

一维卷积神经网络在机械故障特征提取中的可解释性研究

王芳珍¹, 张小丽¹, 赵琦武¹, 王保建²

(1. 长安大学道路施工技术与装备教育部重点实验室, 710064, 西安;
2. 西安交通大学机械基础国家级实验教学示范中心, 710049, 西安)

摘要: 针对一维卷积神经网络在机械故障诊断中的内部决策和推断过程未知, 导致结果可解释性与可信度不足的问题, 从信号特征提取的视角建立信号分析与神经网络之间的相似性联系, 通过提取神经网络卷积层权重, 观察信号时域、频域特征随网络层的变化规律, 从而揭示神经网络特征提取的本质, 并采用实验测试数据和凯斯西储大学轴承公开数据进行验证。结果表明: 卷积核可以等效为有限脉冲滤波器, 最大池化层能够满足简单二分类任务中神经网络的非线性化要求, 此时的卷积层无需添加激活函数; 神经网络能够通过逐层提高频率分辨率, 寻找到接近理论故障特征频率的频率成分, 此行为与傅里叶变换存在相似性; 当频谱范围最终分解到 1~3 倍故障特征频率时, 能够更好地完成识别任务。该研究可为揭示卷积神经网络的“黑盒”机制与可解释性提供新的思路与方法。

关键词: 可解释性; 一维卷积神经网络; 傅里叶变换; 故障诊断; 频域

中图分类号: TH133.33 **文献标志码:** A

DOI: 10.7652/xjtuxb202507003 **文章编号:** 0253-987X(2025)07-0024-12

Study on the Interpretability of One-Dimensional Convolutional Neural Networks in Mechanical Fault Feature Extraction

WANG Fangzhen¹, ZHANG Xiaoli¹, ZHAO Qiwu¹, WANG Baojian²

(1. The Ministry of Education Key Laboratory of Road Construction Technology and Equipment, Chang'an University, Xi'an 710064, China; 2. The National Demonstration Center for Mechanical Foundation Experimental Teaching, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: To address the limited interpretability and reliability caused by the unknown internal decision-making and inference processes of one-dimensional convolutional neural networks (CNNs) in mechanical fault diagnosis, a similarity connection between signal analysis and neural networks is established from the perspective of feature extraction. By extracting the weights of the convolutional layers in the neural network and observing the variations in time/frequency domain features as the network layers change, this study reveals the intrinsic feature extraction behavior of neural networks. Experimental test data and publicly available bearing data from Case Western Reserve University are used for validation. The results indicate that the convolutional kernel can be equivalent to a finite impulse filter, and the max pooling layer can meet the non-linear

requirements of neural networks for simple binary classification tasks, therefore not requiring an activation function in the convolutional layer; the neural network is capable of incrementally increasing frequency resolution layer by layer to identify frequency components close to theoretical fault characteristic frequencies, exhibiting similarities to Fourier transforms. When the spectral range is ultimately decomposed to 1 to 3 times the fault characteristic frequency, the identification task is better accomplished. This study can provide new ideas and methods for revealing the “black box” mechanisms and interpretability of convolutional neural networks.

Keywords: interpretability; one-dimensional convolutional neural networks; Fourier transform; fault diagnosis; frequency domain

卷积神经网络(CNN)因其特征提取能力强、计算量小以及能够自动完成特征提取和故障模式分类的端到端诊断特点,已逐渐成为机械健康监测与智能维护的重要方法^[1-2]。由于卷积神经网络内部存在“黑盒”问题,网络如何从输入数据中提取有用的特征信息从而推断正确结论,以及数据特征流如何在神经网络层中传递变化难以知晓,因此神经网络的可解释性研究成为当前的热点与难点问题。严如强等^[3-4]分别从归因解释技术和先验赋能技术分析了解释性人工智能在智能诊断方面的前景和挑战。Rudin等^[5]综述了当前可解释性研究,提出可解释性基本原则,指出十大可解释挑战问题。Lipton^[6]指出,可解释深度学习可根据模型是否被修改分为事前可解释性和事后可解释性。林京等^[7]进一步指出,事前可解释性主要包括物理知识嵌入、功能化框架嵌入、算法结构等效等方法,旨在对模型施加一定约束从而让模型重点学习指定的知识,而事后可解释性则包含模型相关可解释性和无关可解释性,主要通过显著图、可视化等方法帮助理解神经网络的学习过程。Li等^[8]提出了一种多层梯度加权类激活映射,实验表明其不仅在时域上突出了周期性脉冲,还在频域上强调了故障特征频率。王冉等^[9]采用梯度加权类激活映射算法,对卷积神经网络在基于声像的轴承故障诊断中的可解释性进行了研究,结果表明模型对输入样本激活程度较大的部分均集中在声源附近。Yu等^[10]使用梯度加权类激活映射和基于特征向量的类激活图对 ResNet06 进行可视化解读。Yang等^[11]利用神经元激活最大化和显著图方法对深度神经网络学习到的诊断知识进行可视化,直观地观察到了不同机器健康状态的判别特征。廖才波等^[12]构建变压器故障诊断模型,形成油中溶解气体数据与故障诊断结果的动态关联分析方法,实现了油中溶解气体分析特征量对故障诊

断决策过程及决策结果影响的可视化,提升了变压器故障诊断模型的准确率和解释性。

事后可解释性只能根据结果对模型的学习和推理过程进行推断,验证其是否符合一般的学习和推理逻辑,而不能根据模型本身正向推导结果。为了让模型具有一定的事前可解释性,部分学者采用数据预处理方法,对数据进行部分特征提取或者对信号进行降噪,再交由神经网络完成进一步的特征提取和分类。其中,数据预处理方法大多基于信号特征指标,如包络谱的频率成分和小波变换的峭度^[13]和熵^[14],使其预处理部分存在部分可解释性。Brito等^[15]采用 Shaply 加性解释(SHAP),通过对提取特征的重要程度排序进行无监督故障诊断,使得诊断过程具有依据。Li等^[16]通过 transformer 编码器挖掘各类状态信号之间的关联,通过稀疏约束使注意力权重更有效地关注关键信息。Li等^[17]提出小波卷积层代替第一层普通卷积层,并对卷积核施加约束,使得第一层卷积层为可学习的离散小波变换,之后采用 Mallat 算法将 CNN 的级联卷积结构与小波包分解和重构联系起来,使得其在模型架构上具有可解释性。Zhong等^[18]基于傅里叶逆变换和带通滤波器设计了一个物理信息卷积层,并将其作为卷积神经网络的一部分。

虽然上述研究在可解释性方面均取得了良好效果,但都是通过给网络增加模块或者可视化梯度观察其行为,而不是直接对模型本质及其决策机理进行解释,于是一些学者便开始着眼于解释网络本身的特性和行为。Xu等^[19-20]提出频率原则,指出深度神经网络首先会快速地学习低频成分,之后再相对较慢地学习高频成分,为理解神经网络的学习过程提供了思路。Liao等^[21]采用二次神经元设计卷积神经网络,再从二次神经元中独立推导出注意力机制,使得由二次神经元构成的模型具有内在可解

