DOI:10.16356/j.1005-2615.2025.05.002

基于深度学习的时间序列预测方法综述

潘志松[1],韩 笑[1],黎 维[2]
(1.陆军工程大学指挥控制工程学院,南京 210007;2.陆军装甲兵学院信息通信系,北京 100072)

摘要: 深度学习因能够更好地捕捉时间序列数据中的复杂关系和模式而成为解决时间序列预测的有效方法。典型的做法是单独地学习这些任务,为每个任务训练1个单独的神经网络,在时间序列预测中取得了丰硕的成果。最近的多任务学习技术通过学习共享知识联合处理多个预测任务,在性能、计算和内存占用方面显示出了其优势。本文首先综述了以卷积神经网络、循环神经网络、注意力机制、Transformer和图神经网络为代表的时间序列预测深度模型,包括数据集、模型特点和性能;然后深入分析了深度多任务时间序列预测模型,按照参数共享方式和参数共享(交互)位置进行分类概述,并讨论了一些常见的多任务时间序列预测框架。最后对深度时间序列预测面临的问题和挑战进行了总结,并对未来研究趋势进行了展望。

关键词: 深度学习;时间序列预测;多任务时间序列预测;参数共享;参数交互
中图分类号: TP391 **文献标志码:** A **文章编号:** 1005-2615(2025)05-0799-23

Review of Time Series Forecasting Methods Based on Deep Learning

PAN Zhisong[1], *HAN Xiao*[1], *LI Wei*[2]
(1. College of Command and Control Engineering, Army Engineering University of PLA, Nanjing 210007, China;
2. Department of Information Communication, Army Academy of Armored Forces, Beijing 100072, China)

Abstract: Deep learning has emerged as an effective solution for time series forecasting due to its superior ability to capture complex relationships and patterns within temporal data. A typical approach involves learning these tasks individually and training a separate neural network for each task, which has yielded fruitful results in time series forecasting. Recent advances in multitask learning techniques have demonstrated their advantages in terms of performance, computation, and memory usage by jointly processing multiple prediction tasks through learning shared knowledge. This paper presents the first comprehensive review of methods for multitask time series forecasting. It begins by summarizing deep models for time series forecasting, represented by convolutional neural networks, recurrent neural networks, attention mechanisms, Transformer and graph neural networks, including datasets, model characteristics, and performance. Subsequently, an in-depth analysis of deep multitask time series forecasting models is conducted, categorizing them based on parameter sharing methods and the location of parameter sharing (or interaction), and discussing some common multitask time series forecasting frameworks. Finally, this paper summarizes the challenges faced by deep time series forecasting and offers insights into future research trends.

Key words: deep learning; time series forecasting; multi-task time series forecasting; parameter sharing; parameter interaction

基金项目: 国家自然科学基金(62076251);陆军工程大学基础学科科研基金科研课题培育项目(KYJBJKQTZK23003)。
作者简介: 潘志松,男,教授,主要从事模式识别与机器学习方面的研究,E-mail: hotpzs@hotmail.com。
通信作者: 韩笑,女,讲师,E-mail:h.x.good@163.com。
收稿日期: 2024-03-18;**修订日期:** 2024-07-01

引用格式: 潘志松,韩笑,黎维. 基于深度学习的时间序列预测方法综述[J]. 南京航空航天大学学报(自然科学版),2025,57(5):799-821. PAN Zhisong, HAN Xiao, LI Wei. Review of time series forecasting methods based on deep learning[J]. Journal of Nanjing University of Aeronautics & Astronautics(Natural Science Edition),2025, 57(5):799-821.

时间序列通常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。时间序列预测一直是学术研究的一个关键领域,其在交通预测[1-3]、气象预测[4-7]、医疗保健[8-10]以及金融行业[11-14]等诸多领域有着广泛的应用。时间序列数据因具有多元、顺序等特性,导致其潜在特征之间在空间和时间上存在错综复杂的动态相关性,这使得时间序列预测成为一项具有挑战性的问题。

现代机器学习方法提供了一种以纯粹数据驱动的方式学习时间动态的手段[15],已经成为下一代时间序列预测模型中至关重要的一部分。在计算机视觉[16]、自然语言处理[17]和强化学习[18]等领域取得显著成果的启发下,深度学习在最近几年得到了广泛的关注。因为深度神经网络能够通过使用一系列非线性层来构建中间特征表示而成为时间序列预测的有效方法,所以人们提出了更复杂的深度学习架构[19-20],主要包括单任务学习(Single-task learning, STL)模型和多任务学习[21](Multi-task learning, MTL)模型。面对不同领域时间序列预测问题的多样性和复杂性时,不同的深度学习模型呈现了特有的优势,所以有必要对使用深度神经网络的时间序列预测方法进行梳理和总结。当前比较认可的一些综述文章[20,22-23]对时间序列预测深度学习架构阐述详尽,但是仅聚焦于单任务的时间序列预测,并且缺乏具体的实验对比和更多实际数据集上的验证结果。本文不但对比了单任务时间序列预测模型在不同数据集上的性能表现,而且还专门对多任务时间序列预测模型的特点以及其在不同数据集上的性能进行了综述。

1 基本概念

1.1 时间序列预测

时间序列预测模型可以是单变量(一个时变变量),也可以是多变量(多个时变变量)。尽管单变量和多元系统之间的模型可能有很大的不同,但大多数深度学习模型都可以模糊地处理它们[23]。

对于大量现有的研究,假设观测值在等距时间间隔中可用。在这种情况下,时间序列预测问题可以表述为

$$\hat{\boldsymbol{x}}_{t+h}=f(\boldsymbol{x}_t,\boldsymbol{x}_{t-1},\cdots,\boldsymbol{x}_{t-N+1}) \tag{1}$$

式中:$\boldsymbol{x}_t,\boldsymbol{x}_{t-1},\cdots,\boldsymbol{x}_{t-N+1}$为时间序列数据点;$\hat{\boldsymbol{x}}_{t+h}$为预测结果;$N$为输入数,在一些研究中也称为嵌入维数。时间步$h$可以是1,或其他任何正整数,被称为多步超前预测。

时间序列可以是多元的,即$\boldsymbol{x}_t$表示多个序列在t时刻的值,表达式为

$$\boldsymbol{x}_t=(x_{1t},x_{2t},\cdots,x_{mt})^{\mathrm{T}} \tag{2}$$

式中m为序列的个数。

1.2 多任务时间序列预测

假设共有K个任务,多任务时间序列预测问题定义为

$$\begin{bmatrix}\hat{\boldsymbol{x}}_{t+h}^{(1)}\\ \hat{\boldsymbol{x}}_{t+h}^{(2)}\\ \vdots\\ \hat{\boldsymbol{x}}_{t+h}^{(K)}\end{bmatrix}=f\begin{pmatrix}\{\boldsymbol{x}_{t-i}^{(1)}\}_{i=0}^{N-1}\\ \{\boldsymbol{x}_{t-i}^{(2)}\}_{i=0}^{N-1}\\ \vdots\\ \{\boldsymbol{x}_{t-i}^{(K)}\}_{i=0}^{N-1}\end{pmatrix} \tag{3}$$

式中:$\{\boldsymbol{x}_{t-i}^{(k)}\}_{i=0}^{N-1}=\{\boldsymbol{x}_{t-i}^{(k)}|\,i=0,1,\cdots,N-1\}$为任务$k$的时间序列。

1.3 时间序列预测常用评价指标

时间序列预测评价指标可以分为回归和分类两大类,本文列举了一些常见的预测评价指标。

(1) 平均绝对误差(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

$$\mathrm{MAE}=\frac{1}{N}\sum_{i=1}^{N}|y_i-\hat{y}_i| \tag{4}$$

式中:y_i为真实值,$\hat{y}_i$为模型预测(下同)。MAE的取值范围为$[0,+\infty)$,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。

当不同数据集或不同量纲的预测变量间进行比较时,可使用归一化平均绝对误差(Normalized mean absolute error, NMAE),计算公式为

$$\mathrm{NMAE}=\frac{\mathrm{MAE}}{\sum_{i=1}^{N}|y_i|} \tag{5}$$

(2)平均绝对缩放误差(Mean absolute scaled error, MASE),通过对MAE进行缩放,以使得提出模型误差与基线模型进行比较。定义如下

$$\mathrm{MASE}=\frac{\sum_{i=1}^{N}|y_i-\hat{y}_i|}{\frac{N}{m}\sum_{i=m+1}^{N}|y_i-y_{i-m}|} \tag{6}$$

式中m为季节周期长度(若数据无周期性,则$m=1$)。若MASE<1,则提出模型性能优于基线预测;若MASE>1,则提出模型性能较差。MASE适用于不同时间尺度的数据,不受数量级影响。

(3) 均方误差(Mean square error, MSE),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

$$\mathrm{MSE}=\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2 \tag{7}$$

MSE的取值范围同样是[0，+∞)，计算速度更快，其一直作为时序预测算法的主要评价指标之一。

(4) 均方根误差(Root mean square error，RMSE)，是均方误差进行开方得到，具体公式为

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2} \tag{8}$$

该评价指标在有足够数据可用的情况下常被选用，但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。

归一化均方根误差(Normalized root mean square error，NRMSE)对RMSE进行归一化处理，使其变成无量纲的标量，从而便于在不同尺度或单位的数据集之间比较模型性能。计算公式如下

$$\mathrm{NRMSE}=\frac{\mathrm{RMSE}}{\bar{y}} \tag{9}$$

式中$\bar{y}$为实际值的平均值。

(5) 相对平方误差(Relative squared error，RSE)用于比较一个预测模型的误差与一个简单基准模型(通常是目标变量的平均值)的误差之间的比例。定义如下

$$\mathrm{RSE}=\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2} \tag{10}$$

相对均方根误差(Root relative squared error，RRSE)为

$$\mathrm{RRSE}=\sqrt{\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2}} \tag{11}$$

RRSE是RSE的平方根。它也将模型的误差与均值模型的误差进行比较，但其结果在尺度上更接近于原始数据。

(6) 平均绝对百分比误差(Mean absolute percentage error，MAPE)，衡量的是预测值与实际值之间的绝对百分比误差的平均值，避免了正误差和负误差相互抵消，具体公式为

$$\mathrm{MAPE}=\frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i-\hat{y}_i}{y_i}\right|\times 100\% \tag{12}$$

当MAPE值很小时，表示预测值与实际值的相对偏差较小，也就意味着预测模型的准确性更高。

(7) 加权平均绝对百分比误差(Weighted mean absolute percentage error，WMAPE)是对MAPE的一种改进，计算公式为

$$\mathrm{WMAPE}=\frac{\sum_{i=1}^{N}|y_i-\hat{y}_i|}{\sum_{i=1}^{N}|y_i|}\times 100\% \tag{13}$$

WMAPE不是计算每个点的百分比误差再平均，而是先计算总绝对误差，再将其与总实际值进行比较，避免了MAPE对低实际值敏感的问题。

(8) 对称平均绝对百分比误差(Symmetric mean absolute percentage error，SMAPE)，计算预测值和实际值之间的相对误差，并进行对称归一化，定义如下

$$\mathrm{SMAPE}=\frac{1}{N}\sum_{i=1}^{N}\frac{|y_i-\hat{y}_i|}{\left(|y_i|+|\hat{y}_i|\right)/2}\times 100\% \tag{14}$$

SMAPE适用于具有不同量纲的时间序列数据，归一化的分母可以防止极端情况下误差过大的问题，如实际值接近0。SMAPE的理论范围是0%～200%。0%表示完美模型，预测完全准确；200%是最差情况，通常发生在其中一个值为0，而另一个值不为0时。

(9) 平均平方百分比误差(Mean squared percentage error，MSPE)是计算每个数据点百分比误差的平方的平均值。计算公式如下

$$\mathrm{MSPE}=\frac{1}{N}\sum_{i=1}^{N}\left[\frac{(y_i-\hat{y}_i)}{y_i}\right]^2\times 100\% \tag{15}$$

均方根百分比误差(Root mean squared percentage error，RMSPE)是MSPE的平方根

$$\mathrm{RMSPE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\frac{(y_i-\hat{y}_i)}{y_i}\right]^2\times 100\%} \tag{16}$$

