

基于自注意力机制的时间序列插补

徐磊, 曾艳, 袁俊峰*, 岳鲁鹏, 殷昱煜, 张纪林, 薛梅婷, 韩猛

(杭州电子科技大学计算机学院, 浙江 杭州 310000)

摘要: 船舶轨迹数据作为海上交通的核心数据, 可以用于轨迹预测、预警等任务, 具有非常明显的时序特征, 但海上环境恶劣、通信可靠性差等因素导致收集到的船舶轨迹数据普遍存在数据缺失的问题, 对含有缺失数据的时间序列进行学习会严重影响时间序列分析的准确性。当前主流的解决方案是对缺失数据进行近似插补, 主要基于卷积模型沿着时间轴对时间序列进行重塑, 捕捉时间序列的局部特征, 但对长时间序列的全局特征捕捉能力较弱。Transformer 模型通过其核心的自注意力机制来捕获时间序列各个时间点之间的关系, 从而增强模型对时间序列全局特征的捕捉能力, 但注意力机制是通过矩阵乘计算得到的, 导致其忽视了时间序列自身的时序性, 得到的全局特征权重不具有时间跨度依赖性。因此, 针对长时间序列全局特征捕捉的问题, 提出一种基于自注意力机制的变体网络 GANet。GANet 首先通过自注意力机制获得基础的时间序列点之间的全局特征权重矩阵, 再使用门控循环单元在时间轴上对全局特征权重矩阵进行遗忘与更新, 从而得到具有时间跨度依赖性的全局特征权重矩阵; 然后使用该矩阵进行数据重构, 对缺失数据进行插补。GANet 通过结合自注意力机制与门控机制实现了在捕捉全局特征的同时考虑时间跨度对各个时间点的影响, 使得捕获到的全局特征具有时间跨度依赖性。实验结果表明, 与现有 Autoformer、Informer 等模型相比, GANet 对于 Trajectory、ETT、Electricity 数据集具有较好的插补效果。

关键词: 自注意力机制; 门控循环单元; 全局特征捕捉; 时间跨度依赖性; 时间序列插补

中图分类号: TP391

文献标志码: A

DOI: 10.19678/j.issn.1000-3428.0069678

Time Series Imputation Based on Self-Attention Mechanism

XU Lei, ZENG Yan, YUAN Junfeng*, YUE Lupeng, YIN Yuyu, ZHANG Jilin, XUE Meiting, HAN Meng

(College of Computer Science, Hangzhou Dianzi University, Hangzhou 310000, Zhejiang, China)

【Abstract】 As core data for maritime traffic, ship trajectory data can be used for trajectory prediction, early warning, and other tasks with pronounced temporal characteristics. However, owing to factors such as harsh marine environments and poor communication reliability, missing ship trajectory data is a common problem. Learning from time series containing missing data can affect the accuracy of time series analysis significantly. The current mainstream solution is to approximate the imputation of missing data, mainly based on convolutional models, to reshape the time series along a timeline to capture its local features of the time series. However, the ability to capture the global features of long time series is limited. The Transformer enhances the ability of a model to capture the global features of a time series by capturing the relationships between various time points in the time series through its core self-attention mechanism. However, because its attention is calculated through matrix multiplication, it ignores the temporal nature of the time series, and the obtained global feature weights do not have a time span dependency. Therefore, to address the issue of capturing global features in long time series, this study proposes the GANet, a variant network based on the self-attention mechanism. GANet first obtains the basic global feature weight matrix from the time series points through the self-attention mechanism and then uses gated recurrent units to forget and update this global feature weight matrix on the timeline, thereby obtaining a global feature weight matrix with time span dependency, which is then used for data reconstruction to impute the missing data. GANet combines the self-attention mechanism and gating mechanism to capture global features while considering the impact of the time span on different time points, making the captured global feature time span dependent. Experimental results show that compared with existing models, such as Autoformer and Informer, GANet achieves better interpolation performance on Trajectory, ETT, and Electricity datasets.

【Key words】 self-attention mechanism; gated recurrent unit; global feature capture; time span dependency; time series imputation

收稿日期: 2024-04-01 修回日期: 2024-05-15

基金项目: 国家自然科学基金(62072146); 浙江省“高层次人才特殊支持计划”科技创新领军型人才项目(2022R52043); 浙江省重点研发计划(2023C03194, 2021C03187)。

通信作者 E-mail: * yuanjunfeng@hdu.edu.cn

0 引言

随着全球化的不断推进,海上航运作为我国重要的国际贸易运输方式之一,正面临着船舶数量不断增多所带来的海上交通和船舶安全问题。复杂多变的海上作业环境与各种突发的天气变故进一步影响了海上交通秩序并可能导致海事事故的发生,催生了对船舶轨迹预测的需求。船舶轨迹随时间变化而变化,数据呈现非常明显的时序特征。然而,由于恶劣的海上天气以及卫星信号丢失等原因,采集到的海上船舶轨迹数据普遍存在数据缺失问题,这会严重影响下游任务对船舶轨迹数据的分析准确性。数据缺失问题在其他领域也普遍存在^[1-4],解决数据缺失问题对于下游时间序列分析任务至关重要^[5-8]。数据缺失时间序列的数据插补是当前最有效的方式之一,其通过分析已观测值对时间序列构成模式进行学习,进而完成对缺失数据的重构。

目前,基于深度学习的时间序列插补方法已成为主流。得益于深度模型强大的非线性建模能力,这类方法能够有效捕捉时间序列中复杂的时间变化特征,因此研究者们已提出了多种基于深度学习的插补模型。CAO 等^[9]提出基于递归神经网络的 BRITS,其通过双向递归迭代估算时间序列中的所有变量。雷未等^[10]提出基于卷积神经网络的大坝监测缺失数据插补模型,其通过双向融合插补消除长序列缺失数据插补时间步的累积误差。HE 等^[11]提出时间卷积网络(TCN),其在时间线上对时间序列数据进行卷积,同样可以用于缺失数据重构。上述方法虽然很有效地提取了局部相邻时间点之间的信息,但是通常无法捕获时间跨度较大的时间点之间的长期相关性。随着 Transformer 模型的出现,其采用的自注意力机制有效解决了长时间序列中全局特征难以捕捉的问题。WU 等^[12]提出 Autoformer,其通过对分解后的时间序列进行编码解码来重构时间序列。KITAEV 等^[13]提出 Reformer,其通过局部敏感哈希来近似实现 Transformer 中自注意力机制的点积计算。由于 Transformer 的自注意力机制在时间序列中计算点之间的权重时是通过矩阵乘得到的,虽然捕获了全局特征,但因此忽略了时间序列自身的时序性,时间点之间的时间跨度依赖关系在复杂的时间模式中难以被学习。

