基于Transformer的时间序列预测方法综述

陈嘉俊, 刘波, 林伟伟, 郑剑文, 谢家晨

引用本文

陈嘉俊, 刘波, 林伟伟, 郑剑文, 谢家晨. 基于Transformer的时间序列预测方法综述[J]. 计算机科学, 2025, 52(6): 96-105.

CHEN Jiajun, LIU Bo, LIN Weiwei, ZHENG Jianwen, XIE Jiachen. Survey of Transformer-based Time Series Forecasting Methods [J]. Computer Science, 2025, 52(6): 96-105.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240500043

基于 Transformer 的时间序列预测方法综述

陈嘉俊[1]　刘　波[1,3]　林伟伟[2]　郑剑文[3]　谢家晨[3]
1 华南师范大学人工智能学院　广州 510631
2 华南理工大学计算机科学与工程学院　广州 510640
3 华南师范大学计算机学院　广州 510631
(1046696528@qq.com)

摘　要　时间序列预测作为分析历史数据以预测未来趋势的关键技术,已广泛应用于金融、气象等领域。然而,传统方法如自回归移动平均模型和指数平滑法等在处理非线性模式、捕捉长期依赖性时存在局限。最近,基于 Transformer 的方法因其自注意力机制,在自然语言处理与计算机视觉领域取得突破,也开始拓展至时间序列预测领域并取得显著成果。因此,探究如何将 Transformer 高效运用于时间序列预测,成为推动该领域发展的关键。首先,介绍了时间序列的特性,阐述了时间序列预测的常见任务类别及评估指标。接着,深入解析 Transformer 的基本架构,并挑选了近年来在时间序列预测中广受关注的 Transformer 衍生模型,从模块及架构层面进行分类,并分别从问题解决、创新点及局限性 3 个维度进行比较和分析。最后,进一步探讨了时间序列预测 Transformer 在未来可能的研究方向。

关键词:时间序列;Transformer 模型;深度学习;注意力机制;预测
中图分类号　TP391

Survey of Transformer-based Time Series Forecasting Methods

CHEN Jiajun[1],LIU Bo[1,3],LIN Weiwei[2],ZHENG Jianwen[3] and XIE Jiachen[3]
1 School of Artificial Intelligence,South China Normal University,Guangzhou 510631,China
2 School of Computer Science and Engineering,South China University of Technology,Guangzhou 510640,China
3 School of Computer Science,South China Normal University,Guangzhou 510631,China

Abstract　Time series forecasting,a critical technique for analyzing historical data to predict future trends,has been widely applied in fields such as finance and meteorology. However,traditional methods like the autoregressive moving average model and exponential smoothing face limitations when dealing with nonlinear patterns and capturing long-term dependencies. Recently,Transformer-based approaches,due to their self-attention mechanism,have achieved breakthroughs in natural language processing and computer vision,and have also shown significant promise in time series forecasting. Therefore,exploring how to efficiently apply Transformers to time series prediction has become crucial for advancing this field. This paper first introduces the characteristics of time series data and explains the common task categories and evaluation metrics for time series forecasting. It then delves into the basic architecture of the Transformer model and selects Transformer-derived models that have garnered widespread attention in recent years for time series forecasting. These models are categorized based on their modules and architectures,and are compared and analyzed from three perspectives:problem-solving capabilities,innovations,and limitations. Finally,this paper discusses potential future research directions for the application of Transformers in time series forecasting.

Keywords　Time series,Transformer model,Deep learning,Attention mechanism,Prediction

1　引言

时间序列是一连串按时间顺序排列的数据点集[1],其本质上是一组按时序整合的观测值,这些值可以根据不同的时间间隔获得,如秒、小时、天或月。无论在自然界还是人类社会中,时间序列数据无处不在,其在金融市场股价波动[2]、交通流量预测[3]、疾病传播预警[4]以及能源负荷预测[5]等方面都有所体现。这类数据对于揭示过去模式、诊断当前状态以及预测未来趋势发挥着不可估量的作用。

随着时间序列数据在各领域中被广泛使用,时间序列预

到稿日期:2024-05-10　返修日期:2024-10-13
基金项目:国家自然科学基金面上项目(62072187);广州市开发区国际合作项目(2023GH02)
This work was supported by the National Natural Science Foundation of China(62072187) and Guangzhou Development Zone Science and Technology Project(2023GH02).
通信作者:林伟伟(linww@scut.edu.cn)

测已经成为数据分析和决策制定的重要工具。传统的时间序列预测方法，如自回归移动平均模型（ARMA）[6]和自回归积分移动平均模型（ARIMA）[7]等，已经取得了一定的成果[8]。然而，由于长期依赖建模、非线性关系捕获和多变量关联等问题，这些方法无法处理大量复杂的时间序列数据[9]。

最近，基于 Transformer 模型的时间序列预测方法引起了广泛关注。该模型由谷歌大脑团队 Vaswani 等[10]于 2017 年提出，主要用于自然语言处理[11-12]和计算机视觉领域[13-14]。凭借独特的自注意力机制和并行计算能力，Transformer 在处理序列数据时表现出色。与传统方法相比，基于 Transformer 的模型具备以下优势：首先，该模型能够捕捉时间序列数据中的长程依赖关系，从而更好地建模序列中的趋势和周期性；其次，它能够处理多变量时间序列，从而更好地理解不同变量之间的关系；最后，该模型通过自注意力机制对序列中的关键信息进行加权，从而更准确地预测未来趋势和变化。

现有的时间序列预测相关综述文献主要聚焦于对传统方法、机器学习模型和深度学习模型的分析[15-16]，缺少对 Transformer 类模型以及预测任务本身的深入剖析。为了填补这一空缺，本文从分析时间序列预测数据出发，详细介绍了多种时间序列预测任务。同时，本文创新性地从模块和架构的维度对现有的 Transformer 类模型进行分类，并进行了对比分析和总结。

本文第 2 章对时间序列预测进行概述；第 3 章详细介绍 Transformer 模型的基本原理和核心组件，并分析和比较了基于 Transformer 的衍生模型在时间序列预测任务中所解决的问题、创新点和局限性；第 4 章对基于 Transformer 的方法在时间序列预测领域的未来研究方向进行展望。

2 时间序列预测概述

2.1 时间序列数据的特性

时间序列数据是数据点按照时间顺序排列，每个数据点都与一个或多个成对的时间戳相关联。在时间序列预测中，通过这些按时间顺序记录的数据来预测未来的事件。时间序列预测的概念可表示为：

$$\hat{X}_{t+h|t}=f(X_t,X_{t-1},X_{t-2},\cdots,X_{t-n})+\epsilon \tag{1}$$

其中，$\hat{X}_{t+h|t}$代表在时间点 t 的信息基础上预测未来时间点 $t+h$ 的值；$X_t,X_{t-1},X_{t-2},\cdots,X_{t-n}$表示历史数据点；$f$ 是模型的预测函数；ϵ表示预测误差，通常假设为白噪声。

1）趋势性：时间序列数据通常显示出某种趋势，表明数据随时间推移呈现出上升、下降或保持稳定的模式。趋势性反映了数据随时间变化的长期方向，这种变化可以是线性或非线性的。

