LLGformer: Learnable Long-range Graph Transformer for Traffic Flow Prediction

LLGformer：用于交通流预测的可学习远程图转换器

汇报人：薛彤

2025/9/7

研究背景

局限性

交通流预测

视为时间序列预测问题：难以处理涉及各种远程依赖关系的复杂多样的空间相关性。

视为时空图建模问题：存在许多局限性。
1，现有模型缺乏构建时空图的有效方法。构建高效的全局时空图至关重要，如图1（b）、（c）。
2，现有模型未充分利用历史信息。大多数当前方法仅根据前一小时的数据预测流量，而忽略周期性模式，如图1（c）、（d）。
3，模型架构设计存在局限性。如自回归错误传播、自注意力机制时间信息丢失、编码器和解码器交叉注意力信息冗余。

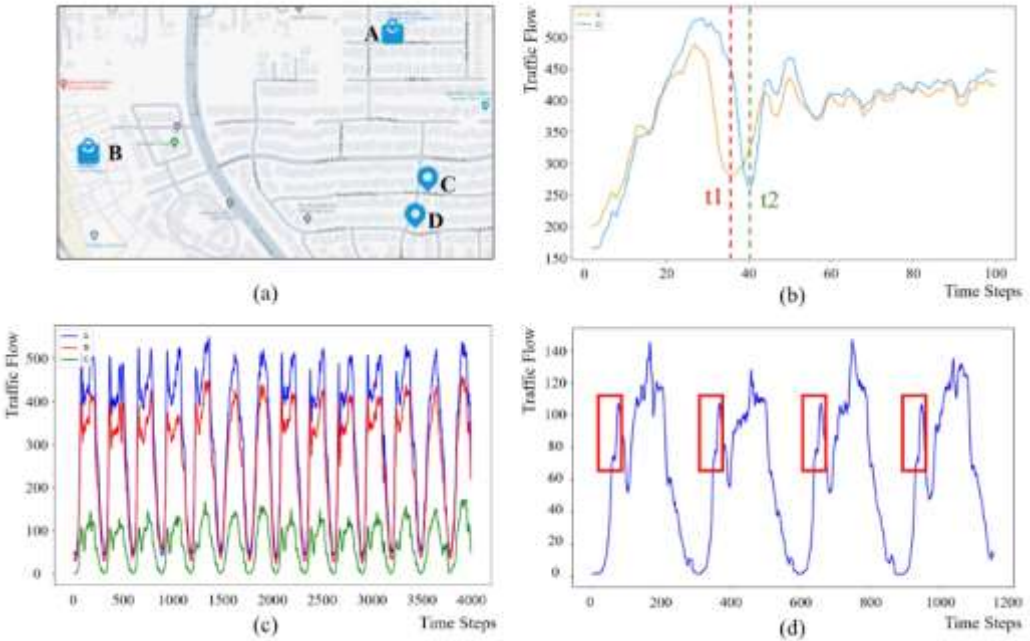


图1：（a）显示了某个地区的路线图。（b）、（c）和（d）显示了A、B、C和 D 四个传感器在不同时间段内记录的交通数据变化。

本文贡献

- 确定了历史信息和远程时空关系在交通预测任务中的重要性，指出了现有方法中图构建和模型结构的不足。
- 我们提出了一种新颖的图嵌入方法和编码技术来学习每个传感器的数据表示，并设计了一个有效的模型来捕获长序列输入中的长程依赖关系。此外，还引入了两种优化策略来提高模型效率。
- 通过实验分析，所提模型可以在四个基准数据集上实现高性能。

问题定义

定义 1（道路网络）

道路网络：$\mathcal{G} = (\mathcal{V}, \mathcal{E}, A)$，作为交通流量预测的输入

节点集：$\mathcal{V} = v_1, v_2, \cdots, v_N\}$

节点数：$|\mathcal{V}| = N$

边集：$\mathcal{E}$

图的邻接矩阵：$A \in \mathbb{R}^{N \times N}$，存储了路网中传感器之间的距离，如果节点i和节点j之间存在边，则 $\mathcal{A}_{i,j} > 0$，否则为 0。

定义 2（交通流张量）

使用 $X^t \in \mathbb{R}^{N \times d}$ 来表示在时间戳 t 观察到的道路网络中 N 个节点的交通流，其中 d 是特征的数量。

使用 $X = (X^1, X^2, \cdots, X^T) \in \mathbb{R}^{T \times N \times d}$ 来表示所有节点在总 T 时间戳上的流量张量。