为解决当前大多数模型对时间序列全局特征的捕捉问题以及基于 Transformer 的模型无法兼顾时间序列时序性的问题,本文提出一种基于自注意力

机制的变体网络 GANet,利用自注意力机制捕捉时间序列的全局特征的过程中加入门控循环单元,使得模型在捕捉时间序列全局特征的同时学习时间点之间的时间跨度依赖关系,并且在嵌入层中综合使用了 3 种嵌入方法来增强模型对时间跨度依赖关系的捕捉。本文的贡献和创新点如下:

1)构建一种基于自注意力机制与门控机制的时间序列学习模型,该模型用于解决时间序列数据缺失问题,兼具了对时间序列全局特征与时间跨度依赖关系的学习能力,有效解决了对缺失数据的插补问题。

2)采用多层独立的门控循环单元捕捉多样化的时序信息,并将捕捉到的时序信息作为新的权重矩阵,使得模型学习到的全局特征中所包含的时间跨度依赖性更加丰富。

3)在 7 个数据集上进行缺失值插补实验,结果证明了 GANet 对缺失值插补的有效性,并且进行消融实验,结果证明了门控循环单元在模型中起到的重要作用,同时采取不同层数的门控循环进行实验,结果进一步展示了门控循环单元对模型捕捉时间序列全局特征的影响。

1 相关工作

目前,时间序列插补方法已取得显著进展,主要可分为以下 3 类:

1)基于统计学习的方法。ANDERSON 等^[14]提出差分自回归移动平均模型。ZHANG 等^[15]提出迭代最小修复(IMR)对异常和缺失数据提供异常检测和修复。但是,许多经典方法都是假设时间变化遵循预定义的模式,而现实世界时间序列的变化通常过于复杂,无法被预定义的模式覆盖,限制了经典方法的实际适用性。

2)基于机器学习的方法。YOON 等^[16]提出基于生成对抗网络的 GAIN,其通过生成器与判别器迭代优化参数,最终得到目标生成器进行缺失数据生成。FENG 等^[17]提出基于支持向量机(SVM)的缺失值插补方法,通过 SVM 对时间序列点进行划分,采用划分最近的值进行插补。PANTANOWITZ 等^[18]提出随机森林算法,通过集成多种不同的插补方法筛选最优结果对缺失数据进行插补。虽然这些方法可以捕捉时序的非线性特征,但仅限于简单的时间序列,对于常见的多种模式叠加的复杂时间序列学习能力较弱。

3)基于深度学习的方法。CHE 等^[19]提出基于循环神经网络的缺失数据插补方法,该方法采用缺

失模式的两种表示形式,即掩码和时间间隔,并将其有效地整合到深度模型架构中,使其不仅捕获了时间序列中的长期时间依赖关系,而且利用缺失模式来获得更好的插补结果。HOCHREITER 等^[20]提出的长短期记忆网络通过串联多个循环神经元的变体进行时间序列的全局特征捕捉,解决了循环神经网络的梯度爆炸与长期依赖问题。郑欣彤等^[21]提出采用基于编码-解码结构的序列-序列深度学习结构 BiLSTM-I,以长短期记忆(LSTM)及其变体网络 LSTM-I 作为编码-解码结构的核心,实现对温度观测缺失数据的高精度插补。HE 等^[11]提出一种 TCN,通过在完整时间序列上进行训练,其能够预测未来多个时间步长的趋势变化并对时间序列中的缺失值进行有效插补。但这些方法大多是对时间序列局部的时间片段进行学习,欠缺对长时间跨度依赖性的学习。

本文综合上述方法提出 GANet,其基于自注意力机制得到全局特征权重矩阵,再结合门控循环单元对得到的全局特征权重矩阵进行遗忘与更新,从而得到新的具有时间跨度依赖性的全局特征权重矩阵,并通过该权重矩阵进行序列重构,使其兼顾了对全局特征的捕捉以及对长时间序列的时间跨度依赖性的学习,最后通过特征融合层融合高维特征得到插补完整的时间序列。

2 GANet 模型

2.1 相关定义

定义 1(多元时间序列) 将在时间 $T = \{t_1, t_2, \dots, t_n\}$ 内观察到的多元时间序列数据表示为一个时间序列 $\mathbf{X} = \{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{n-1}\} \in \mathbb{R}_n^D$,将 $\mathbf{X}$ 在第 i 个时间步内观察到的值表示为 $\mathbf{x}_i$,其中 $\mathbf{x}_i \in \mathbb{R}_n^{1 \times D}$, $\mathbf{x}_i$

包含 D 个特征 $\{x_i^1, x_i^2, \dots, x_i^D\}$, $\mathbf{x}_i$ 可能会缺失。

定义 2(缺失掩码) 缺失掩码矩阵 $\mathbf{M} \in \mathbb{R}^{n \times D}$,通过式(1)表示 $\mathbf{X}$ 中的缺失值:

$$M_i^d = \begin{cases} 0, & x_i^d \text{ 缺失} \\ 1, & x_i^d \text{ 未缺失} \end{cases} \quad (1)$$

根据式(1)得到时间序列 $\mathbf{X}$ 对应的缺失掩码矩阵 $\mathbf{M} = \{\mathbf{m}_0, \mathbf{m}_1, \dots, \mathbf{m}_{n-1}\} \in \{0, 1\}^{n \times D}$,其中, $\mathbf{m}_i$ 代表 $\mathbf{M}$ 的第 i 个向量,对应 $\mathbf{X}$ 的第 i 个观测值, m_i^j 是 $\mathbf{m}_i$ 的第 j 个值,对应 $\mathbf{x}_i$ 的第 j 个特征。时间序列的缺失率计算公式如式(2)所示:

$$R = \frac{\sum_{i=1}^n \sum_{j=1}^d (1 - M_i^j)}{n \times d} \quad (2)$$

根据不同的缺失率对时间序列 $\mathbf{X}$ 进行随机缺失预处理来模拟多元时间序列的完全随机缺失。

定义 3(位置掩码) 位置掩码矩阵 $\mathbf{P} \in \mathbb{R}^{n \times D}$,通过数据下标计算出每个数据的绝对位置和相对位置,将其作为辅助变量嵌入时间序列,对每个时间点进行位置标记,计算公式如式(3)所示:

$$\mathbf{P}_i = [\sin(\omega_0 i), \cos(\omega_0 i), \dots, \sin(\omega_k i), \cos(\omega_k i)] \quad (3)$$

$$\text{式中: } \omega_t = \frac{1}{1000 \frac{t}{d_{\text{model}}^{-1}}}; k = \frac{d_{\text{model}}}{2} - 1.$$

