



智能系统学报

CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

面向车队云协同计算的深度神经网络推理任务拆分与卸载策略研究

邓贻香, 张玉林, 柏龙

引用本文:

邓贻香, 张玉林, 柏龙. 面向车队云协同计算的深度神经网络推理任务拆分与卸载策略研究[J]. 智能系统学报, 2026, 21(5): 1260–1267.

DENG Yixiang, ZHANG Yulin, BAI Long. Task partitioning and offloading strategy for DNN inference in platooning cloud[J]. CAAI transactions on intelligent systems, 2026, 21(5): 1260–1267.

在线阅读 View online: <https://dx.doi.org/10.11992/tis.202512029>

您可能感兴趣的其他文章

面向机器学习的分布式并行计算关键技术及应用

Key technologies and applications of distributed parallel computing for machine learning

智能系统学报. 2021, 16(5): 919–930 <https://dx.doi.org/10.11992/tis.202108010>

面向车规级芯片的对象检测模型优化方法

Object detection model optimization method for car-level chips

智能系统学报. 2021, 16(5): 900–907 <https://dx.doi.org/10.11992/tis.202107057>

一种基于级联神经网络的飞机检测方法

Cascade convolutional neural networks for airplane detection

智能系统学报. 2020, 15(4): 697–704 <https://dx.doi.org/10.11992/tis.201908028>

多智能体分层强化学习综述

A survey on multi-agent hierarchical reinforcement learning

智能系统学报. 2020, 15(4): 646–655 <https://dx.doi.org/10.11992/tis.201909027>

事件驱动的强化学习多智能体编队控制

Event-triggered reinforcement learning formation control for multi-agent

智能系统学报. 2019, 14(1): 93–98 <https://dx.doi.org/10.11992/tis.201807010>

强化学习的地空异构多智能体协作覆盖研究

Air-ground heterogeneous coordination for multi-agent coverage based on reinforced learning

智能系统学报. 2018, 13(2): 202–207 <https://dx.doi.org/10.11992/tis.201609017>

DOI: 10.11992/tis.202512029

网络出版地址: <https://link.cnki.net/urlid/23.1538.tp.20260708.1638.010>

面向车队云协同计算的深度神经网络推理任务 拆分与卸载策略研究

邓贻香¹, 张玉林¹, 柏龙²

(1. 重庆城市科技学院 电气工程与智能制造学院, 重庆 402160; 2. 重庆大学 机械与运载工程学院, 重庆 400044)

摘要: 随着深度神经网络 (deep neural networks, DNN) 在自动驾驶感知任务 (如目标跟踪、语义分割等) 中的广泛应用, DNN 推理对实时性的要求与车载计算资源受限之间的矛盾日益突出, 已成为制约智能驾驶系统性能提升的重要瓶颈。针对该问题, 本文提出一种面向车队云协同计算环境的 DNN 推理任务拆分与卸载策略。以最小化推理时延为目标, 构建层间-层内双维度协同拆分与卸载模型: 在层间维度, 结合车辆异构计算能力与链路状态, 对 DNN 任务进行自适应拆分; 在层内维度, 针对单层计算负载较重的部分, 进一步进行空间划分并分配给多辆车并行处理, 以缓解局部计算瓶颈。将动态任务拆分与卸载联合优化问题建模为分布式部分可观测马尔可夫决策过程 (decentralized partially observable Markov decision process, Dec-POMDP), 并设计基于多智能体深度确定性策略梯度 (multi-agent deep deterministic policy gradient, MADDPG) 的求解方法, 实现车队协同环境下的动态决策优化。仿真结果表明, 在不同车队规模、计算能力和带宽条件下, 所提方法均能有效降低 DNN 推理时延; 与全本地计算、整体卸载和传统层间拆分方案相比, 所提方法在资源受限场景下表现出更优的时延性能。

关键词: 深度神经网络; 自动驾驶; 车队云; 任务拆分; 计算卸载; 双维度拆分; 多智能体强化学习; 资源分配

中图分类号: TP183 **文献标志码:** A **文章编号:** 1673-4785(2026)05-1260-08

中文引用格式: 邓贻香, 张玉林, 柏龙. 面向车队云协同计算的深度神经网络推理任务拆分与卸载策略研究 [J]. 智能系统学报, 2026, 21(5): 1260-1267.

英文引用格式: DENG Yixiang, ZHANG Yulin, BAI Long. Task partitioning and offloading strategy for DNN inference in platooning cloud[J]. CAAI transactions on intelligent systems, 2026, 21(5): 1260-1267.

Task partitioning and offloading strategy for DNN inference in platooning cloud

DENG Yixiang¹, ZHANG Yulin¹, BAI Long²

(1. College of Electrical Engineering and intelligent manufacturing, Chongqing metropolitan College of Science and Technology, Chongqing 402160, China; 2. College of Mechanical and Vehicle Engineering, Chongqing University, Chongqing 400044, China)

Abstract: With the wide deployment of deep neural networks (DNNs) in autonomous driving perception tasks such as object tracking and semantic segmentation, the contradiction between stringent real-time inference requirements and limited onboard computing resources has become a major bottleneck restricting system performance. To address this issue, this paper proposes a DNN task partitioning and offloading strategy for platooning cloud environments. First, a dual-dimensional collaborative partitioning and offloading model is developed with the objective of minimizing inference latency. In the inter-layer dimension, DNN tasks are adaptively partitioned according to heterogeneous vehicle computing capabilities and link conditions. In the intra-layer dimension, the parts with heavy single-layer computational loads are further spatially partitioned and assigned to multiple vehicles for parallel execution, thereby alleviating local computational bottlenecks. Second, the joint optimization problem of dynamic task partitioning and offloading is formulated as a decentralized partially observable Markov decision process (Dec-POMDP), and a multi-agent deep deterministic policy gradient (MADDPG) based solution is designed to achieve dynamic decision optimization in platooning cloud environments. Simulation results show that the proposed method can effectively reduce DNN inference latency under different platoon scales, computing capacities, and bandwidth settings. Compared with full local computing, whole-model offloading, and conventional inter-layer partitioning schemes, the proposed method achieves superior latency performance, especially under resource-constrained conditions.