释性。Magadán 等^[22]采用堆叠自编码器提取特征,确定故障健康指标,发现该指标与某些谐波的幅值高度相关,之后再根据指标和动态时间规整进行相应的故障判断。Chen 等^[23]发现,通过无高斯假设线性判别分析学习权重向量后,能够直接定位轴承故障特征频率及其谐波,为数据驱动的轴承故障诊断方法提供了新见解。Borghesani 等^[24]认为,一维卷积神经网络的单卷积核卷积层可以等效为有限脉冲响应滤波器,通过实验验证,发现频域上能显示出故障特征频率,但无法观察到更加丰富的倍频信息,也未能解释网络层数选取与故障特征频率的关系。

综上所述,现有研究主要通过类激活图等可视化工具,突出神经网络的重点关注内容或嵌入特定模块,从而对神经网络做出部分解释,未从信号处理的视角分析神经网络特征提取与模式识别的本质行为。为了揭示一维卷积神经网络提取信号特征与故障模式识别的机理,本文主要贡献如下。

(1)一维卷积神经网络能够通过堆叠卷积/最大池化块逐层提高频率分辨率,从而寻找目标频率成分,进而完成状态识别任务。其寻找目标频带和特征频率的方式与传统信号处理方法如傅里叶变换具有相似性。

(2)在已知故障特征频率的前提下,卷积/最大池化块的数量可以根据故障特征频率进行设置,当范围细化至 1~3 倍的故障特征频率时,神经网络能够更好地提取特征频率及其倍频信息,从而完成状态识别任务。

(3)建立卷积层与有效脉冲响应滤波器的等效机制,认为卷积层对信号的频域影响可以等效为有限脉冲响应滤波器。

(4)分析最大池化块对于信号的非线性效应,认为最大池化层已经能够满足简单二分类任务的非线性要求,此时卷积层中无需再添加激活函数。

1 卷积神经网络

1.1 一维卷积神经网络经典结构

图 1 给出了一维卷积神经网络在机械故障诊断中的经典结构。首先由卷积层和最大池化层堆叠组成卷积/最大池化块,之后是全连接层和 Softmax 层。图中实线为数据特征提取和分类过程,虚线为反向传播。网络训练过程为输入带有标签的一维振动信号,先由多个卷积/池化层从一维振动信号中逐

渐提取深层特征,之后将特征拉直成一维向量送入到全连接层模块中进行维度转换和特征重要程度的取舍,最后经过 Softmax 层输出预测标签的概率分布,通过损失函数计算损失值,并不断反向传播到全连接层和卷积层中完成参数更新,直到损失值降至规定范围内或者达到最大训练轮数后停止训练,保存模型权重后输入测试集数据对模型进行评估,从而完成从输入信号到状态识别的端到端过程。

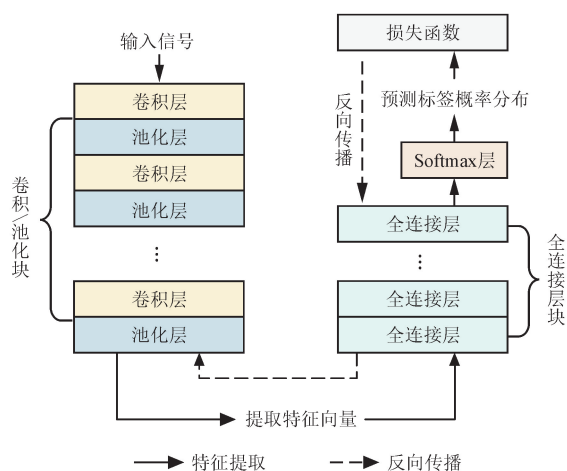


图 1 卷积神经网络经典结构

Fig. 1 Classical architecture of convolutional neural network

1.2 卷积层

一维卷积神经网络实际上是在二维方向上进行操作,卷积核首先沿着时间方向进行卷积,然后沿着通道方向进行计算形成临时特征图,之后再行特征融合形成对应通道的特征图。若输入为 n 个通道,用 k 个卷积核进行卷积,则该卷积层会形成 $n \times k$ 种数据组合,此时不利于揭露数据特征变换过程。由于输入和输出均为单通道时是最简单的卷积操作过程,有利于揭示其数据特征变换过程,因此下文均以 1 个卷积核为例,即从第一层卷积层开始,所有的输入通道和输出通道均为 1,其卷积示操作意图如图 2 所示。以输入长度为 L 的一维振动信号 x 与尺寸为 3×1 的卷积核卷积过程为例,步长设定为 1,不进行补零操作,其中 w 表示卷积层权重, y 表示卷积层输出, $*$ 表示卷积。

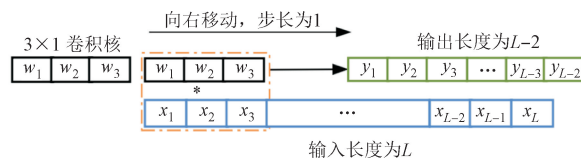


图 2 卷积操作示意图

Fig. 2 Schematic diagram of the convolution process

卷积核沿时间维度与输入进行卷积,其输出长度可写为

$$L_{\text{out}} = \frac{L_{\text{in}} - K + 2L_{\text{pad}}}{S} + 1 \quad (1)$$

式中: L_{out} 为卷积后的输出长度; L_{in} 为卷积层的输入; K 为卷积核长度; L_{pad} 为补零填充长度; S 为移动步长。

一维卷积神经网络的卷积公式为

$$o_n = \sum_{m=0}^{M-1} i_n h_m, \quad n = 0, 1, 2, \dots, L_{\text{in}} - K \quad (2)$$

式中: m, n 分别为输入、输出信号时间步长; o_n, i_n 分别为第 n 个时间步长的输出、输入; h_m 为第 m 个时间步长的卷积核; M 为卷积核长度。

在无补零情况下,卷积层将 L 个时间步长的输入转换为 $L - K + 1$ 个时间步长的输出。卷积层通过卷积核在输入数据上沿着时间维度滑动并计算覆盖区域的加权和,由于此行为和有限脉冲响应滤波器通过将输入与滤波器的冲击响应进行卷积运算实现信号滤波处理的作用非常相似,因此可将这一层在时间维度上看作是有限脉冲响应滤波器。假设此时卷积核的长度为 k ,则等效有限脉冲响应滤波器的构建过程如下。

(1) 将训练集的样本输入到模型中进行训练,到达指定准确率要求或指定轮数后停止训练。

(2) 训练结束后,提取第一层卷积层权重,得到长度与卷积核长度相同的权重向量 w_k 。

(3) 将 w_k 往后补零到奈奎斯特长度得到 w'_k ,将第一层长度的奈奎斯特长度作为输入样本的长度。

(4) 对 w'_k 进行离散傅里叶变换,得到滤波响应的频谱 W_k

$$\begin{cases} W_k = 20 \lg | \text{DFT}\{w'_k\} | \in \mathbf{R}^{1 \times f_{\text{Ny}}} \\ 0 \leq k \leq K \end{cases} \quad (3)$$