2）季节性：季节性是指时间序列数据在固定周期内重复出现的模式。这种周期性波动通常与特定的季节、月份、一周内的某天或一天内的某个小时相关。

3）周期性：许多时间序列具有重复的周期特征。周期性是指数据在季节性模式之外的波动，这些波动通常反映了长期的循环变化，且可能不受季节因素的直接影响。

4）随机性：时间序列数据中经常包含一些不规则的变化，这些变化称为随机性或噪声。随机性是数据中无法预测或无法解释的变化，反映了数据集中的随机波动或异常值。

5）平稳性：平稳性是指时间序列数据的统计属性（如均值、方差）在时间上保持不变。非平稳性的时间序列在进行分析和建模时，需要通过差分、对数转换等方法转化为平稳序列。

2.2 时间序列预测任务的分类

随着数据时代的发展，时间序列预测方法的研究虽然取得了重大成果，但仍然存在很多问题。本文将从不同的时间序列预测任务出发，分析这些任务所面临的问题。

2.2.1 单变量与多变量时间序列预测

单变量时间序列预测是指仅使用单一时间序列变量的数据来进行未来趋势的预测。例如，可以利用过去的股票价格预测未来价格，或根据历史气温预测未来的气温。单变量时间序列预测的主要挑战包括数据的非平稳性、季节性与周期性成分的干扰、长程依赖关系的捕捉，以及异常值和噪声的存在。非平稳性会使模型难以准确捕捉数据的趋势和模式；而季节性与周期性成分可能对预测结果造成干扰；传统模型在面对长程依赖关系时表现较为乏力，尤其当数据量有限时更是如此；此外，异常值和噪声的存在，也会影响模型的训练效果和预测精度。

多变量时间序列预测指利用多个相关的时间序列变量来进行未来趋势的预测。例如，使用股票价格、交易量和市场指数等多种数据来预测股票价格，或使用气温、湿度和风速等多种气象数据来预测未来的气温。多变量时间序列预测面临的主要困难包括变量之间的相互依赖关系、高维数据处理、数据同步问题和特征选择[17]。多变量时间序列数据中的各变量之间可能存在复杂的相互依赖关系，如何有效建模这些关系是一个关键问题。高维数据增加了模型的复杂度和计算负担，可能导致模型过拟合或计算资源不足。各变量可能具有不同的采样频率或时间戳，如何进行数据同步和对齐是一个挑战。在多变量时间序列预测中，选择合适的特征至关重要，过多的无关或冗余特征会增加模型的复杂度，而缺少关键特征则会降低预测效果[18]。

2.2.2 短期与长期时间序列预测

短期时间序列预测是指对未来较短时间范围内的数据进行预测，例如预测未来几小时的电力负荷或未来几天的股票价格。短期预测通常依赖于近期数据，因此能够较好地捕捉到数据中的短期模式和趋势[19]。然而，短期预测也面临一些挑战。首先，数据中的噪声和随机波动可能会对预测结果产生较大影响，因为模型容易对这些短期波动产生过拟合。其次，短期预测需要处理高频数据，这对数据采集和处理提出了更高的要求。此外，短期预测模型需要具备快速响应能力和高实时性，这对计算资源和算法效率也构成了挑战。

长期时间序列预测是指对未来较长时间范围内的数据进行预测。例如，预测未来几个月的气温变化或未来几年的经济增长趋势。长期预测需要捕捉时间序列中的长期模式和趋势，因此通常需要更复杂的模型和更多的历史数据。长期时间序列预测面临的主要困难包括数据的非平稳性、长期依赖

关系的捕捉以及外部因素的影响[20]。长期预测中，数据的非平稳性会导致模型难以准确捕捉长期趋势和变化。长期依赖关系难以被传统模型捕捉，尤其是在数据量较少的情况下。此外，长期预测容易受到外部因素（如政策变化、市场环境变化等）的影响，这些因素难以在模型中准确量化和预测。

2.2.3 平稳与非平稳时间序列预测

平稳时间序列预测是指对平稳时间序列数据进行预测。平稳时间序列的均值和方差在时间上保持恒定，且自相关结构不随时间变化。由于平稳时间序列的统计特性相对稳定，因此预测模型可以更容易地捕捉其规律性和周期性。例如，某些季节性调整后的经济指标或经过差分处理后的股票价格序列即为平稳时间序列。平稳时间序列预测的主要挑战在于如何准确建模其自相关结构和周期性成分。尽管平稳时间序列的特性较为稳定，但在应用中，实际数据仍可能受到随机波动和噪声的影响。此外，模型选择和参数估计的准确性也对预测效果有重要影响。

非平稳时间序列预测是指对非平稳时间序列数据进行预测。非平稳时间序列的均值和方差随时间变化，其自相关结构也可能随时间变化[21]。非平稳时间序列广泛存在于许多实际应用中，如股票价格、气温变化等。非平稳时间序列预测面临的主要困难在于数据的非平稳性和复杂的依赖结构[22]。首先，非平稳时间序列的数据特性随时间变化，因此传统的平稳模型难以适用。其次，非平稳时间序列可能包含趋势、季节性和周期性成分，这些成分需要通过适当的方法进行分解和处理。此外，非平稳时间序列中的突发事件和异常值也会对预测模型产生较大影响。

2.3 评价指标

在对时间序列预测模型进行评估时，研究者通常会使用多种性能衡量标准。本节将对这些评价指标的定义、重要性以及计算方法进行全面概述。

1）均方误差[23]（Mean Squared Error，MSE）是最常见的评价指标之一，其通过计算实际值与预测值之间差值的平方的平均值来量化预测的精度。模型的预测能力越强，即预测结果与实际结果的偏差越小，其 MSE 值就越小。因此，MSE 可以反映出预测值的变化幅度和稳定性。

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

2）平均绝对误差[24]（Mean Absolute Error，MAE）与 MSE 有所不同，该指标提供了预测误差的直接度量，是预测值与实际值之间的差值的平均绝对值。通常认为，MAE 比 MSE 更能直观地反映出模型的预测性能。

$$MAE=\frac{1}{N}\sum_{i=1}^{N}|y_i-\hat{y}_i| \tag{3}$$

3）均方根误差[25]（Root Mean Square Error，RMSE）通过计算预测值与实际值之差的平方的均值后进行平方根运算，对模型的预测准确性进行评估。RMSE 更强调大误差的影响，因此在对预测精度有较高要求的评估模型性能场景下尤其适用。

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2} \tag{4}$$

其中，N 表示观测点的数量，y_i 是第 i 个观察点的真实值，$\hat{y}_i$ 是第 i 个观察点的预测值。

2.4 传统预测方法与机器学习方法

传统的时间序列预测方法，如 ARMA[6] 和 ARIMA[7]，通过分析历史数据中的模式来预测未来。其中，ARIMA 模型还通过差分操作使非平稳时间序列转为平稳。相较之下，机器学习方法，如支持向量机[26] 和随机森林[27]，能够处理更复杂的数据结构和更大的数据集，通过构建多个决策树等方式提高预测的准确性和鲁棒性。