Keywords: deep neural networks; autonomous driving; platooning cloud; task partitioning; computation offloading; dual-dimensional splitting; multi-agent reinforcement learning; resource allocation

收稿日期: 2025-12-16. 网络出版日期: 2026-07-09.

基金项目: 重庆市教委科学技术研究项目 (KJQN202402501); 国家自然科学基金项目 (51975070); 重庆市教委科学技术研究项目 (KJZD-K202202502)。

通信作者: 邓贻香. E-mail: 342717900@qq.com.

车队作为自动驾驶领域的新兴范式, 其独特的驾驶模式和丰富资源潜力为智能交通系统带来了创新机遇^[1]。车队由多辆车组成, 其中一辆车

担任领导角色, 其余车辆则跟随。首先, 车队中的各辆车通过车载传感器收集自身周围环境感知数据, 并将收集的数据通过车车通信 (vehicle to vehicle, V2V) 技术传输给领导车。然后, 领导车将收集和接受到的感知数据进行分析得到驾驶结果, 并通过广播共享驾驶信息给车队的成员, 从而自动确定每个跟随者的转向、速度和制动, 以提高整体交通安全性^[2]。

近年来, 许多驾驶辅助应用 (如行人检测、障碍物回避、增强现实导航等) 逐渐涌现并取得了显著的发展^[3-4]。这些应用通常依赖于深度学习技术 (例如, 深度神经网络 (deep neural networks, DNN)) 来获得更高的性能^[5]。这些应用的计算服务通过预训练的 DNN 模型进行推理。推理时间是评估用户体验的一个关键因素。推理时间不仅对于确保安全和改善驾驶体验至关重要, 还可能影响交通拥堵。然而, 执行 DNN 推理任务需要大量计算资源。这些 DNN 推理任务通常由于车载计算资源的限制而在车辆上表现出较差的用户体验。传统的模型优化方法, 如权重剪枝和精度降低, 虽然可以在一定程度上减轻计算负担, 但在面对如此大规模的感知数据时, 单一车辆的有限计算能力仍然难以胜任。

为应对上述挑战, 任务卸载与模型拆分技术被认为是提升 DNN 推理实时性的有效途径^[6]。现有研究大多采用层间线性拆分策略^[7-8], 即将 DNN 推理过程划分为本地执行段和远端执行段。该方法虽在一定程度上缓解了单车算力不足的问题, 但仍存在两方面局限: 一方面, 在车队云协同环境下, 中间特征数据的传输性能受限于链路带宽, 容易引入较大的通信时延; 另一方面, DNN 推理任务整体计算开销较大, 且部分层的单层计算负载较重, 若仍由单车独立完成, 仍可能形成显著的局部计算瓶颈。因此, 仅依赖层间线性拆分难以同时兼顾通信与计算两方面的优化需求。基于此, 本文进一步结合车队内车辆异构算力与链路状态, 面向车队云协同计算环境, 提出一种层间 (垂直) 与层内 (水平) 协同的双维度任务拆分与卸载机制。在层间维度上, 结合车队内车辆异构算力与链路状态, 对 DNN 推理任务进行自适应拆分; 在层内维度上, 针对单层计算负载较重的部分, 进一步进行空间划分并分配给多辆车并行处理, 以缓解局部计算瓶颈。在此基础上, 本文将以最小化推理时延为目标的任务拆分与卸载联合优化问题建模为分布式部分可观测马尔可夫决策过程 (decentralized partially observable Markov decision process, Dec-POMDP), 并采用

多智能体深度确定性策略梯度算法 (multi-agent deep deterministic policy gradient, MADDPG) 进行求解^[9-11]。仿真结果表明, 所提方法在不同车队规模、计算能力和带宽条件下均能有效降低推理时延, 且在资源受限场景下表现出更优性能优势。

1 网络模型

1.1 场景描述

如图 1 所示, 本文构建了一个 I 个车队列场景。车队中的车辆用索引集 $\{V_i | V_{i,j}, V_{i,0}, V_{i,1}, \dots, V_{i,I_i}\}$ 表示, 其中 $V_{i,0}$ 为领导车, 其他车辆为跟随车辆。车队中每辆车共同感知周围环境, 车辆协同执行推理任务。在此场景下, 每个车队有一个推理任务 (DNN 模型), 车队成员共享相同的通信信道, 由领导车根据车辆的计算能力和网络状况决定 DNN 的水平 and 垂直拆分策略。本文将车队车辆的网络建模为一个无向图 $G_{i,j} = (V_{i,j}, \xi_{i,j})$, 其中 $V_{i,j}$ 是节点集合, $\xi_{i,j}$ 是边集合。任意两个节点 $V_{i,j}, V_{i,j'} \in V_{i,j}$ 之间的链路权重可以通过加权邻接矩阵。邻接矩阵中的链路权重 $V_{i,j}, V_{i,j'} \in V_{i,j}$ 代表了车辆 $V_{i,j}$ 到车辆 $V_{i,j'}$ 的数据传输速率 $R_{i,j,j'}$ 。每个车辆 $V_{i,j}$ 的 CPU 频率为 $f_{i,j}$ 。图 2 给出了一个包含 4 辆车的示例车队网络拓扑图, 每条边的权重 $R_{i,j,j'}$ 表示车辆 $V_{i,j}$ 和车辆 $V_{i,j'}$ 之间的传输速率, 每个顶点由资源向量 $f_{i,j}$ 表征。