式中: DFT 表示离散傅里叶变换; f_{Ny} 表示信号的奈奎斯特频率。

(5) 将频谱 W_k 中最大幅值对应频率作为滤波器的中心频率 f_c 。

(6) 滤波器带宽 B 等于奈奎斯特频率除以卷积核长度。

(7) 根据中心频率 f_c 和带宽 B , 构建等效有效脉冲响应滤波器。

为了更好地显示等效滤波器的建立过程,绘制流程图如图 3 所示。

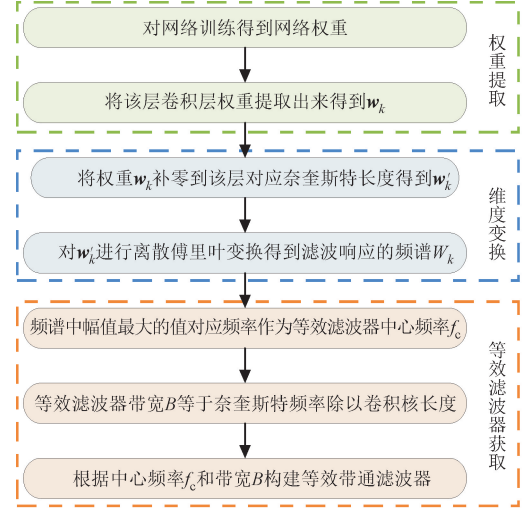


图 3 卷积层等效滤波器构建过程

Fig. 3 Construction process of the convolutional layer equivalent to a filter

1.3 池化层

池化层是卷积神经网络中的重要部分,若选用如式(2)最简单的卷积形式,那么连续两个卷积层的堆叠将没有任何意义,因为两个线性组合叠加过程之后仍然是线性过程,和一个线性过程的参数调整完全一样。最大池化层是机械故障诊断中采用最多的池化层,其数学表达式为

$$o_n^{\max} = \max_{p=0,1,2,\dots,P-1} (i_{n,P+p}), \quad n = 0, 1, 2, \dots, \left\lfloor \frac{L}{P} \right\rfloor - 1 \quad (4)$$

式中: $o_n^{\max}$ 为最大池化输出; p 为池化核步长; P 为池化核长度。

最大池化操作过程是计算移动窗口中的最大值,然后进行下采样,最常见的形式是池化核大小 P 和步长相等,此时移动窗没有任何重叠,也就是在每 P 个原始样本中选取一个最大值,其输出长度的计算依然遵循式(1),只需把卷积核长度和步长更换为池化核长度和步长即可。池化层会将后续样本之间的平均时间间隔增加到 P 倍,即 $\Delta t_{\text{out}} = P \Delta t_{\text{in}}$,其中 Δt_{out} 表示最大池化输出的时间间隔, Δt_{in} 表示最大池化输入的时间间隔。由于采样频率与时间间隔为倒数关系,因此输出信号的采样频率 $f_{s,\text{out}} = f_{s,\text{in}}/P$, $f_{s,\text{out}}$ 和 $f_{s,\text{in}}$ 为对应最大池化输出、输入的采样频率。最大池化对于高频和低频分量的非线性影响有很大不同,对于周期为 T 的时域信号,如果 $P \geq T$,则最大池化几乎等同于包络与降采样,即提取信号的幅值。包络的不完美性导致每个周期的采样最大值不同,如图 4(a) 所示, P 越长,误差越

小,此时多个周期内的最大值不会受到窗口与信号周期不同步的影响。从图 4(b) 可以看出,经过池化后,原信号的频率特征无法保留。

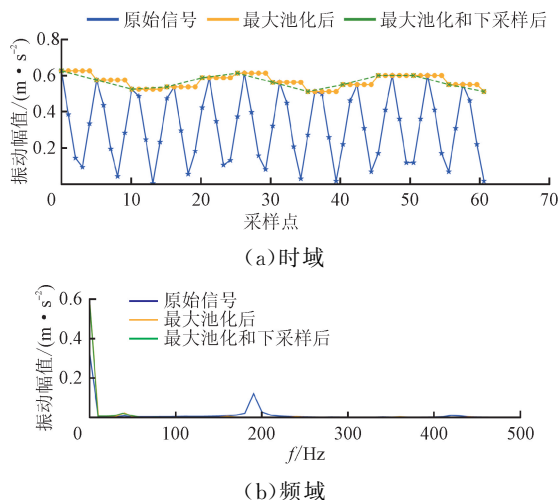


图 4 $P \geq T$ 时最大池化和下采样对时域和频域的影响

Fig. 4 Effect of maximum pooling and downsampling on the time domain frequency domain for $P \geq T$

相反,若 $T \gg P$,最大池化和下采样在时域和频域中的影响如图 5 所示。可以看出,此时原始信号的整体趋势在下采样信号中得到保留,但从图 5(b) 中能够看出,最大池化操作会在频域上产生直流分量(坐标为 0 时的频率分量)、原始成分的谐波以及 T 和 P 组合的小分量。值得注意的是,只有在池化核大小 P 等于池化步长时,才会出现图 4 和图 5 所示情况,否则最大池化和下采样的操作会产生更多的分量,从而为选择相邻和不重叠池化的常见做法提供傅里叶分析的依据。

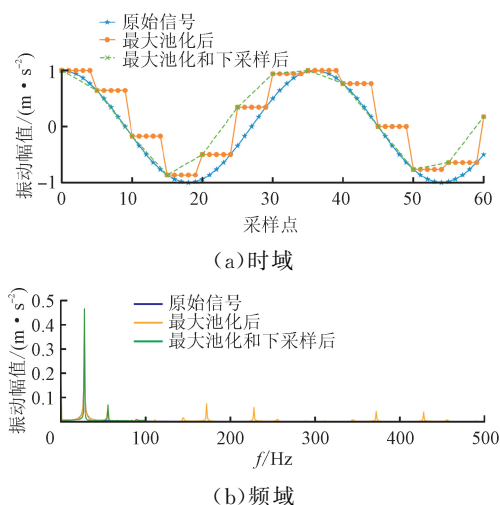


图 5 $T \gg P$ 时最大池化和下采样对时域和频域影响示意图

Fig. 5 Schematic representation of the effect of maximum pooling and downsampling on the time domain and frequency domain at $T \gg P$

通过对比可知,对于 $T \gg P$ 的情况,振动信号最大池化后能更好地保留其信号的频域信息,且产生的干扰分量影响也较小。对于滚动轴承等旋转机械故障诊断来说,因为采样频率普遍较高,一个周期(约为滚动轴承旋转 1 圈)的采样点数要远远大于池化核长度,故最大池化能较好地解决此类问题。

将最大池化层的非线性效应与卷积神经网络中常用的 Relu 激活函数进行定性比较,结果如图 6 所示。可以发现,最大池化层已经能够进行解调,且最大池化能够对于高频载波进行高质量的解调,而激活函数只能实现部分解调。因此,本文后续将不在卷积层中使用激活函数,而是采用卷积层加最大池化层的特征提取组合。

