南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

本科生毕业设计（论文）

题　　目：基于 Transformer 的多变量时序列深度学习模型

姓　　名：黄文杰

学　　号：11812314

系　　别：计算机科学与工程系

专　　业：计算机科学与技术

指导教师：宋　轩

2023 年　6 月　2　日

诚信承诺书

1.本人郑重承诺所呈交的毕业设计（论文），是在导师的指导下，独立进行研究工作所取得的成果，所有数据、图片资料均真实可靠。

2.除文中已经注明引用的内容外，本论文不包含任何其他人或集体已经发表或撰写过的作品或成果。对本论文的研究作出重要贡献的个人和集体，均已在文中以明确的方式标明。

3.本人承诺在毕业论文（设计）选题和研究内容过程中没有抄袭他人研究成果和伪造相关数据等行为。

4.在毕业论文（设计）中对侵犯任何方面知识产权的行为，由本人承担相应的法律责任。

作者签名：

_____年____月____日

基于transformer的多变量时序列深度学习模型

黄文杰

（计算机科学与工程系　指导教师：宋轩）

[摘要]：时间序列预测对于现代社会的生产生活具有非常重要的意义。无论是气象变化还是交通规划，准确的预测结果可以节省通行成本、优化生产建设，创造巨大的利益。然而，现实中的事件变化往往是长期而复杂的。以交通为例，由于交通流量的高度非线性和动态时空依赖性，使得及时、准确、长期的交通预测仍是一个巨大的挑战。近年来，随着深度学习技术的不断发展，越来越多的研究者开始尝试使用深度学习技术来解决时间序列预测问题。另一方面，自 2017 年 Transformer 模型被 Google 团队提出以来，各种利用其自注意力优秀性质的新模型不断被提出。在此基础上，本毕业设计以 Transformer 多头自注意力结构为基础，重构编码嵌入部分，从原本的单一特征提取改为多变量多特征共同输入参与训练，构建出名为 Embedded Transformer 的新模型。实验结果表明，该模型在交通流量预测上取得了优于现有基线的成果，证明了多变量编码结构的有效性与可行性。

[关键词]：深度学习；多变量；交通流量预测；Transformer

[ABSTRACT]: Time series prediction is of great significance to daily life and industrial production in modern society. Whether it is weather change or traffic planning, accurate prediction results can save traffic costs, optimize construction, and create huge benefits. However, real events are often long and complex. Taking traffic as an example, timely, accurate and long-term traffic prediction is still a huge challenge due to the highly nonlinear and dynamic spatio-temporal dependence of traffic flow. In recent years, with the continuous development of deep learning technology, more and more researchers try to use deep learning technology to solve the problem of time series prediction. On the other hand, since the Transformer model was proposed by the Google team in 2017, various new models have been proposed to take advantage of the excellent nature of its self-attention structure. This graduation design takes Transformer self-attention structure as the basis, reconstructs the coding embedding part, changes from the original single feature extraction to multi-variable and multi-feature input to participate in training, and builds a new model known as Embedded Transformer. The experimental results show that the model is better than the existing baseline in traffic flow prediction, which proves the validity and feasibility of the multi-variable encoding structure.

[Keywords]: deep learning; multi-variables; Traffic flow prediction; Transformer

目录

1. 引言

1.1 研究背景

随着城市化进程的加速和人口的不断增长，交通拥堵问题已成为城市发展中的重要问题之一。交通拥堵不仅会影响人们的出行效率和生活质量，还会对城市的经济发展和环境保护产生负面影响。因此，交通流量预测成为解决交通拥堵问题的重要手段之一。

预测未来交通流量的任务可以被视作时间序列预测问题的一种。时间序列预测，具体指根据已获得的按时间顺序排列的数据，分析其特性和变化方向，总结演化趋势，对未来一定时间后的数据进行预测。在现代社会，时间序列预测被广泛用于气候变化、经济发展、交通预测等领域。准确的预测结果可以节省通行成本，优化生产建设，创造巨大的利益。在城市交通领域，时间序列预测可以帮助我们预测未来的交通流量，从而优化交通规划和管理，减少交通拥堵，提高城市的交通效率和生活质量。

需要注意的是，影响交通流量的因素复杂多样，交通流的高度非线性和动态时空依赖性也使得传统线性时间序列模型（如 ARIMA[1]等）即便花费很大代价也很难做出长期准确的交通预测。既有的模型往往忽视了交通流量的时间、空间相关性，并且对于整个路网上交通流量互相影响的因素考虑不足。为了提高长期预测的准确率，我们需要构建新的模型，新模型可以关注全局，并且同时注重时间与空间的影响。

近年来，深度学习技术在时间序列预测领域取得了很大进展。其中，Transformer 模型[2]是一种基于自注意力机制的深度学习模型。它可以并行化训练并且拥有全局信息，相比传统的图卷积神经网络和时空图卷积网络更具优势。后来的工作表明基于 Transformer 的预训练模型(PTMs)在各种任务上实现了最先进的状态。因此，基于 Transformer 的多变量交通流量预测模型具有很大的研究价值和应用前景

1.2 国内外研究现状

国内外学者在交通流量预测方面进行了大量的研究。传统的时间序列预测方法包括 ARIMA、VAR[3]、SARIMA[4]等，这些方法在一定程度上可以预测交通流

量的趋势和变化。但是，这些方法往往需要手动选择模型参数，且对于非线性和动态时空依赖性的交通流量预测效果不佳。

近年来，深度学习技术在交通流量预测领域得到了广泛应用。学者们提出了各种基于深度学习的交通流量预测模型，如基于循环神经网络（RNN）[5]的模型、基于卷积神经网络（CNN）[6] 的模型、基于长短时记忆网络（LSTM）[7]的模型等。这些模型在一定程度上提高了交通流量预测的准确性和效率，但是仍然存在一些问题，如模型训练时间长、模型参数过多等。

