计算机科学与探索
Journal of Frontiers of Computer Science and Technology doi: 10.3778/j.issn.1673-9418.2211108

深度学习应用于时序预测研究综述

梁宏涛，刘硕+，杜军威，胡强，于旭

青岛科技大学 信息科学技术学院，山东 青岛 266061
+ 通信作者 E-mail: liushuo2020@mails.qust.edu.cn

摘　要:时间序列一般是指对某种事物发展变化过程进行观测并按照一定频率采集得出的一组随机变量。时间序列预测的任务就是从众多数据中挖掘出其蕴含的核心规律并且依据已知的因素对未来的数据做出准确的估计。由于大量物联网数据采集设备的接入、多维数据的爆炸增长和对预测精度的要求愈发苛刻，导致经典的参数模型以及传统机器学习算法难以满足预测任务的高效率和高精度需求。近年来，以卷积神经网络、循环神经网络和 Transformer 模型为代表的深度学习算法在时间序列预测任务中取得了丰硕的成果。为进一步促进时间序列预测技术的发展，综述了时间序列数据的常见特性、数据集和模型的评价指标，并以时间和算法架构为研究主线，实验对比分析了各预测算法的特点、优势和局限；着重介绍对比了多个基于 Transformer 模型的时间序列预测方法；最后结合深度学习应用于时间序列预测任务存在的问题与挑战对未来该方向的研究趋势进行了展望。

关键词:时间序列数据; 时间序列预测; 深度学习; Transformer 模型

文献标志码：A　　**中图分类号：**TP181

Review of Deep Learning Applied to Time Series Prediction

LIANG Hongtao, LIU Shuo+, DU Junwei, HU Qiang, YU Xu

School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao, Shandong 266061, China

Abstract: The time series is generally a set of random variables that are observed and collected at a certain frequency in the course of something's development. The task of time series forecasting is to extract the core patterns from a large amount of data and to make accurate estimates of future data based on known factors. Due to the access of a large number of IoT data collection devices, the explosive growth of multidimensional data and the increasingly demanding requirements for prediction accuracy, it is difficult for classical parametric models and traditional machine learning algorithms to meet the high efficiency and high accuracy requirements of prediction tasks. In recent years, deep learning algorithms represented by convolutional neural networks, recurrent neural networks and Transformer models have achieved fruitful results in time series forecasting tasks. To further promote the development of time series prediction technology, common characteristics of time series data, evaluation indexes of data sets and models are reviewed, and the characteristics, advantages and limitations of each prediction

基金项目：国家自然科学基金（61973180；62172249）；山东省产教融合研究生联合培养示范基地项目（2020-19）。

This work was supported by the National Natural Science Foundation of China (61973180), the National Natural Science Foundation of China (62172249) and the Industry-Education Postgraduate Joint Cultivation Demonstration Base Project of Shandong Province (2020-19).

algorithm are experimentally compared and analyzed with time and algorithm architecture as the main research line; several time series prediction methods based on Transformer model are highlighted and compared; finally, deep learning is combined with the application to time Finally, the problems and challenges of deep learning applied to time series prediction tasks are combined to provide an outlook on the future research trends in this direction.
Key words: time series data; time series prediction; deep learning; Transformer model

随着社会中物联网传感器的广泛接入，几乎所有科学领域都在以不可估量的速度产生大量的时间序列数据。传统参数模型和机器学习算法已难以高效准确地处理时间序列数据，因此采用深度学习算法从时间序列中挖掘有用信息已成为众多学者关注的焦点。

分类聚类[1-4]、异常检测[5-7]、事件预测[8-10]、时间序列预测[11-14]是时间序列数据的四个重点研究方向。已有的时序预测综述文章，概括了经典的参数模型以及传统机器学习算法的相关内容，但缺少对Transformer 类算法最新成果的介绍和在各行业常用数据集的实验对比分析。余下内容将以深度学习的视角重点分析阐述有关时间序列预测方向的内容，并在多种 GPU 环境下对不同数据集采用多个评价指标进行实验对比分析。基于深度学习的时间序列预测算法发展脉络如图 1 所示：

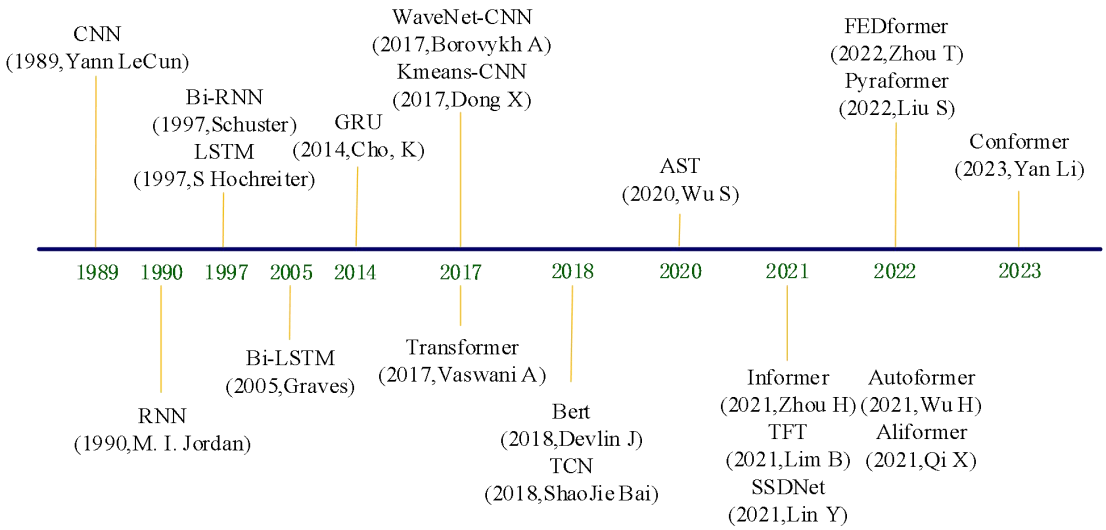


图 1　基于深度学习的时间序列预测算法时间表

Fig.1　Development history of time series prediction algorithms based on deep learning

时间序列预测是时间序列任务中最常见和最重要的应用，通过挖掘时间序列潜在规律，去进行类推或者延展用于解决在现实生活中面临的诸多问题，包括噪声消除[15]、股票行情分析[16-17]、电力负荷预测[18]、交通路况预测[19-20]、流感疫情预警[21]等。

当时间序列预测任务提供的原始数据仅为目标数据的历史数据时，为单变量时间序列预测，当提供的原始数据包含多种随机变量时，为多变量时间序列预测。

时间序列预测任务根据所预测的时间跨度长短，可划分为四类，具体如图 2 所示：

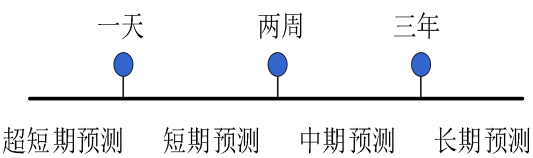


图 2　预测任务以时间跨度分类图

Fig.2　Prediction tasks categorized by time span

文章余下部分主要介绍基于深度学习的时间序列预测算法研究，其中第二节介绍时间序列数据特性，第三节介绍了时间序列预测任务的常用数据集和评价指标，第四节介绍了深度学习的研究进展及在时间序列预测领域的应用，第五节展望未来深度学习在时序预测领域的研究方向。

1 时间序列数据的特性

时间序列预测是对前 $t-1$ 个时刻的历史数据学习分析，来估计出指定未来时间段的数据值。时间序列数据由于其各变量间固有的潜在联系，常表现出一种或多种特性，为对时序预测有更全面的认识，本节将对这些常见特性进行详细介绍。

(1)海量性：随着物联网传感器设备的升级，测量频率的提高，测量维度的增加，时间序列数据爆炸性增长，高维度的时间序列数据占据主流[22]。在数据集层面进行有效的预处理工作，是高质量完成时间序列预测任务的关键。

(2)趋势性：当前时刻数据往往与前一段时刻数据有着密切的联系，该特点暗示了时间序列数据受其他因素影响通常有一定的变化规律，时间序列可能在长时间里展现出一种平稳上升或平稳下降或保持水平的趋势。

(3)周期性：时间序列中数据受外界因素影响，在长时间内呈现出起起落落的交替变化[23]，例如，涨潮退潮，一周内潮水高度不符合趋势性变化，并不是朝着某一方向的近似直线的平稳运动。

(4)波动性：随着长时间的推移和外部多因素影响，时间序列的方差和均值也可能会发生系统的变化，在一定程度上影响时间序列预测的准确度。

(5)平稳性：时间序列数据个别为随机变动，在不同时间上呈统计规律，在方差与均值上保持相对稳定。

(6)对称性：若某段时间周期内，原始的时间序列和其反转时间序列的距离控制在一定的阈值以内，曲线基本对齐，即认定该段时间序列具有对称性[24]，例如港口大型运输车往复作业，起重机抬臂和降臂工作等。

