小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems
DOI:10.20009/j.cnki.21-1106/TP.2025-0091
2026 年 4 月 第 4 期
Vol.47 No.4 2026

边缘环境下基于时空特征融合 Transformer 的交通流量预测

施晓薇[1,2,3],李刘琛[1,2,3],于正欣[4],苗 旺[5],陈哲毅[1,2,3]

[1](福州大学 计算机与大数据学院,福州 350116)
[2](大数据智能教育部工程研究中心,福州 350002)
[3](福建省网络计算与智能信息处理重点实验室(福州大学),福州 350116)
[4](兰卡斯特大学 计算与通信学院,英国 兰卡斯特 LA1 4YW)
[5](埃克塞特大学 计算机科学系,英国 埃克塞特 EX4 4QF)
E-mail:z.chen@fzu.edu.cn

摘 要: 作为一种新兴的计算范式,移动边缘计算(Mobile Edge Computing,MEC)为交通流量预测提供了新的解决思路以更好支持智能交通系统.面对实际 MEC 环境中动态变化的交通流量,现有方法无法有效捕获交通流量中复杂时空依赖关系以实现精准预测.此外,由于边缘服务器资源有限,往往无法及时处理海量交通流量数据,难以满足智能交通系统对实时性的高要求.为解决这些重要挑战,本文提出了一种新颖的边缘环境下基于时空特征融合 Transformer 的交通流量预测(Traffic Flow Prediction based on Transformer with spatio-temporal feature fusion,TFPformer)方法.首先,对原始交通流量数据进行特征嵌入和编码.接着,设计多头卷积低秩分解注意力机制以捕获长期时间依赖关系和获取局部上下文信息.随后,设计注意力图卷积以捕获空间依赖关系.最后,通过门控单元对时空特征进行自适应融合,进而利用前馈神经网络和线性层实现对未来交通流量的精准预测.基于真实的交通流量数据集,通过大量实验全面评估与验证了所提出 TFPformer 方法的有效性.相比于基准方法,TFPformer 方法在不同数据集上均展现出了更加优越的预测精度和效率.

关 键 词: 移动边缘计算;交通流量预测;时空依赖;Transformer;注意力机制
中图分类号: TP393 **文献标识码:** A **文 章 编 号:** 1000-1220(2026)04-0836-08

Traffic Flow Prediction Based on Transformer with Spatio-temporal Feature Fusion in Edge Environment

SHI Xiaowei[1,2,3],LI Liuchen[1,2,3],YU Zhengxin[4],MIAO Wang[5],CHEN Zheyi[1,2,3]

[1](College of Computer and Data Science,Fuzhou University,Fuzhou 350116,China)
[2](Engineering Research Center of Big Data Intelligence,Ministry of Education,Fuzhou 350002,China)
[3](Fujian Key Laboratory of Network Computing and Intelligent Information Processing (Fuzhou University),Fuzhou 350116,China)
[4](School of Computing and Communications,Lancaster University,Lancaster LA1 4YW,UK)
[5](Department of Computer Science,University of Exeter,Exeter,EX4 4QF,UK)

Abstract: As an emerging computing paradigm,Mobile Edge Computing (MEC) offers a new solution for traffic flow prediction to better support Intelligent Transportation Systems (ITS). When facing the dynamic traffic flow in real-world MEC environments,the existing methods cannot effectively capture the complex spatio-temporal dependencies in traffic flows to achieve accurate prediction. Moreover,due to the limited resources on edge servers,massive traffic flow data often cannot be processed in time,which cannot meet the high requirements of ITS for real-time performance. To address these important challenges,this paper proposes a novel Traffic Flow Prediction based on Transformer with spatio-temporal feature fusion (TFPformer) method in edge environments. First,the feature embedding and encoding for raw traffic flow data are carried out. Next,a multi-head low-rank decomposition attention mechanism is designed to capture long-term temporal dependencies and obtain local context information. Then,the attention graph convolution is developed to capture spatial dependencies. Finally,the spatio-temporal features are adaptively fused through the gated unit,and the feedforward neural networks and linear layer are used to accurately predict the future traffic flow. Using the real-world datasets of traffic flow,the effectiveness of the proposed TFPformer method is comprehensively evaluated and verified through a large number of experiments Compared with the benchmark methods,the TFPformer method achieves superior prediction accuracy and efficiency on different datasets.

Keywords: mobile edge computing;traffic flow prediction;spatio-temporal dependence;Transformer;attention mechanism

收稿日期:2025-03-05 收修改稿日期:2025-04-14 基金项目:国家自然科学基金项目(62202103)资助;中央引导地方科技发展项目(2022L3004)资助;福建省科技经济融合服务平台项目(2023XRH001)资助;福厦泉国家自主创新示范区协同创新平台项目(2022FX5)资助.
作者简介:**施晓薇**,女,2002 年生,硕士研究生,CCF 学生会员,研究方向为移动边缘计算、流量预测、图联邦学习;**李刘琛**,女,2003 年生,硕士研究生,研究方向为移动边缘计算、异常流量检测、联邦学习;**于正欣**,女,1993 年生,博士,高级博士后研究员,研究方向为边缘计算、联邦深度学习;**苗 旺**,男,1986 年生,博士,讲师,研究方向为软件定义网络、网络功能虚拟化、移动边缘计算;**陈哲毅**(通信作者),男,1991 年生,博士,教授,博士生导师,CCF 专业会员,研究方向为云边协同计算、资源智能优化.