2018 年提出的扩散卷积循环神经网络（DCRNN）[8]，将交通流量视作有向图上的扩散问题。在此之前的交通预测模型，往往只能处理欧式空间数据或是无向图类型的数据，而 DCRNN 将路网上的传感器作为节点，传感器间的距离换算成权重作为边，依靠扩散性对交通流量进行动态的建模。DCRNN 使用扩散卷积，随机游走学习空间表示，并使用循环神经网络捕捉时间依赖性，取得了可观的提升效果。

Dejiang Kong 和 Fei Wu 还提出了分层时空长短时记忆网络（HST-LSTM）模型[9]。传统的 LSTM 通过细胞状态记忆信息，解决了一般的循环神经网络（RNN）无法处理长期依赖的问题，而 HST-LSTM 将时空要素加入传统的 LSTM 模型中，使用编码器和解码器对用户的历史到访信息进行建模，预测未来用户的运动状态、可能的位置，并在真实轨迹数据集上得到了验证。

2017 年北大的 Bing Yu 等人提出了时空图卷积网络（STGCN）模型[10]。基于 RNN 的迭代卷积方法，随着数据量增加，训练时间会大大延长并且产生梯度消失，错误累计的问题。STGCN 首次将图卷积应用在交通预测问题上，内部由多个时空卷积块组成，时空卷积块内部分别由空间图卷积和时间图卷积来捕获时空依赖。最终 STGCN 用更短的训练时间、更少的参数实现了更好的效果。

近年来 Transformer 在机器学习领域非常热门，国内外很多学者都投身于此，并提出了各具特点的不同 Transformer 变种。去年初，微软亚洲研究院提出的 Graphormer 模型[11]获得了 KDD Cup 的图预测赛道的冠军桂冠，其基于 Transformer，将其强大的表达能力引入了图结构数据中；而在 20 年，Mingxing Xu, Wenrui Dai 等人提出了用于时空序列预测的 STTN 模型[12]，同时捕捉时空信息，结合训练；Haoyi Zhou，Shanghang Zhang 等人提出的 Informer 模型[13]，可以处

理极长输入，解决了 Transformer 不适用长序列时间序列预测(LSTF)的问题。Cai L, Janowicz K, Mai G 等人提出的 Traffic Transformer 模型[14]，则是最早的被应用到交通流量预测的 Transformer 变种。在同样的领域，Zheng C, Fan X, Wang C 等人 20 年提出的 GMAN 模型[15]在当时取得了优秀的效果，本次毕业设计的实验部分将会把 GMAN 的训练结果放在表中一同比较。Zerveas G, Jayaraman S, Patel D 等人在 21 年提出的文章，《A Transformer-based Framework for Multivariate Time Series Representation Learning》[16]中对于多变量时序问题的思考则对本毕业设计的灵感有所启发。

1.3 研究目的和内容

本文旨在研究基于 Transformer 的多变量交通流量预测模型，以提高交通流量预测的准确性和效率。具体研究内容包括：

1. 以 Transformer 为基础，设计多变量交通流量预测模型
2. 对模型结构进行训练和优化
3. 对模型进行评估和分析，包括模型的准确性、效率等
4. 对研究结果进行总结和展望，包括模型的优缺点、未来的研究方向等。

通过本文的研究，可以为交通流量预测提供一种新的思路和方法，为城市交通规划和管理提供有力的支持

2. 相关技术介绍

2.1 传统的深度学习方法

时间序列预测是指通过对历史数据的分析和建模，预测未来一定时间段内的数据趋势、周期和规律。在时间序列预测中，传统的方法包括自回归模型（AR）[17]、移动平均模型（MA）[18]、自回归滑动平均模型（ARMA）[19]和自回归积分滑动平均模型（ARIMA）等。这些模型都利用了时间序列数据的自相关性和平稳性，但是往往只能模拟线性或平稳的时间序列，对于非线性或非平稳的序列预测效果不太理想。

为了解决传统时序预测模型的问题，深度学习方法得到了广泛应用。深度学习是一种基于人工神经网络构建的机器学习方法，具有自主学习和特征提取的能力，可以在处理结构化和非结构化数据上取得出色的成果。目前较为流行的用于

时序列预测的深度学习模型，包括卷积神经网络（Convolutional Neural Networks，CNN）、循环神经网络（Recurrent Neural Networks，RNN）、图神经网络（Graph Neural Networks, GNN）等。

CNN 很早被应用于时间序列预测任务。CNN 可以通过卷积层提取时间序列中的局部特征来发现序列中的模式和趋势。同时，汇聚层可将时间序列降采样，减少噪声和维度，最后通过全连接层实现输出。RNN 是针对按时间顺序排列的事件数据设计的深度学习方法，可以将历史数据和当前的输入共同编码为一个中间状态，以实现对序列整体特征的提取。另外，RNN 还引入了记忆单元，可以捕获时间序列中的长程关系和依赖关系，防止单项信息流与难以捕捉时序信息限制模型性能。GNN 是一种基于图结构数据的深度学习模型，它将节点和边连接的拓扑结构作为输入，通过高效的信息交互和聚合来实现特征的推断和更新，从而简化了对复杂非线性系统的建模和预测。GNN 在处理有关复杂非线性系统建模和预测问题方面，也有着很好的表现。此外，不止一次有人指出 GNN 结构与 Transformer 的相似性。

以上的各种模型结构，虽然也曾在时序列预测领域取得不错的效果，但仍然存在各种问题，比如基于 CNN 结构的模型需要针对特定数据；基于 RNN 结构的模型训练时间过长，梯度爆炸难以忽略等。时间序列预测需要更加强大，更加准确的深度学习结构。它就是 Transformer。