将预测形式化为学习交通预测模型 f，交通流预测表示为：

$$[\mathrm{X}^{(t-T):t}, \mathcal{G}] \xrightarrow{f} \mathrm{X}^{(t+1):(t+T')}. \qquad (1)$$

整体结构

□ 构建步骤

1，提出整合历史交通信息以构建可学习的全局时空图的方法。
2，提出模型架构，该架构由编码、编码器和解码器组件组成，有效地统一了时空信息。
3，引入两种优化策略，以进一步降低计算复杂度并提高模型效率。

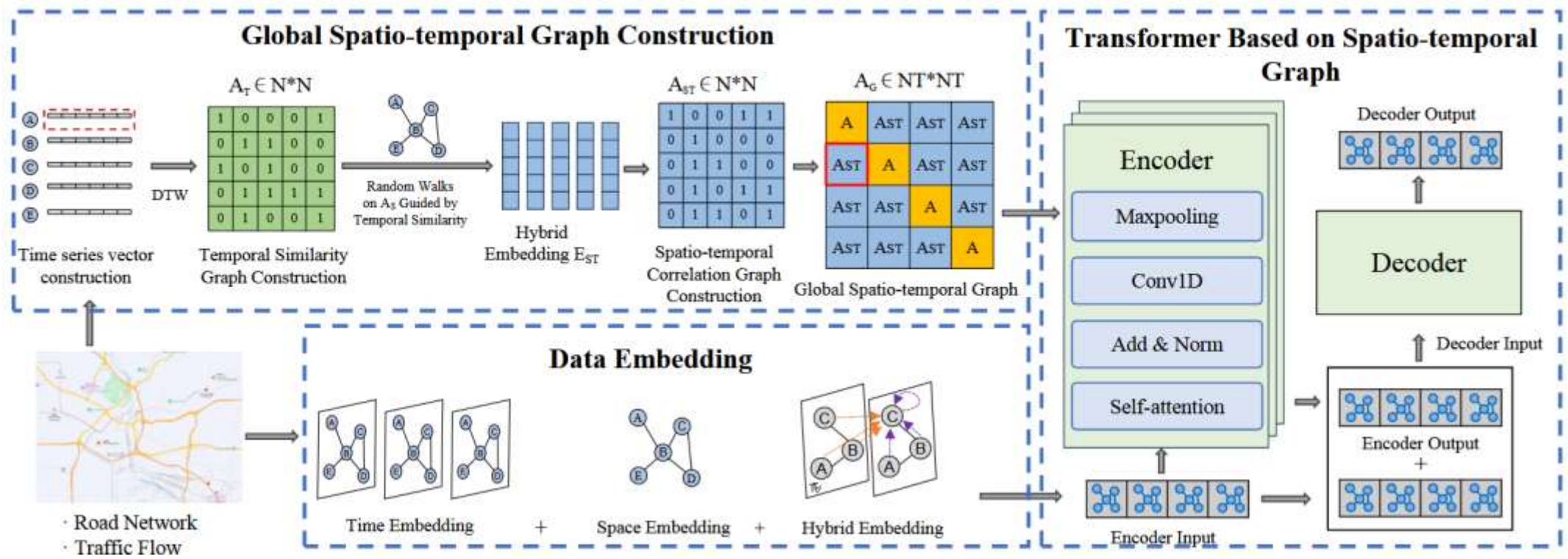


图 2：LLGformer 的整体结构。

全局时空图构建

时间特征向量的构建

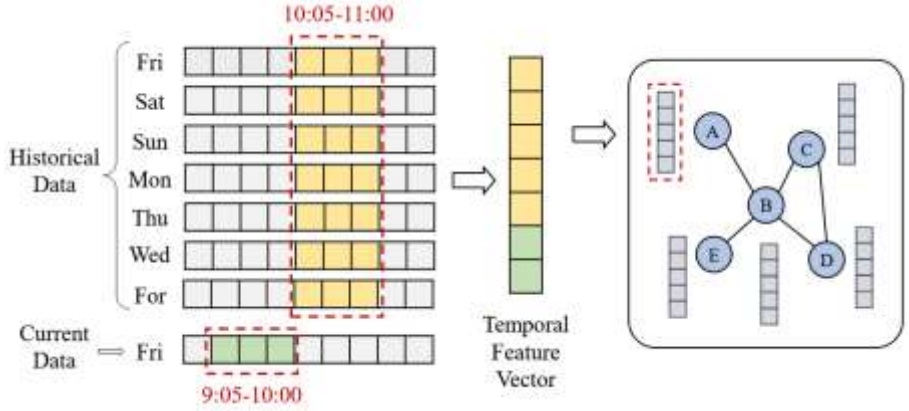


图 3：使用过去一周同一时间段的流量数据作为历史数据（图中黄色），将前一小时的流量数据作为当前数据（图中绿色）。这些历史数据和当前数据组合在一起形成节点的时间特征向量。

时间相似度图

N 个节点的时间特征向量表示为 $X_{\text{time}} = \{X^1_{\text{time}}, X^2_{\text{time}}, \ldots, X^N_{\text{time}}\} \in \mathbb{R}^{N \times T_{hc} \times d}$，其中 T_{hc} 是历史数据和当前数据中时间戳的总数，X^i_{time} 的第 j 个元素表示第 i 个传感器在第 j 个时间戳处的交通流状态。

使用动态时间扭曲（DTW）算法计算两个时间特征向量之间的相似性。对于每个节点，我们选择 k 个最相似的节点作为邻居节点，并通过在节点及其邻居之间分配 1 或 0 的权重来构建一个与时间相关的图 A_T，如下所示：

$$A_T[i,j] = \begin{cases} 1, & \text{if } v_j \text{ is a } k \text{ nearest neighbor of } v_i, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

在图 A_T 中，两个节点之间的边表示相似的城市功能和相应的交通模式。

DTW的核心思想是找到两个时间序列之间的最佳非线性对齐方式，从而计算相似性。它通过“扭曲”时间轴来补偿序列在时间速度上的变化，专注于比较形状而非严格按时间点一一对应。