0 引 言

物联网(Internet-of-Things,IoT)的普及与发展推动了智慧城市的建设进程.作为智慧城市的重要组成部分,智能交通系统(Intelligent Traffic System,ITS)旨在利用先进的智能控制与数据通信等技术为城市交通提供智能化管理[1].在 ITS 中,交通流量预测是不可或缺的环节,在规划出行路线、保障交通安全、缓解交通拥挤等方面起到了至关重要的作用[2].作为 IoT 时代一种新兴的计算范式,移动边缘计算(Mobile Edge Computing,MEC)通过在终端设备附近部署边缘服务器以降低数据传输延迟和带宽消耗,为 ITS 中的交通流量预测问题提供了一种新的解决思路.通过将交通流量数据上传至边缘服务器,进而调用预测模型对历史和实时的交通流量数据进行分析与处理,可实现对未来一段时间段内交通流量的预测.在实际 MEC 环境中,交通流量时刻发生着变化,呈现出高度的动态性与复杂的时空依赖关系,导致交通流量难以被准确高效地预测.因此,如何准确高效地预测交通流量是一个亟待解决的挑战性难题.

交通流量数据通常来源于道路传感器、摄像头和超声波检测设备等多种途径[3],其包含路段流量、车辆速度和道路拥塞率等特征信息.随着 ITS 的发展,交通流量数据量呈现爆炸式增长且形式多样,其具有动态而复杂的时空依赖性与交互性.为了实现准确高效的交通流量预测,需要从原始交通流量数据中提取有效时空特征信息,并设计合适的方法进行预测.现有的预测方法包括实验模拟、基于统计学的方法、基于经典机器学习的方法以及基于深度学习的方法.具体而言,实验模拟对于领域知识储备的要求较高且灵活性较低,在面对情况复杂且影响因素较多的真实交通环境时,难以构建准确的交通流量预测模型.相比之下,基于数据本身的统计学方法对于领域知识的依赖度较低,但对数据质量的要求较高,然而真实交通流量数据往往无法满足这一要求.近年来,机器学习已被应用于交通流量预测,但在处理海量且复杂的交通流量数据时,其无法有效提取交通流量的时空特征和长距离依赖关系,严重影响了预测精度和效率.随着深度学习技术的快速发展,其在学习复杂非线性数据模式方面展现出了强大能力,有望应对复杂的交通流量预测问题,但也面临着以下挑战:

1)时空依赖难以捕获.交通流量数据包含时间和空间维度上复杂的依赖关系,且二者的变化具有高度时空关联性.此外,由于时空数据存在交互,其数据模式具有复杂的非线性特征.因此,如何从交通流量数据中有效捕获时空特征及其依赖性是提高交通流量预测精度的关键挑战.

2)预测实时性难以保证.在实际 ITS 中,交通流量往往呈现高度动态变化,而交通流量预测十分注重实时性和时效性.因此,要求预测模型能够快速处理海量数据并及时给出预测结果,以响应交通流量的实时变化.经典深度学习方法在进行数据处理和流量预测时存在较大延迟,难以满足 ITS 中交通流量预测对实时性的高要求.

3)预测模型资源开销大.随着边缘智能设备和传感器的广泛应用,边缘服务器需要处理海量的交通流量数据.现有基于深度学习的交通流量预测方法通常具有较高的计算复杂度,这对资源有限的边缘服务器造成了巨大的计算压力.因此,需要设计一种高效的交通流量预测方法,在保证预测精度的同时有效降低资源开销.

为了解决上述重要挑战,本文提出了一种新颖的边缘环境下基于时空特征融合 Transformer 的交通流量预测(Traffic Flow Prediction based on Transformer with spatio-temporal feature fusion,TFPformer)方法.本文的主要贡献如下:

1)设计了一种面向 MEC 环境的交通流量预测模型,以应对真实 ITS 中动态变化的交通流量.首先,获取原始交通流量数据并对其进行预处理.随后,进行数据特征嵌入和编码,并提取其时空特征,进而实现对未来交通流量的精准高效预测,以满足真实 ITS 需求.

2)提出了一种精准高效的交通流量预测方法(TFPformer),采用并行结构以捕获交通流量的时空依赖.首先,设计多头卷积低秩分解注意力机制与注意力图卷积分别捕获交通流量的时间和空间特征.特别地,引入因果卷积以获取局部上下文信息,并对注意力矩阵进行低秩分解,从而有效减轻内存负担并大幅提升模型训练效率.随后,采用门控单元对时空特征进行自适应融合.最后,利用前馈神经网络和线性层输出预测结果.预测过程充分提取和保留了原始时空特征信息,避免了潜在的特征丢失.

3)使用 4 个真实的交通流量数据集以评估所提出 TFPformer 方法的优越性.实验结果表明,与基准方法相比,TFPformer 方法在平均绝对误差(MAE)、均方根误差(RMSE)和平均绝对百分比误差(MAPE)等 3 项指标上均表现更优,在不同数据集上精度分别平均提升约 30%、24% 和 28%.此外,消融实验验证了 TFPformer 方法中各组件的有效性,在不同数据集上精度分别平均提升约 17%、16% 和 19%,平均训练时间减少约 1.5%.

1 相关工作

交通流量预测方法大致可分为模型驱动和数据驱动两类.模型驱动的方法依赖于大量领域知识,灵活性较低,因此主流研究方向已从模型驱动转向数据驱动.数据驱动的预测方法可进一步分为经典的方法和基于深度学习的方法[4].本文将从这两个角度回顾并分析交通流量预测的相关工作.

1.1 经典的交通流量预测方法

经典的交通流量预测方法主要包括基于统计学的方法和基于机器学习的方法.典型的统计学方法包括历史平均模型(Historical Average,HA)、自向量回归模型(Vector Auto-Regressive,VAR)、支持向量回归模型(Support Vector Regression,SVR)、自回归差分移动平均模型(Auto-Regressive Integrated Moving Average,ARIMA)和卡尔曼滤波[5]等.这类方法利用历史交通流量数据构建统计模型以预测未来的交通流量,通常具有简单高效的模型,但对数据质量要求较高(如,需要满足某些特定分布或假设线性条件等).然而,现实世界中的交通流量数据通常无法满足这些要求,且这类方法无法有效处理高维度的海量数据.机器学习的出现弥补了统计学方法在处理复杂高维数据方面的不足.例如,Lin 等人[6]提出了一种使用支持向量回归和 K 最近邻(K-Nearest Neighbor,KNN)进行短期交通流量预测的方法.Ramchandra 等人[7]使

用随机森林(Random Forest,RF)对交通流量进行预测.

