文章编号:1671-6833(2026)02-0009-07

基于 iTransformer 的轻量级时序预测模型

周清雷[1], 王宇静[2], 段鹏松[2], 王 超[2], 郑永利[2]

(1. 郑州大学 计算机与人工智能学院,河南 郑州 450001;2. 郑州大学 网络空间安全学院,河南 郑州 450002)

摘 要: 针对时序预测领域难以平衡预测精度与时效性问题,以 iTransformer 模型为基础框架,提出一种轻量级时序预测模型 ILformer。iTransformer 作为基于变量的典型时序预测模型,能有效捕获多变量间复杂交互关系,但其存在计算复杂度较高与参数量较大的局限性,导致在资源受限的实际应用场景中模型难以高效部署。ILformer 针对这些不足展开优化。首先,引入线性注意力机制(Linear Attention)替代传统注意力机制,使输入处理更加灵活,通过线性投影和维度重排,ILformer 在减少参数量的同时,能更好地适应不同输入形状和结构,尤其在处理大规模数据时计算效率较高,并能在不降低模型精度前提下显著减少注意力模块的计算复杂度;其次,通过对注意力机制进行奇异值分解实现矩阵降维,大幅减少了矩阵乘法和加法的计算次数,提升了计算效率,同时降低了模型的过拟合风险;最后,在 8 个不同数据集上进行实验。实验结果表明:ILformer 在保持相同精度的同时,推理速度提高了 40.46%,参数量减少了 78.75%,且计算量减半,展示了优异性能与实用性。

关键词:时序预测; 轻量级; 奇异值分解; 线性注意力机制

中图分类号: TP39;TP183 **文献标志码**:A **doi**:10.13705/j.issn.1671-6833.2026.02.008

时序预测通过分析历史数据之间的关联来预测未来趋势,是目前的研究热点之一。在某些领域中,决策者需根据短期时序数据变化快速作出精准决策,对预测实时性要求极高。如在金融市场中,实时预测可帮助交易者迅速作出买卖决策;在能源行业,实时预测有助于提升电网运行效率和可靠性;在城市交通领域,实时预测能为城市规划和交通管理提供重要决策支持。因此,时序预测模型在保持高精度的同时,还需兼顾时效性[1-2]。

早期的时序预测广泛采用循环神经网络(RNN)[3],能够以自回归方式有效捕捉序列中的规律性。然而,RNN 存在梯度消失和爆炸问题,限制了处理长序列的能力。长短期记忆网络(LSTM)[4]和门控循环单元(GRU)[5]在一定程度上缓解了上述问题,但仍难以同时有效处理长期与短期依赖。2017 年,Transformer 模型[6-7]被提出,凭借其强大的并行计算能力和全局依赖建模优势,在多项时序预测任务中表现卓越,成为研究热点。

根据建模机制的不同,基于 Transformer 的时序预测模型可分为两类:基于块的模型和基于变量的模型。基于块的模型,例如 PatchTST[8],将时序划分为片段,并利用注意力机制挖掘局部时序模式,但在多变量交互建模方面存在不足;基于变量的模型,例如 iTransformer[9],通过变量间的注意力机制显式建模变量相关性,特别适用于高维多元时序预测。然而,这类模型的计算复杂度和序列长度呈指数(平方)增长,严重限制了其在实时场景中的应用。

为降低计算和存储开销,多种模型轻量化技术被提出,如模型剪枝[10]、量化[11]、知识蒸馏[12]和矩阵压缩[13]等。模型剪枝通过移除网络中的冗余连接或非必要结构减小模型规模;量化通过降低数值精度减少存储和计算开销;知识蒸馏通过师生框架提升小模型的表达能力;矩阵压缩,例如奇异值分解(singular value decomposition, SVD)[14]则利用权重

收稿日期:2025-10-18;**修订日期**:2025-12-16

基金项目:河南省科技攻关项目(232102210050, 242102210060);河南省自然科学基金资助项目(222300420295, 242300421474)

作者简介:周清雷(1962—),男,河南郑州人,郑州大学教授,博士,博士生导师,主要从事信息安全、计算复杂性理论研究,E-mail:ieqlzhou@zzu.edu.cn。

通信作者:段鹏松(1983—),男,山西运城人,郑州大学副教授,博士,主要从事车联网安全、无线感知、大数据与人工智能、知识图谱的研究,E-mail:duanps@zzu.edu.cn。

引用本文:周清雷,王宇静,段鹏松,等. 基于 iTransformer 的轻量级时序预测模型[J]. 郑州大学学报(工学版),2026,47(2):9-15,26. (ZHOU Q L, WANG Y J, DUAN P S, et al. Lightweight time series forecasting model based on iTransformer[J]. Journal of Zhengzhou University(Engineering Science),2026,47(2):9-15,26.)

