doi:10.19562/j.chinasae.qcgc.2025.02.009

增强双流Transformer的柴油发动机剩余寿命预测模型*

张 曦[1],杨 颖[1],陈超君[2],王春风[2],杨 磊[3]
(1. 广西大学计算机与电子信息学院,南宁 530004;2. 广西玉柴机器股份有限公司工艺工程部,玉林 537005;
3. 广西科学院,南宁 530007)

[摘要] 基于Transformer的模型在剩余使用寿命(remaining useful life,RUL)预测方面取得了显著的进展。然而,现有Transformer模型主要存在以下不足:模型在提取局部特征方面有所欠缺,且没有同时考虑输入特征的不同时间和不同空间的重要性。针对以上问题,提出一种增强的双流Transformer模型,通过局部特征提取模块和交互融合模块对模型进行增强。首先,通过局部特征提取模块分别在时间流和空间流提取局部特征,以弥补Transformer在局部特征提取方面的不足。然后,使用双流Transformer分别在时间和空间维度提取长期依赖,增强双流分支的互补学习。最后,构建交互融合模块,通过双线性融合方法捕获流级交互,进一步提升预测效果。使用多个模型在某柴油发动机制造商两个真实的数据集上进行实验,其结果表明评价指标RMSE和Score至少分别降低3.23%和5.89%。

关键词:剩余使用寿命预测;Transformer编码器;卷积神经网络;特征融合;滑动窗口

Enhanced Two-Stream Transformer Model for Remaining Useful Life Prediction of Diesel Engines

Zhang Xi[1], Yang Ying[1], Chen Chaojun[2], Wang Chunfeng[2] & Yang Lei[3]
1. *School of Computer, Electronics and Information, Guangxi University, Nanning* 530004;
2. *Process and Engineering Department, Guangxi Yuchai Machinery Co., Ltd., Yulin* 537005;
3. *Guangxi Academy of Science, Nanning* 530007

[**Abstract**] Transformer-based models have made significant progress in Remaining Useful Life (RUL) prediction. However, existing Transformer models have the following limitation of difficulty in local feature extraction and failure to consider the importance of varying temporal and spatial input features. To solve the problems, in this paper, an enhanced two-stream Transformer model is proposed, which is reinforced by the local feature extraction module and the interaction fusion module. Firstly, the local feature extraction module captures local features from both the temporal and spatial streams to compensate for the Transformer's deficiency in local feature extraction. Then, the two-stream Transformer is used to extract long-term dependencies in the temporal and spatial dimensions, enhancing complementary learning between the two streams. Finally, the interaction fusion module is constructed to capture stream-level interaction using bilinear fusion, further improving prediction performance. Experiments using multiple models on two real-world datasets from a diesel engine manufacturer demonstrate that the evaluation metrics RMSE and Score are reduced by at least 3.23% and 5.89%, respectively.

Keywords: remaining useful life prediction; Transformer encoder; convolutional neural network; feature fusion; sliding window

*广西创新驱动发展专项(桂科AA20302002-1)和广西科技基地与人才专项(桂科AD21076002)资助。
原稿收到日期为2024年08月14日,修改稿收到日期为2024年10月16日。
通信作者:杨颖,教授,博士,E-mail:yingy2004@126.com。

前言

故障预测与健康管理(prognostics health management,PHM)是保障工业系统安全运行的关键技术之一,其核心任务之一是基于运行状态信息预测设备的剩余使用寿命(remaining useful life,RUL)[1]。RUL预测通过采集的传感器数据提供有关机械设备健康状态的关键信息,帮助决策者提前规划,从而实现机械设备的科学运维[2]。

RUL预测方法可以分为基于模型的方法和深度学习的方法。虽然基于模型的解决方案能够提供高可解释性的准确估计,但其依赖复杂的先验知识[3]。相比之下,基于深度学习的解决方案不需要先验知识,并展现出较强的预测效果。Li等[4]使用深度卷积神经网络对设备的退化趋势进行建模,以实现RUL预测。Wang等[5]构建了循环卷积网络,挖掘监测序列中的时间依赖性,并记忆退化信息。Shi等[6]提出了一种基于指数平滑的双注意力LSTM轻量级模型用于RUL预测。Zhu等[7]通过将残差网络与自注意力机制深度融合,结合CNN和GRU用于RUL预测。与基于模型的方法相比,深度学习方法无须手动特征工程,能够有效处理多维时间序列,已在工业设备健康管理中得到广泛应用[8-9]。尽管上述方法取得了较好的效果,但CNN由于卷积核的感受野限制,难以捕捉长距离的上下文特征,而基于RNN及其变体的方法则易出现梯度消失问题。

近年来,在机械设备RUL预测领域,Transformer模型凭借其强大的全局信息捕捉能力和并行计算能力[10],展示出优秀的性能。Su等[11]通过改进位置编码层,设计了一个自适应Transformer的端到端框架,用于预测轴承的剩余使用寿命。Gu等[12]将卷积神经网络与Transformer结合,提取局部和全局的时间信息。Xiang等[13]提出了一种基于贝叶斯门控的Transformer模型,平衡短期和长期特征依赖。

尽管Transformer模型在提取全局特征相关性时表现出色,但却在提取局部特征相关性方面效果不佳。因此,理想的预测方法应同时具备提取全局特征和局部特征相关性的能力,以更好地捕捉和利用数据中的多维信息。

此外,多元传感器数据包含与退化状态相关的时间和空间维度信息,这些信息对预测结果的贡献不同,因此须赋予它们不同的权重。为解决这一问题,Jin等[14]提出了一种基于双向LSTM(Bi-LSTM)的双流网络来预测航空发动机的剩余使用寿命,并设计了加法、加权和拼接3种融合方式来组合两个流的特征表示。Gao等[15]则提出了多层Bi-LSTM与卷积神经网络的组合模型,用以提取高级特征,并通过加权求和的融合方式输出预测结果。Zhang等[16]提出了一种基于Transformer的双流自注意力网络模型,分别提取时间和空间特征,并通过拼接融合的方式将结果输入Transformer解码器。然而,上述方法仅对双流输出进行1阶线性组合,未能充分挖掘流级特征的交互。