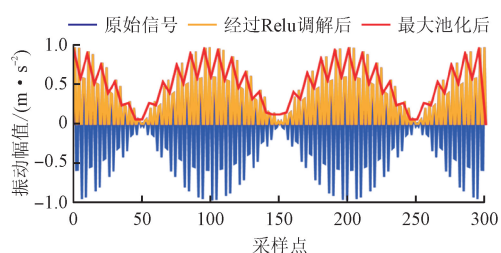


图 6 最大池化和 Relu 解调效果对比

Fig. 6 Comparison of demodulation effects of maximum pooling and Relu

2 基于傅里叶变换的卷积神经网络可解释性分析

基于信号处理的机械故障诊断方法主要是从复杂信号中提取故障特征频率,而其他频率谱线会引起干扰混淆,因此在频域中以高分辨率和强选择性分析信号,能更好地分离相邻谱线分量。由上文可知,卷积层和最大池化层交替的特征提取组合方式能够有效地解决这个问题。作为信号处理的常用基础方法,离散傅里叶变换相当于信号的圆卷积,滤波器的长度等于信号长度 L 。为了获得类似的滤波器选择性,具有单个卷积层的卷积神经网络需要滤波器的长度接近信号本身长度。由于卷积神经网络的非循环特性,不可能实现类似快速傅里叶变换的选择性,即便是接近快速傅里叶变换的选择性,也会导致卷积核长度大约为 $L/2$,在状态监测中很容易达到 $10^5 \sim 10^6$ 的数量级,因此存在训练效率低的问题。与卷积层/最大池化层堆叠方式较类似的信号处理方式是 Zoom 快速傅里叶变换^[25],其利用混叠打破采样条件对于频率分辨率的限制,从而以短信号获得高分辨率。该方法通过对中心频率为 f_c 、半

带宽为 B 的特定目标带进行傅里叶分析的示意如图 7(a)所示。其分析流程为:

(1)通过对原始信号乘以负指数 $\exp(-2\pi j f_c t)$ 进行移频,使目标频带的中心频率偏移到零点;

(2)对上、下截止频率为 $f_c \pm B$ 的原始信号进行低通滤波,从而滤除目标频带外的分量;

(3)对信号进行下采样,将信号长度变为原始信号长度的 $1/P$,在理想情况下,奈奎斯特频率对应带宽 B ;

(4)计算下采样信号的快速傅里叶变换,保留 $[f_L, f_H]$ 范围内的频谱进行理论分析,其中 f_L 为最小频率, f_H 为最大频率。

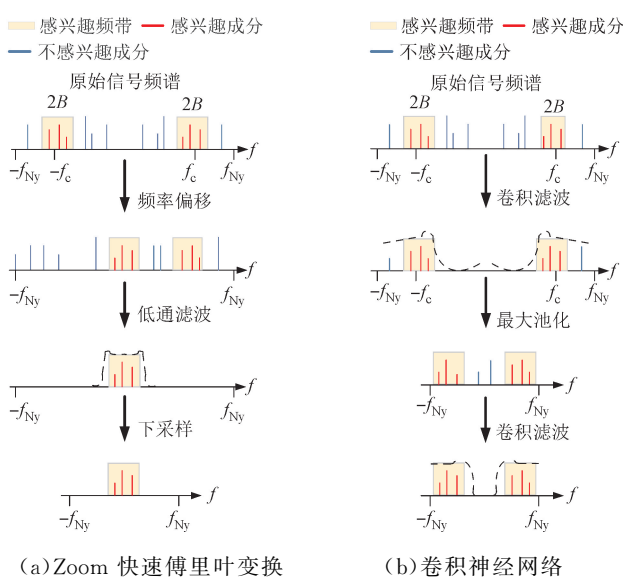


图 7 Zoom 快速傅里叶变换和卷积神经网络处理数据对比
Fig. 7 Comparison of Zoom fast Fourier transform and convolutional neural networks in processing signals

在卷积神经网络中,卷积/最大池化层也采用类似方法,结果如图 7(b)所示。由于通常采用长度很短的卷积核(如长度为 1、3、5 等)在原始信号上进行卷积,第一个卷积层没有足够的过滤选择性来单独提取目标成分,通过带通过滤,只保留目标频率附近频段的成分,然后最大池化层对其输出进行下采样。因此,下一个卷积层将在较短的信号上运算,拥有更高的过滤选择性,从而限制第二个卷积层的通带。

卷积/最大池化块在时域和频域信号中处理信号的过程演示如图 8 所示。以 3 个卷积/最大池化块分解一个长度为 22 的一维振动信号为例,卷积核长度为 3,步长为 1,不进行补零填充,池化核的长度和步长均为 2。时域中信号的变化如图 8(a)所示,

在第一层卷积层进行卷积操作之后信号长度为 20,通过最大池化层后长度变为 10,随后各网络层信号长度分别为 8、4、2、1,信号的时间分辨率逐渐降低,频率分辨率逐渐提高。信号在频域中的变化如图 8(b)所示,假设频谱中包含一个感兴趣的成分如故障特征频率(黄色星形)和一系列干扰分量如噪声和谐波(蓝色圆形),同时假定卷积核有足够的选择性,可以将目前的奈奎斯特频率范围分为 3 部分,且只允许包含感兴趣的成分部分通过,此时干扰分量由 7 个变为 3 个。接着,最大池化层将频谱上半部分的所有分量通过混叠降至下半部分,下采样之后新的奈奎斯特频率降为原来的 $1/2$,频率选择性提高到 2 倍。这使得接下来的卷积操作可以放大并限制目标频率附近的通带,之后再进一步进行卷积和最大池化操作,一步一步缩小通带从而得到目标分量。

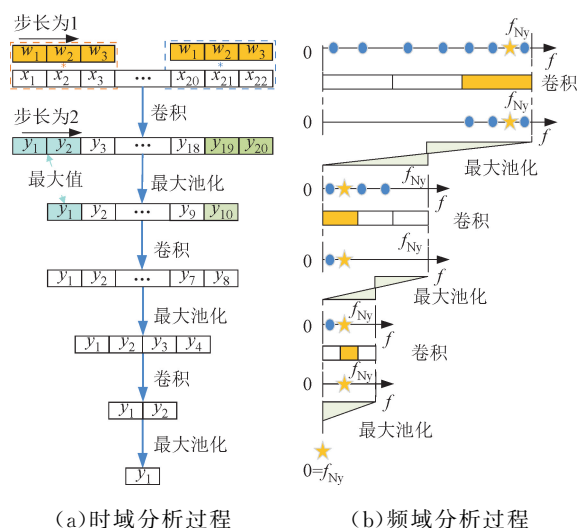


图 8 卷积、最大池化块频率分析示意图

Fig. 8 Schematic of frequency analysis of convolutional/maximum pooling block

3 实验验证分析

上节从傅里叶变换这一传统信号处理视角对卷积神经网络进行了可解释分析,目的不是使用卷积神经网络模拟信号处理工具,而是从时域和频域角度来解释卷积神经网络进行特征提取与模式识别的本质。本节选用健康和外圈故障两种状态,通过滚动轴承故障模拟试验台采集振动信号和凯斯西储大学(CWRU)轴承公开数据集对其进行验证。

