高技术通讯 2024 年 第 34 卷 第 12 期:1330-1340

doi:10.3772/j.issn.1002-0470.2024.12.009

基于双维度注意力残差收缩网络的滚动轴承故障诊断①

陈昊杰② 陈永毅 倪洪杰 张 丹③

(浙江工业大学信息工程学院 杭州 310023)

摘 要 主动运维技术是机械装备提高维修效率、降低维修成本、保证长时间稳定运行的必要保障。本文以滚动轴承作为研究对象,提出了一种基于双维度注意力残差收缩网络的智能故障诊断方法。该方法首先通过引入卷积注意力模块,分别对振动信号在空间与通道维度上的加权赋值来强化重要的特征,使网络的注意力集中在对故障分类更关键的信息上。其次,通过基于混合域注意力的软阈值化模块实现数据降噪以及去除特征中的冗余信息。实验结果表明,所提出的方法能够有效实现故障诊断和分类,且具有良好的鲁棒性。

关键词 滚动轴承;双维度特征提取;软阈值;Swish 激活函数;深度学习

随着工业生产所使用的设备逐渐趋于大型化与自动化,机械装备的主动运维问题也成为了人们日益关注的重点。机械装备的零部件性能一旦退化至失效状态,轻则阻碍设备正常运行,重则造成重大安全事故。滚动轴承作为机械设备最为关键的部件之一,其工作状态正常与否直接影响整个机械设备的可靠性与安全性,关系到整个流水线的生产效率乃至整个企业的产出效率。实际工况中,长时间处于高负载和高强度工作环境下的滚动轴承一旦出现磨损或是异物侵入等问题,极有可能引发多种故障。倘若能够迅速检测出轴承发生的微小磨损并立即采取措施(如隔振、减振、增加润滑液等),便可以减少不必要的损失[1]。因此,开展滚动轴承故障诊断及识别技术研究对于提高机械设备的维修效率、降低维修成本,保证其长时间稳定运行等都具有重大的实际意义。

近年来,滚动轴承故障诊断领域针对振动信号、光信号以及温度信号等方面展开了大量工作。由于光信号和温度信号容易受外界环境因素的干扰,导致诊断结果出现差异化,检测精度下降,因此基于振动信号的故障诊断方法成为了当前的主流方法[1]。伴随着深度学习技术的兴起,基于深度学习的故障诊断方法在机械故障诊断领域得到了广泛的应用。深度学习方法通过构建深度神经网络模型,智能化地诊断出故障类别,并及时将结果反馈给维保人员,既节省了大量人力、物力、财力资源,也确保了生产的安全进行[2]。深度学习模型凭借其出色的特征提取和大数据处理能力在工程领域得到了广泛的应用,而通过卷积神经网络(convolutional neural networks,CNN)对信号进行特征提取则在近几年内成为了一种主流[3]。例如,Ince 等人[4]采用一维卷积神经网络提取信号特征以进行电机故障检测。李东东等人[5]设计了一种基于一维卷积神经网络和 Soft-Max 分类器相结合的双层智能故障诊断架构,能够有效地提取风电机组行星齿轮箱故障特征。Hu 和 Deng[6]提出一种基于无线传感器网络数据融合的滚动轴承故障诊断方法,利用无线传感器网络对机械设备关键部位的数据进行采集,从而改善了传统卷积神经网络故障检测算法采集数据不够准确的问题。

① 国家重点研发计划(2018YFB1403702),国家自然科学基金(61873237)和宁波市科技创新 2025 重大专项资助项目。
② 男,2002 年生,研究方向:基于深度学习的工业互联网装备运维;E-mail:hjchen2002@foxmail.com。
③ 通信作者,E-mail:danzhang@zjut.edu.cn。
(收稿日期:2023-08-17)

此外,基于深度学习思想的模型优化研究也是近几年内的热点方向之一。例如,Zhang 等人[7]提出一种首层宽卷积深度神经网络(deep convolutional neural networks with wide first-layer kernels,WDCNN)用于轴承故障诊断,并使用凯斯西储大学轴承数据的时域信号进行故障算法测试,取得了较好的故障分类结果。He 等人[8]提出了一种深度残差网络用以实现更深层次的神经网络训练。邱大伟等人[9]基于 Transformer 神经网络,结合双向输入样本矩阵处理机制以及训练过程中错误样本权重增强机制,实现了在时间效率以及准确率上的提升。陈永毅等人[10]基于集合经验模态分解(ensemble empirical mode decomposition,EEMD)的方法,在模型收敛速度以及诊断正确率方面均取得了较好的结果。Wu 等人[11]开发了一种深度强化转移卷积神经网络(deep reinforcement transfer convolutional neural network,DRTCNN),实现了利用有限的标注样本完成故障诊断任务的目标。Liu 等人[12]提出了一种小波胶囊生成对抗网络(wavelet capsule generative adversarial network,WCGAN),实现了数据集的平衡与深层特征提取。Wang 等人[13]结合马尔可夫转移场(Markov transition field,MTF)和卷积神经网络,通过将原始一维振动信号转化为具有时间相关性的二维特征图像,实现了对振动信号深层特征的提取。唐纪凯等人[14]提出一种基于同步压缩小波变换(synchrosqueezed wavelet transform,SWT)和卷积神经网络的诊断方法,通过对振动信号经 SWT 处理后得到的时频图进行特征提取,从而实现故障诊断。金江涛等人[15]基于混沌理论,提出采用相空间重构方法用于还原并丰富轴承振动的动力学特性,通过卷积神经网络提取混沌序列中的高级抽象特征,在模型的准确性、鲁棒性以及收敛速度等方面均得到了一定的提升。Peng 等人[16]针对滚动轴承振动信号中故障特征提取费力、分类精度差的问题,提出了一种基于多指标融合的故障特征选择与分类方法,该方法在不同实验数据集的特征提取和故障检测方面具有显著的优势。Zhang 等人[17]基于监督对比学习,提出了一种滚动轴承跨域故障诊断领域自适应网络(supervised contrastive learning based domain adaptation network,SCLDAN),该方法可以准确地识别位于目标域的类边界附近的模糊数据样本,并且显著地提高了诊断准确度。Song 等人[18]采用加宽卷积核的策略,提出了宽卷积核卷积神经网络(convolutional neural network with wide convolution kernels,WKCNN)模型,取得了较好的性能提升。Wang 等人[19]针对实际工况中噪声的干扰,提出了一种用于滚动轴承故障诊断的深度特征增强强化学习方法,在模型准确性以及鲁棒性等方面取得了一定的成果。