基于上述分析,为同时考虑多元传感器数据中的"局部-全局"及"时间-空间"特征,本文提出了一种增强的双流Transformer模型。通过引入局部特征提取模块和流级交互融合模块,进一步提升了双流Transformer的性能。本文的主要贡献如下:

(1)提出双流Transformer架构,分别提取时间流和空间流特征,增强了双流分支的互补学习效果。

(2)设计了局部特征提取模块,通过多尺度卷积神经网络在时间和空间维度上提取特征,不仅有效捕捉局部特征,还进一步挖掘了时间和空间特征,弥补了Transformer在提取局部特征相关性上的不足。

(3)设计了流级交互融合模块,采用双线性融合方法捕捉双流模型中的流级交互,进一步提升了预测性能。

(4)在某柴油发动机制造商的两个真实数据集上的实验结果验证了该模型的有效性。

1 模型设计

针对以上问题,提出一种增强的双流Transformer模型,通过局部特征提取模块和交互融合模块对模型进行增强。

1.1 总体结构

如图1所示,提出的模型框架包括输入层、局部特征提取层、全局特征提取层和交互融合输出层。首先,将时间流和空间流数据分别输入多尺度卷积网络,分别提取不同时间步和不同传感器之间的局部特征。然后,将带有局部特征的信息输入到Transformer编码器中,捕捉全局时间和空间特征相关性。最后,交互融合输出层将时间流和空间流的输出进行交互融合,并输出RUL预测结果。

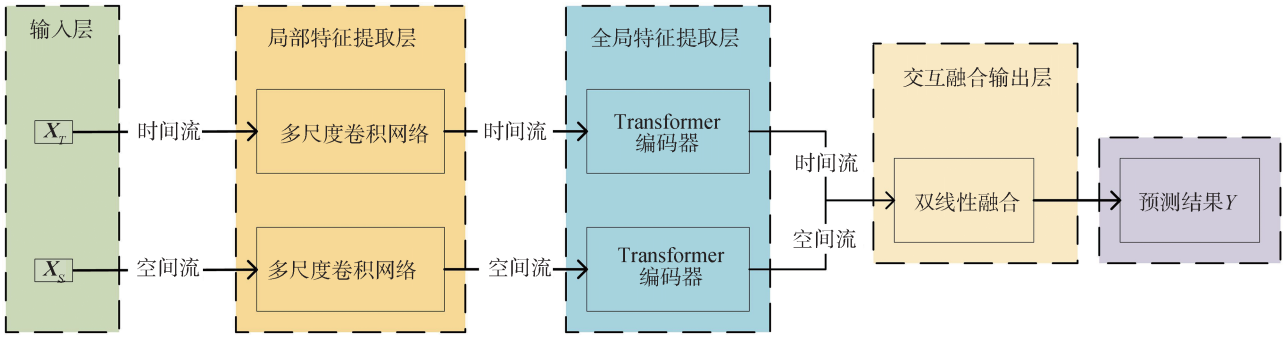
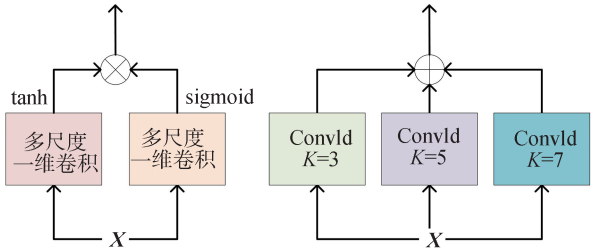


图1 FIT网络流程图

1.2 局部特征提取层

局部特征提取层旨在捕捉数据中的细粒度特征,CNN因其在捕获局部特征方面的鲁棒性而广为人知,可以根据内核大小提取不同尺度的特征。如图2所示,本文利用多尺度卷积网络作为局部特征提取器,以有效提取相邻传感器和相邻时间步的局部特征。通过双流结构分别提取局部时间特征和局部空间特征相关性,确保了双流结构的独立学习能力,避免了单流学习中可能出现的特征混淆。



(a)局部特征提取模块 (b)多尺度一维卷积

图2 局部特征提取层和多尺度一维卷积

1.2.1 局部时间特征提取

局部时间特征提取层旨在通过多尺度卷积网络捕捉沿时间维度的退化特征。该层为不同的时间步分配不同的权重,重点强调对RUL预测影响更大的权重。

时间流数据表示M个时间步的N个传感器数据,大小为$M\times N$,可以表示为$\boldsymbol{X}_T=[T_1,T_2,\cdots,T_M]\in\mathbb{R}^{M\times N}$,$\boldsymbol{T}_i=[t_1^i,t_2^i,\cdots,t_N^i]$,$\boldsymbol{T}_i$表示单个时间步的传感器序列,$t_N^i$表示第$i$个时间步的第$N$个传感器的数据。如图2(b)所示,通过采用多尺度卷积网络中3种不同大小的卷积核($\boldsymbol{KS}=[3,5,7]$),计算得到不同尺度的局部时间特征$\boldsymbol{L}_T^i$:

$$\boldsymbol{L}_T^i=\mathrm{conv}(\boldsymbol{W}_i,\boldsymbol{X}_T) \tag{1}$$

式中:conv(·)表示卷积操作;$\boldsymbol{W}_i$表示可学习的权重。为确保输出相同大小的特征图,在输入数据的两侧进行零填充,从而生成$M\times N$大小的特征图。

然后,将多尺度下提取的局部时间特征融合,计算得到的局部时间特征$\boldsymbol{L}_T$:

$$\boldsymbol{L}_T=\tanh\left(\sum_{i=1}^{3}\boldsymbol{L}_T^i\right)\odot\mathrm{sigmoid}\left(\sum_{i=1}^{3}\boldsymbol{L}_T^i\right) \tag{2}$$

