

ASformer: 具有频谱带通变换和对抗性注意力机制的 Transformer 用于多变量时间序列预测

尹 康, 刘梦赤, 章概煜

(华南师范大学 计算机学院, 广东 广州 510631)

摘要: Transformer 模型已成为时间序列预测的基石, 利用自注意力机制可有效捕捉长期依赖关系。然而, 当应用多变量时间序列时, 传统自注意力机制难以区分关键变量, 导致注意力分布过于均匀, 掩盖了关键的时间和跨变量关系, 还容易受时间序列中噪声信号的影响, 降低准确性。为此, 提出一种基于 Transformer 的新型架构, 包含频谱带通变换 (SBT) 和对抗性双分支注意力机制 (ADBA)。首先, SBT 模块利用实值快速傅里叶变换 (rFFT) 分离指定频带内的主频分量, 基于汉明窗的平滑处理抑制噪声、增强信号保真度; 其次, ADBA 模块引入一种双分支注意力机制: 一个分支将基于雅可比矩阵梯度引导的对抗扰动应用于查询和键向量, 另一个分支则保持标准注意力机制; 最后, 通过可学习的加权方案自适应融合两个分支的输出。实验表明, 所提设计通过强调显著变量和强化多元相关性来优化注意力分配, 减轻了普通自注意力机制固有的同质化问题, 这些模块集成在 Transformer 编码器中协同工作, 提升了模型捕捉复杂时间动态和变量间依赖关系的能力。在 ECL、Traffic 等基准数据集的实验表明, 所提模型的平均 MSE 下降 2%, 在预测准确性方面优于当前先进的方法。

关键词: Transformer; 自注意力机制; 长期依赖关系; 多变量时间序列; 实值快速傅里叶变换

DOI: 10.11907/rjdk.251460

扫描二维码阅读全文: 

中图分类号: TP18

文献标识码: A

文章编号: 1672-7800(XXXX)0XX-0001-07

ASformer: Transformer with Spectral Bandpass Transform and Adversarial Attention for Multivariate Time Series Forecasting

YIN Kang, LIU Mengchi, ZHANG Gaiyu

(School of Computer Science, South China Normal University, Guangzhou 510631, China)

Abstract: Transformer model has become the cornerstone of time series prediction, and utilizing self attention mechanisms can effectively capture long-term dependencies. However, when applying multivariate time series, traditional self attention mechanisms are difficult to distinguish key variables, resulting in an overly uniform distribution of attention, masking key temporal and cross variable relationships, and being easily affected by noise signals in the time series, reducing accuracy. To this end, a novel architecture based on Transformer is proposed, which includes Spectrum Bandpass Transform (SBT) and Adversarial Dual Branch Attention Mechanism (ADB). Firstly, the SBT module uses real valued fast Fourier transform (rFFT) to separate the main frequency components within a specified frequency band, and based on Hamming window smoothing processing, suppresses noise and enhances signal fidelity; Secondly, the ADB module introduces a dual branch attention mechanism: one branch applies adversarial perturbations guided by Jacobian matrix gradients to queries and key vectors, while the other branch maintains the standard attention mechanism; Finally, adaptively fuse the outputs of the two branches through a learnable weighting scheme. Experiments have shown that the proposed design optimizes attention allocation by emphasizing significant variables and strengthening multivariate correlations, alleviating the inherent homogenization problem of ordinary self attention mechanisms. These modules are integrated into the Transformer encoder to work together, enhancing the model's ability to capture complex temporal dynamics and inter variable dependencies. Experiments on benchmark datasets such as ECL and Traffic have shown that the proposed model has an average MSE decrease

收稿日期: 2025-07-14

基金项目: 国家自然科学基金项目 (61672389); 广州市大数据智能教育重点实验室项目 (201905010009)

作者简介: 尹康 (2001-), 男, 华南师范大学计算机学院硕士研究生, 研究方向为基于深度学习的时间序列预测方法研究; 刘梦赤 (1962-), 男, 华南师范大学计算机学院教授, 研究方向为数据库语义推理, 大数据系统; 章概煜 (2001-), 男, 华南师范大学计算机学院硕士研究生, 研究方向为跨语言 Text-to-SQL 语义解析。本文通讯作者: 尹康。

of 2% and outperforms current advanced methods in terms of prediction accuracy.

Key Words: Transformer; self-attention mechanisms; long-time dependencies; multivariate time series; real-valued fast Fourier transform

0 引言

时间序列分析支撑着各种实际应用,例如天气预报、能源优化和社会情绪追踪^[1-3]。这些领域共同的挑战在于如何精准从历史数据中进行学习,预测未来的动态变化。随着物联网、传感器网络和金融市场的蓬勃发展,产生了前所未有的海量、高维时间序列数据,不仅包含每个变量自身随时间演变的复杂动态(时间依赖性),还蕴含了变量间错综复杂的相互影响(跨变量依赖性)。因此,开发能高效捕捉这两种依赖关系,并能扩展到数百甚至数千个变量的模型,已成为当前该领域研究的迫切需求。