1.2 基于深度学习的交通流量预测方法

深度学习可有效捕获数据的动态特征,适用于在海量数据中学习复杂的非线性模式,已被应用在交通流量预测问题.经典的循环神经网络(Recurrent Neural Network,RNN)[8]及其变体门(如,门控循环单元(Gate Recurrent Unit,GRU)[9]和长短期记忆(Long Short-Term Memory,LSTM)[10])也被广泛应用于交通流量预测这一时序预测问题.此外,卷积神经网络(Convolutional Neural Networks,CNN)在时序预测过程中可通过卷积操作提取空间特征.但是,CNN 适用于规则的网格状数据,而真实的交通拓扑结构通常不规则且复杂.不同于CNN,图卷积神经网络(Graph Convolustional Neural Network,GCN)可用于处理非欧式空间结构数据.例如,Anton 等人[11]利用 GCN 对交通流进行短时预测.然而,上述方法忽略了时间和空间特征之间的交互性和关联性,导致许多重要特征信息丢失.为此,一些学者提出时空关系建模方法.例如,Bao 等人[12]提出了一种基于时空复杂图卷积的网络,利用 GCN 和残差单元构建空间特征提取模块,并通过 LSTM 和三维卷积提取时间特征,从而学习时空特征.Xue 等人[13]提出一种基于知识图谱和时空扩散图卷积网络的港口交通流量预测算法,利用知识集成单元将语义信息与港口交通流量数据融合至时空扩散图卷积网络,从而捕获数据的时空依赖.然而,这些方法对时间依赖性的捕获是建立在预定义的静态图之上,无法应对复杂多变的真实交通情况.为此,Jiang 等人[14]提出一种新的基于输入交通信号分解的动态时空融合的交通预测模型,考虑了空间依赖的动态性质.Li 等人[15]提出了一种自适应时空融合图卷积网络来调整时空图结构,以发现不同时空范围对各节点的影响.Shi 等人[16]设计了一种基于注意力的周期性时间神经网络以捕获空间和长短期依赖性.此外,一些研究使用 Transformer[17]模型进行交通流量预测,以增强模型的时空建模能力.例如,Xu 等人[18]提出了一种时空 Transformer 网络,先后考虑了动态定向空间依赖性和长期时间依赖性,但在这过程中存在时空交互信息丢失情况,从而影响预测模型性能.

总体上,现有交通流量预测研究大多使用 RNN、GCN 或其混合模型进行时空建模,其在处理时空信息方面各有优势,但在捕获长期依赖性、模型灵活性等方面尚有欠缺.此外,目前大多研究采用串行结构以捕获时空依赖,但忽略了真实交通流量中存在时间和空间信息的交互性.相比之下,Tansformer 模型展现出了强大的长期依赖捕捉能力,同时其不依赖于递归结构,具备并行化优势,可用于增强预测模型的时空建模能力.然而,经典的 Transformer 模型依赖于注意力机制,存在较高的计算复杂度.如何提升预测效率,降低模型训练开销是一个亟待解决的关键性问题.因此,虽然 Transformer 模型在交通流量预测问题中展现出了巨大潜力,但其在学习能力和训练效率方面仍存在较大提升空间.

2 系统模型与问题定义

为了更好满足 ITS 中交通流量预测的实际应用需求,本文提出了一种面向 MEC 环境的交通流量预测模型,如图 1 所示.首先,边缘服务器从道路上的传感器收集原始交通流量数据并对其进行预处理.随后,调用所提出的 TFPformer 方法对时空特征进行提取与融合,进而实现精准高效的交通流量预测.最后,利用预测结果以支持 ITS 进行智能决策.

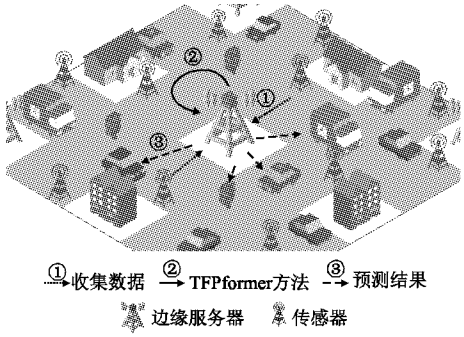

图 1 所提出的面向边缘环境的交通流量预测模型

Fig. 1 Proposed model of traffic flow prediction in edge environments

针对交通流量预测这一时空预测问题,本文首先构建了一个无向图 $G=\{V,E,A\}$ 以表示预测区域的道路拓扑结构.顶点集 $V=\{v_1,v_2,\cdots,v_N\}$ 表示传感器集合,其中 N 为传感器的数量.边集 $E=\{e_1,e_2,\cdots,e_C\}$ 表示不同传感器之间的距离,其中 C 为边的数量.邻接矩阵 $A\in R^{N\times N}$ 表示不同传感器之间的连接关系,其被定义为:

$$A_{ij}=\begin{cases}1, & v_i \text{ 与 } v_j \text{ 连接}\\ 0, & v_i \text{ 与 } v_j \text{ 无连接}\end{cases} \tag{1}$$

接着,训练样本的特征矩阵可表示为 $X\in R^{N\times F\times T}$. $X=[X_0,\cdots,X_t,\cdots,X_T]$,其中 $X_t\in R^{N\times F}$,F 表示交通流量特征的数量,T 表示历史交通流量序列的长度.基于上述定义,交通流量预测问题可被形式化表示为:

$$[X_{T+1},X_{T+2},\cdots,X_{T+p}]=f(G,[X_0,X_1,\cdots,X_T]) \tag{2}$$

其中,f 表示用于预测的映射函数,p 表示未来交通流量预测长度.