矩阵的低秩特性进行更底层的数字重构与参数压缩,实现效率提升。这些方法在视觉和自然语言处理任务中取得了显著进展,然而在时序预测方面仍处于初级阶段。

针对上述问题,本文提出一种基于 iTransformer 的轻量级时序预测模型 ILformer。该模型引入奇异值分解对注意力输出矩阵进行降维,将参数量从 6.40×10^6 减少至 1.36×10^6,显著缓解了长序列预测中的内存瓶颈问题。同时,ILformer 采用改进的线性注意力机制[15],将计算复杂度从指数降至线性,计算量从 57×10^6 降低至 25×10^6,推理时间从 97.1 ms 缩短至 42.0 ms,且展现出较好的性能平衡。

1 ILformer 模型

本文提出的 ILformer 模型整体架构如图 1 所示。首先,在编码层中,对输入数据进行转置处理,将每个特征的时序整体视为一个编码层的 Token;其次,通过多头线性注意力机制重新调整矩阵连乘的计算次序,有效提升推理速度;最后,输出矩阵通过奇异值分解进行降维处理,进一步降低计算量和内存开销。此外,ILformer 沿用了 Transformer 中的层归一化和前馈网络设计,并在最后加入一个线性层作为解码层,生成未来 n 个时间段的预测值。

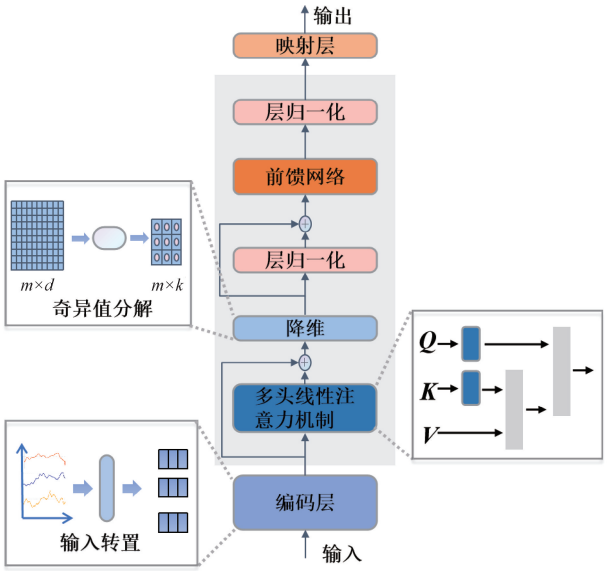

图 1 ILformer 模型架构图

Figure 1 ILformer model architecture

ILformer 算法伪代码如下。

算法 1 ILformer 算法。

输入:时间序列 $\boldsymbol{X}$、SVD 截断秩 k、编码器层数 L;

输出:预测序列 $\boldsymbol{Y}$。

① $\boldsymbol{X}=\mathrm{LayerNorm}(\boldsymbol{X})$;

//输入序列归一化处理

② $\boldsymbol{X}=\boldsymbol{X}.\mathrm{transpose}$//维度转置,变量视为 Token;

//多层感知机对特征进行嵌入编码

③ $\boldsymbol{H}^0=\mathrm{MLP}(\boldsymbol{X})$;

④ for l in $\{1,\cdots,\mathrm{L}\}$:

⑤ $\quad\boldsymbol{Z}=\mathrm{LinearAttention}(\boldsymbol{H}^{l-1})$//线性注意力机制;

⑥ $\quad\boldsymbol{Z}_k=\mathrm{SVD}(\boldsymbol{Z},k)$;

//奇异值分解降维

//残差连接与层归一化

⑦ $\quad\boldsymbol{H}'=\mathrm{LayerNorm}(\boldsymbol{H}^{l-1}+\boldsymbol{Z}_k)$;

//前馈神经网络用于特征提取与融合

⑧ $\quad\boldsymbol{H}^l=\mathrm{LayerNorm}(\boldsymbol{H}'+\mathrm{FeedForward}(\boldsymbol{H}'))$;

⑨ End for

//解码层:将 Token 投影回预测序列长度

⑩ $\boldsymbol{Y}=\mathrm{Projection}(\boldsymbol{H}^L)$;

⑪ $\boldsymbol{Y}=\boldsymbol{Y}.\mathrm{transpose}$;

⑫ Return $\boldsymbol{Y}$

1.1 线性注意力机制

对于 Transformer 类模型,自注意力机制需要为每个时间步生成查询($\boldsymbol{Q}$)、键($\boldsymbol{K}$)和值($\boldsymbol{V}$)矩阵。在传统的 softmax 注意力机制[16]中,假定输入序列长度为 m,嵌入维度为 d,即注意力输入为 $\boldsymbol{X}\in\mathbf{R}^{m\times d}$,分别乘上 3 个权值矩阵 $\boldsymbol{W}^{\boldsymbol{Q}}\in\mathbf{R}^{d\times d_q}$,$\boldsymbol{W}^{\boldsymbol{K}}\in\mathbf{R}^{d\times d_k}$ 及 $\boldsymbol{W}^{\boldsymbol{V}}\in\mathbf{R}^{d\times d_v}$ 后得到 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 矩阵,对于自注意力有 $d_q=d_k=d_v$,因此矩阵可如下表示:

$$\begin{cases}\boldsymbol{Q}=\boldsymbol{X}\boldsymbol{W}^{\boldsymbol{Q}};\\ \boldsymbol{K}=\boldsymbol{X}\boldsymbol{W}^{\boldsymbol{K}};\\ \boldsymbol{V}=\boldsymbol{X}\boldsymbol{W}^{\boldsymbol{V}}。\end{cases}\tag{1}$$