全局时空图构建

时空相关图

需要将时间相似度图 A_T 和道路网络图 $\mathcal{G}$ 整合起来构建时空相关图 A_{ST} 。

具体方法：令 $\mathcal{G} = (N, \mathcal{E})$ 表示道路网络，$A_T = (N, \mathcal{E}_{\text{time}})$ 表示时间相关图。假设随机游走从传感器 N_0 开始，当前位于传感器 N_j 处，路径表示为 $\tau_j = \langle v_0, \ldots, v_{j-1}, v_j \rangle$，下一个要访问的传感器是 v_{j+1}。在这一点上，我们纳入了时间相关性：如果 A_T 中的节点 v_0 和 v_{j+1} 之间存在边，则表明时间相关性强，增加了访问 v_{j+1} 的概率。相反，没有边缘表明时间相关性较弱，从而降低了访问概率。这种抽样策略可以表示为：

$$P(v_{j+1}|\tau_j) \propto \begin{cases} \frac{1}{p}, & \text{if } d = 0 \text{ and } A_T[v_0, v_{j+1}] = 1, \\ 1, & \text{if } d = 1 \text{ and } A_T[v_0, v_{j+1}] = 1, \\ \frac{1}{q}, & \text{if } d = 2 \text{ and } A_T[v_0, v_{j+1}] = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

其中 d 表示 τj 以内的路网 R 中 v j+1 和 v j −1 之间的最短路径距离，p 和 q 是预设的非负常数。下一个传感器 v j+1 有三个选择：要返回传感器 v j −1，传感器 v j+1 和传感器 v j −1 之间的距离保持不变，或者传感器 v j+1 到 v j −1 的距离增加 1。因此，d 的值为 0、1 或 2。

可以得到所有节点的时空混合嵌入EST，既整合了路网的地理信息，又整合了时间相似图的时空相关性。

利用EST构建一个基于学习的时空相关图AST 。我们计算每对节点的嵌入之间的余弦相似度，并选择 K 个最相似的节点作为邻居节点。对于每个节点对，如果它们是相邻节点，则 AST （i， j） = 1，否则为 0。图中两个节点之间的边表示相似的流量模式和空间邻接性，从而有效地统一了单个图中的时空相关性。

