

CBT-AD:CNN-BiLSTM-Transformer混合架构的时序异常检测模型

许建, 陈仕杰, 冯健聪, 杨庚

引用本文

许建, 陈仕杰, 冯健聪, 杨庚. [CBT-AD:CNN-BiLSTM-Transformer混合架构的时序异常检测模型](#)[J]. 计算机科学, 2026, 53(7): 230-241.

XU Jian, CHEN Shijie, FENG Jiancong, YANG Geng. [CBT-AD:CNN-BiLSTM-Transformer Hybrid Model for Time Series Anomaly Detection](#) [J]. Computer Science, 2026, 53(7): 230-241.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[基于分阶段训练策略与多尺度特征融合的目标检测方法](#)

Object Detection Method Based on Phased Training Strategy and Multi-scale Feature Fusion

计算机科学, 2026, 53(6A): 250700088-7. <https://doi.org/10.11896/jsjcx.250700088>

[基于双域注意力和特征融合的航拍图像目标检测模型](#)

Aerial Image Object Detection Model Based on Dual-domain Attention and Feature Fusion

计算机科学, 2026, 53(6A): 250600036-7. <https://doi.org/10.11896/jsjcx.250600036>

[聚焦边界和多尺度特征融合的脑卒中病灶分割](#)

Boundary-focused Multi-scale Feature Fusion Network for Stroke Lesion Segmentation

计算机科学, 2026, 53(2): 264-272. <https://doi.org/10.11896/jsjcx.250300137>

[基于时频域注意力的时间序列异常检测模型](#)

Time-Frequency Attention Based Model for Time Series Anomaly Detection

计算机科学, 2026, 53(2): 161-169. <https://doi.org/10.11896/jsjcx.241200106>

[基于生成式数据增强与Faster-RCNN改进的发动机打刻面缺陷检测](#)

Defect Detection of Engine Engraved Surface Based on Generative Data Augmentation and Improved Faster-RCNN

计算机科学, 2025, 52(11A): 241200025-7. <https://doi.org/10.11896/jsjcx.241200025>

CBT-AD: CNN-BiLSTM-Transformer 混合架构的时序异常检测模型

许建^{1,2} 陈仕杰¹ 冯健聪¹ 杨庚^{1,2}

1 南京邮电大学计算机学院 南京 210023

2 大数据安全与智能处理省高校重点实验室(南京邮电大学) 南京 210023

摘要 现有基于深度学习的时序数据异常检测方法在局部与全局特征协同建模能力以及长序列计算效率方面存在不足。为此,提出一种基于混合架构的异常检测模型。首先,在多尺度特征提取中,采用深度可分离卷积捕获局部细节特征,结合动态池化增强突发异常响应能力。其次,设计双向时序建模机制,通过 BiLSTM 融合双向上下文特征,并采用门控 Dropout 抑制长序列训练过程的过拟合风险。进一步,设计分层稀疏全局注意力模块,利用局部窗口注意力捕捉时频特征,借助多头机制与残差连接优化梯度传播稳定性。最后,提出了一种动态特征融合方法,并结合分块处理框架实现检测精度与计算效率的协同优化。在 4 个公开时序数据集上的实验结果表明,所提模型的各项性能指标较现有方法提升显著,且展现出较好的鲁棒性与泛化能力。

关键词: 时序异常检测;混合神经网络;多尺度特征融合;分层稀疏注意力

中图分类号 TP183

CBT-AD: CNN-BiLSTM-Transformer Hybrid Model for Time Series Anomaly Detection

XU Jian^{1,2}, CHEN Shijie¹, FENG Jiancong¹ and YANG Geng^{1,2}

1 School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China

2 Jiangsu Key Laboratory of Big Data Security and Intelligent Processing(Nanjing University of Posts and Telecommunications), Nanjing 210023, China

Abstract Existing deep learning-based time series anomaly detection methods exhibit limitations in both the collaborative modeling of local and global features, as well as in computational efficiency for long sequences. To address these issues, this paper proposes a hybrid-architecture anomaly detection model. Firstly, for multi-scale feature extraction, depthwise separable convolution is employed to capture local detail features, combined with adaptive pooling to enhance the response capability to burst anomalies. Secondly, a bidirectional temporal modeling mechanism is designed, which integrates bidirectional context features through BiLSTM and utilizes gated Dropout to mitigate the risk of overfitting during long-sequence training. Furthermore, a hierarchical sparse global attention module is designed, leveraging local window attention to capture time-frequency features, while employing multi-head mechanisms and residual connections to optimize gradient propagation stability. Finally, a dynamic feature fusion method is proposed and integrated with a chunk-based processing framework to achieve collaborative optimization of detection accuracy and computational efficiency. Experimental results on four public time-series datasets demonstrate that the proposed model achieves significant improvements across various performance metrics compared to existing methods, while also exhibiting strong robustness and generalization capability.

Keywords Time series anomaly detection, Hybrid neural network, Multi-scale feature fusion, Layered sparse attention

1 引言

时序数据广泛存在于工业监测、金融风控、智能交通等场景中,其动态变化反映了系统状态的演化特征^[1]。时间序列异常检测(Time Series Anomaly Detection, TSAD)旨在识别序列中偏离常规模式的数据点或片段,是保障系统安全与智能决策的关键技术之一^[2]。然而,实际应用中的时序数据通常呈现出非平稳、高维、异常稀缺等特点,给异常检测带来显著挑战。

传统统计和机器学习方法在处理时序数据时,依赖固定假设与手工特征,难以有效建模时序数据中的非线性、长依赖与复杂时空关联^[3]。例如,自回归(AutoRegressive, AR)与移动平均(Moving Average, MA)模型在多维高动态场景下易产生建模误差;孤立森林、支持向量机等方法对长程依赖和局部突变的感知能力有限,限制了模型的检测性能与泛化能力。随着数据复杂性不断提升,传统方法在检测精度、实时响应与可扩展性方面难以满足需求,推动了深度学习在 TSAD 领域的广泛应用。卷积神经网络(Convolutional Neural Network,

到稿日期:2025-06-12 返修日期:2025-10-08

基金项目:国家自然科学基金(62372244)

This work was supported by the National Natural Science Foundation of China(62372244).

通信作者:许建(xuj@njupt.edu.cn)

CNN)、递归神经网络(Recurrent Neural Network, RNN)以及基于自注意力机制的 Transformer 等架构因具有强大的特征学习能力而被广泛应用于时序数据建模^[4]。

然而,这些单一模型在处理复杂时序异常时,各自存在难以克服的局限性,CNN 虽然在提取局部特征方面表现优异,但其有限的感受野难以捕捉长周期信号的全局依赖关系^[5-6]。例如,Jin 等^[7]的研究表明,单一 CNN 对于涉及微观特征与宏观模式复合类型的异常往往存在识别盲区,导致检测性能低下。可见,在单一模型中实现局部与全局动态协同建模是提升检测性能的关键。RNN 虽然在时序依赖建模方面表现优异,但在长序列场景下面临记忆衰减和顺序计算带来的效率瓶颈^[8]。尽管双向长短期记忆网络(Bidirectional Long Short-Term Memory, BiLSTM)通过门控机制缓解了梯度传播问题,但文献[9]显示,当序列超过 1000 步时,模型对早期信息的记忆保持率低于 30%。此外,Liu 等^[10]的实验表明,标准长短期记忆网络(Long Short-Term Memory, LSTM)在 2000 步序列长度下的推理延迟达 120 ms,远超工业物联网设备 50 ms 的实时响应阈值。可见,实现局部与全局特征的动态交互与高效并行化计算,是提升检测性能的另一关键。

为解决上述问题,研究者尝试将多种单一模型融合起来,例如 Zhuang 等^[11]提出的门控并行 CNN-Transformer 模型结合了一维卷积的瞬态扰动感知与注意力机制的多维特征融合能力来捕捉跟踪参数耦合关系,显著增强了多源信号的跨场景泛化性能。该混合架构虽部分缓解了局部与全局信息割裂的问题,但密集自注意力层带来的计算复杂度急剧上升,严重制约了其在长序列上的检测效率。Wang 等^[12]提出的 CNN-BiLSTM 模型通过双向门控单元强化了时序建模能力,但该混合架构参数数量激增近 3 倍,训练耗时增加 40%。Mu 等^[13]的理论分析指出,当序列长度 $N > 500$ 时,模型复杂度将呈 $O(N^2)$ 量级增长。而基于 Transformer 的混合模型在处理原始时序数据时,常因忽略局部波动细节而陷入模型深度与计算成本的权衡困境^[14]。

可见,单一模型难以同时兼顾高精度的局部细节感知、鲁棒的中长程时序依赖建模以及高效的全局模式捕获能力。因此,本文提出将 CNN、BiLSTM 和 Transformer 进行深度融合,构建三级协同架构。CNN 作为局部特征提取器,利用卷积归纳偏置精准捕获短时异常模式;BiLSTM 作为中程时序建模器,通过门控机制和双向结构有效处理上下文依赖的中程动态,克服 CNN 的感受野限制;改进的 Transformer 作为全局编码器,采用稀疏自注意力高效捕捉长周期复杂关联,解决 BiLSTM 在超长序列上的记忆衰减和计算效率问题。这种“局部-中程-全局”设计旨在通过多尺度特征协同建模来提升异常检测精度。但简单堆叠组件面临传统 Transformer 的复杂度 $O(L^2)$ 在长序列上不可行的问题,且现有混合方法还存在多尺度融合缺陷,静态特征融合机制难以适应异常模式的动态演化规律,粗粒度序列分段虽提升了并行效率,却割裂了时序连续性。Pan 等^[15]的实验表明,普通点状注意力难以有效提取时序的局部与全局特征,更无法实现多尺度特征融合。

针对上述问题,本文提出一种混合神经网络架构(CNN-BiLSTM-Transformer for Anomaly Detection, CBT-AD),通过

多尺度特征融合机制实现检测精度与计算效率的协同优化。本文的创新性贡献主要体现在以下方面:

1)提出 CNN-BiLSTM-Transformer 三级架构,解决了多尺度特征割裂问题。针对单一模型难以协同建模局部、中程与全局特征的问题,通过深度可分离卷积、双向门控单元和稀疏注意力层分别捕捉局部、中程与全局特征,实现多尺度协同建模,有效提升对多维时序异常特征的识别精度。

2)引入时域分块与局部窗口注意力机制,解决了长序列 Transformer 计算效率问题。将长序列分割为可并行处理的子序列,通过跨块注意力保持全局关联,构建高效计算框架,显著降低复杂度并提升模型训练效率。