2.2 模型结构

如图 1 所示(彩色效果见《计算机工程》官网 HTML 版,下同),GANet 主要由嵌入层(embedding)与模块化的全局特征捕捉(GFC)模块以及特征融合层(feature fusion)组成,其中多个 GFC 模块通过残差计算相连,以捕获具有时间跨度依赖性的全局特征。

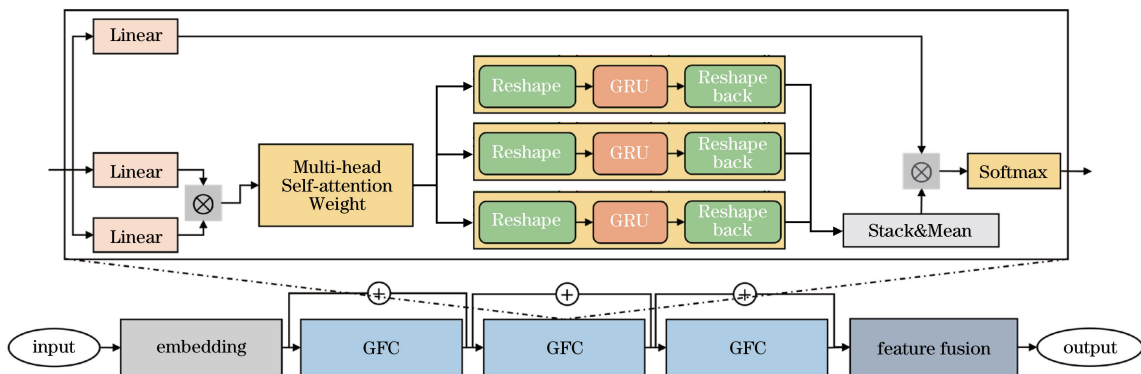


图 1 GANet 模型结构

Fig.1 GANet model structure

2.2.1 嵌入层

受 FEDformer^[22]启发,在嵌入层中通过一维卷积以及日期时间转换等操作,将时间序列本身携带

的时间信息与缺失信息连续化后叠加到时间序列中,以便在后续模块中更好地捕捉时间序列的时序性。

首先,根据时间序列的日期时间数据从位置掩码矩阵 P 中获取对应的时间位置信息,位置信息的计算如式(3)所示。其次,如图 2 所示,通过不同的时间粒度(分钟、小时、天等)对时间序列进行线性映射,从而得到所需时间粒度的时序信息,再将映射后的数值与时间序列叠加,丰富时间序列的时序性。映射过程如式(4)所示:

$$TF(T, t) = T \cdot W_t + b_t \quad (4)$$

式中: T 为需要映射的时间序列; t 代表映射粒度; W_t 为对应的映射权重; b_t 为偏置向量。

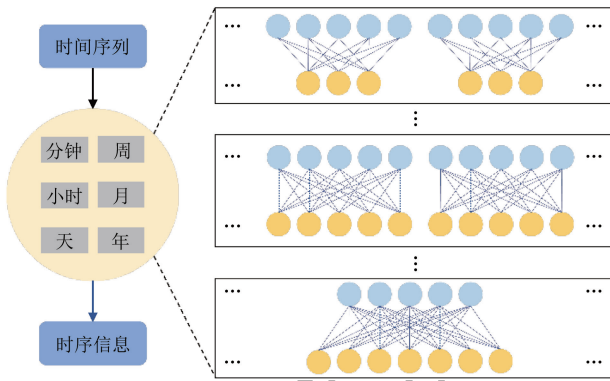


图 2 不同时间粒度的序列映射

Fig.2 Sequence mapping with different time granularities

除此之外,不同于其他大多数模型将缺失掩码与时间序列拼接后再进行线性映射的方式,GANet 先将缺失掩码进行线性映射,再叠加到时间序列上。映射过程如式(5)所示:

$$MF(M) = M \cdot W_m + b_m \quad (5)$$

式中: M 表示时间序列的缺失掩码; W_m 为对应的映射权重; b_m 为偏置向量。

综上所述,在嵌入层将位置信息、缺失信息等时间序列信息进行提取并叠加到时间序列上,以便后续的全局特征捕捉模块更好地学习时间序列的时序性。嵌入层的计算过程如式(6)所示:

$$X = X + R_{PE} + R_{TF} + R_{MF} \quad (6)$$

式中: X 表示输入嵌入层的时间序列, R_{PE} 、 R_{TF} 与 R_{MF} 分别代表对时间序列 X 的位置掩码信息、时间掩码信息与缺失掩码信息的提取。

2.2.2 全局特征捕捉模块

如图 3 所示,通过 3 个线性映射对输入的时间序列进行高维映射分别得到查询矩阵 Q 、键矩阵 K 和值矩阵 V ,使得模型能够更全面、更深入地学习时间序列的全局特征,计算过程如式(7)~式(9)所示:

$$Q = X \cdot W_q + b_q \quad (7)$$

$$K = X \cdot W_k + b_k \quad (8)$$

$$V = X \cdot W_v + b_v \quad (9)$$

将矩阵 Q 与 K 进行点积等计算得到 X 的基础权重矩阵,计算过程如式(10)所示:

$$w_x = \text{Softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \quad (10)$$

式中: d_k 表示特征维度大小。

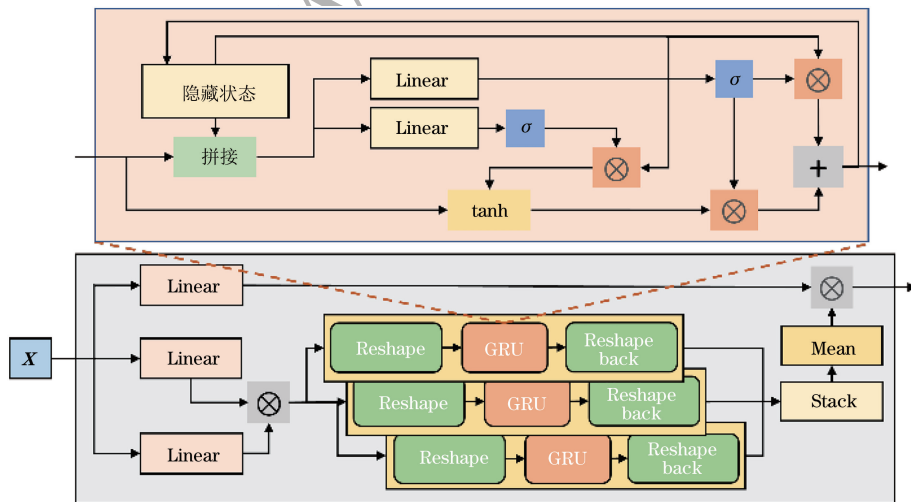


图 3 全局特征捕捉模块

Fig.3 Global feature capture module

此时,权重矩阵中包含了一段时间序列内各个时间点之间的影响权重,但该数值是通过矩阵乘得到的,并未考虑各个时间点在时间轴上随着时间跨度的增加对其他时间点的影响逐渐减弱的情况。因此,通过多层门控循环单元来实现对时间跨度依赖性的捕捉。其中,门控循环单元衔接在自注意力机