全局时空图构建和损失函数

全局时空图

构建全局时空图，具体如下：

$$A_G = \begin{bmatrix} A & A_{ST} & A_{ST} & \cdots & A_{ST} \\ A_{ST} & A & A_{ST} & \cdots & A_{ST} \\ A_{ST} & A_{ST} & A & \cdots & A_{ST} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{ST} & A_{ST} & A_{ST} & \cdots & A \end{bmatrix} \in \mathbb{R}^{NT \times NT}. \quad (4)$$

如式（4）所示，全局时空图 A_G 由 $T \times T$ 子矩阵组成，每个子矩阵的大小为 $N \times N$。对角线子矩阵表示路网 G 的邻接矩阵 A，非对角线子矩阵 $A_G[ti, tj]$ 对应于学习到的时空相关图 A_{ST} 的邻接矩阵，捕获时间戳 ti 和 tj 之间每个节点特征的相互作用。

损失函数

选择均方误差（MSE）作为损失函数。目标函数如下图所示：

$$L(\hat{\mathbf{X}}^{(t+1):(t+T)}; \Theta) = \frac{1}{TN} \sum_{i=1}^{i=T} \sum_{j=1}^{j=N} (\hat{\mathbf{X}}_j^{(t+i)} - \mathbf{X}_j^{(t+i)})^2, \quad (18)$$

其中 Θ 是模型参数。

时空图的数据嵌入

◘ 时空混合嵌入

通过随机游走策略学习的序列结合了来自道路网络的地理数据和来自时间相似图的时间相关性，增强了时空关系的表示。因此，EST用于时空混合编码。为了平衡混合编码中历史信息和当前信息的贡献，我们引入了一个门控模块，该模块融合了混合编码 Xh（基于历史数据）和 Xc（专注于当前信息）：

$$E_{hybrid} = g(\Theta_1 * \mathcal{X}_h + a) \odot \sigma(\Theta_2 * \mathcal{X}_c + b). \quad (5)$$

其中，$E_{hybrid} \in \mathbb{R}^{N \times T \times D}$, Θ_1, Θ_2, a and b 是模型参数，⊙ 是 Hadamard 积，g（·） 是 tanh 函数，σ （·） 是 sigmoid函数。

◘ 时间嵌入

位置嵌入是通过组合不同频率的正弦和余弦函数生成的：

$$E_{(time,2i)} = \sin(time/10000^{2i/D})$$
$$E_{(time,2i+1)} = \cos(time/10000^{2i/D}) \quad (6)$$

其中 $E_{time} \in \mathbb{R}^{T \times D}$，D 是隐藏层的维度，time 是时间位置，T 是时间戳的数量。

时空图的数据嵌入

□ 空间嵌入

空间嵌入是对道路网络结构中空间信息的编码。信息基于路网图和节点的时序相似度提取，计算拓扑图。该拓扑图表示节点之间的空间位置信息，不随时间片的变化而变化。然后，通过矩阵分解，得到一个归一化图拉普拉斯矩阵，并将其对应的特征向量作为节点的空间嵌入，表示为Espace 。拉普拉斯矩阵和特征分解可以表示为：

$$\Delta = \mathrm{I} - D^{-1/2} A D^{-1/2} = U^T \Lambda U, \qquad (7)$$

其中 $U \in \mathbb{R}^{N \times D}$ 是对应于 Laplacian 矩阵的特征向量，表示为 Espace 。这里，A是路网的邻接矩阵，D是度矩阵，I是恒等矩阵，Λ表示拉普拉斯矩阵的特征值。

数据嵌入层的输出 X_{input} 是通过将三种不同形式的嵌入向量相加得到的。

$$X_{input} = E_{time} + E_{space} + E_{hybrid}, \qquad (8)$$

其中 $X_{input} \in \mathbb{R}^{B \times N \times T \times D}$ ，B 是在每个训练批次大小中选择的样本数。 X_{input} 将用作下面编码器-解码器结构的输入。