3.1 滚动轴承故障实验分析

3.1.1 实验平台与数据

故障模拟实验台如图 9 所示,其由电机、变频控

制器、加载装置、ER-12K 测试轴承以及 X、Y 轴传感器组成,测试轴承的参数信息如表 1 所示。其中,变频控制器可以控制电机的转速从而模拟不同工况,加载装置可以调整负载大小,并通过皮带轮加载到主轴上,传感器均为振动加速度传感器,通过动态信号采集分析系统采集信号,电机转频 $f_r=20\text{ Hz}$,采样频率为 25.6 kHz 。

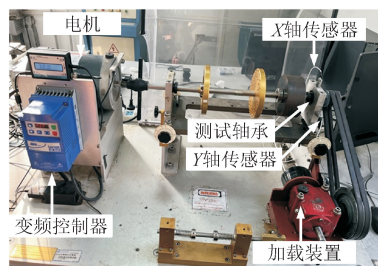


图 9 轴承故障模拟实验台

Fig. 9 Experimental bench for bearing fault simulation

表 1 实验轴承参数

Table 1 Parameters of experimental bearings

内圈 直径/ mm	外圈 直径/ mm	节圆 直径 D / mm	滚动体 数量 n	滚动体 直径 d / mm	接触角 $\alpha/(^\circ)$
25.4	52	33.48	8	7.96	0

轴承外圈故障特征频率 f_o 的计算公式为

$$f_o = \frac{nf_r}{2} \left(1 - \frac{d}{D} \cos \alpha \right) \quad (5)$$

将电机转频 $f_r=20\text{ Hz}$ 和滚动轴承结构参数代入式(5)后,得到外圈故障特征频率 $f_o=61.03\text{ Hz}$ 。

3.1.2 网络结构与实验结果分析

由于数据集采样频率为 25.6 kHz ,而外圈故障特征频率 $f_o=61.03\text{ Hz}$,采样频率相对于故障特征频率较高,因此在第一个卷积/最大池化块中的最大池化层池化核的长度和步长选为 20,这样在其解调时可以将其奈奎斯特频率范围降为 $0\sim640\text{ Hz}$,从而在后续卷积/最大池化层中能够确保其分解到故障特征频率所在频段。之后的池化核长度均取为 2,这样每经过一次最大池化,频率范围均降为原来的一半,以便逐层观察数据特征流的变换过程,卷积/最大池化块数设定为将信号分解到不满足下一次卷积/池化操作为止。为了有足够的卷积/最大池化块数从而提高频率分辨率,将输入信号长度截取为 20 000,此时共有 6 个卷积/最大池化块,最后的网络层数据具有时间分辨率最低、频率分辨率最高

的特点。网络结构及各层输入、输出情况如表 2 所示,其中网络层参数分别表示卷积核/池化核的长度和步长。池化核均采用相同的长度和步长,即所有下采样窗不重叠,由于频谱的对称性,因此仅取频率的正值范围。从表 2 可以看出,随着卷积/最大池化块的堆叠操作,通过第一个卷积/最大池化块将频率范围解调至千赫兹以内。

表 2 网络结构与各层输入输出

Table 2 Network structure with inputs and outputs of each layer

名称	网络层 参数	特征图 尺寸	参数量	频率正值 范围/Hz
输入层		$1\times20\,000$	20	$(0,12\,800)$
第 1 层卷积层	$(20,1)$	$1\times19\,981$	20	$(0,12\,800)$
第 1 层池化层	$(20,20)$	1×999	0	$(0,640)$
第 2 层卷积层	$(20,1)$	1×980	20	$(0,640)$
第 2 层池化层	$(2,2)$	1×490	0	$(0,320)$
第 3 层卷积层	$(20,1)$	1×471	20	$(0,320)$
第 3 层池化层	$(2,2)$	1×235	0	$(0,160)$
第 4 层卷积层	$(20,1)$	1×216	20	$(0,160)$
第 4 层池化层	$(2,2)$	1×108	0	$(0,80)$
第 5 层卷积层	$(20,1)$	1×89	20	$(0,80)$
第 5 层池化层	$(2,2)$	1×44	0	$(0,40)$
第 6 层卷积层	$(20,1)$	1×25	20	$(0,40)$
第 6 层池化层	$(2,2)$	1×12	0	$(0,20)$
全连接层	$(12,2)$	1×2	26	
Softmax 层		1×2	0	

选用 Pytorch 2.0 作为深度学习框架,Python 3.9 作为开发语言,选用 Adam 优化器作为网络优化器,初始学习率设置为 0.001,批次大小设置为 64,二元交叉熵作为损失函数,训练轮数设置为 100 轮,经过多次训练后,识别准确率可达到 93.86%。调整网络结构,首先单独去除第 6 个卷积/最大池化块形成 5 层神经网络结构,并将全连接层输入神经元个数修改为 44,接着再去除第 5 个卷积/池化对形成 4 层网络结构,然后修改全连接层的输入神经元数为 108,再对模型重新训练后,5 层和 4 层网络结构的正确率均可达到 100%,相比之前网络的正确率 93.86%有了较大提升。图 10 给出了模型识别准确率曲线的对比。

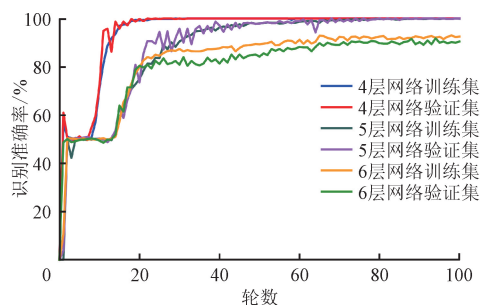


图10 3种网络结构识别准确率曲线对比

Fig. 10 Comparison of recognition accuracy of three network structures

由图10可以看出,4层网络模型的训练集和验证集准确率曲线非常稳定,在20轮左右就已收敛到100%,且几乎没有产生振荡。其次是5层网络,约在70轮左右收敛到100%,但仍然会产生一定振荡,且一直到100轮也无法如4层网络一样使训练集和准确率一直维持在100%。6层网络收敛后的训练集准确率约为92%,验证集准确率约为90%,准确率不够高,且产生了较大振荡,表明此时网络模型难以准确识别故障状态。由于外圈故障特征频率为61.03 Hz,虽然第5层卷积层频谱范围为0~80 Hz,仍能包含其特征频率,但是其经过池化后的频谱范围为0~40 Hz,已然超出特征频率范围,而第六层卷积和最大池化后的频谱范围分别为0~40 Hz和0~20 Hz,频带已不包含特征频率信息,而输入到全连接层或自适应最大池化层的数据正是来自于最后一个卷积/最大池化块。由于网络在第4层卷积层已经能够很好地识别到故障特征频率,之后卷积/最大池化块的增加反而丢掉了故障特征频率信息,模糊了其正确的推断决策,表明网络层数选择与故障特征频率所在的频段应紧密相关。