制计算后,并非添加在自注意力计算内。在图 3 中,每个全局特征捕捉模块中均含有多层门控循环单元。通过门控循环单元的遗忘与更新来捕获时间跨度对时间点权重的影响,门控循环单元内部主要由隐藏状态、遗忘门与更新门组成,在图 3 上半图对门控循环单元的细节描述中,位于上方的线性映射及

衔接的激活函数构成更新门,位于下方的线性映射及衔接的激活函数构成遗忘门。

首先,对当前权重矩阵进行形状变换,将批次维度与序列维度相融合。形状变换过程如式(11)所示:

$$\mathbf{w}'_x = \text{Rearrange}(\mathbf{w}'_x, \text{bnsm} \rightarrow (\text{bn})\text{sm}', n=n) \quad (11)$$

接着,门控循环单元对时间序列的遗忘与更新是通过线性变换与激活函数实现的,计算过程如式(12)、式(13)所示:

$$\mathbf{r}_t = \sigma(\mathbf{W}_r \cdot \mathbf{w}'_x + \mathbf{U}_r \cdot \mathbf{h}_{t-1} + \mathbf{b}_r) \quad (12)$$

$$\mathbf{z}_t = \sigma(\mathbf{W}_z \cdot \mathbf{w}'_x + \mathbf{U}_z \cdot \mathbf{h}_{t-1} + \mathbf{b}_z) \quad (13)$$

式中: $\mathbf{U}$ 和 $\mathbf{W}$ 为对应的权重; $\mathbf{h}_{t-1}$ 为隐藏状态; $\mathbf{r}_t$ 为遗忘结果; $\mathbf{z}_t$ 为更新结果。

然后,将遗忘门得到的遗忘结果再与输入重新结合并激活得到中间隐藏状态,计算公式如式(14)所示:

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}_c \cdot \mathbf{w}'_x, \mathbf{U}_c(\mathbf{r}_t \cdot \mathbf{h}_{t-1})) \quad (14)$$

最后,结合遗忘结果与中间隐藏状态得到新的隐藏状态,也是本次循环门控单元计算所得到的新的权重矩阵,计算公式如式(15)所示:

$$\mathbf{h}_t = \mathbf{z}_t \cdot \mathbf{h}_{t-1} + (1 - \mathbf{z}_t) \cdot \tilde{\mathbf{h}}_t \quad (15)$$

通过多层门控循环单元得到多个不同的新的权重矩阵,综合所有门控循环单元层的学习结果后得到全新的具有时序性的权重矩阵,并将权重矩阵形状还原,记为 $\hat{\mathbf{w}}_x$,再将权重矩阵与值矩阵进行点积得到初步的重构数据,计算公式如式(16)所示:

$$\mathbf{x}' = \hat{\mathbf{w}}_x \cdot \mathbf{V} \quad (16)$$

2.2.3 特征融合层

通过多层全局特征捕捉模块得到的初步重构数据是高维度数据,需要融合降维,如图4所示,通过两次特征映射来融合各个维度的数据,并在两次映射之间加入归一化层使得重构的时间序列更加平滑,最终将高维数据的维度降至与原始数据相同的维度,计算公式如式(17)所示:

$$\hat{\mathbf{x}} = \mathbf{W}_{l_1} \cdot (\text{LayerNorm}(\mathbf{W}_{l_2} \cdot \mathbf{x}') + \mathbf{b}_{l_2}) + \mathbf{b}_{l_1} \quad (17)$$

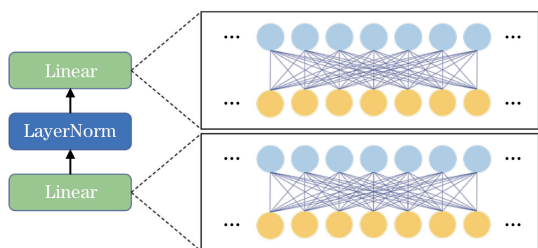


图4 特征融合层

Fig.4 Feature fusion layer

3 实验

3.1 数据集

实验采用东海船舶轨迹数据集来验证模型的效果。另外,为了验证 GANet 的通用性,在电力、天气等领域的6个公开数据集上进行测试。数据集的详细说明如表1所示。所有数据集在预处理时进行了标准化处理,并通过随机采样实现数据集的4种缺失率。

表1 数据集详情

Table 1 Dataset detail

数据集	特征数/ 个	训练集样 本数/个	验证集样 本数/个	测试集样 本数/个	长度	采集 频率
Trajectory	4	8 545	2 881	2 881	96	—
ETTh1, ETTh2	7	8 545	2 881	2 881	96	1 h
ETTm1, ETTm2	7	34 465	11 521	11 521	96	15 min
Electricity	321	18 317	2 633	5 261	96	1 h
Weather	32	36 792	5 271	10 540	96	10 min

Trajectory:来自东海船舶自动识别系统(AIS)设备传回的轨迹数据,为海上真实的船舶历史轨迹数据,该数据集涵盖多个维度的信息,每一条轨迹数据包括经度、纬度等,该类数据集广泛应用于船舶航行状态评估、船舶碰撞概率检测等场景,采集间隔不固定。

ETT^[23]:包含4个数据集,其中2个采集自中国某省的2个地区,分别命名为 ETTm1 和 ETTm2。每条数据由日期时间及7个特征组成,包括油温和6种不同类型的外部电源负载特征。数据采集时间为2016年7月1日至2018年6月30日(共24个月),以15 min为采集间隔。此外,还包含2个开发的小时级变体数据,命名为 ETTh1 和 ETTh2。

Electricity:是 UC Irvine Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014>)提供的一个公共电力数据集,每条数据包含370个用户点的用电量数据。数据采集时间为2016年7月1日至2019年6月30日(共36个月),以1 h为采集间隔,且该数据集无缺失数据。考虑到该数据集中存在大量数据为0的用户,剔除了该数据集中的部分用户数据。

Weather:是 Dokumentation Wetter Station (<https://www.bgc-jena.mpg.de/wetter/>)提供的气象数据集,每条数据包含温度、湿度、风力等32个