各特性具体示例如图 3 所示：

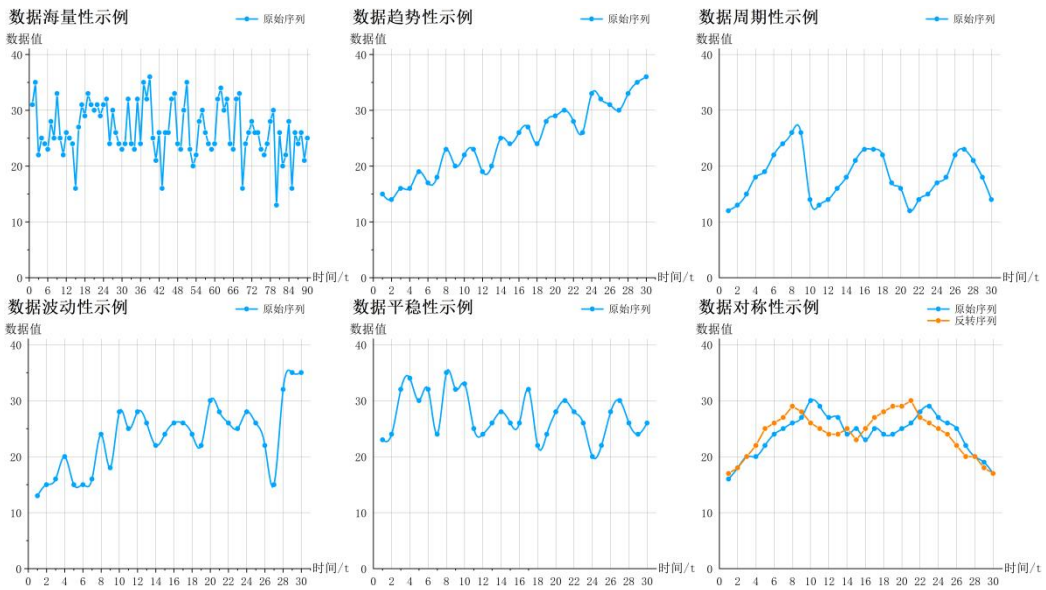


图 3　时间序列数据特性示例图

Fig.3　Example graphs of time series data characteristics

2 时序预测数据集和评价指标

2.1 数据集

权威的数据集一直是衡量不同算法优劣的重要标准，数据集在使用前一般要进行子集选择、噪音处理、缺失值补充和数据类型转换等操作，来保证数据准确性、完整性和一致性。在解决实际任务时，

对于一个给定的数据集，应当根据数据集的情况来选择适当的模型算法进行处理，如果盲目选择经典或最先进算法往往难以得到一个好的预测结果。研究人员可以根据数据集记录条数的数量级和特征变量的多少以及任务要求的预测步长来选定合适的算法。

下文用于衡量各类模型处理不同任务优劣的权威数据集如下：(1)Electricity Load 是一个从电力行业收集的大型电力负荷数据集，其中包含了 2012 至 2014 年超过 140 万条记录，包括目标值“负荷”、位置信息、天气信息、湿度信息和用户数量等多个变量。(2)COVID-19 是一个根据国家发布新冠肺炎感染情况的小数据集，包括从 2020 年 1 月 22 日到 2020 年 6 月 27 日的确诊病例、死亡病例和康复病例数据。(3)ETTh1 是北京航空航天大学收集的中国某县的电力变压器温度数据集，包括从 2016 年 7 月 1 日至 2018 年 6 月 26 日超过 1.7 万条数据记录，以 1 小时为间隔，每条记录包括目标值“油温”和 6 个电力负荷特征组成。(4)Electricity 收集了 321 个电力用户的耗电量，包括从 2012 年 1 月 1 日至 2014 年 12 月 31 日超过 2.6 万条数据记录，以 1 小时为间隔。(5)Weather 包含近 1600 个美国地区的当地气候数据，从 2010 年 1 月 1 日至 2013 年 12 月 31 日超过 3.5 万条数据记录，以 1 小时为间隔，每条记录包括目标值“湿球”和 11 个气候特征组成。

第 3 节将根据上述数据集的规模和不同算法的性能特点进行实验

2.2 评价指标

误差评价指标是衡量一个时间序列预测模型性能的重要方法，一般而言，评价指标计算出的误差越大，则模型预测的准确率越低，进而表示所建立的预测模型性能表现也就越差。目前常用的时间序列预测算法评价指标如下：

(1)平均绝对误差[25](Mean Absolute Error，MAE)，是通过计算每一个样本的预测值和真实值的差的绝对值得出，具体计算公式为：

$$MAE(y,\hat{y})=\frac{1}{m}\sum_{i=1}^{m}(|y_i-\hat{y}_i|) \tag{1}$$

MAE 的取值范围为[0，+∞]，当模型预测完全准确时，所计算出的 MAE 为 0，代表模型预测准确度达到 100%，模型是完美模型。公式中 m 为样本数量，y_i 为真实值，$\hat{y}$ 为模型的预测值，下同。

(2)均方误差[26](Mean Square Error，MSE)，是一个很实用的指标，通过计算每一个样本的预测值与真实值的差的平方再取平均值得出，具体公式为：

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

MSE 的取值范围同样是[0，+∞]，计算速度快。一直作为时序预测算法的主要评价指标之一。

(3)均方根误差[27](Root Mean Square Error，RMSE)，是均方误差进行开方得到，具体公式为：

$$RMSE(y,\hat{y})=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2} \tag{3}$$

RMSE 其取值范围依然是[0，+∞]，最终计算结果容易受数据集中的极端值影响。

(4)平均绝对百分比误差[28](Mean Absolute Percentage Error，MAPE)，是相对误差度量值，避免了正误差和负误差相互抵消，具体公式为：

$$MAPE=\frac{100\%}{m}\sum_{i=1}^{m}\left|\frac{\hat{y}_i-y_i}{y_i}\right| \tag{4}$$

该评价指标在有足够数据可用的情况下常被选用，无法处理真实值存在 0 的数据集，因为会出现分母为 0 的问题，值越小，说明预测模型拟合效果越好。

(5)决定系数 R-squared[29] 又叫可决系数(Coefficient of Determination)也叫拟合优度，其计算结果即为模型预测的准确度，取值范围为[0，1]。R^2 值越接近 1，模型性能越好；该模型等于基准模型时 $R^2=0$，R-squared 其公式为：

$$R^2(y,\hat{y})=1-\frac{\frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2}{\frac{1}{m}\sum_{i=1}^{m}(y_i-\bar{y})^2}=1-\frac{MSE}{Var} \tag{5}$$

上式中，$\bar{y}$ 为 y 的均值，分式的分子可简写为

均方误差 MSE，分式的分母可简写为方差 *Var* 。

上述五种常见评价指标中，由于 MAE、MSE 和 RMSE 等都缺少确定的上限和下限，所以无法有效判断当前预测模型的性能好坏，然而 R-squared 的计算结果位于在[0，1]的区间，使得对预测模型的评价有了更加统一的标准。研究人员在针对预测任务时所提出的各类算法往往采用不同的评价指标来证明算法的先进性。例如，在循环神经网络类算法蓬勃发展时期，研究人员采用的评价指标较为多元化，而到了采用 Transformer 类算法处理时序预测任务时，则更多地使用 MAE 和 MSE 两个评价指标。

3 基于深度学习的时间序列预测方法

最初预测任务数据量小，浅层神经网络训练速度快，但随着数据量的增加和准确度要求的不断提高，浅层神经网络已经远不能满足任务需求。近年来，深度学习引起了各领域研究者的广泛关注，深度学习方法在时间序列预测任务中与传统算法相比表现出了更强劲的性能，得到了长远发展和普遍应用。

深度神经网络与浅层神经网络相比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[30]。本节余下部分将介绍可用于解决时间序列预测问题的三大类深度学习模型。

3.1 卷积神经网络

3.1.1 卷积神经网络

卷积神经网络(Convolutional Neural Networks，CNN)是一类以卷积和池化操作为核心的深层前馈神经网络，在设计之初，其用于解决计算机视觉领域的图片识别问题[31-32]。

卷积神经网络做时间序列预测的原理是利用卷积核的能力，可以感受历史数据中一段时间的变化情况，根据这段历史数据的变化情况做出预测。池化操作可以保留关键信息，减少信息的冗余，卷积神经网络可以有效减少以往算法提取特征的人力资源消耗，同时避免了人为误差的产生。卷积神经网络所需的样本输入量巨大，多用于预测具备空间特性的数据集，其网络结构一般有五层，具体结构如图 4 所示：

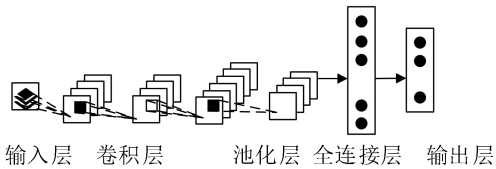


图 4　卷积神经网络结构示意图

Fig.4　CNN structure schematic

2017 年，Li 等[33]通过将时间序列的数值按一定规律排列转化为图像进行处理，再使用 CNN 模型将输入数据进行聚类，再将天气数据等外部影响因素考虑其中，来进行电力负荷预测。

3.1.2 WaveNet-CNN

2017 年，Anastasia Borovykh 等[34]受 WaveNet 这种语音序列生成模型的启发，使用 ReLU 激活函数并采用参数化跳过连接，在结构上进行了简化，改进了 CNN 模型。该模型在金融分析任务中实现了高性能，证明卷积网络不仅更简单更容易训练，同时在有噪声的预测任务上也能有优异的表现。

3.1.3 Kmeans-CNN

2017 年，随着数据集规模越来越大，而 CNN 在处理大数据集中表现不佳，Dong 等[35]选择将可以学习更多有用特征的 CNN 和分割数据的 K 均值聚类算法结合，通过将大数据集中的相似样本聚类，从而分成多个小样本来训练，在百万级大规模电力负荷数据集中表现良好。

3.1.4 TCN

2018 年，Shaojie Bai 等[36]基于 CNN 提出了一种内存消耗更低而且可并行的时间卷积网络架构(Temporal Convolutional Networks，TCN)。TCN 引入因果卷积，保证了未来信息在训练时不会被提前获取到，其反向传播路径与时间方向不同，避免了梯度消失和梯度爆炸问题。为解决 CNNs 在层数过多时导致的信息丢失问题，TCN 引入残差连接使得信息在网络间传递时可以跨层传递。

3.1.5 小结

卷积神经网络类模型在样本数量足够的情况下可用于时间序列短期预测任务，上述算法实验性能对比和总体分析如表 1 及表 2 所示：

表 1 卷积神经网络类算法多变量预测性能对比

Table 1 Comparison of multivariate prediction performance of convolutional neural network-like algorithms

算法	GPU	评价指标	Electricity Load	
			Summer	Winter
CNN[31-32]	GTX TITAN X 12GB	MAE	3.9526	12.5528
		RMSE	0.2502	0.2614
WaveNet-CNN[34]	GTX TITAN X 12GB	MAE	3.1258	8.2356
		RMSE	0.2036	0.2567
Kmeans-CNN[35]	GTX TITAN X 12GB	MAE	3.0554	**7.4102**
		RMSE	0.2194	**0.2399**
TCN[36]	GTX TITAN X 12GB	MAE	**2.9351**	7.7862
		RMSE	**0.1913**	0.2465