由图11和图12的时域图和频域可以看出,健康状态信号的谐波性更为明显,每一层卷积层的信号都能看到较明显的谐波性,且经过第一个最大池化层解调后的信号幅值在卷积前后相差较大,呈现分离状态,其时域幅值在第3层卷积层和第4层卷积层已经相差较大。外圈故障状态因为存在故障冲击,导致各层信号更为复杂,谐波性较弱。而在其原始信号的最高幅值大于健康状态信号的情况下,第2~4层卷积层的时域信号幅值主要在 $\pm 1.5 \text{ m/s}^2$ 范围内波动,因此两种状态的信号在经过卷积/最大池化后的时域幅值有了明显区分。

从频域图来看,图11所示的健康状态信号在第1个卷积/最大池化层解调之后能看到部分频率信息,但在第3层卷积层之后已经很难再看到一些明显的特征频率。在图12所示的第2层卷积层频谱图中,能明显看到外圈故障信号主要聚集在故障特征频率 $f_0=59 \text{ Hz}$ 及其倍频 $2f_0\sim 7f_0$ 处,各频率成分等间隔分布,间隔约为59 Hz,之后在第3卷积层的频谱中仍然可以看到在 $f_0\sim 5f_0$ 处存在频率谱峰。卷积层中提取出的59 Hz与外圈故障特征频率61.03 Hz很接近,虽然存在约2 Hz的偏差,但对于神经网络实现端到端特征提取与模式识别非常重要,表明神经网络通过卷积/最大池化不断地将信号频谱细化,以便自动提取时/频域差异化特征,根据其故障特征频率及倍频信息可以识别不同状态的信号,与傅里叶变换等传统信号处理方法将时域信号转换到频域进行故障特征分析有异曲同工之处。

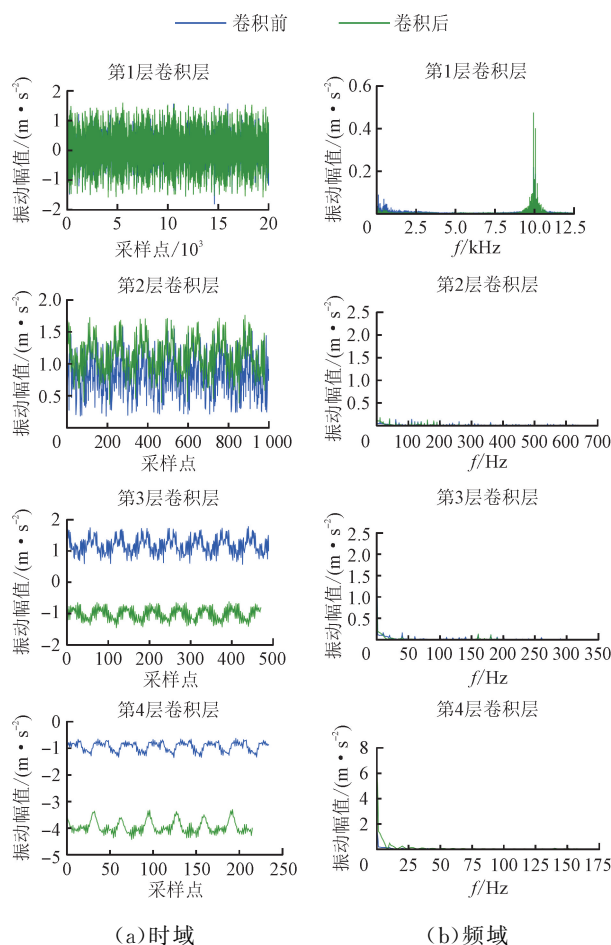


图11 健康信号各卷积层的时域和频域图

Fig. 11 Time-domain and frequency-domain for each convolutional layer in the healthy bearing state of experimental signal

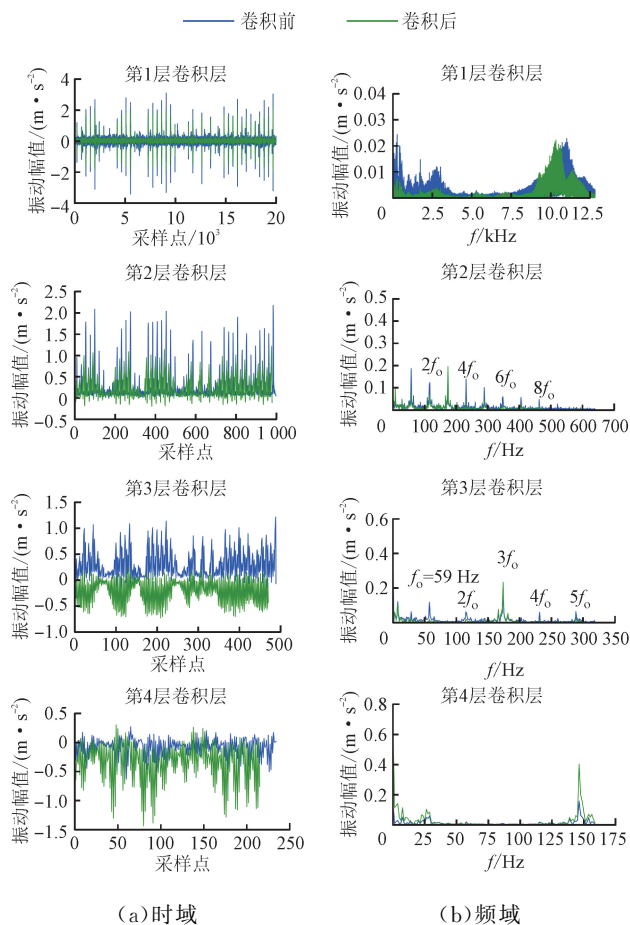


图 12 外圈故障信号各卷积层的时域和频域图

Fig. 12 Time-domain and frequency-domain for each convolutional layers in the outer ring fault bearing state of experimental signal

对网络层数的研究表明:在已知故障特征频率的前提下,卷积/最大池化块的数量可以根据其进行设置,将频率范围细化至 1~3 倍故障特征频率时,神经网络能够更好地提取特征频率及其倍频信息。

3.1.3 参数敏感性分析

为排除因特定参数选取导致网络寻找到故障特征频率结果的偶然性,对网络结构参数进行敏感性分析。将首层池化层的长度和步长由 20 调整为 40,最后一层频率范围依然为 0~160 Hz,此时网络共有 3 层,每个卷积层的频率正值最大值分别为 12.8 kHz、320 Hz 和 160 Hz。将网络重新训练后,网络准确率为 100%,由于健康状态时/频域无明显特征频率信息且与图 11 类似,因此仅展示外圈故障信号在各卷积层卷积前、后的时域和频域,如图 13 所示。由图 13 可以看出,网络依然有明显的特征频率,但频率范围迅速缩小,第 2 层中有故障特征频率及其倍频 $f_0 \sim 5f_0$,且调整后网络