特征。数据采集时间为 2020 年 1 月 1 日至 2020 年 12 月 31 日(共 12 个月),以 15 min 为采集间隔。

3.2 基准模型

考虑到当前时间序列插补的主流模型均基于深度学习,采用当前时间序列领域较为先进的基于 Transformer 的变体模型、基于卷积的时间序列模型以及经典的基于神经网络的模型作为基准模型,具体如下:1) 基于递归神经网络(RNN)的模型,如 LSTM^[20]、LSSL^[24]; 2) 基于 CNN 的模型,如 TCN^[11]; 3) 基于 Transformer 的模型,如 LogTrans^[25]、Reformer^[13]、Informer^[23]、Pyraformer^[26]、Autoformer^[12]、FEDformer^[22]、Non-stationary Transformers^[27] 和 ETSformer^[28]。

3.3 实验设置

实验采用两个指标来评估模型的插补性能:均方误差(MSE, E_{MSE})和平均绝对误差(MAE, E_{MAE})。计算过程如式(18)、式(19)所示:

$$E_{MSE} = \frac{\sum_{d=1}^D \sum_{t=1}^N ((\hat{x}_t^d - x_t^d)^2 \cdot M_t^d)}{\sum_{d=1}^D \sum_{t=1}^N M_t^d} \quad (18)$$

$$E_{MAE} = \frac{\sum_{d=1}^D \sum_{t=1}^N (|\hat{x}_t^d - x_t^d| \cdot M_t^d)}{\sum_{d=1}^D \sum_{t=1}^N M_t^d} \quad (19)$$

式中: $\hat{x}_t^d$ 表示模型插补值; x_t^d 表示真实值; M_t^d 表示缺失值。

在实验参数的详细设置中:在数据相关方面,所有实验统一采用批量大小为 16,时间序列长度为 96;在模型训练相关方面,所有实验学习率初始均为 0.001,默认时间粒度为小时,模型训练轮数为 10,并采用早停策略,早停条件为连续 3 次

MSE 没有下降;在模型优化方面,采用 MSE 作为反向传递的误差值以及 Adam 作为优化器,所有模型均在单个 NVIDIA GeForce RTX 3090 上进行训练。

3.4 实验结果

GANet 在两种不同的缺失率下对船舶轨迹数据的插补效果如图 5 所示。实验结果显示,GANet 对于船舶轨迹数据的插补效果最优,且相比于 Autoformer、Informer 模型具有较大提升,这是由于船舶航行时轨迹变化较缓和,数据的时序性较强,时间跨度对时间点之间的影响力越强,而 GANet 对时间跨度依赖性的捕捉使其对于时间序列全局特征的学习更加准确,因此 GANet 对于船舶轨迹数据的插补能力更强。

另外,为了验证 GANet 的通用性,在 Electricity 等 6 个公开数据集上进行了 12.5%、25.0%、37.5%、50.0% 4 种缺失率的模型插补实验,并计算了 4 种情况下的结果均值。实验结果如表 2 所示,其中,黑体代表最优值,下同。实验结果显示,GANet 在不同缺失率下的 6 个数据集上均展示出了不错的插补性能,其中,在 ETT 数据集与 Electricity 数据集上几乎优于所有基准模型,但在 Weather 数据集上相比于 Autoformer 稍差,这可能是由于该数据集的每个特征是从不同来源收集的,不同特征之间的时间模式可能差别较大,使得总体时间序列的模式更加复杂,同时数据受到的自然因素影响较为严重,在时间上变化较为剧烈,增加了插补难度,使得 GANet 对该数据集的时间跨度依赖性捕捉准确度下降。总体而言,GANet 对 6 个数据集的 4 种缺失情况的插补效果优于大多数基准模型,这验证了 GANet 的通用性。

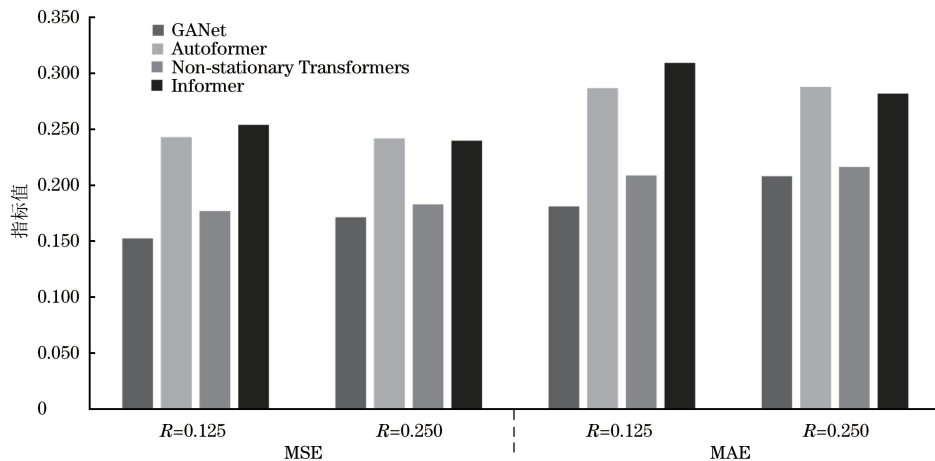


图 5 在不同缺失率下的模型插补结果

Fig.5 Model imputation results under different missing rates

表 2 GANet 与基准模型插补结果的误差均值

Table 2 Mean errors of imputation results among GANet and benchmark models

模型	指标	ETTm1	ETTm2	ETTh1	ETTh2	Electricity	Weather
ETSformer	MSE	0.120	0.208	0.202	0.367	0.214	0.076
	MAE	0.253	0.327	0.329	0.436	0.339	0.171
FEDformer	MSE	0.062	0.101	0.117	0.163	0.130	0.099
	MAE	0.178	0.215	0.246	0.279	0.259	0.203
Non-stationary Transformers	MSE	0.036	0.026	0.094	0.053	0.100	0.032
	MAE	0.126	0.099	0.201	0.152	0.218	0.059
Autoformer	MSE	0.051	0.029	0.103	0.055	0.101	0.031
	MAE	0.151	0.105	0.214	0.156	0.225	0.057
Pyraformer	MSE	0.717	0.465	0.842	1.079	0.297	0.152
	MAE	0.570	0.508	0.682	0.792	0.382	0.235
Informer	MSE	0.071	0.156	0.161	0.337	0.222	0.045
	MAE	0.188	0.292	0.279	0.452	0.328	0.104
LogTrans	MSE	0.050	0.119	0.219	0.186	0.175	0.039
	MAE	0.154	0.246	0.332	0.318	0.303	0.076
Reformer	MSE	0.055	0.157	0.122	0.234	0.200	0.038
	MAE	0.166	0.280	0.245	0.352	0.313	0.087
LSTM	MSE	0.989	1.027	1.225	2.039	0.277	0.365
	MAE	0.786	0.800	0.873	1.114	0.365	0.434
TCN	MSE	0.516	0.266	0.621	0.431	0.582	0.183
	MAE	0.497	0.407	0.571	0.503	0.597	0.291
LSSL	MSE	0.113	0.175	0.424	0.495	0.222	0.045
	MAE	0.243	0.324	0.481	0.475	0.293	0.108
GANet	MSE	0.034	0.025	0.084	0.053	0.095	0.035
	MAE	0.123	0.098	0.193	0.154	0.212	0.065