2.2 Transformer

Transformer 是一种基于自注意力机制的神经网络模型，由 Google 团队在 2017 年提出。Transformer 模型一开始就在自然语言处理（NLP）领域取得了巨大的成功，现已被广泛应用于计算机视觉(CV)和语音处理等领域。后来的工作表明，基于 Transformer 的预训练模型可以在各种任务上实现最先进的状态。因此，Transformer 已经成为 NLP 中的首选体系结构。自 Transformer 提出以来，其拥有的自注意力（Self-Attention）机制，可以并行化训练，处理长序列数据，并且关注全局信息的能力就为学界所关注，并被利用在各种领域，包括时序列预测方面。

相比于其他神经网络模型，Transformer 利用自注意力处理信息并将其聚合到全局信息。自注意力机制是一种强大的序列建模技术，它允许我们将输入序列

中任意两个位置之间的关注程度建模为权重，以便模型根据输入序列中不同位置的表征来执行下游任务。如今，这种技术已被广泛应用于计算机视觉、自然语言处理等领域中。

Transformer 的自注意力机制利用了序列中每个元素的信息，在不同位置上对输入序列进行编码。简单来说，它将每个序列元素表示为一组特征向量，并通过计算相似度矩阵来确定该元素与序列上其他元素之间的关系。

在计算相似度矩阵时，自注意力机制利用了三个函数映射：查询、键和值，对每个位置上的输入向量进行线性变换产生这三个向量。其中查询是用于确定需要关注哪些元素的向量，键和值用于表示序列中的所有元素。通过对它们进行点积打分和归一化操作，可以计算出对于一个特定的查询向量最相关的位置的概率分布。这些位置的值向量的加权和则用于生成最终的输出向量，表示我们关注序列中的哪些元素。通过计算查询向量和所有元素的相关性得分，我们可以得到每个元素的权重，从而为每个元素分配其在处理过程中的权重。最终输出的是一个加权求和的向量，其中每个元素的权重由相应的关注分布确定。这使得每个序列元素都能够得到适当的注意力，进而更好地处理序列上下文信息。

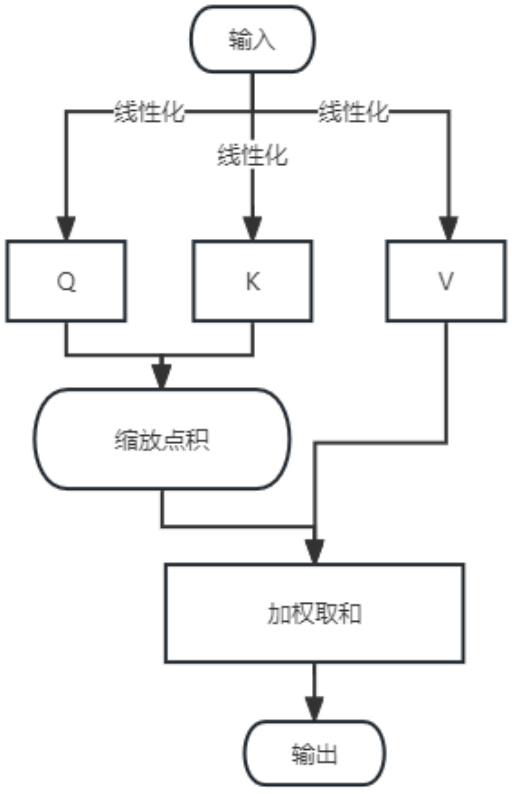

图 1　Attention 结构示意图

自注意力机制作为一种灵活且有效的序列建模技术，具有巨大的开发价值和潜力，目前已经被成功地应用于各种深度学习相关的任务中。

在实际使用自注意力时，为了提升训练效果，并且降低串行使用注意力结构造成的长距离依赖问题，会使用多头注意力（Multi-Head Attention）机制代替单个的注意力头。多头注意力可以看作是对输入序列的“多角度观察”，每个头能够产生一种不同视角的特征表示，类似于同时观察序列中不同区域的特点。

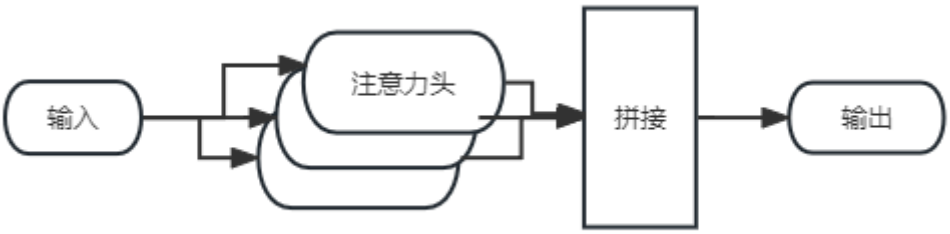

图 2　多头注意力结构示意图

多头注意力中，输入向量首先被划分为多个头（Head），每个头均利用自注意力机制来获取不同的特征表示，再将这些表示拼接起来组成一个更具表达能力和健壮性的向量。形象地来说，就是把图 1 的中间过程同时进行多次，最后的结果拼接为最终输出（如图 2）。通过在不同的子空间上并行地汇总来自各个头的信息，模型能够有效地捕捉到输入序列中的不同层次的关联性，从而提高了模型的泛化能力和精度。

3. 模型

3.1 问题描述

在介绍模型的具体架构之前，我们首先界定将要研究的问题。

在交通流量预测领域，通常将交通网抽象表示为一个无向连通图$G = (V, E, A)$。其中，G 表示交通路网，V 表示总数为 N 的道路传感器的节点集合，E 表示传感器间物理连接的边集合，A 是一个通过高斯内核构造的 N*N 大小的邻接矩阵，用来表示传感器间的欧几里得距离。传感器以固定的步长（时间间隔）记录交通速度，我们可以近似的将其看作所在道路交通流量的大小：速度越快，道路越通畅，反之亦然。针对某一特定路网 G 和某一特定时刻 T，我们有 N 个传感器在过去 M 个步长的历史速度数据$v^{T-M+1}, v^{T-M+2}, \cdots\cdots, v^{T}$，我们希望得到