实际工况中所采集的原始振动信号,在包含大量的数据信息的同时存在一定的冗余与噪声信号,而现有方法在划分并提取其重要信息与冗余信息方面仍存在一些不足之处。如何有效鉴别振动信号中的重要特征与冗余信息,强化有效特征的表达,并将无用的冗余信息去除,避免其对网络模型准确判断造成干扰,是当前故障诊断算法研究中仍未有效解决的挑战性问题。

针对上述问题,本文提出一种基于双维度注意力残差收缩网络的滚动轴承故障诊断方法。首先将原始振动信号输入至卷积注意力机制模块中进行通道、空间维度下的双维度注意力特征提取;之后将提取到的深层特征输入至深度残差收缩网络模块(deep residual shrinkage network block,DRSB)[20]进行去除冗余等操作后进行故障分类;最后,在浙江工业大学电机轴承故障平台上进行了实验验证,结果证明了所提方法的优越性。

1 滚动轴承故障诊断模型

1.1 整体网络模型

本文构建了基于双维度注意力残差收缩网络(dual dimensional attention residual shrinkage network,DDARSN)的滚动轴承故障诊断模型,其具体网络结构如图 1 所示。该模型选用卷积核大小为 64 × 1 的一维卷积层作为模型输入的第 1 层,在达成降低数据维度的同时也起到了将所需的异常波动全部截获的效果。模型的主要核心在于中间的双维度注意力机制特征提取网络和深度残差收缩网络模块,同时

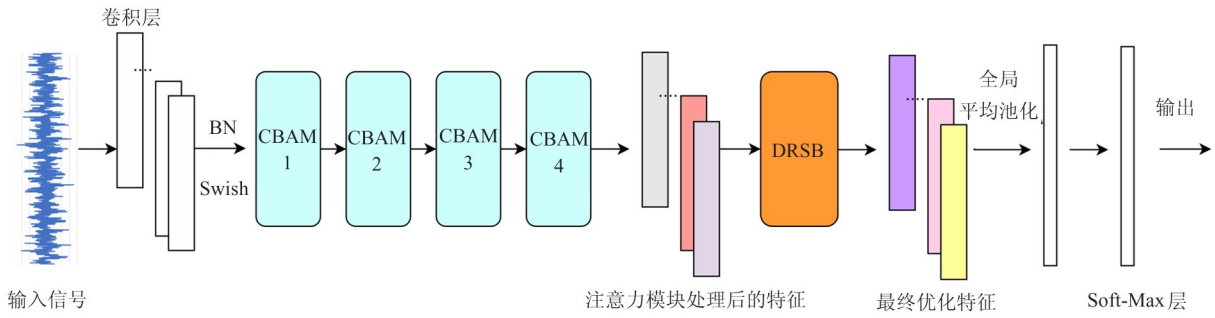

图 1　DDARSN 模型网络结构

选取 Swish 函数作为模型的主要激活函数也是本文方法的一项重要环节。下面将对所提出的 DDARSN 的关键模块作进一步详细说明。

1.2　双维度注意力机制

信号特征提取是滚动轴承故障诊断领域的重要环节之一。滚动轴承原始振动信号在经过模型首层宽卷积核网络层处理后,尚未实现深层特征的提取。基于人类视觉仿生的注意力机制则可以通过分配权重达成关注重点目标区域,抑制其他无用信息的效果,从而实现神经网络对滚动轴承故障关键特征的针对性学习[21]。

在 CNN 中,不同通道所表达的信息具有明显的差异,但实际只有少数通道包含的特征对故障诊断结果影响较大。通过引入一种通道注意力(channel attention)机制,可以自适应地校准信道特征响应,并计算权重以实现有用信道的选择和强调,提取到振动信号中的关键故障信息。其原理图如图 2 所示。

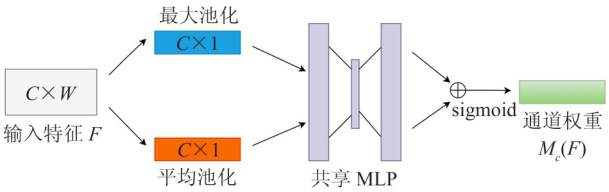

图 2　通道注意力原理

如图 2 所示,给定一个特征图 F 作为输入特征,其通道维度为 C。为了有效计算通道注意力,首先在空间维度上使用全局最大池化(global max pooling,GMP)和全局平均池化(global average pooling,GAP)对输入特征 F 进行压缩处理。并行连接的处理方式相较于单一的池化处理丢失的信息更少,所计算出的最终通道注意力特征也更为精细[22]。使用上述方法聚合输入特征的空间信息,得到 2 个分别表示最大池化和平均池化的特征向量 $\boldsymbol{F}^c_{\max}$ 和 $\boldsymbol{F}^c_{\text{avg}}$,再将这 2 个特征向量输入至共享多层感知器(multi-layer perceptron,MLP)中。在该共享网络中,为了减少参数量,第 1 个隐藏层的神经元数被设置为 C/r[22],激活函数为 ReLU,其中 r 代表下降率(reduction ratio)。在第 2 个隐藏层中,神经元数再恢复为 C。随后将输出得到的 2 个特征向量进行逐元素求和操作,最后通过 sigmoid 函数进行归一化以得到各通道的注意力权重。通道注意力的计算公式为

$$M_C(F) = \sigma(W_1(W_0(\boldsymbol{F}^c_{\text{avg}})) + W_1(W_0(\boldsymbol{F}^c_{\max}))) \tag{1}$$

其中,F 表示输入特征;$M_C(F)$ 表示输出的通道注意力权重;σ 表示 sigmoid 函数;W_0 和 W_1 代表 2 个隐藏层的参数,且 $W_0 \in R^{C/r \times C}$,$W_1 \in R^{C \times C/r}$。