2.5 深度学习方法

与传统算法相比，深度学习方法在时间序列预测任务中表现出了更强大的性能，得到了长远发展和普遍应用[28]。与浅层神经网络相比，深度神经网络有更好的线性和非线性特征提取能力，能够挖掘出浅层神经网络容易忽略的规律，最终满足高精度的预测任务要求。

2.5.1 RNN

循环神经网络[29]（RNN）是处理序列数据的经典深度学习方法之一，能够捕捉时间序列数据中的时间依赖关系。其隐层单元之间的连接形成了一个循环，因此可以保留前一时刻的信息并影响后一时刻的输出。然而，RNN 在处理长序列时容易出现梯度消失或梯度爆炸问题。Lin 等[30] 提出的一种基于 RNN 的长期时间序列预测模型 SegRNN，通过分段迭代和并行多步预测技术有效减少循环迭代次数，从而解决了传统 RNN 在长序列预测任务中的性能和效率低下的问题。

2.5.2 CNN

卷积神经网络[31]（CNN）虽然最初用于图像处理，但在时间序列预测中也有广泛应用。CNN 通过卷积层提取数据的局部特征，然后通过池化层进行降维，最后通过全连接层进行预测。其卷积操作能够有效捕捉时间序列中的局部特征。CNN 通过卷积核在输入数据上滑动，提取局部区域的特征。池化层则通过下采样减少特征图的维度，保留重要特征。最终的全连接层将提取的特征用于预测。Livieris 等[32] 提出了 CNN-LSTM[33] 模型用于黄金价格预测，卷积层用于提取时间序列中的有用特征并学习其内部表示，而 LSTM 层则用于捕捉时间序列中的长期依赖关系。

2.5.3 Transformer

随着 Transformer[10] 在自然语言处理领域的成功，其应用范围逐渐扩展到时间序列预测等其他领域。时间序列预测任务中，数据通常具有高度的时间依赖性和复杂的模式。Transformer 在这一领域的过渡，依赖于其自注意力机制，其能够有效地捕捉时间序列中的长期依赖和季节性变化，超越了传统的 RNN[29] 和 LSTM[33] 的表现。此外，通过引入位置编码，Transformer 能够处理时间序列的顺序信息，进一步提升了模型的预测能力。

3 Transformer 在时间序列预测领域的研究现状

3.1 Transformer 模型概述

Transformer[10] 是一种基于注意力机制的网络架构，完全不依赖于循环神经网络或卷积神经网络[34]。该架构主要由

编码器（Encoder）和解码器（Decoder）两部分组成，如图 1 所示。

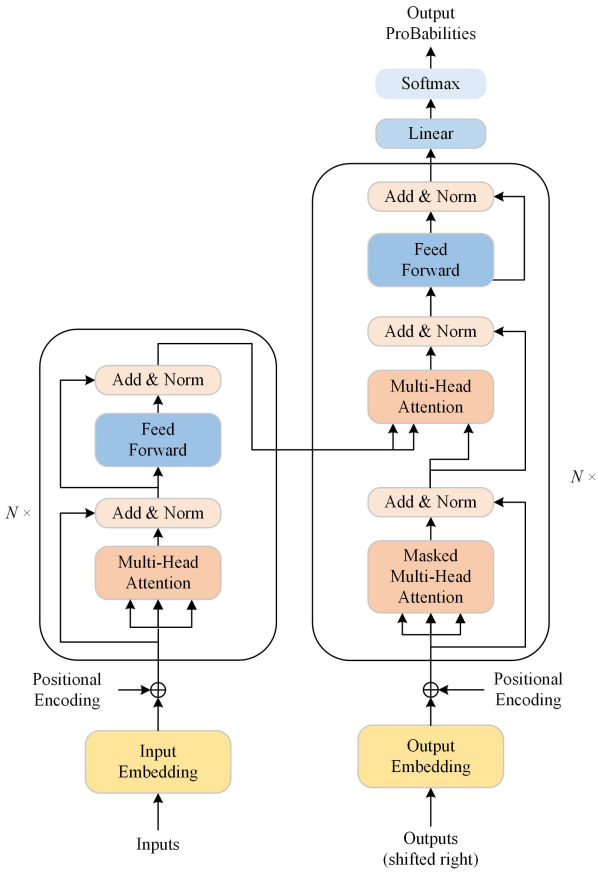

图 1 Transformer 模型结构

Fig. 1 Transformer model structure

3.1.1 编码器

编码器包含多层，每层主要由自注意力机制和前馈神经网络两部分构成。自注意力机制用于捕获输入序列的全局关系，确保每个位置的信息能全面反映整个序列内容；前馈神经网络对这些信息进行处理。为了确保深层网络训练的稳定性，引入了残差连接和层归一化，有效防止了梯度消失问题。编码器从符号化的输入序列开始，输出高维连续表示序列，为后续的解码阶段提供支持。

3.1.2 解码器

解码器负责将编码器的信息转化为目标序列，其包括若干层，每层有 3 个核心机制：自注意力、编码器-解码器注意力机制和前馈神经网络。自注意力机制保证解码依赖前一步的输出，防止信息泄露；编码器-解码器注意力机制捕捉编码器输出与当前解码器输出的关系，确保基于上下文生成正确符号；前馈神经网络进一步处理信息。这些机制相互合作，逐步构建最终序列，保证了输出的连贯性。

3.1.3 多头注意力机制

多头注意力机制通过将 Query，Key 和 Value 分割成多个“头”进行并行处理，能够在不同的表示子空间中综合不同的特征和上下文信息[35]，从而提升模型对数据的理解深度，加强其在捕捉复杂和多样化依赖关系上的能力，并增强模型在序列预测和语言理解等多种任务中的性能和泛化能力。这种机制在自注意力机制的基础上发展而来。多头注意力计算式如下：

$$MultiHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Concat(\boldsymbol{head}_1,\cdots,\boldsymbol{head}_H)\boldsymbol{W}_O \tag{5}$$

$$\boldsymbol{head}_i=Attention(\boldsymbol{QW}_i^Q,\boldsymbol{KW}_i^K,\boldsymbol{VW}_i^V) \tag{6}$$

其中，$\boldsymbol{W}_O$ 表示可学习的输出投影矩阵，每一个头 $head_i$ 的 *Attention* 计算方式为：

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{QK}^{\mathrm{T}}}{\sqrt{D_k}}\right)\boldsymbol{V} \tag{7}$$

其中，$\boldsymbol{W}_i^Q$，$\boldsymbol{W}_i^K$，$\boldsymbol{W}_i^V$ 分别是 Query，Key 和 Value 的投影矩阵，D_k 表示键的维度。

多头注意力结构如图 2 所示。

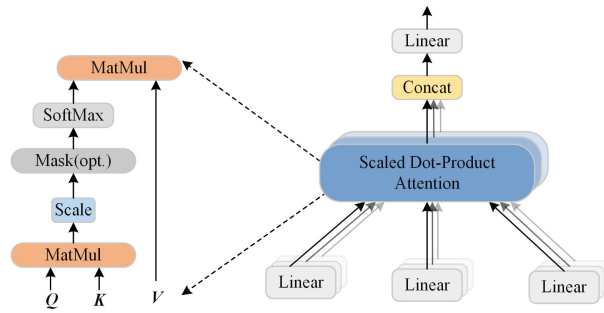

图 2 多头注意力结构

Fig. 2 Multi-head attention structure

3.1.4 前馈和残差网络

前馈网络（Feed-Forward Network，FFN）是一种经典的深度学习模型中的基本组成单元，其通过完全连接的方式实现对数据的传递和处理。在深度学习中，前馈网络的表达式可以被定义为：

$$FFN(\boldsymbol{H}_0)=\mathrm{ReLU}(\boldsymbol{H}_0\boldsymbol{W}_1+\boldsymbol{b}_1)\boldsymbol{W}_2+\boldsymbol{b}_2 \tag{8}$$