给定 3 个矩阵后,自注意力的计算方式如下:

$$\boldsymbol{A}=\mathrm{softmax}(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}/\sqrt{d})\boldsymbol{V}。\tag{2}$$

矩阵 $\boldsymbol{Q}$ 与 $\boldsymbol{K}^{\mathrm{T}}$ 相乘之后除以 $\sqrt{d}$,主要是解决 softmax 激活函数梯度消失的问题。将上式展开得到每行自注意力矩阵的计算方式如下:

$$\boldsymbol{A}_i=\frac{\sum_{j=1}^{N}\exp(\boldsymbol{Q}_i\boldsymbol{K}_j^{\mathrm{T}}/\sqrt{d})\boldsymbol{V}_j}{\sum_{j=1}^{N}\exp(\boldsymbol{Q}_i\boldsymbol{K}_j^{\mathrm{T}}/\sqrt{d})}。\tag{3}$$

在非线性注意力中,要求先计算 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 的矩阵乘法,再计算其与 $\boldsymbol{V}$ 的乘法。可依乘法结合律,将 $(\phi(\boldsymbol{Q})\phi(\boldsymbol{K})^{\mathrm{T}})\boldsymbol{V}$ 改为 $\phi(\boldsymbol{Q})(\phi(\boldsymbol{K})^{\mathrm{T}}\boldsymbol{V})$,即用 $\phi(\boldsymbol{Q})$ 乘上 $\phi(\boldsymbol{K})^{\mathrm{T}}\boldsymbol{V}$[17],如图 2 所示。

输入的 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 通过线性变换进行投影 $\boldsymbol{Q}'=\boldsymbol{Q}\boldsymbol{W}_q$,$\boldsymbol{K}'=\boldsymbol{K}\boldsymbol{W}_k$,$\boldsymbol{V}'=\boldsymbol{V}\boldsymbol{W}_v$,将投影后的查询、键和值矩阵变换为多头注意力的格式,并改变原多头注意

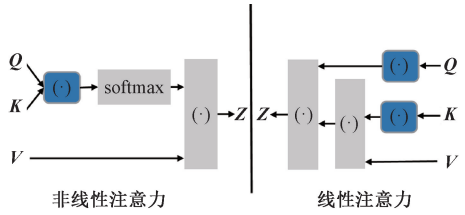

图 2　注意力计算顺序图

Figure 2　Attention computation order diagram

力中的排列顺序。本文采用爱因斯坦求和规则[18]计算 $\boldsymbol{Q}$ 和 $\boldsymbol{V}$ 的相似度,具体如下:

$$scores = \boldsymbol{Q}'(\boldsymbol{K}')^{\mathrm{T}}。\tag{4}$$

在注意力机制中,softmax 激活函数用来实现归一化,因此使用式(5)计算注意力:

$$\boldsymbol{A} = \mathrm{softmax}\left(\frac{scores}{\sqrt{d}}\right)。\tag{5}$$

最终,将注意力输出矩阵 $\boldsymbol{A}$ 与 $\boldsymbol{V}'$ 相乘,再通过输出层权重 $\boldsymbol{W}_{\mathrm{o}}$ 进行映射,得到输出矩阵:

$$\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{V}'\boldsymbol{W}_{\mathrm{o}}。\tag{6}$$

传统注意力机制在处理长时序数据时注意力分布往往较为稀疏,导致模型难以有效捕捉序列中的关键模式[19]。本文通过简化计算,将注意力机制的复杂度从传统的二次复杂度降低至线性,线性注意力机制不仅大幅提高了计算效率,还保持了模型对关键模式的高效建模能力。

1.2　低秩分解

注意力机制中矩阵由输入序列生成的向量构成,用于计算各时间步的相关性和依赖关系。由于需实时从输入数据中生成,且运算消耗大量计算资源和时间,尤其对于维度差异较大的矩阵 $\boldsymbol{V}$ 和 $\boldsymbol{K}$,在乘积计算中易出现大量冗余数据,增加计算开销,因此,对模型中矩阵的维度进行缩减,能够显著降低矩阵乘法的计算量[20]。

为加速计算并减少计算复杂度,本文模型在原始模型的自注意力机制后引入奇异值分解[21]来降低 $\boldsymbol{Z}$ 的维度并保留最重要的特征,分解公式如下:

$$\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}。\tag{7}$$

式中:$\boldsymbol{U} \in \mathbf{R}^{m\times m}$ 为左奇异向量矩阵;$\boldsymbol{\Sigma} \in \mathbf{R}^{m\times d}$ 为对角矩阵;$\boldsymbol{V} \in \mathbf{R}^{d\times d}$ 为右奇异向量矩阵。

矩阵 $\boldsymbol{Z}_p$ 可表示为奇异值和奇异向量的加权和:

$$\boldsymbol{Z}_p = \sum_{i=1}^{p} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}}。\tag{8}$$

式中:σ_i 为奇异值;$\boldsymbol{u}_i$ 和 $\boldsymbol{v}_i$ 分别为 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 的第 i 列;$p = \min(m,d)$。为了降低维度,仅保留前 k 个最大的奇异值及其对应的奇异向量,秩为 k 情况下矩阵 $\boldsymbol{Z}_k$ 的最佳近似满足以下形式:

$$\boldsymbol{Z}_k = \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}} = \boldsymbol{U}_k \boldsymbol{\Sigma}_k \boldsymbol{V}_k^{\mathrm{T}}。\tag{9}$$