一个预测模型 F，可以尽量准确的预测出接下来 P 个步长后的速度信息。用公式表达的话，就是

$$F(G; v^{T-M+1}, v^{T-M+2}, \cdots\cdots, v^{T}) = v^{T+1}, v^{T+2}, \cdots\cdots, v^{T+P}$$

得到一个效果优良的 F，就是既往交通流量预测领域的共同追求了。深度学习领域的发展日新月异，优秀的交通流量预测模型也层出不穷，较新的同样基于 Transformer 的 Graph WaveNet[20]模型和 STTN 模型已经可以做到同时灵敏地捕捉到时间与空间依赖关系，得到出色的长期预测效果，但它们仍然存在一些缺憾，模型训练时考虑的要素集中于时序变化和路网本身，而忽视了现实生活中，同样可能对交通流量产生整体影响的因素。本毕业设计正是基于这个思路出发，探究可能对预测效果产生影响的要素，增加输入变量的维度，最终获得一个优于现有效果的新模型。

3.2 模型设计

在这个部分，我们将详细介绍一个基于 Transformer 的模型，它采用多变量多通道共同参与的方式来增强预测效果，在本研究里将其称之为 Embedded Transformer。如图 3 示，Embedded Transformer 模型的总体架构可以分做两个模块：多变量位置编码、编码器和解码器。下面将分别介绍这两个模块。

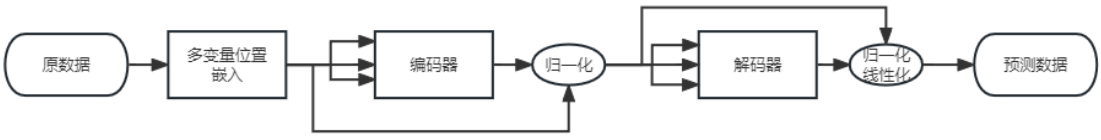

图 3 Embedded Transformer 模型整体架构图

3.2.1 多变量位置嵌入

位置嵌入对于所有基于 Transformer 的模型都很重要，因为它可以帮助模型学习到序列中元素之间的相对距离和顺序信息，从而更好地理解输入序列的结构和含义。在过去的循环神经网络和卷积神经网络等传统模型中，输入数据的顺序是由时间或空间自然顺序决定的，这种自然顺序可以通过序列索引的方式来体现。但对于采用了自注意力机制的 Transformer 来说，所有元素都可以同时运用，不存在任何时序性、空间关系的先后顺序，这就使得模型难以处理输入数据的位置

信息。

为了解决这一问题，Transformer 引入了两种位置嵌入向量，即绝对位置嵌入向量和相对位置嵌入向量。绝对位置嵌入向量是传统的位置编码方法，它根据输入序列中每个元素在位置上的编号生成一个固定的编码向量；而相对位置嵌入向量则根据输入序列中不同元素之间的相对距离生成不同的编码向量。通过将这两种位置嵌入向量相加得到的完整位置编码向量，Transformer 模型就能够有效地利用输入序列的位置信息，并将其纳入模型中进行训练和优化。

本模型主要的创新点就在于这一步，即在通常的位置编码外，增加了三个编码信息，即 day in week（DIW）编码，time in day（TID）编码和 node 编码。

参考图 4 训练数据刚输入时，暂且不考虑 batch，可以看作维度为 N*T 的二维数据（N 为节点数，T 为历史步数，即空间维度和时间维度），通过扩围操作将形状变成 N*T*C（C 代表 Channel，输入的变量种类数量），此时的 C=1，不包含实际信息。随后，相同的数据并行执行 4 种不同的嵌入，获得新的 N*T*C 形状的数据，此时的 C 维中就包含了嵌入后的编码。随后 4 种数据合并，新的数据形状边为 N*T*C，C=4。处理完毕，新数据进入注意力层，正式开始训练。

图中的 input embedding，就是 Transformer 类型本身的位置编码，用于提示数据的前后信息。一般的 NLP 任务或是 CV 任务，拥有位置编码已经足够良好地执行任务了，但交通流量预测作为时序列预测问题中的一种，有着其复杂性和特殊性存在，针对其特点优化往往可以取得事半功倍的效果。在这个思路下，本研究在通常的位置嵌入外额外添加了三种针对真实交通状况的嵌入。

图中第二行和第三行的 DIW 嵌入和 TID 嵌入类似，代表的是现实中时间的周期性。如果只有位置嵌入，模型只会考虑短期的时间相关，六时十分至六时四十分的三十分钟，和十二时二十五分至十二时五十五分的三十分钟，对于模型来说并没有什么区别，如果这两段时间的输入数据相似就会同样给出相似的预测结果。但在实际生活中，我们知道交通流量具有明显的周期性，周三的早八时和周四的早八时很大概率具有相似的路况；同样的，本周的周五和下周的周五，一天中的路况也可以期望有相近的变化。考虑了这些后，便有了两个周期性的数据嵌入，二者的工作类似，在短期的时间顺序数据外，加入分别代表一天中时刻和一周中天数的时间戳，给模型额外的提示信息。之后的实验可以证明，新的嵌入对

于预测效果有着显著提升。

第四行的 node 嵌入，代表的则是传感器节点间，更抽象的说是道路彼此间的空间关系。拥有之前的三个嵌入后，模型已经可以很好地捕获数据的时间依赖关系，但对于空间依赖的关注仍然不够。现实中的交通路网是高度相关的：如果一条道路前方堵车，后方的速度也不会快；熟悉道路的司机会在高峰期转到相近的道路；如果所有道路都比较畅通，车辆就会倾向于走最短的路径。用于训练的数据集并没有单独的道路信息，但可以用传感器来近似模拟真实路网。node 嵌入的工作，就是提取出传感器节点间的空间关系，经过泽维尔初始化为位置编码，嵌入数据中。