其中，$\boldsymbol{H}_0$ 表示前一层的输出结果，$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 是权重矩阵，$\boldsymbol{b}_1$ 和 $\boldsymbol{b}_2$ 是偏置项，这些都是模型中需要通过学习得到的可训练参数。在构建复杂的模型如 Transformer 模型时，通常会在每个模块周围添加残差连接（Residual Connection）和层归一化（Layer Normalization），以增强模型的表达能力并帮助模型更好地收敛。具体来说，这种结构的描述如下：

$$\boldsymbol{H}_0=LayerNorm(SelfAttn(\boldsymbol{X})+\boldsymbol{X}) \tag{9}$$

$$\boldsymbol{H}=LayerNorm(FFN(\boldsymbol{H}_0)+\boldsymbol{H}_0) \tag{10}$$

其中，*SelfAttn* 表示自注意力模块，其是处理序列数据特别有效的一种机制，能够使得模型在处理每个元素时考虑到序列中的其他元素。*LayerNorm* 代表层归一化操作，通过对每一层的输出进行标准化处理，帮助缓解深层网络中的梯度消失或爆炸问题。

3.2 Transformer 在时间序列预测任务中的有效性

3.2.1 捕捉长程依赖关系

时间序列数据常常包含长时间跨度的依赖关系，如某些经济指标的变化可能受到几个月甚至几年前的事件影响。传统的 RNN 和 LSTM 在捕捉长程依赖关系时往往表现不佳，因为这些模型需要逐步处理序列数据，导致梯度消失或梯度爆炸问题。而 Transformer 模型通过自注意力机制，可以直接计算序列中任意两个位置之间的依赖关系，从而更有效地捕捉长程依赖。

具体来说，自注意力机制通过计算 Query，Key 和 Value 之间的点积来生成注意力权重，再使用这些权重对值进行加权求和，以得到新的表示。这种机制使模型能够全局关注输入序列中的信息，而无需逐步处理，从而大大提高了捕捉全局依赖的能力。

3.2.2 并行计算

Transformer 模型的架构允许并行计算，从而大大提高了训练和推理的效率。相比于 RNN 和 LSTM 的顺序处理，Transformer 模型能够更快地处理大规模时间序列数据。这是因为在 Transformer 模型中，所有的输入序列位置可以同时计算，而不是像 RNN 那样逐个时间步地进行处理。

并行计算不仅提高了计算效率，还使得 Transformer 模型能够更好地利用现代硬件（如 GPU 和 TPU）的计算能力，从而在大规模数据集上进行高效训练。

3.2.3 多头注意力机制

多头注意力机制使得模型能够在不同子空间中捕捉时间序列的多种特征，从而提高预测的准确性。具体来说，多头注意力机制通过并行计算多个自注意力头，捕捉不同子空间中的依赖关系，使模型能够关注到输入序列中的不同特征和模式。在时间序列预测任务中，不同的注意力头可以分别关注到长期趋势、季节性变化和短期波动，从而提供更丰富和多样化的特征表示[36]。这种多样化的特征表示，有助于提高模型的预测性能。

3.2.4 位置编码

Transformer 模型不包含递归结构或卷积操作，因此需要显式地引入位置信息。位置编码通过添加固定或可学习的位置向量，使模型能够捕捉输入序列中元素的位置关系。位置编码通常采用正弦和余弦函数生成固定位置编码，或者通过训练得到可学习的位置编码。

位置编码使得 Transformer 模型能够保留时间序列数据的顺序信息，从而更好地理解时间序列中的时序关系。这对于时间序列预测任务尤为重要，因为时间序列数据的顺序关系直接影响到预测结果的准确性。

3.3 Transformer 衍生模型

Transformer 模型在时间序列预测上的成功，促进了 Transformer 衍生模型的发展。根据最近的研究，主要可以将衍生模型划分为模块级变体和架构级变体。模块级变体针对自注意力、前馈网络等单一组件进行优化，以适应时间序列数据。而架构级变体则重塑和优化整个模型结构，如专门为时间序列预测任务设计了子模块。这些变体在处理复杂数据时的表现往往优于原始的 Transformer 模型。

3.3.1 模块级别的变体

1）AST

现有的方法多采用自回归生成模式，在训练时使用真实值，而在推断时使用模型自身的一步预测值，这会导致推断过程中的误差被累积，从而影响对长时间序列的准确预测。

Wu 等[37]提出 Adversarial Sparse Transformer（AST）模型，其创新之处在于稀疏自注意力机制和对抗训练方法。稀疏自注意力机制通过引入稀疏矩阵，显著降低了注意力计算的时间和空间复杂度，使得模型在处理长序列数据时更加高效。该机制通过选择性地关注序列中的关键点，减少计算资源消耗的同时，保持了对重要信息的有效捕捉。对抗训练方法通过生成对抗样本，增强了模型的鲁棒性和泛化能力。对抗训练不仅帮助模型更好地应对时间序列数据中的不确定性和变化，还提高了模型在不同数据集上的适应性。通过对抗样本的引入，模型能够学习到更丰富的特征，从而提升预测的准确性。

2）TFT

当前时间序列预测研究已经从单变量预测转向了多元时间序列建模。这种多元时间序列不仅包含更多变量和特征，还展现出各种复杂的统计分布特性，如缺失数据、趋势、季节性、波动性、漂移和罕见事件等。这种复杂多元时间序列给时间序列预测带来了巨大挑战。

Lim 等[38]提出了 Temporal Fusion Transformer（TFT）模型，其创新之处在于静态协变量编码器、门控特征选择模块和时间自注意力解码器的多尺度预测模型。静态协变量编码器用于处理时间序列中的静态特征，通过特征嵌入和特征选择，确保模型能够有效利用这些对预测结果有重要影响的静态特征。门控特征选择模块通过门控机制动态选择和过滤输入特征，该模块通过学习的门控参数选择重要特征并过滤掉噪声和无关特征，动态调整策略确保模型在不同时间步长上都能选择最重要的特征，从而提升预测的准确性和灵活性。时间自注意力解码器结合多头自注意力机制和多层堆叠架构，捕捉时间序列中的复杂依赖关系，实现多尺度预测。多头自注意力机制捕捉不同时间步长之间的依赖关系，多层堆叠架构增强模型的表达能力，使得模型能够同时处理长短期依赖关系。解码器生成多尺度预测结果，提升了模型在不同时间尺度上的适用性。TFT 通过这些创新设计，显著提升了时间序列预测的准确性和灵活性，并具有较高的可解释性，在多视角时间序列预测任务中表现出色，能够有效处理具有复杂依赖关系的时间序列数据。