式中:$\boldsymbol{U}_k \in \mathbf{R}^{m\times k}$;$\boldsymbol{\Sigma}_k \in \mathbf{R}^{k\times k}$;$\boldsymbol{V}_k \in \mathbf{R}^{d\times k}$。降维后注意力输出矩阵的维度从 $m \times d$ 降低至 $m \times k$。

引入 SVD 后,计算过程包括计算 $\boldsymbol{Z}$ 的截断 $\boldsymbol{S}$ 及后续处理降维后的矩阵,最终复杂度为 $O(md^2 + m^2d + mdk + m^2k)$。由于 $k \ll d$,复杂度显著降低。SVD 方法的降维流程如图 3 所示。

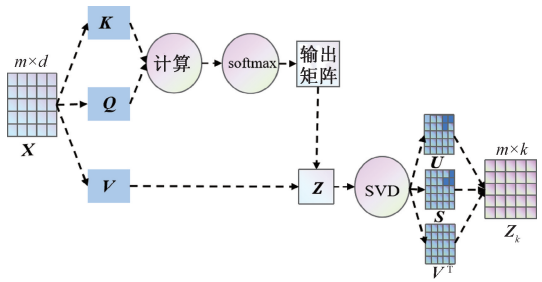

图 3　SVD 方法的降维流程

Figure 3　Dimensionality reduction process using SVD

SVD 通过保留最大奇异值,有效提取注意力矩阵中最重要的信息,减少无关或冗余数据的干扰,并降低了内存需求。对于时序数据而言,通常表现出周期性或趋势性,这些特征在奇异值分解中对应于较大的奇异值及其关联的奇异向量。通过截断 SVD,去除噪声和冗余信息,从而增强模型对核心特征的捕捉能力,提升预测的准确性和鲁棒性。

2　实验设置

2.1　数据集

本文实验在以下 8 个公开基准数据集上进行。其中,ETT[22]记录了电力变压器中与油温相关的运行参数,包含来自 2 个独立变压器共 7 个特征;Traffic[22]记录了美国加州交通部门高速公路网络中 862 个传感器测得的每小时道路占用率;Electricity[23]记录了 321 名用户每小时的用电量(单位:kW·h),涵盖 2012 年至 2014 年的用电记录;Exchange-Rate[24]收集了 1990 年至 2016 年 8 个不同国家每日相对于美元的汇率;Weather[22]包含了 2020 年德国气象站记录的 21 项气象指标。

在数据预处理阶段,首先,对每个数据集的时序特征应用 Z-score 归一化处理,公式表示为

$$x' = \frac{x - \mu_x}{\sigma_x}。\tag{10}$$

式中:μ_x 和 σ_x 分别基于训练集的均值和标准差计算。其次,针对缺失值问题,采用线性插值法进行填补[25],该方法通过相邻观测值的加权平均估算缺失点,有效维护时序的连续性特征。最后,对于具有明

显周期特性的数据集,例如 ETT 和 Weather,通过引入时间戳特征,例如小时序数和星期序数,以增强模型对时序模式的捕捉能力。

数据集被划分为训练集、验证集和测试集,按照7:2:1的比例分割。输入序列长度统一固定为 96 时间步,预测长度设置为多尺度任务{96, 192, 336, 720},覆盖短期至长期预测场景。

2.2 评估指标

为了全面评估模型在时序预测中的表现、复杂度和实时性,本文采用 5 个指标来评估模型性能,包括参数量、均方误差(*MSE*)、平均绝对误差(*MAE*)、计算量以及推理速度。其中,参数量指模型中所有参数的总数,用于反映模型的大小和存储需求。参数量越小,表示模型越轻量化,存储需求也越低。*MSE* 和 *MAE*[26]数值越小表示模型性能越好,*MSE* 和 *MAE* 公式如下:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2; \tag{11}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|。 \tag{12}$$

式中:y_i 为实际值;$\hat{y}_i$ 为预测值;n 为样本数量。计算量衡量一个网络模型的计算复杂度。推理速度指模型每次迭代处理一个批次数据所需的时间。

2.3 实验环境

本文全部实验均在一个高性能虚拟化计算平台上完成。该平台配备了 NVIDIA 虚拟 GPU 技术。具体的硬件配置包括 4 个虚拟 CPU 核心、16 GB 内存和 1 块 NVIDIA VGPU50-12 虚拟显卡。在软件环境方面,本文采用 Python 3.8 作为编程语言,并基于 PyTorch 2.0.0 框架开发。

模型训练采用 Adam 优化器,其学习率为 0.001,批量大小为 256。所有模型均训练 100 个 *epoch*,并在每个 *epoch* 结束后在验证集上评估性能,保存验证损失最小的模型用于最终测试。为充分保证实验的可复现性,本文固定了 PyTorch、NumPy 和 Python 随机数生成器的种子,并对所有超参数进行严格控制和记录。

3 结果分析

在时序数据预测领域,Informer 模型[22]、Reformer 模型[27]和 Flashformer 模型[28]在减轻注意力负担方面具有典型性。其中,Informer 模型通过引入稀疏注意力和蒸馏技术将计算复杂度从二次方降低到接近线性;Reformer 模型通过局部敏感哈希机制使自注意力机制仅关注序列中最相关部分从而降低计算复杂度;Flashformer 模型通过引入门控注意力单元(GAU)来提高处理长序列数据的效率。