与通道维度相似,在特征空间中,对任务相关的区域(即空间维度)投入更多的关注有助于更好地实现故障诊断。空间注意力(spatial attention)机制通过重新标定特征图在空间位置上的权重分布,实现有用信息权重的自适应提高[23]。其原理如图 3 所示。首先基于输入特征的通道轴对其进行最大池化和平均池化操作,得到 2 个特征向量 $\boldsymbol{F}^s_{\max}$ 和 $\boldsymbol{F}^s_{\text{avg}}$,将它们沿通道方向拼接起来生成一个特征描述符;其次使用卷积操作将其映射为空间注意力图;最后通过 sigmoid 函数进行标准化得到最终注意力权重。空间注意力的计算公式为

$$M_S(F) = \text{CONV}(\{\boldsymbol{F}^s_{\max};\ \boldsymbol{F}^s_{\text{avg}}\}) \tag{2}$$

其中,$M_S(F)$ 表示空间权重,CONV 表示卷积操作。

本文设计的双维度注意力模块通过串行连接将

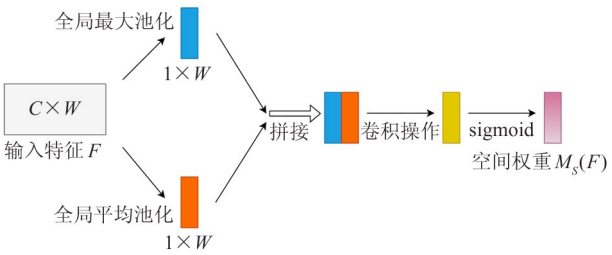

图 3　空间注意力原理

通道注意力和空间注意力集成在一起,使其能够同时兼顾并强化数据集在这 2 个维度上的重要信息,实现对振动信号双维度下的深层特征提取,在获得更多重要信息表达的同时达成了较好的模型性能提升。如图 4 所示,首先对输入特征 F 进行 2 次步长分别为 2 和 1 的卷积运算,其主要目的在于对输入特征作一个初步的精细化处理,以便于网络在后续

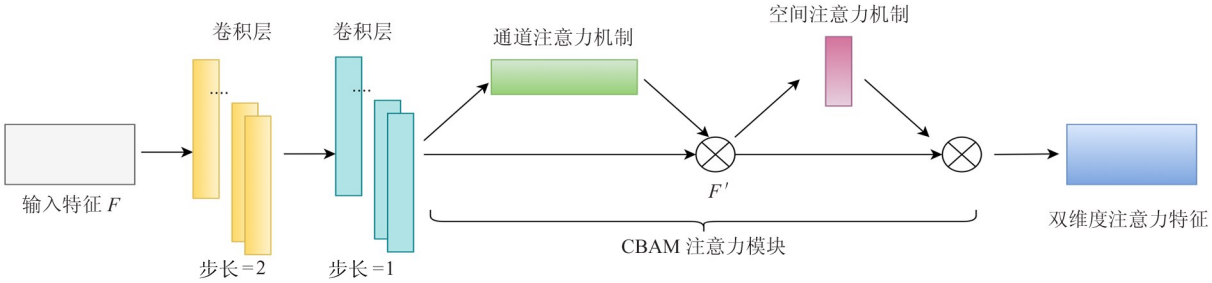

图 4　双维度注意力模块

特征提取上的准确率以及计算效率方面得到一定的提升;随后将得到的输出特征输入至通道域和空间域注意力模块。其中,通道注意力机制计算出通道注意力权重并与卷积层的输出作加权乘法得到中间特征 F';F'再经空间注意力机制加权得到最终的双维度注意力特征。经实验证明,相较于二者并行的连接方式,CBAM(convolutional block attention module)[22]注意力机制使用的先通道后空间的连接方式,在特征提取方面拥有更高的准确率。

在本文所构建的由 4 个双维度注意力模块组成的混合域特征提取网络中,定义其通道注意的缩减率数值恒为 2,每一个模块的卷积层的卷积核数目则逐步增加,从而实现特征提取量的进一步提升。

1.3　深度残差收缩网络

经过 4 层特征提取模块处理后所得到的深层特征,不可避免地含有部分噪声与无用冗余信息,若直接对提取的特征进行分类操作,将在一定程度上对模型的精度造成影响。因此,本文随后引入一个深度残差收缩网络[20]作为该深层特征的处理层。深度残差收缩网络是一种面向含噪声数据的深度学习方法,在保留深度残差网络的模型训练优势上,引入软阈值化模块对输入数据进行降噪处理,从而获得更为优秀的模型[20,24-25]。其基本原理如图 5 所示。

从结构上可见,输入特征经过几次批标准化

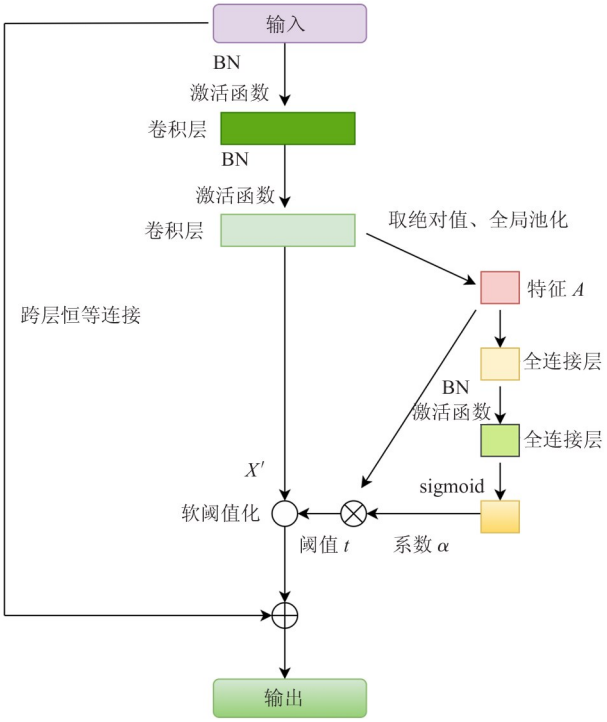

图 5　深度残差收缩网络模块

(batch normalization,BN)、激活函数以及卷积层的处理后得到一个新特征 x',该特征随后被送入到包含一个小型子网络的阈值计算模块。在该模块中,首先对输入特征求绝对值,然后经过全局均值池化获得一个特征 A。在另一条路径中,全局均值池化之后的特征图,被输入到一个小型的全连接网络。