RMSPE可将误差尺度拉到与原始百分比误差更接近的水平，使其更易于解释。

(10) 经验相关系数(Empirical correlation coefficient，CORR))表示两个变量间的线性相关程度，取值介于－1～1，小于0负相关，大于0正相关，绝对值越大相关性越强，计算公式为

$$\mathrm{CORR}=\frac{\sum_{i=1}^{N}(y_i-\bar{y}_i)(\hat{y}_i-\bar{\hat{y}}_i)}{\sqrt{\sum_{i=1}^{N}(y_i-\bar{y}_i)^2(\hat{y}_i-\bar{\hat{y}}_i)^2}} \tag{17}$$

式中$\bar{\hat{y}}$表示预测值的平均值。

(11) 决定系数(Coefficient of determination，记为R^2)也称为拟合优度，反应了因变量的波动有多少百分比能被自变量的波动解释，在0～1之间取值，越大越好。具体公式为

$$R^2=1-\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{N}(y_i-\bar{y}_i)^2}=1-\frac{\mathrm{MSE}}{\sum_{i=1}^{N}(y_i-\bar{y}_i)^2} \tag{18}$$

(12) 分位数损失(Quantile loss),分位数回归损失函数用于预测分位数。分位数表示组中有多少值低于或高于特定阈值的值。它计算跨预测变量(独立)变量值的响应(因)变量的条件中位数或分位数。除了第50个百分位数是MAE,损失函数是MAE的扩展。它不对响应的参数分布作任何假设,甚至为具有非常量方差的残差提供预测区间。具体公式为

$$D_\rho(y_i,\hat{y}_i)=\left(\rho-I_{\{y_i\leqslant\hat{y}_i\}}\right)(y_i-\hat{y}_i) \tag{19}$$

式中:$\rho\in(0,1)$,I为指示函数。

为了总结给定跨度下所有项目的分位数损失,往往考虑分位数损失的归一化总和被称为ρ-risk,计算方法为

$$R_\rho(y_i,\hat{y}_i)=\frac{2\sum_{i=1}^{N}\sum_{t=1}^{T}D_\rho(y_t^{(i)},\hat{y}_t^{(i)})}{\sum_{i=1}^{N}\sum_{t=1}^{T}|y_t^{(i)}|} \tag{20}$$

式中T为数据时间步长。

(13) 准确率(Accuracy):反映分类器对整个样本的判定能力,能将正的判定为正,负的判定为负。

$$\mathrm{Accuracy}=\frac{1}{N}\sum_{i=1}^{N}[\hat{y}_i=y_i]=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{21}$$

式中:TP(True positive)表示正确预测正类的样本(真阳性),FN(False negative)表示错误预测为负类的样本(假阴性),FP(False positive)表示错误的预测为正类(假阳性),TN(True negative)表示正确预测为负类(真阴性)。

(14) 精确率(Precision):在所有被模型预测为正例的样本中,真正的正例所占的比例。计算公式为

$$\mathrm{Precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{22}$$

(15) 召回率(Recall):在所有真实的正例样本中,被模型成功预测为正例样本所占的比例。计算公式为

$$\mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{23}$$

(16) 阳性预测值(Positive predictive value,PPV)。公式为

$$\mathrm{PPV}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{24}$$

(17) 敏感性(Sensitivity):真阴性中被正确预测的比例。公式为

$$\mathrm{Sensitivity}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{25}$$

(18) F_1-score是精确率和召回率的调和平均数。它与单一的分类阈值紧密相关,公式为

$$F_1=2\times\frac{\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \tag{26}$$

(19) 平均精度-召回率曲线下的面积(Area under the precision-recall curve, AUPRC),是用于评价二分类模型性能的指标。AUPRC的范围为0~1,数值越高,表示模型的性能越好。

(20) 受试者工作特征曲线(Receiver operating characteristic curve,ROC),是一种用于展示二分类模型在不同分类阈值下性能的图形化工具。它描绘了模型的敏感性和特异性之间的权衡关系。受试者工作特征曲线下面积(Area under the receiver operating characteristic curve,AUROC)是ROC曲线下的面积。AUROC=1.0表示完美模型,能完全区分正负类;AUROC=0.5表示模型没有区分能力,相当于随机猜测;AUROC<0.5表示模型比随机猜测还差。

本文不仅对各深度预测方法进行了模型的特点分析,而且还通过具体的评价指标对模型的性能进行了定量分析。

2 基于深度学习的时间序列预测

本节总结了以卷积神经网络(Convolutional neural network, CNN)[24]、循环神经网络(Recurrent neural network, RNN)[25]、注意力机制(Attention mechanism)[26]、Transformer[27]和图神经网络(Graph neural network, GNN)[28]为代表的常见时间序列预测深度模型。

2.1 卷积神经网络

2.1.1 卷积神经网络

CNN是一组深度架构,最初为计算机视觉任务而设计[29]。CNN学习使用卷积操作从原始数据中提取有意义的特征。卷积操作是一个创建特征映射的滑动过滤器,旨在捕获不同区域数据的重复模式。这种特征提取过程使得CNN有一个突出特点,即失真不变性,这意味卷积操作可以提取数据中任意位置的特征。上述特点使得CNN适合处理一维数据,如时间序列。基于CNN的模型在时间序列预测文献中并没有被广泛使用,而一些研究提出改进的CNN单独或与RNN一起作为特征提取器以进行预测。表1列出了CNN时间序列预测的相关研究。通过表1可以看出,两个基于CNN改

表 1　CNN 时间序列预测的相关研究

Table 1　Related research on CNN time series forecasting

模型	优于	数据集	模型特点	性能指标		
				最优基线	本文模型	性能提升
CNN＋K-means[30]	CNN	Electricity load	CNN 与 K-means 结合，用 K-means 聚类将数据集划分为训练子集和测试子集，用所有这些子集对 CNN 进行训练以构建预测模型。	MAPE/RMSE/NMAE 分别为 3.953/0.250/0.027	MAPE/RMSE/NMAE 分别为 3.055/0.219/0.024	MAPE/RMSE/NMAE 分别提升为 0.227/0.142/0.125
WaveNet-CNN[31]	LSTM	Financial data	基于 WaveNet 用堆叠的扩张卷积，广泛访问历史范围，并行应用多个卷积滤波器对分离的时间序列执行 ReLU 激活和条件处理，以实现对数据的快速处理和利用多变量时间序列之间的相关结构。	MASE(A/B/C) 分别为 0.82/0.925/0.950 (A，B，C 表示不同的测试周期，下同)	MASE(A/B/C) 分别为 0.693/0.690/0.702	MASE(A/B/C) 分别提升为 0.164/0.254/0.261
CNN-LSTM[32]	LSTM	Electricity load	LSTM 提取长期依赖关系，CNN 捕捉局部趋势模式，但模型只能捕获局部特征。	RMSE 为 1 246.392	RMSE 为 1 134.179	RMSE 提升 0.090

进的时间序列预测模型的性能优于长短期记忆网络(Long short-term memory network，LSTM)[33]。尽管 LSTM 已被广泛用于时间序列分析，但在一些特殊的场景下，不同结构的 CNN 模型也被用作预测工具来提高预测性能。如果时间序列数据中存在明显的局部模式或周期性特征，CNN 能够很好地捕捉这些局部特征。当数据中存在噪声时，CNN 的卷积操作可以起到一定的平滑作用，通过多个卷积层和池化层，能够过滤掉一些高频噪声。LSTM 对噪声比较敏感，因为它在每个时间步都要处理输入信息并更新隐藏状态。过多的噪声可能会干扰 LSTM 对长期依赖关系的学习，并且在处理非平稳数据时 LSTM 可能会过度关注序列中的异常点，对于局部特征的提取也没有 CNN 那么高效，导致预测性能下降。但是 CNN 缺少对序列数据的记忆功能，不能充分挖掘潜在特征之间的依赖关系。

2.1.2　时间卷积神经网络

Bai 等[34]提出 CNN 的一种变体时间卷积神经网络(Temporal convolutional neural network，TCN)，更适应时间序列数据集，在执行时间和内存需求方面直接与 RNN 竞争。TCN 使用膨胀卷积，网络的输出与输入序列具有相同的长度，能够捕获时间序列中长期依赖关系。并且 TCN 通过使用因果卷积来防止信息从未来泄露到过去，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。膨胀因果卷积 TCN 模块结构如图 1 所示。

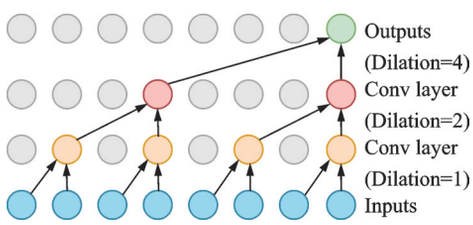

图 1　膨胀因果卷积模块

Fig.1　Dilated causal convolution module

膨胀卷积是一个函数，记为 $F_d(x)$，定义如下

$$F_d(x)=\sum_{i=0}^{K-1} f(i)\,x_{t-di} \tag{27}$$

式中：$f(i)$ 为滤波器，K 为滤波器大小，d 为膨胀因子参数，$t-di$ 为过去的方向。

为了进一步增加网络感受野串联多个 TCN 块，在每个 TCN 块的输出中添加残差连接，可以在很深的体系结构中提高性能。TCN 模型定义如下

$$a_t^l=g(W_a^l F_d a_t^{l-1}+b_a^l+a_t^{l-1}) \tag{28}$$

式中：$F_d(\cdot)$ 为 d 因子的膨胀卷积，a_t^l 为第 l 层神经元在时刻 t 的值，W_a^l 和 b_a^l 为第 l 层对应的权重和偏置，g 为激活函数。

这些特点使得 TCN 成为适合处理复杂时间序列问题的深度学习架构。表 2 展示了 TCN 成功应用于时间序列预测的研究成果。在一些特定的领域中，上述基于 TCN 建立的预测模型比普通的 TCN 性能要好，甚至在一些数据集上捕获数据中的时间依赖性比基于 RNN 类的模型更有效，并且利用 TCN 模型的并行性可以加快训练时间。

表2 TCN时间序列预测的相关研究

Table 2 Related research on TCN time series forecasting

模型	优于	数据集	模型特点	性能指标		
				最优基线	本文模型	性能提升
SCINet[35]	TPA-LSTM[36]	ETT/Solar/PeMS	一种递归下采样-卷积-交互架构。在每层中,用多个卷积滤波器从下采样的子序列或特征中提取不同但有价值的时间特征,从多个分辨率聚合的特征中,建模复杂的时间动态性。但SCINet仅关注先前序列的相关性信息或普适性特征,没有考虑先前序列与未来序列的联系。	RSE/CORR分别为0.183/0.934 7	RSE/CORR分别为0.170 9/0.940 0	RSE/CORR分别提升0.065/0.017
M-TCN[37]	LSTM	Beijing PM2.5	将膨胀网络作为元网络,增加感受野;构建非对称残差块,学习更多有用信息并降低计算成本。但在全连接层,每个特征图被平均,损失了整个特征值的局部信息,同时大大增加参数数量。	RMSE/RRSE/CORR分别为68.07/0.73/0.69	RMSE/RRSE/CORR分别为65.35/0.70/0.72	RMSE/RRSE/CORR分别提升0.04/0.04/0.042
Deep-GLO[38]	STGCN[39]	PeMSD	提出由TCN正则化的矩阵分解模型(Temporal convolution network regularized matrix factorization, TCN-MF),可以在预测期间处理全局依赖关系。将来自TCN-MF模型的预测作为时间卷积网络的协变量,从而使最终模型能够同时关注每个时间序列的局部属性以及全局属性。	MAE/MAPE/RMSE分别为3.57/0.087/6.77	MAE/MAPE/RMSE分别为3.53/0.079/6.49	MAE/MAPE/RMSE分别提升0.011/0.09/0.041

2.2 循环神经网络

RNN专门用于处理序列数据,例如机器翻译相关问题中的单词序列、语音识别中的音频数据或预测问题中的时间序列。RNN将每个时间步与前一个时间步连接起来,以建模数据的时间依赖性。RNN细胞包含1个内部记忆状态,充当过去信息的小结,内存状态递归地在每个时间步用新的观测值更新。RNN模型结构如图2所示。