输入特征$\boldsymbol{X}_T$和局部特征$\boldsymbol{L}_T$通过残差连接形成新的特征$\boldsymbol{F}_T$,然后将$\boldsymbol{F}_T$作为后续全局时间特征提取层的输入。

$$\boldsymbol{F}_T=\boldsymbol{X}_T+\boldsymbol{L}_T \tag{3}$$

1.2.2 局部空间特征提取

局部空间特征提取操作与局部时间特征提取操作类似,但其侧重于沿空间维度提取特征,并对重要的传感器特征分配更高的权重。

空间流表示N个传感器的M个时间步数据,大小为$N\times M$,可以表示为$\boldsymbol{X}_S=[S_1,S_2,\cdots,S_N]\in\mathbb{R}^{N\times M}$,$\boldsymbol{S}_j=[s_1^j,s_2^j,\cdots,s_M^j]$,$\boldsymbol{S}_j$表示单个传感器的时间序列,$s_M^j$表示第$j$个时间步的第$M$个传感器的数据。同样地,通过采用多尺度卷积网络中3种不同大小的卷积核($\boldsymbol{KS}=[3,5,7]$),计算得到不同尺度的局部空间特征$\boldsymbol{L}_S^j$:

$$\boldsymbol{L}_S^j=\mathrm{conv}(\boldsymbol{W}_j,\boldsymbol{X}_S) \tag{4}$$

式中$\boldsymbol{W}_j$表示可学习的权重。

然后,将多尺度下提取的局部空间特征融合,计算得到局部空间特征$\boldsymbol{L}_S$:

$$\boldsymbol{L}_S=\tanh\left(\sum_{i=1}^{3}\boldsymbol{L}_S^i\right)\odot\mathrm{sigmoid}\left(\sum_{i=1}^{3}\boldsymbol{L}_S^i\right) \tag{5}$$

输入特征$\boldsymbol{X}_S$和局部特征$\boldsymbol{L}_S$通过残差连接形成新的特征$\boldsymbol{F}_S$。然后将$\boldsymbol{F}_S$作为后续全局空间特征提取层的输入。

$$\boldsymbol{F}_S=\boldsymbol{X}_S+\boldsymbol{L}_S \tag{6}$$

1.3 全局特征提取层

全局特征提取层旨在捕捉数据中的全局特征依赖关系。如图3所示,该层采用多头注意力机制,将

局部特征提取层的输出作为本层的输入,从多个视角并行捕捉数据的内在相关性和动态变化。结合层归一化和前馈神经网络,此单元不仅使学习过程更加稳定,还增强了模型的表达能力,使其能够捕捉整个序列的全局特征相关性。通过双流结构分别提取全局时间特征和全局空间特征相关性,确保模型在捕捉长距离依赖关系时能够准确区分时间和空间信息,从而增强全局特征的表达能力,避免单流模型在处理复杂依赖时的特征混淆。

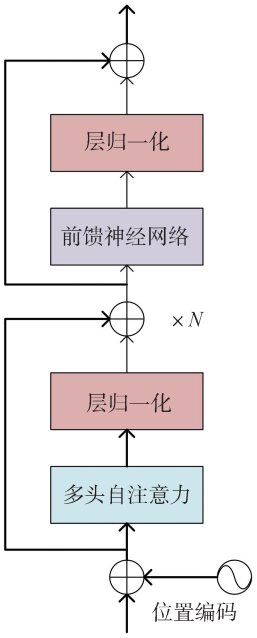


图3　全局特征提取层

1.3.1　全局时间特征提取

为了捕获时间特征的长期依赖性,查询矩阵 $\boldsymbol{Q}_T$、键矩阵 $\boldsymbol{K}_T$ 和值矩阵 $\boldsymbol{V}_T$ 是由输入序列 $\boldsymbol{F}_T$ 进行3个线性层映射得到的3个向量:

$$\boldsymbol{Q}_T = \boldsymbol{F}_T\boldsymbol{W}_T^Q, \boldsymbol{K}_T = \boldsymbol{F}_T\boldsymbol{W}_T^K, \boldsymbol{V}_T = \boldsymbol{F}_T\boldsymbol{W}_T^V \tag{7}$$

然后Transformer编码器沿时间维度进行softmax,得到加权的空间特征对应的权重向量:

$$\text{Attention}(\boldsymbol{Q}_T, \boldsymbol{K}_T, \boldsymbol{V}_T) = \text{softmax}\left(\frac{\boldsymbol{Q}_T\boldsymbol{K}_T^{\mathrm{T}}}{\sqrt{D}}\right)\boldsymbol{V}_T \tag{8}$$

编码器层中的多头自注意力机制使模型能够从不同的表现空间中学习信息,而不局限于单一表现空间。这种机制使模型更加注重数据中的重要信息,减少对外部信息的依赖。其计算公式为

$$head_k = \text{Attention}(\boldsymbol{Q}_T, \boldsymbol{K}_T, \boldsymbol{V}_T)_k \tag{9}$$

$$\text{MutilHead}(\boldsymbol{Q}_T, \boldsymbol{K}_T, \boldsymbol{V}_T) = \text{Concat}(\{head_k\}_{k=1}^h)\boldsymbol{W}_T \tag{10}$$

式中:$\boldsymbol{W}_T$ 表示多头注意力权重矩阵;h 表示注意力头的数量;$head_k$ 表示第 k 个头;Concat函数用于拼接每个头的输出值。

本文采用不同频率的正弦和余弦函数进行位置编码,计算公式如下:

$$PE(t, 2k) = \sin\left(\frac{t}{10000^{\frac{2k}{d}}}\right) \tag{11}$$

$$PE(t, 2k+1) = \cos\left(\frac{t}{10000^{\frac{2k}{d}}}\right) \tag{12}$$

式中:PE 表示位置编码;t 表示位置编码的位置;i 是0~(d/2 − 1)之间的整数值;d 表示每个位置向量的维度。

1.3.2　全局空间特征提取

与全局时间特征提取层类似,将 $\boldsymbol{F}_S$ 输入多头注意力层,计算得到查询矩阵 $\boldsymbol{Q}_S$、键矩阵 $\boldsymbol{K}_S$ 和值矩阵 $\boldsymbol{V}_S$,然后Transformer编码器沿空间维度进行softmax操作,得到加权的空间特征对应的权重向量:

$$\text{Attention}(\boldsymbol{Q}_S, \boldsymbol{K}_S, \boldsymbol{V}_S) = \text{softmax}\left(\frac{\boldsymbol{Q}_S\boldsymbol{K}_S^{\mathrm{T}}}{\sqrt{D}}\right)\boldsymbol{V}_S \tag{13}$$

同样地,全局空间特征的多头注意力的计算公式如下:

$$head_k = \text{Attention}(\boldsymbol{Q}_S, \boldsymbol{K}_S, \boldsymbol{V}_S)_k \tag{14}$$

$$\text{MutilHead}(\boldsymbol{Q}_S, \boldsymbol{K}_S, \boldsymbol{V}_S) = \text{Concat}(\{head_k\}_{k=1}^h)\boldsymbol{W}_S \tag{15}$$