表 2 卷积神经网络类算法总体分析

Table 2 Overall analysis of convolutional neural network class algorithms

算法	改进方式	优势	局限
CNN[31-32]	采用一维卷积操作对时间序列的数值进行处理	可在很少或没有特征工程情况下训练	处理小样本数据集时，参数权重更新困难，预测表现甚至不如传统机器学习，
WaveNet-CNN[34]	使用 ReLU 激活函数和参数化跳过连接的方法	结构简单易于训练，抗噪声能力突出	只能处理有限长度的时间序列，预测过长的时间序列表现不佳
Kmeans-CNN[35]	使用K均值聚类算法切分数据集，使用其子集进行训练	处理有多分类倾向的大数据集时效率和准确度显著提高	在数据量小场景中预测效果不佳，不适合处理没有明显特征用于聚类的数据
TCN[36]	引入了因果卷积，扩张卷积和残差连接	并行性好，梯度稳定，在长序列输入情况下，内存消耗更低	感受野与目前先进模型比仍有差距

从表 1 中可以看出模型在样本量巨大的多变量数据集上处理短期预测任务时，Kmeans-CNN 采用先聚类分类再由模型训练的思路取得了比较理想的预测效果，后续也有不少研究人员在解决时序预测问题时进行类似处理。引入了扩展卷积和残差连接等架构元素的 TCN 能保有更长的有效历史信息，同样达到了不错的预测效果，而且其网络较为简单清晰。

目前，CNNs 的预测精度与循环神经网络等其他网络结构相比已不占优势，难以单独处理步长较长的时序预测问题，但常作为一个功能强大的模块接入其他先进算法模型中用于预测任务。

3.2 循环神经网络

3.2.1 循环神经网络

循环神经网络(Recurrent Neural Networks，RNN)是由 M. I. Jordan 在 1990 年提出的用于学习时间维度特征的深度学习模型[37]。

RNN 的各单元以长链的形式连接在一起按序列发展的方向进行递归，模型的输入是序列数据，可用于处理自然语言处理的各种任务（例如文本情感分类、机器翻译等），也可与 CNN 组合构筑成解决计算机视觉领域问题的模型 RNN 同 CNN 一样，参数是共享的，因此在处理时间序列数据、语音数据时能体现出较强的学习能力，通过识别数据的顺序特征并使用先前的模式来预测，具体结构如图 5 所示：

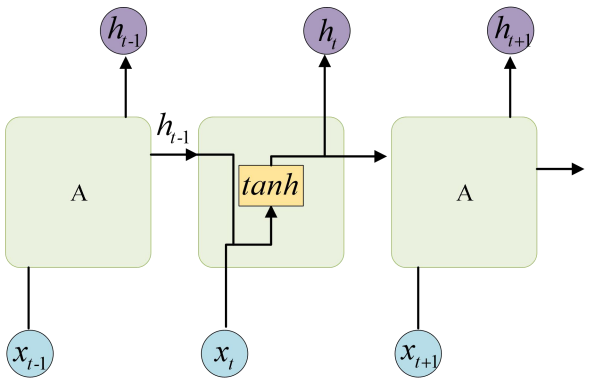


图 5　循环神经网络结构示意图

Fig.5　RNN structure schematic

其中 x_t 表示 t 时刻的输入向量，h_t 表示 t 时刻的隐藏向量，可以看到传统 RNN 神经元会接受上一时刻的隐藏状态 h_{t-1} 和当前输入 x_t。

使用 RNN 训练容易出现很严重程度的梯度消失问题或者是梯度爆炸的问题。梯度的消失问题主要是因为在神经网络模型中位于最前面层的网络权重无法及时进行有效的更新，导致训练失败；梯度爆炸问题是指由于迭代参数的改变幅度太过剧烈，导致学习过程不平衡。随着数据长度的提升，该问题愈加明显，导致 RNN 只能有效捕捉短期规律，即仅具有短期记忆。

1997 年，Mike Schuster 等[38]将常规循环神经网络 RNN 扩展到双向循环神经网络(Bidirectional Recurrent Neural Networks，Bi-RNN)。Bi-RNN 通过同时在前向和后向上训练，不受限制地使用输入信息，直到预设的未来帧，可同时获得过去和未来的特征信息。在人工数据的回归预测实验中，Bi-RNN 与 RNN 训练时间大致相同并取得了更好的预测效果。

3.2.2　长短期记忆网络

长短期记忆网络(Long Short-term Memory，LSTM)于 1997 年被 Hochreiter 提出，用于解决 RNNs 模型的诸多问题[39]。LSTM 循环单元结构如图 6 所示：

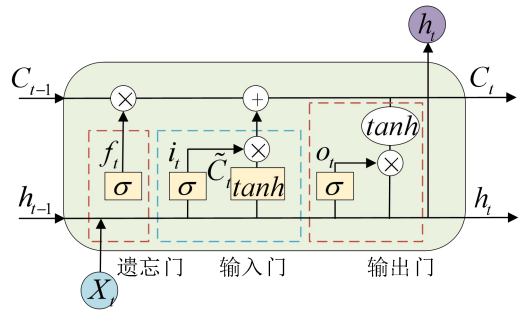


图 6　LSTM 单元结构示意图

Fig.6　LSTM cell structure schematic

LSTM 的神经元在 RNN 的基础上还增加了一个 cell 状态 c_{t-1}，与 RNN 中 h 的作用相似，都是用来保存历史状态信息的。LSTM 采用三个门来选择忘记和记住一些关键信息。

遗忘门和输入门都作用于单元的内部状态，分别控制遗忘多少前一个时间步内部状态的信息和吸收多少当前时刻的输入信息，若门的值为 0，即不遗忘和完全不吸收，若门的值为 1，即完全遗忘和全部吸收。输出门在隐层 h_t 发挥作用，主要决定该单元的内部状态对系统整体状态的影响多少[40]。

王鑫 等[41]提出了一种基于 LSTM 的单变量故障时间序列预测算法应用于航空领域的飞机数据案例，对比多元线性回归模型，支持向量回归等多个模型，最终 LSTM 模型表现出更好的性能。

2005 年，A Graves 等[42]提出的双向长短期记忆网络(Bidirectional Long Short-term Memory，Bi-LSTM)神经网络结构模型结构类似于 Bi-RNN，其由两个独立的 LSTM 拼接而成。Bi-LSTM 的模型设计初衷是克服 LSTM 无法利用未来信息的缺点，使 t 时刻所获得特征数据同时拥有过去和将来的信息[43]。由于 Bi-LSTM 利用额外的上下文而不必记住以前的输入，所以处理较长时间延迟的数据时表现出更强大的能力。经实验表明，没有时间延迟的 LSTM 几乎返回同样的结果，这代表着在部分时间序列数据中向前训练和向后训练两个方向上的上下文同样重要，Bi-LSTM 的特征提取能力明显高于 LSTM。

3.2.3 门控循环单元

门控循环单元(Gated Recurrent Unit，GRU)是由Kyunghyun Cho 等[44]在 2014 年通过改进 LSTM 模型提出的，具体循环单元结构图 7 所示：

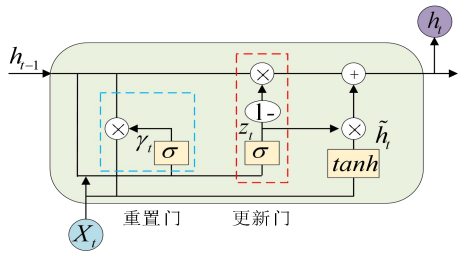


图 7 GRU 单元结构示意图

Fig.7 GRU cell structure diagram

GRU 相较于 LSTM 简化了结构，图中的 z_t 和 r_t 分别表示 GRU 仅有的更新门和重置门。重置门决定着前一状态的信息传入候选状态的比例。更新门是将 LSTM 的遗忘门和输出门的功能组合在一起，用于控制前一状态的信息 h_{t-1} 有多少保留到新状态 h_t 中，GRU 的门的计算方式和 LSTM 类似，因此参数比 LSTM 少得多，从而训练时间更少，而且在多个数据集中的表现证明 GRU 有不亚于 LSTM 的准确度表现。

文献[45]首次将 GRU 应用于交通流量预测并与 LSTM 模型作对比进行实验，在 MAE 评价指标下，GRU 的表现比 LSTM 模型低 5%左右。

对于电子商务中广泛存在的促销销售预测任务，Qi Y 等[46]提出了一种基于 GRU 的算法来明确建模目标产品与其替代产品之间的竞争关系。Xin S 等[47]提出的另一项工作将异构信息融合到修改后的 GRU 单元中，以了解促销活动前的预售阶段的状态。

3.2.4 小结

RNNs 循环神经网络类算法自提出就一直是解决时间序列预测任务的重要方法，常常作为一个模块嵌入到其他算法中来获得更好的预测效果，在 2017 年以前一直作为解决时间序列数据预测问题的主力模型，得到广泛应用。主要循环神经网络类算法实验性能对比和总体分析如表 3 和表 4 所示：

表 3 循环神经网络类算法单变量预测性能对比

Table 3 Comparison of univariate prediction performance of recurrent neural network-like algorithms

算法	GPU	评价指标	COVID-19 Confirmed cases	COVID-19 Death cases 48	COVID-19 Recovered cases
RNN[37]	NVIDIA GTX 1070 8GB	MAE	4.2365	0.0536	12.5231
		RMSE	4.3512	0.0552	15.3524
		R-squared	0.9425	0.9345	0.9651
Bi-RNN[38]	NVIDIA GTX 1070 8GB	MAE	2.1026	0.0272	7.3258
		RMSE	2.2154	0.0231	8.3215
		R-squared	0.9986	0.9752	0.9975
LSTM[39]	NVIDIA GTX 1070 8GB	MAE	**2.0463**	0.0095	7.5628
		RMSE	2.2428	0.0103	8.5216
		R-squared	0.9982	0.9979	**0.9996**
Bi-LSTM[42]	NVIDIA GTX 1070 8GB	MAE	2.1121	**0.0070**	**5.5398**
		RMSE	**2.0635**	**0.0077**	**6.5915**
		R-squared	0.9976	**0.9997**	0.9987
GRU[44]	NVIDIA GTX 1070 8GB	MAE	2.8558	0.0321	7.0486
		RMSE	3.3158	0.0402	8.4009
		R-squared	**0.9989**	0.9981	0.9976