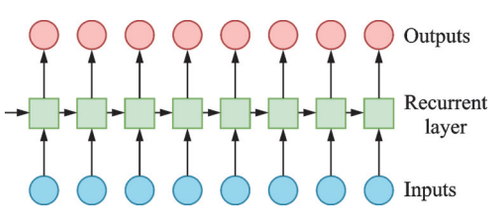

图2 RNN模型

Fig.2 RNN model

传统的RNN在学习数据中的长期依赖关系时可能会遭遇梯度爆炸或者消失梯度[40]的问题,使得RNN只具备短期记忆,从而丢失部分信息。为解决该类问题,出现了RNN的变体结构,如长短期记忆网络LSTM[33]和门控循环单元(Gated recurrent unit, GRU)[41]。

2.2.1 长短期记忆网络

LSTM[33]通过改善网络中的梯度流来解决上述梯度弥散的限制。在LSTM中用隐藏单元存储短期记忆,用细胞状态来存储长期信息,通过一系列门进行调制,即

$$\text{遗忘门:}f_t=\sigma(W_{\mathrm{f}}\cdot[h_{t-1},x_t]+b_{\mathrm{f}}) \tag{29}$$

$$\text{输入门:}i_t=\sigma(W_{\mathrm{i}}\cdot[h_{t-1},x_t]+b_{\mathrm{i}}) \tag{30}$$

$$\text{输出门:}o_t=\sigma(W_{\mathrm{o}}\cdot[h_{t-1},x_t]+b_{\mathrm{o}}) \tag{31}$$

式中:x_t和h_{t-1}分别为当前输入和前一时刻的隐藏状态,$\sigma(\cdot)$为sigmoid激活函数,W_{f}、W_{i}、W_{o}以及b_{f}、b_{i}、b_{o}分别是控制遗忘门f_t、输入门i_t和输出门o_t行为的权重和偏置。

门机制对LSTM的隐藏状态和细胞状态进行如下修正

$$\text{隐藏状态:}h_t=o_t*\tanh(C_t) \tag{32}$$

式中:tanh为激活函数;C_t为细胞状态,表达式为

$$C_t=f_t*C_{t-1}+i_t*\tilde{C}_t \tag{33}$$

$\tilde{C}_t$为候选细胞状态,表达式为

$$\tilde{C}_t=\tanh(W_c\cdot[h_{t-1},x_t]+b_c) \tag{34}$$

式中：W_c和b_c分别为候选细胞状态$\tilde{C}_t$的权重和偏置，tanh为激活函数。

LSTM循环单元结构图如图3所示。LSTM具有通过反馈连接学习长期依赖关系的能力，也是时间序列预测方法中使用最多的一种深度学习技术。表3中列出了可以解决LSTM网络时间序列预测问题的相关研究。这些模型证明了LSTM在从时间序列中提取有意义信息方面优于传统的MLP和RNN。此外，最近的一些研究还提出了更具创新性的改进方案，如混合模型、遗传算法等来优化网络架构。

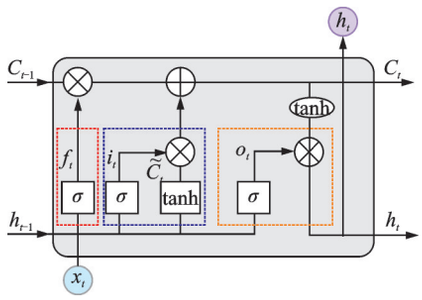

图3　LSTM循环单元结构

Fig.3　Structure of LSTM recurrent unit

表3　LSTM时间序列预测的相关研究

Table 3　Related research on LSTM time series forecasting

模型	优于	数据集	模型特点	性能指标		
				最优基线	本文模型	性能提升
CNN-LSTM[13]	LSTM	Financial data	CNN层从处理后的时间序列数据中提取主要特征。LSTM层计算最终的预测结果，但没有将定量和定性因素作为预测模型的输入。	RMSE为0.022 4	RMSE为0.017 3	RMSE提升0.224 8
LSTM-PSO[42]	LSTM	Oilfield	通过LSTM捕获时间序列数据的依赖关系，采用粒子群优化算法优化LSTM模型的基本结构，但忽略了数据的时空依赖性。	MAPE/MAE/RMSE分别为14.15/2.31/2.73	MAPE/MAE/RMSE分别为9.88/1.6/2.02	MAPE/MAE/RMSE分别提升0.31/0.31/0.26
DLSTM（Deep long-short term memory）[43]	ARIMA	Oilfield	有多个LSTM层，每层包含多个细胞，并使用遗传算法（Genetic algorithm，GA）来配置最佳的DLSTM架构和参数，但仅考虑了单变量时间序列数据的预测。	RMSE/RMSPE分别为0.027/3.731	RMSE/RMSPE分别为0.025/3.496	RMSE/RMSPE分别提升0.074 1/0.063
Multi-head ATT-LSTM[44]	LSTM	16 public datasets	提出了一种LSTM和多头注意力两种深度学习方法的混合模型。	SMAPE为3.25	SMAPE为3	SMAPE提升0.07

2.2.2　门控循环单元

GRU[41]是RNN梯度问题的另一种解决方案。然而，它也可以被视为LSTM单元的简化。GRU单元使用1个隐藏状态，将隐藏状态和细胞状态合并为1个状态，使用1个更新门和1个重置门，将遗忘门和输入门合并为单个更新门。它需要更少的训练时间，且具有更好的网络性能。GRU循环单元结构如图4所示。

GRU中隐藏单元的激活可按如下步骤进行

$$\text{重置门：}r_t=\sigma(W_r\cdot[h_{t-1},x_t]+b_r) \quad (35)$$

$$\text{更新门：}z_t=\sigma(W_z\cdot[h_{t-1},x_t]+b_z) \quad (36)$$

式中：W_r、W_z以及b_r、b_z分别为控制重置门r_t和更新门z_t的权重和偏置。

$$\text{隐藏状态：}h_t=(1-z_t)*h_{t-1}+z_t*\tilde{h}_t \quad (37)$$

其中$\tilde{h}_t$为候选隐藏状态，有

$$\tilde{h}_t=\tanh(W_h\cdot[r_t*h_{t-1},x_t]+b_h) \quad (38)$$

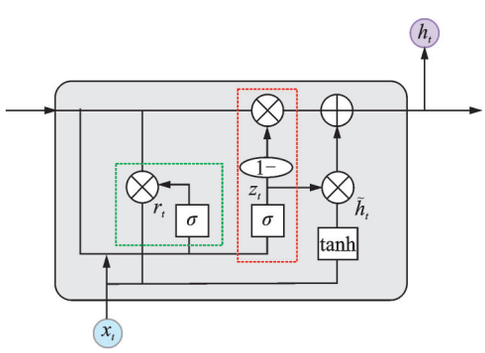

图4　GRU循环单元结构

Fig.4　Structure of GRU recurrent unit

式中W_h和b_h分别为候选隐藏状态的权重偏置。

表4列出了一些用GRU网络解决时间序列预测问题的研究。近期的研究工作通常对标准GRU模型进行修改，以提高相对于其他循环网络的性能。使用GRU模型的数量相对低于使用LSTM网络的数量。

表4 GRU时间序列预测的相关研究

Table 4 Related research on GRU time series forecasting

模型	优于	数据集	模型特点	性能指标		
				最优基线	本文模型	性能提升
DSF[45]	XGBoost[46]	Snack/PG&U	一种序列到序列的预测框架,以循环方式估计预测值。在解码器之上引入了1个残差网络,以统一的方式融合不同的异构特征及其相互作用。	MAE/WMAPE分别为244.46/0.485 9	MAE/WMAPE分别为228.93/0.455 1	MAE/WMAPE分别提升0.064/0.063
XGB-GRU[47]	XGBoost[46]	Heating data	利用XGBoost提取数据中多个控制变量的隐含信息,再利用GRU提取数据中的时序信息,XGB-GRU优于单一的XGBoost和GRU模型,但模型复杂度较高。	RMSE/MAE/R^2分别为10.353/8.217/0.848	RMSE/MAE/R^2分别为8.461/7.177/0.898	RMSE/MAE/R^2分别提升0.183/0.127/0.06
FM-GRU[48]	FC-LSTM[49]	Water quality	用GRU作为编码器和解码器,引入因子分解机(Factorization machine, FM),以解决数据高稀疏性和高维特征交互问题。用双重注意力机制解决数据长周期和时间跨度问题。但需进一步提高FM的高维潜在信息捕获能力。	MAE/MSE/RMSE/NRMSE分别为1.2/2.26/1.50/0.39	MAE/MSE/RMSE/NRMSE分别为0.57/0.64/0.77/0.16	MAE/MSE/RMSE/NRMSE分别提升0.53/0.72/0.49/0.59

2.3 注意力机制

在时间序列预测中,深度学习模型的一个研究方向是捕捉动态时间序列数据中的时间模式[50],基于RNN的预测模型开始流行起来,如DeepAR[51]。尽管LSTM在一定程度上克服了梯度消失或爆炸的问题,但基于RNN的模型仍然不能很好地对长期依赖关系进行建模[33]。注意力机制作为当前深度学习领域的重要组件之一,已经成为解决序列到序列问题最成功的方法。其在众多的输入信息中聚焦于对当前任务更为关键的信息,降低了对其他信息的关注度[52]。注意力结构如图5所示。

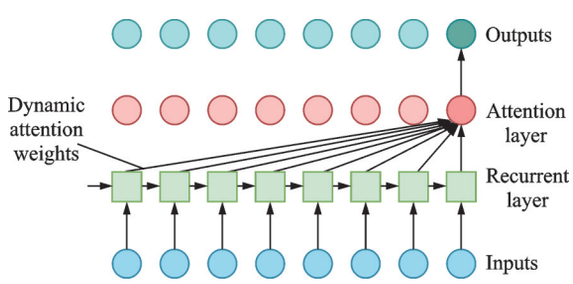

图5 注意力结构图

Fig.5 Attention structure diagram

Bahdanau等[53]提出的注意力属于软注意力(Soft-attention),是常见的一种注意机制,它使用所有key的加权平均值来构建上下文向量,而Xu等[54]提出的硬注意力通过随机取样的key来计算上下文向量。Soft-attention的注意力模块相对于输入是可微分的,因此整个系统仍然可以通过标准的反向传播方法进行训练[55]。Soft-attention通过学习得到1个权重向量,用于对输入序列中的每个元素进行加权求和。这个权重向量可以根据当前任务的需求,自动地调整不同元素的重要性。自注意力机制(Self-attention)作为Soft-attention的一种变体,是一种将单个序列的不同位置关联起来以计算同一序列的表示的注意力机制。当在序列内部引入Self-attention后,可以将序列中任意两个位置直接联系起来,就更容易捕获特征之间的依赖关系,不再受中间距离的限制。为了能够更好地了解Soft-attention在时间序列预测中的应用,将相关的模型列在表5中。表5中的工作证明了在时间序列预测应用中使用Soft-attention的优势,在可比较的循环网络上具有更好的性能。