为评估 ILformer 模型中注意力机制改进带来的效果,本文选取以上 3 个模型作为 baseline。同时,为进一步验证 ILformer 模型的有效性,还将其与 iTransformer 和 Transformer 模型进行对比实验。

3.1 时效性分析

为了验证 ILformer 模型在推理过程中的优势,本文在 ETT 数据集上对比了不同算法在不同预测长度下的推理表现,实验结果见表 1。从实验结果可知,不同模型的增长速度存在显著差异,ILformer 的推理时间在不同预测长度下表现出平稳的增长趋势,从 96 步预测时的 11.2 ms 增加到 720 步预测时的 42.0 ms,总增幅约为 274.1%。相比之下,Informer 和 Transformer 在较长预测长度下的推理时间显著增加,特别是在 720 步预测时,推理时间超过 200 ms,Transformer 增幅为 387.8%,表现出较高的计算延迟。Flashformer 在 720 步预测时推理时间达到 355.2 ms,表明其对预测长度变化具有高敏感性。在预测长度为 96 步时,ILformer 相比原模型 iTransformer 推理时间缩短 30.0%。总体而言,ILformer 在各预测长度下推理时间平均减少约 30 ms,减少比例超过 40.46%。这些结果表明,ILformer 在推理过程中的高效性和对预测长度变化的鲁棒性,使其在实时性要求较高的任务中具有显著优势。

表 1 不同算法的推理时间对比

Table 1 Inference time comparison of different algorithms

模型	推理时间/ms			
	96	192	336	720
iTransformer	16.0	45.2	72.6	97.1
Informer	67.4	88.6	120.1	205.4
Transformer	66.4	93.8	143.1	324.0
Reformer	35.9	77.8	156.6	291.0
Flashformer	98.3	156.3	230.7	355.2
ILformer	11.2	32.8	38.0	42.0

3.2 预测误差分析

本文对比了 ILformer、iTransformer 及经典轻量化模型 Reformer 在 8 个数据集上的评估结果,选取 *MSE* 和 *MAE* 作为评价指标,结果如表 2 所示,实验结果中,对 *MSE* 和 *MAE* 的最优值进行了下划线标注。

根据表 2 实验结果可知,其与 iTransformer 模型、Reformer 模型相比,ILformer 在多个数据集上的预测精度表现更加优秀,取得了最好的效果。具体而言,在 ETTh2 和 ETTm2 电力能源数据集上,ILformer 模型在大多数预测步长中均取得了最佳结

表 2　不同算法的 *MSE*/*MAE* 对比

Table 2　*MSE*/*MAE* comparison of different algorithms

数据集	指标	ILformer				iTransformer				Reformer			
		96	192	336	720	96	192	336	720	96	192	336	720
ETTh1	*MSE*	0.430	0.487	0.543	0.698	0.440	0.480	0.487	0.503	0.698	0.710	0.827	0.866
	MAE	0.448	0.487	0.523	0.616	0.454	0.465	0.477	0.491	0.653	0.675	0.692	0.749
ETTh2	*MSE*	0.186	0.223	0.260	0.324	0.196	0.380	0.428	0.427	0.358	0.533	0.722	1.609
	MAE	0.292	0.323	0.352	0.397	0.299	0.400	0.432	0.445	0.451	0.476	0.639	1.017
ETTm1	*MSE*	0.385	0.438	0.509	0.591	0.370	0.377	0.426	0.491	0.369	0.472	0.538	0.655
	MAE	0.408	0.437	0.475	0.531	0.400	0.422	0.425	0.459	0.402	0.578	0.642	0.792
ETTm2	*MSE*	0.122	0.155	0.192	0.247	0.123	0.250	0.311	0.412	0.173	0.249	0.511	1.310
	MAE	0.236	0.269	0.299	0.339	0.236	0.264	0.309	0.407	0.307	0.355	0.495	0.823
Traffic	*MSE*	0.482	0.518	0.626	0.688	0.443	0.489	0.522	0.690	0.732	0.733	0.742	0.755
	MAE	0.329	0.354	0.373	0.429	0.302	0.314	0.359	0.413	0.423	0.420	0.420	0.432
Weather	*MSE*	0.182	0.231	0.287	0.361	0.173	0.233	0.278	0.368	0.689	0.752	0.639	1.130
	MAE	0.223	0.262	0.355	0.351	0.214	0.245	0.296	0.324	0.596	0.638	0.698	0.792
Electricit-y	*MSE*	0.181	0.198	0.235	0.267	0.163	0.189	0.234	0.291	0.157	0.236	0.534	0.641
	MAE	0.264	0.323	0.365	0.396	0.252	0.350	0.373	0.394	0.277	0.378	0.651	0.739
Exchang-e	*MSE*	0.085	0.156	0.247	0.579	0.087	0.177	0.241	0.823	0.098	0.192	0.378	0.723
	MAE	0.205	0.254	0.373	0.571	0.209	0.285	0.374	0.698	0.235	0.389	0.567	0.853