表 4　循环神经网络类算法总体分析

Table 4　Overall analysis of recurrent neural network class algorithms

算法	改进方式	优势	局限
RNN[37]	节点按照链式连接	参数共享，能有效捕捉短期记忆	无法捕捉长期规律，存在严重的梯度消失和梯度爆炸问题
Bi-RNN[38]	同时在正序和负序上不受限制地使用输入信息训练	训练时间与 RNN 大致相同，并取得了更好的预测结果	梯度消失梯度爆炸问题严重
LSTM[39]	通过输入门、输出门和遗忘门在单元内部建立内循环	适合处理时间序列任务中间隔长的任务，准确率更高	参数过多，梯度消失和梯度爆炸问题仍然存在，仅能使用过去信息
Bi-LSTM[42]	由两个不同的 LSTM 隐藏层组成	过去和未来的信息均可被利用，泛化性强	训练参数过多，不能并行处理，梯度消失和梯度爆炸问题仍然存在
GRU[44]	将 LSTM 的遗忘门和输入门整合为更新门	同条件下所得出的预测精度与 LSTM 相当，训练参数更少，训练速度更快	去掉了跟踪中间数据值的单元，GRU 不能像 LSTM 那样有效地控制数据流，序列过长时梯度消失和梯度爆炸问题同样可能发生

表 3 可以看出，GRU 和 LSTM 在性能上相当，但都受限于只能从一个方向上学习训练，在预测精度上要低于可以从两个方向上获取信息的 Bi-LSTM 模型。Bi-LSTM 在解决短期时序预测任务时的优势包括所需的样本数量少，拟合速度快，预测精度高，如今依然有众多学者研究使用。

循环神经网络类方法可以捕获并利用长期和短期的时间依赖关系来进行预测，但在长序列时间序列预测任务中表现不好，并且 RNNS 多为串行计算，导致训练过程中对内存的消耗极大，而且梯度消失和梯度爆炸问题始终没有得到彻底解决。

3.3Transformer 类模型

介绍 Transformer 模型之前先要介绍一下注意力机制，人类眼睛的视角广阔，但局限于视觉资源，往往重点关注视线中的特定部分，注意力机制就是以此为灵感提出，重点关注数据中的更有价值的部分[48-49]。

Transformer 所采用的自注意力机制所解决的情况是：神经网络的输入是很多大小不一的向量，不同时刻的向量往往存在着某种潜在联系，实际训练的时候无法充分捕捉输入之间的潜在联系而导致模型训练结果较差。

一个自注意力模块接收 n 个输入，然后返回 n 个输出，其中的所有输入都会彼此作用，挖掘出其中作用明显的注意力点，这些相互作用的聚合和注意力分数即为模块给出的输出。自注意力机制的输入(Query，Key)，计算公式为：

$$A(Q,K,V)=Softmax(\frac{QK^{\mathrm{T}}}{\sqrt{d}})V \tag{6}$$

其中，$Q\in R^{L_Q\times d},K\in R^{L_K\times d},V\in R^{L_V\times d}$，$d$ 表示输入维度，第 i 个 Query 的注意力系数的概率公式是：

$$A(q_i,K,V)=\sum_j\frac{k(q_i,k_i)}{\sum_l k(q_i,k_l)}v_j=E_{p(k_i|q_i)}[V_j] \tag{7}$$

其中，$p(k_i,q_i)=\frac{k(q_i,k_i)}{\sum_l k(q_i,k_l)}$，$k(q_i,k_i)$ 选择非对称指数 $exp(\frac{q_i k_j^{\mathrm{T}}}{\sqrt{d}})$。

3.3.1　Transformer

Vaswani 等[50]提出了 Transformer 这种与以往的 CNNs 或者 RNNs 结构不同的新的深度学习框架。Transformer 是完全依赖注意力机制来表征模型的输入和输出之间的全局依赖关系，具体结构如图 8 所示：

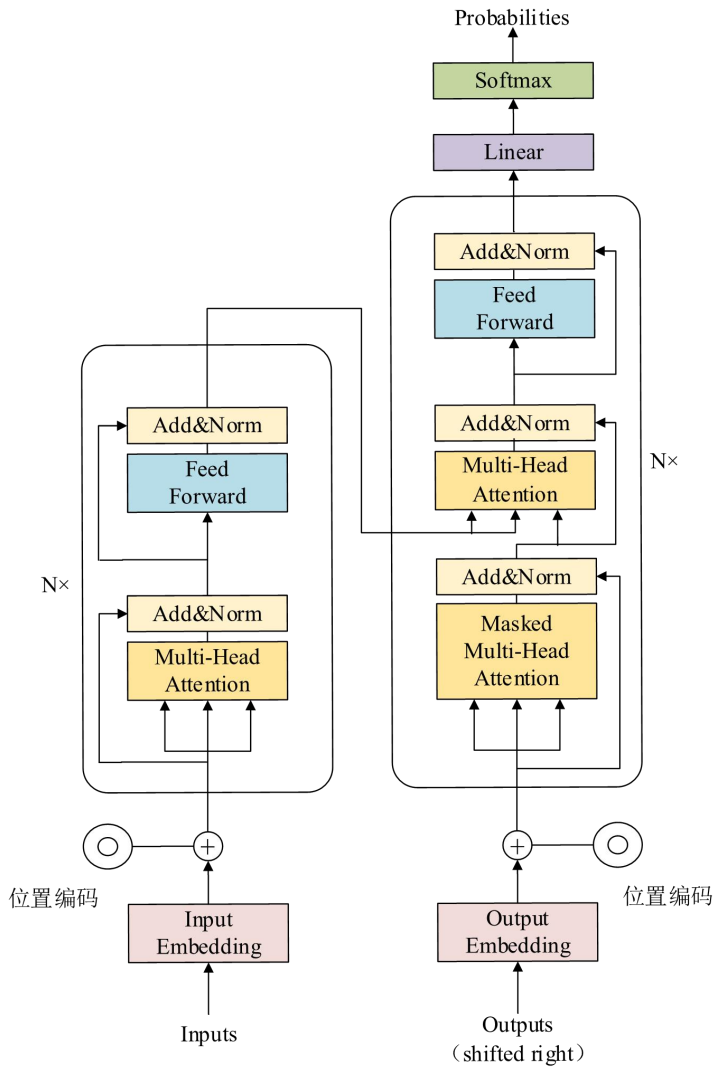


图 8 Transformer 结构示意图

Fig.8 Transformer structure schematic

图 8 中的 N 是一个超参数，表示编码器和解码器部分是由多个相同的层叠起来。

Transformer 的核心是自注意力模块，它可以被视为一个完全连接层，其权重是基于输入模式的成对相似性而动态生成的。其参数数量少，同条件下所需计算量更少，使其适合建模长期依赖关系[51]。

相较于 RNNs 的模型，就算使用 LSTM 和 GRU 也不能避免梯度消失和梯度爆炸的问题：随着网络往后训练，梯度越来越小，要走 $n-1$ 步才能到第 n 个词，而 Transformer 的最长路径仅为 1，解决了长期困扰 RNNs 的问题。Transformer 捕捉长期依赖和彼此交互的突出能力对于时间序列建模任务有巨大吸引力，能在各种时间序列任务中表现出高性能[52]。

3.3.2 BERT

2018 年 10 月，Google 的 BERT(Bidirectional Encoder Representation from Transformers)模型[53]横空出世，并横扫自然语言处理领域 11 项任务的最佳成绩，随后 Transformer 模型运用于各大人工智能领域。

2021 年，KH Jin 等[54]为克服交通流量预测所需道路天气数据繁杂，通用性差和应用局限等缺点，提出了 trafficBERT 这种适用于各种道路的模型。该模型通过多头自注意力来代替预测任务常用的

RNNs 来捕获时间序列信息，还通过分解嵌入参数化来更有效地确定每个时间步之前和之后状态之间的自相关性，只需要有关交通速度和一周内几天的道路信息，不需要当前时刻相邻道路的流量信息，应用局限性小。

3.3.3 AST

2020 年，Sifan Wu 等[55]应用生成对抗思想在 Sparse Transformer[56] 基础上提出了对抗稀疏 Transformer(Adversarial Sparse Transformer，AST)。

大多数点预测模型只能预测每个时间步的准确值，缺乏灵活性，难以捕捉数据的随机性，在推理过程中常常被网络自己的一步超前输出代替，导致推理过程中的误差累积，由于误差累积，它们可能无法预测长时间范围内的时间序列。大多数时间序列预测模型会优化特定目标，例如最小化似然损失函数或分位数损失函数，然而这种强制执行步级精度的精确损失函数无法处理时间序列中的真实随机性，从而导致性能下降。

AST 模型通过对抗训练和编码器-解码器结构可以更好地表示时间序列，并在序列级别以更高的保真度预测时间序列的多个未来步骤来缓解上述问题，并使用鉴别器来提高序列级别的预测性能。实验表明，时间序列步骤之间的依赖关系具有一定的稀疏性，AST 采用的对抗训练可以从全局角度改善时间序列预测，基于编码器-解码器的 Transformer 的性能优于仅采用自回归解码器的 Transformer。

3.3.4 Informer

2021 年，北京航空航天大学的 Haoyi Zhou 等[57]在经典的 Transformer 编码器-解码器结构的基础上提出了 Informer 模型来弥补 Transformer 类深度学习模型在应用于长序列时间预测问题时的不足。在此之前解决预测一个长序列的任务往往采用多次预测的方法，而 Informer 可以一次给出想要的长序列结果，Informer 具体结构如图 9 所示：