表 5　注意力机制时间序列预测的相关研究

Table 5　Related research on attention time series forecasting

注意力机制	模型	优于	数据集	模型特点	性能指标		
					最优基线	本文模型	性能提升
软注意力	AASTH-GCN[56]	ADGAT[57]	ACL18	采用噪声感知的时空注意力机制过滤无效关联，并结合时间注意力捕捉股票序列的时空模式；通过自适应节点嵌入将股票内在关联映射到可训练的稠密矩阵；将图卷积操作从静态图扩展到自适应超图以探索动态关联；使用时间感知的级联卷积提取细粒度时间特征。	Accuracy/F_1 分别为 0.520/0.601	Accuracy/F_1 分别为 0.534/0.558	Accuracy/F_1 分别提升 0.026/0.077
	MARNN[58]	DARNN[59]	Air quality	一种基于多注意力机制的循环神经网络，分别利用输入注意力和自注意力获得相关的编码器隐藏状态和关联属性的编码器隐藏状态。再使用时序-卷积注意力神经网络来处理编码器隐藏状态并捕获长程时序模式。	MAE/RMSE 分别为 0.054/0.071	MAE/RMSE 分别为 0.023/0.036	MAE/RMSE 分别提升 0.568/0.487
自注意力	DSANet[60]	TPA-LSTM[36]	Gas station data	首先用 2 个并行卷积结构来捕获全局和局部时间模式的复杂混合，然后利用 1 个自注意模块来建模多个时间序列之间的依赖关系。	RRSE/MAE 分别为 0.809/0.437	RRSE/MAE 分别为 0.771/0.410	RRSE/MAE 分别提升 0.047/0.062
	SAnD[61]	LSTM	MIMIC-Ⅲ	模型完全基于掩码自注意力机制，完全消除了递归。自注意力模块捕获序列中限制在邻域内的依赖关系，并使用多头注意力进行设计。此外，使用位置编码和密集插值嵌入技术将时间顺序合并到序列表示中。	MSE/MAPE 分别为 42 165/235.9	MSE/MAPE 分别为 39 918/157.8	MSE/MAPE 分别提升 0.053/0.331
	Attn-Embed[62]	PatchTST[63]	Traffic	使用全局地标和局部窗口构建的注意力图作为数据点的稳健核表示，能够抵抗噪声和分布偏移，将注意力权重作为时间序列数据的主要特征表示。	MSE/MAE 分别为 0.487/0.308	MSE/MAE 分别为 0.447/0.282	MSE/MAE 分别提升 0.082/0.084

2.4　Transformer

Transformer[30] 主要应用于自然语言处理等序列到序列的任务，由编码器（Encoder）和解码器（Decoder）组成，每个编码器模块由 1 个多头自注意力模块和 1 个位置前馈网络组成，而每个解码器模块在多头自注意力模块和位置前馈网络之间插入交叉注意力模型。Transformer 对软注意力提出了很多改进[54]，使得在没有循环网络单元的情况下进行序列到序列建模成为可能。Transformer 通过 Self-attention 捕获长程依赖和交互的能力对时间序列建模特别有吸引力，使得各种时间序列应用有了显著进展。在 Transformers 的变体中设计新的注意力模块占比最大[64]。但是传统的 Transformer 中的自注意模块存在较高的时间和内存复杂度，这在处理长序列时成为计算瓶颈。为了降低二次复杂度，人们提出了许多高效的 Transformer 变体。本文总结了基于注意力模块改进的 Transformer 的变体，相关模型具体见表 6。通过表 6 可以看出，大多数开发的时间序列 Transformer 模型保持了传统 Transformer 的架构，研究者引入各种时间序列归纳偏差来设计新的模块，大幅降低了时间和空间的复杂度。同时上述模型减小了实现信号传递遍历的最大路径长度，可以更好地捕捉时间序列地长期依赖。

表6 Transformer时间序列预测的相关研究

Table 6 Related research on Transformer time series forecasting

模型	注意力模块	优于	数据集	模型特点	性能指标		
					最优基线	本文模型	性能提升
AST[65]	稀疏注意力权重	DeepAR[51]	Electricity	在注意力头中用α-entmax替换softmax,学习稀疏注意力权重,以更好地关注与预测有关的历史时间步,并使用生成对抗编码器解码器框架训练模型,提高模型序列级预测性能,降低误差累积。	$R_{0.5}/R_{0.9}$分别为0.075/0.040	$R_{0.5}/R_{0.9}$分别为0.042/0.025	$R_{0.5}/R_{0.9}$分别提升0.44/0.375
Informer[66]	稀疏$\boldsymbol{Q}$矩阵	DeepAR[51]	ETT	通过允许每个key只关注主要query来获得ProbSparse自注意力,以解决Transformer时间复杂度高、内存占用高等问题。	MSE/MAE分别为0.162/0.327	MSE/MAE分别为0.158/0.319	MSE/MAE分别提升0.025/0.025
TFT[67]	多头注意聚合共享	Conv-Trans[68]	Electricity	为了学习不同尺度下的时序关系,TFT使用循环层进行局部处理,通过加性聚合多头注意力进行共享。TFT使用专门的组件来选择相关的特征和一系列的门控层来抑制不必要的组件。	$R_{0.5}/R_{0.9}$分别为0.059/0.034	$R_{0.5}/R_{0.9}$分别为0.055/0.027	$R_{0.5}/R_{0.9}$分别提升0.068/0.206
Pyraformer[69]	金字塔分层注意	Informer[66]	Electricity	设计金字塔注意力模块总结原始时间序列不同尺度的表示,形成一个C叉树,通过尺度内相邻节点的连接,较粗尺度更容易捕获长程依赖。比单独使用单一的最精细的尺度模型捕获的方式更能减轻模型的计算负担。	MSE/MAE分别为0.745/0.266	MSE/MAE分别为0.719/0.256	MSE/MAE分别提升0.035/0.038
FEDformer[70]	频域中的自注意力	Autoformer[71]	Electricity	提出了一种频率增强的分解Transformer,以捕捉时间序列的全局属性,并混合了专家进行季节性趋势分解。通过Fourier增强块和Wavelet增强块,替代自注意力模块和交叉注意力模块,能够通过频域映射在时间序列中捕获重要结构。	MSE/MAE分别为0.201/0.317	MSE/MAE分别为0.183/0.297	MSE/MAE分别提升0.09/0.063
Crossformer[72]	两阶段注意力	FEDformer	ETT	将数据按维度分割嵌入2D向量数组中,以保留时间和维度信息。然后提出两阶段注意来捕获跨时间和跨维度的依赖关系。建立一个分层编码器-解码器,以利用不同尺度的信息进行最终的预测。	MSE/MAE分别为0.318/0.384	MSE/MAE分别为0.305/0.367	MSE/MAE分别提升0.041/0.044
Robformer[73]	分解的Transformer	FEDformer	Electricity	由3个新的内部模块组成,分解架构和季节成分调整块,以缓解趋势项和周期项的突然波动和偏移对长期预测的影响。通过鲁棒的趋势预测块来提取长期时间序列的多种趋势模式。	MSE/MAE分别为0.190/0.305	MSE/MAE分别为0.184/0.292	MSE/MAE分别提升0.032/0.043

续表

模型	注意力模块	优于	数据集	模型特点	性能指标		
					最优基线	本文模型	性能提升
CATS[74]	Cross-attention	Time-mixer[75]	Electricity	仅包含交叉注意力的时间序列Transformer架构，用交叉注意力机制代替自注意力并优化参数共享，并通过设置未来预测期限作为查询和增强参数共享来优化模型，提高了长期预测准确性，同时减少了参数数量和内存使用。	MSE/MAE分别为0.157/0.247	MSE/MAE分别为0.149/0.237	MSE/MAE分别提升0.051/0.040

2.5　图神经网络

多变量时间序列建模长期以来是经济、金融和交通等领域的研究热点。多元时间序列预测的一个基本假设是其变量之间相互依赖。Jin等[76]综述了GNN在时间序列分析中的应用，包括预测、分类、异常检测和数据插补。GNN由于其排列不变性、局部连通性和组合性等特点，凸显了其在处理时间序列数据中的优势，成为对多变量时间序列数据进行建模很有效的方法，而传统的方法和其他基于深度神经网络的方法在这方面却难以做到。根据GNN中节点特征的聚合方式可以分为图卷积网络（Graph convolutional network，GCN）[77]和图注意力网络（Graph attention network，GAT）[78]。

2.5.1　图卷积网络

传统的CNN可以获取局部空间特征，但其只能在欧氏空间中使用，如图像、规则网格等。GCN是一种针对图结构数据泛化的特殊CNN，将卷积神经网络推广到可处理任意图结构数据。给定邻接矩阵和特征矩阵，GCN模型在傅里叶域构造的滤波器作用于图的节点。首先通过其一阶邻域捕获节点间的空间特征来学习节点嵌入，然后通过堆叠多个卷积层构建GCN模型。GCN在聚合过程中显式地为邻居节点分配一个非参数权重，如图6所示。

对于时间序列预测，开发了一些专门的架构，表7展示了针对不同时间序列预测问题使用GCN的相关研究。

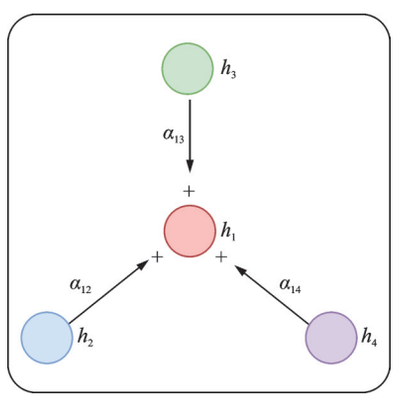

图6　GCN节点特征聚合

Fig.6　GCN node feature aggregation

表7　GCN时间序列预测的相关研究

Table 7　Related research on GCN time series forecasting

模型	优于	数据集	模型特点	性能指标		
				最优基线	本文模型	性能提升
T-GCN[3]	GRU	SZ-taxi/Los-loop	集成了GCN和GRU，利用GCN捕获数据拓扑结构，进行空间依赖性建模。GRU用于捕获数据的动态变化，以模拟时间依赖性。T-GCN模型也可用于其他时空预报任务，但只学习多个时间序列之间稳定的空间关系，限制了捕获具有多模式的空间依赖关系的能力。	RMSE/MAE/Accuracy/R^2分别为4.00/2.6/0.72/0.83	RMSE/MAE/Accuracy/R^2分别为3.93/2.71/0.73/0.85	RMSE/MAE/Accuracy/R^2分别提升0.018/0.045/0.007/0.026
STNN[79]	DC-GRU[80]	PeMSD4/PeMSD8	通过空间注意力网络来建模复杂和动态的空间相关性，无需昂贵的矩阵操作；利用Temporal Transformer建模跨多个时间步的长程时态依赖关系，可捕获历史数据的周期性依赖，但对计算和存储有较高的要求。	MAE/RMSE分别为17.86/27.82	MAE/RMSE分别为17.74/27.63	MAE/RMSE分别提升0.358/0.358
STGCN[39]	DC-GRU[80]	BJER4/PeMSD7	架构包含若干个时空卷积块，它们是图卷积层和卷积序列学习层的组合，用于建模空间和时间依赖关系，但依赖预定义的图结构。	MAE/MAPE/RMSE分别为3.84/9.31/5.22	MAE/MAPE/RMSE分别为3.78/9.11/5.20	MAE/MAPE/RMSE分别提升0.016/0.022/0.004

续表

模型	优于	数据集	模型特点	性能指标		
				最优基线	本文模型	性能提升
Graph WaveNet[81]	GGRU[82]	Metr-La/Pems-Bay	模型将图卷积层与自适应邻接矩阵和空洞因果卷积相结合,以捕获时空依赖关系,但没有考虑数据中的动态空间依赖关系。	MAE/RMSE/MAPE分别为2.71/5.24/0.070	MAE/RMSE/MAPE分别为2.69/5.15/0.069	MAE/RMSE/MAPE分别提升0.007 4/0.017/0.013
TPGCN[83]	SCID[84]	PeMSD8	TPGCN时间感知离散图结构估计(Time-aware discrete graph structure estimation, TADG)和动态个性化图卷积(Dynamic personalized graph convolutional, DPGC)两个组件。TADG通过动态输入推断受交互特性变化影响的图结构,DPGC在考虑演变模式和外部因素的同时,建模变量效应。在信息聚合过程中动态融合这两种信息。	MAE/RMSE/MAPE分别为14.20/23.49/9.28	MAE/RMSE/MAPE分别为13.45/22.98/8.79	MAE/RMSE/MAPE分别提升0.053/0.022/0.053

2.5.2 图注意力网络

GAT通过注意力机制为不同的相邻节点自适应地分配权重,以提高图模型的归纳学习能力。具体来说,引入注意力的思想,在通过堆叠这些隐藏的自注意力层获得全局信息之前,并行计算相邻节点的重要性。GAT以隐式地方式捕获权重α_{ij},使得更重要的节点获得更大的权重,如图7所示。为了实现更稳定的学习过程,GAT使用多头自注意力机制来捕获不同图层次的节点信息。此外,GAT结构可以使用反向传播算法进行训练,成为大规模深度学习框架的重要组成部分。