果,相较于原模型 iTransformer,*MSE* 分别下降 30.1%和 13.5%,*MAE* 分别下降 34.7%和 6.0%;同样地,在 Exchange-e 汇率数据集上,ILformer 的各项评价指标在大多数步长中领先。无论是 *MSE* 还是 *MAE*,ILformer 的性能均优于经典轻量化模型 Reformer,进一步验证了所提模型的有效性。

然而,在 Traffic 交通流量数据集上,ILformer 的表现相对一般。主要原因在于 Traffic 数据集具有高度的非平稳性和较强的噪声特性,交通流量易受多种不确定因素的影响,如突发事件、天气变化等。在这些因素共同作用下,数据呈现出显著的随机波动和短期变化。这种复杂的时序特征增加了模型在捕捉短期依赖关系时的难度。总体而言,ILformer 在大多数数据集上表现优异,证明其在时序预测任务中的性能优势。

为验证性能差异的统计显著性,本文采用配对 t 检验(显著性水平 $\alpha = 0.05$)对 ILformer 与 baseline 模型(iTransformer、Reformer 及 Flashformer)的 *MSE*/*MAE* 进行 10 次重复实验分析,统计显著性检验结果如表 3 所示。由表 3 可知,ILformer 与 iTransformer 在 96 步预测的 *MSE* 差异的 p 值为 0.003 2(< 0.01),具有高度统计显著性。所有比较的 p 值均小于 0.05,表明性能改进不是随机波动所致。

表 3　统计显著性检验结果

Table 3　Statistical significance test results

对比模型	步长	p 值	
		MSE	*MAE*
iTransformer	96	0.003 2	0.004 1
Reformer	192	0.012 0	0.008 3
Flashformer	336	0.001 1	0.002 4

3.3　参数量与计算复杂度分析

本文还对比了不同算法的参数量和计算量,结果如表 4 所示。由表 4 可知,ILformer 模型在参数量和计算量方面表现出显著优势。Transformer、Informer 和 Flashformer 等模型的参数量较大,导致在训练和推理过程中占用更多内存,增加了硬件资源需求和计算负担。而 ILformer 的参数量仅为 1.36×10^6,是对比模型中最小的,为原模型的 1/5 左右,大幅降低了内存占用。ILformer 的计算量,远低于其他模型,体现了其轻量化设计的有效性。

表 4　不同算法的参数量和计算量对比

Table 4　Parameter counts and operations comparison of different algorithms

模型	参数量/10^6	计算量/(10^6 FLOPs)
iTransformer	6.40	57
Informer	11.37	7 860
Transformer	10.59	572
Reformer	5.82	254
Flashformer	10.59	436
ILformer	1.36	25

实验结果表明,ILformer 不仅在存储效率上显著优化,还极大地降低了计算复杂度,为高效的时序预测任务提供了更优的解决方案。

3.4　消融实验

为验证 ILformer 的结构有效性,本文设计了相

应的消融实验。表5展示了从ILformer中移除两个关键方法后得到的实验结果,其中ILformer W/O SVD指未对注意的输入进行矩阵分解降维的模型;ILformer W/O LA指采用普通多头注意力机制替代线性注意力机制的模型。所有模型均在ETT数据集上使用相同的超参数进行评估。从实验结果可知,随着时间步长的增加,所有模型的预测误差(包括*MSE*和*MAE*)普遍上升,表明长时间预测更具挑战性。在不同时间步长下,ILformer的预测误差始终较低,尤其在短时间步长下表现尤为显著,验证了线性注意力机制和降维方法在提升模型预测精度方面的协同作用。去除线性注意力机制(ILformer W/O LA)时,推理时间略有增加,同时预测精度有所下降。这表明线性注意力机制不仅提升了预测精度,还显著加快了推理速度。在去除矩阵分解降维(ILformer W/O SVD)后,尽管推理速度有所提升,但预测精度略有下降,进一步说明降维方法对模型精度的积极影响。根据实验结果,ILformer和ILformer W/O LA在大多数时间步长下的预测误差低于不采用降维的模型,证明降维不会影响预测精度。线性注意力机制与降维方法的结合,既能保持较高的计算效率,又能不影响预测精度。总之,消融实验结果充分证明了线性注意力机制和降维方法对ILformer模型性能的关键贡献。这些改进不仅提升了模型的预测精度和推理效率,还为未来的模型优化和实际应用提供了重要的参考依据。

表5 消融实验

Table 5 Ablation experiment

步长	ILformer			ILformer W/O SVD			ILformer W/O LA		
	MSE	*MAE*	推理时间/ms	*MSE*	*MAE*	推理时间/ms	*MSE*	*MAE*	推理时间/ms
96	0.430	0.448	11.2	0.436	0.452	10.9	0.431	0.449	12.0
192	0.487	0.487	32.8	0.501	0.496	31.8	0.487	0.488	34.7
336	0.543	0.523	38.0	0.554	0.531	36.9	0.544	0.524	39.8
720	0.698	0.616	42.0	0.721	0.627	40.2	0.703	0.619	44.1