3)设计动态门控特征融合通道,解决了静态融合机制适应性差的问题。通过可微分门控机制自适应调节局部、全局特征的贡献,增强模型对复杂数据分布的适应性和鲁棒性。

4)在公开数据集上进行实验验证,证明了所提出的模型相较于现有方法在性能上的提升,并通过消融实验揭示了多模块协同机制的优势,在噪声注入与长序列场景下验证了模型鲁棒性。

2 相关工作

时序异常检测是保障系统安全与智能决策的关键技术。随着深度学习的发展,神经网络方法逐步取代传统手段,在建模复杂时序数据的动态性、多尺度特征与长依赖方面展现出优势^[16]。然而,单一模型难以兼顾局部与全局信息,在复杂异常检测中存在局限^[17]。因此,亟需系统梳理现有方法,探索组合模型以实现多模态特征的协同建模。

2.1 基于单模型的时序异常检测方法

基于单模型的时序异常检测方法可按照异常判断标准分为 5 类:密度估计、聚类、预测、重构和对比学习^[18]。本节系统梳理各类方法的代表性工作,并分析其优缺点。

基于密度估计的方法通过计算数据密度分布情况来检测异常样本。Moore 等^[19]的 conDENSE 利用条件密度估计替代传统的重构误差,通过掩码自回归流动态建模时间序列条件分布,实现了对异常的精准识别。Goswami 等^[20]则通过无监督方式评估不同密度估计器在时序异常检测中的性能差异,提升检测稳定性。但此类方法假设时间步独立,难以建模时序依赖,面对动态变化或长期依赖时检测性能较弱。

基于聚类的方法通过计算数据间的相似性或距离识别异常。Schmidl 等^[21]提出的 AutoTSAD 结合模糊 C-均值聚类与加权动态时间规整,在低维隐空间中通过簇距判断异常,有效兼顾了精度与效率。Ghojogh 等^[22]基于高斯混合模型和 MAR(Masked AutoRegressive)框架,引入概率聚类,从而实现在无需先验分布假设的情况下,对高维时序数据进行鲁棒的异常检测。但聚类方法对簇数与距离度量敏感,性能依赖于超参数的精确选取,且聚类本质上仍视每个时间点为独立样本,忽略了序列内部的时序连续性和上下文信息,易在局部异常下误检或漏检。

基于预测的方法通过学习时序动态模式预测未来值,并以预测误差判断异常。例如,Zhou 等^[23]的 Informer 利用改进的稀疏自注意力机制显著提升了长序列预测效率,但在捕捉局部突发异常方面存在灵敏度不足的问题。Xu 等^[24]的

AnomalyTransformer 则设计了关联差异度量机制,通过注意力放大异常与正常的差异,增强对复杂依赖的感知。但此类方法依赖数据分布平稳性,难以应对概念漂移,且计算开销大,不适用于实时或资源受限场景。

基于重构的方法利用自动编码器或生成对抗网络构建低维表示,并通过重构误差识别异常。Tuli 等^[25]的 TranAD 模型在多变量时序数据上结合对抗训练和自调节机制实现较高的重构精度,但对时序依赖刻画有限。Yue 等^[26]的 TS2Vec 通过分层 Transformer 提取多尺度特征,尽管在特征提取上具备优势,但计算复杂度高,难以满足工业实时性要求。但此类方法通常依赖于强大的生成模型(如 GAN 或扩散模型),训练不稳定且过度关注重构误差,容易忽略序列全局结构,难以识别细微异常。

近年来的对比学习方法中,Yang 等^[27]的 DCdetector 设计了双重注意力对比损失,通过对正常与异常样本对进行多尺度编码对比,显著降低了误报率并提升了检测精度。Darban 等^[28]提出的 CARLA 则通过异常注入生成负样本,并结合时序邻域信息进行对比表示学习,有效解决了负样本稀缺带来的表征塌缩问题。Lee 等^[29]的 SoftCLT 方法则在对比学习中引入软分配机制,在样本不平衡的场景下自适应地平衡正负样本,从而提升模型整体的鲁棒性。但该方法对正负样本构造极为敏感,不当采样易导致过拟合,削弱对未知异常的泛化能力。

2.2 基于组合模型的时序异常检测方法

为突破单一模型的性能瓶颈,近年来的研究集中于融合不同架构的优势,构建具有多尺度特征提取能力的组合模型来实现局部特征与全局依赖的协同建模,以及时序依赖与并行化计算的平衡优化。

在实现局部特征与全局依赖的协同建模方面,Zhao 等^[30]构建改进的 VMD-XGBoost-BiLSTM 模型。该模型融合自适应赋权、注意力机制和优化算法,使得光伏预测误差较单一模型下降超 20%。Ning 等^[31]的 MCFMAE 利用多尺度卷积和自注意力捕捉不同时间特征和上下文信息,并引入

记忆模块识别正常模式。Liu 等^[32]的 EdgeConvFormer 通过融合动态图卷积与 Transformer 结构,综合提取局部和全局时空信息,实现了复杂时序的分解式建模。这些方法均在多个数据集上展现出优越性能,表明多尺度卷积、图结构等混合模块能有效增强模型对不同尺度和层次特征的感知能力,从而更好地识别短时突发异常和长周期变化。

在时序依赖与并行化计算的平衡优化方面,Zhang 等^[33]的 BiLSTM-Attention 采用 BiLSTM 网络来学习时间序列数据的特征和固有规律,并集成了注意力机制,为不同的时间步动态分配权重,突出重要区间并减少低有效点的影响,从而提高预测准确性,但仍存在计算开销大的问题。Zhou 等^[34]的 ProbSparse 自注意力和蒸馏机制,通过稀疏化注意力计算与层次化特征压缩,在保持预测精度的同时压缩了长序列计算量,有效解决了传统 Transformer 在超长序列建模中的内存与效率问题。Bai 等^[35]的 PAFormer 引入并行窗口与全局注意力模块,分别建模局部模式与全局上下文,通过并行的窗口注意力和全局注意力分支,同时捕获局部模式与长程依赖,有效提升了检测性能并将检测时间缩短 30%。Liu 等^[36]的 EdgeConvFormer 并行融合图卷积与 Transformer,兼顾多尺度特征和高效计算,实现多尺度建模与高效检测。

总而言之,现有方法多基于时间上下文依赖提升表示能力,但往往忽略了依赖差异性,且混合架构常用静态拼接或简单加权,难以应对复杂动态变化。为此,本文提出 CBT-AD 模型,通过构建深度可分离卷积-双向门控单元-分层稀疏注意力的三级协同架构,创新性地引入局部窗口注意力与时域分块并行机制,并设计动态门控权重函数,在保持多尺度优势的同时,有效解决长依赖建模、计算效率和动态特征融合等核心问题。

3 模型架构

CBT-AD 模型通过多尺度特征融合实现端到端时序异常检测。整体框架如图 1 所示,模型包含局部特征提取、时序建模、全局编码、动态特征融合以及分类器 5 个部分。

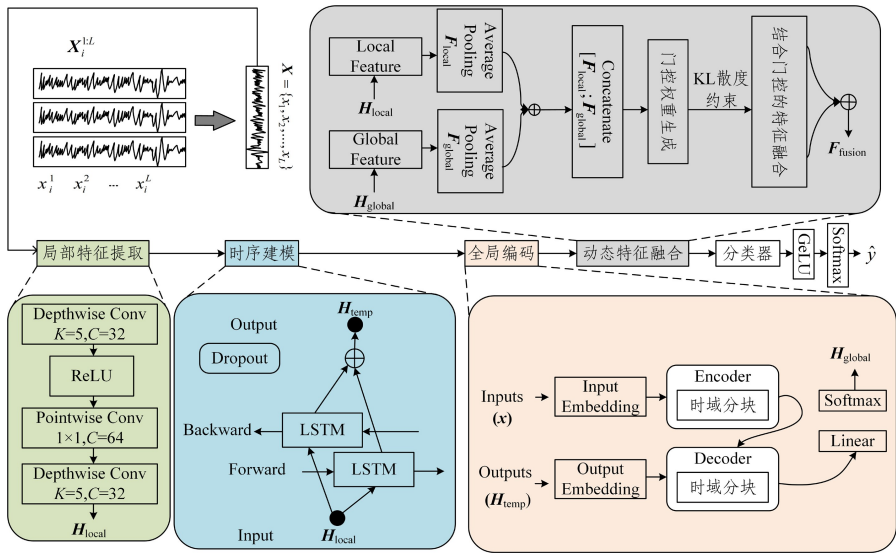


图 1 CBT-AD 整体架构

Fig. 1 Overall framework of CBT-AD

1)局部特征提取:采用深度可分离卷积捕获局部细节特征,结合动态池化增强对突发异常的响应能力。2)时序建模:通过双向 LSTM 融合正反向上下文依赖,并引入门控 Dropout 抑制长序列训练中的过拟合风险。3)全局编码:基于局部窗口注意力捕捉时频特征,借助多头机制与残差连接优化梯度传播稳定性。4)动态特征融合:设计可微分门控函数自适应调节局部与全局特征的权重比例,结合 KL 散度约束防止门控值过早收敛。5)分类器:通过层次化非线性映射(GE-LU 激活)与类别平衡损失函数,实现高鲁棒性异常判定。

3.1 问题描述

将时序数据表示为 $\mathbf{X}_i^{1:L} = \{x_i^1, x_i^2, \dots, x_i^L\}$, 其中 $i \in \{1, 2, \dots, N\}$ 为样本索引, N 为样本总数, L 为序列长度。每个时间步的观测值 $x_i^l \in \mathbb{R}^M$ 或 $\mathbb{R}^1$ 表示第 i 个样本在时刻 t 的多维或单维时序数据, M 为多维序列的特征维度。本文旨在设计一种混合神经网络架构 $f_\theta: \mathbb{R}^{B \times M \times L} \rightarrow \mathbb{R}^{B \times D \times L}$ (其中 B 为批次大小, D 为隐空间维度), 通过时域分块策略将输入序列分割为 $K = \lfloor \frac{L}{S} \rfloor$ 个连续子块 $\{\mathbf{X}_i^{(1)}, \mathbf{X}_i^{(2)}, \dots, \mathbf{X}_i^{(K)}\}$, 并将每个子块 $\mathbf{X}_i^{(K)}$ 映射为高阶特征表示 $\mathbf{Z}^{(k)} \in \mathbb{R}^{D \times LK}$, 使其能够通过局部窗口注意力捕获短时瞬态波动, 通过双向门控循环单元建模中程时序依赖, 并通过跨块稀疏注意力捕捉全局长周期关联, 最终通过动态门控机制自适应融合多尺度特征, 实现复杂时序异常模式的精准检测。