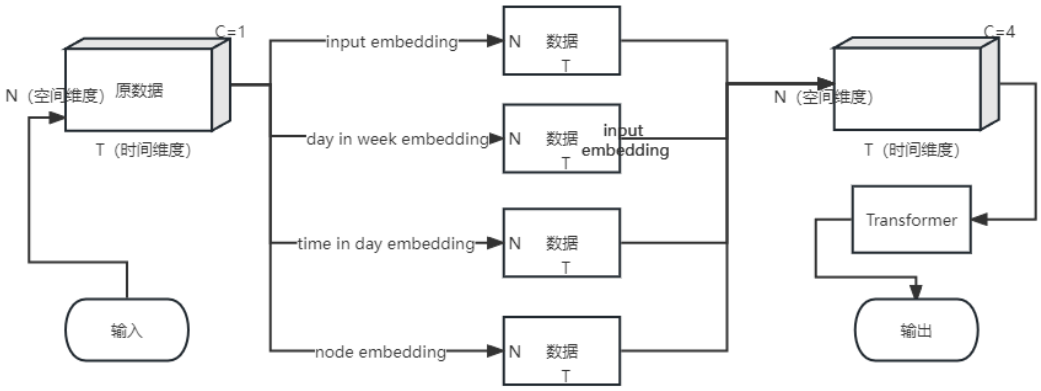

图 4 嵌入部分结构图

这种多通道的方法允许模型同时考虑多个嵌入器的贡献，并将不同类型的数据映射到不同通道上。通过这种方式，模型可以更好地捕捉不同特征之间的关系，进一步提高预测的准确性。

3.2.2 编码器和解码器

编码器由 N 个相同的层堆叠而成，其中每层包含两个子层：一个是多头自注意力机制，另一个是全连接前馈网络（Feed Forward Network）。多头自注意力的概念在 2.1 部分已经比较详细的介绍过了。本模型中，从上个部分多变量嵌入后得到的数据 X，在这里映射为 Q、K、V，进行点乘和加权求和操作。表达成公式的话，就是

$$softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

K^T 表示 K 的转置，d_k 表示 K 的维数。得到的结果，再输入全连接层中。全

连接层由两个线性变换和一个 ReLU 激活函数组成，是用于处理多头自注意单元输出向量的传统前馈神经网络。

为避免梯度消失或爆炸问题，在每个子层输出前再添加一个残差连接，并进行归一化。完成所有操作后，将编码器的输出传入解码器中。

解码器结构与编码器基本相同，都是由相同的 N 层堆叠而来。但有所不同的是，在解码器阶段需要正式对未来进行预测，这就需要区分过去，现在和将来的不同数据，以避免出现利用未来的数据预测当下的情况。Transformer 巧妙地使用了掩码机制，将读取到的解码器的输出屏蔽掉未来时刻的数据，再进入自注意力层提取特征。

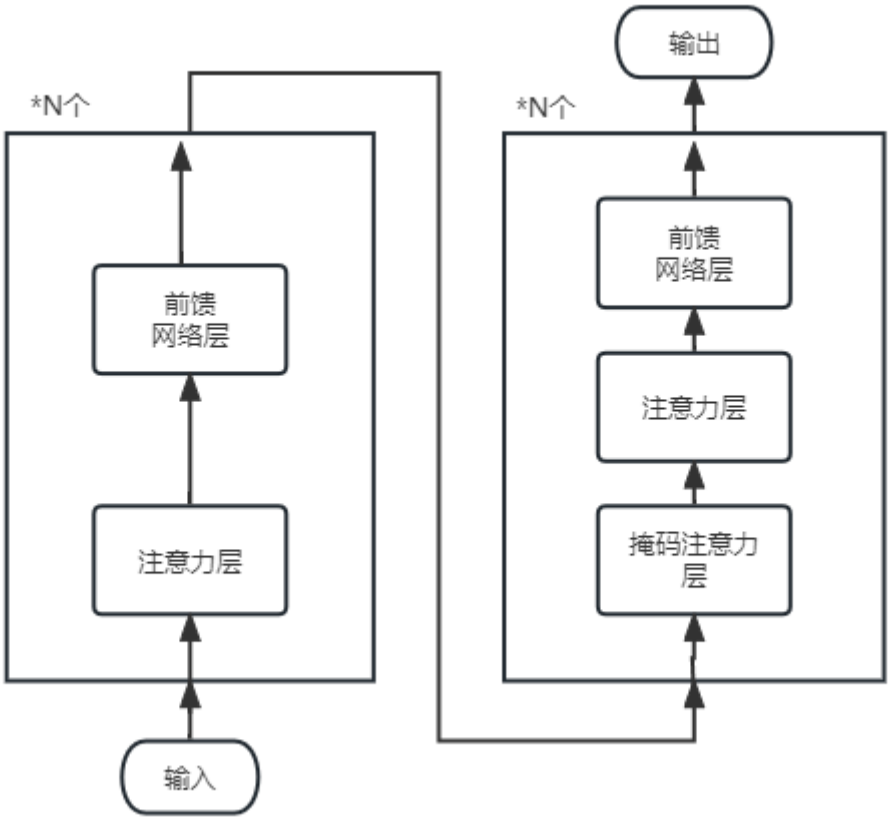

图 5 编码器和解码器（图中已省略归一化等部分）

Transformer 的编码器-解码器结构特点可以在交通流量预测领域大显身手。正如前面提到的，现实中的交通情况复杂多变，而 Transformer 可以并行化处理数据的特点正适用于数据量庞大的交通数据集。同时，Transformer 关注全局信息的能力可以很好地利用本模型在嵌入结构中提供的多变量编码，提取时间、空间特征以达到优秀的预测效果。