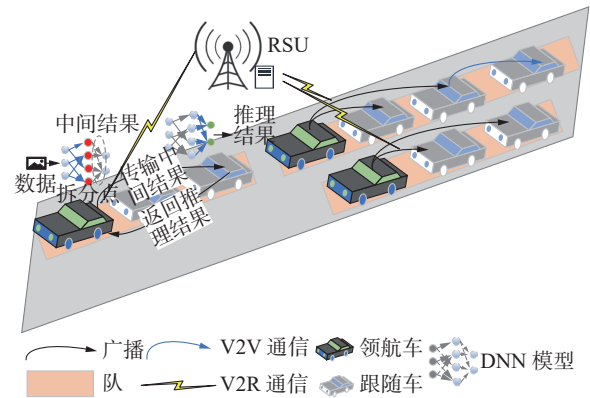


图 1 车队内车辆间协同推理的模型分割

Fig. 1 Model splitting for collaborative inference among vehicles in a fleet

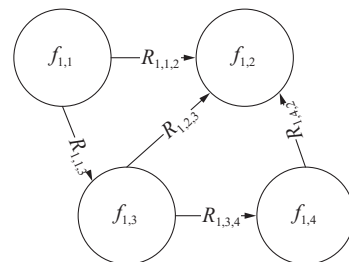


图 2 单车队的网状网络拓扑示例

Fig. 2 Example of a mesh network topology for a platoon

给定一个 DNN 模型, 记 $L = \{1, \dots, l, \dots, L\}$ 为其层索引集合^[12]。对于卷积神经网络 (convolutional neural network, CNN) 模型, 每层可以为卷积层、池化层、平坦层或密集层。用配置向量 $(\psi, s, c_{in}, c_{out}, p, H)_l$ 表示第 $l \in L$ 层的结构特性, 即卷积核大小 ψ_l 、步幅 s_l 、填充 p_l 、输入通道数 $c_{in,l}$ 、输出通道数 $c_{out,l}$ 以及该层输入特征图的高度 H_l 。进一步地, 记输入特征图的高度和宽度分别为 H 和 W ^[13]。在水平维度上, 本文采用基于空间维度的特征图划分方式。具体而言, 对某一层输入特征图仅沿高度和宽度 2 个空间维度进行切分, 而每个空间分区均保留该分区对应位置上的全部通道信息。因此, 本文的层内拆分属于空间划分而非通道划分。

以前的 CNN 只按层拆分, 其中每层都是一个分区 (即神经网络有向无环图 (directed acyclic graph, DAG) 中的一个顶点代表一个层)。给定一个层数为 $L = \{1, \dots, l, \dots, L\}$, 生成的分区图将具有 $L = \{\lambda_1, \lambda_2, \dots, \lambda_L\}$ 个节点依次连接^[14]。与先前的方案不同, 本文细化了任务拆分, 不再限制拆分只能包含整个层, $\Lambda = \{\lambda_1, \dots, \lambda_k, \dots, \lambda_A\}$ 为 DNN 的分区集, 其中分区 λ_1 由一对 (γ_1, ω_1) 定义, γ_1 和 ω_1 为分区 λ_1 的高度和宽度。特别地, 分区 λ_k 覆盖了层的部分输入 γ_k , 并且每个分区 λ_k 跨越了多个连续的层 ω_k 。为了确保模型 (或子模型) 中层之间的正确执行顺序以及它们之间的中间数据流, 分区集合 Λ 必须遵循由描述的图中的方向。图 3 给出了神经网络分区示例, 其中 CNN 被分割为一组分区 $\{\lambda_1, \lambda_2, \dots, \lambda_5\}$, 这些元素通过有向边连接形成一个 DAG, 其中 5 个分区的任务分配给 4 个用户执行。分区间数据量 $D_{k,k'}$ 为分区 λ_k 传输到 $\lambda_{k'}$ 的数据量。所有分区的并集等效于总分区集合, 即: $\bigcup_{k=1}^A \{\lambda_k\} = \Lambda$ 。同样, 所有分区的交集等效于一个空集, 即: $\bigcap_{k=1}^A \{\lambda_k\} = \emptyset$ 。

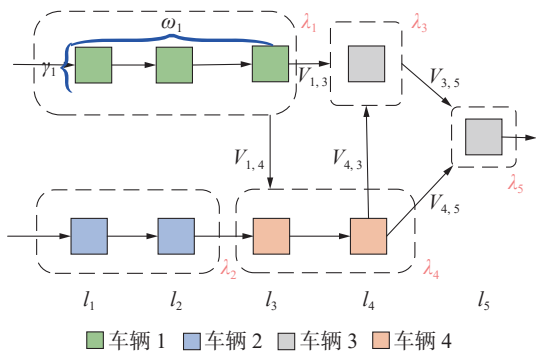


图 3 多车协作推理的工作流程

Fig. 3 Workflow of multi-vehicle collaborative inference

每个分区最多执行一次并且分区的计算之间存在顺序, 因此 CNN 模型表示为一个有向无环图 DAG $G_\Lambda = (\lambda, \varepsilon)$, 其中 λ 为顶点集, ε 为边集。加权邻接矩阵 D 的元素 $D_{k,k'}$ 是任意两个顶点 $\lambda_k, \lambda_{k'} \in \lambda$ 之间的链路权重, 其值表示由分区 λ_k 产生的输出并作为输入提供给分区 $\lambda_{k'}$ 的数据量 $D_{k,k'}$ 。邻接矩阵描述了神经网络 DAG 中任意两个分区之间是否存在数据交换以及需要传输的数据量:

$$D_{k,k'} = \begin{cases} D_{k,k'}, & (k, k') \in \varepsilon \\ 0, & (k, k') \notin \varepsilon \end{cases} \quad (1)$$

$\mathbf{s}_i = (s_{i,1}, \dots, s_{i,A}) \in V^{i,A}$ 为匹配向量。如果 $s_{i,k}$ 表示分区 λ_k 分配给 V_{i,λ_k} , 则 $s_{i,k} = 1$; 否则, $s_{i,k} = 0$ 。在完成分区 λ_k 所需的计算后, 执行分区 λ_k 的车辆需要将中间数据发送到执行下一个分区的车辆, 直到推理任务完成。每层的输入特征都是前一层的输出特征, 因此推导出车辆在计算分区后需要发送的输出数据的确切大小。以卷积层 l 为例, 设分区 λ_k 是该层在空间维度上的一个子分区, 其分区宽度为 ω_k , 配置向量为 $(\psi, s, c_{in}, c_{out}, p, H)_l$ 。若输入到该分区的特征图大小为 (γ, W, c_{in}) , 其中 γ_k 表示该分区覆盖的输入高度, W_k 表示对应输入宽度, 则该分区的输出特征图大小可表示为

$([\gamma - \psi + 2p/s] + 1, [W - \psi + 2p/s] + 1, c_{out})$ (高度、宽度、通道数)。可以看出, 在本文所采用的水平拆分方式中, 分区后的输出特征图在空间维度上被划分, 而通道维度保持完整, 不对输入通道或输出通道进行进一步切分。图 4 给出了不同车辆执行多个空间分区时的数据流示意。其中, 分区之间主要交换的是由空间划分引起的中间特征数据和边界填充数据, 图中给出了各分区输入高度及填充数据规模的示例。