时空图编码器和解码器

时空图编码器

对于流量图中的全局信息，首先使用前面构建的全局时空图 A_G 获取不同注意力头的Q、K 和 V 矩阵：

$$Q_i = A_G W_i^Q, K_i = A_G W_i^K, V_i = A_G W_i^V \qquad (9)$$

其中 W_i^Q, W_i^K and $W_i^V \in \mathbb{R}^{NT \times d_i}$ 。di 是 Q 和 K 矩阵的维度。

编码器的特征表示源自全连接的自注意力机制，并与编码器的输入 X_{input} 串联形成 X_i 。然后对 X_i 进行蒸馏操作，涉及一维卷积，然后进行最大池化。这种池化作将每一层的输入序列长度减少一半，从而减小输出特征维度并为下一层生成串联注意力特征图。从i层到i+1层的蒸馏过程如下：

$$X_{i+1} = \text{MaxPooling}(\text{ELU}(\text{1D-Conv}([X_i]))) \qquad (13)$$

式中[·]是全连接自注意力层中的操作，ELU表示激活函数。通过蒸馏作，具有主导注意力的特征被赋予更高的权重，从而能够更有效地处理更长的输入序列。

时空图解码器

将解码器替换为简单的线性层。编码器采用自注意力机制来捕获交通数据中的时间和空间依赖关系，而线性层则直接将编码器的学习特征映射到预测序列。计算过程如下：

$$\mathrm{X}_{de} = \text{Linear}(\mathrm{X}_{input} + \mathrm{X}_{en}). \qquad (14)$$

其中 Xen 是编码器的输出。线性层实现了更简洁、更高效的计算，更适合长序列的时空图预测任务。

降低复杂性的策略

经典 Transformer 自注意力模块的复杂度为 $O(T^2)$。我们将全局时空图应用于 Transformer 模型，使复杂度为 $O(\bar{N}^2 \times \bar{T}^2)$。为了降低计算复杂度，提出了两种优化策略。

◘ 稀疏掩码矩阵策略

仅考虑序列中的元素子集来减少相互依赖关系的计算，从而从稀疏矩阵的角度减少计算负载。
在构建时空相关图时，我们选择 k 个最相似的节点作为每个节点的邻居，通过调整 k 来控制图的稀疏性。我们创建一个具有最优值 k′ 的新时空相关矩阵，作为掩码矩阵 W_{mask}。然后使用以下公式屏蔽注意力矩阵：

$$A_{\text{spare}} = W_{\text{mask}} \odot (Q \times K), \tag{15}$$

其中 $\odot$ 是 Hadamard 乘积。将掩码应用于全连接的自注意力矩阵后，我们得到了稀疏注意力矩阵 A_{spare}，显著降低了计算负载。此调整将计算复杂度降低到 $O(E \times T^2)$。

◘ Memsizer 优化策略

提出了一种新的机制来取代自注意力模块，实现递归推理计算。在这里，注意力分量计算为 $\alpha = f(QK^T)$ 和 $X_{\text{out}} = \alpha V$，其中 K 和 V 被认为是原始向量 Xs 的逐点投影。这三个矩阵可以用以下新形式表示：

$$\begin{aligned} Q &= X^t, \quad K = \Phi, \\ V &= \text{LN}(W_l(X_s)^T)\text{LN}(X_s W_r). \end{aligned} \tag{16}$$

K 是跨实例共享的可训练矩阵，而不是从输入的线性变换派生而来。该矩阵明显小于原始 Transformer 生成的 K，降低了注意力机制的计算复杂度。

实验结果

数据集。我们使用四个真实世界数据集评估所提出模型的性能：PEMS04、PEMS08、METR-LA 和 PEMSBAY。对于 METR-LA 和 PEMS-BAY 数据集，预测了未来 15、30 和 60 分钟的交通状况。对于 PEMS04 和 PEMS08 数据集，预测了未来一小时的交通状况。