式中 $\boldsymbol{W}_S$ 表示多头注意力权重矩阵。

1.4　双线性融合输出层

通过双流局部特征提取层和全局特征提取层得到的时间流输出 $\boldsymbol{O}_T$ 和空间流输出 $\boldsymbol{O}_S$:

$$\boldsymbol{O}_T = \text{MHSA}(\boldsymbol{F}_T) \tag{16}$$

$$\boldsymbol{O}_S = \text{MHSA}(\boldsymbol{F}_S) \tag{17}$$

如前所述,现有工作[14-16]大多采用求和或拼接作为融合层,但这些操作无法捕捉流级特征交互。受到CV领域广泛研究的双线性池化启发[17-18],如图4所示,本文提出了一个交互聚合层来融合流输出并实现流级特征交互,计算公式如下:

$$Y = b + \boldsymbol{W}_1^{\mathrm{T}}\boldsymbol{O}_T + \boldsymbol{W}_2^{\mathrm{T}}\boldsymbol{O}_S + \boldsymbol{O}_T^{\mathrm{T}}\boldsymbol{W}_3\boldsymbol{O}_S \tag{18}$$

式中:$b \in \mathbb{R}$,$\boldsymbol{W}_1 \in \mathbb{R}^{d_1 \times 1}$,$\boldsymbol{W}_2 \in \mathbb{R}^{d_2 \times 1}$,$\boldsymbol{W}_3 \in \mathbb{R}^{d_1 \times d_2}$ 是可学习的权重;d_1 和 d_2 分别表示 $\boldsymbol{O}_T$ 和 $\boldsymbol{O}_S$ 的维度;Y 是最后输出的RUL预测值。

双线性项 $\boldsymbol{O}_T^{\mathrm{T}}\boldsymbol{W}_3\boldsymbol{O}_S$ 对 $\boldsymbol{O}_T$ 和 $\boldsymbol{O}_S$ 之间的2阶相互作用进行建模,当 $\boldsymbol{W}_3$ 为单位矩阵时,该项即为点积。当将 $\boldsymbol{W}_3$ 为零矩阵时,就退化为线性加权融合,即 $b + [\boldsymbol{W}_1, \boldsymbol{W}_2]^{\mathrm{T}}[\boldsymbol{O}_T, \boldsymbol{O}_S]$。

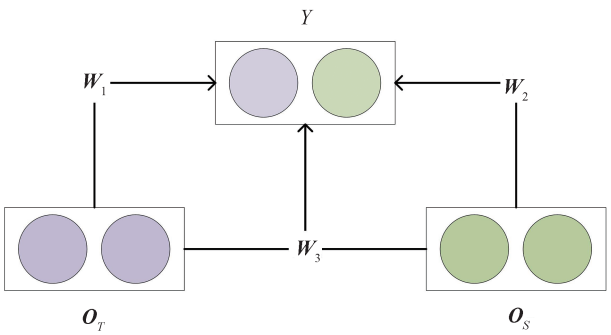


图 4 双线性融合输出层

2 实验设置

2.1 数据集

本次实验使用玉柴发动机质量大数据平台提供的YCS07系列中的两个柴油发动机数据集,分别为YCS07-A和YCS07-B。数据集信息如表1所示,每个数据集包含训练集和测试集。训练集包括100台柴油发动机从正常状态到故障状态的全程数据,测试集则包含100台柴油发动机从正常状态到某一特定时间点的数据。

表 1 数据集统计

数据集	YCS07-A	YCS07-B
训练发动机数量	100	100
测试发动机数量	100	100
故障模式	1	1

每条数据包括发动机编号、运行周期和12个传感器的监测信号,监测信号参数如表2所示。

表 2 柴油发动机监测信号参数

传感器编号	参数名称	单位
1	功率	kW
2	油耗率	g/(kW·h)
3	进气温度	℃
4	排气温度	℃
5	进水温度	℃
6	排水温度	℃
7	机油温度	℃
8	燃油温度	℃
9	转速	r/min
10	转矩	N·m
11	活塞漏气量	L/min
12	机油压力	kPa

2.2 数据预处理

2.2.1 数据平滑

由于12个传感器参数稳定性差且噪声较大,本文采用指数平滑法对其进行降噪。图5展示了指数平滑法的降噪效果,其数据来自YCS07-A训练集中1号柴油发动机的9号传感器在前100次运行周期的记录。

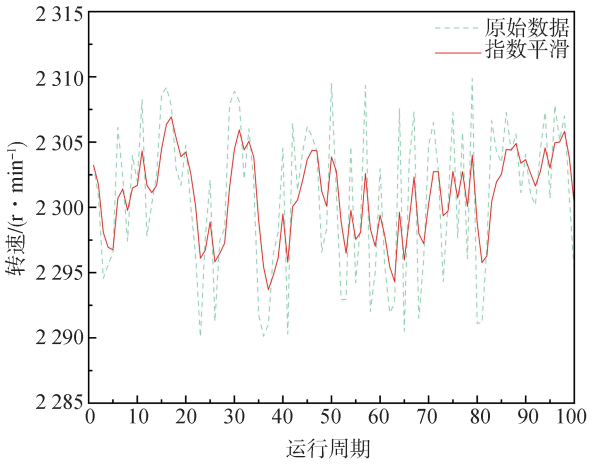


图 5 指数平滑法的降噪效果

由图5可知,指数平滑法不仅可以保留原始数据的变化趋势,还可以对原始数据进行降噪平滑处理,是一种有效的时序数据降噪技术。

2.2.2 归一化

由于单个传感器数据的不同取值范围会影响模型预测精度,因此需要对样本数据进行归一化。本文使用最小-最大归一化方法将每个传感器数据缩放到范围[0,1],具体来说,对于传感器数据 $X_i=\{X_1,X_2,\cdots,X_T\}$。

$$\hat{X}_i=\frac{X_i-\min(X_i)}{\max(X_i)-\min(X_i)} \tag{19}$$

式中:$\hat{X}$是归一化数据;$\max(X_i)$和 $\min(X_i)$分别代表最大值和最小值。

YCS07-A训练集中1号柴油发动机的归一化传感器数据如图6所示。

2.2.3 滑动时间窗口

如图7所示,为使模型能够从时间序列数据中获取尽可能多的有价值特征信息,采用固定长度的时间窗对原始时间序列进行分割。时间窗口以固定步幅在整个时间序列上滑动,每滑动一步取一个时间序列片段。然后,将分割得到的所有片段汇总起来形成新的数据集。