3.5 消融实验

为了验证模型嵌入层与全局特征捕捉模块的效果,本节进行两个消融实验以及循环门控单元层数的实验。在这些实验中仅对 GANet 模型内容进行修改,不涉及对实验参数的修改,以及实验采用的指标仍为 MAE 与 MSE,数据集也与第 3.4 节中的实验一致,且实验的误差结果均基于测试集。

3.5.1 嵌入层消融实验

为了验证嵌入层中 3 种不同的嵌入方式对时间序列特征捕捉的影响,本节分别消去了位置掩码信息的嵌入、时间粒度信息的嵌入以及缺失掩码信息的嵌入,并分别记为 NPE 模型、NTF 模型与 NMF 模型。在 12.5%、25.0%、37.5%、50.0% 4 种缺失率下进行了实验,并计算 4 种情况的误差均值,实验结果如图 6 所示,时间粒度信息的嵌入对模型特征捕捉影响最大,对模型效果有极大提升,而位置掩码信息与缺失掩码信息的嵌入的影响相对较小,这里需要说明的是,NPE 与 NMF 在 12.5%与 25.0%的

缺失率下对模型效果有较明显的提升,而在 37.5%与 50.0%的缺失率下却对模型效果有较大的削弱,这可能是由于较高的缺失率下两种嵌入对模型特征捕捉的准确率影响较大,所学习到的特征与真实特征相差更大。因此,NPE 与 NMF 在 4 种缺失率下的误差均值几乎与 GANet 持平。由于在高缺失率下缺失掩码信息与位置掩码信息的嵌入可能会存在负作用,因此为了保证对比实验的准确性,在第 3.4 节的实验中,不针对缺失率对嵌入层进行调整,在不同缺失率下均同时使用 3 种嵌入方式。

3.5.2 门控循环单元消融实验

为了验证 GANet 中门控循环单元对时间序列全局特征捕捉的优化效果,本节进行了剔除门控循环单元的实验以及不同层数的门控循环单元的实验。

如图 7、图 8 所示,相比于 GANet,剔除门控循环单元的模型 Abolition 在几乎所有情况下的插补精度都有所降低,或是与 GANet 持平,这证明了 GANet 的合理性。