由于Transformer模型在自然语言处理、计算机视觉领域都取得了巨大成功,逐渐发展成了遵循尺度定律的基础模型,彻底改变了时间序列分析这一领域,因此利用自注意力机制捕捉序列数据中的长距离依赖关系,处理时间序列任务的效率较高^[4-8]。近期,Pyraformer采用分层注意力结构建模具有线性复杂度的多尺度时间依赖关系^[9]。Autoformer将季节性趋势分解与自相关机制相结合,以增强模型提取序列间关系能力^[10]。尽管,这些模型取得了成功,但在处理包含大量变量的多变量时间序列(862个维度的Traffic数据集)时,仍面临巨大挑战。在这种情况下,传统自注意力机制往往会将注意力权重均匀地分布在各个变量之间,但关键变量间依赖关系的关注度有所减少。具体表现为:注意力图缺乏明显的峰值,例如Transformer、iTransformer的注意力得分较低,无法集中在关键特征上^[11]。

为此,本文提出一种基于Transformer的新型架构ASformer,主要包含频谱带通变换(Spectrum Bandpass Transform, SBT)和对抗性双分支注意力(Adversarial Dual Branch Attention, ADBA)这两个专用模块。这些模块共同作用使ASformer克服传统Transformer注意力均匀分布的问题,尤其在高维环境中可减轻噪声影响。本文贡献主要体现在3个方面:①提出一种专为长期多变量预测量身定制的Transformer架构ASformer,旨在解决高变量场景下注意力分配不足的难题;②SBT模块通过隔离关键频率成分、平滑序列来提升数据质量,ADBA模块利用对抗性学习优化注意力焦点,以共同提升变量间相关性的捕捉能力;③通过严格的实验验证模型优势,并将其定位为优秀的时间序列预测解决方案。

1 相关工作

1.1 基于频率的策略及其挑战

频域技术用于丰富时间序列的表示形式,常将其与各

种基础模型集成。例如,SMF将序列分解为频带,以进行多视野预测;StemGNN将图神经网络与傅里叶变换相结合,以捕捉序列间和序列内的依赖关系^[16,17];FEDformer利用频域稀疏性分析季节性趋势^[18];FreTS使用频域多层感知器(Multilayer Perceptron, MLP)强调全局依赖性^[19];TiDE利用谱嵌入改进长期预测^[20];CoST、ATFN专注于季节性趋势分离和动态周期模式捕捉^[21-24]。这些方法凸显了频率分析的价值,但一个常见的缺陷是无法有效处理分解过程中引入的噪声,可能会放大不相关的成分,从而降低下游表征质量。

为此,本文提出ASformer模型,在频谱带通变换(SBT)模块中通过带通滤波、汉明窗平滑解决这个问题,以确保信号更清晰,预测精度更高。

1.2 注意力驱动的时间序列建模创新

随着自然语言处理和计算机视觉领域的重大进步,各种复杂的Transformer变体已被开发出来,用于处理广泛的时间序列预测任务^[25,26]。相较于传统的时间卷积网络(Time Convolutional Network, TCN)和基于循环神经网络(Recurrent Neural Network, RNN)的模型,Transformer展现出了卓越的序列建模能力和强大的可扩展性,推动了深度优化时间序列预测的热潮^[27-31]。

早期,Informer利用ProbSparse自注意力机制高效处理长序列^[12]。Reformer则采用局部敏感哈希抑制内存需求^[13]。Pyraformer引入一种分层注意力机制处理跨多个时间分辨率的模型依赖关系。ContiFormer将注意力机制与神经微分方程相结合,处理不规则时间序列^[21]。近期,PatchTST将时间序列分割成块以揭示局部相关性^[14]。Crossformer强调跨维度的交互作用^[15]。这些方法旨在优化计算效率或增强局部模式识别,但忽略了高维环境下的一个关键限制:标准注意力机制倾向于将权重均匀地分布在各个变量之间,削弱了对关键变量间关系的关注度。同时,这种均匀分布无法高效建模复杂的多变量动态,而本文的ASformer正试图通过其对抗性双分支注意力(Adversarial Dual Branch Attention, ADBA)模块来解决这个问题。

ASformer独特整合了SBT和ADBA,以应对注意力均匀分布和噪声放大的挑战。相较于以块为中心的方法(PatchTST)而言,ADBA采用对抗性学习增强注意力聚焦。通过计算注意力矩阵的基尼系数,清晰反映不同变量子体间注意力机制的关注程度。实验表明,PatchTST、iTransformer和普通Transformer的基尼系数分别为0.098、0.078和0.081,而ASformer的显著值达到0.146,表明ASformer具有更集中的注意力分布特征,能更有效地聚焦时间序列中的关键变量,缓解注意力权重分布过于均匀问题。SBT的定向频率选择和去噪功能则使其有别于更广泛的分解策略(FEDformer),使ASformer成为了多变量时间序列预测