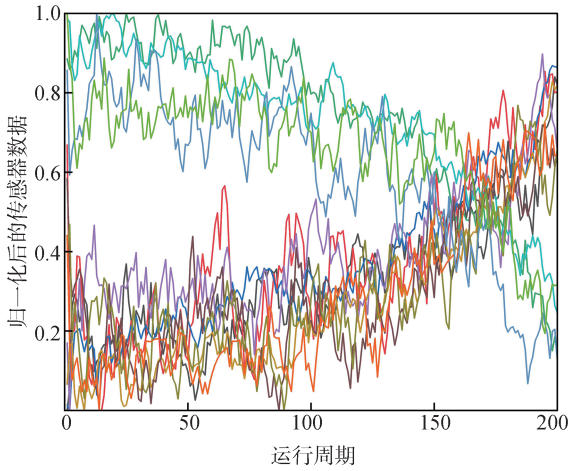


图 6　YCS07-A 训练集中 1 号柴油发动机的归一化数据

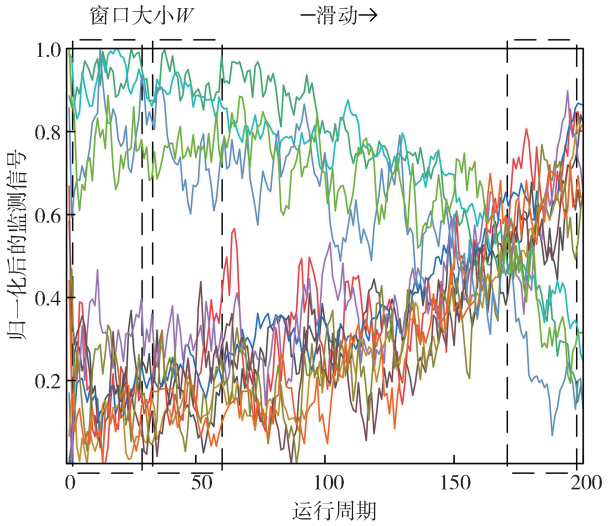


图 7　滑动时间窗口

本文将步幅设置为 1,以获取尽可能多的时间序列片段。通过滑动窗口分割出来的训练样本数量和测试样本数量如表 3 所示。

表 3　两个数据集的训练样本和测试样本数

数据集	时间窗口大小	训练样本数量	测试样本数量
YCS07-A	40	18 374	8 741
YCS07-B	40	18 967	8 701

2.2.4　RUL标签

本文提出的 FIT 方法采用有监督的训练方式,因此须合理确定输入数据对应的预期输出。借鉴 Jin 等[19]的经验,使用分段线性 RUL 方法代替实际的 RUL,如图 8 所示。

2.3　评价指标

为客观地评估 FIT 模型的预测性能,本文采用

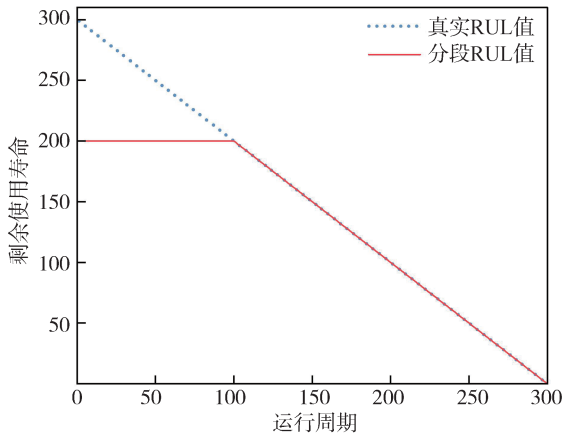


图 8　RUL 分段线性退化

均方根误差(RMSE)和评分函数(Score)这两种广泛认可的评估指标。

(1) 均方根误差(RMSE)是回归任务中常用的评估指标之一,其计算公式为

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2} \tag{20}$$

式中:N 代表样本数;y_n 和 $\hat{y}_n$ 分别表示第 n 个样本的实际 RUL 值和预测 RUL 值。

(2) 评分函数(Score)中提前预测和延迟预测的计算方式不同,其公式为

$$Score = \begin{cases} \sum_{n=1}^{N}\left(e^{-\frac{d_n}{13}} - 1\right), & d_n < 0 \\ \sum_{n=1}^{N}\left(e^{\frac{d_n}{10}} - 1\right), & d_n \geqslant 0 \end{cases} \tag{21}$$

式中 $d_n = \hat{y}_n - y_n$。

与 *RMSE* 不同,*Score* 对延迟预测的惩罚比对提前预测更强。

2.4　超参数

为确定 FIT 模型的超参数,采用网格搜索方法来确定模型的最优超参数,以增强 RUL 预测性能。训练过程中使用 Adam 优化器,编码器的丢弃率设置为 0.2,学习率设置为 0.001,批量大小设置为 128,迭代次数最多为 100 轮。完整参数如表 4 所示。

3　实验结果与分析

3.1　实验结果

FIT 在 YCS07-A 和 YCS07-B 数据集上的 RUL 预测结果如图 9 所示。首先,虽然少数点存在较大偏差,但整体上预测的趋势与真实值的变化趋势一

表 4 FIT 模型的超参数

超参数	数值
YCS07-A 滑动窗口大小	40
YCS07-B 滑动窗口大小	40
优化器	Adam
批量大小	128
学习率	0.001
迭代次数	100
丢弃率	0.2
多尺度一维卷积核大小	[3,5,7]
Transformer 编码器块数	2
Transformer 编码器注意力头数	3

致。然后,当 RUL 值较大时,预测的 RUL 值与真实值之间的偏差较大,这是因为模型在早期阶段捕捉到的退化信息较少,导致预测的不确定性增加。最后,当 RUL 值较小时,预测的 RUL 值与真实的 RUL 值更加接近,这是由于随着发动机的老化和磨损,模型可以更容易捕捉到退化信息。