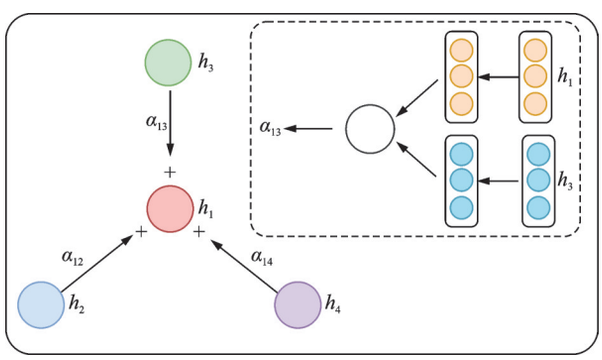

图7 GAT节点特征聚合

Fig.7 GAT node feature aggregation

目前,GAT已经成为链路预测、节点分类和交通时间序列建模等诸多领域最活跃的方法之一[85]。同时,它在工业领域的预测问题中也取得了很大的成功[86],期望为捕捉工业多变量时间序列内部交互特征提供新颖的解决方案。表8给出了使用GAT网络解决时间序列预测问题的相关研究。

表8 GAT时间序列预测的相关研究

Table 8 Related research on GAT time series forecasting

模型	优于	数据集	模型特点	性能指标		
				最优基线	本文模型	性能提升
TC-GATN[87]	DARNN[59]	Industrial data	通过MTS上的因果分析来学习有向图,改进的GAT模型在图邻域空间中引入并行的GRU编码器,以捕获数据中的非线性相关性。但从现有数据集中学习到的图结构是固定的,使得网络的边连接无法处理不同变量之间因果关系发生变化的情况。	RMSE/MAE/CORR分别为0.083/0.052/0.98	RMSE/MAE/CORR分别为0.074/0.046/0.99	RMSE/MAE/CORR分别提升0.114/0.114/0.007
GATCN[88]	STGCN[39]	PEMS-BAY	基于GAT和TCN的预测框架,通过GAT处理空间特征,通过TCN处理时间特征,通过GAT和TCN的融合层学习时空特征,同时考虑外生因素。此外,图中的节点可以通过堆叠多个层来捕捉其邻域的信息,但对动态交通状况的适应性较弱。	MAE/MAPE/RSME分别为2.25/0.053/4.04	MAE/MAPE/RSME分别为1.93/0.041/3.18	MAE/MAPE/RSME分别提升0.142/0.222/0.213
Weight-bounded GAT[89]	AST-GCN[90]	US stock-prices	为多变量时间序列预测开发了一种新的范数有界GAT,通过对GAT模型中每层权重的Frobenius范数设定上界来实现最优性能。	MSE/MAE分别为7.14/36.77	MSE/MAE分别为2.04/30.61	MSE/MAE分别提升0.714/0.168

通过表7、8可以看出，上述基于图神经网络的时间序列预测模型大多数用于交通数据的预测。交通数据的预测是城市交通管理和规划中的关键问题，而传统预测方法在面对数据稀疏性、非线性关系和复杂动态性等挑战时表现不佳。图神经网络是一种基于非欧结构数据的深度学习方法，在各种复杂网络建模和预测任务中得到广泛应用[91]。近来，很多方法将图神经网络与RNN或CNN集成在一起，在反映城市道路网络的复杂拓扑结构的同时满足中长期预测任务的要求。研究人员提出的此类时空图神经网络，相较之前的预测模型有显著的进步。

3　深度多任务时间序列预测

现有的单任务时间序列预测模型是对每个任务单独进行训练，各个任务之间的模型空间是相互独立的（图8(a)），没有很强的任务泛化能力。受到MTL[21]这一机器学习范式的启发，多个任务之间的模型空间是共享的（图8(b)），通过任务间知识共享提高每个任务的性能，降低每个任务的过拟合风险。例如，在医学指标预测中，在某一特定时间步为某一任务（如预测败血症发病时间）学习的知识可能对在稍后时间步学习另一任务（如预测死亡率）有用；在交通预测中，交通速度的预测和交通流量的预测显然是相互影响的任务，同时进行两个任务的学习有助于提高任务预测性能。因此，一些深度多任务学习模型被提出来用于时间序列的预测。本节回顾了深度多任务时间序列预测的架构，分别从网络体系结构信息共享机制和信息共享（交互）位置两个层面对不同的方法进行分类，并分析了这些模型的设计特点。

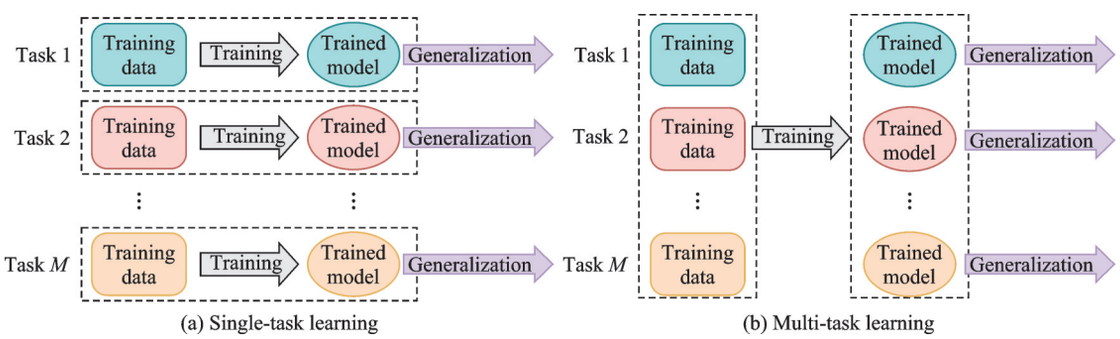

图8　单任务学习和多任务学习网络结构图

Fig.8　Network architecture diagrams for single-task learning and multi-task learning

3.1　硬参数共享和软参数共享

3.1.1　硬参数共享

硬参数共享是深度学习[92]中最常用的MTL方法之一。所有任务共享公共的神经网络底层（通常是网络的前面几层）的参数，用于提取不同任务之间的共性特征，而每个任务在网络的上层（通常是后面几层）有自己独立的参数，用于处理特定任务相关的特征。硬共享结构如图9所示。

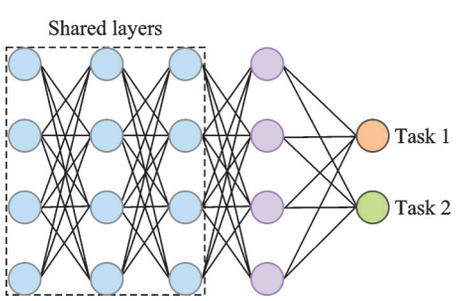

图9　硬参数共享的多任务学习结构

Fig.9　Multi-task learning structure for hard parameter sharing

通过硬共享方法最大限度地降低了训练过程中过拟合的风险，但在相关性较低的任务中会遇到困难，共享参数可能会导致任务之间的干扰，降低模型性能。另外，当所有任务都依赖于相同的共享参数时，对于那些具有特殊要求的任务，可能无法很好地适应。表9中列出通过硬共享机制进行深度多任务时间序列预测的相关研究。

表9　硬共享MTL时间序列预测的相关研究

Table 9　Related research on hard shared MTL time series forecasting

模型	优于	数据集	模型特点	性能指标		
				最优基线	本文模型	性能提升
AECRNN[93]	MTCNN[94]	Industrial data	CNN和自动编码器提取时间序列的鲁棒特征，然后分别输入到解码器和RNN进行序列重构和预测。	RMSE/MAPE分别为0.6/14.463	RMSE/MAPE分别为0.57/14.22	RMSE/MAPE分别提升0.05/0.0166

续表

模型	优于	数据集	模型特点	性能指标		
				最优基线	本文模型	性能提升
MURAT[95]	STNN[79]	BJS-Pickup/NYCTrip	将链接信息和时空信息嵌入学习空间中,并在嵌入向量上应用图拉普拉斯正则化执行先验和周期性的时空平滑度,经过残差网络再馈送到各任务网络。	MAPE/MAE/MARE分别为0.24/149.4/0.23	MAPE/MAE/MARE分别为0.22/139.44/0.21	MAPE/MAE/MARE分别提升0.080/0.067/0.074
LSTM-based[96]	XGBoost[46]	Wireless communication	通过线性变换和LSTM对不同任务数据进行特征提取,将提取的特征矩阵进行融合,最后分别输入到不同预测任务头。	MSE/MAE/MAPE分别为8.32/1.51/17.81	MSE/MAE/MAPE分别为6.46/1.32/16.63	MSE/MAE/MAPE分别提升0.224/0.126/0.066 3
MDL[97]	Conv-LSTM[98]/MRF[99]	TaxiBJ/TaxiNYC	MDL用于节点和边流量的预测,两个任务均用三流全卷积网络(Three-stream fully convolutional networks, 3S-FCN)捕获接近流、周期流和趋势流的时间相关性,然后用Concat进行特征融合,再运用卷积输入到两个任务网络,最后利用门控机制与外部因素融合后进行预测。	Task 1:RMSE/MAE分别为66.57/18.3 Task 2:RMSE/MAE分别为55.7/18.28	Task 1:RMSE/MAE分别为53.68/13.98 Task 2:RMSE/MAE分别为47.44/14.63	Task 1:RMSE/MAE分别提升0.193 6/0.236 1 Task 2:RMSE/MAE分别提升0.145 2/0.199 7

3.1.2 软参数共享

软参数共享不强制所有任务共享完全相同的参数。它为每个任务都设置了独立的参数集,但是会添加一些约束来鼓励不同任务参数之间的相似性。这种相似性可以通过一些正则化方法来实现,例如对不同任务参数之间的距离进行惩罚。软共享架构如图10所示。软参数共享的多任务学习结构每个任务都有自己独立的参数,能够更好地适应不同任务的特殊需求,避免了硬参数共享中可能出现的任务冲突问题。对于任务差异较大的情况,软参数共享可以更好地平衡不同任务之间的学习,同时仍然能够利用不同任务之间可能存在的共性。在软共享多任务时间序列预测模型中,任务间可以进行参数共享、特征共享和注意力共享,是一种比较灵活的共享方式。

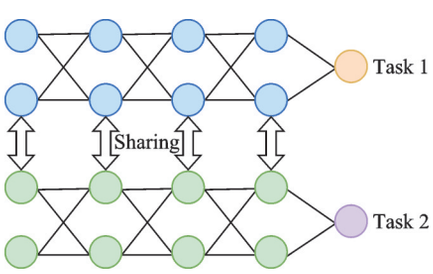

图10 软参数共享的多任务学习结构

Fig.10 Multi-task learning structure for soft parameter sharing

表10中列出了通过软共享机制进行深度多任务时间序列预测的相关研究。基于软共享的多任务学习方法明显依赖于预定义的共享结构,模型对新任务的泛化性能可能较差。由于每个任务都有自己独立的参数集,软参数共享的模型参数数量通常比硬参数共享多,这可能需要更多的计算资源和数据来进行训练。软共享机制需要设计合适的正则化方法来控制不同任务参数之间的相似性,训练过程相对复杂,需要仔细调整正则化参数等超参数。

表10 软共享MTL时间序列预测的相关研究

Table 10 Related research on soft shared MTL time series forecasting

模型	优于	数据集	模型特点	性能指标		
				最优基线	本文模型	性能提升
MTL-Trans[100]	SSP-MTL[101]	TRA-MI	模型通过设置外部公共多传感头注意力函数来捕获和存储跨不同任务的自注意力信息,然后通过拼接加权的方式进行共享。	CORR/RMSE/SMAPE分别为0.85/0.11/0.31	CORR/RMSE/SMAPE分别为0.87/0.11/0.30	CORR/RMSE/SMAPE分别提升0.018/0.020/0.021
Deep-TTE[102]	RNN-TTE[102]	Chengdu/Beijing traffic	基于CNN和LSTM的时空组件学习数据中的局部空间依赖性和时间依赖性及外部因素嵌入,进行局部路径旅行时间的估计。并且通过多因素注意力机制学习不同局部路径的隐藏表示和外部因素的权重进行整个路径旅行时间的估计。	MAPE/RMSE/MAE分别为15.65/358.74/246.52	MAPE/RMSE/MAE分别为11.89/282.55/186.93	MAPE/RMSE/MAE分别提升0.240/0.212/0.241