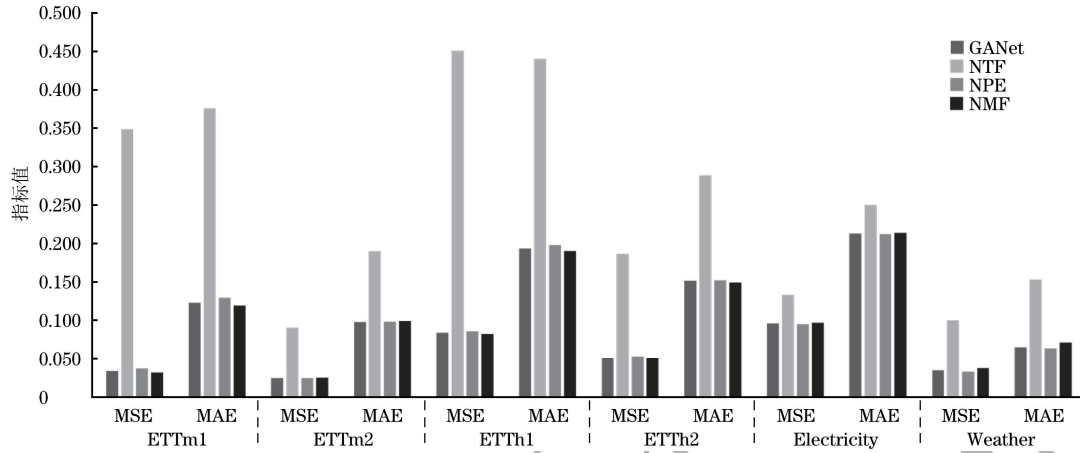


图 6 GANet、NPE、NTF、NMF 插补结果的误差均值

Fig.6 Mean errors of imputation results for GANet, NPE, NTF and NMF

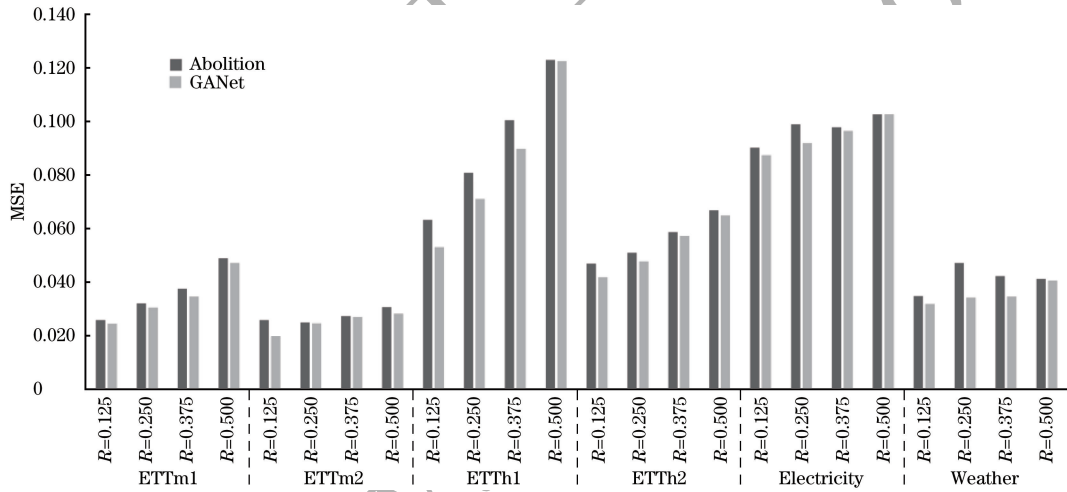


图 7 Abolition 与 GANet 的 MSE 值

Fig.7 MSE values of Abolition and GANet

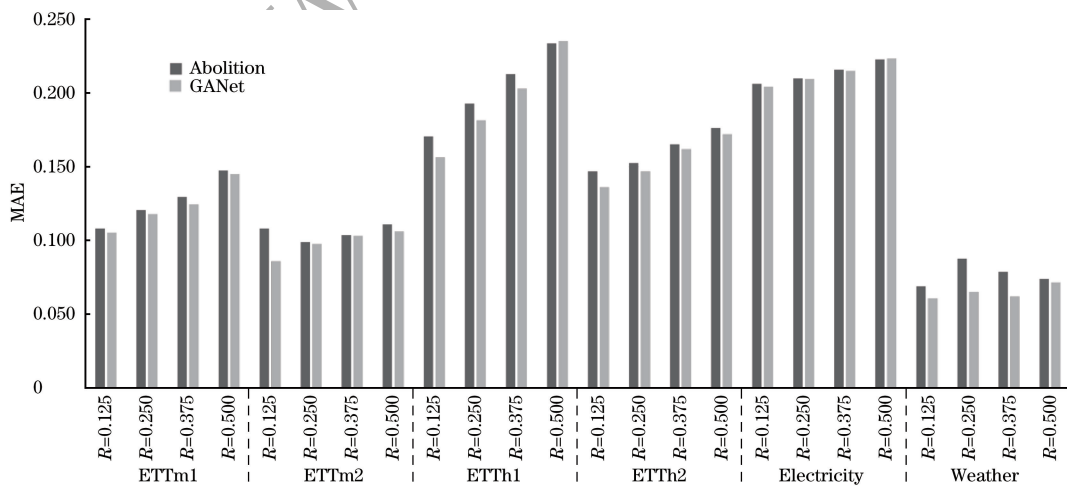


图 8 Abolition 与 GANet 的 MAE 值

Fig.8 MAE values of Abolition and GANet

3.5.3 门控循环单元层数实验

为了进一步研究门控循环单元对时间序列时序性学习的效果,为 GANet 设置了不同层数的门控循环单元进行实验。如图 9、图 10 所示,

其中 i 表示循环层数。当循环层数为 3、4 时, GANet 在 Electricity 数据集上的性能稍有提升,而在其他 5 个数据集上性能稍有下降。另外,在部分数据集上更高的门控循环单元层数的实验

结果表明,当缺失率为 12.5%与 25.0%时,较高的门控循环单元层数有较大的性能提升,但当缺失率为 37.5%与 50.0%时却有较大性能的下降,可能是较高的缺失率导致时间序列的时序性

丢失严重,以及模型的复杂度增加,削弱了 GANet 对于时间序列全局特征的正确捕捉以及门控循环单元对于时间跨度依赖性的捕捉能力,导致模型训练过早停止。

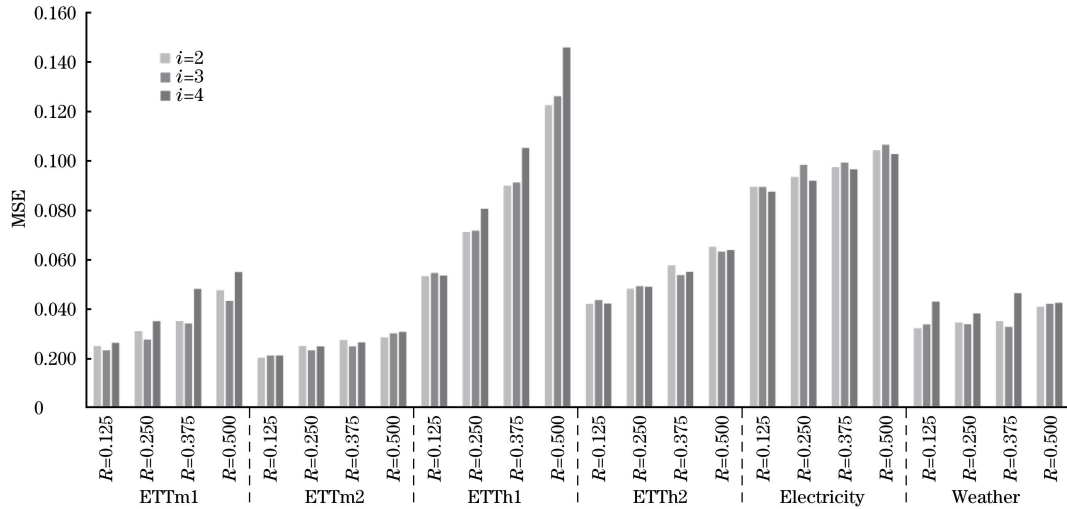


图 9 GANet 不同循环层数的 MSE 值

Fig.9 MSE values of GANet with different number of loop layers

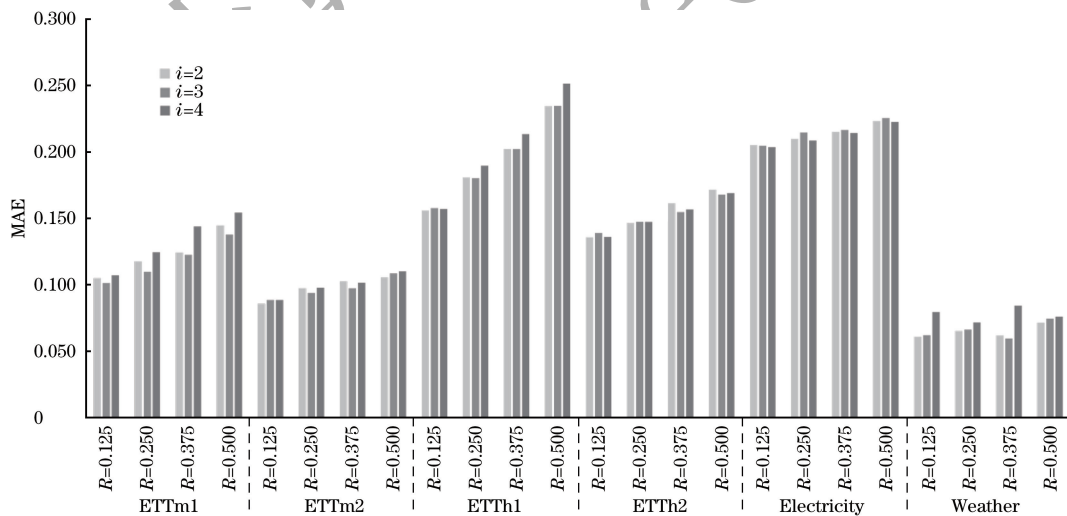


图 10 GANet 不同循环层数的 MAE 值

Fig.10 MAE values of GANet with different numbers of loop layers

4 结束语

针对长时间序列全局特征捕捉的问题,本文提出了一种基于自注意力机制的时间序列插补模型 GANet。该模型通过自注意力机制构建时间序列的查询、键、值矩阵,并在此基础上加入门控循环单元捕捉时间序列的时序性,使得模型能够在捕捉时间序列全局特征的同时兼顾对长时间序列的时间跨度依赖性的学习。为了验证 GANet 的有效性,在东海船舶轨迹数据集以及 Electricity 等公开数据集上进行了插补实验,实验结果表明,相比于基准模型,GANet 对于船舶轨迹数据具有更好的插补效