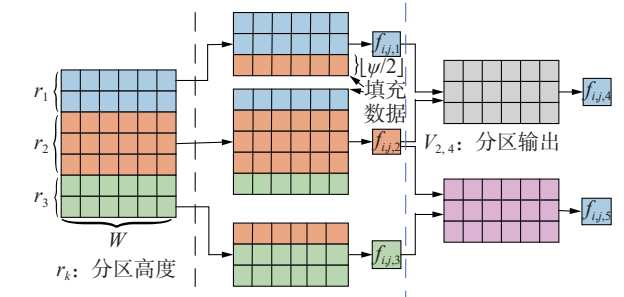


图 4 分区数据的示例场景

Fig. 4 Example scenarios of partitioned data

设一个返回在第 l 层的划分 λ_k 的高度函数 $\gamma_k(l)$ 。对于第 l 层, 所有部分覆盖该层的划分高度之和必须等于该层的输入大小 Γ_l , 即

$$\sum_{k: \lambda_k \in \Lambda} \gamma_k(l) = \Gamma, \quad \forall l \in L \quad (2)$$

设 γ 为从 DAG G_Λ 的起始划分到结束划分之

间的所有可能路径的集合。设 $v \in \mathcal{V}$ 为从起始到结束划分之间的一条路径, 其中 v 为 DAG G_A 的一个子图, 即 $v = (G_v, E_v)$ 。需要注意的是, G_v 表示路径中的顶点集, 而 E_v 表示路径中的边集。在任何给定路径中, 该路径中所有划分的宽度 ω_k 之和等于 DNN 模型的层数, 即

$$\sum_{k: \lambda_k \in v} \omega_k = |L|, \forall v \in \mathcal{V} \quad (3)$$

为了使划分可接受, 分配在不同 $V_{i,k}$ 和 $V_{i,k'}$ 上的任何两个分区 λ_k 和 $\lambda_{k'}$ 之间的任何边必须位于 $V_{i,k}$ 和 $V_{i,k'}$ 之间现有的通信边上, 即

$$\forall \{\lambda_k, \lambda_{k'}\} \in \Lambda: V_k \neq V_{k'} \Rightarrow \{V_k, V_{k'}\} \in \xi \quad (4)$$

1.2 通信模型

网络的带宽被划分为 M 个正交的子信道, 由集合 $\mathcal{K} \triangleq \{1, \dots, m, \dots, M\}$ 表示, 每个子信道带宽都是 B Hz。车队可被分配到一个或多个子信道。因此, 车队 i 中车辆通过信道 m 传输数据到车辆的传输速率:

$$R_{i,k,k'} = B \sum_{m=1}^M \beta_{i,m} \log_2 \left(1 + \frac{P_0 |h_{k,k}^{(m)}|^2}{\sigma^2 + \sum_{i'=1, i' \neq i}^M \beta_{i',m} \cdot P_0 |h_{k,k}^{(m)}|^2} \right) \quad (5)$$

式中: $|h_{i,k,k'}^{(m)}|$ 是从车辆 $V_{i,k}$ 通过子信道 m 传输数据到车辆 $V_{i,k'}$ 的信道增益, P_0 为车辆 $V_{i,k}$ 的传输功率, $\beta_{i,m}$ 表示信道 m 是否分配给车队 i 和 σ^2 是噪声功率。

为了计算分区 λ_k , 需要从其他分区输入或填充数据。定义邻域内 $N^-(\lambda_k) = \{\lambda_{k'} \in \Lambda: (\lambda_k, \lambda_{k'}) \in \epsilon\}$ 作为向分区 λ_k 发送分区间数据的分区集合。本文定义邻域外 $N^+(\lambda_k) = \{\lambda_{k''} \in \Lambda: (\lambda_k, \lambda_{k''}) \in \epsilon\}$ 作为从分区 λ_k 接收填充数据的集合。 $D_{k,k'}$ 取决于 (γ_k, ω_k) 和 $(\gamma_{k'}, \omega_{k'})$, 以及分区 λ_k 和 $\lambda_{k'}$ 所覆盖的层的构型向量。因此, 在车辆 $V_{i,k}$ 上从邻居 $N^+(\lambda_k)$ 接收分区数据所需的接收时间:

$$T_{i,k}^r = \sum_{\lambda_{k'} \in N^-(\lambda_k)} \frac{S_{i,k} D_{k,k'}}{R_{i,k,k'}} \quad (6)$$

将填充数据发送到车辆 $V_{i,k}$ 上所需的需要传输时延:

$$T_{i,k}^t = \sum_{\lambda_{k''} \in N^+(\lambda_k)} \frac{S_{i,k} D_{k'',k}}{R_{i,k'',k}} \quad (7)$$

传输分区 λ_k 的中间数据所需的总时间:

$$T_{i,k}^l = T_{i,k}^r + T_{i,k}^t \quad (8)$$

1.3 计算模型

如果分区 λ_k 的输入特征图大小为 $(\gamma_k, W_k, c_{in,k})$, 输入到分区 λ_k 的数据量 $D_k = \gamma_k \cdot W_k \cdot c_{in,k}$ 。因此, 在车辆 $V_{i,j}$ 上执行分区 λ_k 所需的计算时间:

$$T_{i,k}^c = \frac{S_{i,k} D_k}{F_{i,k}} \quad (9)$$

式中 $F_{i,j}$ 为 $V_{i,j}$ 的 CPU 频率。

1.4 问题建模

在通信和计算资源的约束下, 通过联合优化模型拆分和子信道分配, 以最小化总推理时延。问题描述为

$$\min \sum_{k=1}^K T_{i,k} = \sum_{k=1}^K (T_{i,k}^r + T_{i,k}^c) \quad (10)$$

s.t. (2), (3), (4)

2 分布式多智能体强化学习算法

上述问题是非确定性多项式问题, 本文将该优化问题通过分布式多智能体强化学习算法求解。考虑车队 i 的本地观测能力有限, 仅能观测到自身任务信息、位置信息和车队内车辆计算资源信息等部分环境信息, 本文将问题进一步建模成一个有限时域下的 Dec-POMDP 问题^[15], 其中各车队的领导车可仅根据本地观测信息即可独立做出其信道选择和模型拆分动态决策。