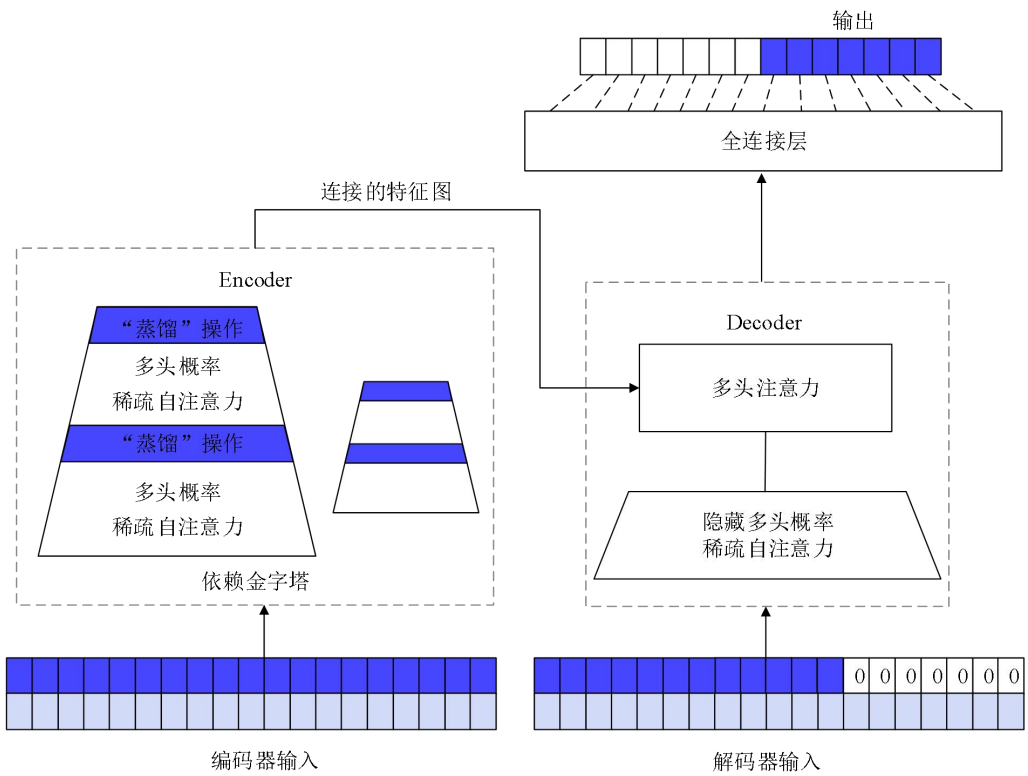


图 9　Informer 结构示意图

Fig.9　Informer structure schematic

Informer 具有三个显著特点：(1)ProbSparse 自注意力机制：在 Informer 的整体结构图中，编码器部分采用多头稀疏自注意力替换了 Transformer 模型传统的自注意力，可以有效处理较长的序列输入。(2)自注意力提炼：蓝色梯形部分是提取主导注意力的自注意力蒸馏部分，大大减少了网络的层数，并且提高了层堆叠部分的鲁棒性。(3)生成式解码器：解码器部分将预测序列及之后的数据置为 0 来进行遮挡，分析特征图的注意力权重，随后生成预测的结果，序列输入只需要一个前向步骤，有效避免了误差的累计。

Informer 在自我注意模型中引入了稀疏偏差，以及 Logsparse 掩码，从而将传统 Transformer 模型的计算复杂度从 $O(L^2)$ 降低到 $O(LlogL)$，它没有显式引入稀疏偏差，而是根据查询和关键相似性选择 $O(LlogL)$ 占主导地位的查询，从而在计算复杂度上实现较好的改进。长序列的预测在极端天气的预警和长期能源消耗规划等实际应用中尤为重要，Informer 能在长时间序列任务上表现出的优越的性能。

3.3.5 TFT

2021 年，Lim B 等[58]提出的 TFT(Temporal Fusion Transformers)设计了一个包含静态协变量编码器、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。

已经提出的几种深度学习方法，通常都是“黑盒”模型，没有阐明它们如何使用实际场景中存在的全部输入。TFT 编码不仅各种协变量信息中选择有用的信息来执行预测，它还保留了包含全局、时间依赖性和事件的可解释性。

3.3.6 SSDNet

2021 年，Yang L 等[59]提出的空间状态空间分解神经网络(State Space Decomposition Neural Network，SSDNet)，将 Transformer 深度学习架构和状态空间模型(State Space Models，SSM)相结合，兼顾了深度学习的性能优势和 SSM 的可解释性。

SSDNet 采用 Transformer 架构来学习时间模式并直接估计 SSM 的参数，为了便于解释，使用固定形式的 SSM 来提供趋势和季节性成分以及 Transformer 的注意力机制，以识别过去历史的哪些部分对预测最重要。

评估 SSDNet 在太阳能、电力、交易所等五个数据集的时间序列预测任务上的性能，结果表明，SSDNet 比最先进的深度学习模型 DeepAR[60]、DeepSSM[61]、LogSparse Transformer、Informer 和 N-BEATS[62]以及统计模型 SARIMAX[63]和 Prophet[64]的预测准确度更高。

3.3.7 Autoformer

2021 年，Wu H 等[65]提出的 Autoformer 设计了一种简单的季节性趋势分解架构。Autoformer 继承使用 Transformer 的编码器-解码器结构，通 Autoformer 采用的独特内部算子能够将变量的总体变化趋势与预测的隐藏变量分离，这种设计可以使模型在预测过程中交替分解和细化中间结果，其采用独特的自相关机制，这种逐级机制实现了长度-L 系列的 $O(LlogL)$ 复杂度，并通过将逐点表示聚合扩展到子序列级别来打破信息利用瓶颈，在多个公开数据集中表现出优异的性能。

3.3.8 Aliformer

电子商务中，产品的趋势和季节性变化很大，促销活动严重影响销售导致预测难度较大对算法要求更高。

2021 年，阿里巴巴的 Qi X 等[66]为解决电子商务中准确的时间序列销售预测问题，提出基于双向 Transformer 的 Aliformer 利用历史信息、当前因素和未来知识来预测未来的数值。Aliformer 设计了一个知识引导的自注意力层，使用已知知识的一致性来知道时序信息的传输，并且提出未来强调训练策略，使模型更加注重对未来知识的利用。

对四个公共基准数据集(ETTh1、ETTm1、ECL2、Kaggle-M53)和一个大规模的天猫商品销售数据集- TMS 进行的广泛实验表明，Aliformer 在销售预测问题中可以比最先进的时间序列预测方法表现更好。

3.3.9 FEDformer

2022 年，Tian Zhou 等[67]提出的 FEDformer

(Frequency Enhanced Decomposed Transformer)设计了两个注意模块，分别用傅里叶变换[68]和小波变换[69]处理频域中应用注意力操作。

FEDformer 将广泛用于时间序列分析的季节性趋势分解方法[70]融入到基于 Transformer 的方法中，还将傅里叶分析与基于 Transformer 的方法结合起来，没有将 Transformer 应用于时域，而是将其应用于频域，这有助于 Transformer 更好地捕捉时间序列的全局特征。

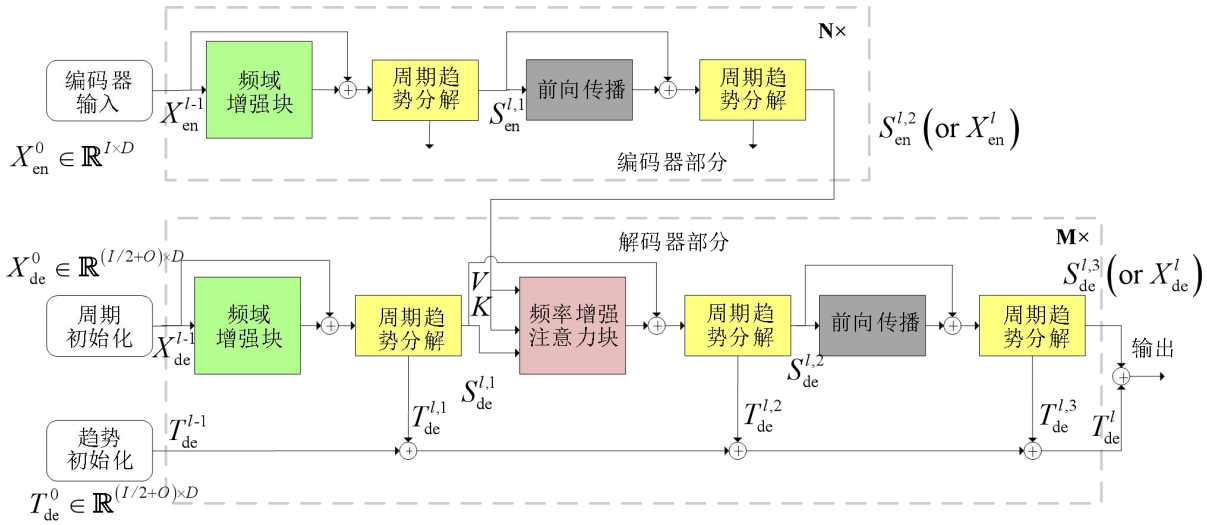


图 10　FEDformer 结构示意图

Fig.10　FEDformer structure schematic

图 10 中频率增强块(Frequency Enhanced Block，FEB)和频率增强注意力(Frequency Enhanced Attention，FEA)，二者用于在频域中进行表示学习，周期趋势分解块 MOE 用于从输入数据中提取周期趋势特征。

编码器部分采用多层结构 $X_{\text{en}}^{l}=\text{Encoder}\left(X_{\text{en}}^{l-1}\right)$，其中 $l\in\{1,\cdots,N\}$ 表示该层编码器的输出，$X_{\text{en}}^{0}\in\mathbb{R}^{I\times D}$ 是嵌入式历史时间序列，编码器可形式化为如下公式：

$$S_{\text{en}}^{l,1},\text{-}=\text{MOEDecomp}\left(\text{FEB}\left(X_{\text{en}}^{l-1}\right)+X_{\text{en}}^{l-1}\right) \tag{8}$$

$$S_{\text{en}}^{l,2},\text{-}=\text{MOEDecomp}\left(\text{FeedForward}\left(S_{\text{en}}^{1}\right)+S_{\text{en}}^{1}\right) \tag{9}$$

$$X_{\text{en}}^{l}=S_{\text{en}}^{l,2} \tag{10}$$

其中 $S_{\text{en}}^{l,i},i\in\{1,2\}$ 分别表示第 l 层中第 i 个分解块之后的季节性分量。对于 FEB 模块，它有两个不同的版本(FEB-f 和 FEB-w)，分别通过离散傅里叶变换[71]和离散小波变换[72]机制来实现，可以无缝地替换自注意力模块。