的先驱解决方案。

2 ASformer

2.1 问题表述

多变量时间序列预测是一项关键的科学与工程问题,核心在于对一组随时间共同演变的变量进行联合建模,以精准预测未来状态。该任务的根本挑战不仅在于捕捉每个单一变量内在的时间动态(趋势性、季节性),更在于揭示并利用不同变量间复杂且随时间变化的相互依赖关系(跨变量动态)。多变量时间序列由多个变量在一系列时间步上的观测值组成,包含 T 个时间步和 N 个变量,每个时间步都记录了所有 N 个变量的观测值。时间序列预测的主要目标是利用历史数据预测未来连续 S 个时间步内所有变量的取值。在每一个时间步,模型以最近的 L 个时间步的观测数据作为输入,包括所有变量在过去 L 个时间步的观测值,模型基于这些输入来预测未来 S 个时间步内所有 N 个变量的数值。

2.2 模型架构

ASformer 是一种新颖的基于 Transformer 的框架,旨在解决高维数据中注意力分布过于均匀和噪声干扰的问题,提升多变量时间序列的预测性能,模型架构如图 1 所示。

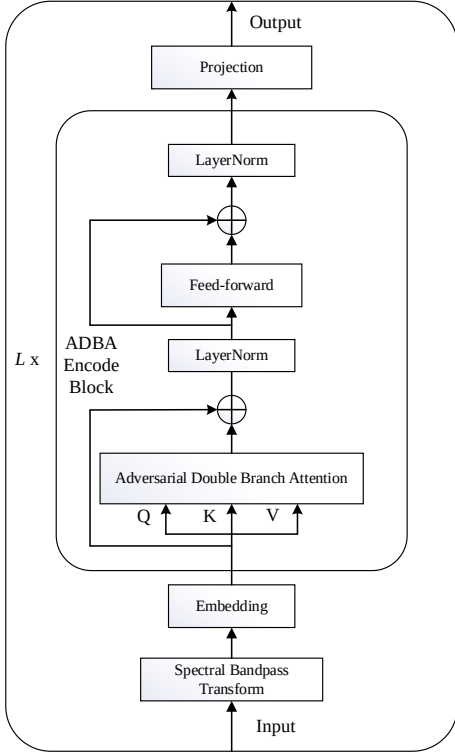


Fig. 1 Overall architecture of the model

图 1 模型整体架构

由此可见,该架构由 4 个无缝集成的模块组成:频谱带通变换(SBT)模块、嵌入层(Embedding Layer)、一组由对抗性双分支注意力(ADBA)编码器块堆叠而成的编码器及

投影层(Projection Layer)。首先,SBT 模块利用实值快速傅里叶变换(Real valued Fast Fourier Transform, rFFT)进行带通滤波,结合 Hamming 窗平滑对输入时间序列进行去噪处理,在抑制噪声的同时保留关键的时序特征^[22]。其次,经过处理的序列先被嵌入为变量特定的向量,再输入到多层堆叠的 ADBA 编码器块中。这些编码器块采用双分支注意力机制,将标准注意力与对抗性增强注意力相结合,以聚焦关键的变量间依赖关系,并辅以层归一化和前馈网络实现更强的时序建模能力。最后,投影层将丰富的表示解码为预测结果,ASformer 通过紧密协作的设计以有效捕捉复杂的多变量相关性,在复杂预测任务中效果优于传统的 Transformer 模型。

2.3 频谱带通变换

时间序列去噪对增强噪声对预测精度的影响至关重要,在多变量环境中的效果明显。为此,本文提出一种基于频谱带通变换模块的新型去噪策略,将时间序列数据转换到频域,选择性地保留关键频率成分,抑制噪声和不相关的波动。具体过程如算法 1 所示。

算法 1 频谱带通变换算法

输入: Time series $X \in \mathbb{R}^{T \times N}$, length T , Variates N 。

输出: Filtered and smoothed series $X_h \in \mathbb{R}^{T \times N}$ 。

1. A Hamming Window w_n sized w ; the frequency band $[f_{low}, f_{high}]$
2. For $n = 1$ to N do
3. $X_{fn} = rFFT(x_n)$ Fast Fourier Transform
4. $X_{fn} = Bandpass(X_{fn}, f_{low}, f_{high})$ Select the frequency band
5. $x_{ifn} = IrFFT(X_{fn})$ Convert to the time domain
6. $x_{pn} = Reflective(x_{ifn}, w_n)$ Pad the signal
7. $x_{hn} = Apply(x_{pn}, w_n)$ Smooth with Hamming window
8. End
9. $X_h = Concat(x_{hn})$ Concatenate N univariate series
10. Return: X_h
11. END

首先,利用实值快速傅里叶变换将长度为 T 的单变量时间序列 ($x \in \mathbb{R}^T$) 转换为其频域表示 X_f 。

$$X_f[k] = \sum_{t=0}^{T-1} x[t] e^{-2\pi i \frac{kt}{T}}, k = 0, \dots, \lfloor T/2 \rfloor \quad (1)$$

式中: $X_f[k]$ 表示频域系数; k 表示奈奎斯特频率 $\lfloor T/2 \rfloor$ 以内的离散频率分量。

为了保留最有用的频率,应用带通滤波器保留指定范围 $[f_{low}, f_{high}]$ 内的分量, f_{low}, f_{high} 是可定义的频率索引 ($0 \leq f_{low} < f_{high} \leq \lfloor T/2 \rfloor$)。滤波后的频域信号 X'_f 为:

$$X'_f[k] = \begin{cases} X_f[k], & \text{if } f_{low} \leq k \leq f_{high}, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