续表

模型	优于	数据集	模型特点	性能指标		
				最优基线	本文模型	性能提升
HUMTL-CGRUG[103]	HUMTL-CGRU[103]	Energy loads	用CNN提取特征与不同结构的GRU连接，建模时间动态性；再在不同网络上进行同方差不确定性多任务学习分析，并通过梯度提升回归树自动确定每个模型的权重进行集成。	Task 1：MAPE/RMSE分别为1.428/1.081 Task 2：MAPE/RMSE分别为1.349/0.305 Task 3：MAPE/RMSE分别为1.129/0.709	Task 1：MAPE/RMSE分别为1.396/1.061 Task 2：MAPE/RMSE分别为1.287/0.294 Task 3：MAPE/RMSE分别为1.085/0.687	Task 1：MAPE/RMSE分别提升0.022/0.018 5 Task 2：MAPE/RMSE分别提升0.046 0/0.036 Task 3：MAPE/RMSE分别提升0.039 0/0.031

3.2　以编码器为中心和以解码器为中心的多任务时间序列预测

多任务网络被分为硬参数共享技术或软参数共享技术。然而，最近的多任务学习工作从这两种共享技术中获得了灵感，共同解决了多个密集预测任务。因此，软参数共享与硬参数共享范式是否仍被应用作MTL架构分类的主要框架是有争议的。为了更好地区分不同的体系结构，出现了一种替代分类法，即根据任务之间交互或共享信息网络中的位置来进行分类。然而，在多任务时间序列预测方面，这种方法尚未得到深入研究。

以编码器为中心的架构（图11(a)）在使用独立的任务专用头解码每个任务之前，仅使用硬参数或软参数来共享编码器中的信息。不同的是，以解码器为中心的架构（图11(b)）也在解码阶段交换信息。

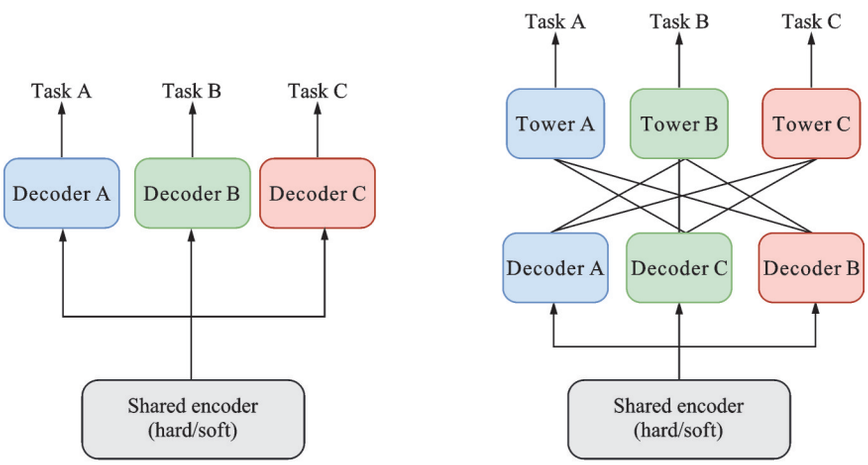

图11　以编码器为中心和以解码器为中心的模型结构

Fig.11　Model structures centered around the encoder and the decoder

3.2.1　以编码器为中心的架构

以编码器为中心的架构是指使用共享编码器或多个编码器来学习多个相关任务的时间序列预测模型。在这种方法中，编码器将输入序列转换为共享的特征表示，解码器将特征表示转换为每个任务的预测输出。这种方法的主要思想是通过共享特征表示来提高模型的效率和准确性，同时保留每个任务的独特特征表示。

在以编码器为中心的多任务时间序列预测架构中，有一种常见的分支结构：共享编码器作为骨干网络学习共享特征，然后分支到不同的特定任务，类似图9所示。如MTRL[104]，该模型依赖于共享网络来学习场景的通用表示，通过参数的硬共享将编码器的特征用于任务特定的头，以获得每个任务的预测。这个简单的模型在所有任务中共享完整的编码器，但受到“十字绣网络”[105]（Cross-stitch）和多任务注意力网络[106]（Multi-task attention network，MTAN）的启发，最近的工作已经考虑了在编码器中应该在哪里以及如何进行特征共享。例如，在Transformer的基础上，MTL-Trans[100]展示了两种不同的多任务间共享注意力架构，如图12(a,b)所示。通过设置外部公共

多传感头注意力函数来捕获和存储跨不同任务的自注意力信息。观察图12(a)可以清晰看出，MTL-Trans的混合注意力共享架构在形式上和“十字绣网络”是类似的。“十字绣网络”是使用多任务学习来学习卷积网络中的共享表示，通过端对端的学习来自动决定共享层，通过一个线性变换来“缝合”不同任务间的特征以实现特征共享。图12(a)架构是通过共享注意力模块进行不同任务特定编码器特征的“缝合”，从而实现任务间特征共享。这种注意力共享架构通过将任务特定编码器的输出反馈到共享的多头注意力层，可以加强共享知识的学习。MTAN将共享骨干网络与编码器中的任务特定注意力模块结合使用，每个特定任务的注意力模块通过应用软注意力掩码从共享网络中选择特征。类似地，MTL-Trans的另外一种通用的全局注意力共享架构由特定任务注意力编码器层和共享注意力层组成。与MTAN共享方式不同的是，MTL-Trans共享多头注意力层捕获所有任务的共享信息，通过与任务特定编码器输出进行拼接再计算加权平均信息，以使来自不同任务的信息充分交互。

“十字绣网络”的缺点是网络的大小随着任务的数量线性增加。此外，还不清楚十字绣单元应插入何处可以最大限度地提高其有效性。而水闸网络[107]可以支持子空间和跳跃连接的选择性共享，从而扩展了这项工作。以编码器为中心的多任务时间序列预测相关研究如表11所示。

3.2.2 以解码器为中心的架构

编码器结构遵循一个共同的模式：它们在1个处理周期中直接预测来自同一输入的所有任务输出。这样它们就无法捕捉到任务之间的共性和差异，而这些任务之间很可能相得益彰，这可能是以编码器为中心的方法对MTL只实现了适度的性能改进的原因。为了缓解这一问题，经过编码器处理后得到共享的特征表示在解码阶段也共享或交换信息，然后再分别预测每个任务的输出序列。将这样的模型分类为以解码器为中心的架构，如图12(b)所示。

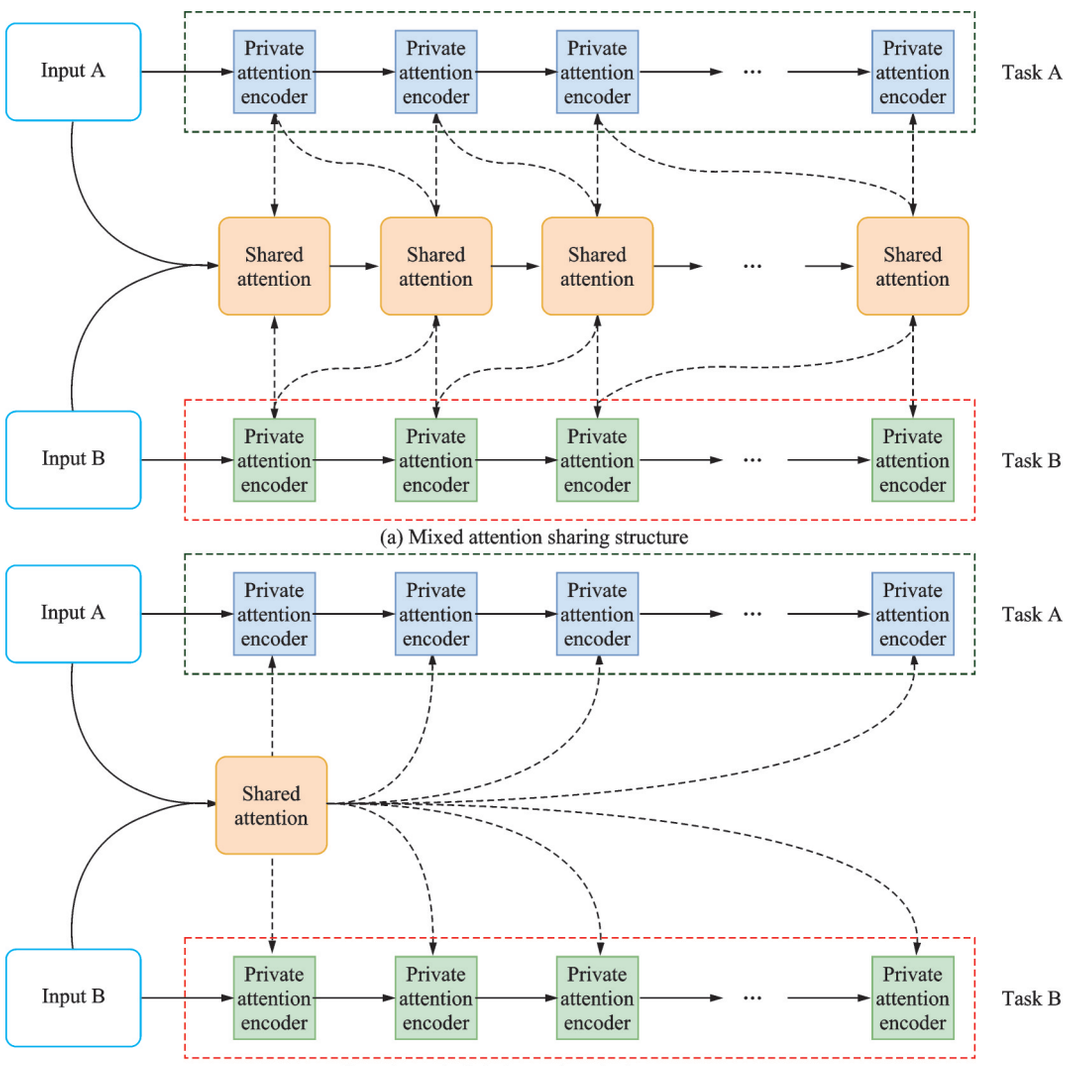

图12 MTL-Trans两种共享注意力网络结构图

Fig.12 Two shared attention network architectures of MTL-Trans

表 11　以编码器为中心的多任务时间序列预测的相关研究

Table 11　Related studies on encoder-centred multi-task time series forecasting

模型	优于	数据集	模型特点	性能指标		
				最优基线	本文模型	性能提升
MTL-RNN[108]	LSTM	MIMIC-Ⅲ v1.4	任务网络共享一个 LSTM 编码器提取时序特征。然后用于循环解码重构原始序列和预测记录的医院死亡率。并且，在每个任务的学习中都纳入注意力机制，分别实现充分运用编码器知识和为每个时间步设置不同的重要性级别。	AUPRC/PPV/Sensitivity 分别为 0.49/0.48/0.49	AUPRC/PPV/Sensitivity 分别为 0.52/0.49/0.503	AUPRC/PPV/Sensitivity 分别提升 0.068/0.029/0.033
CL-based[109]	TS2Vec[110]	ETT	数据预处理为上下文、时间和转换一致性 3 个自监督任务生成正负对，反馈到共享编码器。然后执行每个自监督任务并计算对比损失。通过同方差的不确定性权衡多个对比损失函数来推导出多任务损失。	MAE 分别为 0.712	MAE 分别为 0.690	MAE 分别提升 0.031
MTRL[104]	TriNet[111]	UCR archive	共享网络由深度小波分解网络和残差网络组成，提取隐藏在不同时域和频域中的信息。然后分别输入到任务特定网络进行学习任务的特定表示。	Accuracy 分别为 0.87	Accuracy 分别为 0.882	Accuracy 分别为 0.013 8

在以解码器为中心的架构中，解码器信息交互的方式也有所不同。如 MSJF[112]使用注意力机制来衡量共享信息和任务私有信息对自身预测任务的贡献，然后该模型将这些信息与它们的权重进行组合，得到最优的组合。而 TP-AMTL[113]采用了一种新的非对称知识转移概率公式，其中知识转移的数量取决于特征层面的不确定性。该框架利用特征层面的不确定性进行任务间和跨时间步的知识转移，从而同时利用任务相关性和时间依赖性。以解码器为中心的多任务时间序列预测相关研究如表 12 所示。

