾驶效率。

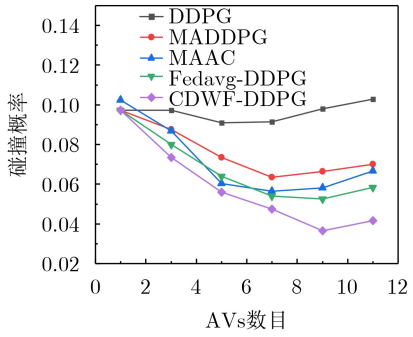

图 5 车辆群体平均碰撞概率

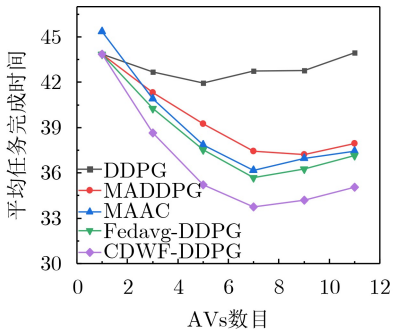

图 7 车辆群体平均任务完成时间

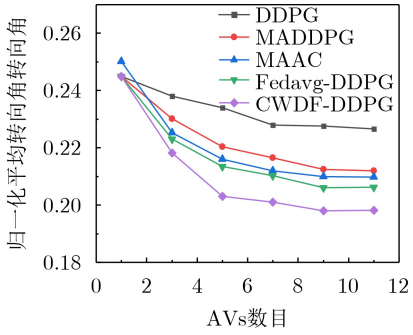

图 6 车辆群体归一化平均转向角

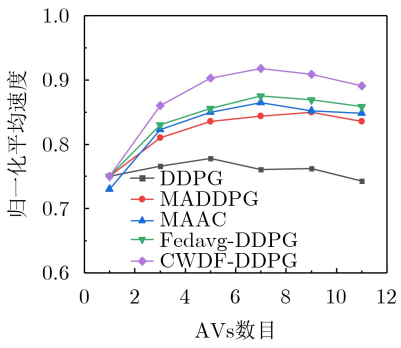

图 8 车辆群体归一化平均速度

参 考 文 献

[1] KIRAN B R, SOBH I, TALPAERT V, *et al*. Deep reinforcement learning for autonomous driving: A survey[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2022, 23(6): 4909–4926. doi: 10.1109/TITS.2021.3054625.

[2] LI Yanqiang, MING Yu, ZHANG Zihui, *et al*. An adaptive ant colony algorithm for autonomous vehicles global path planning[C]. 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD), Dalian, China, 2021: 1117–1122. doi: 10.1109/CSCWD49262.2021.9437682.

[3] ZHOU Jian, ZHENG Hongyu, WANG Junmin, *et al*. Multiobjective optimization of lane-changing strategy for intelligent vehicles in complex driving environments[J]. *IEEE Transactions on Vehicular Technology*, 2020, 69(2): 1291–1308. doi: 10.1109/TVT.2019.2956504.

[4] ZHU Gongsheng, PEI Chunmei, DING Jiang, *et al*. Deep deterministic policy gradient algorithm based lateral and longitudinal control for autonomous driving[C]. 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), Harbin, China, 2020: 740–745. doi: 10.1109/ICMCCE51767.2020.00163.

[5] SHI Dian, DING Jiahao, ERRAPOTU S M, *et al*. Deep *Q*-network-based route scheduling for TNC vehicles with passengers' location differential privacy[J]. *IEEE Internet of Things Journal*, 2019, 6(5): 7681–7692. doi: 10.1109/JIOT.2019.2902815.

[6] KHALIL A A and RAHMAN M A. FED-UP: Federated deep reinforcement learning-based UAV path planning against hostile defense system[C]. 2022 18th International Conference on Network and Service Management (CNSM), Thessaloniki, Greece, 2022: 268–274. doi: 10.23919/CNSM55787.2022.9964907.

[7] LI Yijing, TAO Xiaofeng, ZHANG Xuefei, *et al*. Privacy-preserved federated learning for autonomous driving[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2022, 23(7): 8423–8434. doi: 10.1109/TITS.2021.3081560.

[8] 唐伦, 文明艳, 单贞贞, 等. 移动边缘计算辅助智能驾驶中基于高效联邦学习的碰撞预警算法[J]. 电子与信息学报, 2023, 45(7): 2406–2414. doi: 10.11999/JEIT220797.

TANG Lun, WEN Mingyan, SHAN Zhenzhen *et al*. Collision warning algorithm based on efficient federated learning in mobile edge computing assisted intelligent driving[J]. *Journal of Electronics & Information Technology*, 2023, 45(7): 2406–2414. doi: 10.11999/JEIT220797.

[9] KARRAS A, KARRAS C, GIOTOPOULOS K C, *et al*. Peer to peer federated learning: Towards decentralized machine learning on edge devices[C]. 2022 7th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM), Ioannina, Greece, 2022: 1–9. doi: 10.1109/SEEDA-CECNSM57760.2022.9932980.

[10] SHEN Gaoqing, LEI Lei, LI Zhilin, *et al*. Deep reinforcement learning for flocking motion of multi-UAV systems: Learn from a digital twin[J]. *IEEE Internet of Things Journal*, 2022, 9(13): 11141–11153. doi: 10.1109/JIOT.2021.3127873.

[11] GLAESSGEN E and STARGEL D. The digital twin paradigm for future NASA and U. S. air force vehicles[C]. 53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference, Honolulu, Hawaii, 2012: 1818. doi: 10.2514/6.2012-1818.

[12] TAO Fei, ZHANG He, LIU Ang, *et al*. Digital twin in industry: State-of-the-art[J]. *IEEE Transactions on Industrial Informatics*, 2019, 15(4): 2405–2415. doi: 10.1109/TII.2018.2873186.

[13] 唐伦, 贺兰钦, 谭颀, 等. 基于深度确定性策略梯度的虚拟网络功能迁移优化算法[J]. 电子与信息学报, 2021, 43(2): 404–411. doi: 10.11999/JEIT190921.

TANG Lun, HE Lanqin, TAN Qi, *et al*. Virtual network function migration optimization algorithm based on deep deterministic policy gradient[J]. *Journal of Electronics & Information Technology*, 2021, 43(2): 404–411. doi: 10.11999/JEIT190921.

[14] LIN Qifeng and LING Qing. Byzantine-robust federated deep deterministic policy gradient[C]. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022: 4013–4017. doi: 10.1109/ICASSP43922.2022.9746320.

[15] MA Xu, SUN Xiaoqian, WU Yuduo, *et al*. Differentially private byzantine-robust federated learning[J]. *IEEE Transactions on Parallel and Distributed Systems*, 2022, 33(12): 3690–3701. doi: 10.1109/TPDS.2022.3167434.

唐 伦：男，教授，博士生导师，研究方向为新一代无线通信网络、异构蜂窝网络、软件定义无线网络等.

戴 军：男，硕士生，研究方向为自动驾驶的路径规划、数字孪生、联邦学习优化等.

成章超：男，硕士生，研究方向为车联网、数字孪生、深度强化学习等.

张鸿鹏：男，硕士生，研究方向为车联网中的网络切片，数字孪生，资源分配策略等.

陈前斌：男，教授，博士生导师，研究方向为个人移动通信、多媒体信息处理与传输、下一代移动通信网络、异构蜂窝网络等.

责任编辑：马秀强