3.2 局部特征提取

局部特征提取模块采用深度可分离卷积 (Depthwise Separable Convolution, DSC) 来提取输入序列的局部时序特征。与标准卷积相比, DSC 将一个卷积操作分解为深度卷积和逐点卷积两步, 从而显著降低计算量。具体地, 首先对每个输入通道独立应用一维深度卷积。为覆盖典型突发异常的短时波动周期, 同时避免过大卷积核引入噪声, 并在保证特征多样性的前提下控制计算量, 将深度卷积核大小设为 $K=5$ 、输出通道为 $C=32$, 并接 ReLU 激活函数增强非线性表达能力。其数学形式如式(1)所示:

$$\mathbf{Z}_{\text{dw}}^{(c)} = \text{ReLU}(\mathbf{W}_{\text{dw}}^{(c)} * \mathbf{X}_{\text{in}}^{(c)} * \mathbf{b}_{\text{dw}}^{(c)}) \quad (1)$$

其中, $\mathbf{Z}_{\text{dw}}^{(c)}$ 为第 c 个通道的深度卷积输出, $\mathbf{X}_{\text{in}}^{(c)}$ 为输入第 c 个通道的特征序列, 它们的维度均为 $\mathbb{R}^{L \times 1}$; $\mathbf{W}_{\text{dw}}^{(c)}$ 为深度卷积核的权重, 维度为 $\mathbb{R}^{K \times 1}$, $\mathbf{b}_{\text{dw}}^{(c)} \in \mathbb{R}$ 是深度卷积偏置项。逐点卷积通过 1×1 卷积融合跨通道信息。为了平衡信息密度与模型复杂度, 将输出特征维度压缩至 $d=64$ 。其数学表达如式(2)所示:

$$\mathbf{H}_{\text{local}} = \sum_{c=1}^C (\mathbf{W}_{\text{pw}}^{(c)} \cdot \mathbf{Z}_{\text{dw}}^{(c)}) + \mathbf{b}_{\text{pw}} \quad (2)$$

其中, $\mathbf{H}_{\text{local}}$ 是维度为 $\mathbb{R}^{L \times 64}$ 的局部特征输出, $\mathbf{W}_{\text{pw}}^{(c)}$ 是维度为 $\mathbb{R}^{1 \times 32}$ 的逐点卷积核的权重, $\mathbf{b}_{\text{pw}}$ 是逐点卷积偏置量。通过以上两步分解步骤, 深度可分离卷积相较于普通卷积减少约 79.70% 的参数数量, 从而显著降低了计算开销, 并保留了对局部时序模式的有效提取能力。

3.3 时序建模

时序建模模块通过前向和后向两个 LSTM 单元分别捕捉时间序列中的正向和反向依赖关系。其数学表达如式(3)一

式(5)所示:

$$\tilde{\mathbf{h}}_t = \text{LSTM}_{\text{forward}}(\mathbf{H}_{\text{local}}^{(t)}, \tilde{\mathbf{h}}_{t-1}) \quad (3)$$

$$\tilde{\mathbf{h}}_t = \text{LSTM}_{\text{backward}}(\mathbf{H}_{\text{local}}^{(t)}, \tilde{\mathbf{h}}_{t+1}) \quad (4)$$

$$\mathbf{H}_{\text{temp}} = [\tilde{\mathbf{h}}_t; \tilde{\mathbf{h}}_t] \in \mathbb{R}^{128} \quad (5)$$

式(3)和式(4)分别为前向传播和反向传播的计算式, 其中 $\tilde{\mathbf{h}}_t \in \mathbb{R}^{64}$ 表示前向 LSTM 在时间步 t 的隐藏状态, $\mathbf{H}_{\text{local}}^{(t)} \in \mathbb{R}^{64}$ 表示局部特征提取模块在时间步 t 的输出; $\tilde{\mathbf{h}}_{t-1} \in \mathbb{R}^{64}$ 表示前向 LSTM 在时间步 $t-1$ 的隐藏状态; $\tilde{\mathbf{h}}_t \in \mathbb{R}^{64}$ 表示反向 LSTM 在时间步 t 的隐藏状态, $\tilde{\mathbf{h}}_{t+1} \in \mathbb{R}^{64}$ 表示前向 LSTM 在时间步 $t+1$ 的隐藏状态。反向 LSTM 的计算过程与前向 LSTM 类似, 区别在于其时间步顺序为从终点到起点。前向和反向 LSTM 的输出经拼接融合, 生成双向依赖特征, 如式(5)所示。 $\mathbf{H}_{\text{temp}}$ 为 BiLSTM 的输出, 同时包含过去与未来上下文信息, 其值等于前向和后向 LSTM 在时间步 t 的隐藏状态的向量拼接。

同时, 为防止过拟合, 在 BiLSTM 的输出层引入 Dropout 机制, 以 0.2 的概率随机屏蔽神经元, 从而增强模型的泛化能力。BiLSTM 通过结合两端时序信息, 能够同时利用历史与未来上下文信息, 有效增强了对时序动态模式的建模能力。

3.4 全局编码

全局编码模块基于改进的稀疏自注意力机制来捕获长距离依赖关系, 并通过局部注意力窗口降低计算复杂度。输入为时序建模模块输出的双向特征 $\mathbf{H}_{\text{temp}} \in \mathbb{R}^{B \times L \times 128}$, 输出为局部注意力编码后的特征 $\mathbf{H}_{\text{attn}} \in \mathbb{R}^{B \times L \times 128}$ 。针对传统 Transformer 在长序列中因全局自注意力机制导致的计算复杂度高的问题, 本文提出了一种基于时域分块 (Temporal Blocking, TB) 与稀疏局部窗口注意力 (Sparse Local Window Attention, SLWA) 的高效并行化框架。

首先将输入序列 $\mathbf{H}_{\text{temp}} \in \mathbb{R}^{B \times L \times 128}$ 分割为 $K = \lfloor \frac{L}{S} \rfloor$ 个连续子块 $\{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \dots, \mathbf{X}^{(K)}\}$, 实现时域分块处理。分割公式如式(6)所示:

$$\mathbf{X}^{(k)} = \mathbf{X}_{kB+1:(k+1)B}, k=0, 1, \dots, K-1 \quad (6)$$

分块后, 各子块可独立输入至稀疏自注意力模块进行局部特征建模。块长 S 的取值等于批次大小 B , 确保分块后无信息丢失。子块数 K 由 $K = \lfloor \frac{L}{S} \rfloor$ 自动确定, 确保分块后无信息丢失。将序列分割为 K 个连续子块后, 对于第 k 个子块 $\mathbf{X}^{(k)}$, 其查询矩阵 $\mathbf{Q}^{(k)} \in \mathbb{R}^{B \times d}$ 、键矩阵 $\mathbf{K}^{(k)} \in \mathbb{R}^{B \times d}$ 与值矩阵 $\mathbf{V}^{(k)} \in \mathbb{R}^{B \times d}$ 的注意力计算范围限制在窗口半径 R 内, 计算式如式(7)所示:

$$\text{LocalAttention}(\mathbf{Q}^{(k)}, \mathbf{K}^{(k)}, \mathbf{V}^{(k)}) = \text{softmax}\left(\frac{\mathbf{Q}^{(k)} \mathbf{K}_{[i-R:i+R]}^{(k)\top}}{\sqrt{d}}\right) \mathbf{V}_{[i-R:i+R]}^{(k)} \quad (7)$$

其中, i 为当前查询位置; R 为窗口半径, 为遵循覆盖完整局部模式的设定原则, 设置 $R=512$; d 为投影维度。此设计将单块复杂度从 $O(B^2 d)$ 降至 $O(BRd)$, 显著提升了计算效率。

为维持全局依赖关系, 进一步建模跨块依赖, 本文提出跨块注意力连接 (Cross-Block Attention, CBA), 允许相邻子块

间的特征交互,如图 2 所示。

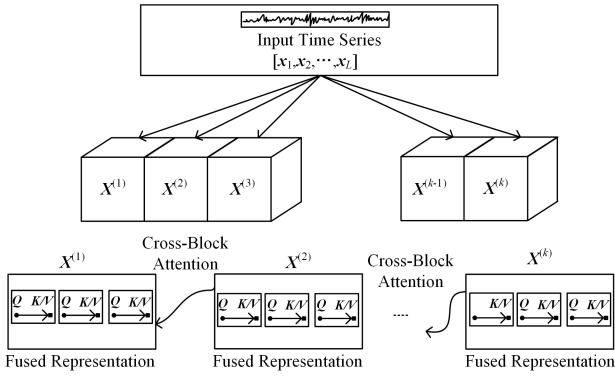


图 2 时域分块与跨块注意力机制示意图

Fig. 2 Schematic diagram of the time-domain chunking and cross-block attention mechanisms

第 k 个子块的查询矩阵 $\mathbf{Q}^{(k)}$ 不仅关注本块内窗口,还关注前一子块的键值矩阵 $\mathbf{K}^{(k-1)}$ 和 $\mathbf{V}^{(k-1)}$, 计算式如式(8)所示:

$$\text{CrossBlockAttention}(\mathbf{Q}^{(k)}, \mathbf{K}^{(k-1)}, \mathbf{V}^{(k-1)}) = \text{Softmax}\left(\frac{\mathbf{Q}^{(k)} \mathbf{K}^{(k-1)T}}{\sqrt{d}}\right) \mathbf{V}^{(k-1)} \quad (8)$$

其中, $\mathbf{K}^{(k-1)}, \mathbf{V}^{(k-1)} \in \mathbb{R}^{B \times d}$ 用于存储前一子块的上下文信息; d 为投影维度,用于控制注意力权重的分布稳定性。将跨块注意力输出与本块局部注意力结果相加,形成最终特征表示,计算式如式(9)所示:

$$\mathbf{H}^{(k)} = \text{LocalAttention}(\mathbf{Q}^{(k)}, \mathbf{K}^{(k)}, \mathbf{V}^{(k)}) + \gamma \cdot \text{CrossBlockAttention}(\mathbf{Q}^{(k)}, \mathbf{K}^{(k-1)}, \mathbf{V}^{(k-1)}) \quad (9)$$