果,同时在 ETT 与 Electricity 数据集上有较好的表现,但在 Weather 数据集上略差,这可能是 Weather 数据集中不同特征之间的时间序列模式相差较大且更加复杂,导致 GANet 对该数据集的时间跨度依赖性捕捉准确度较低。

此外,为了证明 GANet 设计的合理性,本文进行了消融实验,实验结果证明了嵌入层的 3 种嵌入方法与门控循环单元在模型中起到了重要作用,而门控循环单元部分不同层数的实验结果表明,在较低缺失率下,更高的层数对于复杂的数据集有更好的插补效果,但缺失率较高会导致模型学习效果下降。未来将继续研究 GANet 对于不同特征模式之

间相关性较低的时间序列的时间跨度依赖性捕捉问题,提高对于该种时间序列的模式学习能力,以及平衡门控循环单元层数与缺失率之间的影响。另外,将研究 GANet 在其他时间序列分析任务上的性能,以进一步验证该模型在时间序列建模领域的通用性。

参考文献

- [1] SILVA I, MOODY G, SCOTT D J, et al. Predicting in-hospital mortality of ICU patients: the PhysioNet/computing in cardiology challenge 2012 [J]. *Computing in Cardiology*, 2012, 39: 245-248.
- [2] YI X, ZHENG Y, ZHANG J, et al. ST-MVL: filling missing values in geo-sensory time series data [C] // *Proceedings of the 25th International Joint Conference on Artificial Intelligence*. Palo Alto, USA: AAAI Press, 2016: 2704-2710.
- [3] DU W J, COTE D, BARBER C, et al. Forecasting loss of signal in optical networks with machine learning [J]. *Journal of Optical Communications and Networking*, 2021, 13(10): 109-121.
- [4] BRUNI R, DARAIO C, AUREL D. Imputation techniques for the reconstruction of missing interconnected data from higher Educational Institutions [J]. *Knowledge-Based Systems*, 2021, 212: 106512.
- [5] 于明霞. 基于深度学习的交通数据插补与预测方法研究 [D]. 长沙: 湖南大学, 2022.
- [6] 杨正泽. 轨道检测数据异常检测及缺失数据插补算法研究 [D]. 北京: 北京交通大学, 2023.
- [7] SHAO Z Z, ZHANG Z, WANG F, et al. Pre-training enhanced spatial-temporal graph neural network for multivariate time series forecasting [C] // *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. New York, USA: ACM Press, 2022: 1567-1577.
- [8] MIRZA M, OSINDERO S. Conditional generative adversarial nets [EB/OL]. [2024-03-05]. <https://arxiv.org/abs/1411.1784>.
- [9] CAO W, WANG D, LI J, et al. BRITS: bidirectional recurrent imputation for time series [EB/OL]. [2024-03-05]. <https://arxiv.org/abs/1805.10572v1>.
- [10] 雷未, 王建, 吉同元, 等. 基于深度学习框架的长序列大坝监测缺失数据插补模型 [J]. *水利水电科技进步*, 2023, 43(6): 82-88.
- [11] HE Y D, ZHAO J B. Temporal convolutional networks for anomaly detection in time series [J]. *Journal of Physics: Conference Series*, 2019, 1213(4): e042050.
- [12] WU H X, XU J H, WANG J M, et al. Autoformer: decomposition Transformers with auto-correlation for long-term series forecasting [EB/OL]. [2024-03-05]. <https://arxiv.org/abs/2106.13008v5>.
- [13] KITAEV N, KAISER L, LEVSKAYA A. Reformer: the efficient Transformer [EB/OL]. [2024-03-05]. <https://arxiv.org/pdf/2001.04451>.
- [14] ANDERSON O D, KENDALL M. Time-series [J]. *The Statistician*, 1976, 25(4): 308.
- [15] ZHANG A Q, SONG S X, WANG J M, et al. Time series data cleaning [J]. *Proceedings of the VLDB Endowment*, 2017, 10(10): 1046-1057.
- [16] YOON J, JORDON J, SCHAAR M. GAIN: missing data imputation using generative adversarial nets [C] // *Proceedings of International Conference on Machine Learning*. [S. l.]: PMLR, 2020: 5689-5698.
- [17] FENG H H, CHEN G S, YIN C, et al. A SVM regression based approach to filling in missing values [M]. Berlin, Germany: Springer, 2005.
- [18] PANTANOWITZ A, MARWALA T. Missing data imputation through the use of the random forest algorithm [M]. Berlin, Germany: Springer, 2009.
- [19] CHE Z P, PURUSHOTHAM S, CHO K, et al. Recurrent neural networks for multivariate time series with missing values [J]. *Scientific Reports*, 2018, 8(1): 6085.
- [20] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. *Neural Computation*, 1997, 9(8): 1735-1780.
- [21] 郑欣彤, 边婷婷, 张德强, 等. 基于深度学习的温度观测数据长时间缺失值插补方法 [J]. *计算机系统应用*, 2022, 31(4): 221-228.
- [22] ZHENG X T, BIAN T T, ZHANG D Q, et al. Interpolation of long time missing values of temperature based on deep learning [J]. *Computer Systems and Applications*, 2022, 31(4): 221-228. (in Chinese)
- [23] ZHOU T, MA Z Q, WEN Q S, et al. FEDformer: frequency enhanced decomposed Transformer for long-term series forecasting [EB/OL]. [2024-03-05]. <https://arxiv.org/abs/2201.12740v3>.
- [24] GU A, GOEL K, RÉ C. Efficiently modeling long sequences with structured state spaces [EB/OL]. [2024-03-05]. <https://arxiv.org/abs/2111.00396v3>.
- [25] LI S Y, JIN X Y, XUAN Y, et al. Enhancing the locality and breaking the memory bottleneck of Transformer on time series forecasting [EB/OL]. [2024-03-05]. <https://arxiv.org/abs/1907.00235v3>.
- [26] LIU S, YU H, LIAO C, et al. Pyraformer: low-complexity pyramidal attention for long-range time series modeling and forecasting [EB/OL]. [2024-03-05]. <https://openreview.net/forum?id=0EXmFzUn5I>.
- [27] LIU Y, WU H X, WANG J M, et al. Non-stationary Transformers: exploring the stationarity in time series forecasting [J]. *Advances in Neural Information Processing Systems*, 2022, 35: 9881-9893.
- [28] WOO G, LIU C H, SAHOO D, et al. ETSformer: exponential smoothing Transformers for time-series forecasting [EB/OL]. [2024-03-05]. <https://arxiv.org/abs/2202.01381v2>.

编辑 陆燕菲