3.3 模型评估

为了评估模型的性能，本研究采用了均方根误差（RMSE），平均绝对误差（MAE）和平均绝对百分比误差（MAPE）作为评估指标。模型的训练过程中，采用了早停法，当模型在验证集上的损失函数连续 20 个 epoch 没有下降时，停止训练。为了直观评估模型效果，本研究除了预测效果表外还绘制了模型预测值与 Ground Truth 的拟合曲线图。

4. 实验

4.1 实验设置

本实验使用的数据集有两个，分别是 METR-LA 和 PEMS-BAY。METR-LA 和 PEMS-BAY 是两个常用的深度学习数据集，用于检验和比较交通流量预测模型的效果和准确率。METR-LA 数据集包含从洛杉矶市域 207 个传感器收集的前后四个月的交通流量信息，数据较小。PEMS-BAY 数据集则来自加利福尼亚运输部（California Transportation Agencies）的性能测量系统（Performance Measurement System, PeMS），包含了湾区交通网络上 325 个传感器前后六个月的交通流量数据，数据较大。两个数据集都采用了训练：测试：评估=7：2：1 的比例进行模型训练和效果评估。图 3 展现的是两个数据集的热力图，反应了不同传感器间的空间依赖关系。

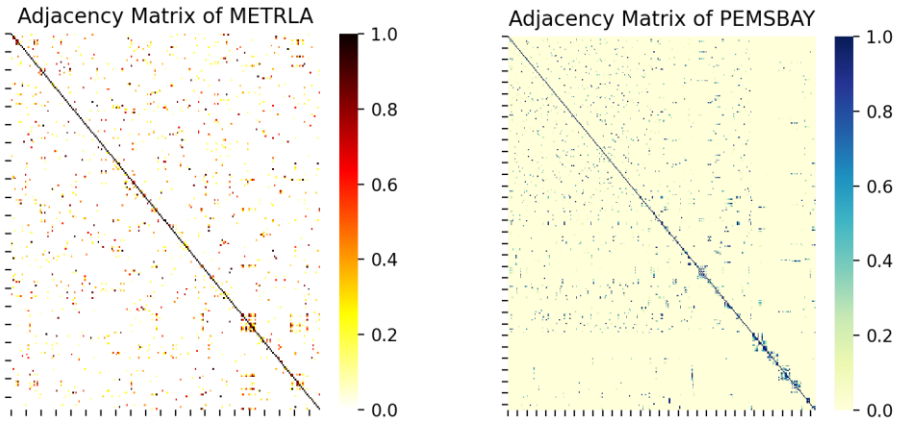

图 6　数据集热力图

本研究使用了 Python 编程语言和 PyTorch 深度学习框架来实现模型。实验在一台配备了 Intel(R) Xeon(R) Gold 6240 CPU @ 2.60GHz 和 GeForce RTX 2080

Ti GPU 的服务器上运行。模型的训练时间，METR-LA 约为 2 小时，PEMS-BAY 约为 4 小时。

模型的超参数设置如下。模型的训练过程中，采用了批量训练的方式，每个批次包含了 64 个样本。模型层数为 3，以 5 分钟为一步输入输出步长均为 12，头数为 4，dropout 率为 0.1，学习率为 0.001，优化器为 Adam。

4.2 实验结果

表 1-表 3 分别展示了不同的模型在 METR-LA 和 PEMS-BAY 两个数据集上，短期（15 分钟）、中期（30 分钟）和长期（60 分钟）的预测效果对比。选取的模型包括几个传统预测模型，以及较新锐的作为基线的 STTN 与 Graph WaveNet 模型，和我们的 Embeded Transformer。

表 1 的短期结果中可以看出，Embeded Transformer 在较小的数据集上远优于基线，在较大的数据集上与基线相差仿佛；而从表 2 的中期效果来看，Embeded Transformer 无疑的在小数据集上优于基线，而在大的数据集上也取得了略好以往的效果；观察表 3 的长期效果，这个优势继续保持了下来。

总体来说，在数据量较小的情况下，无论是短期预测还是长期预测 Embeded Transformer 都取得了极佳的成绩。而面对更复杂，更庞大的数据集时，Embeded Transformer 仍然可以在长期预测上得到更让人满意的结果。

表 1　3 步/15 分钟短期预测效果表

数据集	模型	RMSE	MAE	MAPE
METR-LA	HistoricalAverage	14.73744	11.01257	23.34%
	CopyLastSteps	14.21547	6.79927	16.73%
	GMAN	9.54627	4.20179	11.67%
	STTN	7.64516	3.29858	8.27%
	Graph WaveNet	7.44325	3.21183	7.85%
	Embeded Transformer	5.54743	2.81850	7.47%
PEMS-BAY	HistoricalAverage	6.68711	3.33339	8.10%
	CopyLastSteps	7.02243	3.05215	6.84%
	GMAN	4.14808	1.78238	4.35%
	STTN	2.85262	1.34919	2.87%

	Graph WaveNet	2.76617	1.32569	2.81%
	Embeded Transformer	2.89708	1.35488	2.85%

表 2　6 步/30 分钟中期预测效果表

数据集	模型	RMSE	MAE	MAPE
	HistoricalAverage	14.73686	11.01017	23.34%
	CopyLastSteps	14.21443	6.79874	16.73%
METR-LA	GMAN	10.52446	4.59798	12.56%
	STTN	9.56039	4.04627	10.33%
	Graph WaveNet	9.42638	3.94608	9.87%
	Embeded Transformer	6.31877	3.10897	8.66%
	HistoricalAverage	6.68584	3.33254	8.10%
	CopyLastSteps	7.01596	3.04928	6.84%
PEMS	GMAN	4.11041	1.78295	4.31%
-BAY	STTN	3.84405	1.69725	3.85%
	Graph WaveNet	3.75055	1.66450	3.80%
	Embeded Transformer	3.75193	1.65229	3.68%