1) 智能体: 各车队的领导车 $V_{i,0}$ 均作为智能体完成其各自的信道选择和模型拆分决策。

2) 本地观测: 各车队观测 o_i 可表示为 $[h_{i,k,k'}^{(m)}, F_{i,k}]$ 。

3) 动作: 各车队独立选择信道选择和模型拆分决策, 则车队 i 的动作车队 i 的动作 a_i 可表示为 $[S_{i,k}, \lambda_k]$ 。

4) 系统奖励: 各车队所获得的来自环境的系统奖励 r_i , 即: $r_i = - \sum_{k=1}^K T_{i,k}$

本文采用了 MADDPG 算法^[16]。DDPG 的核心思想是基于 Actor-Critic 架构。Actor 网络负责生成动作, 通过最大化预期回报来更新网络权重; Critic 网络则负责估计状态-动作值函数, 通过最小化 TD 误差来更新网络权重。这 2 个网络通过梯度下降的方式进行优化, 使得智能体的行为逐渐接近最优策略。MADDPG 框架由环境和 I 个智能体组成。在多智能体环境中, 每个智能体在独立更新策略的过程中可能会导致训练不稳定。为了解决多智能体环境的这一不稳定性并实现智能体协作, 本文采用集中式训练和分布式执行的方式, 其中每个智能体有两个阶段: 集中式训练和分布式执行。

在集中式离线训练阶段, Critic 网络根据所有智能体的动作和观察全局状态信息计算集中式动作价值函数 $Q(s, a_1, a_2, \dots, a_k; \theta^c)$ 。集中式 Q 函数从全局视角评估智能体的动作, 并利用这一点来

指导智能体选择更好的动作。然后, Critic 网络通过最小化损失来更新参数

$$\text{Loss}(\theta^c) = E[(Q(s, a_1, a_2, \dots, a_I; \theta) - y)^2] \quad (11)$$

式中 $y = r + \gamma Q^*(s', a'_1, a'_2, \dots, a'_I; \theta')$, θ 是 Critic 网络的参数, $s = (o_1, o_2, \dots, o_I)$ 。同时, Actor 网络根据由 Critic 计算的集中式 Q 函数及其自身的观察信息更新网络参数 θ^a 并输出动作 a 。具体来说, Actor 网络在 $\nabla_{\theta^a} J(\pi)$ 方向上直接调整网络参数 π , 由式给出

$$\nabla_{\theta^a} J(\theta^a) \approx E[\nabla_a Q(s, a_1, a_2, \dots, a_I; \theta) \nabla_{\theta^a} \mu(s, \pi)] \quad (12)$$

式中 θ^a 是 Actor 网络的参数。

在分布式执行阶段, Actor 网络根据其自身状态输出动作。在这个执行过程中, 每个智能体都可以获得接近全局最优的动作 $a_i = \mu(o_i, \pi)$, 而无需了解其他智能体的信息。基于 MADDPG 的训练与执行过程如算法 1 所示。

算法 1 基于集中式训练、分布式执行的 MADDPG 训练与执行过程

A 集中式训练阶段

- 1) 初始化训练回合 J , 训练步数 T ; 软更新因子 τ ; 回放缓冲区 Z 。
- 2) for 回合 $j=1, 2, \dots, J$ do
- 3) 初始化车队环境;
- 4) for 步数 $t=1, 2, \dots, T$ do
- 5) for $i=1, 2, \dots, I$ do
- 6) 收集本地观测 $o_{i,t}$, 根据 Actor 网络选择动作;
- 7) end for
- 8) 执行动作 $a_{i,t}$;
- 9) 所有车队得到环境所反馈的系统奖励 r_t 和新的本地观测 $o_{i,t+1}$;
- 10) 经验 $[o_{i,t}, a_{i,t}, r_t, o_{i,t+1}]$ 存储到共享经验池 Z ;
- 11) end for
- 12) 从共享经验池采样获得经验集合;
- 13) 通过最小化损失函数 L 更新在线网络参数;
- 14) 在目标网络中软更新 Critic 和 Actor 网络的参数。

$$\theta^* \leftarrow \tau \theta + (1 - \tau) \theta^* \quad \pi^* \leftarrow \tau \pi + (1 - \tau) \pi^*$$

15) end for

B 分布式执行阶段

- 16) 对于每个智能体 i , 仅保留训练好的 Actor 网络;
- 17) 在在线执行阶段, 智能体 i 根据本地观测 $o_{i,t}$ 独立输出动作 $a_{i,t}$;
- 18) 各智能体无需访问其他智能体观测, 也无需调用集中式 Critic, 即可完成分布式协同决策。

3 仿真结果及分析

为验证本文提出的层间-层内双维度拆分 (inter-intra-layer splitting, IILS) 方案的有效性, 设置以下基准方案进行对比分析:

- 1) 本地计算 (local-computing, LC): 所有 DNN 推理任务在领航车辆上执行, 无任务卸载;
- 2) 全卸载计算 (offloading-computing, OC)^[17]: 将完整 DNN 模型卸载至随机选择的跟随车执行, 无任务拆分;
- 3) 分层卸载 (layer-splitting, LS)^[18]: 仅采用层间拆分, 不执行层内拆分。

实验采用 Python 语言实现, 主要基于 TensorFlow 和 Numpy 搭建深度神经网络推理与多智能体强化学习训练环境。仿真场景设置为单小区车队云协同计算环境, 车队车辆在路边单元覆盖范围内行驶, 所有车辆均位于距离路侧单元 500 m 的范围内随机分布。各车队沿道路同向行驶, 车辆行驶速度统一设为 60 km/h。车队内车辆通过 V2V 链路进行中间特征数据和分区结果的传输, 车辆之间的最大通信范围设置为 70 m, 当两车间距离超过该范围时, 认为二者无法建立直接通信链路。系统总带宽 B 的取值范围为 150 ~ 400 MHz。除非特别说明, 其他仿真参数设置如表 1 所示。