其中, γ 为可学习的跨块权重系数,用于动态调整相邻块间的信息贡献比例。经过以上步骤,本文方法的总复杂度为块内局部注意力机制和跨块注意力两部分的复杂度之和,其中块内局部注意力的复杂度为 $O(KBRd)$,跨块注意力的复杂度为 $O(KB^2d)$,因此改进后的总复杂度为 $O(KBd(R+B))$ 。当 $R \ll B$ 且 $B \ll L$ 时,复杂度近似为 $O(L(R+B)d)$,而传统 Transformer 的全局自注意力复杂度为 $O(L^2d)$,因此,改进后的复杂度较传统 Transformer 显著降低。

稀疏局部窗口注意力机制通过限制每个位置仅关注局部邻域内的上下文,相较于传统 Transformer 的自注意力机制计算复杂度由 $O(L^2)$ 降至 $O(LRd)$ 。核心公式如式(10)所示:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}_{[i-R:i+R]} \mathbf{K}_{[i-R:i+R]}^T}{\sqrt{d}}\right) \mathbf{V}_{[i-R:i+R]} \quad (10)$$

其中, $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{L \times d}$ 分别由 $\mathbf{H}_{\text{temp}}$ 通过线性投影得到(投影矩阵 $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{128 \times 64}$), $\mathbf{Q}_{[i-R:i+R]} \in \mathbb{R}^{(2R+1) \times d}$ 表示查询矩阵中位置 i 的局部窗口,半径 $R=512$; $d=128$ 为特征维度缩放因子,用于稳定梯度计算。通过选择邻近窗口,局部注意力能够聚焦关键邻域信息,同时避免无关噪声干扰。

多头注意力机制将 $\mathbf{H}_{\text{attn}} \in \mathbb{R}^{B \times L \times 128}$ 分割为 4 个头(每头维度 $d_h=32$),融合后输出 $\mathbf{H}_{\text{multi}} \in \mathbb{R}^{B \times L \times 128}$ 。在保持总计算量的同时提升信息密度。其核心公式如式(11)、式(12)所示:

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) \mathbf{W}^O \quad (11)$$

$$\text{head}_i = \text{Attention}(\mathbf{Q} \mathbf{W}_i^Q, \mathbf{K} \mathbf{W}_i^K, \mathbf{V} \mathbf{W}_i^V) \quad (12)$$

其中, $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V \in \mathbb{R}^{64 \times 16}$ 为第 i 个头的投影矩阵, $\mathbf{W}^O \in \mathbb{R}^{64 \times 128}$ 为输出融合矩阵。每个注意力头独立学习不同子空间的依赖关系(如高频波动、低频趋势),增强模型对复杂模式的表征能力。最终通过残差连接和层归一化融合时序特征 $\mathbf{H}_{\text{temp}}$ 和多头注意力的输出 $\mathbf{H}_{\text{multi}}$, 得到全局特征 $\mathbf{H}_{\text{global}} \in \mathbb{R}^{B \times L \times 128}$, 计算式如式(13)所示:

$$\mathbf{H}_{\text{global}} = \text{LayerNorm}(\mathbf{H}_{\text{temp}} + \mathbf{H}_{\text{multi}}) \quad (13)$$

层归一化公式如式(14)所示:

$$\text{LayerNorm}(x) = \gamma \odot \frac{x - \mu}{\sigma + \epsilon} + \beta \quad (14)$$

其中, $\mu, \sigma \in \mathbb{R}$ 分别为特征的均值和标准差; $\gamma, \beta \in \mathbb{R}^{128}$ 为可学习的缩放与偏移参数; $\epsilon=1\text{E-}6$ 为数值稳定性常数。该设计通过局部聚焦与全局融合平衡计算效率与建模能力,缓解梯度消失并抑制噪声干扰。

本文提出的高效并行化框架通过时域分块与跨块注意力机制,在保证全局依赖建模能力的同时降低计算复杂度,为长序列异常检测提供了理论支撑与技术实现路径。

3.5 动态特征融合

动态特征融合模块通过门控机制自适应地融合局部与全局特征,实现了局部特征与全局特征的高效融合,解决了静态融合在复杂异常时性能低下的问题。本模块包含双路径输入、门控权重生成、特征融合及正则化约束 4 个核心部分。

双路径输入分为局部路径输入和全局路径输入。局部路径的输入是局部特征 $\mathbf{H}_{\text{local}}$, 经平均池化降维后得到 $\mathbf{F}_{\text{local}} \in \mathbb{R}^{B \times 64}$; 全局路径的输入是全局特征 $\mathbf{H}_{\text{global}}$, 经平均池化降维后得到 $\mathbf{F}_{\text{global}} \in \mathbb{R}^{B \times 128}$ 。平均池化降维的公式如式(15)、式(16)所示:

$$\mathbf{F}_{\text{local}} = \frac{1}{L} \sum_{t=1}^L \mathbf{H}_{\text{local}}^{(t)} \quad (15)$$

$$\mathbf{F}_{\text{global}} = \frac{1}{L} \sum_{t=1}^L \mathbf{H}_{\text{global}}^{(t)} \quad (16)$$

平均池化操作沿时间维度取均值,将特征映射至样本级表示,保留关键统计信息。本模块设计平均池化降维的原因是原始特征 $\mathbf{H}_{\text{local}}$ 和 $\mathbf{H}_{\text{global}}$ 包含细粒度时序信息,但将维度为 64 和 128 的它们直接融合会让维度达到 192,导致维度爆炸,所以需要平均池化通过聚合时间步信息来提取高阶统计特征。

门控权重生成的输入是经过平均池化后的局部特征 $\mathbf{F}_{\text{local}}$ 和全局特征 $\mathbf{F}_{\text{global}}$, 通过控制局部与全局特征的融合比例,得到动态门控权重 $g \in [0, 1]^B$, 计算式如式(17)所示:

$$g = \sigma(\mathbf{W}_g \cdot \text{GELU}([\mathbf{F}_{\text{local}}; \mathbf{F}_{\text{global}}]) + \mathbf{b}_g) \quad (17)$$

其中, $\mathbf{W}_g \in \mathbb{R}^{192 \times 1}$, $\mathbf{b}_g \in \mathbb{R}$ 为可学习参数; GELU 函数用于将池化后的局部特征和全局特征拼接起来,并增强非线性表达能力; σ 为 sigmoid 函数,将输出压缩至 $[0, 1]$ 。这样的设计可以使得门控权重根据输入特征动态生成,且 GELU 激活函数相比 ReLU 具有更平滑的梯度特性,从而能缓解门控值的饱和问题。

在得到局部特征、全局特征、门控权重这 3 个值后,特征融合模块让这 3 个输入经过计算后得到输出的融合特征 $\mathbf{F}_{\text{fusion}}, \mathbf{F}_{\text{fusion}} \in \mathbb{R}^{B \times 128}$, 计算式如式(18)所示:

$$\mathbf{F}_{\text{fusion}} = \mathbf{g} \odot (\mathbf{W}_{\text{local}} \mathbf{F}_{\text{local}}) + (1 - \mathbf{g}) \odot (\mathbf{W}_{\text{global}} \mathbf{F}_{\text{global}}) \quad (18)$$

其中, $\mathbf{W}_{\text{local}} \in \mathbb{R}^{64 \times 128}$ 和 $\mathbf{W}_{\text{global}} \in \mathbb{R}^{128 \times 128}$ 为特征投影矩阵, 通过线性投影 $\mathbf{W}_{\text{global}}$ 和 $\mathbf{F}_{\text{global}}$, 将局部与全局特征映射至同一维度空间(128 维), 确保融合操作的可加性。利用广播机制, 使门控权重 $\mathbf{g}$ 与特征维度对齐, 并执行逐元素乘法。

为防止门控权重 $\mathbf{g}$ 过早收敛至极端值(0 或 1), 导致特征融合僵化, 引入了 KL 散度正则化项。正则化损失计算式如式(19)所示:

$$\mathcal{L}_{\text{reg}} = \lambda \cdot \frac{1}{B} \sum_{i=1}^B \left(\mathbf{g}_i \log \frac{\mathbf{g}_i}{0.5} + (1 - \mathbf{g}_i) \log \frac{1 - \mathbf{g}_i}{0.5} \right) \quad (19)$$

其中, 为避免因正则化项过小而导致门控权重两极分化以及因过大而抑制特征选择灵活性, 设置正则化强度系数 $\lambda = 0.1$ 。假设门控权重服从均匀分布, 鼓励特征多样性, 从而保证融合过程的稳定性。通过上述设计, 动态特征融合模块有效解决了局部与全局特征割裂问题, 为复杂时序异常检测任务提供了鲁棒的特征表示基础。

3.6 分类器与损失函数

分类模块采用两层全连接网络实现层次化特征映射, 将融合特征向量 $\mathbf{F}_{\text{fusion}}$ 作为输入, 第一层通过非线性变换计算隐层特征, 计算式如式(20)所示:

$$\mathbf{Z} = \text{GELU}(\mathbf{W}_1 \mathbf{F}_{\text{fusion}} + \mathbf{b}_1) \quad (20)$$

其中, $\mathbf{W}_1 \in \mathbb{R}^{128 \times 64}$ 为权重矩阵, $\mathbf{b}_1 \in \mathbb{R}^{64}$ 为偏置项。GELU 激活函数能在增强模型非线性表达能力的同时保持梯度稳定性。第二层通过 Softmax 归一化生成类别概率预测, 计算式如式(21)所示:

$$\hat{\mathbf{y}} = \text{Softmax}(\mathbf{W}_2 \mathbf{Z} + \mathbf{b}_2) \quad (21)$$

其中, $\mathbf{W}_2 \in \mathbb{R}^{C \times 128}$ 将隐层特征映射至 C 个目标类别。

损失函数采用加权交叉熵来缓解类别不平衡问题。模型总损失函数由加权交叉熵分类损失 $\mathcal{L}_{\text{cls}}$ 、KL 散度正则化损失 $\mathcal{L}_{\text{reg}}$ 、特征一致性对齐损失 $\mathcal{L}_{\text{con}}$ 三部分构成。针对类别不平衡问题, 定义类别权重 w_c 与总样本数 N_c 成反比, 计算式如式(22)所示:

$$w_c = \frac{N_{\text{total}}}{C \cdot N_c} \quad (22)$$

加权交叉熵损失计算式如式(23)所示, 该设计通过放大少数类样本的梯度贡献来减小数据偏斜导致的模型偏差。

$$\mathcal{L}_{\text{cls}}(\theta) = - \sum_{c=1}^C w_c y_c \log \hat{y}_c \quad (23)$$