图 10 展示了 YCS07-A 中的 20 号测试发动机和 YCS07-B 中的 30 号测试发动机的 RUL 预测结果。

在退化初期,预测 RUL 值与真实 RUL 值之间存在较大的误差,这是因为设备刚开始运行时工况良好,退化信息较少。在退化曲线拐点处,预测 RUL 值与真实 RUL 值的误差存在一些波动,这是因为拐点处发动机的工况变化较为剧烈。在退化后期,预测 RUL 值与真实 RUL 值的误差相对较小,这是因为设备的性能退化趋势变得更加明显和规律,模型能够更好地捕捉到这种趋势。

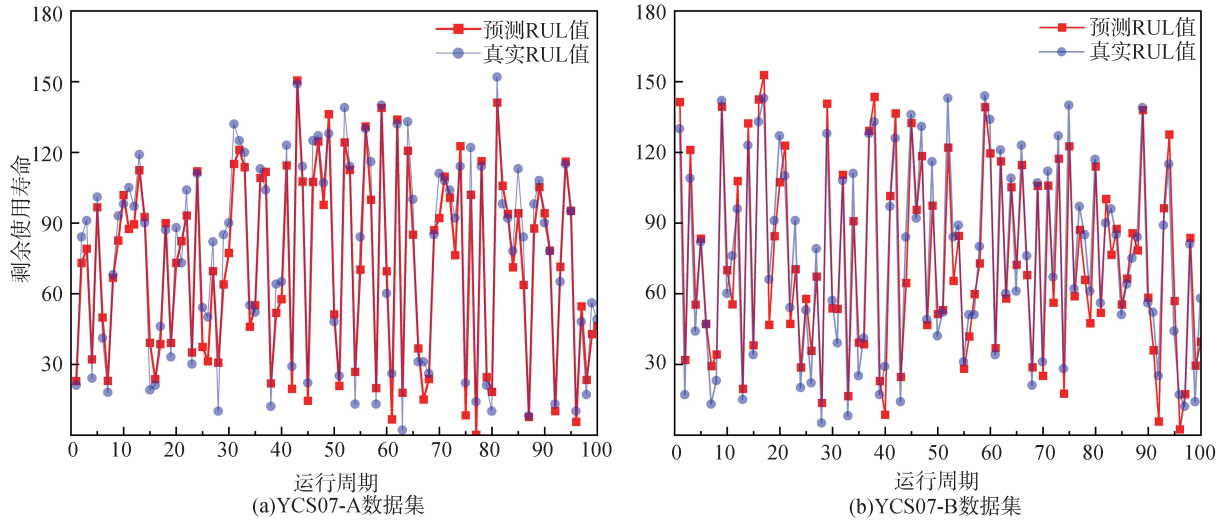


图 9 两个数据集的预测值和真实值

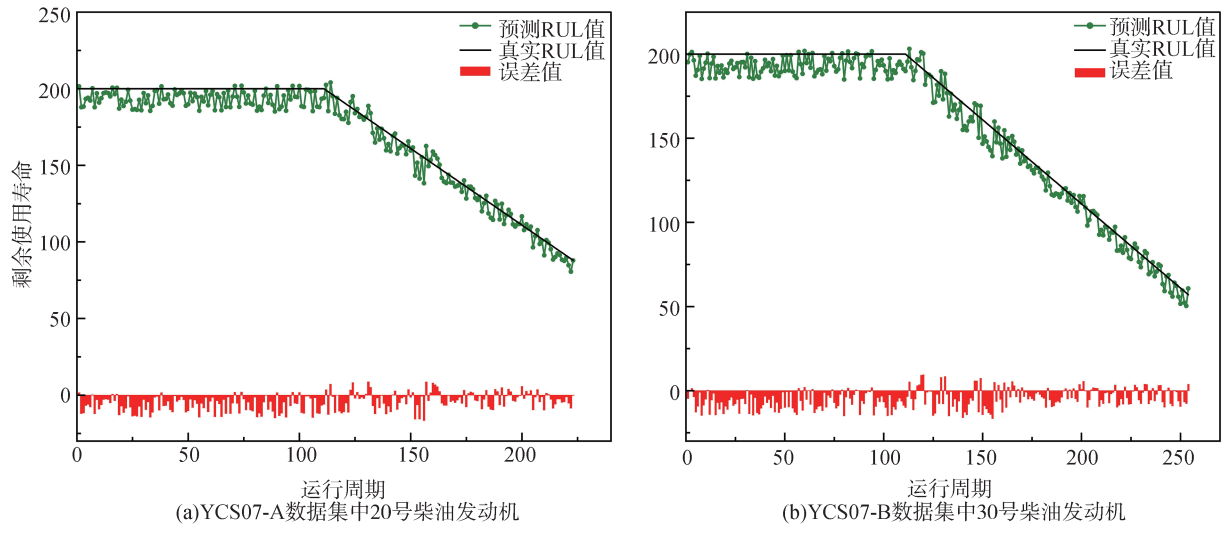


图 10 两个数据集中单台测试发动机的预测值和真实值

3.2　与其他模型对比

为验证所提出方法的性能,本文将FIT模型与以下模型的评价指标进行对比。

(1) 基线方法:CNN、RNN、Transformer。

(2) 基于CNN/RNN变体的方法:DA-LSTM[6]、Res-HAS[7]。

(3) 基于双流网络的方法:DCFA[15]、DAST[16]。

(4) 基于Transformer的改进方法:AT[11]、CNN-Transformer[12]、BGT[13]。

为减轻随机性的影响,本文采用10次重复实验的平均*RMSE*和*Score*作为实验结果,实验结果如表5所示。

表5　与其他模型对比

类别	方法	YCS07-A		YCS07-B	
		RMSE	*Score*	*RMSE*	*Score*
基线方法	CNN	23.71	1126.39	28.28	1630.49
	RNN	18.51	414.69	17.47	371.74
	Transformer	13.56	198.09	14.2	225.44
基于CNN/RNN变体的方法	DA-LSTM	12.61	183.49	13.87	207.94
	Res-HAS	12.63	176.96	13.45	197.48
基于双流网络的方法	DCFA	12.31	171.92	13.66	199.16
	DAST	11.77	166.76	13.59	207.27
基于Transformer的改进方法	AT	11.65	150.45	12.59	176.52
	CNN-Transformer	11.31	144.51	12.28	174.37
	BGT	<u>11.24</u>	<u>144.33</u>	<u>12.06</u>	<u>165.16</u>
本文模型	FIT	**10.32**	**129.27**	**11.67**	**155.43**