交通流量预测的目标是最小化预测值与真实值之间的误差.具体而言,交通流量的真实值和预测值分别记为 Y_t 和 $\hat{Y}_t$,loss 为损失函数.因此,交通流量预测问题的优化目标可表示为:

$$\min loss(Y_t,\hat{Y}_t) \tag{3}$$

为了全面评估预测性能,本文采用了平均绝对误差(Mean Absolute Error,MAE)、均方根误差(Root Mean Square Error,RMSE)和平均绝对百分比误差(Mean Absolute Percentage Error,MAPE)等指标.3 项指标的取值范围均为[0,1],越接近 0 表示预测性能越好.具体定义为:

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|\hat{Y}_i-Y_i| \tag{4}$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}|\hat{Y}_i-Y_i|} \tag{5}$$

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{Y}_i-Y_i}{Y_i}\right|\times 100\% \tag{6}$$

3 所提出的 TFPformer 方法

3.1 概览

如图 2 所示,所提出的 TFPformer 方法主要由 3 个部分

组成,包括输入层、编码层和输出层,其中编码层主要由多头卷积低秩分解注意力机制、注意力图卷积和门控单元构成. 首先,输入层对历史交通流量进行特征嵌入和位置编码处理. 接着,将处理后的交通流量特征同时输入多头卷积低秩分解注意力机制和注意力图卷积以分别捕获时间和空间依赖,进而通过门控单元对其进行自适应融合以获得交通流量的时空特征. 随后,将融合结果输入前馈神经网络以进一步提取有效时空特征,并引入两个残差连接以避免梯度消失问题. 最后,通过输出层以输出未来交通流量预测结果.

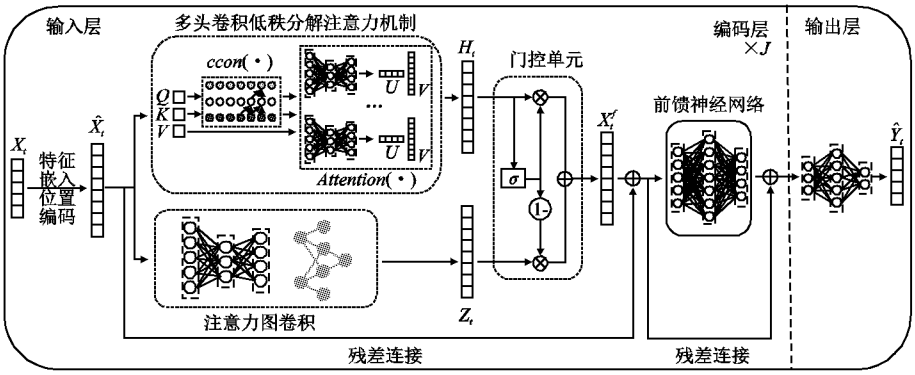

图 2　所提出的 TFPformer 方法

Fig. 2　Overview of the proposed TFPformer method

3.2　多头卷积低秩分解注意力机制

本文设计了一种多头卷积低秩分解注意力机制以用于捕获输入特征的时间依赖. 经典的 Transformer 模型采用的是一种全局注意力机制,其难以对局部上下文信息进行有效建模. 针对该问题,本文将因果卷积与多头注意力机制相结合,在增强局部建模能力的同时保持了数据在时间上的前后依赖关系,进而提高了模型对时间依赖的学习能力. 在经典的多头注意力机制中,通常采用点积的方式来计算每个输入的位置与其他位置的相似度,导致了大量内存消耗. 为了缓解该问题,本文针对注意力矩阵引入了低秩分解以降低其计算复杂度,进而减轻内存负担并提升模型训练效率.

首先,为获取局部上下文信息,在计算注意力矩阵之前,需对输入的查询 Q 和键 K 进行因果卷积,其被定义为:

$$ccon(X_t)=\sum_{k=0}^{m}X_{t-k}\times\Phi_k \tag{7}$$

其中,Φ_k 表示卷积核,m 表示卷积核的大小,$W^V\in R^{d_k\times d_k}$ 表示可学习参数.

接着,针对输入的查询 Q、键 K 和值 V,计算注意力矩阵并进行低秩分解. 该过程被定义为:

$$Attention(Q,K,V)=\left(soft\ \max\left[\frac{QK^T}{\sqrt{d_k}}\right]V\right)UL \tag{8}$$

其中,d_k 表示缩放因子,$U\in R^{d_k\times d_k}$ 和 $L\in R^{d_k\times d_k}$ 表示低秩矩阵.

对于单头自注意力机制,其平均化抑制了来自不同位置的表示子空间信息. 相比而言,多头自注意力机制能够有效地解决该问题,其被定义为:

$$H_t=concact(head_1,head_2,\cdots,head_h)W^O \tag{9}$$

$$head_h=Attention(ccon^Q(Q),ccon^K(K),VW^V) \tag{10}$$

其中,$concact(\cdot)$ 表示串联拼接操作,h 表示注意力头的数量,W^O、W^V 表示可学习参数,得到最终输出 H_t.

3.3　注意力图卷积

本文设计了一种注意力图卷积以用于捕获输入特征的空间依赖关系. 经典的图卷积通常利用预定义的图邻接矩阵来聚合各邻居节点,存在一定局限性. 因为其假设各邻居节点具有相同的贡献,其灵活性较差且无法有效表示数据中复杂的空间关系. 针对该问题,本文基于经典的图卷积,引入了缩放点积来计算各节点之间的相似性. 该设计允许模型自适应调整各邻居节点的权重,从而有效学习节点特征以捕获节点复杂的空间依赖关系.

首先,针对输入特征,计算其注意力分数. 随后,进行图卷积操作,其被定义为:

$$S=dropout\left(soft\ \max\left(\frac{XX^T}{\sqrt{d_k}}\right)\right) \tag{11}$$

$$Z_t=relu(\Theta(\hat{A}SX_t)) \tag{12}$$

其中,$\hat{A}=D^{-1}A$,D 为 A 的度矩阵,Θ 为线性投影函数.

