故障诊断模型的可解释性增强方法研究

赵施斌　李浩然
上海航天电子技术研究所　上海　201109

摘　要：本文针对深度学习模型在旋转机械故障诊断中的“黑箱”问题，提出融合注意力机制与Grad-CAM的解释框架。通过注意力权重可视化关键故障特征，结合机械振动理论验证模型决策逻辑。研究构建多维度可解释性评估指标体系，量化解释方法对诊断结果可信度的提升效果，为故障诊断模型在工业场景的应用提供可靠支撑。

关键词：故障诊断模型；可解释性增强；注意力机制；Grad-CAM；机械振动理论

引言：深度学习模型在故障诊断领域应用广泛，但“黑箱”特性导致决策逻辑难以理解，无法满足工业场景对决策信任的需求。旋转机械故障诊断中，信号冗余与噪声干扰使传统模型诊断偏差大。本文聚焦该问题，提出融合注意力机制与Grad-CAM的解释框架，结合机械振动理论验证，构建评估指标体系，旨在提升模型可解释性，为工业应用提供更可靠方案。

1 可解释性增强方法的核心思路

1.1 注意力机制与故障特征可视化

在故障诊断场景中，注意力机制的适用性体现在能够模拟人类对关键信息的聚焦能力，尤其适配工业设备复杂的运行环境。工业设备故障信号常伴随大量冗余噪声，传统模型易受干扰导致诊断偏差，而注意力机制可通过动态分配权重，优先捕捉与故障关联紧密的信号片段，契合故障诊断对关键特征精准识别的需求[1]。在关键故障特征提取与可视化路径上，注意力机制先对输入的故障信号进行多维度分解，这里的故障信号包括振动信号等类型，通过计算不同信号片段的重要性权重，筛选出对故障分类起决定性作用的特征分量；随后借助热力图等可视化工具，将抽象的权重分布转化为直观的图像呈现，使研究人员能够清晰观察模型对故障特征的关注重点，明确模型决策所依赖的关键信息来源，为后续验证模型合理性提供直观依据。

1.2 梯度加权类激活映射（Grad-CAM）的融合

Grad-CAM在故障分类任务中具备突出的局部解释能力，其核心优势在于无需修改模型结构，即可通过计算目标类别对卷积层特征图的梯度，定位对分类结果贡献最大的局部区域，大幅降低解释方法的应用门槛。在故障诊断中，该技术能精准标记出信号或图像中与特定故障对应的关键区域，常见的特定故障包括轴承磨损、齿轮断齿等，清晰揭示模型判断故障类别的局部依据。注意力机制与Grad-CAM的协同作用机制体现在两者的互补性上，注意力机制从全局层面筛选关键故障特征，解决信号冗余问题；Grad-CAM则从局部层面细化特征贡献，定位核心故障区域。两者结合时，注意力机制筛选的全局关键特征可为Grad-CAM提供聚焦方向，避免局部解释陷入无关区域；Grad-CAM的局部定位结果又能反向验证注意力机制权重分配的合理性，形成全局与局部相互支撑的解释体系，显著提升模型解释的全面性与精准度。

1.3 机械振动理论的验证逻辑

频谱分析等机械振动理论为模型决策提供了重要的物理约束，是连接数据驱动模型与实际工业场景的关键桥梁。在旋转机械故障诊断中，不同故障类型对应特定的振动频率特征，例如轴承内圈故障对应特定倍数的转频，这些基于物理原理的特征规律构成了模型决策的客观标准。通过频谱分析提取故障信号的频率成分，可与模型输出的故障类别及关注特征进行比对，判断模型决策是否符合机械振动的物理规律，避免模型因数据偏差产生违背物理常识的错误判断。理论验证与模型解释的闭环反馈设计则进一步强化解释的可靠性，首先利用模型解释结果明确模型决策依据，常用的模型解释结果包括注意力权重、Grad-CAM热力图等，再通过机械振动理论验证该依据的物理合理性；若验证发现偏差，则反向调整模型参数或解释方法，比如优化注意力权重计算方式，直至模型解释与物理理论一致，形成“模型解释-理论验证-方法优化”的闭环，确保解释结果兼具数据驱动合理性与物理理论支撑性，满足工业场景对决策可靠性的严格要求。

2 解释框架的设计与实现

2.1 框架总体架构

框架采用分层递进式结构，各层级功能明确且协同联动。输入层聚焦多传感器振动信号预处理，针对旋转机械不同监测位置采集的振动信号，通过去趋势化、滤波去噪等操作，消除环境干扰与设备运行背景噪声，保