为提升训练稳定性和收敛性能, 还设计了动态权重调度策略。具体地, 采用算法 1 所示的多参数退火调度方法, 动态调整学习率、正则化权重和特征一致性权重。

算法 1 动态权重退火调度算法

输入: 初始学习率 η_0 , 最小学习率 η_{min} , 最大训练轮次 T_{max} , 初始正则化

权重 λ_1^{init} , 衰减周期 T_{decay} , 初始特征一致性权重 λ_2^{init} , 最终特征一致性权重 λ_2^{final}

输出: 当前轮次的学习率 η_t , 正则化权重 $\lambda_1^{(t)}$, 特征一致性权重 $\lambda_2^{(t)}$

- 1. for $t = 1$ to T_{max} do
- 2. $\eta_t \leftarrow \eta_{\text{min}} + 0.5 * (\eta_0 - \eta_{\text{min}}) * (1 + \cos(\pi * t / T_{\text{max}}))$ //学习率退火
- 3. $\lambda_1^{(t)} \leftarrow \lambda_1^{\text{init}} * \exp(-t / T_{\text{decay}})$ //正则化权重退火
- 4. $\lambda_2^{(t)} \leftarrow \lambda_2^{\text{init}} * ((\lambda_2^{\text{final}} - \lambda_2^{\text{init}})t / T_{\text{max}})$ //特征一致性权重递增
- 5. return $\eta_t, \lambda_1^{(t)}, \lambda_2^{(t)}$ //返回当前轮次的参数
- 6. end

其中, 初始学习率 η_0 通过余弦退火平滑衰减至 η_{min} , 使前期快速下降后期逐渐稳定; 正则化权重 $\lambda_1^{(t)}$ 按指数衰减, 使训练初期允许较大多样性, 后期收敛更严格; 特征一致性权重 $\lambda_2^{(t)}$ 线性递增, 使模型先专注分类任务, 后期逐步加强局部和全局特征对齐。

4 实验与结果分析

4.1 数据集与预处理

为全面评估模型性能, 本文采用 4 组权威时序数据集: 网络安全领域的 UNSW-NB15(网络流统计特征)和 CIC-IDS2017(深度包解析), 工业控制领域的 SWaT(水处理厂传感器数据), 以及服务器监控领域的 SMD(服务器集群指标)。各数据集特性如表 1 所列。

表 1 数据集核心属性
Table 1 Core attributes of data set

	UNSW-NB15	CICIDS2017	SWaT	SMD
发布时间	2015	2017	2015	2018
总样本量	2540044	2830743	946722	708420
攻击类别数	9 类(含 Fuzzers, Analysis 等)	7 类(含 Botnet, DDoS 等)	41 类(含泵阀故障, 水质污染等)	29 类(含 CPU 过载, 内存泄漏等)
特征维度	49	80	51	38
正常样本占比/%	87.72	83.26	88.93	92.47
攻击复杂性	单阶段攻击为主	多阶段复合攻击	物理系统攻击	硬件故障、软件异常
典型应用场景	企业内网入侵检测	云环境 APT 攻击检测	工业控制系统检测	服务器集群监控检测

预处理流程包含标准化处理、窗口分割和类别平衡 3 步。首先对时序数据 $\mathbf{X} \in \mathbb{R}^{N \times D}$ 的每个特征维度进行 Z-score 标准化, 消除量纲差异, 其中 N 为序列长度, D 为特征维度。随后, 采用滑动窗口法分割序列, 设置窗口长度 $L = 128$, 步长为 $\lfloor 0.75L \rfloor$ (25% 重叠率), 既保留边界信息又增强时序依赖建模能力。分割后每个窗口的维度为 $\mathbb{R}^{L \times D}$, 总窗口数为 $K = \lceil \frac{N-L}{0.75L} \rceil + 1$, 每个窗口独立处理。滑动窗口分割示意图如图 3 所示。

在对输入数据预处理后, 采用局部特征提取、时序建模、全局编码、动态特征融合 4 个模块对数据特征进行提取。模型通过 PyTorch 框架实现, 并且所有的实验均在单个 NVIDIA GeForce GTX1050Ti 4GB 显卡上运行。

数据集预处理中, 对 UNSW-NB15 数据集执行特征降维(从 49 维减少至 17 维), 采用动态分层下采样缓解类别失衡, 保留攻击时序连续性; CICIDS2017 从 80 维特征降至 30 维, 结合主成分分析与排列测试优化特征维度, 并通过自适应过采样平衡类别分布。SWaT 通过随机森林重要性分析(阈值

大于 0.85)从 51 维特征降至 25 维,SMD 使用主成分分析从 38 维特征降至 20 维。

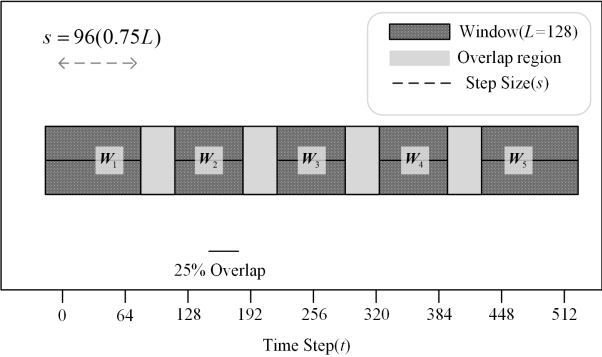


图 3 滑动窗口分割示意图

Fig. 3 Schematic diagram of sliding window segmentation

4.2 评价指标

为了全面评估模型的性能,本文选择了准确率(Accuracy)、精确率(Precision)、召回率(Recall)指标、F1 分数以及 AUC-ROC 作为评价指标。

准确率(Accuracy):模型正确分类的样本占总样本的比例。计算式如式(24)所示:

Accuracy= (TP+TN)/(TP+TN+FN+FP) (24)

精确率(Precision):模型预测为正类的样本中实际为正类的比例。计算式如式(25)所示:

Precision= TP/(TP+FP) (25)

召回率(Recall):实际为正类的样本中被模型正确预测为正类的比例。计算式如式(26)所示:

Recall= TP/(TP+FN) (26)

F1 分数(F1-Score):精确率和召回率的调和平均值,用于平衡两者的表现。计算式如式(27)所示:

F1-Score= (2 * Precision * Recall)/(Precision + Recall) (27)

AUC-ROC(Area Under the Receiver Operating Characteristic Curve):表征模型在所有分类阈值下对正负样本的全局区分能力,其值为 ROC 曲线与坐标轴围成的面积。

以上公式中,真阳性(TP)是指模型正确检测到的攻击样本;假阳性(FP)是指模型认为是攻击样本但其实是良性攻击样本;真阴性(TN)是指模型正确检测到的良性样本;假阴性(FN)是指模型认为是良性样本但其实是攻击样本。

4.3 结果分析

4.3.1 对比实验

为验证 CBT-AD 模型的性能,本文选取 8 类主流时序异常检测模型(包括 CNN, BiLSTM, Transformer 及其组合架构)作为基线方法进行对比。

如表 2 所列, CBT-AD 模型在网络安全、工业控制与服务器监控场景下均展现出显著性能优势。在 UNSW-NB15 和 CICIDS2017 网络安全数据集上, CBT-AD 的 F1 分数分别达到 0.9976 和 0.9956, 较次优模型 CNN-Transformer 提升 1.04 个百分点和 1.6 个百分点;在工业控制场景的 SWaT 数据集上, 面对泵阀故障、水质污染等 41 类物理系统攻击, CBT-AD 的 F1 分数达 0.9634, 较 CNN-Transformer 提升 4.75 个百分点;在服务器监控场景的 SMD 数据集上, 针对 CPU 过载、内存泄漏等 29 类硬件故障, CBT-AD 的 F1 分数达 0.9815, 较 CNN-Transformer 提升 3.58 个百分点。这种跨领域性能优势源于多模态特征协同机制:深度可分离卷积精准提取局部突发特征,尤其在 SWaT 中有效捕捉阀门状态跳变等瞬态物理信号;BiLSTM 建模双向时序依赖,成功解析 SMD 中 CPU 过载→内存泄漏的故障演进逻辑;稀疏注意力捕捉全局模式,显著增强对 CICIDS2017 多阶段 APT 攻击潜伏期行为的识别能力;动态门控机制自适应融合多尺度特征,在复杂工业噪声环境中保持鲁棒性。相比之下,单一模型如 Informer 因缺乏局部特征提取能力,对 SWaT 设备瞬态扰动灵敏度不足;静态融合模型如 BiLSTM-Attention 难以适应 SMD 多源异构故障的动态分布规律。

表 2 本文模型与基线模型的实验结果比较

Table 2 Comparison of experimental results of the proposed model and the baseline models

Datasets	Methods	Accuracy	Precision	Recall	F1-Score	AUC-ROC
UNSW-NB15	CNN	0.9532	0.9387	0.9471	0.9423	0.9754
	LSTM-ED	0.9614	0.9498	0.9546	0.9521	0.9832
	CNN-BiLSTM	0.9687	0.9623	0.9639	0.9628	0.9891
	Informer	0.8825	0.8649	0.8583	0.8614	0.9087
	TranAD	0.8924	0.8765	0.8683	0.8721	0.9157
	TS2Vec	0.9843	0.9786	0.9802	0.9793	0.9961
	BiLSTM-Attention	0.9772	0.9695	0.9738	0.9716	0.9937
	CNN-Transformer	0.9915	0.9862	0.9883	0.9872	0.9982
CICIDS2017	CBT-AD	0.9976	0.9976	0.9976	0.9976	0.9999
	CNN	0.9217	0.9054	0.9008	0.9029	0.9413
	LSTM-ED	0.9296	0.9152	0.9107	0.9128	0.9529
	CNN-BiLSTM	0.9385	0.9263	0.9237	0.9249	0.9621
	Informer	0.8523	0.8376	0.8289	0.8327	0.8854
	TranAD	0.9674	0.9589	0.9543	0.9565	0.9817
	TS2Vec	0.9758	0.9693	0.9675	0.9684	0.9895
	BiLSTM-Attention	0.9712	0.9627	0.9619	0.9612	0.9883
SMD	CNN-Transformer	0.9856	0.9804	0.9789	0.9796	0.9958
	CBT-AD	0.9956	0.9956	0.9956	0.9956	0.9999

(续表)