3.4　门控单元

在经典的 GRU 中,门控机制被用于控制信息传播,可用于融合多个特征信息,决定每条路径对最终输出的贡献权重. 鉴于此,本文引入门控单元对所捕获的时空依赖进行自适应融合,以有效平衡时间与空间信息之间的交互. 通过利用门控单元,能够自适应地决定每条路径对最终输出的贡献,从而提升模型性能. 该过程被定义为:

$$X_t^f=(gate\cdot H_t+(1-gate)\cdot Z_t)W^f+b^f \tag{13}$$

$$gate=\frac{1}{1+e^{-H_t}} \tag{14}$$

其中,$W^f\in R^{d_k\times d_k}$ 和 $b^f\in R^{d_k\times 1}$ 表示可学习参数.

算法 1. 所提出的 TFPformer 方法

输入:历史交通流量 $X=[X_0,X_1,\cdots\cdots,X_t]$、图邻接矩阵 A、训练轮次 M、训练批次 B

输出:预测时间步的交通流量 $X=[X_{t+1},X_{t+2},\cdots\cdots,X_{t+p}]$

1. 初始化参数,选择 MAE 损失函数和 Adam 优化器;
2. for *epoch* = 1,2,⋯,M do
3.　　for i = 1,2,⋯,B do
4.　　进行特征嵌入和编码:$\hat{X}_t\leftarrow X_t$;
5.　　通过多头卷积低秩分解注意力机制提取时间特征:
　　$H_t\leftarrow MultiHead(\hat{X}_t)$;

6. 通过注意力图卷积提取空间特征：$Z_t \leftarrow \varphi(\hat{X}_t)$；
7. 利用门控单元融合时空特征：$X_t^f \leftarrow fusion(H_t, Z_t)$；
8. 前馈神经网络编码：$m_t \leftarrow FNN(X_t^f)$；
9. 预测未来交通流量：$\hat{Y} \leftarrow linear(m_t)$；
10. 使用 MAE 计算损失：$loss \leftarrow L_1(Y_t, \hat{Y}_t)$；
11. 利用 Adam 优化器更新模型参数；
12. end for
13. end for

3.5 模型训练

本文所提出 TFPformer 方法的关键步骤如算法 1 所示. TFPformer 方法的输入包括历史交通流量 X 和图邻接矩阵 A. 首先，初始化 TFPformer 方法中的参数、损失函数和优化器（第 1 行）. 在每个训练轮次中，针对每个批次的样本进行特征嵌入和编码，得到嵌入向量 $\hat{X}_t$（第 4 行）. 接着，将嵌入向量输入多头卷积低秩分解注意力机制和注意力图卷积，以提取交通流量的时间特征 H_t（第 5 行）和空间特征 Z_t（第 6 行）. 接着，采用门控单元对时间和空间特征进行融合以有效捕获交通流量中的时空依赖关系，并得到融合后的特征 X_t^f（第 7 行）. 随后，将其输入至前馈神经网络以学习更高层次的非线性特征表示（第 8 行），得到 m_t. 进一步地，通过线性层输出对未来一定时间内交通流量的预测结果 $\hat{Y}_t$（第 9 行）. 最后，使用 MAE 损失函数来度量预测值与真实值之间的差距（第 10 行），并利用 Adam 优化器以更新和优化模型参数（第 11 行）.

4 性能评估

4.1 数据集与实验设置

本文实验在一台高性能工作站上开展，其配备了 Intel（R）Core（TM）i5-12600KF CPU 和 NVIDIA GeForce RTX 4060 GPU，CUDA 驱动版本为 11.6. 基于深度学习框架 PyTorch 2.3.1 + cu121，实现了所提出的 TFPformer 方法. 本文在 4 个真实交通流量数据集上进行了大量实验以验证 TFPformer 方法的有效性和优越性. PEMS 系列数据集记录了加利福尼亚州高速公路交通流量[19]，具体信息如表 1 所示. 原始数据由道路传感器每隔 5 分钟采集一次得到，其包含流量、占有率和速度等特征信息. 在实验中，本文根据流量特征和道路传感器地理位置信息构建图数据.

表 1 数据集描述

Table 1 Description of datasets

数据集	传感器数量	历史序列长度
PEMS03	358	26208
PEMS04	307	16992
PEMS07	883	28224
PEMS08	170	17856

本文首先对原始交通流量数据进行预处理，包括数据标准化、平滑处理等，以避免部分异常值或错误值对实验造成影响. 随后，将数据集划分为训练集、验证集和测试集，其比例为 6∶2∶2. 此外，设置训练轮次 M 为 100，训练批次 B 为 8（PEMS03、PEMS04、PEMS08）和 4（PEMS07），学习率为 0.001，采用 4 个编码模块，8 个注意力头，数据嵌入维度为 64，并利用 12 个历史时间步（1 个小时）预测未来 12 个时间步的交通流量[20].

为了评估所提出 TFPformer 方法的优越性，本文与以下 8 种基准方法进行了大量对比实验：

1）HA[21]是一种统计学方法，通过计算历史平均值来预测未来值，常用于具有周期性或季节性规律的预测.

2）LSTM[22]是 RNN 的一种经典变体，专门用于处理时间序列预测问题.

3）STGCN[23]不使用传统的 CNN 和递归单元，而是将问题描述为图，使用完全卷积结构来建立预测模型.

4）MTGNN[24]是一种专门针对多变量时间序列数据而设计的通用 GNN 框架.

5）AGCRN[25]引入节点自适应参数学习、自适应图生成和递归网络，以捕获序列中的时空相关性.

6）DCRNN[26]使用双向随机游走和具有预定采样的编码器-解码器架构，进而分别捕获空间和时间依赖.

7）Graph-WaveNet[27]引入了一种自适应依赖矩阵，进而结合节点嵌入以捕获数据中隐藏的空间依赖性.

8）ASTGNN[20]通过自注意机制和动态图卷积以分别捕获数据中的时间和空间依赖.