这种带通滤波可分离出重要的时间模式,例如周期性和趋势性,排除高频噪声和低频漂移。同时,本文应用逆实值快速傅里叶变换(Inverse real valued Fast Fourier Transform, IrFFT)将滤波后的信号变回时间域,得到去噪

后的时间序列 x_f 。

$$x_f[t] = \frac{1}{T} \sum_{k=0}^{\lfloor T/2 \rfloor} X'_f[k] e^{2\pi i \frac{kt}{T}}, t = 0, \dots, T-1 \quad (3)$$

在获得去噪后的时间序列 x_f 后,通过实施平滑技术来进一步提升质量,减轻频域截断可能带来的边缘效应。具体而言,定义一个钟形汉明窗,大小为 w (w 为偶数),窗函数为:

$$w[n] = 0.54 - 0.46 \cos\left(\frac{2\pi n}{w}\right), n = 1, \dots, w \quad (4)$$

式中: n 为窗口内的样本点的索引。

为了确保窗口大小 w 与序列长度一致,反射性地填充 x_f 的值:

$$x'_f[t] = \begin{cases} x_f\left[\frac{w}{2} - n\right], & 1 \leq n \leq \frac{w}{2} \\ x_f[t], & \frac{w}{2} < t < T - \frac{w}{2} \\ x_f\left[T - 1 - \left(n - \left(T - \frac{w}{2}\right)\right)\right], & T - \frac{w}{2} \leq n \leq T + \frac{w}{2} \end{cases} \quad (5)$$

最后,利用汉明窗进行卷积运算来平滑序列,得到最终去噪和平滑的时间序列 x_s 。

$$x_s[t] = \frac{\sum_{n=1}^w x'_f[t+n-1] \cdot w[n]}{\sum_{n=1}^w w[n]}, t = 1, \dots, T \quad (6)$$

式中:平滑后的序列 $x_s[t]$ 是每个时间点 t 对序列取加权平均值; T 表示序列总长度。

对于多变量时间序列 $X \in \mathbb{R}^{T \times V}$ (包含 V 个变量),分别对每个变量应用 SBT,得到去噪且平滑的序列 $X_s \in \mathbb{R}^{T \times V}$ 。该过程有效实现了去噪和平滑,为后续基于 Transformer 的处理提供了高质量的输入。

2.4 对抗性双分支注意力

增强注意力分配对提升多变量时间序列的预测准确性至关重要,尤其是在处理包含大量变量的数据集时效果更好。本文提出一种对抗性双分支注意力机制,旨在优化注意力权重的分布并优先考虑关键变量间的相关性,解决 Transformer 模型中的注意力分布不均问题,如图 2 所示。

具体而言,本文提出一种双分支架构,将标准注意力分支与对抗性增强分支相结合,并进行可学习的融合以优化注意力聚焦。首先,描述标准分支,该分支基于缩放的点积相似度计算注意力分数;其次,详细描述对抗分支,该分支通过扰动和归一化来细化特征表示。对于一组查询、键 $K \in \mathbb{R}^{B \times S \times H \times E}$ 和值 $V \in \mathbb{R}^{B \times S \times H \times D}$ 而言,标准分支 $S_1 = Q \cdot K^T \in \mathbb{R}^{B \times H \times L \times S}$, S_1 表示注意力分数, B 为批量大小, L , S 为序列长度, H 为头的数量, E 为嵌入维度; D 为值维度。

式(6)中的分数通过因子 $scale = \frac{1}{\sqrt{E}}$ 进行缩放,并使

用 softmax 函数进行归一化,得到注意力权重:

$$A_1 = \text{softmax}(scale \cdot S_1, dim = -1) \quad (7)$$

由此可得,标准分支的输出为:

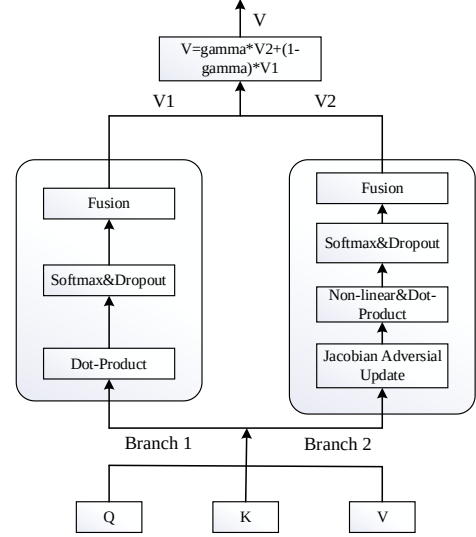


Fig. 2 ADBA attention map

图2 ADBA 注意力图

$$V_1 = A_1 \cdot V \in \mathbb{R}^{B \times L \times H \times D} \quad (8)$$

为了增强注意力分配,在对抗性分支中引入扰动来优化 Q 、 K 。具体为,定义一个相似度函数 $f(Q, K) = \Sigma(Q \cdot K^T)$,并计算其关于 Q 、 K 的雅可比矩阵。

$$J_Q = \frac{\partial f}{\partial Q}, J_K = \frac{\partial f}{\partial K} \quad (9)$$