这个全连接网络以 sigmoid 函数作为最后一层,将输出归一化到 0 和 1 之间,获得一个系数 α。最终的阈值 t 可以表示为 $\alpha \times A$。

软阈值化结果如式(3)所示[26]:

$$y = \begin{cases} x' - t & x' > t \\ 0 & -t \leqslant x' \leqslant t \\ x' + t & x' < -t \end{cases} \tag{3}$$

其中,y 代表软阈值化结果,x'代表软阈值化模块的输入特征,t 代表输入对应的阈值正参数。

将软阈值模块嵌入残差网络中,使其通过注意力机制根据各个样本的情况自适应地设置阈值,避免了人工设置阈值所需的繁琐计算过程。通过这种方式,每个样本都能够获得一组阈值,使得该模块适用于各个样本的噪声以及冗余信息含量不同的情况。结合所得阈值对所提取的深层特征作软阈值化处理,从而保留输入数据中对目标任务重要的特征,删除不重要的冗余和噪声信息,使轴承数据在不同状态下的特征区分更加清晰,完成最终的特征选择。

通过堆叠一定数量的基本模块以及卷积层、批标准化、激活函数、全局均值池化以及全连接输出层等,便可构建一个专门用于处理噪声数据以及冗余信息的深度残差收缩网络模块,其中的软阈值化功能模块能够自动根据输入数据的噪声或冗余含量设定不同的阈值,从而使模型具有一定的鲁棒性。

1.4　Swish 激活函数

在深度卷积神经网络中,激活函数是模型训练过程中不可或缺的一部分。其主要作用是引入非线性元素,解决线性模型表达、分类能力不足的问题,为整个模型提供充足的非线性扭曲力,使得神经网络能够任意逼近任何非线性函数,进而帮助网络更好地理解和拟合复杂函数模型,完成计算任务[27]。

由于本实验使用的滚动轴承振动信号数据不乏负区间数值,为了避免大量数据被 ReLU 函数屏蔽以及更好地提升网络性能,本文引入 Swish 激活函数[28]用于替代本文主要网络模块所默认使用的 ReLU 函数。其定义表达式为

$$Swish(x) = x \cdot \frac{1}{1 + e^{-x}} \tag{4}$$

其中,x 表示输入,e 表示自然常数,函数曲线如图 6 所示。

由式(4)及图 6 可知,Swish 函数结合了 sigmoid 函数与线性函数,在保留 sigmoid 函数平滑性特征的基础上,线性函数的引入使得该函数获得了负下界。对比 ReLU 函数,Swish 函数在具备上无界特性的同时显得更为平滑。上无界性可以有效避免慢速训练期间,梯度逐渐接近 0 并导致饱和的现象;平滑性则在优化和泛化中起了重要作用;在负区间输出不为 0 也很好地解决了 ReLU 的神经元死亡问题;同时 Swish 具有下边界,这使其对负数噪声具有较高的鲁棒性[20]。

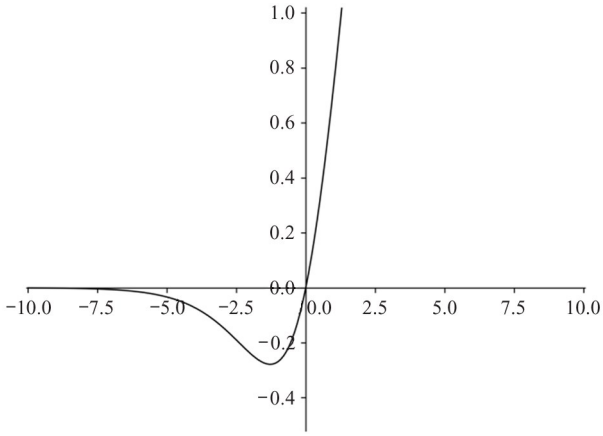

图 6　Swish 函数图像

1.5　故障诊断流程

本文所构建的 DDARSN 模型故障诊断流程如图 7 所示。首先将滚动轴承振动信号数据划分为训练集和测试集,训练集数据经预处理后输入至 DDARSN 模型;经过首层宽卷积核网络层作初步的特征提取,进入由 4 个双维度注意力模块组成的混合域特征提取网络实现双维度下的深层特征提取;将提取到的深层特征输入至深度残差收缩网络模块进行去除冗余等操作后进行故障分类。其次采用训练集对模型进行多次训练,经一定数量迭代后得到训练好的网络模型,模型输出为预测标签值。然后利用训练好的网络模型对测试集进行故障诊断,通过故障分类准确率和召回率来衡量模型的性能是否达标。特别地,故障分类准确率表示测试集中故障样本被正确分类数目占总样本的比例,召回率指在某一类故障类别下,预测对的正样本在在整个真实

正样本中的比例。最后对所有类别的召回率做算术平均计算。

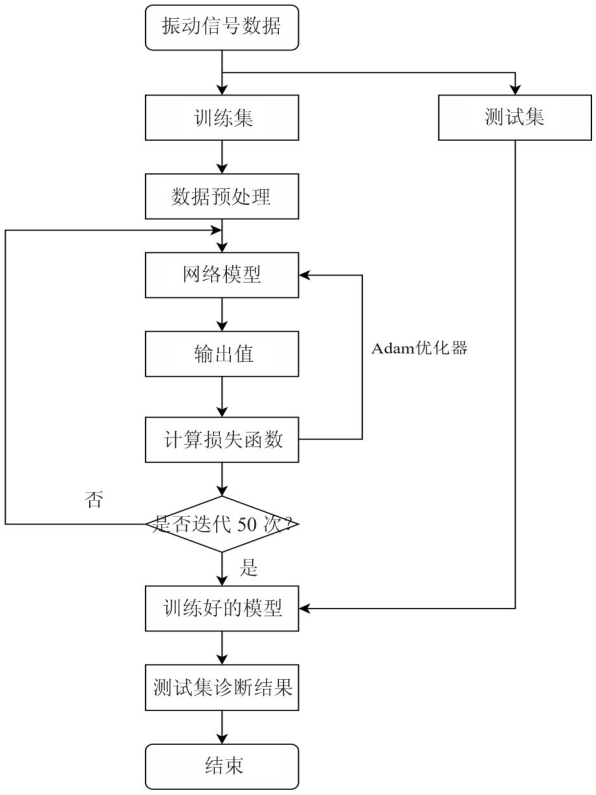

图 7　基于 DDARSN 模型的滚动轴承故障诊断流程图

2　模型训练方法及其主要参数

本文实验在 Windows 10 操作系统下进行,选用 i5-1035G1 处理器和 NVIDIA GeForce MX330 显卡,实验算法则由 Python 3.8 下的 keras 2.9.0 框架实现。