基于本文所提出的方法,与基线方法、基于CNN/RNN变体的方法、基于双流网络的方法和基于Transformer改进的方法进行对比,其结果如表5所示。本文提出的FIT方法在RUL性能上表现最优,表明其在提取数据中的局部和全局退化信息方面具有更好的能力。

在基线方法中,Transformer的性能优于CNN和RNN,而基于Transformer的改进方法则优于基于CNN/RNN变体的方法。这表明,Transformer通过自注意力机制能够更有效地捕捉长距离依赖关系,而CNN和RNN在处理长距离依赖时效率较低。

与基于CNN/RNN变体的方法相比,FIT方法的*RMSE*降低了至少13.23%,*Score*降低了至少21.29%;与基于双流网络的方法相比,FIT方法的*RMSE*降低了至少12.31%,*Score*降低了至少21.95%;与基于Transformer的改进方法相比,FIT方法的*RMSE*降低了至少3.23%,*Score*降低了至少5.89%。

3.3　消融实验

为验证局部特征提取模块和交互融合层模块的有效性,本文将FIT模型与以下变体进行比较。

(1) DualTransformer:简单地将两个Transformer编码器作为两个流,分别提取时间流和空间流的特征。

(2) IT:在FIT基础上不使用局部特征提取模块。

(3) FIT-Sum:在FIT中使用求和融合。

(4) FIT-Concat:在FIT中使用拼接融合。

(5) FIT-Hadamard:在FIT中使用哈达玛积融合。

消融研究结果如表6所示。当删除局部特征提取模块或用其他常用的融合操作替换双线性融合时,预测性能明显下降,这验证了本文所提出的特征选择和双线性融合模块的有效性。此外,由于替换局部特征选择模块导致的性能下降幅度更大,因此局部特征选择模块比双线性融合模块起着更为重要的作用。另外,双线性融合方式相比于使用加权求和、拼接、元素级融合方式的预测性能更好,说明了双线性融合方式的有效性。

表6　消融实验结果

方法	YCS07-A		YCS07-B	
	RMSE	*Score*	*RMSE*	*Score*
DualTransformer	12.11	158.24	13.88	217.31
IT	11.55	150.03	13.18	191.02
FIT-Sum	10.75	135.28	12.79	188.49
FIT-Concat	10.66	134.22	12.55	<u>169.82</u>
FIT-Hadamard	<u>10.58</u>	<u>131.41</u>	<u>12.44</u>	184.22
FIT	**10.32**	**129.27**	**11.67**	**155.43**

3.4 计算复杂度对比

为验证FIT模型在实际应用中的计算复杂度和计算成本,本文使用“FLOPs”、“参数量”、“训练时间”和“模型大小”4个综合评价指标。

其中:FLOPs是每秒浮点运算次数;参数量是模型的参数数量;训练时间是从模型开始训练到结束所用的时间差;模型大小表示训练后的模型文件所占存储空间大小。“FLOPs”和“参数量”可以由thop库计算得来。

将FIT模型和表5中性能较好的Transformer模型在YCS07-A数据集上进行计算复杂度对比实验,实验结果如表7所示。

表7 计算复杂度对比结果

方法	FLOPs/$\times10^6$	参数量/$\times10^6$	训练时间/min	模型大小/MB
DAST	**7.09**	**0.374**	**4.7**	**1.58**
AT	7.24	0.415	5.4	1.79
CNN-Transformer	7.28	0.397	4.9	1.63
BGT	7.42	0.402	5.6	1.66
FIT	7.17	0.387	4.8	1.61

由表7可知,FIT模型和其他4个模型具有相近的计算复杂度。尽管DAST模型计算复杂度较低,但是其准确性与FIT模型相比较差。虽然FIT次优,但却在相近的计算复杂度下,有着更好的性能。

在实际工业应用中,过于复杂的模型虽然可能提高计算准确率,但也会显著增加计算资源和时间成本,导致效率低下和成本上升;而过于简单的模型则可能因无法充分捕捉数据特征而降低准确率,影响决策效果。因此,在实际应用中,需要权衡模型复杂性,找到既能满足准确率要求,又能控制计算成本的最佳平衡点,以确保模型的高效、稳定和可靠运行。

4 结论

本文提出了一种增强的柴油发动机剩余使用寿命预测模型FIT。模型通过设计多尺度卷积网络来提取时间流和空间流中的局部特征,从而弥补了Transformer在捕捉局部特征方面的不足。此外,模型采用双线性融合模块,实现时间流和空间流之间的流级交互融合,从而提升整体预测性能。实验结果表明,FIT模型在某柴油发动机制造商提供的两个真实数据集上的*RMSE*和*Score*至少分别降低3.23%和5.89%。

未来的工作将考虑进一步降低Transformer编码器的计算复杂度,并采用自注意力蒸馏学习减少参数维度和参数量,以提高FIT模型在生产应用中的鲁棒性。

参考文献

[1] LIU H, SUN Y, DING W, et al. Enhancing non-stationary feature learning for remaining useful life prediction of aero-engine under multiple operating conditions [J]. Measurement, 2024, 227: 114242.

[2] FU S, LIN L, WANG Y, et al. MCA-DTCN: a novel dual-task temporal convolutional network with multi-channel attention for first prediction time detection and remaining useful life prediction [J]. Reliability Engineering & System Safety, 2024, 241: 109696.

[3] HU J, SUN Q, YE Z S, et al. Joint modeling of degradation and lifetime data for RUL prediction of deteriorating products [J]. IEEE Transactions on Industrial Informatics, 2020, 17(7): 4521-4531.

[4] LI X, DING Q, SUN J Q. Remaining useful life estimation in prognostics using deep convolution neural networks [J]. Reliability Engineering & System Safety, 2018, 172: 1-11.

[5] WANG B, LEI Y, YAN T, et al. Recurrent convolutional neural network: a new framework for remaining useful life prediction of machinery [J]. Neurocomputing, 2020, 379: 117-129.