从表 11 和表 12 可以看出，在以信息共享（交互）位置对深度多任务时间序列预测进行分类时，以编码器为中心进行信息交互的研究模型相对较

表 12　以解码器为中心的多任务时间序列预测的相关研究

Table 12　Related studies on decoder-centred multi-task time series forecasting

模型	优于	数据集	模型特点	性能指标		
				最优基线	本文模型	性能提升
MSJF[112]	FC	Stock data	每个任务有特定的编码器学习任务特定潜在特征，然后通过注意力机制进行私-共享特征的信息交互，学习优化的表示以供每个任务的执行。	MAPE/MAE/MARE 分别为 2.51/0.011/1.19	MAPE/MAE/MARE 分别为 2.13/0.010/1.05	MAPE/MAE/MARE 提升 0.149/0.10/0.112
TP-AMTL[113]	SAnD[61]	MIMICⅢ/PhysioNet	模型共享低层网络，然后基于特征的不确定性来学习知识转移量，将知识加权组合进行跨任务和时间步非对称性知识转移，最后用组合特征执行每个预测任务。	AUROC 分别为 0.860 7	AUROC 分别为 0.874 3	AUROC 提升 0.015 8
Interpre-MTL[114]	KNNR[114]	Industrial	模型采用矩阵分解技术和注意力机制增强的编解码器架构。利用 MLP 作为编码器进行特征提取供解码器使用。然后将 3 个不同的解码器输出融合用于时间序列重建和预测两个任务。	MSE 分别为 4.95	MSE 分别为 4.77	MSE 提升 0.035

多,即先在编码器阶段先获取共享信息,然后用于不同的特定任务。这是比较简单的一种多任务学习共享架构,在任务相关性较强的时候比较适用。架构相对复杂的以解码器为中心进行信息交互的方法,可以在学习阶段获取与每个任务强相关的特征,如通过注意力机制、对比学习等方式,这在一定程度上可以抑制信息的负迁移。

4 结 论

本文首先概述了以CNN、RNN、Attention、Transformer以及GNN为代表的常见深度时间序列预测的方法,包括模型的组件构成、使用的数据集、已知模型改进方式和局限。总结发现,基于深度学习的时间序列预测方法具有一定的性能优势,但仍需要进一步的提升和完善。本文给出了时间序列预测领域的重点问题和进一步的研究方向。

(1)深度神经网络的性能在很大程度上取决于其各自架构和超参数设置的配置。然而,由于学习过程的网络复杂性和沉重的计算成本,手动寻找最优的网络配置极具挑战性。神经架构搜索(Neural architecture search,NAS)是一种自动设计神经网络架构的技术,具有出色的全局搜索能力,通过自动遍历给定任务的架构搜索空间,可以识别最佳的学习配置以及发现创新的网络结构。因此,在未来的研究中采用NAS来寻找最优深度神经网络配置将成为时间序列预测研究的热点之一。

(2)深入挖掘出时空序列数据中隐藏的动态变化的时间和空间相关性,更全面地提取出数据中的时间特征和空间特征,是时空序列分析和建模过程中面临的一个重点和难点。图神经网络因其处理不规则图结构数据的卓越能力而被引入。最近的时空图神经网络通常依赖于复杂的机制来捕捉动态依赖关系,这可能会引入过多的参数,并面临过度拟合的高风险。此外,其中一些模型还严重依赖于领域动态因素(如道路占用率和天气条件),在一定程度上失去了不同应用中的鲁棒性和泛化性。因此,平衡模型的动态性和鲁棒性是一个仍需深入研究的问题。

(3)神经网络模型在多任务学习方面展现出了广阔的前景,其重点是学习共享层以提取共同的任务变量特征。深度多任务学习中跨多个任务的特征共享的可能仍然主要是线性加权求和。如何在所提方法中融入非线性是把握更多任务相关性的未来方向。另外,在大多数现有方法中,提取的共享特征很容易受到特定任务特征或其他任务带来的噪声污染,如何有效缓解共享特征和私有特征的互相干扰也值得进一步深入研究。

(4)多任务学习旨在同时解决一系列相关任务,通过共享知识来提高单个任务的表现。因此,了解一组任务中的相似性是多任务学习的一个重要方面。之前的研究已经将相似性信息以显式方式(例如,每个任务的加权损失)或隐式方式(例如,用于特征自适应的对抗损失)融入其中,以取得良好的实证性能。然而,关于任务相似性知识的理论研究往往是缺失或不完整的,仍需进一步研究。

参考文献:

[1] ZHANG K, ZHENG L, LIU Z, et al. A deep learning based multitask model for network-wide traffic speed prediction[J]. Neurocomputing, 2020, 396: 438-450.

[2] WU Y, TAN H, QIN L, et al. A hybrid deep learning based traffic flow prediction method and its understanding[J]. Transportation Research Part C: Emerging Technologies, 2018, 90: 166-180.

[3] ZHAO L, SONG Y, ZHANG C, et al. T-GCN: A temporal graph convolutional network for traffic prediction[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 21(9): 3848-3858.

[4] MUDELSEE M. Trend analysis of climate time series: A review of methods[J]. Earth-Science Reviews, 2019, 190: 310-322.

[5] HAN J, LIU H, ZHU H, et al. Joint air quality and weather prediction based on multi-adversarial spatiotem poral networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2021: 4081-4089.

[6] XU X, YONEDA M. Multitask air-quality prediction based on LSTM-autoencoder model[J]. IEEE Transactions on Cybernetics, 2021, 51(5): 2577-2586.

[7] CHEN C W S, CHIU L M. Ordinal time series forecasting of the air quality index[J]. Entropy, 2021, 23(9): 1167.

[8] MA L, GAO J, WANG Y, et al. Adacare: Explainable clinical health status representation learning via scale-adaptive feature extraction and recalibration [C]//Proceedings of the AAAI Conference on Artificial Intelligence.[S.l.]: AAAI, 2020: 825-832.

[9] NGUYEN A T, JEONG H, YANG E, et al. Clinical risk prediction with temporal probabilistic asymmetric multi-task learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2021: 9081-9091.

[10] HARUTYUNYAN H, KHACHATRIAN H, KALE D C, et al. Multitask learning and benchmark-

ing with clinical time series data[J]. Scientific Data, 2019, 6(1): 96.

[11] VULETIĆ M, PRENZEL F, CUCURINGU M. Fin-GAN: Forecasting and classifying financial time series via generative adversarial networks[J]. Quantitative Finance, 2024, 24(2): 175-199.

[12] CHENG D, YANG F, XIANG S, et al. Financial time series forecasting with multi-modality graph neural network[J]. Pattern Recognition, 2022, 121: 108218.

[13] WIDIPUTRA H, MAILANGKAY A, GAUTAMA E. Multivariate CNN-LSTM model for multiple parallel fi nancial time-series prediction[J]. Complexity, 2021, 2021: 1-14.

[14] 赵洪科, 吴李康, 李徽, 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(8): 1621-1631.
ZHAO Hongke, WU Likang, LI Zheng, et al. Dynamic prediction of internet finance market based on deep neural network structure[J]. Journal of Computer Research and Development, 2019, 56(8): 1621-1631.

[15] AHMED N K, ATIYA A F, GAYAR N E, et al. An empirical comparison of machine learning models for time series forecasting[J]. Econometric Reviews, 2010, 29(5/6): 594-621.

[16] 卢宏涛,罗沐昆.基于深度学习的计算机视觉研究新进展[J].数据采集与处理, 2022, 37(2): 247-278.
LU Hongtao, LUO Mukun. Recent advances in computer vision research based on deep learning[J]. Journal of Data Acquisition and Processing, 2022, 37(2): 247-278.

[17] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S.l.]: ACL, 2019: 4171-4186.

[18] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484-489.

[19] YAMAK P T, YUJIAN L, GADOSEY P K. A comparison between ARIMA, LSTM, and GRU for time se ries forecasting[C]//Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence. Sanya, China: ACM, 2019: 49-55.

[20] HAN Z, ZHAO J, LEUNG H, et al. A review of deep learning models for time series prediction[J]. IEEE Sensors Journal, 2021, 21(6): 7833-7848.

[21] CARUANA R. Multitask learning[J]. Machine Learning, 1997, 28: 41-75.

[22] LIM B, ZOHREN S. Time-series forecasting with deep learning: A survey[J].Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2021, 379(2194): 20200209.

[23] 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.

[24] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[25] MEDSKER L R, JAIN L C. Recurrent neural networks: Design and applications[M]. Boca Raton: CRC Press, 2001.

[26] KIM Y, DENTON C, HOANG L, et al. Structured attention networks[EB/OL]. (2025-01-20). https://arxiv.org/abs/1702.00887.

[27] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of Advances in Neural Information Processing Systems. [S.l.]: [s.n.],2017: 5998-6008.

[28] SCARSELLI F, GORI M, TSOI A C, et al. The graph neural networkmodel[J]. IEEE Transactions on Neural Networks, 2008, 20(1):61-80.

[29] FUKUSHIMA K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position[J]. Biological Cybernetics, 1980, 36(4): 193-202.

[30] DONG X S, QIAN L J, LEI H. Short-term load forecasting in smart grid: A combined CNN and K-means clustering approach[C]//Proceedings of 2017 IEEE International Conference on Big Data and Smart Computing (BigComp). Jeju Island, South Korea: IEEE, 2017: 119-125.

[31] BOROVYKH A, BOHTE S, OOSTERLEE C W. Conditional time series forecasting with convolutional neural networks[J]. The Journal of Computational Finance, 2022, 25(4): 69-89.

[32] TIAN C, MA J, ZHANG C, et al. A deep neural network model for short-term load forecast based on long short-term memory network and convolutional neural network[J]. Energies, 2018, 11(12): 3493.

[33] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[34] BAI S, KOLTER J Z, KOLTUN V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling[EB/OL]. (2018-04-18).https://doi.org/10.48550/arXiv.1803.01271.

[35] 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.

[36] SHIH S Y, SUN F K, LEE H. Temporal pattern attention for multivariate time series forecasting[J]. Ma chine Learning, 2019, 108: 1421-1441.

[37] WAN R, MEI S, WANG J, et al. Multivariate temporal convolutional network: A deep neural net works approach for multivariate time series forecasting[J]. Electronics, 2019, 8(8): 876.

[38] SEN R, YU H F, D HILLON I S. Think globally, act locally: A deep neural network approach to high-dimensional time series forecasting[C]//Proceedings of Advances in Neural Information Processing Systems. [S.l.]:[s.n.],2019: 4837-4846.

[39] YU B, YIN H T, ZHU Z X. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden: IJCAI Organization, 2018: 3634-3640.

[40] LECUN Y, BENGIO Y, HINTON G. Deep learning [J]. Nature, 2015, 521(7553): 436-444.

[41] CHUNG J, GULCEHRE C, CHO K H,et al.Empirical evaluation of gated recurrent neural networks on sequence modeling[J].Eprint Arxiv, 2014.DOI:10.48550/arXiv.1412.3555.

[42] SONG X, LIU Y, XUE L, et al. Time-series well performance prediction based on long short-term memory (LSTM) neural network model[J]. Journal of Petroleum Science and Engineering, 2020, 186: 106682.

[43] SAGHEER A, KOTB M. Time series forecasting of petroleum production using deep LSTM recurrent net works[J]. Neurocomputing, 2019, 323: 203-213.

[44] ABBASIMEHR H, PAKI R. Improving time series forecasting using LSTM and attention models[J]. Journal of Ambient Intelligence and Humanized Computing, 2022, 13(1): 673-691.

[45] 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, China: ACM, 2019: 299-308.

[46] CHEN T, GUESTRIN C. XGBoost: A scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.]: ACM,2016: 785-794.

[47] ZHAI N, YAO P, ZHOU X. Multivariate time series forecast in industrial process based on XGBoost and GRU[C]//Proceedings of 2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC). [S.l.]: IEEE, 2020: 1397-1400.

[48] XU J, WANG K, LIN C, et al. FM-GRU: A time series prediction method for water quality based on seq2seq framework[J]. Water, 2021, 13(8): 1031.

