文章编号：1006-3080(2025)06-0817-10 DOI：10.14135/j.cnki.1006-3080.20250225003

基于双尺度自适应令牌注意力的交通流量预测

郭津延[1]， 郑 红[1]， 杜佳宇[1]， 罗俞建[1]， 李鹏威[1]， 单蓉胜[2]
（1. 华东理工大学信息科学与工程学院, 上海 200237； 2. 上海交通大学网络空间安全学院, 上海 200240）

摘要：针对现有交通流量预测方法存在计算复杂度高、实时性差以及局部与全局特征整合不足等缺点，本文提出了一种基于双尺度自适应令牌注意力的交通流量预测模型。该模型结合双尺度自适应令牌注意力机制提取复杂时空特征：通过双尺度可学习池化得到的令牌分别捕获数据的长期和短期特征，并利用自适应令牌注意力机制整合全局依赖关系，提升预测准确性和效率。在两个公开数据集上进行实验验证，结果表明该方法在预测精度和计算效率上优于现有主流模型，适用于实时交通流量预测场景，为智能交通系统提供了一种高效、精准的解决方案。

关键词：智能交通；交通流预测；注意力机制；Transformer；深度学习

中图分类号：U495;TP18 **文献标志码：**A

交通流量预测是交通管理与城市智能交通系统的基础和核心，随着机动车数量的激增，传统的交通流量预测方法已难以应对日益复杂的交通状况，因此预测精度更准确、实时性更优秀以及计算资源要求更低的预测方法是目前研究的热点。交通流预测方法分为基于统计分析、基于传统机器学习和基于深度学习 3 类[1-3]。基于统计分析的交通流预测是该领域最初的研究方向，研究者主要采用经典统计方法[4-5]。后续研究者将基于传统机器学习的方法引入交通流预测领域来处理交通流数据的复杂非线性关系[6-7]。随着深度学习技术的发展，采用深度学习进行交通流预测能更有效地提取数据特征、挖掘数据之间的关联性，表现出更高的预测精准度[8-9]。近些年，如 Weng 等[10] 提出了一种模式匹配动态记忆网络 PM-DMNet 来提取交通流数据中的代表性模式以进行预测。Fang 等[11] 提出了基于离散小波变换和高效谱图注意力的 STWave 模型。Fan 等[12] 提出了一种基于 Seq2Seq 体系结构的时空图神经网络 PDG2Seq，通过学习到的周期特征和交通信号构建周期动态图结构，以此捕捉道路节点之间的动态关系。Weng 等[13] 提出的 DDGCRN 模型通过一种新的动态图生成方式在没有先验知识的情况下生成动态图来提取空间特征，并区分正常与异常交通信号后进行预测。Fan 等[14] 提出的 RGDAN 模型通过随机图注意力机制和自适应矩阵构建的图扩散模块动态调整邻接权重，并融合局部与全局空间依赖以提升预测精准度。

深度学习领域中原本用于自然语言处理与计算机视觉任务的注意力机制[15] 以及使用注意力机制作为核心的 Transformer 模型[16] 也在交通流预测领域优秀表现。如 Jiang 等[17] 提出了一种包含传播延迟感知模块的 Transformer 变体模型 PDFormer，通过动态空间注意力和延迟感知模块捕捉动态时空依赖性和长距离空间关联。Liu 等[18] 提出基于自适应嵌入的 Transformer 变体模型 STAEformer，在简单结构下实现有效预测。Chen 等[19] 提出 DTRformer模型通过跨时空注意力机制融合动态趋势与静态图信息，有效捕捉节点间的时空关联。Bai 等[20] 提出基于 Transformer 的时空预测模型 T-Graphormer，通过融合时间编码和全局注意力机制在统一框架中直接建模节点间的时空依赖关系。另外，针对注意力机制

收稿日期： 2025-02-25
基金项目： 上海市 2024 年度“科技创新行动计划”资助(24BC3200500, 24BC3200300)
作者简介： 郭津延(1999—)，男，云南人，硕士生，主要研究方向为人工智能和智能交通。E-mail: Y80220103@mail.ecust.edu.cn
通信联系人： 郑 红，E-mail: zhenghong@ecust.edu.cn
引用本文： 郭津延，郑 红，杜佳宇，等. 基于双尺度自适应令牌注意力的交通流量预测 [J]. 华东理工大学学报(自然科学版)，2025，51(6)：817-826.
Citation: GUO Jinyan, ZHENG Hong, DU Jiayu, *et al*. Traffic Flow Prediction Based on Dual-Scale Adaptive Token Attention[J]. Journal of East China University of Science and Technology, 2025, 51(6): 817-826.

的改进在交通流预测领域同样重要。如 Han 等[21]提出了一种新的注意力范式 Agent Attention，在计算机视觉领域以较低计算成本实现了较强的模型表达能力。Zhu 等[22]提出了 Bi-Level Routing Attention，通过查询感知筛选相关区域，在计算机视觉领域实现了高效长程依赖建模，降低计算复杂度并提升了性能。

上述研究在以下几个方面仍存在不足，首先，当前 Transformer 变体模型中对注意力机制的研究仍有较高的计算复杂度，限制了模型在实时交通流预测中的应用；其次，当前研究要求性能较强的计算硬件资源，当处理大规模交通网络时容易出现计算瓶颈，影响预测效率；此外，现有方法在整合局部与全局交通特征方面仍不够充分，难以全面反映复杂的交通流动态。因此，如何在保证预测准确性的同时降低计算开销，提升模型的实时性，仍是亟待解决的挑战。

针对上述问题，本文提出了一种基于双尺度自适应令牌注意力(Dual-Scale Adaptive Token Attention, DSATA)的交通流量预测模型(Dual-Scale Adaptive Token Attention Transformer, DSAFormer)。DSAFormer 模型通过 DSATA 能够高效地捕捉交通流数据中复杂的时空依赖关系，从而提升预测精准度。同时 DSATA 具有近线性的计算复杂度，解决了传统注意力机制在计算资源和时间上的高消耗问题，提升了模型预测的实时性，并减少了对高性能计算设备的依赖。

1 基于双尺度自适应令牌注意力的交通流量预测模型

1.1 模型架构

DSAFormer 模型结构如图 1 所示，包含输入层、嵌入层、时间编码模块、空间编码模块以及线性预测模块。其中，时间编码模块和空间编码模块均分别由 L_{Layer} 层相同的编码器层堆叠组成。每层编码器层都包含 DSATA 模块、前馈网络以及残差连接与层归一化等基本组件。在 DSATA 模块中，首先由两套并行的可学习自适应池化分别得到交通流数据的长期和短期令牌，再通过双尺度令牌聚合得到最终的自适应令牌，随后通过深度卷积(Depth-Wise Convolution, DWC)[23]处理输入注意力计算的值向量以增强局部领域信息，最后通过多头自注意力(Multi-Head Self-Attention)机制捕获全局时间依赖关系。另外，前馈网络能够有效处理和转换注意力机制捕捉到的特征信息，提升模型的拟合能力。残差连接与层归一化则用来缓解深层网络中的梯度消失问题，并提升模型的稳定性和鲁棒性。