4.2 实验结果与分析

4.2.1 对比实验

为了验证所提出 TFPformer 方法的优越性，本文对比了不同方法在 4 个数据集上的表现，实验结果如表 2 ~ 表 5 所示. 具体来说，DCRNN 方法在各数据集上的性能几乎均是最差的，在 PEMS03 数据集上的 MAE、RMSE 和 MAPE 分别为 52.49、38.82 和 74.91%. 这是因为 DCRNN 方法的模型结构

表 2 不同方法在 PEMS03 数据集上的对比

Table 2 Comparison of different methods on PEMS03

方法	MAE	RMSE	MAPE
HA	31.42	49.89	37.17%
LSTM	21.55	37.47	24.68%
STGCN	18.02	27.36	19.17%
MTGNN	16.05	25.78	15.51%
AGCRN	15.63	27.63	14.96%
DCRNN	38.82	52.49	74.91%
Graph-WaveNet	14.70	25.16	16.82%
ASTGNN	14.68	25.10	16.03%
TFPformer	**10.88**	**20.87**	**11.80%**
Improvement	**25.87%**	**16.86%**	**21.18%**

复杂，其通过扩散卷积的方式来建模时空依赖. 然而现实场景中，模型对新数据的敏感度较高且适应能力较弱，导致其无法有效应对训练与测试数据之间的差异. 其余 5 种方法同时考虑了时空特征，预测效果显著优于经典的 HA 和 LSTM 方法. 相比 HA 和 LSTM 方法，STGCN 方法取得了明显的性能提升，但其采用图卷积和 1-D 卷积，难以充分捕获复杂的时空依赖关系，尤其是在面对长时间预测时存在性能瓶颈. AGCRN 和 Graph-WaveNet 方法均采用了自适应机制来提取数据中的时空特征. 其中，AGCRN 方法在各数据集上均表现良好，在 PEMS03 数据集上取得了最优的 MAPE，Graph-WaveNet 方法在 PEMS04、PEMS07 和 PEMS08 数据集上均取得了两项最优指标，这是因为自适应机制有助于增强模型的表达能力. 但

是,这两种方法对时空特征采用了先后提取的方式,一定程度上弱化或丢失了时空之间的关联,从而影响时空特征质量和降低预测性能. ASTGNN 方法在 PEMS03 数据集上的 MAE 和 RMSE 分别为 14.68 和 25.10,在其余 3 个数据集上均取得最优的 MAE 或 MAPE,其同样分阶段对数据进行时空特征提取,同时引入了自注意力机制,因此计算复杂度较高,增加了模型训练的计算开销.

表 3　不同方法在 PEMS04 数据集上的对比

Table 3　Comparison of different methods on PEMS03

方法	MAE	RMSE	MAPE
HA	40.51	59.05	33.81%
LSTM	28.25	46.53	20.90%
STGCN	24.87	37.71	18.74%
MTGNN	20.37	32.05	14.07%
AGCRN	19.72	32.26	13.23%
DCRNN	38.67	53.04	36.30%
Graph-WaveNet	19.01	30.22	13.66%
ASTGNN	19.27	31.44	12.50%
TFPformer	**13.47**	**22.82**	**10.19%**
Improvement	**29.18%**	**24.49%**	**18.49%**

表 4　不同方法在 PEMS07 数据集上的对比

Table 4　Comparison of different methods on PEMS07

方法	MAE	RMSE	MAPE
HA	45.62	66.55	23.57%
LSTM	31.04	50.41	16.54%
STGCN	27.22	41.04	13.23%
MTGNN	24.11	37.37	10.27%
AGCRN	21.27	35.17	9.00%
DCRNN	39.51	54.92	34.35%
Graph-WaveNet	20.79	33.58	8.80%
ASTGNN	20.67	33.83	9.20%
TFPformer	**14.24**	**24.82**	**5.85%**
Improvement	**31.13%**	**26.09%**	**34.10%**

表 5　不同方法在 PEMS08 数据集上的对比

Table 5　Comparison of different methods on PEMS08

方法	MAE	RMSE	MAPE
HA	31.88	46.49	23.08%
LSTM	24.00	41.31	19.20%
STGCN	21.14	31.98	17.34%
MTGNN	16.36	25.55	10.39%
AGCRN	16.35	25.87	10.58%
DCRNN	39.96	55.86	32.42%
Graph-WaveNet	14.79	23.54	10.01%
ASTGNN	15.44	24.94	9.76%
TFPformer	**9.64**	**17.02**	**6.14%**
Improvement	**34.83%**	**27.69%**	**37.14%**

相比于其他方法,所提出的 TFPformer 方法的 3 项性能在 4 个数据集上均取得了最优结果. 具体而言,TFPformer 方法的 MAE、RMSE 和 MAPE 在 4 个数据集上分别平均降低了 30.25%、23.78% 和 27.73%. 这得益于 TFPformer 方法采用了并行的多头卷积低秩分解注意力机制和注意力图卷积,可独立地捕获时间和空间依赖并保持了时空信息的交互性,避免了潜在的特征冲突,进而最大程度地保留了重要时空信息. 进一步地,通过门控单元自适应地融合时空信息,提高模型的鲁棒性和适应性.

4.2.2　消融实验

为验证所提出 TFPformer 方法中各组件的有效性,本文进行了消融实验. 相比于 TFPformer 方法,TFPformer-A 方法不包含注意力图卷积,TFPformer-B 方法不包含多头卷积低秩分解注意力机制,TFPformer-C 方法不包含注意力图卷积和多头卷积低秩分解注意力机制. 各方法在不同数据集上的性能表现如图 3 ~ 图 6 所示.