识别的特征频率 f_0 仍为 59 Hz,排除了结果偶然性的发生。

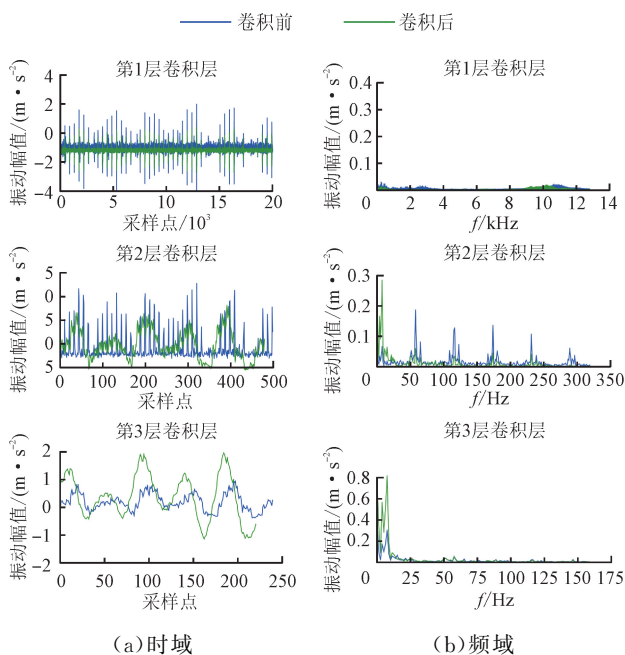


图 13 首层池化核长度为 40 时外圈故障信号的时域和频域信息

Fig. 13 Time-domain and frequency-domain for each convolutional layers for the outer-ring fault state when the length of the first pooling kernel is 40

3.2 凯斯西储大学轴承公开数据集验证分析

3.2.1 数据集介绍

凯斯西储大学轴承公开数据集是目前深度学习领域广泛使用的数据集,其通过电火花损伤加工模拟了不同故障损伤程度的多种故障状态。为进一步验证所提方法的有效性,采用健康状态和外圈故障数据进行验证分析,数据集采样频率为 48 kHz,故障尺寸为 0.177 8 mm(0.007 英寸),主轴转速为 1 797 r/min(29.95 Hz),无负载,轴承的结构参数如表 3 所示。将转频和结构参数代入式(5)后,得到 CWRU 轴承公开数据集外圈故障 $f_0 = 107.36$ Hz。

表 3 CWRU 轴承参数

Table 3 Parameters of CWRU bearings

D/mm	n	d/mm	$\alpha/(\text{^\circ})$
39.04	9	7.94	0

3.2.2 实验结果分析

根据 3.1.2 小节所提理论,在已知故障特征频率的前提下,卷积/最大池化块的数量可以根据其进行设置。由于外圈故障特征频率 $f_0 = 107.36$ Hz,因此将卷积/最大池化层的数量设置为 4,此时频率

正值范围为 0~300 Hz。网络结构和表 2 前 4 层所示一致,只需将第 5 和第 6 层网络删除同时修改全连接层个数为 108 即可,4 个卷积层频率正值范围分别为 24 kHz、1.2 kHz、600 Hz 和 300 Hz。经过训练后,最终的识别准确率为 100%,表明此时网络拥有较好的性能,验证了在已知故障特征频率时网络层数设置的有效性。

图 14 展示了外圈故障信号在各卷积层卷积前后的时域和频域图。可以看到,从第 2 层开始,频域中存在明显的特征频率 $f_0 \sim 7f_0$,一直到网络的最后一层仍然能够保留其特征 f_0 ,约为 108 Hz,与理论计算的外圈故障特征频率 107.36 Hz 非常接近,再次验证了本文所提方法的有效性。

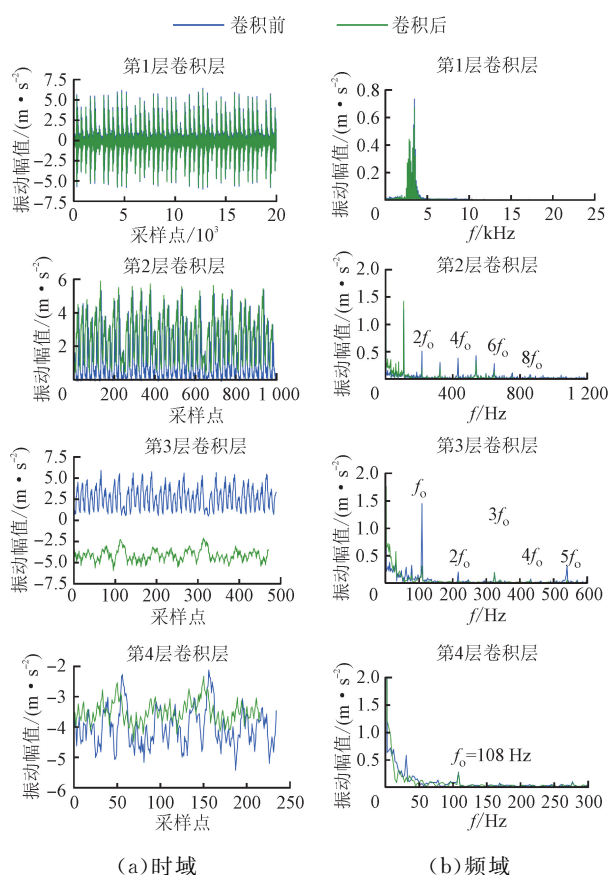


图 14 CWRU 轴承外圈故障状态各卷积层时域和频域图
Fig. 14 Time-domain and frequency-domain for each convolutional layer of the outer ring fault state at CWRU

4 结 论

针对卷积神经网络存在的“黑盒”现象与难解释问题,分析卷积神经网络中的卷积层、池化层对信号时域和频域的影响,以及卷积层/最大池化层堆叠结构与傅里叶变换在寻找目标频带和频率的相似性,

并利用实验测试信号与 CWRU 数据进行验证。在不对网络结构进行改进调整的情况下,对卷积神经网络各层行为进行解释,一定程度上揭示了卷积神经网络在机械故障诊断中的运算过程与本质规律,取得如下主要结论。

(1)卷积层/最大池化层堆叠的结构和傅里叶变换寻找目标频带和特征频率时存在一定的相似性,都是通过类似的滤波、下采样操作从而逐渐将频谱细化到目标频带进而寻找其目标频率。

(2)通过实验测试信号和轴承公开数据集在各卷积层前后的时域和频域的表现可以发现,卷积神经网络通过逐层提高频率分辨率从而逐渐寻找到故障特征频率及其倍频,可实现端到端故障模式识别。

(3)卷积/最大池化块数量并非越多越好,因为过多卷积/最大池化块会将频谱细化到超出故障特征频率及其倍频的区间,从而使得网络丢失重要判定依据。在已知故障特征频率时,卷积/最大池化块将频谱细化到 1~3 倍故障特征频率是个较好的选择,此时不仅包含了故障特征频率信息,也能很好地包含倍频信息,从而增强网络的准确性。

(4)最大池化层已经能够满足简单二分类任务中神经网络的非线性化要求,此时卷积层无需添加激活函数。

本文基于傅里叶变换对一维卷积神经网络以及卷积层/最大池化层在时域、频域中的影响机制进行了分析,今后可以从更加宽广的角度对神经网络的可解释性进行分析,同时对其他网络结构如长短期记忆网络等也可进行相关研究。

参考文献:

- [1] 卞文彬, 邓艾东, 刘东川, 等. 基于改进深度残差收缩网络的风电机组滚动轴承故障诊断方法 [J]. 机械工程学报, 2023, 59(12): 202-214.
BIAN Wenbin, DENG Aidong, LIU Dongchuan, et al. Fault diagnosis method of wind turbine rolling bearing based on improved deep residual shrinkage network [J]. Journal of Mechanical Engineering, 2023, 59(12): 202-214.
- [2] ZHANG Xiaoli, WANG Fangzhen, ZHOU Yongqing, et al. The optimized local sparse parallel multi-channel deep convolutional neural network-LSTM and the application in bearing fault diagnosis under noise and variable working condition [J]. Proceedings of the Institution of Mechanical Engineers: Part C Journal of Mechanical Engineering Science, 2025, 239(1): 205-218.