Datasets	Methods	Accuracy	Precision	Recall	F1-Score	AUC-ROC
SWaT	CNN	0.8623	0.8451	0.8712	0.8579	0.9027
	LSTM-ED	0.8765	0.8572	0.8824	0.8696	0.9115
	CNN-BiLSTM	0.9026	0.8864	0.9087	0.8974	0.9289
	Informer	0.8512	0.8326	0.8613	0.8467	0.8923
	TranAD	0.8923	0.8746	0.8987	0.8864	0.9196
	TS2Vec	0.8732	0.8558	0.8794	0.8674	0.9087
	BiLSTM-Attention	0.9137	0.8972	0.9186	0.9078	0.9362
	CNN-Transformer	0.9315	0.9153	0.9268	0.9159	0.9428
SMD	CBT-AD	0.9630	0.9592	30.9683	0.9634	0.9811
	CNN	0.9014	0.8932	0.9026	0.8978	0.9321
	LSTM-ED	0.9173	0.9086	0.9142	0.9114	0.9418
	CNN-BiLSTM	0.9357	0.9263	0.9328	0.9295	0.9564
	Informer	0.8887	0.8794	0.8912	0.8852	0.9216
	TranAD	0.9284	0.9192	0.9263	0.9227	0.9492
	TS2Vec	0.9126	0.9038	0.9104	0.9071	0.9387
	BiLSTM-Attention	0.9423	0.9336	0.9398	0.9367	0.9603
SMD	CNN-Transformer	0.9512	0.9428	0.9486	0.9457	0.9651
	CBT-AD	0.9822	0.9800	0.9830	0.9815	0.9910

如图 4 和图 5 所示,在 UNSW-NB15 数据集上,模型仅需 9 个训练周期即稳定收敛,且训练和验证损失同步下降,无过拟合振荡;在 CICIDS2017 数据集上,模型通过动态门控避开局部极小陷阱,验证损失最终收敛于相近量级。

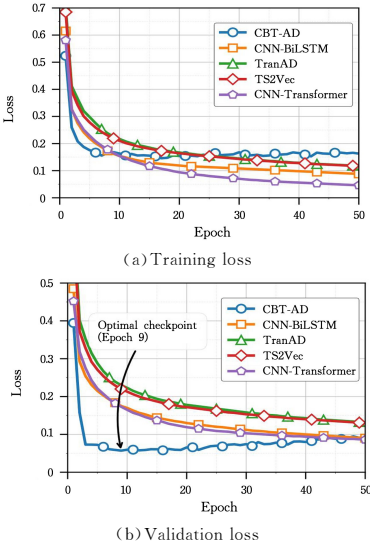


图 4 在 UNSW-NB15 数据集上的损失变化对比
Fig. 4 Comparison of loss changes on UNSW-NB15 dataset

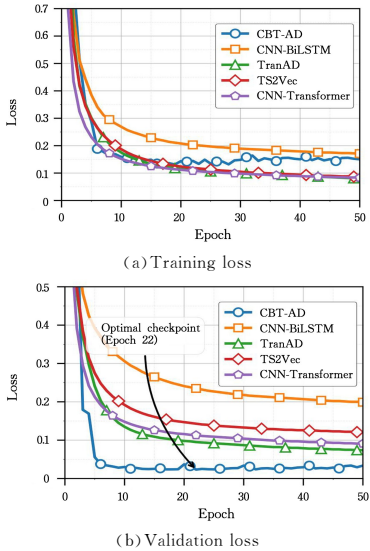


图 5 在 CICIDS2017 数据集上的损失变化对比
Fig. 5 Comparison of loss changes on CICIDS2017 dataset

训练周期内实现稳定收敛,训练损失与验证损失同步平滑下降,展现出对工业传感器噪声环境的高度鲁棒性;在 SMD 数据集上,模型同样快速收敛,尽管在第 8 轮出现轻微梯度波动(源于门控机制对突发负载的自适应权重调整),但验证损失最终迅速稳定于低水平区间,验证了其对于服务器故障链长程依赖关系建模的有效性。

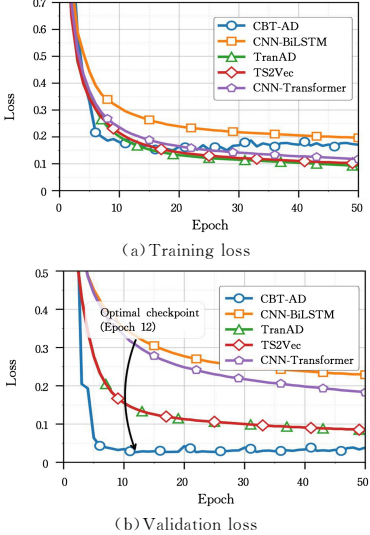


图 6 在 SWaT 数据集上的损失变化对比
Fig. 6 Comparison of loss changes on SWaT dataset

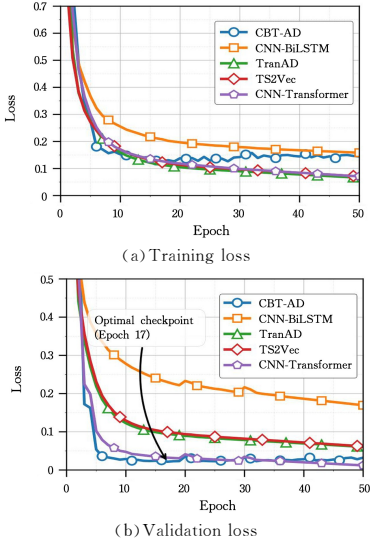


图 7 在 SMD 数据集上的损失变化对比
Fig. 7 Comparison of loss changes on SMD dataset

如图 6 和图 7 所示,在 SWaT 数据集上,模型在 12 个训

4.3.2 消融实验

为了验证各模块的有效性,本节设计消融实验,依次移除局部特征(NoCNN)、时序建模(NoBiLSTM)、全局编码(No-Transformer)及动态融合(NoDynamicFusion),并在数据集上进行实验验证。

如表 3 所列,完整 CBT-AD 模型在网络安全、工业控制与服务器监控 4 类数据集上均展现出显著优势。移除局部特征提取模块后,模型 UNSW-NB15, CICIDS2017, SWaT, SMD 上的 F1 值分别下降 2.61 个百分点、1.03 个百分点、3.86 个

百分点以及 1.33 个百分点,证实了深度可分离卷积对突发异常的捕捉能力具有跨场景普适性,尤其在 SWaT 物理设备突发故障中展现出强瞬态响应能力。移除时序建模模块后,模型在 UNSW-NB15 上的召回率降低 1.63 个百分点,在 CICIDS2017, SWaT, SMD 上的 F1 值分别下降 3.01 个百分点、2.59 个百分点和 3.90 个百分点,体现出双向门控对中程时序依存关系的建模价值,尤其在 SMD 服务器故障链(如 CPU 过载→内存泄漏→服务崩溃)中具有不可替代性。

表 3 模块消融实验结果
Table 3 Module ablation experimental results

Datasets	Methods	Accuracy	Precision	Recall	F1-Score	AUC-ROC
UNSW-NB15	NoCNN	0.9713	0.9727	0.9713	0.9715	0.9971
	NoBiLSTM	0.9713	0.9733	0.9713	0.9715	0.9981
	NoTransformer	0.9467	0.9533	0.9467	0.9495	0.9906
	NoDynamicFusion	0.9401	0.9482	0.9401	0.9407	0.9990
	CBT-AD	0.9976	0.9976	0.9976	0.9976	0.9998
CICIDS2017	NoCNN	0.9843	0.9853	0.9843	0.9853	0.9998
	NoBiLSTM	0.9656	0.9666	0.9656	0.9656	0.9977
	NoTransformer	0.9751	0.9761	0.9751	0.9753	0.9983
	NoDynamicFusion	0.9536	0.9577	0.9542	0.9546	0.9998
	CBT-AD	0.9956	0.9956	0.9956	0.9956	0.9999
SWaT	NoCNN	0.9318	0.9224	0.9357	0.9290	0.9487
	NoBiLSTM	0.9427	0.9372	0.9453	0.9412	0.9625
	NoTransformer	0.9459	0.9418	0.9496	0.9457	0.9683
	NoDynamicFusion	0.9476	0.9443	0.9512	0.9477	0.9701
	CBT-AD	0.9632	0.9586	0.9621	0.9663	0.9742
SMD	NoCNN	0.9621	0.9587	0.9662	0.9624	0.9782
	NoBiLSTM	0.9375	0.9329	0.9418	0.9373	0.9594
	NoTransformer	0.9683	0.9652	0.9721	0.9686	0.9825
	NoDynamicFusion	0.9712	0.9688	0.9743	0.9715	0.9849
	CBT-AD	0.9748	0.9723	0.9786	0.9754	0.9867

全局编码模块缺失对模型性能的影响最显著:在 UNSW-NB15 上 F1 值骤降 4.82 个百分点,AUC-ROC 下降 0.92 个百分点,在 CICIDS2017, SWaT 和 SMD 上 F1 值分别降低 2.03 个百分点、2.13 个百分点和 0.69 个百分点。这表明分层稀疏注意力对复杂协同行为的捕捉效能,在多阶段 APT 攻击潜伏期、水质污染攻击等长周期关联场景中具有关键作用。采用静态融合替代动态门控后,模型在 UNSW-NB15 上的 F1 值下降 5.70 个百分点,在 CICIDS2017 上的召回率降低 4.15 个百分点;在 SWaT 与 SMD 上的召回率分别下降 1.57 个百分点和 0.40 个百分点,揭示了动态门控在多频段异常适应中的核心优势,尤其在工业传感器噪声与服务器硬件抖动环境中展现出强鲁棒性。

消融实验表明,CBT-AD 通过深度卷积提取局部瞬态特征(如 SWaT 阀门状态跳变);BiLSTM 构建双向中程依赖(支撑 SMD 故障演进逻辑);稀疏注意力捕获全局长周期模式(解析 CICIDS2017 攻击潜伏期);动态门控实现多尺度自适应融合。该协同架构有效弥补了传统方法中局部-全局特征割裂、计算效率与建模能力失衡的缺陷,为复杂时序异常检测建立了理论框架与实践基准。

以 UNSW-NB15 数据集为例,图 8 展示了本文模型消融实验中的损失曲线变化,通过损失曲线分析验证模块协同机制。基于梯度收敛轨迹对比发现:移除局部特征提取(CNN)后,在 UNSW-NB15 上的初始验证损失上升 16%,稳态稳定在 0.15 ± 0.02 ,表明局部特征感知能力退化;BiLSTM 缺失使验证损失从 0.4973 跃升至 0.5889,验证了时序建模模块