3）Informer

首先，传统 Transformer 模型中的自注意力机制的计算复杂度随序列长度呈平方级增长，这对于处理长序列任务来说计算开销较大。其次，Transformer 模型通常需要堆叠多层网络结构，导致内存占用较高，限制了模型规模的扩展。此外，Transformer 采用的 step-by-step 的序列解码方式相比于并行推理，速度也更慢。这些问题都在一定程度上限制了 Transformer 模型的应用。

Zhou 等[39]提出了 Informer 模型，采用了创新性的 ProbSparse 自注意力机制、自注意力蒸馏机制和生成式解码器来优化长序列处理。ProbSparse 机制通过基于 Query 与 Key 的注意力分布相似度及其与均匀分布的 Kullback-Leibler 散度，精准选取潜在高得分的 Query，再通过随机采样挑选出关键且稀疏的 Query，大幅减少了计算量和内存需求。自注意力蒸馏机制在每个注意力层后添加一维卷积和最大池化层，对输出进行减半处理，更集中于关键信息，有效处理长输入序列。生成式解码器允许模型在一次计算中预测所有输出，提升长序列预测的效率。这些技术综合优化了模型的时间和空间复杂度，显著提高了性能和效率，能应对传统 Transformer

在长短期时间序列预测中的多重挑战。Informer 模型架构图如图 3 所示。

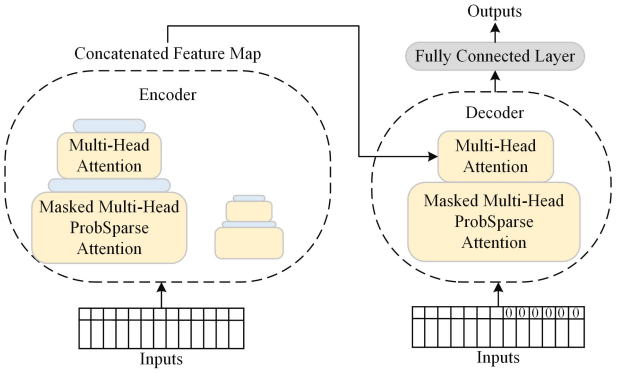

图 3 Informer 模型架构图

Fig. 3 Informer model architecture

4）Autoformer

以往基于 Transformer 的预测模型主要通过采用稀疏版本的自注意力机制来提高效率，虽然性能有所提升，但该方法仍使用逐点表示聚合，降低了信息利用率，成为长期时间序列预测的瓶颈。

Wu 等[40]提出了 Autoformer 模型，其创新之处在于独特的分解架构和高效的自相关性机制。深度分解架构通过将时间序列拆分为趋势和季节性两部分，能更精确地把握数据的长期趋势和周期性波动。该分解不仅提升了预测准确度，也增强了对不同时间尺度变化的适应性。自相关机制通过自相关层自动识别序列内部的相关性，简化了自注意力计算，特别是在处理长序列时显著降低了计算成本。该层通过分析各时间点间的相关性来识别重复模式，并基于这些模式预测未来数据点。这种方法充分利用了时间序列的内在统计特性，降低了模型的复杂度，并提升了处理长时间序列的效率。Autoformer 在处理具有明显季节性和趋势性的长时间序列预测任务中表现出卓越的性能。

5）PatchTST

最近，一个简单的线性模型 DLinear[41]在各种常见基准数据集上取得了优于之前所有 Transformer 模型的性能，这对 Transformer 在时间序列预测中的有效性提出了挑战。

Nie 等[42]在 Transformer 的基础上做出改进，提出了 PatchTST 模型，该模型通过数据分块、通道独立性和自监督表征学习有效应对长序列处理的挑战。首先，通过数据分块将时间序列分割成子序列级别的“patch”单元，减少模型需处理的 token 数量，降低计算成本，同时提高长期依赖处理能力和对时间序列动态的敏感度。其次，PatchTST 的通道独立设计允许单独处理每个特征通道，增强了处理独特特征的效率和准确性，简化了模型结构，易于训练。最后，通过自监督预训练，在大型未标注数据集上学习到丰富且泛化的特征表示，不仅提升了模型的迁移学习能力和微调效果，还加快了模型适应新任务的速度。这些创新，使 PatchTST 在复杂时间序列预测中展现出高效性和准确性，模型取得了优异的预测成果。

3.3.2 架构级别的变体

1）Aliformer

在电商领域，准确的时间序列销售预测（TSSF）可以显著提高经济效益。然而，TSSF 面临产品趋势和季节性变化大、促销活动影响销售等难题。此外，除了历史统计数据外，还可以利用一些未来的已知信息，这些信息可能反映未来促销活动对当前销售的影响，并有助于提高预测准确性。然而，大多数现有的方法仅基于历史信息进行预测。

Qi 等[43]提出了 Aliformer 模型，其创新之处在于基于双向 Transformer 架构引入未来信息进行预测。Aliformer 模型通过历史信息、当前因素和未来信息来预测未来的销售，特别适用于电商中的时间序列销售预测。其双向 Transformer 在处理时间序列数据时，不仅关注历史数据，还结合未来已知信息，从而提升预测的准确性和模型的理解能力。知识引导的自注意力层通过利用已知知识的一致性来指导时序信息的传输，增强了模型对未来促销活动等因素的预测能力。该训练策略使模型更加注重对未来信息的利用，从而提高了预测的准确性。在 4 个公共基准数据集和天猫提出的一个大规模工业数据集上的广泛实验中，Aliformer 表现出显著优于现有最先进的时间序列预测方法的性能。

2）FEDformer

基于 Transformer 的时间序列预测方法虽然在长期预测中的效果有所改善，但还存在计算成本高昂以及无法有效捕捉时间序列的全局特征等问题。

Zhou 等[44]提出了 FEDformer 模型，其创新之处在于结合了频域增强机制、序列分解架构和多尺度建模机制。频域增强机制通过傅里叶变换将时间序列数据从时域转换到频域，提取重要的频率成分并去除噪声，然后通过逆傅里叶变换将处理后的数据转换回时域。这一机制能够更好地捕捉时间序列中的周期性模式，提高预测的准确性。序列分解架构通过将时间序列数据分解为趋势、季节性和残差 3 个部分，使得模型能够更精确地把握数据的长期趋势和周期性波动。这种分解方法不仅提升了预测准确度，也增强了对不同时间尺度变化的适应性。此外，FEDformer 引入了多尺度建模机制，通过多尺度卷积和多尺度注意力机制，处理不同时间尺度上的特征，从而同时捕捉短期和长期的依赖关系，进一步提高了对复杂时间序列的建模能力。

这种结合频域信息和序列分解技术的方法，充分利用了时间序列的内在统计特性，将计算成本降低为线性复杂度，并提高了处理长时间序列的效率。在多个基准数据集上的实验结果表明，FEDformer 在处理复杂时间序列预测任务中表现出了卓越的性能。

3）Crossformer

以往的研究主要聚焦于时间维度关系的建模，采用自注意力机制来建立不同时间步之间的依赖，忽略了跨维度的依赖关系。然而，在多元时间序列预测中，各变量之间的关系同样重要。