1.2 DSATA

传统的注意力机制，尤其是传统 Transformer 架

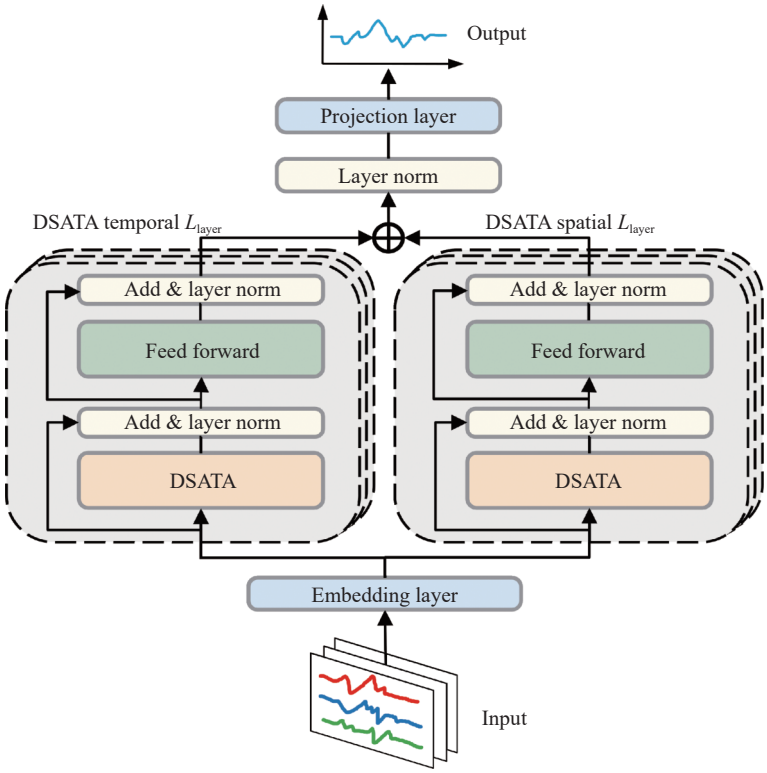

图 1 DSAFormer 模型框架

Fig. 1 Architecture of the DSAFormer model

构中的自注意力机制因能有效捕捉长距离依赖关系而被广泛应用于各种领域。但是标准的自注意力机制需要将每个输出位置与所有输入位置进行配对计算, 导致其计算复杂度与输入序列长度的平方成正比。而交通流量预测是典型的时空预测任务, 既有空间节点又有时间步长, 若使用包含标准自注意力机制的传统 Transformer 模型处理规模较大的交通流数据集需要巨大的资源消耗与较长的训练、推理时长。传统的使用 Softmax 函数的自注意力机制的时间复杂度为 $O(L^2d)$, 其中 L 为序列长度, d 为嵌入向量维度。通过改写相似性函数为分解形式的线性注意力机制的时间复杂度为 $O(Ld^2)$, 虽然其降低了计算的时间复杂度, 但也导致部分全局信息丢失, 模型表达力减弱。

本文提出 DSATA 并应用于 Transformer 变体模型以解决上述问题, DSATA 模块的结构如图 2 所示。DSATA 集成了 Softmax 注意力和线性注意力的优点, 首先通过双尺度自适应池化分别得到代表短期和长期数据特征的令牌, 然后通过双尺度自适应聚合得到最终令牌后进行近线性的注意力计算, 同时使用卷积增强特征表达, 使模型在交通流预测任务中的计算复杂度和计算开销降低并提升了模型的预测精度。图中虚线框为 DSATA 模块的双尺度自适应令牌聚合(Dual-Scale Adaptive Token Aggregation, DSA)组件。

传统注意力机制中计算查询向量 $\boldsymbol{Q}$、键向量 $\boldsymbol{K}$、值向量 $\boldsymbol{V}$ 的全局交互时, 复杂度为 $O(L^2d)$, 这种全连接的计算代价十分高昂, 因此 DSATA 模块中引入了双尺度自适应令牌 $\boldsymbol{T}$ 以减少计算量。$\boldsymbol{T}$ 是通过 DSATA 的 DSA 组件生成的。具体而言, 首先通过两组并行的 MLP、Softmax 激活函数以及加权求和对查询向量 $\boldsymbol{Q}$ 进行处理, 分别得到短期和长期尺度下的令牌, 最后通过聚合以及层归一化得到最终的自适应令牌 $\boldsymbol{T}$。根据不同数据集特点调整短期与长期尺度令牌数量的比例, 得到不同尺度下的特征信息 $\boldsymbol{T}$, 可表示为 $\boldsymbol{T}=\mathrm{DSA}(\boldsymbol{Q})$。其中 $\boldsymbol{T}\in\mathbb{R}^{B\times n\times d}$, B 表示批次大小, n 是自适应令牌的数量, DSA(·) 表示双尺度自适应池化与聚合过程。在 DSATA 中, 通过 $\boldsymbol{T}$ 从 $\boldsymbol{K}$ 和 $\boldsymbol{V}$ 中提取全局信息, 然后再将这些信息回传给查询向量。具体来说分为两步: 首先将 $\boldsymbol{T}$ 视为查询, 与 $\boldsymbol{K}$ 和 $\boldsymbol{V}$ 交互计算提取特征, 从所有值中聚合得到令牌特征 V_T。然后, 将 $\boldsymbol{T}$ 作为键向量, V_T 作为值向量, 与 $\boldsymbol{Q}$ 进行第 2 次计算, 得到最终输出 $\boldsymbol{O}_T$。其公式可表示为:

$$\boldsymbol{V}_T=\mathrm{Softmax}(\boldsymbol{T}\boldsymbol{K}^{\top})\cdot\boldsymbol{V} \tag{1}$$