[6] SHI J, ZHONG J, ZHANG Y, et al. A dual attention LSTM lightweight model based on exponential smoothing for remaining useful life prediction [J]. Reliability Engineering & System Safety, 2024, 243: 109821.

[7] ZHU J, JIANG Q, SHEN Y, et al. Res-HSA: residual hybrid network with self-attention mechanism for RUL prediction of rotating machinery [J]. Engineering Applications of Artificial Intelligence, 2023, 124: 106491.

[8] GAN F, SHAO H, XIA B. An adaptive model with dual-dimensional attention for remaining useful life prediction of aero-engine [J]. Knowledge-Based Systems, 2024, 293: 111738.

[9] WANG B, LEI Y, LI N, et al. Multiscale convolutional attention network for predicting remaining useful life of machinery [J]. IEEE Transactions on Industrial Electronics, 2020, 68(8): 7496-7504.

[10] DU X, JIA L, HAQ I U. Fault diagnosis based on SPBO-SDAE and transformer neural network for rotating machinery [J]. Measurement, 2022, 188: 110545.

[11] SU X, LIU H, TAO L, et al. An end-to-end framework for remaining useful life prediction of rolling bearing based on feature pre-extraction mechanism and deep adaptive transformer model [J]. Computers & Industrial Engineering, 2021, 161: 107531.

(下转第325页)

6 结论

本文从结合整车应用场景和电驱转矩过0防抖目标出发,分析了整车的各种防抖控制功能需求。根据防抖功能需求,设计出防抖的软件模块架构及控制策略,并同时给出相应控制阶段的目标计算推导,并进行了仿真分析。然后结合实际项目,将防抖控制算法应用到整车上,通过靠齿和转矩过0的实测数据表明基本达到设计的控制效果。本文通过对防抖控制理论的初步探讨、控制目标推导、仿真分析以及实车测试,证明了该控制策略不仅有效实现电驱转矩过0的防抖功能,同时根据实车测试出来的车辆加速度曲线及电驱转速波动范围可知该控制策略还保证了车辆优异的防抖性能。目前该产品应用车型已经成功上市且在驾驶性上取得了非常好的评价。另外,计算仿真和理论探讨的结果虽然对防抖软件实车的标定控制基本保持一致,但针对过0的极限梯度边界限制仍靠工程经验及测试数据来指导标定控制。而深层次地探讨电驱间隙及过0的真实抖动机理和边界获取可作为后续的基础研究,借助于CAE物理仿真及测试设备来得到准确的防抖转矩极限梯度边界。

参考文献

[1] 魏敦烈.基于电机波动补偿的电动车防抖控制设计[J].汽车零部件,2022,2:39-42.
WEI Dunlei. Design of anti-shake control of pure electric vehicle based on motor speed fluctuation compensation [J]. Automobile Parts, 2022,2:39-42.

[2] 纪历,马雪晴,陈震民.磁悬浮高速电机转子低频振动机理及补偿方法[J].中国机械工程,2022,33(17):2053-2060.
JI Li, MA Xueqing, CHEN Zhenmin. Low frequency vibration mechanism for amb high-speed motor rotor systems and its compensation strategy [J]. China Mechanical Engineering, 2022,33(17):2053-2060.

[3] 张剑锋,叶先军.电动化车辆主动防抖控制策略的研究[J].上海汽车, 2020,10(1):4-7.
ZHANG Jianfeng, YE Xianjun. A research on active anti jerk control strategy for electric vehicles [J]. Shanghai Auto, 2020, 10(1):4-7.

[4] 严周栋,杭鹏,陈重璞,等.分布式电驱动装载机驱动防滑控制[J].汽车工程,2023,45(10):1943-1953.
YAN Zhoudong, HANG Peng, CHEN Chongpu, et al. Drive anti-skid control of distributed electric drive loaderr [J]. Automotive Engineering, 2023,45(10):1943-1953.

[5] 赵治国,王晨,张彤,等.纯电Tip-In/Tip Out工况的前馈校正与主动阻尼防抖控制[J].汽车工程,2018,40(1):19-27.
ZHAO Zhiguo, WANG Chen, ZHANG Tong, et al. Anti-shake control with feed-forward correction and active damping control in Tip-In/Out phases of pure electric driving [J]. Automotive Engineering, 2018,40(1):19-24.

[6] 余卓平,冯源,熊璐,等.分布式驱动电动汽车动力学控制发展现状综述[J].机械工程学报,2013,49(8):105-111.
YU Zhuoping, FENG Yuan, XIONG Lu, et al. Review on vehicle dynamics control of distributed drive electric vehicle [J]. Journal of Mechanical Engineering, 2013, 49(8):105-111.

[7] 余志生.汽车理论[M].北京:机械工业出版社,2018:17-19.
YU Zhisheng. Automotive theory [M]. Beijing: Machinery Industry Press, 2018:17-19.

(上接第300页)

[12] GU X, SEE K W, LI P, et al. A novel state-of-health estimation for the lithium-ion battery using a convolutional neural network and transformer model[J]. Energy, 2023, 262: 125501.

[13] XIANG F, ZHANG Y, ZHANG S, et al. Bayesian gated-transformer model for risk-aware prediction of aero-engine remaining useful life [J]. Expert Systems with Applications, 2024, 238: 121859.

[14] JIN R, CHEN Z, WU K, et al. Bi-LSTM-based two-stream network for machine remaining useful life prediction[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-10.

[15] GAO H, LI Y, ZHAO Y, et al. Dual channel feature attention-based approach for RUL prediction considering the spatiotemporal difference of multisensor data [J]. IEEE Sensors Journal, 2023, 23(8): 8514-8525.

[16] ZHANG Z, SONG W, LI Q. Dual-aspect self-attention based on transformer for remaining useful life prediction[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-11.

[17] HOSSAIN S, UMER S, ROUT R K, et al. Fine-grained image analysis for facial expression recognition using deep convolutional neural networks with bilinear pooling[J]. Applied Soft Computing, 2023, 134: 109997.

[18] WANG D, WANG J, REN Z, et al. DHBP: a dual-stream hierarchical bilinear pooling model for plant disease multi-task classification [J]. Computers and Electronics in Agriculture, 2022, 195: 106788.

[19] JIN Y, HOU L, CHEN Y. A time series transformer based method for the rotating machinery fault diagnosis [J]. Neurocomputing, 2022, 494: 379-395.