进行模型训练前,先对数据集进行预处理,由于数据标签代表的是各故障类别,其作为离散特征的取值之间没有大小的意义。因此将其标签转化为独热码(one-hot)的形式,使得标签输出为 0 ~ 9 的标准格式以便于后续运算的进行。相应地,本文采用交叉熵损失函数(cross entropy loss function)计算模型损失值。模型训练具体方法则选用 Adam 梯度下降法,初始学习率为 0.001,该方法参数的更新不受梯度的伸缩变换影响,可以在训练过程中自动调整学习率,使得计算更为高效。

本文模型主要由一维神经网络搭建而成,其中各网络模块的一维卷积层相关参数如表 1 所示。算法流程如算法 1 所示。

表 1　各模块一维卷积层相关参数

编号	网络层	卷积核大小	数目	步长
C1	卷积层 1	64 × 1	16	16
CBAM1	卷积层 2-1	3 × 1	16	2
	卷积层 2-2	3 × 1	16	1
CBAM2	卷积层 3-1	3 × 1	32	2
	卷积层 3-2	3 × 1	32	1
CBAM3	卷积层 4-1	3 × 1	64	2
	卷积层 4-2	3 × 1	64	1
CBAM4	卷积层 5-1	3 × 1	128	2
	卷积层 5-2	3 × 1	128	1
DRSB	卷积层 6-1	3 × 1	128	1
	卷积层 6-2	3 × 1	128	1

算法 1　DDARSN 模型算法

Input:
X_T:训练集的一维轴承振动信号
Y_T:训练集的标签
N: 训练迭代总次数
t: 当前迭代次数
s: 训练批量大小
Output: DDARSN 故障诊断模型
Repeat
　for 来自 X_T 中的 s 个样本:do
　　首层宽卷积核与双维度注意模块提取深层特征;
　　DRSB 模块中软阈值模块删除上述特征的冗余信息;
　　计算 Soft-Max 层发生每个故障的概率;
　　通过交叉熵损失函数计算交叉熵损失;
　end for
Until: 迭代次数 t 达到 N

3　实验及结果分析

3.1　实验数据集

本实验电机平台如图 8 所示,无线智能传感器 CMSONE-WMS1 放置在电机风扇端轴承座上方用于振动信号采集,电机转速控制器用于调节电机转速,DWZ-3 盘式涡流制动器则用于实现扭转载荷的调节。本实验使用了正常轴承和包含 9 种不同损伤类型的轴承。轴承的损伤位置包括滚球、外圈、内圈 3 处,每处位置的损伤程度按照其损伤直径尺寸将其细分为 0.1 mm、0.3 mm、0.5 mm 3 类。轴承具体损

伤位置与程度如表 2 所示。数据采集环节中,将不同故障状态的滚动轴承依次安装到电机中,轴承座上方的无线智能传感器以 16 000 Hz 的采样频率采集振动信号。

实验过程中,通过调节电机转速与负载,分别设置了如下 3 种工况。

工况 1:转速 20 Hz,扭转载荷 0 A;

工况 2:转速 30 Hz,扭转载荷 1 A;

工况 3:转速 40 Hz,扭转载荷 2 A。

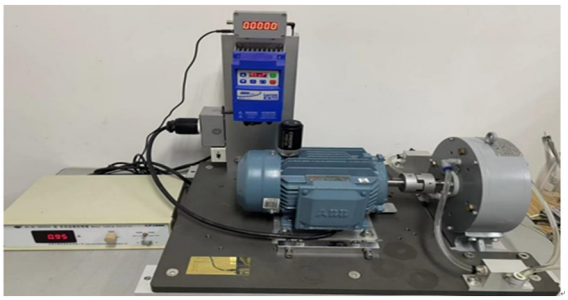

图 8　滚动轴承数据采集实验台

表 2　滚动轴承故障标签

标签	故障位置	故障直径尺寸/mm
0	无	无
1	滚球	0.1
2	滚球	0.3
3	滚球	0.5
4	外圈	0.1
5	外圈	0.3
6	外圈	0.5
7	内圈	0.1
8	内圈	0.3
9	内圈	0.5

数据集编号如表 3 所示。其中每种故障类的采样时间为 2 min。最终在每种故障类型中选取了 1 200个样本,总共 12 000 条数据。在后续的模型训练中,选取 10 500 条样本的数据集作为训练集,1 500条样本的数据集则作为测试集。实际电机平台采集到的振动信号样本如图 9 所示。

表 3　数据集编号

转速/Hz	扭转载荷/A	数据集编号
20	0	1
30	1	2
40	2	3

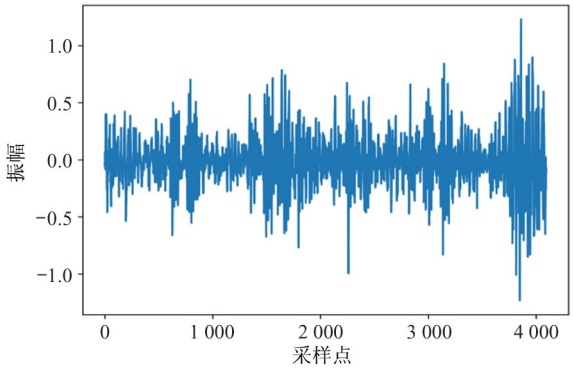

图 9　实验平台采集的振动信号样本

3.2　模型性能评估

利用训练好的网络模型,对 10 种不同状态的滚动轴承数据进行故障检测仿真实验。选用倒三角和方块用于区分训练集和验证集曲线,图 10 为训练集和测试集的准确率随迭代次数变化的结果图,可以看到当迭代次数为 10 时,准确率趋于最大值,随后趋于稳定。以交叉熵损失函数作为损失函数,得到训练集和测试集的损失率曲线如图 11 所示。