Dataset	Models	15 min			30 min			60 min		
		MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)
METR-LA	ARIMA	3.99±0.12	8.21±0.16	9.60±0.10	5.15±0.22	10.45±0.25	12.70±0.10	6.90±0.00	13.23±0.32	17.40±0.15
	AGCRN	3.67±0.00	9.58±0.04	8.45±0.02	4.75±0.11	12.10±0.05	10.77±0.27	6.13±0.13	14.86±0.09	13.46±0.16
	ASTGCN	2.96±0.05	5.71±0.00	7.81±0.13	3.44±0.09	6.62±0.12	9.33±0.23	3.85±0.19	7.79±0.04	10.88±0.21
	STFGNN	3.21±0.04	6.52±0.02	8.14±0.09	3.51±1.13	6.62±0.09	9.77±1.10	3.86±0.09	7.65±0.15	10.89±0.13
	STSGCN	3.43±0.23	6.57±0.19	9.73±0.25	3.60±0.17	6.96±0.20	10.35±0.07	3.95±0.10	7.77±0.16	11.65±0.09
	Graph Wavenet	2.69±0.00	5.15±0.04	6.93±0.02	3.07±0.10	6.22±0.05	8.37±0.05	3.53±0.11	7.37±0.08	10.01±0.10
	Trafformer	2.78±0.05	5.35±0.02	7.32±0.04	3.05±0.08	6.18±0.05	8.67±0.10	3.41±0.10	7.17±0.13	9.96±0.11
	LLGformer	**2.42±0.00**	**4.91±0.03**	**6.64±0.02**	**3.01±0.11**	**6.02±0.09**	**8.14±0.13**	**3.15±0.00**	**6.89±0.03**	**9.38±0.01**
PEMS-BAY	ARIMA	1.82±0.08	3.30±0.11	3.50±0.06	2.33±0.23	4.76±0.19	5.40±0.15	3.38±0.32	6.51±0.28	8.34±0.19
	AGCRN	2.14±0.11	4.85±0.09	4.65±0.13	1.76±0.20	3.97±0.17	3.82±0.17	1.39±0.09	2.98±0.11	4.20±0.14
	ASTGCN	1.92±0.03	3.98±0.03	4.27±0.02	1.82±0.12	3.95±0.15	4.16±0.20	2.04±0.23	4.65±0.30	4.22±0.26
	STFGNN	2.25±0.17	4.35±0.20	5.41±0.19	2.42±0.31	4.25±0.26	5.88±0.15	2.54±0.21	4.89±0.24	5.71±0.19
	STSGCN	2.54±0.18	4.47±0.23	5.88±0.20	2.61±0.15	4.93±0.15	6.03±0.21	2.71±0.18	5.28±0.25	6.39±0.22
	Graph Wavenet	1.32±0.05	2.74±0.03	2.73±0.03	1.63±0.17	3.70±0.10	3.67±0.15	1.95±0.09	4.52±0.17	4.63±0.14
	Trafformer	1.88±0.20	4.38±0.19	4.59±0.24	1.61±0.06	3.74±0.03	3.82±0.03	1.31±0.11	2.83±0.14	2.92±0.13
	LLGformer	**1.16±0.07**	**2.68±0.10**	**2.54±0.05**	**1.59±0.13**	**3.26±0.09**	**3.42±0.09**	**1.22±0.02**	**2.61±0.00**	**2.77±0.06**