本文使用可学习参数 $\alpha \in \mathbb{R}$ (初始化为一个小的随机标量),更新查询和键向量:

$$Q_{adv} = Q + \alpha \cdot J_Q, K_{adv} = K + \alpha \cdot J_K \quad (10)$$

为了增强特征显著性,还应用了非线性变换。

$$Q' = \text{sign}(Q_{adv}) \cdot \sqrt{|Q_{adv}| + \epsilon}, K' = \text{sign}(K_{adv}) \cdot \sqrt{|K_{adv}| + \epsilon} \quad (11)$$

式中: $\epsilon = 10^{-6}$ 确保数值稳定性。

随后,将变换后的向量归一化为单位长度:

$$Q_{adv} = \frac{Q'}{\|Q'\|_2}, K_{adv} = \frac{K'}{\|K'\|_2} \quad (12)$$

再经过缩放和标准化得到注意力权重。

$$A_2 = \text{softmax}(scale \cdot S_2, dim = -1) \quad (13)$$

对抗分支的输出为:

$$V_2 = A_2 \cdot V \in \mathbb{R}^{B \times L \times H \times D} \quad (14)$$

最后,使用可学习的权重 $\gamma \in [0, 1]$ 融合两个分支的输出。

$$V = \gamma \cdot V_2 + (1 - \gamma) \cdot V_1 \quad (15)$$

对于包含 V 个变量的多变量时间序列,ADBA 可独立应用于各个头和变量,产生更清晰的注意力分布,突出关键的变量间模式。

这种双分支设计有效缓解了 vanilla Transformers 中注意力分布的均匀性,使 ASformer 能捕捉复杂的多变量相关性,提升了模型的预测性能。

3 实验结果与分析

3.1 参数设置

3.1.1 数据集

为了全面评估所提方法的有效性,选用 5 个广泛使用的多变量时间序列预测标准数据集 ETTh2、Weather、ECL、Exchange 和 Traffic,涵盖电力、气象、能源消耗、汇率和交通流量等多个实际应用场景,具有变量数量多、序列长度长等特点。

ETTh2 来自 Electricity Transformer Temperature (ETT) 数据集的子集,记录了变压器在不同时间段的温度等相关指标,采样频率为一次/h,用于评估模型在电力系统监测中的时序预测能力,变量数量(维度)为 7。Weather 为收集自美国某地区的气象观测数据,包含 21 个气象变量(温度、湿度、风速等)的连续观测值,适用于复杂环境下的多变量预测任务,变量数量(维度)为 21。Electricity Consuming Load (ECL)数据集收录了英国某地区 321 个客户的用电负荷数据,采样周期为一次/h,广泛用于电力负荷预测研究,变量数量(维度)为 321。Exchange 包含 8 种主要货币兑美元的汇率数据,采样频率为一次/日,适用于金融领域的多变量时间序列建模与预测,变量数量(维度)为 8。Traffic 数据集由美国洛杉矶高速公路系统下的交通流量监测点采集而成,包含 862 个传感器的道路交通流量数据,采样频率为一次/h,是智能交通管理中常用的时序预测数据集,变量数量(维度)为 862。

本文将上述数据集按照常规方法划分为训练集、验证集和测试集,具体比例及预处理方式与主流文献保持一

致,以确保实验结果的公平性和可复现性。

3.1.2 基线模型

将 ASformer 与 6 个先进模型进行比较,包括:①基于 Transformer 的模型 iTransformer、PatchTST、FEDformer 和 Autoformer;②基于 TCN 的模型 TimesNet^[25];③基于 MLP 的模型 DLinear^[26]。

3.1.3 评估指标

本文采用两种广泛使用的评估指标,包括均方误差 (Mean Squared Error, MSE)、平均绝对误差 (Mean Absolute Error, MAE)量化所有比较方法的预测准确性。

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \tag{16}$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \tag{17}$$

3.1.3 实验细节

本文在单个 NVIDIA RTX4090 (24 GB) 上基于 PyTorch 进行实验,使用 ADAM 优化器优化 L2 损失,并从 $\{1 \times 10^{-3}, 5 \times 10^{-4}, 1 \times 10^{-4}\}$ 中选择初始学习率,通过网格搜索策略得出最优超参数 f_low、f_high。

3.2 比较实验

为了验证所提 ASformer 模型的有效性,在 5 个标准多变量时间序列预测数据集 (ETTh2、Weather、ECL、Exchange 和 Traffic) 上与多种主流 Transformer 系列方法进行实验,主要包括 iTransformer、DLinear、PatchTST、TimesNet、FEDformer 和 Autoformer。实验采用 MSE 与 MAE 作为评价指标,数值越低代表模型预测性能越好,预测步长分别为 96、192、336 和 720。表 1 为各模型在不同数据集和预测步长下的实验结果。

Table 1 Comparison results of different prediction step sizes with the current advanced baseline models on five publicly available datasets