留与故障相关的有效信号成分，为后续特征提取奠定数据基础。特征提取层实现注意力机制与深度学习模型的深度结合，深度学习模型负责对预处理后的信号进行多尺度特征挖掘，注意力机制则嵌入模型关键层，通过动态计算特征通道与空间维度的重要性权重，强化故障特征表征，抑制冗余信息干扰[2]。解释层依托Grad-CAM技术生成故障区域热力图，基于特征提取层输出的高维特征图，计算目标故障类别对应的梯度信息，将梯度权重与特征图进行加权融合，转化为直观的热力图，清晰标记模型决策依赖的关键区域。验证层通过频谱分析匹配模型输出与物理特征，对输入信号进行频谱变换，提取故障对应的特征频率，将其与模型解释结果进行比对，验证模型决策是否符合机械振动物理规律。

2.2 关键技术实现

注意力权重分配策略注重与故障特征的关联性分析，采用基于通道注意力与空间注意力结合的混合策略。通道注意力通过计算不同特征通道与故障类别的相关性，强化故障敏感通道的权重；空间注意力则聚焦信号时域或频域中的故障特征集中区域，提升该区域的权重分配，确保权重分布与故障特征高度匹配。Grad-CAM热力图与机械振动频带的对齐方法，先将热力图标记的关键区域对应到原始振动信号片段，对该片段进行频谱分析，获取其频带分布；再结合不同故障类型的典型特征频带，建立热力图关键区域与特征频带的映射关系，实现两者在物理意义上的对齐。动态阈值调整机制针对不同故障类型的特征差异设计，通过分析历史故障数据中不同故障的特征强度分布，建立阈值与故障类型的关联模型；在框架运行过程中，根据模型初步识别的故障类型，自动调用对应阈值，对热力图与频谱分析结果进行筛选，确保仅保留高可信度的特征信息，提升框架对不同故障类型的适应性。

3 可解释性评估指标体系构建

3.1 现有评估方法的局限性

在故障诊断模型可解释性评估领域，现有方法存在明显短板。传统评估中常用的准确率指标，仅能衡量模型诊断结果与实际故障类型的吻合程度，却无法反映模型得出诊断结论的过程是否可解释，即无法判断模型是基于关键故障特征做出决策，还是依赖数据中的噪声或无关信息，导致仅以准确率为依据时，难以全面评估模型在工业场景中的实用价值。此外，现有评估还存在主观评价与客观量化结合缺失的问题。部分评估依赖专家经验进行主观判断，虽能结合领域知识给出定性反馈，但评价结果易受专家个人认知差异影响，缺乏统一标准；而单纯的客观量化评估，又难以充分融入机械故障诊断的专业理论，无法精准捕捉解释结果与物理规律的关联，使得评估结论要么过于主观、缺乏说服力，要么过于机械、脱离实际应用场景。

3.2 多维度评估指标设计

为弥补现有评估的不足，从三个核心维度设计可解释性评估指标。其一为逻辑一致性指标，以模型输出与振动理论匹配度为核心，通过对比模型识别出的故障特征与机械振动理论中已知故障类型对应的物理特征，判断模型决策逻辑是否符合客观物理规律，若二者匹配度高，则说明模型决策具有可靠的理论支撑，解释结果更具可信度。其中，模型识别的故障特征包括特征频率、振动幅值变化等关键信息。其二是特征显著性指标，聚焦注意力权重与故障频带的关联强度，通过分析注意力机制分配的高权重区域是否集中在机械振动理论明确的故障频带范围内，量化模型对关键故障特征的捕捉能力，关联强度越高，表明模型越能精准聚焦核心故障信息，解释结果的针对性越强。其三为决策可信度指标，重点衡量解释方法对诊断结果置信度的提升效果，通过对比引入解释方法前后，用户或系统对模型诊断结论的信任程度变化，评估解释方法在消除“黑箱”疑虑、增强决策信任方面的作用，若解释后置信度显著提升，则证明解释方法有效改善了模型的可接受度。

3.3 量化评估方法

为确保评估指标的有效应用，设计配套的量化评估方法。一方面，构建基于对比实验的指标验证流程，选取未引入可解释性增强方法的基础故障诊断模型作为对照组，引入融合注意力机制与Grad-CAM的模型作为实验组，在相同的轴承故障数据集上开展实验，分别计算两组模型在各评估指标上的数值，通过对比差异验证指标对可解释性差异的区分能力，同时排除数据分布、实验环境等无关因素对评估结果的干扰[3]。另一方面，建立评估指标的权重分配与综合评分模型，结合工业场景对各维度指标的需求优先级，采用层次分析法等科学方法确定逻辑一致性、特征显著性、决策可信度的权重，再通过加权求和计算模型可解释性的综合评分，将多维度指标转化为直观的量化结果，为不同故障诊断模型的可解释性对比提供统一、全面的评价标准。