解码器部分同样采用多层结构 $X_{\text{de}}^{l},T_{\text{de}}^{l}=\text{Decoder}\left(X_{\text{en}}^{l-1},T_{\text{de}}^{l-1}\right)$，其中 $l\in\{1,\cdots,M\}$ 表示第 l 层解码器的输出。解码器可形式化为如下公式：

$$S_{\text{de}}^{l,1},T_{\text{de}}^{l,1}=\text{MOEDecomp}\left(\text{FEB}\left(X_{\text{de}}^{l-1}\right)+X_{\text{de}}^{l-1}\right) \tag{11}$$

$$S_{\text{de}}^{l,2},T_{\text{de}}^{l,2}=\text{MOEDecomp}\left(\text{FEA}\left(S_{\text{de}}^{l,1},X_{\text{en}}^{N}\right)+S_{\text{de}}^{l,1}\right) \tag{12}$$

$$S_{\text{de}}^{l,3},T_{\text{de}}^{l,3}=\text{MOEDecomp}\left(\text{FeedForward}\left(S_{\text{de}}^{l,2}\right)+S_{\text{de}}^{l,1}\right) \tag{13}$$

$$X_{\text{de}}^{l}=S_{\text{de}}^{l,3} \tag{14}$$

$$T_{\text{de}}^{l}=T_{\text{de}}^{l-1}+W_{l,1}\cdot T_{\text{de}}^{l,1}+W_{l,2}\cdot T_{\text{de}}^{l,2}+W_{l,3}\cdot T_{\text{de}}^{l,3} \tag{15}$$

其中 $S_{\text{en}}^{l,1},T_{\text{en}}^{l,1},i\in\{1,2,3\}$ 分别表示第 l 层中第 i 个分解块之后的周期性和趋势性分量。$W_{l,i},i\in\{1,2,3\}$ 表示第 i 个提取到的趋势 $T_{\text{de}}^{l,i}$。FEA 同样有两个不同的版本(FEA-f 和 FEA-w)，分别通过离散傅里叶变换和离散小波变换机制投影实现，并具有注意力设计可以替换交叉注意力模块。最终预测结果是两个精细分解分量的总和，即 $W_S\cdot X_{\text{de}}^{M}+T_{\text{de}}^{M}$，其中 W_S 可将深度变换的季节分量 X_{de}^{M} 投影到目标维度。

FEDformer 通过傅里叶变换中的随机模式部分实现了线性复杂度，部分相关算法复杂度分析如表 5 所示：

表 5 不同预测模型的复杂度分析

Table 5 Complexity analysis of different forecasting models

算法	Training		Testing
	Time	Memory	Steps
FEDformer[67]	$O(L)$	$O(L)$	1
Pyraformer[73]	$O(L)$	$O(L)$	1
Autoformer[65]	$O(L\log L)$	$O(L\log L)$	1
Informer[57]	$O(L\log L)$	$O(L\log L)$	1
Transformer[50]	$O(L^2)$	$O(L^2)$	L
GRU[44]	$O(L)$	$O(L)$	L
LSTM[39]	$O(L)$	$O(L)$	L

需要指出的是，自 FEDformer 提出以来，时间序列数据在频域或时频域中的独特属性在时间序列预测领域中引起了广泛的关注。

3.3.10 Pyraformer

2022 年，Shizhan Liu 等[73]提出的 Pyraformer，这是一种基于金字塔注意力的新型模型，可以有效地描述短期和长期时间依赖关系，且时间和空间复杂度较低。

Prayformer 首先利用更粗尺度构造模块(coarser scale construction module，CSCM)构造多分辨率 C 叉树，然后设计金字塔注意模块以跨尺度和尺度内的方式传递消息，当序列长度 L 增加时，通过调整 C 和固定其他参数，Pyraformer 可以达到理论 $O(L)$ 复杂度和 $O(1)$ 最大信号遍历路径长度。实验结果表明，Prayformer 模型在单步和多步预测任务中都优于最先进的模型，而且计算时间和内存成本更少。

3.3.11 Conformer

2023 年，Yan Li 等[74]为解决有明显周期性的长序列预测任务的效率和稳定性问题，提出了一种针对多元长周期时序预测的 Conformer 模型。

该模型采用快速傅里叶变换对多元时间做处理，以此来提取多元变量的相关性特征，完成了多个变量之间关系的建模，以及月、周、天、小时等不同频率下规律性的提取。为了提升长周期预测的运行效率，Conformer 采用了滑动窗口的方法，即每个位置只和附近一个窗口内的邻居节点结算 attention，牺牲了全局信息提取和复杂序列模型建模能力，从而将时间复杂度降低到 $O(L)$。Conformer 又提出了静止和即时循环网络模块，使用 GRU 编码输入时间序列，来提取全局信息弥补滑动窗口方法造成的全局信息损失。

为解决高位多元时间序列联合建模所形成的分布复杂的问题，Conformer 采用标准化流操作，即用 GRU 产出的全局信息和解码器信息进行标准化流的初始化，然后进行一系列映射得到目标分布后进行预测。

3.3.12 小结

Transformer 类算法如今广泛用于人工智能领域的各项任务，在 Transformer 基础上构建模型可以打破以往算法的能力瓶颈，可以同时具备良好的捕捉短期和长期依赖的能力，有效解决长序列预测难题，并且可以并行处理。上述算法性能对比和总体分析如表 6 和表 7 所示：

表 6 Transformer 类算法多变量预测性能对比

Table 6 Comparison of multivariate prediction performance of Transformer class algorithms

算法	GPU	评价指标	ETTh1			Electricity			Weather		
			48	192	720	48	192	720	48	192	720
AST[55]	NVIDIA V100 32GB	MSE	1.436	2.651	3.021	0.388	0.442	0.542	0.441	0.543	0.687
		MAE	1.227	1.559	1.631	0.402	0.493	0.637	0.465	0.556	0.756
Informer[57]	NVIDIA V100 32GB	MSE	1.575	2.251	2.903	0.287	0.296	0.373	0.388	0.545	0.615
		MAE	1.086	1.625	1.410	0.378	0.386	0.439	0.421	0.567	0.675
TFT[58]	NVIDIA V100 32GB	MSE	1.654	1.924	2.931	0.366	0.456	0.512	0.368	0.568	0.685
		MAE	1.125	1.852	1.658	0.412	0.496	0.553	0.411	0.577	0.721

Autoformer[65]	NVIDIA V100 32GB	MSE	**0.703**	**1.495**	**1.624**	0.189	0.222	0.254	**0.226**	0.307	0.419
		MAE	**0.671**	**1.102**	**1.403**	0.269	0.334	0.361	**0.241**	0.367	**0.428**
FEDformer[67]	NVIDIA V100 32GB	MSE	0.752	1.524	1.923	**0.185**	**0.201**	**0.246**	0.247	**0.274**	**0.399**
		MAE	0.735	1.228	1.452	0.261	0.315	0.355	0.289	**0.331**	0.435
Pyraformer[73]	Titan Xp 12GB	MSE	0.801	1.926	2.201	0.256	0.288	0.371	0.286	0.385	0.432
		MAE	0.736	1.255	1.684	**0.254**	**0.285**	**0.346**	0.391	0.415	0.465

表 7　Transformer 类算法总体分析

Table 7　Overall analysis of Transformer class algorithms

算法	改进方式	优势	局限
trafficBERT[54]	在 BERT 的双向 Transformer 结构基础上，通过分解嵌入参数化改进	更有效确定每个时间步前后状态之间的相关性，泛化性强，便于迁移学习	在真实值出现剧烈下降趋势时，预测能力稍显不足，没法准确预测下降幅度
AST[55]	引入对抗损失函数，将 Transformer 嵌入对抗神经网络，使用 sparse Transformer 模型，采用 alpha_entmax 计算稀疏注意力权重	对抗训练可以从全局的角度改善时间序列预测，能捕捉到时间序列数据稀疏性倾向的依赖关系，提高了模型鲁棒性	注意力图中参数值影响较大，分配过于稀疏的注意力来学习底层关系会降低性能，分配过于密集的注意力会浪费注意力在不相关的步骤上，导致性能较差
Informer[57]	采用 ProbSparse 自关注机制、自注意力提炼和生成式解码器	降低了计算复杂度，降低了内存消耗量，更适用于长序列的预测	在预测曲线的波峰波谷处仍有较大的误差，难以应对长序列越来越高的精度要求，降低计算复杂度但牺牲了信息利用率
TFT[58]	采用可解释性更好的时间自注意力解码器，利用特定的组件来选择相关特征，并利用门控层来抑制不必要的特征	能有效应对多元异构的输入，逐层筛选非必要的特征，去噪能力明显，可以更好地捕捉到长期依赖在实现高性能的同时兼顾可解释性	在数据曲线的在波动性较低时，对过去的输入给予同等的注意力权重，对于关键信息的特征提取能力受限
SSDNet[59]	采用 Transformer 架构来学习时间模式，提取潜在组件并估计 SSM 的参数，它应用 SSM 生成具有非平稳趋势和周期性成分的可解释预测结果	结合了无需密集特征工程以及从时间序列推断共享模式的优势和SSM模型的可解释性，准确度高	
Autoformer[65]	增加了时序拆解模块，修改了 self-attention 模块，提出了能更好挖掘数据规律的自相关机制	明显降低计算复杂度并可以更好捕捉特征信息，在明显没有周期性的数据集中预测准确率改进较大	过度依赖寻找时序数据的周期特性，不适合对周期性较弱的数据集上训练
Aliformer[66]	通过添加知识引导分支来修改注意力图以最小化噪声的影响，通过在序列中间添加跨度掩码来强调未来知识的重要性	不仅在拟合趋势方面有效，而且在处理曲线剧烈变化方面也很有效，在销售预测领域达到了新的最先进的性能	模型注重对未来知识的挖掘应用，对于未来信息难以预估和获取的任务并不友好
FEDformer[67]	提出了周期项趋势项分解混合专家机制的频率增强的分解 Transformer，用 Fourier 增强模	在长期序列预测任务中，可以捕捉到其他基于 Transformer 算法无法捕捉到	对于频率很高的时间序列，该算法的整体计算成本较高，同时会产生大量的冗余信息