表 1 实验参数
Table 1 Experimental parameters

参数	值
噪声功率/dBm	-114
V2V 传输功率/dBm	23
每个队列最多车辆数量	10
车队最大的计算能力/GHz	4

图 5 给出了不同车队数量条件下各方案的推理时延变化趋势。随着道路中车队数量的增加, IILS、OC 和 LS 方案的推理时延均呈上升趋势, 而 LC 方案基本保持稳定。其原因在于, 随着车队数量增加, 系统中并发推理任务数量同步上升, 多个车队需要竞争有限的无线通信资源, 从而导致中间特征数据传输时延增加, 最终推高整体推理时延。相比之下, LC 方案始终由领导车本地完成推理, 不涉及车队间或车队内的中间结果传输, 因此基本不受通信资源竞争影响^[19]。进一步观察可知, IILS 方案在所有车队规模下均保持最低推理时延, 说明所提双维度拆分机制在高并发协同场景下具有更强的资源适应能力。具体而言, LS 方案虽然能够通过层间拆分缓解部分单车算力不足问题, 但其拆分粒度仍局限于层级边

界,无法进一步处理单层内部的重负载计算;OC方案则将完整模型交由单一节点处理,虽然减少了拆分管理开销,但容易受限于单节点计算能力和整模型传输负担^[20]。与之相比,IILS同时结合层间拆分与层内空间划分,一方面能够依据链路状态和计算能力选择更优的层间卸载位置,另一方面能够将单层重负载部分进一步分散至多辆车并行处理,从而在车队数量增加、通信竞争加剧的条件下,仍能有效抑制推理时延增长^[21]。该结果表明,所提方法不仅适用于低负载场景,在多车队并发推理环境下同样具有较好的可扩展性。

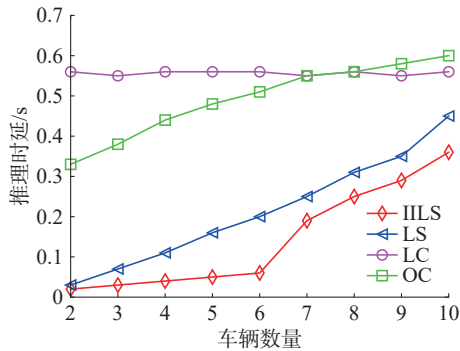


图5 不同方案下的推理时延

Fig. 5 Inference latency under different schemes

图6给出了不同车队总计算能力条件下各方案的推理时延变化。可以看出,随着车队总计算能力的提高,IILS、OC和LS方案的推理时延均明显下降。这说明在当前协同推理场景中,计算资源仍是影响系统性能的重要因素之一:更高的计算能力能够缩短分区执行时间,从而降低整体任务完成时延。LC方案同样会从本地算力提升中受益,但由于其始终依赖单车本地执行,无法利用车队内其他车辆的可用算力,因此整体性能始终劣于协同计算类方案。IILS在各算力区间内均优于LS和OC,说明其性能提升并非仅来源于“更多卸载”,而是来源于更细粒度的计算任务组织方式^[22]。OC方案虽然借助远端节点执行完整模型,但忽略了带宽受限条件下的中间数据传输代价;LS方案虽然具备层间任务迁移能力,但当某些层本身计算负载较重时,仅依赖层间边界调整难以消除单层内部的局部计算瓶颈^[23]。相比之下,IILS方案在层间拆分基础上进一步引入层内并行处理,使得高计算负载层能够被多个节点同时承担,因此在中低算力区间内优势尤为明显^[24]。此外,当计算能力超过3.5 GHz后,IILS与LS之间的差距逐渐缩小,这表明系统主导瓶颈正在发生转移:在低算力条件下,单层计算瓶颈更为突出,层内拆分能够带来显著收益;而当车队整体计算能力持续增强后,单层计算压力被明显削

弱,系统性能逐步由“计算主导”转向“通信与调度共同主导”,此时层内进一步细粒度拆分所带来的边际收益开始减小^[25]。因此,该结果说明所提双维度拆分机制在算力受限场景下优势最为显著,而在高算力场景下则依然保持稳定优于基线方案的性能表现。

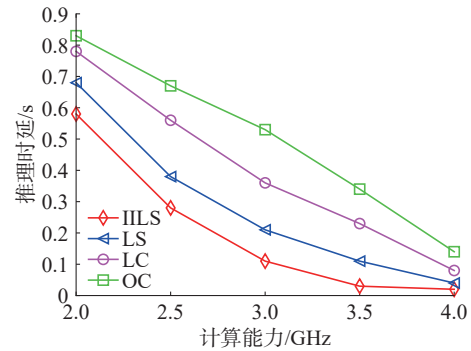


图6 不同方案下的推理时延

Fig. 6 Inference latency under different schemes

图7给出了不同带宽条件下各方案推理时延的变化趋势。随着带宽增加,IILS、OC和LS方案的推理时延均呈下降趋势,而LC方案基本保持不变。其根本原因在于,带宽提升能够直接改善中间特征数据与分区间结果的传输效率,从而降低协同推理过程中的通信时延;而LC不涉及车间数据传输,因此不受带宽变化影响。该现象说明,对于协同推理类方案,通信资源依然是影响总推理时延的重要外部因素^[24]。进一步比较可见,IILS在整个带宽区间内始终保持最低推理时延,说明所提方法的优势并不仅仅来自链路条件改善,而在于其能够更有效地协调“传输开销”与“计算开销”之间的关系。对于OC方案,尽管带宽增加能够降低模型整体卸载带来的传输代价,但其仍受限于单节点计算瓶颈;对于LS方案,带宽改善虽然增强了层间卸载效率,但其无法进一步缓解单层内部的重负载计算问题,因此性能提升仍然有限^[26]。相比之下,IILS通过在层间维度选择更合适的拆分节点、在层内维度进一步并行处理重负载层,使得通信资源的增长可以被更充分地转化为整体时延收益^[27]。此外,从曲线变化趋势还可以看出:在低带宽区间,IILS相较LS和OC的优势更加明显,表明当系统处于通信受限状态时,双维度拆分机制能够更有效地压缩不必要的传输代价,并降低由不合理拆分导致的通信放大效应^[28-29];而在高带宽区间,各方案差距有所缩小,但IILS仍保持最优,这进一步说明所提方法并非仅在通信受限场景下有效,其在通信条件改善后仍可凭借更细粒度的任务划分维持性能优