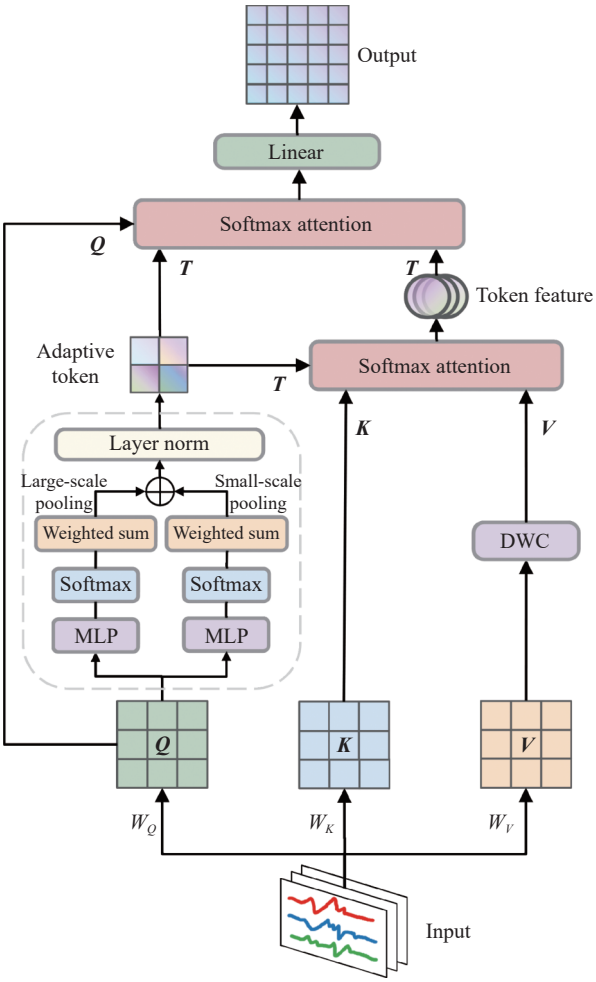

图 2　双尺度自适应令牌注意力模块框架

Fig. 2　Architecture of the DSATA module

$$\boldsymbol{O}_T=\mathrm{Softmax}(\boldsymbol{Q}\boldsymbol{T}^{\top})\cdot\boldsymbol{V}_T \tag{2}$$

$$\boldsymbol{O}_T=\mathrm{Softmax}(\boldsymbol{Q}\boldsymbol{T}^{\top})\cdot(\mathrm{Softmax}(\boldsymbol{T}\boldsymbol{K}^{\top})\cdot\boldsymbol{V}) \tag{3}$$

其中, Softmax(·) 表示标准的 Softmax 函数。进一步简化式(3), 将其改写为 2 次 Softmax 操作的嵌套, 并定义两个线性映射函数如下:

$$\varphi_q(\boldsymbol{Q})=\mathrm{Softmax}(\boldsymbol{Q}\boldsymbol{T}^{\top}) \tag{4}$$

$$\varphi_k(\boldsymbol{K})=\mathrm{Softmax}(\boldsymbol{T}\boldsymbol{K}^{\top})^{\top} \tag{5}$$

将式(4)和式(5)代入式(3), 得到:

$$\boldsymbol{O}_T=\varphi_q(\boldsymbol{Q})\cdot(\varphi_k(\boldsymbol{K})^{\top}\boldsymbol{V}) \tag{6}$$

由此可以看出, DSATA 通过两个 Softmax 注意力实现了一种广义的线性注意力, 将原本的 $O(L^2d)$ 复杂度降为 $O(Lnd)$, 通常情况下 $n\ll L$。但上述广义线性注意力机制仍存在特征表达能力不足的问题。为解决这个问题, DSATA 在注意力计算中的 $\boldsymbol{V}$ 使用 DWC 对其特征进行局部滤波, 再将卷积后的 $\boldsymbol{V}$ 与注意力权重相乘, 完成全局加权求和。该设计既保留了注意力的全局依赖, 又通过 DWC 引入了局部

特征信息,有效提升了特征表示能力。完整的DSATA计算过程可以表示为:

$$\boldsymbol{O}_T = \mathrm{Softmax}(\boldsymbol{Q}\boldsymbol{T}^{\top})\cdot(\mathrm{Softmax}(\boldsymbol{T}\boldsymbol{K}^{\top})\cdot \mathrm{DWC}(\boldsymbol{V})) \quad (7)$$

其中,$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\in\mathbb{R}^{B\times L\times d}$,$\boldsymbol{T}\in\mathbb{R}^{B\times n\times d}$。DSATA模块在接近线性计算时间复杂度前提下高效捕获交通流数据中时间与空间维度的特征,同时通过DSATA中的DWC补充额外的局部特征,进一步增强预测的精准性。

1.3 数据嵌入层

先将数据嵌入层中来源于交通流量监测器的原始输入 $X\in\mathbb{R}^{B\times S\times N\times d_{\mathrm{input}}}$ 经过一个全连接变换映射到低维空间,得到初步的特征表示 $G\in\mathbb{R}^{B\times S\times N\times d_g}$。其中,$S$为时间步,$N$为节点数,$d_g$为投影后的特征维度,该过程可表示为

$$G = \varnothing(X) \quad (8)$$

其中,$\varnothing(\cdot)$表示全连接层的映射关系。为了充分捕捉时间信息,该层包含两个嵌入模块,分别编码每日与每周。具体来说,时间嵌入将一天内的多个时刻编码为d_{time}维度的向量,可表示为

$$\boldsymbol{H}_{\mathrm{time}} = \mathrm{Embed}_{\mathrm{time}}(\mathrm{time}) \quad (9)$$

星期嵌入则将一周七天的信息映射为d_{week}维度的向量,可表示为

$$\boldsymbol{H}_{\mathrm{week}} = \mathrm{Embed}_{\mathrm{week}}(\mathrm{week}) \quad (10)$$

然后将这两个嵌入特征在最后一个维度上进行拼接便得到了统一的嵌入表示 $\boldsymbol{Z}\in\mathbb{R}^{B\times S\times N\times d_z}$,其中

$$d_z = d_g + d_{\mathrm{time}} + d_{\mathrm{week}} \quad (11)$$

而嵌入表示$\boldsymbol{Z}$的计算可表示为

$$\boldsymbol{Z} = G \,\|\, \boldsymbol{H}_{\mathrm{time}} \,\|\, \boldsymbol{H}_{\mathrm{week}} \quad (12)$$

该数据嵌入层设计将原始数值信息和时间周期性信息有效地整合到一个高维特征表示中,为后续的时空特征提取和注意力计算提供了充足的上下文信息。

1.4 预测输出层

在预测输出层中,模型首先得到时间与空间维度DSATA模块输出并融合后的特征表示 $\boldsymbol{X}\in\mathbb{R}^{B\times S\times N\times d'}$,其中$B$为批次大小,$d'$为融合后特征维度。随后预测输出层会将$\boldsymbol{X}$进行转置和重构,将时间步与特征维度合并形成新表示 $\widehat{\boldsymbol{X}}\in\mathbb{R}^{B\times N\times(S\times d)}$,然后再通过线性映射将其转换为目标预测维度,将每个节点的长向量映射到预测输出空间,可表示为

$$\boldsymbol{Y}' = \mathrm{Linear}\left(\widehat{\boldsymbol{X}}\right)\in\mathbb{R}^{B\times N\times(T'\times d_{\mathrm{out}})} \quad (13)$$