	块和 Wavelet 增强模块替代自注意力模块和交叉注意力模块	的时间序列的全局视图，且计算成本低，鲁棒性强	
Pyraformer[73]	设计了一个带有二叉树跟随路径的分层金字塔注意模块，以捕获具有线性时间和内存复杂度的不同范围的时间依赖性	内存消耗小，能利用小内存设备高效完成单步和超长多步预测，且批次训练时间很短	
Conformer[74]	融入傅里叶变换、多频率序列采样、周期项趋势项分解、标准化流等多个优化	提升了长周期预测的效率和稳定性，时间序列预测以生成的方式产生，抗噪声能力强	难以捕捉到结构化较差的时间序列数据中的时间模式

从表 6 可以看出 Transformer 类算法为避免过拟合需要大量数据来进行自身的训练，在中期和长期预测任务上都有着不错的性能表现。

目前，部分 Transformer 类算法在保留编码器–解码器架构的同时，开始重新审视注意力机制的作用，因为在错综复杂的长序列预测任务中自注意力机制可能不可靠。Informer 等在降低复杂度的同时选择牺牲了一部分的有效信息，Conformer 使用局部注意力与全局的 GRU 进行功能互补。

Pyraformer 在相对较低的配置下依然表现出不错的性能，一定程度上缓解了 Transformer 类算法设备要求高的问题，适合在欠发达地区普及使用。

4 总结与展望

文章在对时间序列数据、经典时间序列参数模型和算法评价指标简单介绍后，系统总结了基于深度学习的时间序列预测算法，其中以基于 Transformer 的模型为主，深入分析了 Transformer 类算法的网络架构优缺点，在注意力机制被提出以来，时间序列预测任务发展进入快车道取得了令人瞩目的成果。下面列出了时间序列预测领域的重点问题和进一步的研究方向，以促进时间序列预测算法的研究和完善。

但是终归是调参工作

(1)采用随机自然启发优化算法优化深度学习模型的多个超参数。深度学习算法愈发复杂，需要处理的超参数越来越多，超参数的选择往往决定着算法能不能突破局部最优陷阱达到全局最优。随机自然启发优化算法灵感来自群体智能的各种现象、为动物的自然行为、物理定律以及进化定律。优化算法首先基于问题的约束随机生成一定数量的可解解，然后利用算法的各阶段重复寻找全局最优解，在限制范围内寻找最优的超参数以提升模型预测能力。因此，采用随机自然启发优化算法用于模型最优超参数寻找，将成为未来研究热点之一。

(2)研究适合时间间隔不规则的小数据集的网络架构。现有 Transformer 模型架构复杂，参数多，在周期性好的数据集上表现出优越的性能，但在数据量小，时间间隔不规则的数据集中表现不理想。Transformer 类模型为在小数据上的过拟合问题值得进一步思考和解决。处理时间间隔不规则的数据集时，在模型架构中引入重采样、插值、滤波或其他方法是处理时间序列数据和任务特征的新思路，会是未来一个新的研究方向。

(3)引入图神经网络(graph neural network，GNN)用于多变量时序预测建模。由于多变量时序预测任务的潜在变量相关性十分复杂，且在现实世界中的数据相关性是变化的，导致准确多变量预测具有挑战性。最近不少学者采用时间多项式图神经网络将动态变量相关性表示为动态矩阵多项式，可以更好地理解时空动态和潜在的偶然性，在短期和长期多变量时序预测上都达到了先进的水平。因此 GNN 对多变量时序预测的强大建模能力值得进一步研究。

(4)研究同时支持精确形状和时间动态的可微损失函数作为评价指标。在时间序列预测领域中已经使用了许多测量度量，并且基于欧氏距离的点误差损失函数，例如 MSE，被广泛用于处理时间序列

数据，但是其逐点映射，对形状和时间延后失真不具有不变性。损失函数不仅要最小化预测和目标时间序列之间的差距还应该考虑整个输出序列和基本事实之间的相关性，从而帮助模型生成更及时、更稳健和更准确的预测，而不是仅仅逐点优化模型。如果损失函数能在曲线形状和时间感知上对模型进行评价能更有利于训练出高效准确的时间序列预测模型。

5 结束语

数据维度扩张，数据量级别增大，应用场景需求变换依旧给时间序列预测任务带来巨大的挑战。基于深度学习的时间序列预测算法，目前看来具有一定的性能优势，但仍需要进一步的提升和完善。本文以时序数据特性、常用数据集和评价指标为引，以基于深度学习时序预测算法发展时间线为主线，将卷积神经网络类算法、循环神经网络类算法和 Transformer 类算法进行性能分析和优缺点综述，最后对深度学习应用于时间序列预测算法的发展趋势进行了总结与展望。

参考文献：

[1] Yuan Y, Lin L. Self-supervised pretraining of transformers for satellite image time series classification[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 474-487.

[2] 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, Singapore, Aug 14-18, 2021. New York: ACM, 2021: 2114-2124.

[3] 张国豪, 刘波. 采用 CNN 和 Bidirectional GRU 的时间序列分类研究[J]. 计算机科学与探索, 2019, 13(06): 916-927.
ZHANG G H, LIU B. Research on Time Series Classification Using CNN and Bidirectional GRU[J]. Journal of Frontiers of Computer Science and Technology, 2019, 13(06): 916-927.

[4] 张雅雯, 王志海, 刘海洋, 等. 基于多尺度残差 FCN 的时间序列分类算法[J]. 软件学报, 2022, 33(02): 555-570.
ZHANG Y W, WANG Z H, LIU H Y, et al. Time Series Classification Algorithm Based on Multiscale Residual Full Convolutional Neural Network. Ruan Jian Xue Bao/Journal of Software, 2022, 33(2): 555-570.

[5] Ruff L, Kauffmann J R, Vandermeulen R A, et al. A unifying review of deep and shallow anomaly detection[J]. Proceedings of the IEEE, 2021, 109(5): 756-795.

[6] Meng H, Zhang Y, Li Y, et al. Spacecraft anomaly detection via transformer reconstruction error[C]// International Conference on Aerospace System Science and Engineering 2019, Toronto, Jul 30-Aug 1, 2019. Singapore: Springer, 2020: 351-362.

[7] Chen Z, Chen D, Zhang X, et al. Learning graph structures with transformer for multivariate time series anomaly detection in iot[J]. IEEE Internet of Things Journal, 2022, 9(12): 9179-9189.

[8] Shchur O, Türkmen A C, Januschowski T, et al. Neural temporal point processes: A review[J]. arXiv preprint arXiv:2104.03528, 2021.

[9] Zhang Q, Lipani A, Kirnap O, et al. Self-attentive Hawkes process[C]//International conference on machine learning, virtual, Jul 13-18, 2020. PMLR , 2020: 11183-11193.

[10] Zuo S, Jiang H, Li Z, et al. Transformer hawkes process[C]//International conference on machine learning, virtual, Jul 13-18, 2020. PMLR, 2020: 11692-11702.

[11] Esling P, Agon C. Time-series data mining[J]. ACM Computing Surveys (CSUR), 2012, 45(1): 1-34.

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

[13] Torres J F, Hadjout D, Sebaa A, et al. Deep learning for time series forecasting: a survey[J]. Big Data, 2021, 9(1): 3-21.

[14] 洪申达, 尹宁, 邱镇 等. SPG-Suite:面向伪周期时间序列的预测方法[J]. 计算机科学与探索, 2014, 8(10): 1153-1161.
HONG S D, YIN N, QIU Z, et al. SPG-Suite: Forecasting Method Towards Pseudo Periodic Time Series[J]. Journal of Frontiers of Computer Science and Technology, 2014, 8(10): 1153-1161.

[15] Gao J, Sultan H, Hu J, et al. Denoising nonlinear time series by adaptive filtering and wavelet shrinkage: a comparison[J]. IEEE signal processing letters, 2009, 17(3): 237-240.

[16] Rojo-Álvarez J L, Martínez-Ramón M, de Prado-Cumplido M, et al. Support vector method for robust ARMA system identification[J]. IEEE transactions on signal processing, 2004, 52(1): 155-164.

[17] 赵洪科, 吴李康, 李徵, 等. 基于深度神经网络结构的互联网金融市场动态预测[J]. 计算机研究与发展, 2019, 56(08): 1621-1631.
ZHAO H K, WU L K, LI Z, et al. Predicting the Dynamics in Internet Finance Based on Deep Neural Network Structure[J]. Journal of Computer Research and Development, 2019, 56(08): 1621-1631.

[18] Hong T, Fan S. Probabilistic electric load forecasting: A tutorial review[J]. International Journal of Forecasting,

2016, 32(3): 914-938.

[19] Shao H, Soong B H. Traffic flow prediction with long short-term memory networks (LSTMs)[C]//2016 IEEE region 10 conference (TENCON). Singapore, Nov 22-25, 2016. IEEE, 2017: 2986-2989.

[20] 王永恒, 高慧, 陈炫伶. 采用变结构动态贝叶斯网络的交通流量预测[J]. 计算机科学与探索, 2017, 11(04): 528-538.
WANG Y H, GAO H, CHEN X L. Traffic Prediction Method Using Structure Varying Dynamic Bayesian Networks[J].Journal of Frontiers of Computer Science and Technology, 2017, 11(04): 528-538.

[21] 郑月彬, 朱国魂. 基于 Twitter 数据的时间序列模型在流行性感冒预测中的应用[J]. 中国预防医学杂志, 2019, 20(09): 793-798.
ZHENG, Y B, ZHU G H. Application of Twitter time series model in influenza prediction[J]. Chinese Preventive Medicine, 2019, 20(09): 793-798.