为全面验证本文模型的有效性,在相同条件下,将本文提出的 DDARSN 模型与几种现有深度学习方法进行比较。在实验过程中,在测试集的原始数据中加入 −4 ~ 6 dB 不同信噪比的噪声信号用以模拟不同的工作环境,随后将其输入至训练好的模型进行对比实验。信噪比(signal to noise ratio,SNR)定义表达式如下:

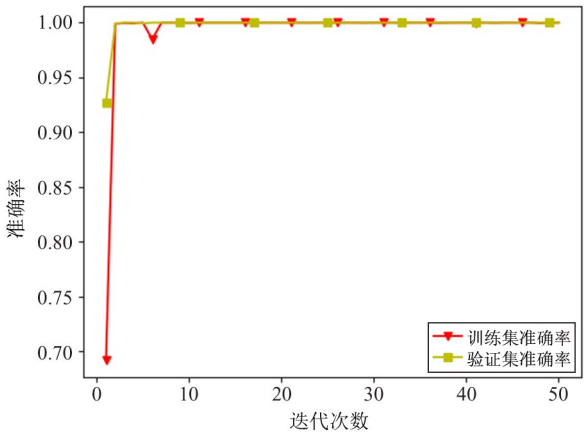

图 10　故障准确率曲线

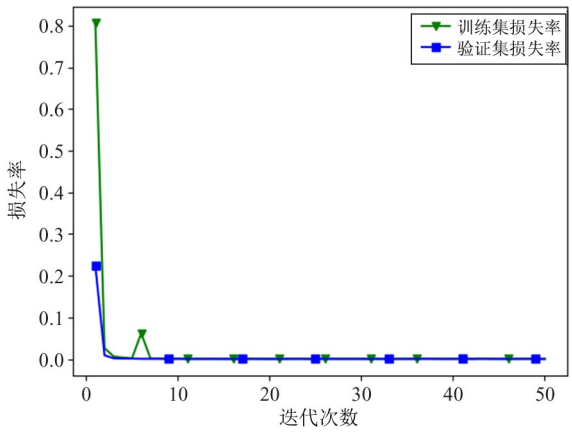

图 11　故障损失率曲线

$$SNR = 10\log_{10}\left(\frac{s}{n}\right) \tag{5}$$

其中 s 和 n 分别表示信号和噪声的有效功率。

针对本文所提出的问题,选取了如下对比方法。

MTF-CNN[13]:该方法利用马尔可夫转移场首先对采集到的振动信号进行编码,并生成相应特征图;之后将其输入二维卷积神经网络完成特征的提取并进行故障分类。在保证网络参数与本文 DDARSN 模型基本一致的前提下,将上述 2 种方法的卷积核大小设置为 3×3,卷积核数目以及步长等参数则与表 1 保持一致。

STW-CNN[14]:该方法通过同步压缩小波变换方法获得振动信号时频图,随后将频域图输入至二维卷积神经网络完成后续故障分类等操作。该二维神经网络参数与 MTF-CNN 保持一致。

WKCNN[18]:该网络结构与 WDCNN 基本相似,但其第 1 层卷积核大小设置为 64×1,第 2 层卷积核大小为 16×1,此后添加的卷积层核大小为 5×1,卷积核数目以及步长等参数则与表 1 保持一致。

DFEQN (deep feature enhanced Q network)[19]:该方法结合了注意力模块与强化学习机制,实现了强化数据集的重要特征与忽略噪声与冗余信号的功能,具有较好的故障识别精度。对其深度 Q 网络(deep Q network,DQN)模块中的 CNN,设置其卷积核大小为 3×1,同样按照表 1 的参数构建网络层。

WDCNN[7]:其第 1 卷积层卷积核大小设置为 64×1,通过添加 3×1 个小卷积核的多个卷积层以及对应的池化层来加深网络。

ResNet(residual network)[8]:为了使比较更加公平合理,该模型以本文 DDARSN 模型增加通道数的方式叠加网络层。

各模型的训练损失函数、训练方法以及迭代次数均与本文模型保持一致。

在本文设置的 3 种工况下,上述模型的准确率及召回率如表 4~6 所示。针对表中 7 种模型所具备的不同特征,下文将结合其实验结果进行分析。

实验结果表明,WDCNN 和 ResNet 2 种方法相较于其他几种对比方法存在一定差距,且其性能在 SNR 值低于 0 dB 的情况下出现了大幅下降。原因在于,上述二者相比其他几种方法缺少如注意力机制、图像编码等方法的辅助,难以准确提取信号的重要特征,且这种缺陷在噪声环境的影响下进一步得到放大;MTF-CNN 与 STW-CNN 在正常环境下表现都十分优秀,但随着噪声干扰强度的增加,其性能也在一定程度上受到影响。这是由于采集振动信号时,噪声信号会不可避免地被导入到时频图像中,从而影响到后续特征提取;WKCNN 在噪声干扰较小的环境下有着较好的表现,但其性能同样随着噪声干扰的增加而有所下降,这表明单纯提升卷积核宽度尚不足以应对噪声环境;DFEQN 在不同噪声环境下均发挥了优秀的效果,但该方法主要通过注意力机制实现重要特征提取,忽略输入中的噪声和冗余,这尚不足以应对噪声干扰过强的环境。因此在 $SNR = -4$ dB 的情况下,该方法略逊于本文方法。

本文所提出的 DDARSN 模型,选用一维网络作为主要框架,这避免了将一维振动信号转换为二维图像过程中所可能造成的破坏原始序列中空间相关性、丢失重要的特征信息等影响。这一点在与 MTF-CNN 与 STW-CNN 二者的实验对比结果中得到了充分体现;双维度注意力模块的引入,使得模型能够同时兼顾并强化数据集在通道与空间 2 个维度上的重要信息,充分实现了在通道与空间维度上对振动信号提取并强调更有效的信息特征,忽略噪声及其他无用信息的干扰,提升了模型的精度与鲁棒性;通过引入软阈值模块,自适应地学习所提取的重要特征并获取相应的参数,将其作为阈值进行软阈值化处理,从而实现冗余以及噪声信号的删除,完成最终的特征选择。由于冗余信息的去除,模型在低 SNR 环境下的精度也得到有效提升,实验结果显示,在SNR

表 4　工况 1 的实验结果(%)