其中,T'表示预测的未来时间步长,d_{out}表示每个节点在每个预测时间步上输出的特征维度,该值在本文中为1,即表示只预测交通流量这一个值。最后将$\boldsymbol{Y}'$重塑并转置后得到最终的预测输出 $\boldsymbol{Y}\in\mathbb{R}^{B\times T'\times N\times d_{\mathrm{out}}}$。

通过上述预测层模型,能够在对融合后的高维特征进行整体建模,然后由一个全连接层将信息映射到所需输出空间。

2 实验结果与分析

本文在两个交通流量预测任务常用数据集上进行了DSAFormer模型性能评估实验,并与基线模型进行了对比分析;同时还对DSAFormer模型进行了计算资源消耗实验和消融实验。

2.1 数据集

本文使用了交通流量预测领域常用的两个基准数据集来验证DSAFormer的有效性,两个数据集均为加利福尼亚州交通数据采集系统(PeMS)收集的实时交通流量数据,分别是PeMS04、PeMS08数据集,数据集信息详见表1。实验中的两个数据集均按照时间顺序对训练集、验证集和测试集以6∶2∶2的比例进行划分。此外,实验过程中使用过去连续的12个时间步数据来预测未来12个时间步的交通流量。

2.2 基线模型与评价指标

为了完整评估DSAFormer模型的性能,本文使用包含经典统计方法、传统机器学习、深度学习和Transformer变体模型在内的多个基线模型进行对比。

实验中使用交通流量预测领域中的3个常用评价指标来评估模型的性能,分别是:平均绝对误差(MAE)、均方根误差(RMSE)和平均绝对百分比误差(MAPE),计算公式如下:

$$\mathrm{MAE}(y,\widehat{y}) = \frac{1}{N}\sum_{i=1}^{N}\left|y_i-\widehat{y}_i\right| \quad (14)$$

$$\mathrm{RMSE}(y,\widehat{y}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\widehat{y}_i)^2} \quad (15)$$

$$\mathrm{MAPE}(y,\widehat{y}) = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i-\widehat{y}_i}{y_i}\right| \quad (16)$$

其中,N为样本数量,$\widehat{y}_i$为预测值,y_i为真实值。

2.3 实验设置

DSAFormer模型在Python 3.8和PyTorch 1.11环境下开发实现,本文中的实验均在同一个实验平台进行,其搭载了24 GB显存的NVIDIA GeForce RTX 4090D GPU、Intel Xeon E5-2683 V4 16核CPU以及32 GB内存。在模型训练过程中采用Huber Loss作

表 1　数据集详细信息

Table 1　Details of datasets

Datasets	Node	Edge	Time step	Time interval/min	Time range
PeMS04	307	340	16992	5	2018/01/01—2018/02/28
PeMS08	170	295	17856	5	2016/07/01—2016/08/31

为损失函数, 并使用 Adam 优化器进行参数优化, 初始学习率设置为 0.001。模型的输入特征维度为 3, 输出特征维度为 1, 时间步为 12, 预测步为 12, 注意力总头数为 4, DSAFormer 中时间与空间 DSATA 层的数量均为 3, 时间维度 DSATA 层的自适应令牌数量为 12, 且长、短期令牌数量比例为 1∶1, 空间维度 DSATA 层的自适应令牌数量为 48, 且长、短期令牌数量比例为 2∶1, 模型中 Dropout 比例为 0.1。

2.4　模型性能比较

2.4.1　模型预测性能比较　为了验证 DSAFormer 模型的性能, 本文在 PeMS04 与 PeMS08 两个数据集上进行性能对比实验, 具体预测结果见表 2, 表现最好的结果用黑体表示。从表 2 中可以看出, 本文提出的 DSAFormer 模型在两个数据集中的所有指标都有明显降低, 优于其他基线模型, 证明了模型的有效性。在 PeMS04 数据集中, DSAFormer 在 MAE、RMSE 和 MAPE 指标上分别比基准模型中最优秀方法 PM-DMNet(R) 降低了 1.20%、2.44% 和 0.33%。在 PeMS08 中 DSAFormer 在 MAE、MAPE 指标上分别比 PM-DMNet(R) 降低了 0.30% 和 1.24%, 在 RMSE上略微上升 0.09%。

表 2　DSAFormer 模型与基线模型性能比较

Table 2　Performance comparison of DSAFormer and baseline models

Dataset	PeMS04			PeMS08		
	MAE	RMSE	MAPE/%	MAE	RMSE	MAPE/%
HA	38.03	59.24	27.88	34.86	59.11	25.24
ARIMA	33.73	48.80	24.18	31.09	44.32	22.73
VAR	24.54	38.61	17.24	19.19	29.81	13.10
GraphWaveNet	24.89	39.66	17.29	18.28	30.05	12.15
ASTGNN	18.60	30.91	12.36	15.00	24.70	9.50
STWave	18.50	30.39	12.43	13.42	23.40	8.90
DDGCRN	18.45	30.51	12.19	14.40	23.75	9.40
PM-DMNet(P)	18.34	30.36	12.05	13.55	23.35	9.04
PM-DMNet(R)	18.37	30.68	12.01	13.40	23.22	8.87
PDG2Seq	18.58	31.02	12.36	13.54	**23.19**	8.89
STAEformer	18.22	30.18	11.98	13.46	23.25	8.88
PDFormer	18.32	29.97	12.10	13.58	23.51	9.05
DSAFormer	**18.15**	**29.93**	**11.97**	**13.36**	23.24	**8.76**

表 2 中的结果表明, 经典统计分析方法 HA、ARIMA 和 VAR 仅能处理简单、平稳的线性数据, 对于复杂的非线性数据预测效果较差。而基于图神经网络的模型如 GraphWaveNet 和 STWave 在一定程度上考虑了空间特征, 但其对时序特征的捕捉仍存在不足。DDGCRN 通过引入动态空间依赖性和双向时序建模, 表现优于前述模型。然而, 现有模型在捕捉双尺度时空依赖性和动态关注关键区域方面仍存在显著的局限性, 未能充分挖掘交通数据中的复杂时空特征。这些不足导致其在复杂交通环境下的预测性能和适应性受到限制。