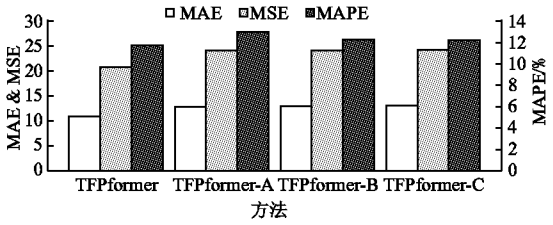

图 3　在 PEMS03 数据集上对 TFPformer 的消融实验

Fig. 3　Ablation study for TFPformer on PEMS03

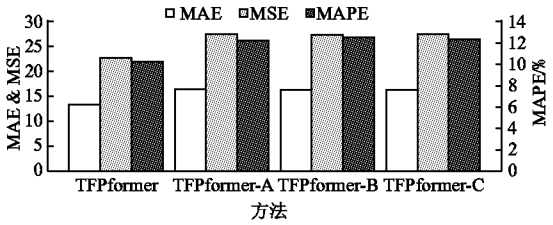

图 4　在 PEMS04 数据集上对 TFPformer 的消融实验

Fig. 4　Ablation study for TFPformer on PEMS04

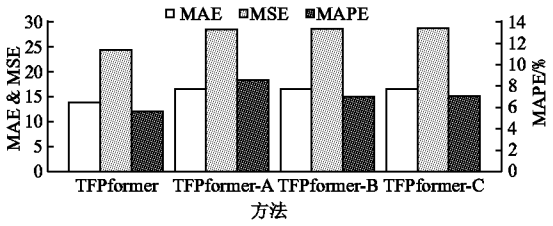

图 5　在 PEMS07 数据集上对 TFPformer 的消融实验

Fig. 5　Ablation study for TFPformer on PEMS07

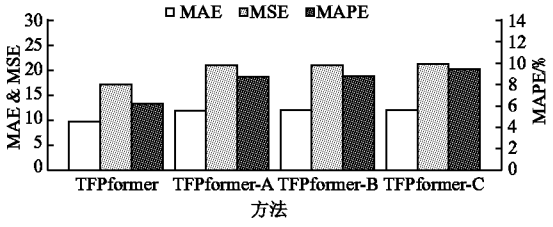

图 6　在 PEMS08 数据集上对 TFPformer 的消融实验

Fig. 6　Ablation study for TFPformer on PEMS08

在 PEMS08 数据集上,TFPformer-A 方法的 MAE、RMSE 和 MAPE 分别为 11.91、20.84 和 8.70%,TFPformer-B 方法的 3 项指标分别为 12.06、21.02 和 8.72%,TFPformer-C 方法的 3 项指标分别为 12.11、21.11 和 9.43%. 相较于 TFPform-

er-C 方法,TFPformer-A 和 TFPformer-B 方法在 3 项指标上分别提升了 1.65%、1.25%、7.73% 和 0.41%、0.41%、7.46%,性能小幅提升. TFPformer 有效结合了注意力图卷积和多头卷积低秩分解注意力机制,相较于 TFPformer-C 方法,其在 3 项评价指标上分别提升了 20.37%、19.35% 和 34.91%,性能明显提升. 这验证了 TFPformer 方法中两个组件的有效性,也体现了 TFPformer 方法中并行结构的有效性. 单纯依赖时间或空间信息难以有效捕捉时空依赖,因此对模型性能的提升效果有限. 通过并行学习时间和空间依赖,可有效提升模型对时空特征的建模能力,进而显著提升预测效果. 在 PEMS03 数据集上,相较于 TFPformer-C 方法,TFPformer 方法在 3 项指标上分别提升了 16.58%、14.49% 和 4.23%. 在 PEMS04 数据集上,TFPformer 方法在 3 项指标上分别提升了 17.07%、16.49% 和 17.89%. 在 PEMS07 数据集上,TFPformer 方法在 3 项指标上分别提升了 15.72%、14.46% 和 18.46%. 上述实验充分验证了所提出 TFPformer 方法中各组件的有效性.

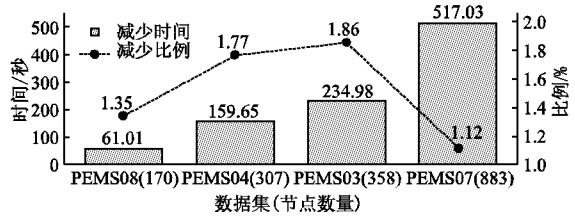

图 7 低秩分解对提升训练效率的有效性

Fig. 7 Effectiveness of the low-rank decomposition for enhancing training efficiency

进一步地,本文评估了所提出 TFPformer 方法中多头卷积低秩分解机制在提升模型训练效率方面的有效性. 具体而言,TFPformer-D 方法不包含多头卷积低秩分解机制中的低秩分解部分,进而对比其与 TFPformer 方法训练 100 轮次所用的时间. 如图 7 所示,相较于 TFPformer-D 方法,TFPformer 方法在 PSEM08、PSEM04、PSEM03 和 PSEM07 数据集上的训练时间分别减少了 61.01s、159.65s、234.98s 和 517.03s,减少比例分别为 1.35%、1.77%、1.86% 和 1.12%,与数据集中节点数量大致成正比. 从实验结果可以看出,TFPformer 方法中所设计的低秩分解有助于缩短模型训练时间和提升模型训练效率.

5 总 结

本文提出了一种新颖的面向边缘环境的交通流量预测方法(TFPformer). 该方法基于 Transformer 模型,采用并行结构以有效捕获交通流量中的时空依赖,进而提高交通流量预测精度和效率. 首先,在输入层对进行特征嵌入和编码;随后,利用编码层中的并行结构分别捕获时间和空间依赖并进行自适应融合;最后,通过输出层输出未来交通流量的预测结果. 基于 4 个真实交通数据集和对比 8 种基准方法,大量实验验证了所提出 TFPformer 方法的优越性. 实验结果表明,与其他方法相比,TFPformer 方法在不同性能指标上均取得了最优的预测精度. 此外,通过消融实验全面验证了 TFPformer 方法中各组件对于提升交通流量预测模型精度和效率方面的有效性.

References:

[1] Medina Salgado B, Sánchez DelaCruz E, Pozos Parra P, et al. Urban traffic flow prediction techniques: a review[J]. Sustainable Computing: Informatics and Systems, 2022, 35: 100739, doi: 10.1016/j.suscom.2022.100739.