[22] 宋亚奇, 周国亮, 朱永利. 智能电网大数据处理技术现状与挑战[J]. 电网技术, 2013, 37(04): 927-935.
SONG Y Q, ZHOU G L, ZHU Y L. Present Status and Challenges of Big Data Processing in Smart Grid[J]. Power System Technology, 2013, 37(04): 927-935.

[23] Wen Q, He K, Sun L, et al. RobustPeriod: Robust time-frequency mining for multiple periodicity detection[C]//Proceedings of the 2021 International Conference on Management of Data. 2021: 2328-2337.

[24] 李盼盼, 宋韶旭, 王建民. 时间序列对称模式挖掘[J]. 软件学报, 2022, 33(03): 968-984.
LI P P, SONG S X, WANG J M. Time Series Symmetric Pattern Mining. Ruan Jian Xue Bao/Journal of Software, 2022, 33(3): 968−984.

[25] 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.

[26] 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.

[27] 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.

[28] De Myttenaere A, Golden B, Le Grand B, et al. Mean absolute percentage error for regression models[J]. Neurocomputing, 2016, 192: 38-48.

[29] Cameron A C, Windmeijer F A G. An R-squared measure of goodness of fit for some common nonlinear regression models[J]. Journal of econometrics, 1997, 77(2): 329-342.

[30] 万晨,李文中,丁望祥, 等.一种基于自演化预训练的多变量时间序列预测算法[J].计算机学报, 2022, 45(03): 513-525.
WAN C, LI W Z, DING W X, et al. A Multivariate Time Series Forecasting Algorithm Based on Self-Evolutionary Pre-training[J]. Chinese Journal of Computer, 2022, 45(03): 513-525.

[31] Goodfellow I, Bengio Y, Courville A, et al. Deep Learning (vol. 1) Cambridge[J]. 2016: 326-366.

[32] Gu J, Wang Z, Kuen J, et al. Recent advances in convolutional neural networks[J]. Pattern recognition, 2018, 77: 354-377.

[33] Li L, Ota K, Dong M. Everything is image: CNN-based short-term electrical load forecasting for smart grid[C]//2017 14th International Symposium on Pervasive Systems, Algorithms and Networks & 2017 11th International Conference on Frontier of Computer Science and Technology & 2017 Third International Symposium of Creative Computing (ISPAN-FCST-ISCC),Exeter, Jun 21-23, 2017. IEEE, 2017: 344-351.

[34] Borovykh A, Bohte S, Oosterlee C W. Conditional time series forecasting with convolutional neural networks[J]. arXiv preprint arXiv:1703.04691, 2017.

[35] Dong X, Qian L, Huang L. Short-term load forecasting in smart grid: A combined CNN and K-means clustering approach[C]//2017 IEEE international conference on big data and smart computing (BigComp), Jeju, Feb 13-16, 2017. IEEE,2017: 119-125.

[36] Bai S, Kolter J Z, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling[J]. arXiv preprint arXiv:1803.01271, 2018.

[37] Goodfellow I, Bengio Y, Courville A. Deep learning[M]. MIT press, 2016: 363-405.

[38] Schuster M, Paliwal K K. Bidirectional recurrent neural networks[J]. IEEE transactions on Signal Processing, 1997, 45(11): 2673-2681.

[39] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[40] Yu Y, Si X, Hu C, et al. A review of recurrent neural networks: LSTM cells and network architectures[J]. Neural computation, 2019, 31(7): 1235-1270.

[41] 王鑫, 吴际, 刘超, 等. 基于 LSTM 循环神经网络的故障时间序列预测[J]. 北京航空航天大学学报, 2018, 44(04): 772-784.
WANG X, WU J, LIU C, et al. Fault time series prediction based on LSTM recurrent neural network[J]. Journal of Beijing University of Aeronautics and Astronautics, 2018, 44(04): 772-784.

[42] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural networks, 2005, 18(5-6): 602-610.

[43] Siami-Namini S, Tavakoli N, Namin A S. The performance of LSTM and BiLSTM in forecasting time series[C]//2019 IEEE International Conference on Big Data (Big Data), Los Angeles, Dec 9-12, 2019. IEEE, 2019: 3285-3292.

[44] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint

arXiv:1406.1078, 2014.

[45] Fu R, Zhang Z, Li L. Using LSTM and GRU neural network methods for traffic flow prediction[C]//2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC), Wuhan, Nov 11-13, 2016. IEEE, 2016: 324-328.

[46] Qi Y, Li C, Deng H, et al. A deep neural framework for sales forecasting in e-commerce[C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Beijing, Nov 3-7, 2019. New York: ACM, 2019: 299-308.

[47] Xin S, Ester M, Bu J, et al. Multi-task based sales predictions for online promotions[C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management, Beijing, Nov 3-7, 2019. New York: ACM, 2019: 2823-2831.

[48] 王文冠, 沈建冰, 贾云得. 视觉注意力检测综述[J]. 软件学报, 2019, 30(02): 416-439.
WANG W G, SHEN J B, JIA Y D. Review of Visual Attention Detection. Ruan Jian Xue Bao/Journal of Software, 2019, 30(2): 416−439.

[49] Chorowski J K, Bahdanau D, Serdyuk D, et al. Attention-based models for speech recognition[J]. Advances in neural information processing systems, 2015, 28, 577-585.

[50] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30, 5998-6008.

[51] Li S, Jin X, Xuan Y, et al. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting[J]. Advances in neural information processing systems, 2019, 32, 5243-5253.

[52] Wen Q, Zhou T, Zhang C, et al. Transformers in time series: A survey[J]. arXiv preprint arXiv:2202.07125, 2022.

[53] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[54] Jin K H, Wi J A, Lee E J, et al. TrafficBERT: Pre-trained model with large-scale data for long-range traffic flow forecasting[J]. Expert Systems with Applications, 2021, 186: 115738.

[55] 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.

[56] Child R, Gray S, Radford A, et al. Generating long sequences with sparse transformers[J]. arXiv preprint arXiv:1904.10509, 2019.

[57] 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, virtual, Feb 2-9, 2021. California: AAAI, 2021, 35(12): 11106-11115.

[58] Lim B, Arık 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.

[59] Lin Y, Koprinska I, Rana M. SSDNet: State space decomposition neural network for time series forecasting[C]//2021 IEEE International Conference on Data Mining (ICDM), Auckland, Dec 7-10, 2021. IEEE, 2021: 370-378.

[60] 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.

[61] Rangapuram S S, Seeger M W, Gasthaus J, et al. Deep state space models for time series forecasting[J]. Advances in neural information processing systems, 2018, 31, 7785-7794.

[62] Oreshkin B N, Carpov D, Chapados N, et al. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting[J]. arXiv preprint arXiv:1905.10437, 2019.

[63] Vagropoulos S I, Chouliaras G I, Kardakos E G, et al. Comparison of SARIMAX, SARIMA, modified SARIMA and ANN-based models for short-term PV generation forecasting[C]//2016 IEEE International Energy Conference (ENERGYCON), Leuven, Apr 4-8, 2016. IEEE, 2016: 1-6.

[64] Gibran K, Bushrui S B. The prophet: A new annotated edition[M]. Simon and Schuster, 2012.

[65] 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.

[66] Qi X, Hou K, Liu T, et al. From known to unknown: Knowledge-guided transformer for time-series sales forecasting in Alibaba[J]. arXiv preprint arXiv:2109.08381, 2021.

[67] Zhou T, Ma Z, Wen Q, et al. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting[J]. arXiv preprint arXiv:2201.12740, 2022.

[68] Bracewell R N, Bracewell R N. The Fourier transform and its applications[M]. New York: McGraw-Hill, 1986.

[69] Zhang D. Wavelet transform[M]//Fundamentals of image data mining. Springer, Cham, 2019: 35-44.

[70] Wen Q, Gao J, Song X, et al. RobustSTL: A robust seasonal-trend decomposition algorithm for long time series[C]//Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, Jan 27-Feb 1, 2019. California: AAAI, 2019, 33(01): 5409-5416.

[71] Wang Z. Fast algorithms for the discrete W transform and for the discrete Fourier transform[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1984, 32(4): 803-816.

[72] Shensa M J. The discrete wavelet transform: wedding the a trous and Mallat algorithms[J]. IEEE Transactions on signal processing, 1992, 40(10): 2464-2482.

[73] 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, virtual, Apr 25-29, 2022. ICLR, 2021, 1-20.
[74] Li Y, Lu X, Xiong H, et al. Towards Long-Term Time-Series Forecasting: Feature, Pattern, and Distribution[J]. arXiv preprint arXiv:2301.02068, 2023.

梁宏涛（1979—），男，山东济宁人，博士，副教授，CCF 高级会员，主要研究方向为数据挖掘、能源互联网等。

LIANG Hongtao, born in 1979, Ph.D., associate professor, senior member of CCF. His research interests include data mining, Internet of energy, etc.

刘硕（1998—），男，山东青岛人，硕士研究生，CCF 学生会员，主要研究方向为数据挖掘、能源互联网等。

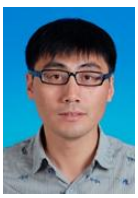
LIU Shuo, born in 1998, M.S.candidate, student member of CCF. His research interests include data mining,Internet of energy, etc.

杜军威（1974—），男，山东威海人，博士，教授， CCF 专业会员，主要研究方向为数据挖掘、知识图谱等。

DU Junwei, born in 1974, Ph.D., professor, professional member of CCF. His research interests include data mining, knowledge graph, etc.

胡强（1980—），男，山东邹城人，博士，副教授，CCF 专业会员，主要研究方向为数据挖掘、软件形式化验证等。

HU Qiang, born in 1980, Ph.D., associate professor, professional member of CCF. His research interests include data mining, software formal verification, etc.



于旭（1982—），男， 山东青岛人，博士，副教授，CCF 高级会员，主要研究方向为推荐系统、迁移学习等。

YU Xu, born in 1982, Ph.D., associate professor, senior member of CCF. His research interests include recomendation system, transfer learning, etc.