针对上述问题, 本文提出的 DSAFormer 模型综合考虑了交通数据的长期与短期尺度差异性、时空相关性以及空间异质性, 并且分别沿时间与空间维度获取数据的短期与长期代表性特征, 利用 DSATA 进行建模, 有效地捕捉了不同尺度上的时空依赖关系, 提升了模型对复杂交通模式的理解能力。这些改进使得 DSAFormer 在 PeMS04 和 PeMS08两个数据集上的绝大多数指标优于其他基线模型, 特别是在 PeMS04 数据集上, 提升幅度更为显著, 验证了其在复杂交通环境中的优越性能。

在 PeMS08 和 PeMS04 数据集中随机选择一个节点(27 号节点)的前 8000 个时间步进行可视化分析, 结果如图 3 和图 4 所示。由图可知, DSAFormer 模型在交通流量较小(即非高峰时间段)有较好的拟合效果, 在高峰期虽然不能完全复现真实数据的短期震荡, 但整体也能较好地拟合真实数据, 以上表明 DSAFormer 模型能够准确捕捉到交通流数据的内在联系, 并在不同尺度和周期下进行较为精准地预测, 验证了 DSAFormer 模型的有效性和先进性。

从表 2 中选择部分具有代表性的基线模型与 DSAFormer 模型进行性能可视化对比分析, 结果如图 5 和图 6 所示。由图可知, 在 PeMS04 和 PeMS08 两个数据集上, 相较其他基线模型 DSAFormer 在不同评价指标下始终保持较低的预测误差, 尤其是在预测步数较大即长预测序列时优势更为明显, 说明

DSAFormer在捕捉复杂时空关联与缓解长期累积误差上有更强的建模能力。其他基线模型虽在短期预测上与DSAFormer差距不大,但随着预测步数的增加,其误差上升速度更快,说明DSAFormer拥有更好的泛化能力与稳定性,能够有效地应对随时间不断增长的不确定性。

综上所述,DSAFormer在降低模型计算时间复杂度的前提下,克服了传统模型在捕捉复杂时空依赖性和注意力关注关键区域方面的不足,在预测精度上取得了显著提升。

2.4.2 模型计算复杂度比较 DSAFormer模型的计算时间复杂度主要受节点数 N 和时间步长 S 的影响。通过多次实验,并综合考虑模型预测精度、训练与推理时长以及显存占用后,发现当时间维度DSATA的自适应令牌数取值 $n_t=S$,空间维度DSATA的自适应令牌数取值 n_s=48时模型综合性能达到最优。因此,时间维度DSATA的计算时间复杂度为 $O(Sn_td)=O(S^2d)$,与标准自注意力复杂度相同,并未实现时间复杂度降低;空间维度DSATA的计算时间复杂度为 $O(Nn_sd)$,由于 $n_s \ll N$ 且 n_s 固定为49,在复杂度表达式中退化为常数项,使时间复杂度变为 $O(Nd)$,相比标准自注意力显著降低了计算时间复杂度。模型整体计算复杂度为 $O(S^2d+Nd)$,由于预测时间步长 S 固定为12,同样在式中退化为常数项,所以DSAFormer的实际计算时间复杂度为 $O(Nd)$,即与节点数 N 呈线性关系。