SNR/dB	模型准确/召回率						
	WDCNN	ResNet	MTF-CNN	STW-CNN	WKCNN	DFEQN	DDARSN
-4	78.01/76.22	75.63/72.38	90.99/88.11	87.22/84.99	84.32/83.79	99.38/99.15	99.68/99.64
-2	90.82/88.22	87.85/85.41	92.30/91.58	90.82/89.26	91.69/90.92	99.62/99.60	99.72/99.70
0	96.11/95.35	91.64/90.88	98.78/98.72	97.09/96.47	97.81/97.20	99.64/99.64	99.82/99.82
2	97.60/97.56	93.19/92.64	99.33/98.92	97.23/96.68	98.10/98.10	99.69/99.69	99.87/99.87
4	98.33/98.33	94.77/94.77	99.36/99.31	97.30/97.29	98.94/98.94	99.72/99.72	99.88/99.88
6	98.72/98.72	95.28/95.28	99.46/99.46	99.02/99.02	99.16/99.16	99.77/99.77	99.88/99.88

表 5　工况 2 的实验结果(%)

SNR/dB	模型准确/召回率						
	WDCNN	ResNet	MTF-CNN	STW-CNN	WKCNN	DFEQN	DDARSN
-4	78.51/76.63	76.53/72.68	90.81/88.07	87.29/85.07	85.16/84.22	99.37/99.15	99.70/99.64
-2	90.62/88.17	88.36/85.70	92.20/91.43	90.85/89.30	92.07/91.55	99.60/99.60	99.77/99.76
0	96.25/95.41	91.71/90.99	98.69/98.55	97.15/96.52	97.90/97.40	99.63/99.64	99.83/99.83
2	98.10/97.24	93.29/92.73	99.29/98.84	97.24/96.71	98.21/98.21	99.68/99.68	99.87/99.87
4	98.88/98.40	94.80/94.80	99.35/99.28	97.31/97.29	98.96/98.96	99.70/99.70	99.91/99.91
6	98.90/98.54	95.31/95.31	99.45/99.45	99.05/99.05	99.19/99.19	99.72/99.72	99.91/99.91

表 6　工况 3 的实验结果(%)

SNR/dB	模型准确/召回率						
	WDCNN	ResNet	MTF-CNN	STW-CNN	WKCNN	DFEQN	DDARSN
-4	79.30/76.89	75.58/72.69	90.77/88.50	87.42/84.99	83.87/82.59	99.43/99.24	99.66/99.64
-2	90.42/88.64	89.11/87.82	92.40/91.60	90.60/89.26	91.39/90.52	99.61/99.60	99.75/99.75
0	96.34/95.51	91.82/90.98	98.88/98.81	96.83/96.47	97.78/97.10	99.64/99.64	99.80/99.80
2	97.60/96.14	93.41/92.76	99.38/98.97	97.13/98.68	98.00/98.00	99.74/99.74	99.85/99.85
4	98.33/98.33	94.82/94.82	99.41/99.31	97.19/97.29	98.72/98.72	99.74/99.74	99.88/99.88
6	98.72/98.72	95.30/95.30	99.47/99.47	98.12/98.12	98.96/98.96	99.74/99.74	99.90/99.90

值达到 -4 dB 的条件下,对比方法的性能均在一定程度上受到了影响,但是本文模型则依然维持着较好的精度。

综上所述,实验结果证明,本文方法在不同噪声环境下都表现出了很大的优势。在信号含噪量较小的条件下,相比现有深度学习模型得到了约 1% ~ 8% 的提升。在信号受噪声干扰较为严重时,其准确率与召回率也依旧维持在一个较高的水准,这表明本文所采用方法具备良好的性能与鲁棒性。

3.3　混淆实验

考虑到本文选用 Swish 激活函数替代 ReLU 激活函数,为证明其有效性,设置了一组以激活函数为混淆变量的混淆实验。在工况 1 的数据集下,选用 DDARSN 作为验证模型,计算在不同信噪比噪声影响下模型的故障诊断准确率。具体数据见表 7。

表 7　激活函数的混淆实验

SNR/dB	模型准确率/%	
	ReLU	Swish
-4	97.92	99.68
-2	98.24	99.72
0	98.67	99.82
2	99.14	99.87
4	99.45	99.88
6	99.60	99.88

如表 7 所示,使用 Swish 激活函数后,故障诊断准确率整体相比使用 ReLU 函数的组别表现更为优秀。而在信噪比为负值时,Swish 相比 ReLU 表现出了更为明显的优势,这也反映了 Swish 对于负数噪声具有更高的鲁棒性。

同理,选取工况 1 的数据,设置一组以深度残差收缩模块为混淆变量的实验。控制激活函数为 Swish,计算在不同信噪比噪声影响下模型的故障诊断准确率,具体数据见表 8。

表 8　DRSB 的混淆实验

SNR/dB	模型准确率/%	
	无 DRSB	DDARSN
-4	85.62	99.68
-2	93.20	99.72
0	96.73	99.82
2	97.22	99.87
4	98.17	99.88
6	98.91	99.88

如表 8 数据所示,DRSB 模块的引入使得模型在正常环境下整体准确率上升了约 2% 左右,在噪声干扰环境下本文模型则表现出了更为优秀的抗噪性能。由此得以证明在本文模型中使用 DRSB 模块的必要性。

另外,通过更改工况 1 训练集与测试集比例,本文额外设置了一组对比实验,结果如表 9 所示。

如表 9 数据所示,更改训练集与测试集比例之后,模型的准确率依然维持在一个较高的水准,这在一定程度上也反映了所提方法的优越性。

表 9　数据划分比例对比实验

训练集与测试集比例	模型准确率/%
7∶1	99.82
7∶3	99.64
8∶2	99.77

4　结　论

在滚动轴承故障诊断过程中,对其振动信号特征的提取和优化对整个故障诊断系统精确度的提高和复杂度的减小具有重要意义。本文提出的 DDARSN 模型,通过注意力机制对振动信号进行双维度下的深层特征提取,随后进一步利用深度残差收缩网络的软阈值化模块删除冗余信号与噪声信号。使用 Swish 激活函数优化各网络模块的策略经本文实验也验证了其优越性。最后与现有方法对比,表明本文方法能够有效实现不同的故障类型的诊断和分类。在今后工作中,将针对信号预处理方法以及小样本情况下的诊断方法作出进一步研究,使得模型得以应对更为复杂多变的环境。

参考文献

[1] 闵浩,刘惠康,唐超. 基于间谐波的双馈风力发电系统故障诊断[J]. 高技术通讯,2019,29(5):480-486.