表 1 在五个公开数据集上与当前先进基线模型不同预测步长的比较结果

模型		ASformer		iTransformer		Dlinear		PatchTST		TimesNet		FEDformer		Autoformer	
数据集	步长	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh2	96	0.297	0.346	0.297	0.349	0.333	0.387	0.302	0.348	0.340	0.374	0.358	0.397	0.346	0.388
	192	0.376	0.394	0.380	0.400	0.477	0.476	0.388	0.400	0.402	0.414	0.429	0.439	0.456	0.452
	336	0.422	0.429	0.428	0.432	0.594	0.541	0.426	0.433	0.452	0.452	0.496	0.487	0.482	0.486
	720	0.426	0.442	0.427	0.445	0.831	0.657	0.431	0.446	0.462	0.468	0.463	0.474	0.515	0.511
Weather	96	0.173	0.212	0.174	0.214	0.196	0.255	0.177	0.218	0.172	0.220	0.217	0.296	0.266	0.336
	192	0.221	0.254	0.221	0.254	0.237	0.296	0.225	0.259	0.219	0.261	0.276	0.336	0.307	0.367
	336	0.278	0.296	0.278	0.296	0.283	0.335	0.278	0.297	0.280	0.306	0.339	0.380	0.359	0.395
	720	0.357	0.348	0.358	0.349	0.345	0.381	0.354	0.348	0.365	0.359	0.403	0.428	0.419	0.428
ECL	96	0.148	0.239	0.148	0.240	0.197	0.282	0.181	0.270	0.168	0.272	0.193	0.308	0.201	0.317
	192	0.163	0.252	0.162	0.253	0.196	0.285	0.188	0.274	0.184	0.289	0.201	0.315	0.222	0.334
	336	0.178	0.268	0.178	0.269	0.209	0.301	0.204	0.293	0.198	0.300	0.214	0.329	0.231	0.338
	720	0.208	0.296	0.225	0.317	0.245	0.333	0.246	0.324	0.220	0.320	0.246	0.355	0.254	0.361
Exchange	96	0.092	0.212	0.086	0.206	0.088	0.218	0.088	0.205	0.107	0.234	0.148	0.278	0.197	0.323
	192	0.176	0.300	0.177	0.299	0.176	0.315	0.176	0.299	0.226	0.344	0.271	0.315	0.300	0.369
	336	0.322	0.411	0.331	0.417	0.313	0.427	0.301	0.397	0.367	0.448	0.460	0.427	0.509	0.524
	720	0.841	0.690	0.847	0.691	0.839	0.695	0.901	0.714	0.964	0.746	1.195	0.695	1.447	0.941
Traffic	96	0.396	0.268	0.395	0.268	0.650	0.396	0.462	0.295	0.593	0.321	0.587	0.366	0.613	0.388
	192	0.409	0.276	0.417	0.276	0.598	0.370	0.466	0.296	0.617	0.336	0.604	0.373	0.616	0.382
	336	0.423	0.282	0.433	0.283	0.605	0.373	0.482	0.304	0.629	0.336	0.621	0.383	0.622	0.337
	720	0.454	0.297	0.467	0.302	0.645	0.394	0.514	0.322	0.640	0.350	0.626	0.382	0.660	0.408

表2为所有预测步长的平均实验结果(加粗斜体表示最佳)。由此可知,本文模型在所有5个基准测试中均取得了最优的综合表现,无论是在ETTh2、Weather、ECL还是Traffic数据集上,本文模型的平均MSE、MAE最低,预测精度卓越、泛化能力强,相较于各类先进的基线模型性能更优。此外,ASformer模型在长时序预测任务中表现较好,当预测步长增加时,许多模型的性能会显著下降,但本文模型仍能保持高精度。例如,当最大预测步长设置为720时,本文模型在ETTh2、ECL和Traffic数据集上的MSE、MAE最低,能有效捕捉并外推序列的长期依赖关系,缓解

长时序预测中的精度衰减问题。同时,所提模型相较于iTransformer、PatchTST等近期表现优异的Transformer变体模型也具明显优势。在ECL数据集上,当预测步长为720时,本文模型的MSE(0.208)、MAE(0.296)相较于次优的iTransformer(MSE为0.225,MAE为0.317)分别降低7.6%、6.6%。在ETTh2数据集上,当预测步长为336时,本文模型相较于表现不俗的PatchTST的MSE、MAE下降明显。数据表明,本文模型在关键性能指标上优于当前的主流方法,充分证明了SBT、ADBA模块在提升预测准确性方面的有效性和通用性。

Table 2 Comparison results of different prediction step sizes with the current advanced baseline models on 5 public datasets