对于捕获长期依赖的重要性;移除 Transformer 模块导致验证损失终值增长约 83%,并出现验证-训练损失倒置现象,验证了全局注意力对复杂模式识别的必要性;移除动态融合模块导致验证损失在 0.15~0.20 之间周期性波动,揭示了静态融合的泛化缺陷并验证了动态融合模块能够提高模型对复杂数据分布的适应性和鲁棒性。

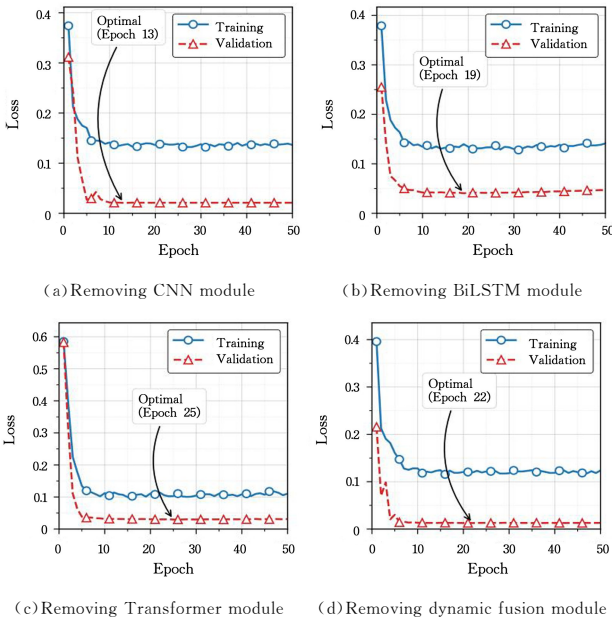


图 8 UNSW-NB15 数据集上消融实验的损失变化
Fig. 8 Loss change of ablation experiment on UNSW-NB15 dataset

4.3.3 超参数实验

为检验设置不同超参数对 CBT-AD 模型训练的性能影响,以 UNSW-NB15 数据集为例,本节设计了滑动窗口大小 L 、分块块长 S 及门控权重 g 的测试实验。其中第 1 个实验固定其他参数不变,只分析滑动窗口大小 L 的变化对模型的影响;第 2 个实验固定其他参数不变,只分析分块块长 S 的变化对模型的影响;第 3 个实验固定其他参数不变,只分析门控权重 g 的变化对模型的影响。

窗口尺寸 L 是影响上下文依赖建模范围的关键参数,对模型性能与计算效率存在显著非线性影响。如图 9 所示,当 $L=128$ 时,模型在 UNSW-NB15 数据集上取得最优 F1 分数 0.9976。当 L 从 64 增大至 256 时,F1 分数呈现先增后降趋势。在 UNSW-NB15 数据集上,窗口尺寸为 128 时 F1 分数达到峰值 0.9976,较窗口 64 的 0.9880 提升 0.96 个百分点;当窗口增至 256 时,F1 分数下降至 0.9948,降幅达 0.28 个百分点。这一现象揭示了过小的窗口会导致局部特征捕捉不充分,过大的窗口则会引入冗余时序噪声并加剧计算负担。实验结果表明, $L=128$ 的窗口设置可有效平衡局部瞬态特征与长周期模式捕获,验证了本文采用 128 作为默认窗口长度的合理性。

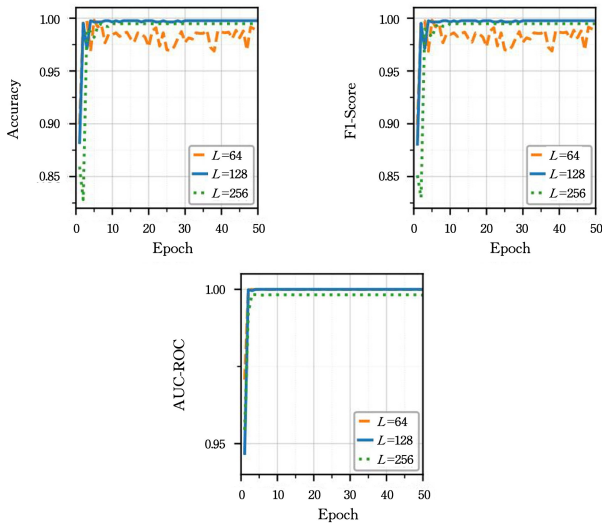


图 9 不同 L 值对模型性能影响对比

Fig. 9 Comparison of the effects of different L values on model performance

分块块长 S 是控制时域分块策略的并行化粒度的重要参数。如图 10 所示,当 $S=64$ 时,在 UNSW-NB15 上的 F1 分数为 0.9921,块内局部注意力覆盖范围不足,难以捕捉跨块的全局关联;增大至 $S=128$ 时,F1 分数提升至 0.9976,块内窗口注意力与跨块全局交互的协同作用增强;继续增大至 $S=256$ 时,F1 分数下降至 0.9853。该现象表明,块内注意力窗口过长会导致局部瞬态特征提取能力退化,而适中的分块粒度(如 128)能平衡局部细节与全局趋势的建模需求。实验结果表明, $S=128$ 的分块设置可有效平衡局部瞬态特征与全局长周期关联的协同建模,验证了本文采用 128 作为默认分块块长的合理性。

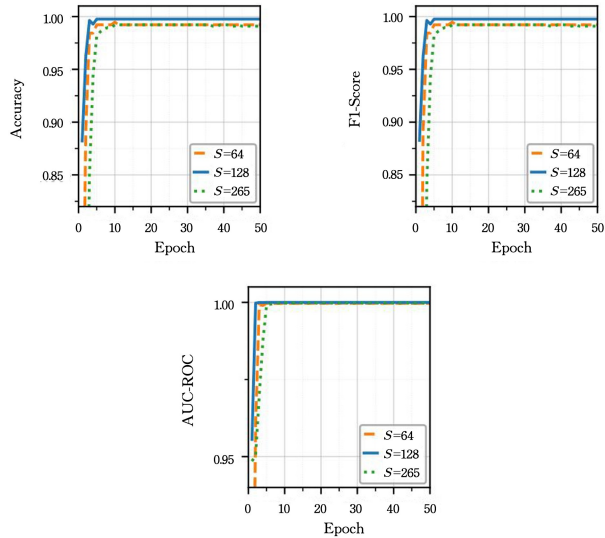


图 10 不同 S 值对模型性能影响对比

Fig. 10 Comparison of the effects of different S values on model performance

动态门控系数 g 调节跨块注意力对特征融合的贡献比例。图 11 表明,当 $g=0.1$ 时,在 UNSW-NB15 上的 F1 分数为 0.9821,跨块注意力贡献过低,全局依赖建模不足;当 $g=0.5$ 时,F1 分数提升至 0.9938,局部与全局特征的动态协同初步显现;当 $g=0.3$ 时,F1 分数达到峰值 0.9976,模型自适应融合局部细节与跨块趋势,检测性能最优。这验证了动态门控机制在多频段异常检测中的必要性。适中的门控权重可有效平衡局部突发与全局长周期特征的互补性。实验结果表明,门控权重 $g=0.3$ 能够有效平衡局部突发特征与全局长周期模式的互补性,验证了本文采用 0.3 作为门控权重的合理性以及动态门控机制在多频段异常检测中的必要性。

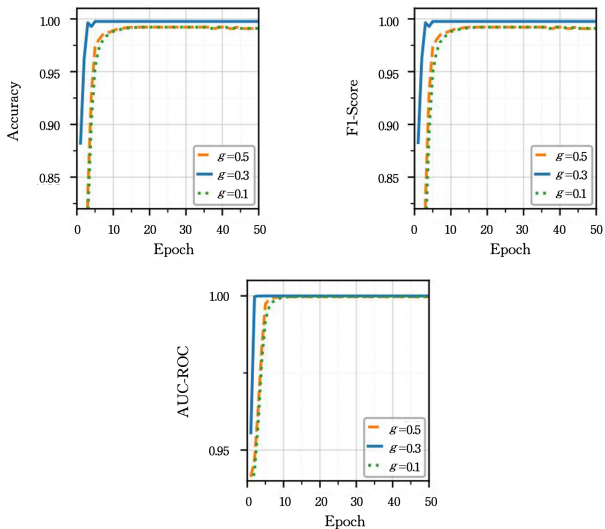


图 11 不同 g 值对模型性能影响对比

Fig. 11 Comparison of the effects of different g values on model performance

4.3.4 模型大小与计算效率分析

为验证 CBT-AD 模型的计算效率优势,本节设计实验对

模型的计算量(FLOPs)、参数量(Params)、推理时间(FPS)以及内存占用(Memory)进行分析。以 UNSW-NB15 数据集为例,从中取一个窗口样本作为输入示例,其维度为[64,128,17],3个取值分别对应批量大小、窗口长度和特征维度。将

输入示例输入模型进行推理并计算出对应的模型计算量、参数量、推理时间,与次优模型 CNN-Transformer 以及 CNN-BiLSTM,BiLSTM-Attention,TranAD,TS2Vec 进行对比,结果如表 4 所列。

表 4 模型大小与计算效率分析
Table 4 Analysis of model sizes and computational efficiency

模型	FLOPs	Params	FPS/ms	Memory/MB
TranAD	$(6.83 \pm 0.21) \times 10^9$	$(3.15 \pm 0.13) \times 10^6$	17.4 ± 0.9	1245 ± 32
TS2Vec	$(5.49 \pm 0.18) \times 10^9$	$(4.22 \pm 0.17) \times 10^6$	16.3 ± 0.4	1587 ± 18
BiLSTM-Attention	$(3.91 \pm 0.15) \times 10^9$	$(2.34 \pm 0.11) \times 10^6$	24.3 ± 0.7	893 ± 24
CNN-BiLSTM	$(3.12 \pm 0.14) \times 10^9$	$(1.63 \pm 0.08) \times 10^6$	35.1 ± 1.5	612 ± 12
CNN-Transformer	$(2.78 \pm 0.12) \times 10^9$	$(1.33 \pm 0.05) \times 10^6$	42.7 ± 0.7	674 ± 15
CBT-AD	$(1.76 \pm 0.07) \times 10^9$	$(2.34 \pm 0.05) \times 10^6$	14.1 ± 0.4	432 ± 9