[2] Lu B, Gan X Y, Jin H M, et al. Spatiotemporal adaptive gated graph convolution network for urban traffic flow forecasting[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM), 2020: 1025-1034.

[3] Miglani A, Kumar N. Deep learning models for traffic flow prediction in autonomous vehicles: a review, solutions, and challenges [J]. Vehicular Communications, 2019, 20: 100184, doi: 10.1016/j.vehcom.2019.100184.

[4] Yin X Y, Wu G Z, Wei J Z, et al. Deep learning on traffic prediction: methods, analysis, and future directions[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(6): 4927-4943.

[5] Wang X Y, Ma Y, Wang Y Q, et al. Traffic flow prediction via spatial temporal graph neural network[C]//Proceedings of the Web Conference, 2020: 1082-1092.

[6] Lin G C, Lin A J, Gu D L. Using support vector regression and K-nearest neighbors for short-term traffic flow prediction based on maximal information coefficient[J]. Information Sciences, 2022, 608: 517-531, doi: 10.1016/j.ins.2022.06.090.

[7] Ramchandra N R, Rajabhushanam C. Machine learning algorithms performance evaluation in traffic flow prediction[J]. Materials Today: Proceedings, 2022, 51(Part 1): 1046-1050.

[8] Fang W, Chen Y P, Xue Q Y. Survey on research of RNN-based spatio-temporal sequence prediction algorithms[J]. Journal on Big Data, 2021, 3(3): 97.

[9] Sun P, Boukerche A, Tao Y J. SSGRU: a novel hybrid stacked GRU-based traffic volume prediction approach in a road network [J]. Computer Communications, 2020, 160: 502-511, doi: 10.1016/j.comcom.2020.06.028.

[10] Bi J, Zhang X, Yuan H T, et al. A hybrid prediction method for realistic network traffic with temporal convolutional network and LSTM[J]. IEEE Transactions on Automation Science and Engineering, 2021, 19(3): 1869-1879.

[11] Agafonov A. Traffic flow prediction using graph convolution neural networks[C]//10th International Conference on Information Science and Technology (ICIST), 2020: 91-95.

[12] Bao Y X, Huang J S, Shen Q Q, et al. Spatial-temporal complex graph convolution network for traffic flow prediction[J]. Engineering Applications of Artificial Intelligence, 2023, 121(3): 106044.

[13] XUE G X, WANG H, ZHOU W F, et al. Port traffic flow prediction based on knowledge graph and spatio-temporal diffusion graph convolutional network[J]. Journal of Computer Applications, 2024, 44(9): 2952-2957.

[14] JIANG T, YANG L, LIU Y L, et al. Traffic flow prediction based on the dynamic spatial-temporal decomposition framework[J]. Science Technology and Engineering, 2025, 25(7): 3007-3017.

[15] Li S W,Ge L,Lin Y Q,et al. Adaptive spatial-temporal fusion graph convolutional networks for traffic flow forecasting[C]//International Joint Conference on Neural Networks(IJCNN),2022:1-8.

[16] Shi X M,Qi H,Shen Y M,et al. A spatial-temporal attention approach for traffic prediction[J]. IEEE Transactions on Intelligent Transportation Systems,2020,22(8):4909-4918.

[17] Jiang J,Han C,Zhao W X,et al. Pdformer:propagation delay-aware dynamic long-range transformer for traffic flow prediction[C]//Proceedings of the AAAI Conference on Artificial Intelligence,2023:4365-4373.

[18] Xu M X,Dai W R,Liu C M,et al. Spatial-temporal transformer networks for traffic flow forecasting[J]. arxiv preprint arxiv:2001.02908,2020.

[19] Song C,Lin Y F,Guo S N,et al. Spatial-temporal synchronous graph convolutional networks:a new framework for spatial-temporal network data forecasting[C]//Proceedings of the AAAI Conference on Artificial Intelligence,2020:914-921.

[20] Guo S N,Lin Y F,Wan H Y,et al. Learning dynamics and heterogeneity of spatial-temporal graph data for traffic forecasting[J]. IEEE Transactions on Knowledge and Data Engineering,2021,34(11):5415-5428.

[21] Shi H F,Pan C S,Yang L,et al. AGG:a novel intelligent network traffic prediction method based on joint attention and GCN-GRU[J]. Security and Communication Networks,2021,(1):7751484,doi:10.1155/2021/7751484.

[22] Li F X,Feng J,Yan H,et al. Dynamic graph convolutional recurrent network for traffic prediction:benchmark and solution[J]. ACM Transactions on Knowledge Discovery from Data,2023,17(1):1-21.

[23] Yu B,Yin H T,Zhu Z X. Spatio-temporal graph convolutional networks:a deep learning framework for traffic forecasting[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence(IJCAI),2018:3634-3640.

[24] Wu Z H,Pan S R,Long G D,et al. Connecting the dots:Multivariate time series forecasting with graph neural networks[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,2020:753-763.

[25] Bai L,Yao L N,Li C,et al. Adaptive graph convolutional recurrent network for traffic forecasting[J]. Advances in Neural Information Processing Systems,2020,33:17804-17815,doi:10.48550/arXiv.2007.02842.

[26] Li Y G,Yu R,Shahabi C,et al. Diffusion convolutional recurrent neural network:data-driven traffic forecasting[C]//International Conference on Learning Representations(ICLR),2018:1-16.

[27] Wu Z H,Pan S R,Long G D,et al. Graph waveNet for deep spatial-temporal graph modeling[C]//28th International Joint Conference on Artificial Intelligence(IJCAI),2019:1907-1913.

附中文参考文献:

[13] 薛桂香,王　辉,周卫峰,等. 基于知识图谱和时空扩散图卷积网络的港口交通流量预测[J]. 计算机应用,2024,44(9):2952-2957.

[14] 蒋　挺,杨　柳,刘亚林,等. 基于分解动态时空分解框架预测交通流量[J]. 科学技术与工程,2025,25(7):3007-3017.