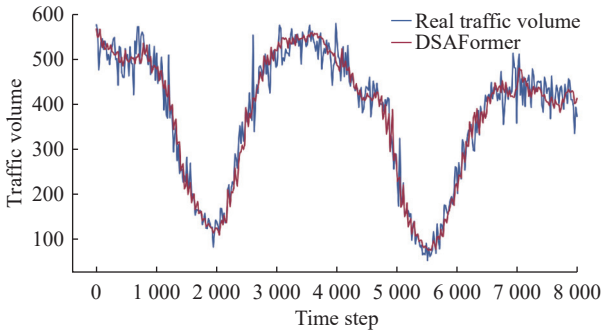

图3 DSAFormer预测值与PeMS08真实值对比

Fig. 3 Comparison of DSAFormer predicted values and PeMS08 true values

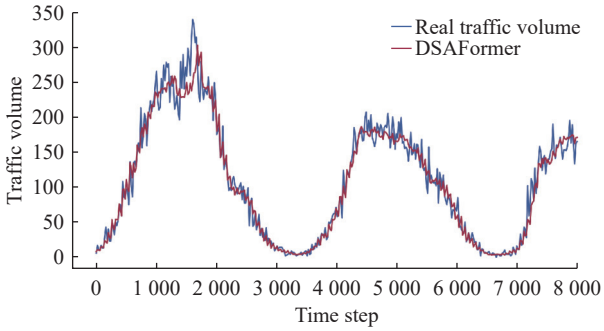

图4 DSAFormer预测值与PeMS04真实值对比

Fig. 4 Comparison of DSAFormer predicted values and PeMS04 true values

为了验证DSAFormer模型在计算复杂度与计算资源开销方面的优化与提升,本文在前述数据集上

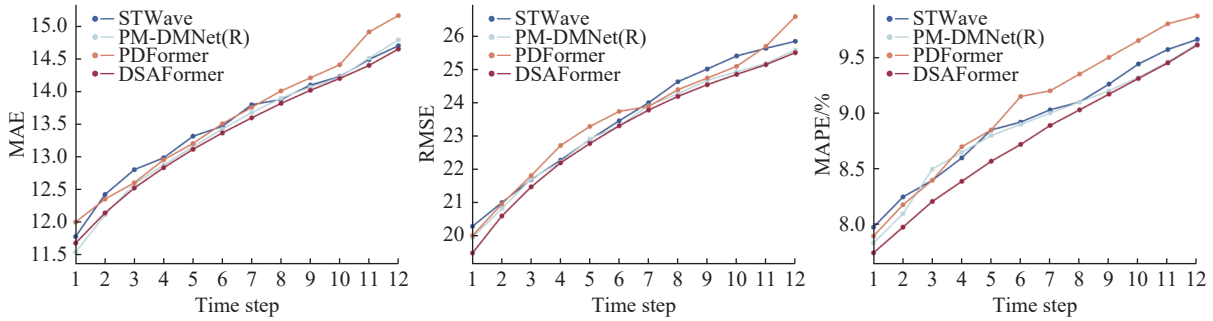

图5 在PeMS08上每个预测时间步的预测误差

Fig. 5 Prediction errors at each forecast time step on PeMS08

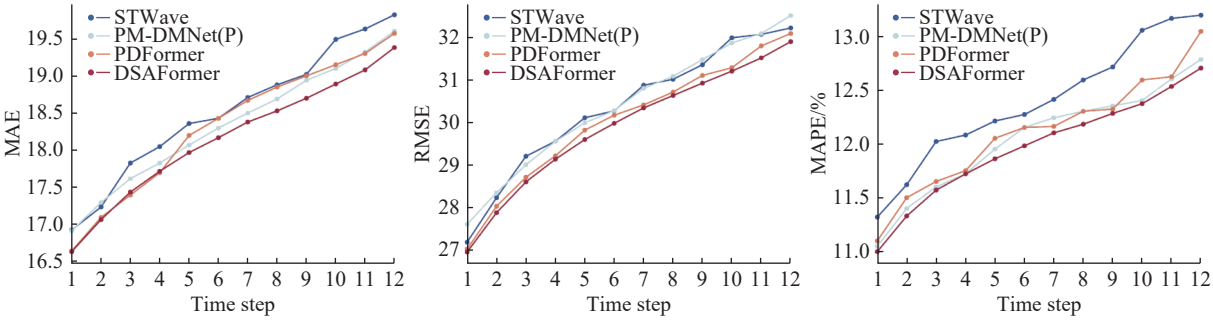

图6 在PeMS04上每个预测时间步的预测误差

Fig. 6 Prediction errors at each forecast time step on PeMS04

进行对比实验。由于 DSATA 改进了传统注意力机制, 故选取利用交通流预测任务中的 Transformer 变体模型 DTRFormer、PDFormer 以及非 Transformer 模型 DDGCRN 作为对照进行对比实验。为进一步分析 DSATA 模块对模型复杂度的影响, 构建了将标准自注意力机制代替 DSATA 的对比模型 DSAFormer (SA)。为了比较模型的计算时间复杂度与计算资源开销, 选取模型的单轮训练时长、推理时长与显存占用大小作为评价指标, 实验结果如表 3 所示。

表 3　DSAFormer 模型与基线的计算复杂度比较

Table 3　Comparison of computational complexity between the DSAFormer and the baseline model

Dataset	PeMS04			PeMS08		
	Training time/ (s·epoch^{-1})	Inference time /s	GPU memory usage/MB	Training time/ (s·epoch^{-1})	Inference time/s	GPU memory usage/MB
DDGCRN	367.29	65.36	10717	104.95	4.71	6051
PDFormer	128.32	8.92	10885	69.49	3.82	5469
DTRformer	104.12	19.44	8789	57.24	14.67	4999
DSAFormer(SA)	64.46	5.63	7051	46.48	3.04	4853
DSAFormer	**56.32**	**4.73**	**6255**	**37.45**	**2.85**	**4637**

相较于其他基于 Transformer 架构的模型, 本文提出的 DSAFormer 模型在两个数据集上的训练时长、推理时长以及显存占用上都有显著降低, 优于其他基线模型, 证明了模型的高效性与实时性。在 PeMS04 中, DSAFormer 在训练时长、推理时长和显存占用指标上分别比先进方法 DTRformer降低了 45.91%、75.67% 和28.83%。在PeMS08 中,DSAFormer 在训练时长、推理时长和显存占用指标上分别比先进方法 DTRformer 降低了 34.57%、80.57%和 7.24%。可以看出, 相较于其他基于注意力或 Transformer 的交通流预测模型, 通过 DSATA 近线性计算时间复杂度 DSAFormer 在保证预测精准度的前提下大幅降低了模型计算复杂度, 使得模型训练、推理时长和显存占用相比得到大幅降低, 提高了模型预测的实时性, 降低了模型对计算设备性能的要求。

由上述可知, 本文提出的 DSAFormer 模型通过 DSATA、沿着时间与空间维度在近线性的时间复杂度进行注意力计算, 减少了计算时间以及显存占用。同时 DSATA 捕获了长期与短期不同尺度下的代表特征进行注意力计算, 有效地对数据的时空特征进行了建模。

2.5　消融分析

为了进一步评估 DSAFormer模型中不同组件对模型预测性能的影响, 设计了以下 5 个 DSAFormer 变体模型进行消融分析实验, 分别如下:

(1) T-SA_S-DSATA: 使用标准自注意力模块 (Self-Attention, SA)替代时间维度 DSATA 模块, 以研究 DSATA 模块处理时间维度特征对模型预测的影响。

(2) T-DSATA_S-SA: 使用 SA 替代空间维度 DSATA 模块, 以研究 DSATA 模块处理空间维度特征对模型预测的影响。

(3) T-SA_S-SA: 采用 SA 分别替代处理时间、空间维度的 DSATA 模块, 作为对照以研究 DSATA 模块对模型预测的影响。

(4) w/o DSA: 移除 DSATA 中的 DSA 组件, 用单一尺度的可学习池化替代, 以研究双尺度聚合对模型预测结果的影响。

(5) w/o DWC: 移除 DSATA 中的 DWC, 以研究 DWC 捕捉局部信息对模型预测结果的影响。

除上述模型结构差异之外, 所有的变体模型在参数设置上与 DSAFormer 保持一致, 实验结果如表 4 所示。由表可以看出, 移除模型中任一维度的 DSATA 模块均会对模型的预测精度产生较大影响, 这证明了 DSATA 模块在处理时间与空间维度特征方面的有效性。具体而言, T-SA_S-DSATA 模型在 PeMS04 上的预测精度明显低于完整模型, 这表明去除处理时间维度的 DSATA 模块削弱了模型对时间特征的捕捉能力。然而, T-DSATA_S-SA 模型的性能下降幅度更大, 说明去除处理空间维度的 DSATA 模块对模型的空间特征捕捉产生了更显著的负面影响。因此, 在当前的实验设置下 DSATA 模块在空间维度上的作用比时间维度更为重要。另外 T-SA_S-SA 模型的性能亦有所下降, 进一步强调了 DSATA 模块在综合处理时间和空间特征中的重要性。在 PeMS08 数据集上, 虽然 3 种变体模型的性能下降幅度较小, 但整体趋势一致, 并且空间维度的性能下降同样显著。

分别移除 DSA 组件与 DWC 组件的变体模型

表4 不同模块的消融实验

Table 4 Ablation experiment of different modules

Dataset	PeMS04			PeMS08		
	MAE	RMSE	MAPE/%	MAE	RMSE	MAPE/%
T-SA_S-DSATA	18.22	29.87	11.98	13.38	23.01	8.90
T-DSATA_S-SA	18.46	30.21	12.19	13.42	**22.99**	8.90
T-SA_S-SA	18.31	30.27	12.01	13.37	23.05	8.88
w/o DSA	18.37	30.13	12.07	13.40	23.27	8.85
w/o DWC	18.32	30.20	12.00	13.38	23.22	8.83
DSAFormer	**18.12**	**29.84**	**11.94**	**13.36**	23.24	**8.76**