[2] 许勇,蔡云泽,宋林. 基于数据驱动的核电设备状态评估研究综述[J]. 上海交通大学学报,2022,56(3):267-278.

[3] DING X X, HE Q B. Energy-fluctuated multiscale feature learning with deep convnet for intelligent spindle bearing fault diagnosis[J]. IEEE Transactions on Instrumentation and Measurement, 2017, 66(8): 1926-1935.

[4] INCE T, KIRANYAZ S, EREN L, et al. Real-time motor fault detection by 1D convolutional neural networks[J]. IEEE Transactions on Industrial Electronics, 2016, 63(11): 7067-7075.

[5] 李东东,王浩,杨帆,等. 基于一维卷积神经网络和 Soft-Max 分类器的风电机组行星齿轮箱故障检测[J]. 电机与控制应用,2018,45(6):80-87.

[6] HU J, DENG S. Rolling bearing fault diagnosis based on wireless sensor network data fusion[J]. Computer Communications, 2022, 181: 404-411.

[7] ZHANG W, PENG G, LI C, et al. A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals[J]. Sensors, 2017, 17(2): 425.

[8] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016: 770-778.

[9] 邱大伟,刘子辰,周一青,等. 基于 Transformer 神经网络的滚动轴承故障类型识别[J]. 高技术通讯,2021,31(1):1-11.

[10] 陈永毅,张丹,薛焕新. 基于集合经验模态分解的滚动轴承智能故障诊断方法[J]. 高技术通讯,2021,31(5):559-568.

[11] WU Z, JIANG H, LIU S, et al. A deep reinforcement transfer convolutional neural network for rolling bearing fault diagnosis[J]. ISA Transactions, 2022, 129: 505-524.

[12] LIU Y, JIANG H, LIU C, et al. Data-augmented wavelet

capsule generative adversarial network for rolling bearing fault diagnosis[J]. Knowledge-Based Systems, 2022, 252: 109439.

[13] WANG M, WANG W, ZHANG X, et al. A new fault diagnosis of rolling bearing based on Markov transition field and CNN[J]. Entropy, 2022, 24(6): 751.

[14] 唐纪凯,卢一相,柏壮壮,等. 基于同步压缩小波变换和 CNN 的滚动轴承故障诊断[J]. 传感器与微系统, 2022,41(6):130-133.

[15] 金江涛,许子非,李春,等. 基于深度学习与混沌特征融合的滚动轴承故障诊断[J]. 控制理论与应用, 2022,39(1):109-116.

[16] PENG C, OUYANG Y, GUI W, et al. A multi-indicator fusion-based approach for fault feature selection and classification of rolling bearings[J]. IEEE Transactions on Industrial Informatics, 2023,19(8):8635-8643.

[17] ZHANG Y, REN Z, ZHOU S, et al. Supervised contrastive learning based domain adaptation network for intelligent unsupervised fault diagnosis of rolling bearing. IEEE/ASME Transactions on Mechatronics, 2022, 27(6):5371-5380.

[18] SONG X, CONG Y, SONG Y, et al. A bearing fault diagnosis model based on CNN with wide convolution kernels[J]. Journal of Ambient Intelligence and Humanized Computing, 2022,13(8):4041-4056.

[19] WANG R, JIANG H, ZHU K, et al. A deep feature enhanced reinforcement learning method for rolling bearing fault diagnosis[J]. Advanced Engineering Informatics, 2022,54: 101750.

[20] ZHAO M, ZHONG S, FU X, et al. Deep residual shrinkage networks for fault diagnosis[J]. IEEE Transactions on Industrial Informatics, 2019,16(7):4681-4690.

[21] 李国强,付乐,查琳琳,等. 基于多特征融合注意力机制的交通标志检测[J]. 高技术通讯,2022,32(11): 1178-1187.

[22] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision. Munich, Germany: Springer, 2018:3-19.

[23] 张万玉,肖锋,张文娟,等. 基于池化的双维度视觉注意力模型[J]. 西安工业大学学报,2021,41(6): 673-683.

[24] 车畅畅,王华伟,倪晓梅,等. 基于深度残差收缩网络的滚动轴承故障诊断[J]. 北京航空航天大学学报, 2021,47(7):1399-1406.

[25] CHEN Y, ZHANG D, ZHANG H, et al. Dual-path mixed-domain residual threshold networks for bearing fault diagnosis[J]. IEEE Transactions on Industrial Electronics, 2022,69(12):13462-13472.

[26] DONOHO D L. Denoising by soft-thresholding[J]. IEEE Transactions on Information Theory, 1995,41(3):613-627.

[27] 杜圣杰,贾晓芬,黄友锐,等. 面向 CNN 模型图像分类任务的高效激活函数设计[J]. 红外与激光工程, 2022,51(3):483-491.

[28] RAMACHANDRAN P, ZOPH B, LE Q V. Searching for activation functions[EB/OL]. (2017-10-16)[2023-06-09]. http://arxiv.org/pdf/1710.05941.

Fault diagnosis of rolling bearing based on dual-dimensional attention residual shrinkage network

CHEN Haojie, CHEN Yongyi, NI Hongjie, ZHANG Dan

(College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023)

Abstract

Active operation and maintenance technology is a necessary guarantee for mechanical equipment to improve its maintenance efficiency, reduce its maintenance cost, and ensure its long-term stable operation. In this paper, an intelligent fault diagnosis method based on dual dimensional attention residual shrinkage network is proposed, taking rolling bearing as the research object. In the proposed method, firstly, a convolution attention module is introduced to weight the processed signal in its spatial dimension and channel dimension to strengthen important features, focusing the network's attention on information which is more critical for fault classification. Secondly, the functions of data denoising and removing redundant information in features are realized through an adopting soft thresholding module based on mixed-domain attention. Experimental results show that the proposed method can realize the effective diagnosis and classification of rolling bearing faults, and has good robustness.

Key words: rolling bearing, dual-dimensional feature extraction, soft threshold, Swish activation function, deep learning