Zhang 等[45]针对传统 Transformer 架构的局限进行了改进，推出了 Crossformer 模型。该模型主要包含维度分段嵌入（DSW）、两阶段注意力（TSA）以及层次编码器-解码器结构（HED）三大创新。DSW 嵌入通过细分不同维度的时间序列数据并生成特征向量，有效提升了模型对多维数据间联系的捕捉能力，突破了长序列处理的障碍。TSA 层在优化模型时

采用了微观至宏观的处理策略，先单独处理每个维度内的信息，再整合各段数据，以此应对时间序列中的复杂相关性和变化。HED结构利用分层次的方式加强了模型梳理多尺度数据的能力，每个编码器层次不仅处理当前级别信息，还与之前层次的输出进行信息融合，有效提高了长距离依赖信息捕捉的效率和准确性。这三大创新的结合，使Crossformer在对时间序列的分析和预测方面表现出卓越的性能。

4）Pyraformer

时间序列预测的主要挑战在于构建一个既强大又简洁的模型，以紧凑地捕捉不同范围的时间依赖性。为了平衡模型容量与复杂度，学者们提出了多种Transformer衍生模型。然而，这些方法虽然在降低复杂度方面有所突破，但在显著缩短最大路径长度上仍面临困难。

Liu等[46]提出了Pyraformer模型，其创新之处在于金字塔式注意力机制，主要包括金字塔注意力模块（PAM）和粗尺度构建模块（CSCM）两部分。PAM通过层次化的注意力机制，在各个层级上捕捉时间序列的关键特征，并允许模型重点关注序列的重要部分，同时吸收来自其他层的信息。这种机制为模型提供了在不同时间尺度上识别和处理时间序列特征的有效方法。在较低层次，PAM集中处理局部细节，在较高层次则覆盖更广泛的时间范围，以此捕捉时间序列的多尺度动态。CSCM在金字塔结构中负责从精细到粗糙尺度的信息转换，实现从底层细节到顶层概括的过渡。通过下采样操作，CSCM将底层的高分辨率信号转换为上层的低分辨率信号，旨在尽可能地保持信息的完整性。这使得模型在更高层可以基于较为粗略的尺度对时间序列的长期依赖性进行有效建模，减轻了顶层模块处理底层细节信息的重负。CSCM对于多尺度信息整合起着至关重要的作用，其通过创建一个较为粗糙的时间尺度抽象，为分析整体趋势和长期依赖提供了基础，从而增强了模型的预测能力和效率，并达到$O(L)$的复杂度和$O(1)$的最大路径长度。Pyraformer模型架构图如图4所示。

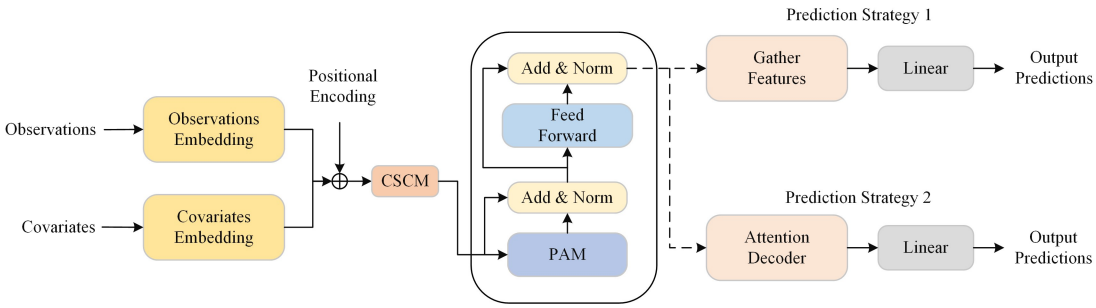

图4 Pyraformer模型架构图

Fig. 4 Pyraformer model architecture

3.3.3 小结

Transformer的衍生模型对时间序列预测的发展具有重要意义。本小节从模块级别和架构级别两个角度对Transformer衍生模型进行分类，深入研究并分析不同变体的特点，如表1所列。

表1 Transformer衍生模型的比较分析

Table 1 Comparative analysis of Transformer-derived models

模型	创新	所解决的问题	局限性
AST[37]	通过对抗性训练增强模型的鲁棒性，通过稀疏自注意力机制降低计算复杂度	解决因自回归生成模式在推理过程中的错误累积而导致无法预测较长时间范围内的时间序列的问题	对抗训练方法增加了训练的复杂性和时间复杂度，在训练过程中需要调整对抗样本的生成策略和损失函数的权重等超参数
TFT[38]	静态协变量编码器提取有效静态特征，门控特征选择模块过滤掉噪声和无关特征，时间自注意力解码器捕捉时间序列中的复杂依赖关系	解决了多元异构输入数据处理问题，去除了无关特征和噪声	对于没有明确静态特征可用的任务，关键信息提取受限
Informer[39]	提出ProbSparse自注意力机制来有效捕获长程依赖性，使用自注意力蒸馏来突出主导的注意力，生成式解码器提高了长序列预测的推理速度	解决了传统Transformer的高时间复杂度和高内存消耗的问题	面对长时间序列预测任务时表现不佳，虽降低了计算复杂度，但造成了信息的丢失
Autoformer[40]	引入分解架构对序列信息进行分解精炼，设计基于自动相关性机制在子序列级别上发现并聚合时间序列周期性	更注重对季节项的建模，比传统的Transformer模型具有更高的信息利用率	模型过度依赖数据的周期性，不适合在周期性不显著的数据集上训练
PatchTST[42]	通过Patching将时间序列数据分割成子序列级别的块，通道独立性分别处理每个特征通道的信息，自监督表征学习提升模型的泛化能力	通过对时间序列数据进行分段和通道独立，可以处理更长的时间序列，还显著降低了时间复杂度和空间复杂度	Patch的大小作为超参数难以确定，过大时不易处理长期依赖关系，过小时可能导致信息丢失
Aliformer[43]	通过知识引导的自注意力层，利用已知知识的一致性来指导时间信息的传递，未来强调的训练策略注重对未来信息的利用，双向Transformer架构同时利用历史信息、当前因素和未来信息进行预测	大多数现有的方法仅根据历史信息预测未来，该方法弥补了未来信息的遗漏	在没有未来信息或未来信息不准确的任务上表现不佳，限制了模型在某些实际应用场景中的适用性

（续表）

模型	创新	所解决的问题	局限性
FEDformer[44]	频域增强机制提高了对周期性模式的捕捉和预测准确性；序列分解架构将时间序列分解，提升了预测准确度和时间尺度适应性；多尺度建模机制处理不同时间尺度上的特征，捕捉短期和长期依赖关系	将 Transformer 应用于频域，应用于时域时比以往方法能更好地捕捉时间序列的全局依赖	模型性能依赖于时间序列的频率信息，对于频率特征不明显的数据集效果不佳
Crossformer[45]	使用 DSW 将时间序列分割成段，并嵌入为特征向量，提出 TSA 层，有效捕获 2D 向量中的跨时间和跨维度依赖关系，建立层次编码器-解码器，以捕获更粗粒度尺度上的依赖关系	解决了以往方法忽略了不同变量之间的依赖关系的问题	在处理高维度数据时，全连接的跨维度依赖性可能引入噪声
Pyraformer[46]	利用金字塔式注意力机制在不同分辨率上提取特征，并通过内部尺度的相邻连接来建模不同范围的时间依赖关系	有效地捕获短期和长期依赖关系，同时具有较低的时间复杂度和空间复杂度	模型提取多尺度、多层级的重要信息，在结构化程度差的数据集中预测表现不佳