表 4 结果显示,尽管 CBT-AD 融合了 CNN,BiLSTM 和 Transformer 组件,但其在计算量、推理速度和内存占用上均显著优于多个基线模型,甚至优于结构相对简单的 CNN-BiLSTM 和 BiLSTM-Attention。这一看似“反直觉”的结果,恰恰源于本文针对效率瓶颈的关键优化设计。

1)局部特征提取模块:严格采用深度可分离卷积替代标准卷积,大幅降低了卷积运算的计算复杂度和参数量,奠定了高效的局部特征提取基础。

2)时域分块与稀疏局部窗口注意力:这是突破 Transformer 效率瓶颈的核心。它通过将长序列分割为子块并在块内限制注意力窗口,有效降低全局自注意力的复杂度。该设计完全规避了标准 Transformer(如 CNN-Transformer 所用)及其全局注意力变体(如 BiLSTM-Attention 中的注意力层)的平方级计算开销,是 FLOPs 低、FPS 高、Memory 小的最主要原因。

3)动态门控特征融合:采用平均池化将时序特征降至至样本级表示,并通过轻量级门控网络学习融合权重,生成固定维度(128 维)的融合特征 F_{fusion} 。这有效避免了简单拼接不同模块特征导致的高维特征膨胀(如从 192 维降至 128 维),控制了后续分类器的计算负担。

因此,虽然 CBT-AD 集成了多类组件,但其通过高效模块选择、注意力复杂度优化以及避免维度膨胀的融合策略,成功地将整体计算和内存开销控制在极低水平,实现了检测精度与计算效率的协同优化。表 4 的数据实证了这些优化措施的有效性。本文模型在参数量上多于 CNN-Transformer,主要是由于模型架构中多模态特征协同机制引入了动态门控参数和跨块注意力投影矩阵。

结束语 本研究针对时序数据异常检测领域中局部特征提取与全局依赖建模之间协同不足,以及在长序列上的计算效率低的问题,提出了一种基于深度学习的混合神经网络架构 CBT-AD。该架构创新性地融合深度可分离卷积、双向门控循环单元及稀疏自注意力机制,通过构建跨层次特征协同框架,有效解决了传统方法在检测精度与计算效率之间的权衡困境。实验结果表明,在 UNSW-NB15, CICIDS2017, SwaT, SMD 这 4 个公开时序数据集上, CBT-AD 均取得了高于基线模型的 F1 分数,在细微异常特征的识别和抗噪性能上均表现出强大的鲁棒性,这为在复杂数据场景中应用该模型提供了坚实的理论依据和实践支持。同时,在长序列数据处理问题上做出了显著改进,通过引入改进的稀疏自注意力

机制,有效降低了传统全连接自注意力模块在序列长度增加时带来的计算复杂度。CBT-AD 模型在检测精度、鲁棒性、计算效率等方面均显著优于现有主流方法,为时序数据异常检测提供了一种新型的高效解决方案。

尽管 CBT-AD 模型展现出优越性能,但本研究仍存在一些局限性,这为未来的探索提供了方向。

1)在线适应性概念漂移处理。当前模型采用离线批量训练,难以动态适应实时概念漂移和未知异常模式。未来应探索在线增量学习或持续学习机制,实现高效模型更新和知识记忆,提升动态环境适应能力。

2)降低标注依赖。本文模型的训练依赖标注数据,异常样本稀缺时成本高昂。未来应融合半监督和自监督学习,通过时序自监督任务利用无标签数据,减少标注需求,拓展在资源有限场景中的应用。

3)领域深度优化与应用。由于通用设计在特定领域优化空间大,未来应结合领域知识定制模型,验证其在噪声下进行微弱异常识别和实时检测中的效能,加速工业落地。

参 考 文 献

[1] REN L, JIA Z, LAILI Y, et al. Deep Learning for Time-Series Prediction in IIoT: Progress, Challenges, and Prospects [J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 35(11):15072-15091.

[2] ZAMANZADEH DARBAN Z, WEBB G I, PAN S, et al. Deep learning for time series anomaly detection: A survey [J]. ACM Computing Surveys, 2024, 57(1):1-42.

[3] YAN P, ABDULKADIR A, LULEY P P, et al. A comprehensive survey of deep transfer learning for anomaly detection in industrial time series: Methods, applications, and directions [J]. IEEE Access, 2024, 12(1):3768-3789.

[4] MOHAMMADI FOUMANI N, MILLER L, TAN C W, et al. Deep learning for time series classification and extrinsic regression: A current survey [J]. ACM Computing Surveys, 2024, 56(9):1-45.

[5] YANG Y P, WANG S T. Research on malicious traffic classification algorithm based on CNN combined with BiGRU [J]. Computer Science, 2024, 51(11A):231100106-9.

[6] YE T, QIN W, LI Y, et al. Dense and small object detection in UAV-vision based on a global-local feature enhanced network [J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71:1-13.

[7] JIN M,LIU Y,ZHENG Y,et al. Anemone:Graph anomaly detection with multi-scale contrastive learning[C]//Proceedings of the 30th ACMInternational Conference on Information & Knowledge Management. 2021;3122-3126.

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

[9] ZHANG Y,WANG J,CHEN Y,et al. Adaptive memory networks with self-supervised learning for unsupervised anomaly detection[J]. IEEE Transactions on Knowledge and Data Engineering,2022,35(12):12068-12080.

[10] LIU P,QIAN W,ZHANG H,et al. Automatic sleep stage classification using deep learning: signals, data representation, and neural networks [J]. Artificial Intelligence Review, 2024, 57(11):301-369.

[11] ZHUANG X B,NIU B,LIN Z J,et al. Multi-parameter deception jamming detection method of parallel CNN-Transformer neural network based on gating mechanism[J/OL]. <http://kns.cnki.net/kcms/detail/11.4494.TN.20250402.1524.004.html>

[12] WANG H,DI X,WANG Y,et al. An Intelligent Digital Twin Method Based on Spatio-Temporal Feature Fusion for IoT Attack Behavior Identification[J]. IEEE Journal on Selected Areas in Communications,2023,41(11):3561-3572.

[13] MU X T,CHENG K,SONG A X,et al. Privacy-preserving federated learning against Byzantine attacks[J]. Chinese Journal of Computers,2024,47(4):842-861.

[14] PIAO X,CHEN Z,MURAYAMA T,et al. Fredformer:Frequency debiased transformer for time series forecasting[C]//Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2024;2400-2410.

[15] PAN J,WANG X H. Time series prediction model integrating multi-scale features and attention mechanism [J]. Computer Science,2026,53(2):180-186.

[16] CHEN W,LI L,LI J. Multimodal selective state space model-based time series classification for electricity theft detection[J]. Expert Systems with Applications,2025,278:127364.

[17] YIN T,ZHANG Z,HOU S,et al. MSAnomaly: Time Series Anomaly Detection with Multi-scale Augmentation and Fusion [C]//International Conference on Advanced Data Mining and Applications. Singapore:Springer,2024:356-371.

[18] CARMONA C U,AUBET F X,FLUNKERT V,et al. Neural contextual anomaly detection for time series[J]. arXiv:2107.07702,2021.

[19] MOORE A,MORELLI D. conDENSE:Conditional Density Estimation for Time Series Anomaly Detection[J]. Journal of Artificial Intelligence Research,2024,79:801-824.

[20] GOSWAMI M,CHALLU C,CALLOT L,et al. Unsupervised model selection for time-series anomaly detection[J]. arXiv:2210.01078,2022.

[21] SCHMIDL S,NAUMANN F,PAPENBROCK T. AutoTSAD: Unsupervised Holistic Anomaly Detection for Time Series Data [J]. Proceedings of the VLDB Endowment,2024,17(11):2987-3002.

[22] GHOJOGH B,TOUTOUNCHIAN M A. Probabilistic Classification by Density Estimation Using Gaussian Mixture Model and Masked Autoregressive Flow[J]. arXiv:2310.10843,2023.

[23] ZHOU H,ZHANG S,PENG J,et al. Informer:Beyond efficient transformer for long sequence time-series forecasting[C]//Pro-

ceedings of the AAAI Conference on Artificial Intelligence. 2021;11106-11115.

[24] XU J,WU H,WANG J,et al. Anomaly transformer: Time series anomaly detection with association discrepancy[J]. arXiv:2110.02642,2021.

[25] TULI S,CASALE G,JENNINGS N R. Tranad:Deep transformer networks for anomaly detection in multivariate time series data[J]. arXiv:2201.07284,2022.

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

[27] YANG Y,ZHANG C,ZHOU T,et al. Dcdetector: Dual attention contrastive representation learning for time series anomaly detection[C]//Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2023;3033-3045.

[28] DARBAN Z Z,WEBB G I,PAN S,et al. CARLA: Self-supervised contrastive representation learning for time series anomaly detection[J]. Pattern Recognition,2025,157:110874.

[29] LEE S,PARK T,LEE K. Soft contrastive learning for time series[J]. arXiv:2312.16424,2023.

[30] ZHAO B C,MA J J,CUI L,et al. Anomaly detection of photovoltaic power generation based on improved VMD-XGBoost-BiLSTM combined model [J]. Computer Engineering, 2024, 50(3):306-316.

[31] NING Z,MIAO H,JIANG Z,et al. Using Multi-Scale Convolution Fusion and Memory-Augmented Adversarial Autoencoder to Detect Diverse Anomalies in Multivariate Time Series [J]. Tsinghua Science and Technology,2024,30(1):234-246.

[32] LIU J,LI Q,AN S,et al. EdgeConvFormer:Dynamic Graph CNN and Transformer based Anomaly Detection in Multivariate Time Series[J]. arXiv:2312.01729,2023.

[33] ZHANG S,DU M,ZHANG G J. Research on Temperature Forecasting Model Based on BiLSTM-Attention Algorithm [C]//International Conference on Image, Vision and Intelligent Systems. Singapore:Springer,2025:461-472.

[34] ZHOU T,MA Z,WEN Q,et al. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting[C]//International Conference on Machine Learning. PMLR, 2022: 27268-27286.

[35] BAI N,WANG X,HAN R,et al. PAFormer: anomaly detection of time series with parallel-attention transformer [J]. IEEE Transactions on Neural Networks and Learning Systems,2023, 36(2):3315-3328.

[36] LIU J,LI Q,AN S,et al. EdgeConvFormer:Dynamic Graph CNN and Transformer based Anomaly Detection in Multivariate Time Series[J]. arXiv:2312.01729,2023.

XU Jian, born 1980, Ph. D, associate professor, is a member of CCF (No. P5359M). His main research interests include artificial intelligence and cyberspace security.

(责任编辑:何杨)