在PeMS04和PeMS08上的MAE、RMSE以及MAPE均出现不同程度的上升，说明去除两者中任一组件都会削弱DSAFormer模型对时空特征的捕捉能力，导致预测精度下降，表明DSA与DWC在捕捉多尺度特征、提升模型的预测精度方面存在重要作用。与部分替换掉DSATA模块的变体模型T-SA_S-DSATA、T-DSATA_S-SA相比，w/o DSA与w/o DWC性能下降幅度较小，说明DSATA模块本身对于同时建模时间和空间维度特征至关重要，而DSA与DWC模块则在进一步增强模型对时空依赖的捕捉、降低预测误差方面起到显著的辅助作用。对比w/o DSA与w/o DWC，去除DSA在两个数据集上会造成更明显的性能下降，尤其是MAE和MAPE评价指标上，说明DSA捕捉尺度时空特征在提升模型整体预测精度方面扮演了更关键的角色，而DWC虽然对局部特征提取与高效计算也同样重要，但其缺失导致的误差上升幅度略小。

消融分析实验表明DSAFormer中的DSATA模块在处理时间特征和空间特征时都具有关键作用。但是在空间维度上，DSATA模块对提升模型预测性能的贡献更为显著，表明DSATA模块在空间维度特征处理方面的优良性能，说明DSATA中的DSA组件捕捉不同尺度时空特征对模型预测精准度起关键作用。

2.6 超参数分析

DSATA模块中针对交通流数据长期与短期的双尺度设计是该模块的核心，即同时采用长期尺度和短期尺度的代理令牌来捕捉不同粒度的交通流数据特征。因此长期与短期尺度代理令牌的总数及其分配比例对模型能否有效捕获并融合双尺度特征起着重要作用。为验证不同令牌总数以及分配比例对模型预测效果的影响，本文分别通过调整DSATA模块中时间与空间维度的超参数来分析超参数值设置对模型预测性能的影响。该实验中分析的超参数为自适应令牌总数N'、长期与短期尺度令牌数量及其比值ta-ratio。

由于在2.4.1节中时间与空间维度的DSATA默认超参数设置不同，因此本节实验分为时间维度DSATA超参数实验与空间维度DSATA超参数实验两个部分。在两部分实验中同时调整令牌总数N'和长短期尺度令牌分配比例ta-ratio，具体实验方案如表5所示。根据超参数实验方案分别在PeMS04与PeMS08数据集上进行超参数实验，实验中除指定超参数外其余参数设置与2.3节设置相同。

根据上述方案进行实验后对实验结果进行可视化分析，分析结果如图7与图8所示。由图可知，当时间维度DSATA模块中N'=12、ta-ratio=1时，即短期与长期令牌均衡分配时能取得最低MAE；而当空间维度DSATA模块中N'=48、ta-ratio=2时，即长期尺度令牌数量是短期尺度的两倍时得到最低MAE值。这说明在时间序列的特征提取上，均衡分配短期与长期信息最有利于捕捉时序依赖；而在空间结构的特征提取上，更偏向于长期尺度信息有助于提炼全局空间模式，同时保留一定数量的短期尺度令牌来补充局部细节。当令牌总数过少或过多时，模型分别会面临信息表达不足或冗余过载的问题，从而导致MAE增大；并且ta-ratio过小或过大时，也难

表5 超参数实验方案

Table 5 Experimental setup for hyperparameters

Method	N'	ta-ratio	Number of long-timescale token	Number of short-timescale token
Temporal DSATA	8	0.5	3	5
	8	1	4	4
	8	2	5	3
	12	0.5	4	8
	12	1	6	6
	12	2	8	4
	16	0.5	5	11
	16	1	8	8
	16	2	11	5
Spatial DSATA	36	0.5	12	24
	36	1	18	18
	36	2	24	12
	48	0.5	16	32
	48	1	24	24
	48	2	32	16
	60	0.5	20	40
	60	1	30	30
	60	2	40	20

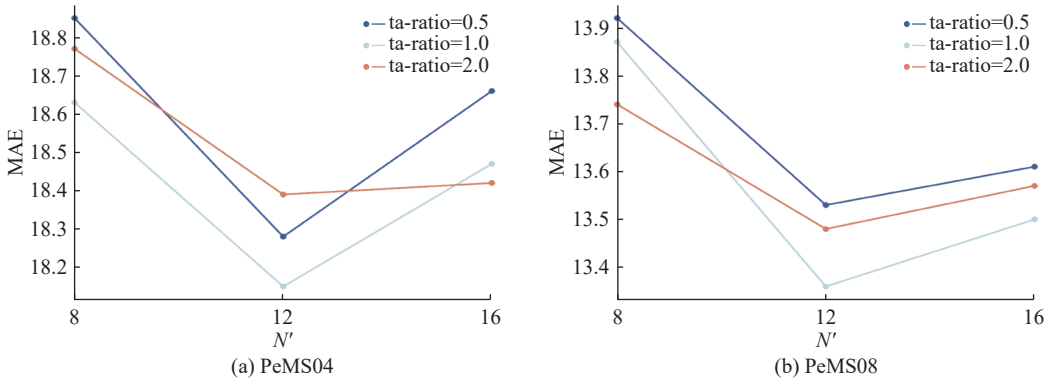

图 7　时间维度 DSATA 的超参数分析

Fig. 7　Hyperparameter analysis of DSATA in temporal dimension

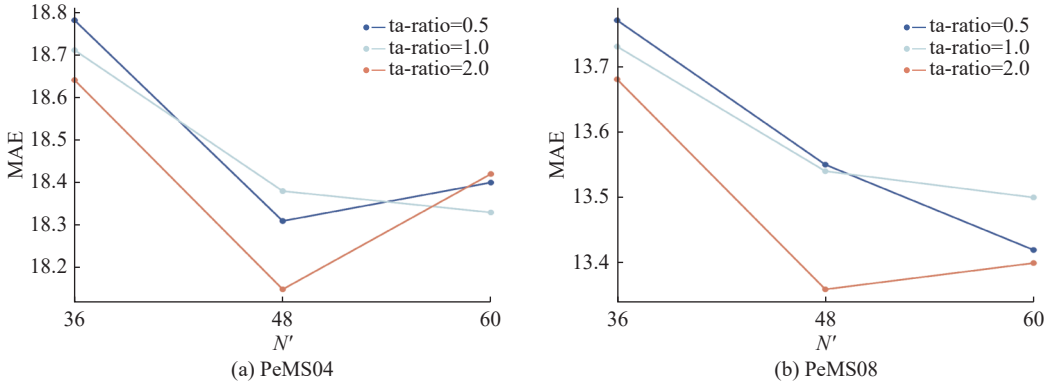

图 8　空间维度 DSATA 的超参数分析

Fig. 8　Hyperparameter analysis of DSATA in spatial dimension

以全面兼顾多尺度特征。这些结果验证了 DSATA 模块的多尺度代理设计思路: 在时间与空间两条路径上, 分别选择与其特征分布相匹配的令牌数量和分配比例, 能最大化发挥多尺度信息的优势, 提高模型的预测精度。