4 研究展望

4.1 增强模型的可解释性

Transformer 及其衍生模型由于具有优异的性能，在时间序列预测中获得了广泛应用。但目前的大部分架构仍然属于“黑箱”模型，其预测结果是由其内部众多参数之间的复杂非线性交互所决定的[38]。Transformer 模型的自注意力机制具有一定的可解释性，通过分析注意力权重，可以了解模型在预测过程中所关注的关键时间点和特征。通过可视化注意力权重和特征重要性分析，展示模型在不同时间步的关注点，帮助理解模型的预测逻辑，增强模型的可解释性。未来的研究应专注于开发新的技术方法，以解释 Transformer 模型的内部机制和预测逻辑。这可能包括改进的可视化工具[47-48]，使模型的注意力机制和数据流向更容易理解，或开发新的算法来解码模型的决策过程[49]。

4.2 时间序列预测与多模态数据融合

现实世界中的时间序列预测问题远比处理单一模态数据复杂得多。通常情况下，预测任务需要考虑多种模态的数据，如文本、图像、声音、数值型时间序列等。多模态数据融合的优势在于不同模态的数据可以提供互补的信息，共同构成对目标现象更全面的描述，从而提高模型的预测能力和泛化能力[50]。在处理多模态数据时，自注意力机制可以有效地整合来自不同模态的数据。各模态数据作为不同输入，自注意力机制通过计算各模态之间的相关性，实现数据的融合。例如，在电力负荷预测中，可以同时输入历史电力负荷数据、气象数据和经济指标，通过自注意力机制整合这些数据，提升预测的准确性。这种多模态融合的方法可能会成为未来的研究热点之一。

4.3 时间序列预测与大模型的结合

Transformer 模型天然具备捕捉长距离依赖的能力，这使得其在处理包含复杂时间关系和模式的时间序列数据时表现出色。结合大模型，尤其是那些已经在海量数据上预训练过的模型，可以进一步提升系统对时间序列数据的语义理解能力，从而更准确地捕捉到数据的时间依赖性及其隐含的趋势和模式。Transformer 模型在处理序列数据方面的先天优势，加之大模型提供的庞大参数空间和强大的泛化能力，使得这种组合在时间序列预测上显示出巨大的潜力。首先，直接利用预训练大模型。预训练模型在自然语言处理和计算机视觉领域取得了巨大的成功[51]，但这些模型在时间序列分析领域进展缓慢，研究者可以使用如 GPT[52] 和 Llama[53] 等模型来进行时间序列预测[54]。其次，发展专门针对时间序列预测的大模型，包括对传统的时间序列预测方法进行扩展和优化，使其能够处理更大规模的数据和更复杂的时序关系。

4.4 模型轻量化设计

Transformer 模型通常包含数百万到数十亿个参数，训练和推理需要大量计算资源，对硬件要求很高。在资源有限的环境中（如移动设备、嵌入式系统），这些需求难以满足。大规模 Transformer 模型需要大量内存存储参数和中间计算结果，在内存受限的设备（如智能手机、物联网设备）上，可能无法部署或运行效率低下。

为了使基于 Transformer 的时间序列预测模型能够广泛应用于这类环境，研究人员可以选择模型压缩和优化策略。举例来说，通过参数剪枝，能够简化模型结构，剔除冗余参数[55]；权重共享的策略能够有效减少模型所需的参数量，降低存储与计算负担[56]；而低秩分解则在确保预测准确性的同时，大幅压缩了模型规模及计算需求[57]。另外，Lin 等[58]通过交叉周期稀疏预测技术将数据周期和趋势性有效解耦，并且极大地压缩了模型参数的大小，降低了对计算资源的需求。这些技术不仅显著降低了模型的体积，还保障了优秀的预测效果。通过模型压缩与优化，基于 Transformer 的时间序列预测模型得以在计算能力和存储空间较为有限的移动设备上部署。

结束语 学术界已经见证了 Transformer 模型在处理序列数据方面的强大能力。尽管目前已经取得巨大的进展，但使用 Transformer 进行时间序列预测仍然是一个快速发展和不断进步的研究领域。未来，随着技术的进步和计算资源的增加，期待出现更多优化方案和创新方法，以应对效率和数据规模的挑战，并推进 Transformer 在时间序列预测方面的深入应用。

参 考 文 献

[1] LI Z X，LIU H Y. A multivariate time series forecasting method incorporatingglobaland serial features[J]. Journal of Computing，2023，46(1)：70-84.

[2] BARRA S，CARTA S M，CORRIGA A，et al. Deep learning and time series-to-image encoding for financial forecasting[J]. IEEE/CAA Journal ofAutomatica Sinica，2020，7(3)：683-692.

[3] MA C，DAI G，ZHOU J. Short-term traffic flow prediction for

urban road sections based on time series analysis and LSTM_BILSTM method[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(6): 5615-5624.

[4] SHARMA R R, KUMAR M, MAHESHWARI S, et al. EVDHM-ARIMA-based time series forecasting model and its application for COVID-19 cases[J]. IEEE Transactions on Instrumentation and Measurement, 2020, 70: 1-10.

[5] AMJADY N. Short-term hourly load forecasting using time-series modeling with peak load estimation capability[J]. IEEE Transactions on Power Systems, 2001, 16(3): 498-505.

[6] BOX G E P, JENKINS G M, REINSEL G C, et al. Time series analysis: forecasting and control[M]. John Wiley & Sons, 2015.

[7] BOX G E P, PIERCE D A. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models[J]. Journal of the American statistical Association, 1970, 65(332): 1509-1526.

[8] XIE Y, JIN M, ZOU Z, et al. Real-time prediction of docker container resource load based on a hybrid model of ARIMA and triple exponential smoothing[J]. IEEE Transactions on Cloud Computing, 2020, 10(2): 1386-1401.

[9] LAI G, CHANG W C, YANG Y, et al. Modeling long-and short-term temporal patterns with deep neural networks[C]// The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018: 95-104.

[10] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 6000-6010.

[11] GALASSI A, LIPPI M, TORRONI P. Attention in natural language processing[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(10): 4291-4308.

[12] ZHANG T, GONG X, CHEN C L P. BMT-Net: Broad multitask transformer network for sentiment analysis[J]. IEEE Transactions on Cybernetics, 2021, 52(7): 6232-6243.

[13] HU Y, JIN X, ZHANG Y, et al. Rams-trans: Recurrent attention multi-scale transformer for fine-grained image recognition [C]// Proceedings of the 29th ACM International Conference on Multimedia. 2021: 4239-4248.

[14] CHEN H, WANG Y, GUO T, et al. Pre-trained image processing transformer[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 12299-12310.

[15] MILLER J A, ALDOSARI M, SAEED F, et al. A survey of deep learning and foundation models for time series forecasting[J]. arXiv: 2401. 13912, 2024.