势^[30]。换言之, ILS 同时具备较好的弱通信鲁棒性和高带宽适应性。

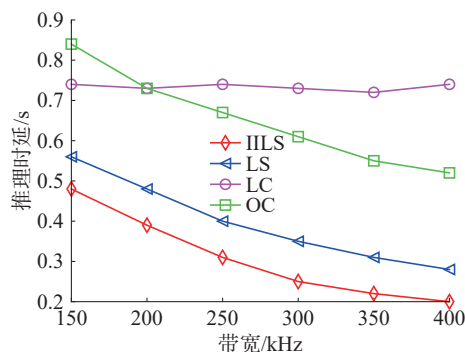


图 7 不同带宽下的推理时延

Fig. 7 Inference latency under different bandwidth

4 结束语

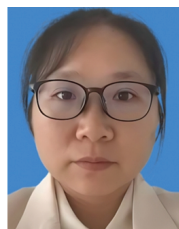
本文面向车队协同计算环境下的 DNN 推理任务优化问题, 提出了一种层间-层内双维度协同拆分与卸载方法。该方法在层间维度上依据计算与通信代价实现任务的自适应拆分, 在层内维度上针对单层重负载部分进行并行处理, 从而有效缓解单车算力受限和局部计算瓶颈问题。随后, 本文将任务拆分与卸载联合优化问题建模为 Dec-POMDP, 并采用多智能体强化学习方法进行求解。仿真结果表明, 所提方法在不同车队数量、计算能力和带宽配置下均能有效降低推理时延, 且在资源受限条件下相较于基准方案具有更显著的性能优势。未来研究将进一步面向车队高速移动、网络状态动态变化和异构计算资源协同场景, 研究更加鲁棒、低开销且可在线自适应的 DNN 推理任务拆分、卸载与资源分配优化方法。

参考文献:

- [1] 李克潮, 张继成. 多阶特征向量融合的深度学习产品推荐算法[J]. 计算机工程与设计, 2025, 46(6): 1742-1749.
LI Kechao, ZHANG Jicheng. Deep neural network product recommendation algorithm based on multi-level feature vector fusion[J]. Computer engineering and design, 2025, 46(6): 1742-1749.
- [2] XU Yuanyuan, ZHU Kun, XU Hu, et al. Deep reinforcement learning for multi-objective resource allocation in multi-platoon cooperative vehicular networks[J]. IEEE transactions on wireless communications, 2023, 22(9): 6185-6198.
- [3] 杨孟璋, 农丽萍, 李然, 等. 基于一维卷积神经网络的雷达个体识别算法[J]. 计算机工程与设计, 2025, 46(5): 1281-1288.
YANG Mengzhang, NONG Liping, LI Ran, et al. Radar individual recognition algorithm based on one-dimensional convolutional neural networks[J]. Computer engineering and design, 2025, 46(5): 1281-1288.
- [4] 徐钰龙, 李君, 李正权, 等. 基于深度强化学习的多用户蜂窝网络能效优化[J]. 计算机工程与设计, 2025, 46(3): 734-740.
XU Yulong, LI Jun, LI Zhengquan, et al. Energy efficiency optimization of multi-user cellular networks based on deep reinforcement learning[J]. Computer engineering and design, 2025, 46(3): 734-740.
- [5] 于晶, 鲁凌云, 李翔. 车联网中基于 DDQN 的边云协作任务卸载机制[J]. 计算机工程, 2022, 48(12): 156-164.
YU Jing, LU Lingyun, LI Xiang. Edge-cloud collaborative task offloading mechanism based on DDQN in vehicular networks[J]. Computer engineering, 2022, 48(12): 156-164.
- [6] 刘先锋, 梁赛, 李强, 等. 基于深度强化学习的云边协同 DNN 推理[J]. 计算机工程, 2022, 48(11): 30-38.
LIU Xianfeng, LIANG Sai, LI Qiang, et al. Cloud-edge collaborative DNN inference based on deep reinforcement learning[J]. Computer engineering, 2022, 48(11): 30-38.
- [7] 咸琳涛, 刘晓兰, 王淦, 等. 异构环境感知的分布式神经网络训练模型[J]. 计算机工程与设计, 2024, 45(9): 2821-2827.
XIAN Lintao, LIU Xiaolan, WANG Gan, et al. Distributed neural network training model for heterogeneous environment perception[J]. Computer engineering and design, 2024, 45(9): 2821-2827.
- [8] 赵婵婵, 吕飞, 石宝, 等. 面向边缘智能的协同推理方法研究综述[J]. 计算机工程与应用, 2025, 61(3): 1-20.
ZHAO Chanchan, LÜ Fei, SHI Bao, et al. Review of collaborative inference methods for edge intelligence[J]. Computer engineering and applications, 2025, 61(3): 1-20.
- [9] SHAO Jiawei, ZHANG Jun. Communication-computation trade-off in resource-constrained edge inference[J]. IEEE communications magazine, 2020, 58(12): 20-26.
- [10] XIAO Yilin, XIAO Liang, WAN Kunpeng, et al. Reinforcement learning based energy-efficient collaborative inference for mobile edge computing[J]. IEEE transactions on communications, 2023, 71(2): 864-876.
- [11] 周燕, 许业文, 蒲磊, 等. 自动驾驶场景下的图像三维目标检测研究进展[J]. 计算机科学, 2024, 51(11): 133-147.
ZHOU Yan, XU Yewen, PU Lei, et al. Research progress of image 3D object detection in autonomous driving scenario[J]. Computer science, 2024, 51(11): 133-147.
- [12] LI Xian, BI Suzhi. Optimal AI model splitting and resource allocation for device-edge co-inference in multi-user wireless sensing systems[J]. IEEE transactions on wireless communications, 2024, 23(9): 11094-11108.
- [13] ZENG Liekang, CHEN Xu, ZHOU Zhi, et al. CoEdge: cooperative DNN inference with adaptive workload partitioning over heterogeneous edge devices[J]. IEEE/ACM transactions on networking, 2021, 29(2): 595-608.
- [14] 王奕峰, 周婷, 徐天衡. 一种 NOMA 系统中的反馈深度神经网络检测算法[J]. 电讯技术, 2023, 63(5): 611-617.
WANG Yifeng, ZHOU Ting, XU Tianheng. A feedback deep neural network detection algorithm in NOMA systems[J]. Telecommunication engineering, 2023, 63(5): 611-617.
- [15] BOUTIBA K, BAGAA M, KSENTINI A. Multi-agent deep reinforcement learning to enable dynamic TDD in a