3　结　论

本文提出了一种基于双尺度自适应令牌注意力的交通流预测模型——DSAFormer。在双尺度自适应令牌注意力机制中, 首先通过双尺度自适应池化模块有效捕捉交通流数据的长期与短期特征, 生成两种尺度下的代表性令牌, 并通过聚合操作获得参与注意力计算的少量高效令牌。随后, 模型在时间与空间两个维度上分别进行近线性的注意力计算, 以充分挖掘时空特征之间的依赖关系。本文在两个公开交通流数据集上进行了充分的实验, 其结果表明 DSAFormer 在预测精度与计算复杂度两方面的各项评估指标中均优于现有的主流方法, 证明了其在交通流预测任务中的有效性与优越性。

参考文献:

[1] 熊章友, 李卫军, 朱晓娟, 等. 基于深度学习的短时交通流预测研究综述 [J/OL]. 计算机工程与应用, (2024-01-19) [2024-12-11]. https://kns.cnki.net/kcms/detail/11.2127.TP.20241210.1603.006.html.

[2] 乔少杰, 薛骐, 杨国平, 等. 基于动态自适应时空图的多元时序预测模型 [J]. 计算机学报, 2024, 47(12): 2925-2937.

[3] 李云, 高雅, 姚枝秀, 等. 面向数据稀缺场景的智能交通流量预测 [J/OL]. 软件学报, (2025-01-15) [2025-01-22]. https://doi.org/10.13328/j.cnki.jos.007239.

[4] WILLIAMS M B, HOEL A L. Modeling and forecasting vehicular traffic flow as a seasonal arima process: Theoretical basis and empirical results[J]. Journal of Transportation Engineering, 2003, 129(6): 664-672.

[5] ZIVOT E, WANG J. Vector Autoregressive Models for Multivariate Time Series[M]. NewYork, USA: Springer-Verlag, 2006.

[6] QI Y, ISHAK S. A hidden markov model for short term prediction of traffic conditions on freeways[J]. Transportation Research Part C, 2014, 4395-111.

[7] WU C, HO J, LEE D. Travel-time prediction with support vector regression[J]. IEEE Transaction on Intelligent Transportation Systems, 2004, 5(4): 276-281.

[8] 倪庆剑, 彭文强, 张志政, 等. 基于信息增强传输的时空图

神经网络交通流预测 [J]. 计算机研究与发展, 2022, 59(2): 282-293.
[9] 申岩松, 李琳, 黄传明. 全局和局部感知的交通速度预测模型 [J]. 电子学报, 2024, 52(9): 3195-3205.
[10] WENG W, WU M, JIANG H, *et al*. Pattern-matching dynamic memory network for dual-mode traffic prediction[EB/OL]. (2024-08-12) [2025-02-25]. https://doi.org/10.48550/arxiv.240807100.
[11] FANG Y, QIN Y, LUO H, *et al*. Spatio-temporal meets wavelet: Disentangled traffic flow forecasting via efficient spectral graph attention network[EB/OL]. (2021-12-06) [2025-01-20]. https://doi.org/10.48550/arXiv.2112.02740.
[12] FAN J, WENG W, CHEN Q, *et al*. PDG2Seq: Periodic dynamic graph to sequence model for traffic flow prediction[J]. Neural Networks: The Official Journal of the International Neural Network Society, 2024, 183: 106941.
[13] WENG W, FAN J, WU H, *et al*. A decomposition dynamic graph convolutional recurrent network for traffic forecasting[J]. Pattern Recognition, 2023, 142: 109670.
[14] FAN J, WENG W, TIAN H, *et al*. RGDAN: A random graph diffusion attention network for traffic prediction[J]. Neural Networks, 2024, 172: 106093.
[15] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[EB/OL]. (2014-09-01)[2025-01-21]. https://doi.org/10.48550/arXiv.1409.0473.
[16] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need[C]//Advances in Neural Information Processing Systems, NeurIPS. USA: Curran Associates Inc, 2017.
[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. Washington, DC, USA: AAAI, 2023, 37(4): 4365-4373.
[18] LIU H, DONG Z, JIANG R, *et al*. Spatio-temporal adaptive embedding makes vanilla transformer sota for traffic forecasting[C]//Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. Birmingham, United Kingdom: ACM, 2023: 4125-4129.
[19] CHEN J, YE H, YING Z, *et al*. Dynamic trend fusion module for traffic flow prediction[EB/OL]. (2025-01-18)[2025-02-21]. https://doi.org/10.48550/arXiv.2501.10796.
[20] BAI H Y, LIU X. T-graphormer: Using transformers for spatiotemporal forecasting[EB/OL]. (2025-01-22)[2025-02-21]. https://doi.org/10.48550/arXiv.2501.13274.
[21] HAN D, YE T, HAN Y, *et al*. Agent attention: On the integration of softmax and linear attention[C]//European Conference on Computer Vision. Milan, Italy: Springer Nature, 2024: 124-140.
[22] ZHU L, WANG X, KE Z, *et al*. Biformer: Vision transformer with bi-level routing attention[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, BC, Canada: IEEE, 2023: 10323-10333.
[23] CHOLLET F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017: 1251-1258.

Traffic Flow Prediction Based on Dual-Scale Adaptive Token Attention

GUO Jinyan[1], ZHENG Hong[1], DU Jiayu[1], LUO Yujian[1], LI Pengwei[1], SHAN Rongsheng[2]
(1. School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China; 2. School of Cyber Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: To address the shortcomings of existing methods in traffic flow prediction concerning computational complexity, real-time performance, and the integration of local and global features, this paper proposes a traffic flow prediction model that employs dual-scale adaptive token attention. The model incorporates a dual-scale adaptive token attention mechanism designed to extract complex spatio-temporal features while optimizing computational efficiency. Through dual-scale learnable pooling operations, the resulting tokens effectively capture both long-term and short-term temporal features of the data. Furthermore, the adaptive token attention mechanism integrates global dependencies to enhance prediction accuracy and operational efficiency. Experimental results on two public datasets demonstrates that the proposed method outperforms existing mainstream models in both prediction accuracy and computational efficiency. Particularly suitable for real-time traffic flow prediction scenarios, this approach provides an efficient and accurate solution for intelligent transportation systems, exhibiting significant theoretical and practical implications.

Key words: intelligent transportation; traffic flow prediction; attention mechanism; Transformer; deep learning

（责任编辑：王晓丽）