表 3　12 步/60 分钟长期预测效果表

数据集	模型	RMSE	MAE	MAPE
	HistoricalAverage	14.73564	11.00500	23.33%
	CopyLastSteps	14.21361	6.79803	16.72%
METR-LA	GMAN	12.32955	5.54962	14.89%
	STTN	11.52183	4.95472	12.86%
	Graph WaveNet	11.48736	4.85927	12.44%
	Embeded Transformer	7.32344	3.46039	9.98%
	HistoricalAverage	6.68527	3.33186	8.10%
	CopyLastSteps	7.00541	3.04445	6.83%
PEMS	GMAN	5.02758	2.17516	5.24%

-BAY	STTN	4. 66030	2. 03159	4. 77%
	Graph WaveNet	4. 51695	1. 97937	4. 73%
	Embeded Transformer	4. 45585	1. 93960	4. 48%

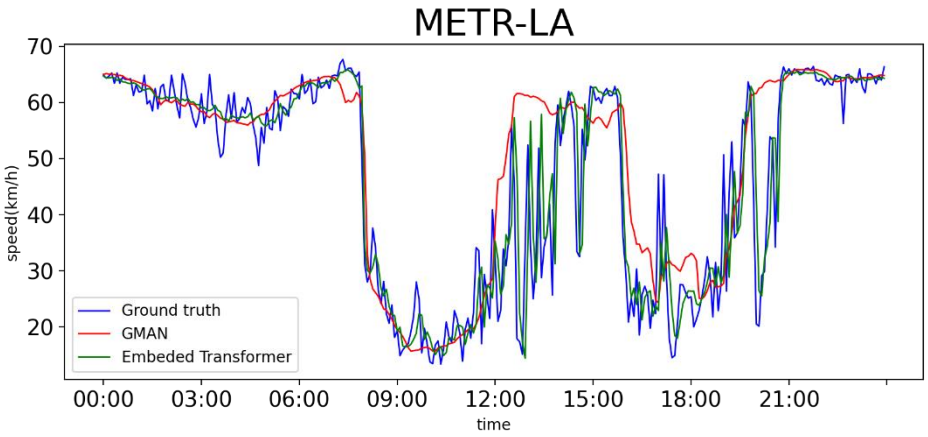

图 7 METR-LA 上的预测效果拟合曲线

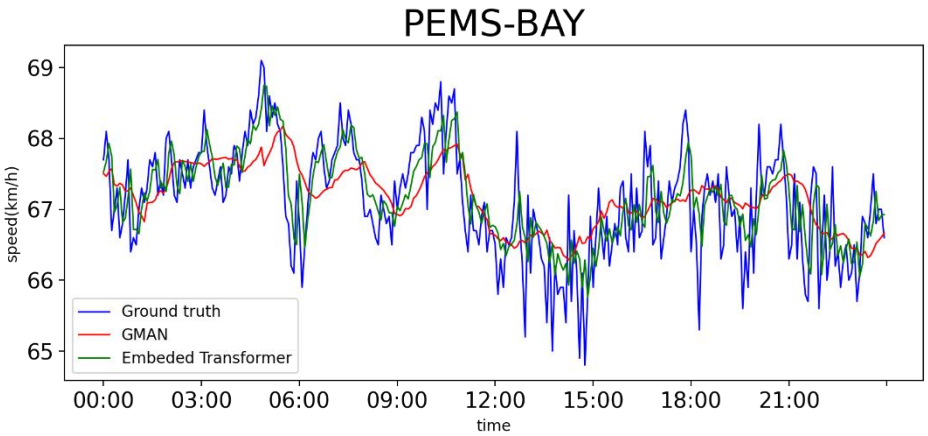

图 8 PEMS-BAY 上的预测效果拟合曲线

图 7 和图 8 截取了 Ground Truth、GMAN 和 Embeded Transformer 在两个数据集上某段时间的 24 小时数据曲线。可以看出，传统的 GMAN 模型拟合效果一般，而 Embeded Transformer 在两个数据集上都拟合的很好。

除了不同模型的效果对比以印证 Embeded Transformer 的优势外，本研究还做了 Embeded Transformer 的消融实验。表 4-表 6 展现了不同结构在 METR-LA 数据集上短期、中期和长期效果对比。当删去所有本研究添加的嵌入结构时，我们就得到了一个单纯的 Transformer，并可以之为基线。随后，我们对表 1-表 3 结果表中的 Embeded Transformer 进行修改。首先删去 node embedding，我们就得到了 Embeded Transformer-3。随后在 Embeded Transformer-3 的基础上，颠倒了 DIW embedding 和 TID embedding 的顺序得到 Embeded Transformer-2；或者改

变原本并行输入的结构，把两个时间嵌入叠加串行输入，这样就得到了 Embeded Transformer-1。

从以下 3 个表中的效果来看，任何一个 Embeded Transformer 结构都优于没有改进的原 Transformer，说明了本研究起到的作用。另外，几种结构中 Embeded Transformer 与 Embeded Transformer-3 的效果差异最大，体现了提取空间依赖对于提升模型准确度的巨大意义。其他几种模型的对比则决定了本研究最终选用的模型结构。

表 4　不同结构对比实验（3 步/15 分钟）表

模型	RMSE	MAE	MAPE
Transformer	6.12478	3.10645	8.72%
Embeded Transformer-1	6.05218	3.11680	8.59%
Embeded Transformer-2	5.92700	3.00697	8.11%
Embeded Transformer-3	5.89486	2.99059	8.01%
Embeded Transformer	5.54743	2.81850	7.47%

表 5 不同结构对比实验（6 步/30 分钟）表

模型	RMSE	MAE	MAPE
Transformer	7.29443	3.60299	10.39%
Embeded Transformer-1	7.33146	3.65082	10.38%
Embeded Transformer-2	7.20166	3.56942	10.27%
Embeded Transformer-3	7.19388	3.56508	10.22%
Embeded Transformer	6.31877	3.10897	8.66%