4 实验设计与验证

4.1 实验环境与数据集

实验环境搭建需满足故障诊断模型训练与解释框架运行的稳定性需求，硬件方面选用具备高效计算能力的处理器与显卡，保障大规模数据处理与模型迭代效率；

软件方面搭配主流深度学习框架与数据处理工具，确保模型构建、信号分析与结果可视化的顺畅衔接。数据集选取聚焦轴承故障场景，选取标准需覆盖多类典型轴承故障类型，同时包含不同故障程度、不同运行工况下的样本数据，以体现数据的多样性与代表性，满足复杂工业场景的模拟需求。数据集预处理环节，针对原始振动信号中的环境噪声与干扰信号，采用滤波算法进行降噪处理；对信号长度、采样频率不一致的样本进行标准化调整；通过时域或频域转换将原始信号转化为模型可识别的特征形式，为后续模型训练与特征提取奠定基础。对比组设置方面，基准模型选取未引入可解释性增强方法的传统深度学习故障诊断模型，解释框架组则采用融合注意力机制与Grad-CAM的模型，两组模型基于相同的网络结构与训练参数开展实验，以排除无关变量对实验结果的干扰。

4.2 实验流程

实验流程分为三个核心阶段。第一阶段为模型训练与解释框架部署，先将预处理后的数据集划分为训练集、验证集与测试集，用于模型参数训练、超参数优化与性能验证；对基准模型与解释框架组模型分别进行训练，直至模型在验证集上的诊断性能趋于稳定；随后在解释框架组模型中部署注意力机制权重可视化模块与Grad-CAM热力图生成模块，确保解释功能正常运行。第二阶段开展解释结果与振动理论的交叉验证，从测试集中随机选取不同故障类型的样本，通过解释框架获取注意力权重分布与故障区域热力图，提取其中标注的关键故障特征；结合机械振动理论中的频谱分析方法，从同一样本的振动信号中提取理论故障特征；对比两组特征的一致性，验证解释结果是否符合物理规律。第三阶段进行评估指标的量化计算与对比分析，依据前文构建的可解释性评估指标体系，分别计算基准模型与解释框架组模型在逻辑一致性、特征显著性、决策可信度三个维度的指标数值；通过横向对比两组指标差异，明确解释框架对模型可解释性的提升效果。

4.3 预期验证结果

预期验证结果将从两个关键维度体现解释框架的有效性。其一，解释框架能实现对关键故障特征的识别准确率提升，相较于基准模型仅依赖数据驱动的特征提取方式，解释框架通过注意力机制聚焦核心故障信息，结合Grad-CAM精准定位故障区域，可减少对冗余信息与噪声的依赖，使模型在测试集中对各类轴承故障关键特征的识别更精准，降低因特征误判导致的诊断偏差。其二，诊断结果可信度在评估指标下呈现量化增长，在逻辑一致性指标上，解释框架提取的故障特征与振动理论特征的匹配度高于基准模型；在特征显著性指标上，注意力权重与故障频带的关联强度显著提升；在决策可信度指标上，引入解释方法后，用户或系统对诊断结论的信任程度明显高于基准模型，三个维度的指标数值均能证明解释框架有效改善了模型的决策信任度，为工业场景应用提供可靠支撑。

结束语

通过构建融合注意力机制与Grad-CAM的解释框架，结合机械振动理论实现决策逻辑验证，并建立多维度评估指标体系，有效缓解了深度学习模型的“黑箱”问题。设计的实验方案与预期验证结果表明，该解释框架能够提升关键故障特征的识别精度，同时促进诊断结果可信度的提升。后续研究可进一步拓展数据集范围，覆盖更多旋转机械故障类型，同时优化解释框架的实时性表现，使其更适配工业现场的动态监测需求，为故障诊断技术向智能化与可信化方向发展提供更全面的支持，助力解决工业场景中模型决策信任度不足的实际问题。

参考文献

[1]陈泽,刘文泽,王康德,等.光伏阵列故障诊断的可解释性智能集成方法[J].电力自动化设备,2024,44(6):18-25.

[2]张昊,王海茹,马继东.基于数据增强的可解释旋转机械故障诊断[J].电子测量技术,2025,48(8):105-115.

[3]韩谯,刘京,何国林,等.面向工业机器人关键部件的多源融合感知智能故障诊断方法研究[J].振动工程学报,2025,38(6):1252-1259.