4 结论

本文设计并实现了一种轻量级时序预测模型ILformer,解决了Transformer类时序模型难以平衡计算效率和预测精度的问题。通过线性化注意力机制的优化,ILformer推理速度提高了40.46%,满足了时序预测对实时性的要求。ILformer的分解降维处理有效减少了模型参数量,参数量减少了78.75%,大幅度降低了存储需求,使模型更加轻量化。在线性化注意力和降维优化的共同作用下,ILformer的计算量降至原模型的一半,有效减轻了计算负担。此外,ILformer使用奇异值分解进行降维,通过对信号能量的重新分布,能够精准地从数据中蒸馏出序列的长期趋势与周期性核心模式。但是,ILformer在处理高波动时序时表现相对一般,主要因交通流量易受突发事件和天气干扰,呈现非平稳性。同时,模型对短序列敏感,因降维会损失细节特征,例如消融实验所示,SVD移除时*MSE*略有上升。此外,ILformer依赖全局注意力机制,在极长序列(大于1 000步)下计算效率仍受限;ILformer在720步预测时推理时间达42.0 ms,虽优于baseline,但对实时性要求较高场景,例如毫秒级金融交易,仍需提升。

参考文献:

[1] MENDIS K, WICKRAMASINGHE M, MARASINGHE P. Multivariate time series forecasting: a review[C]// Proceedings of the 2024 2nd Asia Conference on Computer Vision, Image Processing and Pattern Recognition. New York: ACM, 2024: 1-9.

[2] 梁宏涛, 刘硕, 杜军威, 等. 深度学习应用于时序预测研究综述[J]. 计算机科学与探索, 2023, 17(6): 1285-1300.
LIANG H T, LIU S, DU J W, et al. Review of deep learning applied to time series prediction[J]. Journal of Frontiers of Computer Science & Technology, 2023, 17(6): 1285-1300.

[3] ZAREMBA W, SUTSKEVER I, VINYALS O. Recurrent neural network regularization[EB/OL]. (2014-09-08)[2025-08-13]. https://doi.org/10.48550/arXiv.1409.2329.

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

[5] HE Z S, ZHANG X, LI M, et al. A novel solar radiation forecasting model based on time series imaging and bidirectional long short-term memory network[J]. Energy Science & Engineering, 2024, 12(11): 4876-4893.

[6] HARISH NAYAK G H, ALAM M W, AVINASH G, et al. Transformer-based deep learning architecture for time se-

ries forecasting[J]. Software Impacts,2024,22: 100716.

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

[8] NIE Y Q, NGUYEN N H, SINTHONG P, et al. A time series is worth 64 words: long-term forecasting with transformers[EB/OL]. (2022-11-27)[2025-08-13]. https://doi.org/10.48550/arXiv.2211.14730.

[9] LIU Y, HU T G, ZHANG H R, et al. iTransformer: inverted transformers are effective for time series forecasting [EB/OL]. (2023-10-10)[2025-08-13]. https://doi.org/10.48550/arXiv.2310.06625.

[10] 姜晓勇, 李忠义, 黄朗月, 等. 神经网络剪枝技术研究综述[J]. 应用科学学报, 2022, 40(5): 838-849.
JIANG X Y, LI Z Y, HUANG L Y, et al. Review of neural network pruning techniques[J]. Journal of Applied Sciences, 2022, 40(5): 838-849.

[11] 高杨, 曹仰杰, 段鹏松. 神经网络模型轻量化方法综述[J]. 计算机科学, 2024, 51(增刊 1): 11-21.
GAO Y, CAO Y J, DUAN P S. Lightweighting methods for neural network models: a review[J]. Computer Science, 2024, 51(S1): 11-21.

[12] 王改华, 李柯鸿, 龙潜, 等. 基于知识蒸馏的轻量化 Transformer 目标检测[J]. 系统仿真学报, 2024, 36(11): 2517-2527.
WANG G H, LI K H, LONG Q, et al. Object detection of lightweight Transformer based on knowledge distillation [J]. Journal of System Simulation, 2024, 36(11): 2517-2527.

[13] SAHA R, SRIVASTAVA V, PILANCI M. Matrix compression via randomized low rank and low precision factorization[EB/OL]. (2023-10-17)[2025-08-13]. https://doi.org/10.48550/arXiv.2310.11028.

[14] YANG X H, LIU W F, LIU W, et al. A survey on canonical correlation analysis[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 33(6): 2349-2368.

[15] HAN K, WANG Y H, CHEN H T, et al. A survey on vision transformer[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1): 87-110.

[16] 孟祥福, 石皓源. 基于 Transformer 模型的时序数据预测方法综述[J]. 计算机科学与探索, 2025, 19(1): 45-64.
MENG X F, SHI H Y. Survey of Transformer-based model for time series forecasting[J]. Journal of Frontiers of Computer Science & Technology, 2025, 19(1): 45-64.

[17] 连家诚. 面向长序列 Transformer 的计算简化[D]. 合肥: 中国科学技术大学, 2023.
LIAN J C. Computational simplification for long sequence Transformer[D]. Hefei: University of Science and Technology of China, 2023.

[18] ANTONELLI M, LAGO U D, DAVOLI D, et al. An arithmetic theory for the poly-time random functions[EB/OL]. (2023-01-27)[2025-08-13]. https://doi.org/10.48550/arXiv.2301.12028.

[19] TAY Y, DEHGHANI M, BAHRI D, et al. Efficient transformers: a survey[J]. ACM Computing Surveys, 2022, 55(6): 1-28.