[16] LIM B, ZOHREN S. Time-series forecasting with deep learning: a survey[J]. Philosophical Transactions of the Royal Society A, 2021, 379(2194): 20200209.

[17] HAN Z, ZHAO J, LEUNG H, et al. A review of deep learning models for time series prediction[J]. IEEE Sensors Journal, 2019, 21(6): 7833-7848.

[18] YIN J, RAO W, YUANM, et al. Experimental study of multivariate time series forecasting models[C]// Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019: 2833-2839.

[19] AK R, FINK O, ZIO E. Two machine learning approaches for short-term wind speed time-series prediction[J]. IEEE Transactions on Neural Networks and Learning Systems, 2015, 27(8): 1734-1747.

[20] LINDEMANN B, MÜLLER T, VIETZ H, et al. A survey on long short-term memory networks for time series prediction[J]. Procedia Cirp, 2021, 99: 650-655.

[21] CHENG C, SA-NGASOONGSONG A, BEYCA O, et al. Time series forecasting for nonlinear and non-stationary processes: a review and comparative study[J]. Iie Transactions, 2015, 47(10): 1053-1071.

[22] ARIK S O, YODER N C, PFISTER T. Self-adaptive forecasting for improved deep learning on non-stationary time-series[J]. arXiv: 2202. 02403, 2022.

[23] WANG Z, BOVIK A C. Mean squared error: Love it or leave it? A new look at signal fidelity measures[J]. IEEE Signal Processing Magazine, 2009, 26(1): 98-117.

[24] WILLMOTT C J, MATSUURA K. Advantages of the mean absolute error(MAE) over the root mean square error(RMSE) in assessing average model performance[J]. Climate Research, 2005, 30(1): 79-82.

[25] CHAI T, DRAXLER R R. Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature[J]. Geoscientific Model Development, 2014, 7(3): 1247-1250.

[26] CORTES C, VAPNIK V. Support-vector networks[J]. Machine learning, 1995, 20: 273-297.

[27] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45: 5-32.

[28] LIANG H T, LIU S, DU J W, et al. A review of deep learning applications in time series forecasting[J]. Journal of Frontiers of Computer Science & Technology, 2023, 17(6): 1285.

[29] MEDSKER L R, JAIN L C. Recurrent neural networks: design and applications[M]. CRC press, 1999.

[30] LIN S, LIN W, WU W, et al. Segrnn: Segment recurrent neural network for long-term time series forecasting[J]. arXiv: 2308. 11200, 2023.

[31] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[32] LIVIERIS I E, PINTELAS E, PINTELASP. A CNN-LSTM model for gold price time-series forecasting[J]. Neural Computing and Applications, 2020, 32: 17351-17360.

[33] GRAVES A. Long Short-term Memory[J]. Supervised Sequence Labelling with Recurrent Neural Networks, 2012, 385: 37-45.

[34] WEN Q, ZHOU T, ZHANG C, et al. Transformers in time series: A survey[J]. arXiv: 2202. 07125, 2022.

[35] TAO C, GAO S, SHANG M, et al. Get The Point of My Utterance! Learning Towards Effective Responses with Multi-Head Attention Mechanism[C]// IJCAI. 2018: 4418-4424.

[36] BENIDIS K, RANGAPURAM SS, FLUNKERT V, et al. Deep learning for time series forecasting: Tutorial and literature sur-

vey[J]. ACM Computing Surveys,2022,55(6):1-36.
[37] WU S,XIAO X,DING Q,et al. Adversarial sparse transformer for time series forecasting[J]. Advances in Neural Information Processing Systems,2020,33:17105-17115.
[38] LIM B,ARIK S Ö,LOEFF N,et al. Temporal fusion transformers for interpretable multi-horizon time series forecasting[J]. International Journal of Forecasting,2021,37(4):1748-1764.
[39] 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.
[40] 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.
[41] 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.
[42] NIE Y,NGUYEN N H,SINTHONG P,et al. A time series is worth 64 words:Long-term forecasting with transformers[J]. arXiv:2211.14730,2022.
[43] QI X,HOU K,LIU T,et al. From known to unknown:Knowledge-guided transformer for time-series sales forecasting in alibaba[J]. arXiv:2109.08381,2021.
[44] ZHOU T,MA Z,WEN Q,et al. Fedformer:Frequency enhanced decomposed transformer for long-term series forecasting[C]// International Conference on Machine Learning. PMLR, 2022: 27268-27286.
[45] ZHANG Y,YAN J. Crossformer:Transformer utilizing cross-dimension dependency for multivariate time series forecasting [C]// The Eleventh International Conference on Learning Representations. 2022.
[46] LIU S,YU H,LIAO C,et al. Pyraformer:Low-complexity pyramidal attention for long-range time series modeling and forecasting[C]// International Conference on Learning Representations. 2021.
[47] YEH C,CHEN Y,WU A,et al. Attentionviz:A global view of transformer attention[J]. IEEE Transactions on Visualization and Computer Graphics,2024,30(1):262-272.
[48] BRAŞOVEANU A M P,ANDONIE R. Visualizing transformers fornlp:a brief survey[C]// 2020 24th International Conference Information Visualisation(IV). IEEE,2020:270-279.
[49] CORNIA M,BARALDI L,CUCCHIARA R. Explaining transformer-based image captioning models:An empirical analysis [J]. AI Communications,2022,35(2):111-129.
[50] CHEN H,JIANG D,SAHLI H. Transformer encoder with multi-modal multi-head attention for continuous affect recognition[J]. IEEE Transactions on Multimedia, 2020, 23: 4171-4183.
[51] LIU Z,RODRIGUEZ-OPAZO C,TENEY D,et al. Image retrieval on real-life images with pre-trained vision-and-language models[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:2125-2134.
[52] RADFORD A,NARASIMHAN K,SALIMANS T,et al. Improving language understanding by generative pre-training[J/OL]. https://cdn. openai. com/research-covers/language-unsupervised/language_understanding_paper. pdf.
[53] JIN M,WANG S,MA L,et al. Time-llm:Time series forecasting by reprogramming large language models[J]. arXiv:2310.01728,2023.
[54] ZHOU T,NIU P,SUN L,et al. One fits all:Power general time series analysis by pretrained lm[J]. arXiv:2302.11939,2023.
[55] HE Y,DONG X,KANG G,et al. Asymptotic soft filter pruning for deep convolutional neural networks[J]. IEEE Transactions on Cybernetics,2019,50(8):3594-3604.
[56] HE H,CAI J,LIU J,et al. Pruning self-attentions into convolutional layers in single path[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2024,46(5):13.
[57] DING Z,FU Y. Dual low-rank decompositions for robust cross-view learning[J]. IEEE Transactions on Image Processing, 2018,28(1):194-204.
[58] LIN S,LIN W,WUW,et al. SparseTSF:Modeling Long-term Time Series Forecasting with 1k Parameters[J]. arXiv:2405.00946,2024.

CHEN Jiajun, born in 2001, postgraduate. His main research interests include time series forecasting and so on.

LIN Weiwei, born in 1980, Ph.D, professor, is a distinguished member of CCF (No. 37313D). His main research interestes include cloud computing, big data technology and AI application technology.

(责任编辑:柯颖)