[49] SUTSKEVER I, VINYALS O, LE Q V.Sequence to sequence learning with neural networks[J].Advances in Neural Information Processing Systems, 2014. DOI:10.48550/arXiv.1409.3215.

[50] KIM J, KIM H, KIM H G, et al. A comprehensive survey of deep learning for time series forecasting: Architectural diversity and open challenges[J]. Artificial Intelligence Review, 2025, 58(7): 1-95.

[51] SALINAS D, FLUNKERT V, GASTHAUS J, et al. DeepAR: Probabilistic forecasting with autoregressive recur rent networks[J]. International Journal of Forecasting, 2020, 36(3): 1181-1191.

[52] 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(2): 416-439.
WANG Wenguan, SHEN Jianbing, JIA Yunde. Visual attention detection: A survey[J]. Journal of Software, 2019, 30(2): 416-439.

[53] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and trans late[J]. Computer Science, 2014. DOI: 10.48550/arXiv.1409.0473.

[54] XU K, BA J, KIROS R,et al. Show, attend and tell: Neural image caption generation with visual attention[J]. Computing Research Repository, 2015, 3: 2048-2057.

[55] NIU Z, ZHONG G, YU H. A review on the attention mechanism of deep learning[J]. Neurocomputing, 2021, 452: 48-62.

[56] SU H, WANG X, QIN Y, et al. Attention based adaptive spatial-temporal hypergraph convolutional networks for stock price trend prediction[J]. Expert Systems with Applications, 2024, 238: 121899.

[57] CHENG R, LI Q. Modeling the momentum spillover effect for stock prediction via attribute-driven graph attention networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2021: 55-62.

[58] YIN X, HAN Y, SUN H, et al. A multivariate time series prediction schema based on multi-attention in recurrent neural network[C]//Proceedings of 2020 IEEE Symposium on Computers and Communications (ISCC). Rennes, France: IEEE, 2020: 1-7.

[59] QIN Y, SONG D, CHEN H, et al. A dual-stage attention-based recurrent neural network for time series prediction[C]//Proceedings of the 26th International

Joint Conference on Artificial Intelligence. Melbourne, Australia: IJCAI Organization, 2017: 2627-2633.

[60] HUANG S, WANG D, WU X, et al. DSANet: Dual self-attention network for multivariate time series forecasting[C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Beijing, China: ACM, 2019: 2129-2132.

[61] SONG H, RAJAN D, THIAGARAJAN J, et al. Attend and diagnose: Clinical time series analysis using atten tion models[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2018: 4091-4098.

[62] NIÜ P S, ZHŌU T, WÁNG X, et al. Attention as robust representation for time series forecasting[EB/OL].[2024-12-25]. https://arxiv.org/abs/2402.05370v1.

[63] NIE Y, NGUYEN N H, SINTHONG P, et al. A time series is worth 64 words: Long-term forecasting with transformers[C]//Proceedings of The Eleventh International Conference on Learning Representations. Kigali, Rwanda: ICLR, 2023.

[64] WEN Q, ZHOU T, ZHANG C, et al. Transformers in time series: A survey[C]//Proceedings of the 32nd International Joint Conference on Artificial Intelligence. Macao, China: IJCAI Organization, 2023: 6778-6787.

[65] WU S, XIAO X, DING Q, et al. Adversarial sparse transformer for time series forecasting[C]//Proceedings of Advances in Neural Information Processing Systems. [S.l.]: Curran Associates, Inc., 2020: 17105-17115.

[66] 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. [S.l.]: AAAI, 2021: 11106-11115.

[67] 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.

[68] LI S, JIN X, XUAN Y, et al. Enhancing the locality and breaking the memory bottleneck of Transformer on time series forecasting[C]//Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Vancouver, Canada: [s.n.], 2019.

[69] LIU S, YU H, LIAO C, et al. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting[EB/OL].(2025-01-20). https://openreview.net/forum? id=0EXmFzUn5I.

[70] ZHOU T, MA Z, WEN Q, et al. FEDformer: Frequency enhanced decomposed Transformer for long-term series forecasting[C]//Proceedings of the 39th International Conference on Machine Learning. Baltimore, Maryland, USA: PMLR, 2022: 27268-27286.

[71] 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.

[72] ZHANG Y, YAN J. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting[C]//Proceedings of the Eleventh International Conference on Learning Representations. Kigali, Rwanda: ICLR, 2023: 6778-6787.

[73] YU Y, MA R, MA Z. Robformer: A robust decomposition transformer for long-term time series fore casting[J]. Pattern Recognition, 2024, 153: 110552.

[74] KIM D, PARK J, LEE J, et al. Are self-attentions effective for time series forecasting?[EB/OL]. (2025-01-20). https://arxiv.org/abs/2405.16877.

[75] WANG S, WU H, SHI X, et al. TimeMixer: Decomposable multiscale mixing for time series forecasting[C]//Proceedings of the Twelfth International Conference on Learning Representations. Vienna, Austria: ICLR, 2024.

[76] JIN M, KOH H Y, WEN Q, et al. A survey on graph neural networks for time series: Forecasting, classification, imputation, and anomaly detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023. DOI: 10.1109/TPAMI. 2024. 3443141.

[77] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[C]//Proceedings of the 5th International Conference on Learning Representations. Toulon, France: ICLR, 2017.

[78] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks[C]//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: ICLR, 2018: 1-12.

[79] YIN X, LI F, WU G, et al. STNN: A spatial-temporal graph neural network for traffic prediction[C]//Proceedings of 2021 IEEE 27th International Conference on Parallel and Distributed Systems (ICPADS). Beijing, China: IEEE, 2021: 146-152.

[80] LI Y, YU R, SHAHABI C, et al. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting[C]//Proceedings of the 6th international Conference on Learning Representations. Vancouver, Canada: ICLR, 2018.

[81] WU Z, PAN S, LONG G, et al. Graph WaveNet for

deep spatial-temporal graph modeling[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China: IJCAI Organization, 2019: 1907-1913.

[82] YAN S, XIONG Y, LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1): 7444-7452.

[83] LI Z, GAO Z, ZHANG X, et al. Time-aware personalized graph convolutional network for multivariate time series forecasting[J]. Expert Systems with Applications, 2024, 240: 122471.

[84] SHAO Z, ZHANG Z, WANG F, et al. Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting[C]//Proceedings of the 31st ACM International Conference on Information & Knowledge Management. Atlanta, GA, USA: ACM, 2022: 4454-4458.

[85] YAN Y, REN W, HU X, et al. SRGAT: Single image super-resolution with graph attention network[J]. IEEE Transactions on Image Processing, 2021, 30: 4905-4918.

[86] ZHANG M, HU H, LI Z, et al. Proposal-based graph attention networks for workflow detection[J]. Neural Processing Letters, 2022, 54(1): 101-123.

[87] LI J, SHI Y, LI H, et al. TC-GATN: Temporal causal graph attention networks with nonlinear paradigm for multivariate time series forecasting in industrial processes[J]. IEEE Transactions on Industrial Informatics, 2023, 19(10): 10195-10204.

[88] GUO G, YUAN W. Short-term traffic speed forecasting based on graph attention temporal convolutional networks[J]. Neurocomputing, 2020, 410: 387-393.

[89] ZHANG Z, LI W, LIU H. Multivariate time series forecasting by graph attention networks with the oretical guarantees[C]//Proceedings of International Conference on Artificial Intelligence and Statistics. [S.l.]: PMLR, 2024: 2845-2853.

[90] GUO S, LIN Y, FENG N, et al. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33: 922-929.

[91] 汪维泰, 王晓强, 李雷孝, 等. 时空图神经网络在交通流预测研究中的构建与应用综述[J]. 计算机工程与应用, 2024, 60(8): 31-45.
WANG Weitai, WANG Xiaoqiang, LI Leixiao, et al. Construction and application of spatial-temporal graph neural network in traffic flow prediction: A survey[J]. Computer Engineering and Applications, 2024, 60(8): 31-45.

[92] CARUANA R. Multitask learning: A knowledge-based source of inductive bias[C]//Proceedings of the Tenth International Conference on Machine Learning. Amherst, MA, USA: University of Massachusetts, 1993.

[93] CIRSTEA R G, MICU D V, MURESAN G M, et al. Correlated time series forecasting using deep neural networks: A summary of results[EB/OL]. (2018-04-29). https://doi.org/10.48550/arXiv.1808.09794.

[94] PANG N, YIN F, ZHANG X, et al. A robust approach for multivariate time series forecasting[C]//Proceedings of the 8th International Symposium on Information and Communication Technology. New York, NY: ACM, 2017: 106-113.

[95] LI Y, FU K, WANG Z, et al. Multi-task representation learning for travel time estimation[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, United Kingdom: ACM, 2018: 1695-1704.

[96] CAO K, HU T, LI Z, et al. Deep multi-task learning model for time series prediction in wireless com munication[J]. Physical Communication, 2021, 44: 101251.

[97] ZHANG J, ZHENG Y, SUN J, et al. Flow prediction in spatio-temporal networks based on multitask deep learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 32(3): 468-478.

[98] SHI X, CHEN Z, WANG H, et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting[C]//Proceedings of Advances in Neural Information Processing Systems. [S.l.]: [s.n], 2015: 28.

[99] HOANG M X, ZHENG Y, SINGH A K. FCCF: Forecasting citywide crowd flows based on big data [C]//Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. Burlingame, California, USA: ACM, 2016: 1-10.

[100] CHEN Z, JIAZE E, ZHANG X, et al. Multi-task time series forecasting with shared attention[C]//Proceedings of 2020 IEEE International Conference on Data Mining Workshops (ICDMW). Sorrento, Italy: IEEE, 2020: 917-925.

[101] LIU P, QIU X, HUANG X. Recurrent neural network for text classification with multi-task learning [C]//Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, California, USA: AAAI, 2017: 4203-4209.

[102] WANG D, ZHANG J, CAO W, et al. When will you arrive? Estimating travel time based on deep neural networks[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2018: 2500-2507.

[103]WANG X, WANG S X, ZHAO Q Y, et al. A multi-energy load prediction model based on deep multi-task learning and ensemble approach for regional integrated energy systems[J]. International Journal of Electrical Power & Energy Systems, 2021, 126: 106583.

[104]CHEN L, CHEN D, YANG F, et al. A deep multi-task representation learning method for time series classification and retrieval[J]. Information Sciences, 2021, 555: 17-32.

[105]MISRA I, SHRIVASTAVA A, GUPTA A, et al. Cross-stitch networks for multi-task learning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016: 3994-4003.

[106]LIU S, JOHNS E, DAVISON A J. End-to-end multi-task learning with attention[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 1871-1880.

[107]RUDER S, BINGEL J, AUGENSTEIN I, et al. Latent multi-task architecture learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2019: 4822-4829.

[108]YU R, ZHENG Y, ZHANG R, et al. Using a multi-task recurrent neural network with attention mechanisms to predict hospital mortality of patients[J]. IEEE Journal of Biomedical and Health Informatics, 2020, 24(2): 486-492.

[109]CHOI H, KANG P. Multi-task self-supervised time-series representation learning[J]. Information Sciences, 2024, 671: 120654.

[110]YUE Z, WANG Y, DUAN J, et al. Ts2vec: Towards universal representation of time series[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2022: 8980-8987.

[111]WANG J, WANG Z, LI J, et al. Multilevel wavelet decomposition network for interpretable time series analysis[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, United Kingdom: ACM, 2018: 2437-2446.

[112]MA T, TAN Y. Multiple stock time series jointly forecasting with multi-task learning[C]//Proceedings of 2020 International Joint Conference on Neural Networks (IJCNN). Glasgow, UK: IEEE, 2020: 1-8.

[113]NGUYEN A T, JEONG H, YANG E, et al. Clinical risk prediction with temporal probabilistic asymmetric multi-task learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2021: 9081-9091.

[114]YEH C H, FAN Y C, PENG W C. Interpretable multi-task learning for product quality prediction with attention mechanism[C]//Proceedings of 2019 IEEE 35th International Conference on Data Engineering (ICDE). Macao, China: IEEE, 2019: 1910-1921.

（编辑：刘彦东）