[20] CHEN B D, DAO T, WINSOR E, et al. Scatterbrain: unifying sparse and low-rank attention approximation[EB/OL]. (2021-10-28)[2025-08-13]. https://doi.org/10.48550/arXiv.2110.15343.

[21] BOYAPATI M, AYGUN R. Semanformer: semantics-aware embedding dimensionality reduction using transformer-based models[C]//2024 IEEE 18th International Conference on Semantic Computing. Piscataway: IEEE, 2024: 134-141.

[22] ZHOU H Y, ZHANG S H, PENG J Q, et al. Informer: beyond efficient transformer for long sequence time-series forecasting[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(12): 11106-11115.

[23] 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.

[24] QIN Y, SONG D J, CHENG H F, 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. New York: ACM, 2017: 2627-2633.

[25] NI Z L, YU H, LIU S Z, et al. BasisFormer: attention-based time series forecasting with learnable and interpretable basis[EB/OL]. (2023-10-31)[2025-08-13]. https://doi.org/10.48550/arXiv.2310.20496.

[26] KONG X J, CHEN Z H, LIU W Y, et al. Deep learning for time series forecasting: a survey[J]. International Journal of Machine Learning and Cybernetics, 2025, 16(7): 5079-5112.

[27] KITAEV N, KALEV L, LEVSKAYA A, et al. Reformer: the efficient transformer[EB/OL]. (2020-01-13)[2025-08-13]. https://doi.org/10.48550/arXiv.2001.04451.

[28] DAO T, FU D Y, ERMON S, et al. Flashattention: fast and memory-efficient exact attention with io-awareness [C]//36th Conference on Neural Information Processing Systems. Cambridge:MIT,2022:16344-16359.

(下转第 26 页)

Surrogate-assisted Multi-population Differential Evolution Algorithm Based on Region Decomposition

YU Mingyuan, PAN Wanli, LIANG Jing, YUE Caitong

(School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou 450001, China)

Abstract: In expensive optimization problems, if the optimal solution of the problem was not unique, such problems were referred to as expensive multimodal optimization problems. However, it was extremely difficult to obtain multiple optimal solutions with limited computational resources. Moreover, existing surrogate-assisted evolutionary algorithms paid less attention to multimodal attributes. In view of this, a surrogate-assisted multi-population differential evolution algorithm based on region decomposition was proposed to solve expensive multimodal optimization problems. Firstly, in the population individual initialization stage, the correlation between inter-individual distances and objective values was used to detect potential sub-regions, and sub-populations were divided to explore multiple optimal solutions. Secondly, in the early stage of evolution, the differential evolution algorithm was used to perform global search in each sub-population to capture multiple optimal solutions. After multiple optimal individuals were obtained in the early stage of evolution, the covariance matrix adaptive evolution strategy was adopted to carry out local search on them to improve the quality of optimal solutions. In addition, an infilling criterion was proposed, which could adaptively select appropriate individuals for real evaluation according to specific parameters to improve the accuracy and generalization ability of the surrogate model. Finally, the proposed algorithm was compared with seven other algorithms on 20 test functions. The results showed that the proposed algorithm achieved optimal performance on 13 functions with the *PR* metric, and was slightly inferior to the comparison algorithms on at most 5 functions. Overall, the proposed algorithm exhibited excellent performance in solving expensive multimodal optimization problems.

Keywords: expensive multimodal optimization; differential evolution; local search; surrogate-assisted evolution algorithm

(上接第15页)

Lightweight Time Series Forecasting Model Based on iTransformer

ZHOU Qinglei[1], WANG Yujing[2], DUAN Pengsong[2], WANG Chao[2], ZHENG Yongli[2]

(1. School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou 450001, China; 2. School of Cyber Sincence and Engineering, Zhengzhou University, Zhengzhou 450002, China)

Abstract: To address the challenge of balancing prediction accuracy and efficiency in time series forecasting, in this paper a lightweight time series forecasting model named ILformer was proposed, built upon the iTransformer architecture. As a representative variable-based model for temporal data, iTransformer effectively captured complex inter-variable dependencies. However, it was constrained by high computational complexity and a substantial parameter footprint, limiting its practical deployment in resource-constrained scenarios. To mitigate these limitations, ILformer incorporated the following enhancements: the model first introduces a Linear Attention mechanism to replace the traditional attention mechanism, allowing for more flexible input processing. By leveraging linear projection and dimension rearrangement, ILformer significantly reduced the number of parameters while better adapting to varying input shapes and structures. It achieved high computational efficiency, particularly when handling large-scale datasets, and drastically lowered the computational complexity of the attention module without compromising model accuracy. Furthermore, singular value decomposition (SVD) was incorporated into the attention mechanism to achieve matrix dimensionality reduction. This approach substantially decreased the number of matrix multiplications and additions, improving computational efficiency and mitigating the risk of overfitting. Experimental results on eight diverse datasets demonstrated that ILformer achieved a 40.46% improvement in inference speed on average while maintaining the same level of accuracy. Additionally, the number of parameters was reduced by 78.75%, and the operations were halved, underscoring its superior performance and practical applicability.

Keywords: time series forecasting; lightweight; singular value decomposition; linear attention mechanism