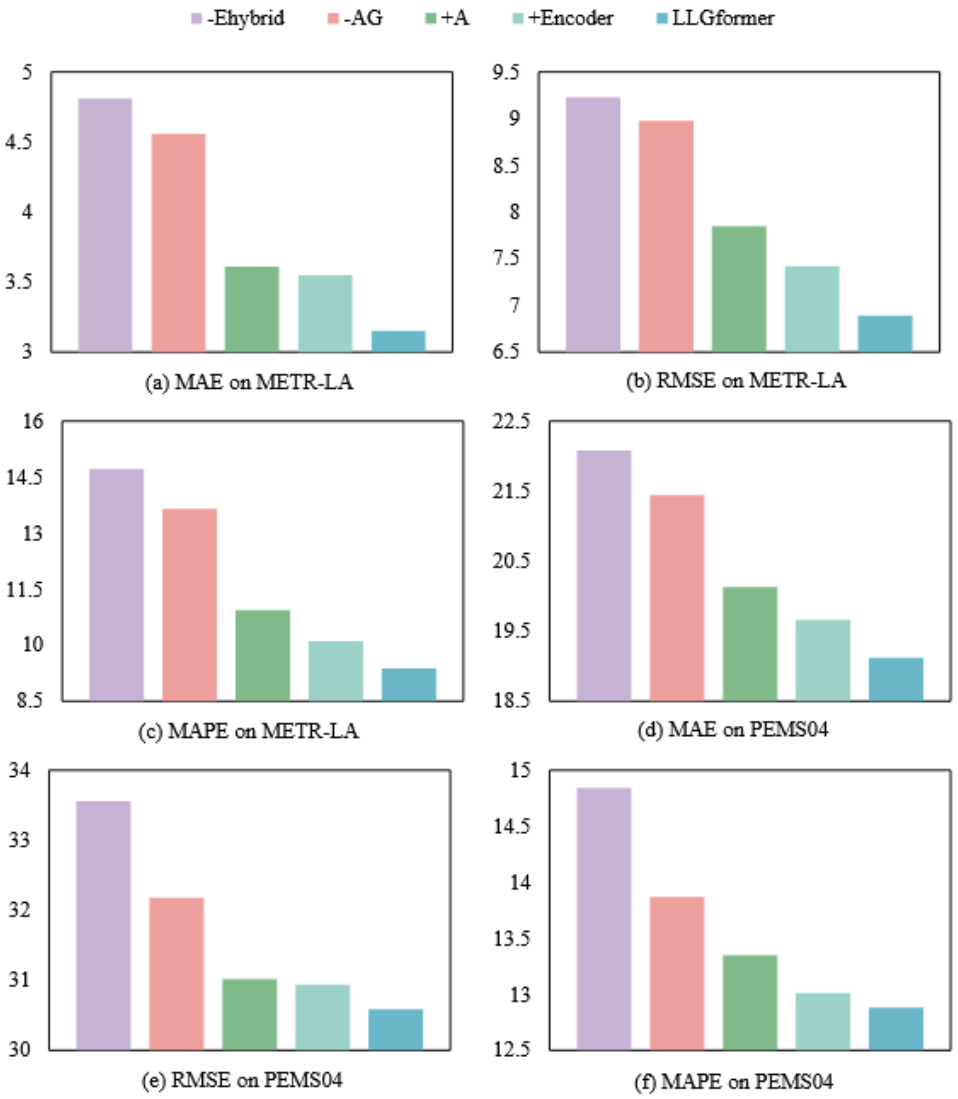
Models	PEMS04			PEMS08		
	MAE	RMSE	MAPE(%)	MAE	RMSE	MAPE(%)
ARIMA	23.71±0.11	36.88±0.17	17.61±0.14	19.02±0.09	29.88±0.14	13.35±0.14
AGCRN	19.74±0.09	32.01±0.03	12.98±0.08	15.92±0.14	25.31±0.08	10.30±0.11
ASTGCN	22.90±0.32	32.59±0.27	16.75±0.22	18.72±0.17	28.99±0.20	12.53±0.17
STFGNN	19.68±0.03	31.85±0.09	13.07±0.04	15.87±0.13	24.98±0.15	10.41±0.15
STSGCN	21.19±0.07	33.65±0.04	13.90±0.05	17.13±0.12	26.80±0.12	10.96±0.07
Graph Wavenet	19.91±0.21	31.06±0.24	13.62±0.18	15.57±0.11	24.32±0.09	10.32±0.12
Trafformer	19.26±0.21	30.67±0.25	12.96±0.15	15.27±0.09	24.33±0.15	10.19±0.20
LLGformer	**19.12±0.09**	**30.59±0.09**	**12.88±0.14**	**15.16±0.06**	**24.21±0.15**	**10.08±0.08**

METR-LA数据集上的时间消耗

Models	MAE	Training	inference
LLGformer	3.15	877.43	58.96
LLGformer-M	4.09	310.24	13.58
LLGformer-S-ST	3.64	586.97	26.88
LLGformer-S-A	4.16	658.73	29.82

评估指标为MAE。M表示Memsizer优化策略，S表示稀疏掩码矩阵策略，ST表示使用时空相关矩阵作为掩码矩阵，A表示使用邻接矩阵作为掩码矩阵。

消融研究



为了评估学习到的时空融合图，创建了两个变体：一种使用传统的注意力机制，表示为“−AGST”，另一种使用连接道路网络的邻接矩阵形成全局时空图，表示为“+A”。删除了混合编码，仅依赖于传统的时间和空间编码，称为“−Ehybrid”。解码器的迭代输出结构不使用线性层，表示为“+编码器”。

结论

本文提出了一种基于可学习远程图的用于时空交通预测的新型LLGformer模型。

- 设计了一种新的图嵌入方法和编码方案来学习每个传感器的数据表示，使模型能够通过历史信息捕获交通数据中的周期性模式。
- 提出了一种简单而有效的方法来捕获输入序列之间的长程依赖关系，并学习特征和预测序列之间的映射。
- 此外，为了降低模型的计算复杂度，引入了LLGformer模型的两种变体，以提高训练效率。
- 在四个真实数据集上的实验证实了LLGformer模型在交通预测任务中的有效性。

北京工業大學

谢 谢！