- [3] 严如强, 商佐港, 王志颖, 等. 可解释人工智能在工业智能诊断中的挑战和机遇: 先验赋能 [J]. 机械工程学报, 2024, 60(12): 1-20.
YAN Ruqiang, SHANG Zuogang, WANG Zhiying, et al. Challenges and opportunities of XAI in industrial intelligent diagnosis: priori-empowered [J]. Journal of Mechanical Engineering, 2024, 60(12): 1-20.
- [4] 严如强, 周峥, 杨远贵, 等. 可解释人工智能在工业智能诊断中的挑战和机遇: 归因解释 [J]. 机械工程学报, 2024, 60(12): 21-40.
YAN Ruqiang, ZHOU Zheng, YANG Yuangui, et al. Challenges and opportunities of XAI in industrial intelligent diagnosis: attribution interpretation [J]. Journal of Mechanical Engineering, 2024, 60(12): 21-40.
- [5] RUDIN C, CHEN Chaofan, CHEN Zhi, et al. Interpretable machine learning: fundamental principles and 10 grand challenges [EB/OL]. (2021-07-10)[2024-10-12]. <https://arxiv.org/abs/2103.11251>.
- [6] LIPTON Z C. The mythos of model interpretability [J]. Communications of the ACM, 2018, 61(10): 36-43.
- [7] 林京, 焦金阳. 可解释机械智能诊断技术的研究进展与挑战 [J]. 机械工程学报, 2023, 59(20): 215-224.
LIN Jing, JIAO Jinyang. Research progress and challenges of interpretable mechanical intelligent diagnosis [J]. Journal of Mechanical Engineering, 2023, 59(20): 215-224.
- [8] LI Sinan, LI Tianfu, SUN Chuang, et al. Multilayer grad-CAM: an effective tool towards explainable deep neural networks for intelligent fault diagnosis [J]. Journal of Manufacturing Systems, 2023, 69: 20-30.
- [9] 王冉, 石如玉, 胡升涵, 等. 基于声成像与卷积神经网络的轴承故障诊断方法及其可解释性研究 [J]. 振动与冲击, 2022, 41(16): 224-231.
WANG Ran, SHI Ruyu, HU Shenghan, et al. An acoustic fault diagnosis method of rolling bearings based on acoustic imaging and convolutional neural network [J]. Journal of Vibration and Shock, 2022, 41(16): 224-231.
- [10] YU Shihang, WANG Min, PANG Shanchen, et al. Intelligent fault diagnosis and visual interpretability of rotating machinery based on residual neural network [J]. Measurement, 2022, 196: 111228.
- [11] YANG Huixin, LI Xiang, ZHANG Wei. Interpretability of deep convolutional neural networks on rolling bearing fault diagnosis [J]. Measurement Science and Technology, 2022, 33(5): 055005.
- [12] 廖才波, 杨金鑫, 邱志斌, 等. 一种基于夏普利值及油中溶解气体分析的可解释变压器故障诊断方法 [J]. 电网技术, 2024, 48(4): 1752-1761.
LIAO Caibo, YANG Jinxin, QIU Zhibin, et al. Interpretable transformer fault diagnosis based on SHAP value and dissolved gas analysis of transformer oil [J]. Power System Technology, 2024, 48(4): 1752-1761.
- [13] JIA Sixiang, SUN Dingyi, NOMAN K, et al. Lifting wavelet-informed hierarchical domain adaptation network: an interpretable digital twin-driven gearbox fault diagnosis method [J]. Reliability Engineering & System Safety, 2025, 254(Part B): 110660.
- [14] GAO Lei, LIU Zhihao, GAO Qinhe, et al. Dual data fusion fault diagnosis of transmission system based on entropy weighted multi-representation DS evidence theory and GCN [J]. Measurement, 2025, 243: 116308.
- [15] BRITO L C, SUSTO G A, BRITO J N, et al. An explainable artificial intelligence approach for unsupervised fault detection and diagnosis in rotating machinery [J]. Mechanical Systems and Signal Processing, 2022, 163: 108105.
- [16] LI Yasong, ZHOU Zheng, SUN Chuang, et al. Variational attention-based interpretable transformer network for rotary machine fault diagnosis [J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(5): 6180-6193.
- [17] LI Sinan, LI Tianfu, SUN Chuang, et al. WPConvNet: an interpretable wavelet packet kernel-constrained convolutional network for noise-robust fault diagnosis [J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(10): 14974-14988.
- [18] ZHONG Jingshu, ZHENG Yu, RUAN Chengtao, et al. M-IPISincNet: an explainable multi-source physics-informed neural network based on improved SincNet for rolling bearings fault diagnosis [J]. Information Fusion, 2025, 115: 102761.
- [19] XU Z Q J, ZHANG Yaoyu, LUO Tao. Overview frequency principle/spectral bias in deep learning [J/OL]. Communications on Applied Mathematics and Computation. (2024-09-04)[2024-12-20]. <https://doi.org/10.1007/s42967-024-00398-7>.
- [20] XU Zhiqin, ZHANG Yaoyu, XIAO Yangang. Training behavior of deep neural network in frequency domain [C]//Neural Information Processing. Cham: Springer International Publishing, 2019: 264-274.
- [21] LIAO Jingxiao, DONG Hangcheng, SUN Zhiqi, et al. Attention-embedded quadratic network (qttnet) for effective and interpretable bearing fault diagnosis [J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 1-13.
- [22] MAGADÁN L, RUIZ-CÁRCEL C, GRANDA J C, et al. Explainable and interpretable bearing fault classification

- cation and diagnosis under limited data [J]. *Advanced Engineering Informatics*, 2024, 62(Part D): 102909.
- [23] CHEN Yikai, WANG Dong, HOU Bingchang, et al. Gaussian assumptions-free interpretable linear discriminant analysis for locating informative frequency bands for machine condition monitoring [J]. *Mechanical Systems and Signal Processing*, 2023, 199: 110492.
- [24] BORGHESANI P, HERWIG N, ANTONI J, et al. A Fourier-based explanation of 1D-CNNs for machine condition monitoring applications [J]. *Mechanical Systems and Signal Processing*, 2023, 205: 110865.
- [25] HOYER E, STORK R. The zoom FFT using complex modulation [C]//ICASSP '77. *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Piscataway, NJ, USA: IEEE, 1977: 78-81.
- (编辑 亢列梅 李慧敏)