表 6 不同结构对比实验（12 步/60 分钟）表

模型	RMSE	MAE	MAPE
Transformer	9.14297	4.53832	13.54%
Embeded Transformer-1	9.27428	4.63652	13.67%
Embeded Transformer-2	8.92037	4.40846	13.41%
Embeded Transformer-3	8.92332	4.39582	13.35%

Embeded Transformer	7.32344	3.46039	9.98%

4.3 结果分析

本研究所提出的 Embeded Transformer 模型在交通流量预测任务上取得了较好的预测效果，这主要得益于 Transformer 的自注意力机制和多变量共同参与训练的特点。具体来说，本研究所提出的多变量位置编码嵌入结构，比以往更好的关注到了不同变量对于交通流量产生的影响，而自注意力机制可以关注全局，良好的运用了额外的编码信息，出色地捕捉到了时间、空间依赖关系，最终取得了优于现有基线的预测效果。

5. 结论与展望

5.1 研究结论

本文基于 Transformer 模型，提出了一种多变量交通流量预测模型 Embeded Transformer，并在真实的交通数据集上进行了实验。实验结果表明，本文提出的模型在交通流量预测任务上具有优异的性能，相比传统的基线模型，Embeded Transformer 取得了更好的效果。同时，本文还探究了多变量共同参与训练的效果，结果表明，多变量共同参与训练可以提高模型的预测精度。

5.2 不足和展望

本研究还存在一些不足之处。首先，本研究只考虑了少量的输入变量，如一天和一周的时刻，节点间的空间关系等，还可以考虑其他因素，如节假日、天气等。其次，本文使用的是历史交通数据进行预测，对于突发事件等未知因素的影响，模型的预测能力仍有待提高。最后，本研究仅在两个交通数据集上进行了实验，还可以考虑使用更多的数据集来验证模型的泛化能力。

参考文献

[1] Newbold P. ARIMA model building and the time series analysis approach to forecasting[J]. Journal of forecasting, 1983, 2(1): 23-35.

[2] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[3] Zivot E, Wang J. Vector autoregressive models for multivariate time series[J]. Modeling financial time series with S-PLUS®, 2006: 385-429.

[4] Dabral P P, Murry M Z. Modelling and forecasting of rainfall time series using SARIMA[J]. Environmental Processes, 2017, 4(2): 399-419.

[5] Lv Z, Xu J, Zheng K, et al. Lc-rnn: A deep learning model for traffic speed prediction[C]//IJCAI. 2018, 2018: 27.

[6] Zhang W, Yu Y, Qi Y, et al. Short-term traffic flow prediction based on spatio-temporal analysis and CNN deep learning[J]. Transportmetrica A: Transport Science, 2019, 15(2): 1688-1711.

[7] Tian Y, Pan L. Predicting short-term traffic flow by long short-term memory recurrent neural network[C]//2015 IEEE international conference on smart city/SocialCom/SustainCom (SmartCity). IEEE, 2015: 153-158.

[8] Li Y, Yu R, Shahabi C, et al. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting[J]. arXiv preprint arXiv:1707.01926, 2017.

[9] Kong D, Wu F. HST-LSTM: A hierarchical spatial-temporal long-short term memory network for location prediction[C]//IJCAI. 2018, 18(7): 2341-2347.

[10] Yu B, Yin H, Zhu Z. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting[J]. arXiv preprint arXiv:1709.04875, 2017.

[11] Ying C, Cai T, Luo S, et al. Do transformers really perform badly for graph representation?[J]. Advances in Neural Information Processing Systems, 2021, 34: 28877-28888.

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

[13] Zhou, Haoyi, et al. "Informer: Beyond efficient transformer for long sequence time-series forecasting." Proceedings of the AAAI conference on artificial intelligence. Vol. 35. No. 12. 2021.

[14]Cai L, Janowicz K, Mai G, et al. Traffic transformer: Capturing the continuity and periodicity of time series for traffic forecasting[J]. Transactions in GIS, 2020, 24(3): 736-755.

[15]Zheng C, Fan X, Wang C, et al. Gman: A graph multi-attention network for traffic prediction[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(01): 1234-1241.

[16]Zerveas G, Jayaraman S, Patel D, et al. A transformer-based framework for multivariate time series representation learning[C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021: 2114-2124.

[17]Meek C, Chickering D M, Heckerman D. Autoregressive tree models for time-series analysis[C]//Proceedings of the 2002 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2002: 229-244.

[18]Durbin J. Efficient estimation of parameters in moving-average models[J]. Biometrika, 1959, 46(3/4): 306-316.

[19]McLeod A I, Li W K. Diagnostic checking ARMA time series models using squared-residual autocorrelations[J]. Journal of time series analysis, 1983, 4(4): 269-273.

[20]Wu Z, Pan S, Long G, et al. Graph wavenet for deep spatial-temporal graph modeling[J]. arXiv preprint arXiv:1906.00121, 2019.

致谢

在论文写作过程中，很多人对我给予了帮助和支持，在此对这些人表示衷心的感谢。

首先，感谢我的导师/指导教师（或其他批评和鼓励我们的人），是他们在整个论文写作过程中给予了我宝贵的指导和帮助。

其次，感谢实验室里所有的老师和同学，感谢他们提供了良好的研究环境和工作条件。与他们的讨论和交流使我获益匪浅。

还要感谢所有提供数据集/代码/设备等帮助的人，为论文做出了重要贡献。

最后，感谢我们的家人和朋友们一直以来对我们的支持和鼓励。没有他们的支持和理解，我们无法克服论文写作中遇到的困难。

总之，这篇论文的成功离不开所有人士的支持和帮助。再次表达我们的诚挚感谢，并希望我们的研究可以对相关领域做出有意义的贡献。