表2 在5个公开数据集上与当前先进基线模型不同预测步长的比较结果

模型		ASformer		iTransformer		Dlinear		PatchTST		TimesNet		FEDformerr		Autoformer	
数据集	步长	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh2	AVG	0.380	0.402	0.383	0.407	0.559	0.515	0.387	0.407	0.414	0.427	0.437	0.449	0.450	0.459
Weather	AVG	0.257	0.277	0.258	0.278	0.260	0.317	0.259	0.281	0.259	0.287	0.309	0.360	0.338	0.382
ECL	AVG	0.174	0.263	0.178	0.270	0.212	0.300	0.205	0.290	0.192	0.295	0.214	0.327	0.227	0.338
Exchange	AVG	0.357	0.403	0.360	0.403	0.354	0.414	0.367	0.404	0.416	0.443	0.519	0.429	0.613	0.539
Traffic	AVG	0.420	0.280	0.428	0.282	0.625	0.383	0.481	0.304	0.620	0.336	0.610	0.376	0.628	0.379

分析发现,本文模型相较于基线模型效果表现最好的原因为:①SBT模块通过频域带通滤波和汉明窗卷积反射填充有效实现了去噪和平滑,为后续基于Transformer的处理提供了高质量的输入,减少了噪声干扰;②ADBA模块将标准注意力与对抗性增强注意力相结合,聚焦于关键的变量间依赖关系,使模型能更好地捕捉关键变量间的特征依赖关系;③SBT、ADBA协同优化了从数据预处理到注意力分配的全流程,且对抗性机制和参数可调性提升了模型在复杂环境中的适应性,鲁棒性和泛化能力更强。

3.3 消融实验

本文在Weather、ETTh2、Traffic数据集上进行了消融实验,从3个角度深度验证SBT、ADBA模块的不可或缺性:①Proposed代表完整的ASformer模型;②wo/ADBA代表将ASformer中的对抗性双分支注意力替换为普通自注意力机制;③wo/SBT代表移除频谱带通变换模块的模型。实验结果如表3所示。由此可知,移除任一模块组件都会导致性能下降(MSE、MAE指标升高)。具体为,缺少SBT模块会使模型更容易受到环境噪声或白噪声干扰,难以准确捕捉时间序列数据中的时序模式,对预测精度产生负面影响;使用普通自注意力替代ADBA模块也会导致预测误差增大,表明在处理具有众多变量的多元时间序列建模时,普通自注意力机制的效果不如对抗性双分支注意力机制。综上,SBT、ADBA模块可共同提升模型在时间序列预测任务中的性能表现。

4 结语

本文为解决多变量时间序列建模中注意力权重分布分散的难题,提出一种基于Transformer的创新架构ASformer。首先提出SBT模块,利用频域滤波和时间平滑技

Table 3 Ablation experiment results

表3 消融实验结果

模型		Proposed		wo / ADBA		wo / SBT	
数据集	步长	MSE	MAE	MSE	MAE	MSE	MAE
ETTh2	96	0.297	0.346	0.300	0.347	0.298	0.349
	192	0.376	0.394	0.377	0.393	0.379	0.399
	336	0.422	0.429	0.422	0.428	0.425	0.434
	720	0.426	0.442	0.428	0.441	0.431	0.448
	AVG	0.380	0.402	0.381	0.402	0.383	0.407
Weather	96	0.173	0.212	0.181	0.221	0.173	0.212
	192	0.221	0.254	0.227	0.259	0.222	0.255
	336	0.278	0.296	0.280	0.298	0.278	0.296
	720	0.357	0.348	0.357	0.349	0.347	0.348
	AVG	0.257	0.277	0.261	0.281	0.257	0.278
Traffic	96	0.396	0.268	0.394	0.270	0.396	0.268
	192	0.409	0.276	0.413	0.277	0.409	0.276
	336	0.423	0.282	0.424	0.282	0.422	0.282
	720	0.454	0.297	0.459	0.300	0.456	0.299
	AVG	0.420	0.280	0.423	0.282	0.421	0.281

术去噪并增强时间序列数据的平滑度;其次提出ADBA模块,将标准注意力机制与对抗分支相结合,以优化注意力权重分布,使模型优先考虑信息量最大的变量及其相关性。

在ETTh2、Weather、ECL、Exchange、Traffic等5种数据集上的实验表明,ASformer在效率和准确性方面优势明显,SBT、ADBA模块共同增强了模型捕捉、利用复杂多变量依赖关系的能力,尤其是在长期预测场景中时间序列任务非常有效。消融实验表明,SBT、ADBA模块在模型中不可或缺,对提升模型性能具有积极作用。

未来,将专注于扩展ASformer,使其能适应大规模真实数据集,并在SBT中融入自适应频带选择以进一步优化模型去噪能力;或计划推动其投入公共服务,以便在实际预测领域实现更广泛的应用。

参考文献:

- [1] HUANG Y, ZHOU Z, WANG Z, et al. Timesnet-pm2. 5: interpretable timesnet for disentangling intraperiod and interperiod variations in PM2. 5 prediction[J]. Atmosphere, 2023, 14(11): 1604.
- [2] DONG J, WU H, ZHANG H, et al. Simmtm: a simple pre-training framework for masked time-series modeling[J]. Advances in Neural Information Processing Systems, 2023, 36: 29996–30025.
- [3] O'CONNOR B, BALASUBRAMANYAN R, ROUTLEDGE B, et al. From tweets to polls: linking text sentiment to public opinion time series [C]// Proceedings of the International AAAI Conference on Web and Social Media, 2010: 122–129.
- [4] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [DB/OL]. <https://arxiv.org/abs/1706.03762>.
- [5] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877–1901.
- [6] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale [DB/OL]. <https://arxiv.org/abs/2010.11929>.
- [7] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling laws for neural language models [DB/OL]. <https://arxiv.org/abs/2001.08361>.
- [8] WEN Q, ZHOU T, ZHANG C, et al. Transformers in time series: a survey [DB/OL]. <https://arxiv.org/abs/2202.07125>.
- [9] LIU S, YU H, LIAO C, et al. Pyraformer: low-complexity pyramidal attention for long-range time series modeling and forecasting [EB/OL]. <https://iclr.cc/virtual/2022/oral/6828>.
- [10] WU H, XU J, WANG J, et al. Autoformer: decomposition transformers with auto-correlation for long-term series forecasting [J]. Advances in Neural Information Processing Systems, 2021, 34: 22419–22430.
- [11] LIU Y, HU T, ZHANG H, et al. iTransformer: inverted transformers are effective for time series forecasting [DB/OL]. <https://arxiv.org/abs/2310.06625>.
- [12] ZHOU H, ZHANG S, PENG J, et al. Informer: beyond efficient transformer for long sequence time-series forecasting [C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2021: 11106–11115.
- [13] KITAEV N, KAISER Ł, LEVSKAYA A. Reformer: the efficient transformer [DB/OL]. <https://arxiv.org/abs/2001.04451>.
- [14] NIE Y, NGUYEN N H, SINTHONG P, et al. A time series is worth 64 words: Long-term forecasting with transformers [DB/OL]. <https://arxiv.org/abs/2211.14730>.
- [15] ZHANG Y, YAN J. Crossformer: transformer utilizing cross-dimension dependency for multivariate time series forecasting [EB/OL]. <https://openreview.net/pdf?id=vSVLM2j9eie>.
- [16] ZHANG L, AGGARWAL C, QI G J. Stock price prediction via discovering multi-frequency trading patterns [C]// Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017: 2141–2149.
- [17] CAO D, WANG Y, DUAN J, et al. Spectral temporal graph neural network for multivariate time-series forecasting [J]. Advances in Neural Information Processing Systems, 2020, 33: 17766–17778.
- [18] ZHOU T, MA Z, WEN Q, et al. Fedformer: frequency enhanced decomposed transformer for long-term series forecasting [C]// International Conference on Machine Learning, 2022: 27268–27286.
- [19] YI K, ZHANG Q, FAN W, et al. Frequency-domain MLPs are more effective learners in time series forecasting [J]. Advances in Neural Information Processing Systems, 2023, 36: 76656–76679.
- [20] DAS A, KONG W, LEACH A, et al. Long-term forecasting with tide: time-series dense encoder [DB/OL]. <https://arxiv.org/abs/2304.08424>.
- [21] CHEN Y, REN K, WANG Y, et al. Contiformer: continuous-time transformer for irregular time series modeling [J]. Advances in Neural Information Processing Systems, 2023, 36: 47143–47175.
- [22] MOTTAGHI-KASHTIBAN M, SHAYESTEH M G. New efficient window function, replacement for the hamming window [J]. IET Signal Processing, 2011, 5(5): 499–505.
- [23] WOO G, LIU C, SAHOO D, et al. Cost: contrastive learning of disentangled seasonal-trend representations for time series forecasting [DB/OL]. <https://arxiv.org/abs/2202.01575>.
- [24] YANG Z, YAN W, HUANG X, et al. Adaptive temporal-frequency network for time-series forecasting [J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(4): 1576–1587.
- [25] WU H, HU T, LIU Y, et al. Timesnet: temporal 2D-variation modeling for general time series analysis [EB/OL]. https://openreview.net/pdf?id=ju_Uqw3840q.
- [26] ZENG A, CHEN M, ZHANG L, et al. Are transformers effective for time series forecasting? [C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2023: 11121–11128.
- [27] LIU M, ZENG A, CHEN M, et al. Scinet: time series modeling and forecasting with sample convolution and interaction [J]. Advances in Neural Information Processing Systems, 2022, 35: 5816–5828.
- [28] BAI S, KOLTER J Z, KOLTUN V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling [DB/OL]. <https://arxiv.org/abs/1803.01271>.
- [29] ZHAO Z, CHEN W, WU X, et al. LSTM network: a deep learning approach for short-term traffic forecast [J]. IET Intelligent Transport Systems, 2017, 11(2): 68–75.
- [30] RANGAPURAM S S, SEEGER M W, GASTHAUS J, et al. Deep state space models for time series forecasting [EB/OL]. <https://www.semanticscholar.org/paper/Deep-State-Space-Models-for-Time-Series-Forecasting-Rangapuram-Seeger/ae4df460a413f3b1d9a0dfa47917751af9db2597>.
- [31] SALINAS D, FLUNKERT V, GASTHAUS J, et al. DeepAR: probabilistic forecasting with autoregressive recurrent networks [J]. International Journal of Forecasting, 2020, 36(3): 1181–1191.

(责任编辑:刘嘉文)