- multi-cell environment[J]. *IEEE transactions on mobile computing*, 2024, 23(5): 6163–6177.
- [16] 刘锋, 张爽, 黄渝昂. 融合压缩采样与深度神经网络的直接序列扩频参数估计[J]. *电讯技术*, 2022, 62(9): 1248–1253.
- LIU Feng, ZHANG Shuang, HUANG Yuang. Estimation of DSSS signal parameter based on compressive sampling and deep neural networks[J]. *Telecommunication engineering*, 2022, 62(9): 1248–1253.
- [17] 鲁富荣, 原之安, 钱宇华. 融合模体图神经网络和自编码器的链路预测[J]. *计算机科学与探索*, 2023, 17(1): 209–216.
- LU Furong, YUAN Zhian, QIAN Yuhua. Link prediction combining motif graph neural network and auto-encoder[J]. *Journal of frontiers of computer science & technology*, 2023, 17(1): 209–216.
- [18] WANG Song, ZHANG Xinyu, UCHIYAMA H, et al. HiveMind: towards cellular native machine learning model splitting[J]. *IEEE journal on selected areas in communications*, 2022, 40(2): 626–640.
- [19] 孟祥福, 崔江燕, 邓敏超. 基于图卷积神经网络的最短路径距离估计方法[J]. *智能系统学报*, 2024, 19(6): 1518–1527.
- MENG Xiangfu, CUI Jiangyan, DENG Minchao. Road network shortest distance estimation method based on graph convolutional networks[J]. *CAAI transactions on intelligent systems*, 2024, 19(6): 1518–1527.
- [20] 沈明辉, 刘宇杰, 陈婧, 等. 基于改进 YOLOv8s 轻量化网络的 PCBA 缺陷检测算法[J]. *计算机工程*, 2026, 52(6): 326–338.
- SHEN Minghui, LIU Yujie, CHEN Jing, et al. Defect detection algorithm for PCBA based on improved YOLOv8s lightweight network[J]. *Computer engineering*, 2026, 52(6): 326–338.
- [21] 李沂杨, 陆声链, 王继杰, 等. 基于 Transformer 的 DETR 目标检测算法综述[J]. *计算机工程*, 2026, 52(4): 62–81.
- LI Yiyang, LU Shenglian, WANG Jijie, et al. Review of DETR object detection algorithm based on transformer[J]. *Computer engineering*, 2026, 52(4): 62–81.
- [22] 李子奇, 苏宇轩, 孙俊, 等. 基于深度学习的多聚焦图像融合方法前沿进展[J]. *计算机科学与探索*, 2024, 18(9): 2276–2292.
- LI Ziqi, SU Yuxuan, SUN Jun, et al. Critical review of multi-focus image fusion based on deep learning method[J]. *Journal of frontiers of computer science & technology*, 2024, 18(9): 2276–2292.
- [23] 吴一全, 蔡佳琦. 自动驾驶中深度学习的三维目标检测方法综述[J]. *智能系统学报*, 2026, 21(2): 297–320.
- WU Yiquan, CAI Jiaqi. Deep learning-based 3D object detection for autonomous driving: a comprehensive review[J]. *CAAI transactions on intelligent systems*, 2026, 21(2): 297–320.
- [24] 田春伟, 宋明键, 左旺孟, 等. 卷积神经网络在图像超分辨上的应用[J]. *智能系统学报*, 2025, 20(3): 719–749.
- TIAN Chunwei, SONG Mingjian, ZUO Wangmeng, et al. Application of convolutional neural networks in image super-resolution[J]. *CAAI transactions on intelligent systems*, 2025, 20(3): 719–749.
- [25] 张浩哲, 杨智博, 焦绪国, 等. 基于自优化神经网络的船舶运动模型辨识[J]. *智能系统学报*, 2025, 20(3): 571–583.
- ZHANG Haozhe, YANG Zhibo, JIAO Xuguo, et al. Identification of ship motion model based on self-optimizing neural network[J]. *CAAI transactions on intelligent systems*, 2025, 20(3): 571–583.
- [26] 袁英华, 金英然, 高赞. 基于双维度特征增强的 Transformer 跟踪器[J]. *计算机工程*, 2025, 51(12): 96–108.
- YUAN Yinghua, JIN Yingran, GAO Yun. Transformer tracker based on dual-dimensional feature enhancement[J]. *Computer engineering*, 2025, 51(12): 96–108.
- [27] 毕阳阳, 郑远帆, 史彩娟, 等. 基于深度学习的图像全景分割综述[J]. *计算机科学与探索*, 2023, 17(11): 2605–2619.
- BI Yangyang, ZHENG Yuanfan, SHI Caijuan, et al. Survey on image panoptic segmentation based on deep learning[J]. *Journal of frontiers of computer science and technology*, 2023, 17(11): 2605–2619.
- [28] XU Xiaolong, TANG Sizhe, QI Lianying, et al. CNN partitioning and offloading for vehicular edge networks in Web3[J]. *IEEE communications magazine*, 2023, 61(8): 36–42.
- [29] XIAO Yang, SONG Yuqian, LIU Jun. Collaborative multi-agent deep reinforcement learning for energy-efficient resource allocation in heterogeneous mobile edge computing networks[J]. *IEEE transactions on wireless communications*, 2024, 23(6): 6653–6668.
- [30] 车飞虎, 张大伟, 邵朋朋, 等. 基于四元数门控图神经网络的脚本事件预测[J]. *智能系统学报*, 2023, 18(1): 138–143.
- CHE Feihu, ZHANG Dawei, SHAO Pengpeng, et al. Script event prediction based on a quaternion-gated graph neural network[J]. *CAAI transactions on intelligent systems*, 2023, 18(1): 138–143.

作者简介:



邓贻香, 高级实验师, 主要研究方向为模式识别与人工智能教学与研究。E-mail: 342717900@qq.com。



张玉林, 教授, 主要研究方向为机构运动及动力学。E-mail: 253199722@qq.com。



柏龙, 教授, 主要研究方向为医疗机器人及其自动化、仿生关节与智能部件、轻质多孔智能点阵结构。E-mail: bailong@cqu.edu.cn。