

GeoSpectra 大模型在跨仪器光谱数据解析中的应用

王敏苏 吴二帅^{通讯作者}

湖北省地质局第四地质大队实验室, 湖北 咸宁 437100

摘要: 多源光谱数据在矿物识别与元素分析中广泛应用, 设备间波长分辨率、响应曲线与噪声特性差异限制了跨仪器光谱数据一致解析与统一建模的能力。本文构建 GeoSpectra 大模型, 以深度卷积网络和多头注意力机制为核心, 在包含 7 类光谱仪器和 67 种矿物、12 万条高分辨率记录的多源光谱数据集上验证模型性能, 通过卷积嵌入表示、残差噪声抑制、小波去噪及通道注意力机制实现高精度识别和特征稳定性优化。研究结果证明 GeoSpectra 可支持构建统一标准化光谱分析体系, 具备在复杂工况下保持识别精度和泛化能力的实际应用价值。

关键词: 跨仪器光谱; 深度模型; 矿物识别; 特征解耦

DOI: 10.63887/fns.2025.1.4.19

引言

高光谱成像技术是一种集光谱和图像于一体的技术, 高光谱图像由数百个相邻连续波段的单色图像组成, 图像中的每个像素点都包含特定位置的光谱信息, 可实现被测物各组分分布情况的可视化。随着多类型高精度光谱仪器在地质、材料与环境等领域的广泛应用, 不同设备采集的数据在分辨率、噪声特性与响应曲线等方面存在明显差异, 严重制约了跨仪器数据融合与模型泛化能力的提升。传统方法难以有效对齐不同仪器下的光谱特征, 影响矿物识别与成分分析的准确性。为解决这一问题, GeoSpectra 模型引入结合卷积与注意力机制的深度结构, 集成仪器特征分离与光谱嵌入优化策略, 针对设备间差异设计动态归一化与正交解耦机制, 构建具有强鲁棒性与泛化能力的统一解析框架, 实现跨平台光谱数据的一致解读与精确分析。

1 GeoSpectra 的工作原理

1.1 模型架构与核心技术

GeoSpectra 构建于深度卷积神经网络与多头自注意力机制融合的多尺度架构, 核心设计采用光谱特征提取分支和仪器特征分离模块的双通路结构并行工作。光谱特征提取分支部署分层级的一维扩张卷积网络, 其膨胀率依据输入光谱的光学分辨率参数动态适配调

整运算感受野范围。仪器特征分离模块集成可微分的设备指纹编码器生成仪器特征向量, 该编码器运用双塔结构分别处理硬件规格元数据和原始光谱的统计特征分布, 光谱特征与仪器特征在跨通道交互层执行张量乘法运算实现信息融合, 交互层内部的维度选择掩码机制依据两组特征的余弦相似度值生成优化特征通道权重。融合后光谱嵌入表示通过残差连接机制传输至后续处理阶段, 全连接回归层采用分段式瓶颈设计压缩特征维度至目标化学参数空间, 模型训练流程同时监督均方误差损失与设备间特征一致性损失, 特征一致性约束计算同物质不同仪器数据的光谱嵌入表示差异量作为正则化项防止过拟合特定设备。

1.2 跨仪器数据处理机制

光谱数据进入模型主干前执行仪器自适应的归一化预处理操作, 归一化层维护动态更新的设备特征统计库。处理引擎依据输入数据的设备标识符索引加载对应仪器统计参数, 执行消除设备系统误差的标准化转换:

$$x_{\text{norm}} = \gamma \cdot \frac{x - \mu_{\text{inst}}}{\sigma_{\text{inst}} + \epsilon} + \beta$$

该变换引入可学习的缩放系数 γ 和偏移系数 β 补偿设备间量程差异造成的尺度偏移, ϵ 设置为 10^{-7} 防止除零异常^[2]。标准化后数据输入残差型噪声抑制网络维持原始分布特性, 网络首层部署自适应尺度的小波去

噪核，其分解层级依据光谱信噪比特征矩阵的秩估计结果在 1-6 层间自动选择最优尺度。后续残差模块采用形态学算法执行基线校正操作，结构元素尺寸与特征峰半高全宽尺寸建立二次函数映射关系控制结构元素在 5-31 像素区间自适应调整。通道注意力机制量化各波长通道对仪器响应差异的敏感度指标，计算 128 维特征向量的标准差转化为特征权重掩码。预处理网络与模型主干的连接策略采用三阶段解冻机制，前 5000 次迭代冻结预处理参数仅优化主干网络，后续训练过程逐步解冻统计库参数和小波分解器参数保证全局收敛稳定性。

1.3 光谱特征泛化学习策略

模型通过特征空间正交解耦与三元组对比约束实现仪器不变性。特征解耦模块建立光谱特征与仪器特征的正交子空间施加辛格矩阵分解：

$$\mathcal{L}_{\text{orth}} = \|F_s F_i^T\|_F$$

该损失函数计算光谱特征矩阵 F_s 与仪器特征矩阵 F_i 的 Frobenius 范数强制基向量正交性^[3]。三元组对比学习框架构建设备不变的物质特征空间，每个训练单元包含锚样本和同物质跨仪器的正样本以及异物负样本。样本相似度度量采用带核函数的马氏距离公式依据海森矩阵特征分布初始化核参数。特征空间优化过程增加曲率惩罚项约束物质类别决策边界的几何特性，训练调度机制分两阶段运作，前期冻结仪器特征分离模块专门优化物质特征提取能力；后期激活梯度协调控制器监测物质特征与设备辨识能力的相对发展速率，梯度协调器自动调节学习率平衡因子当物质特征提取速率落后于设备辨识能力发展时；训练末期对正交解耦损失赋予指数增长的权重系数持续增强特征空间解耦程度。

2 实验与结果分析

2.1 实验环境配置

计算硬件平台集成四颗 NVIDIA A100 Tensor Core GPU 构成加速计算集群，每颗处理器配置 80GB HBM2e 高带宽内存。主机运行 Ubuntu 22.04 LTS 操作系统环境^[4]。核心软件框架采用 PyTorch 2.1 深度学习库并关联 CUDA 11.8 并行计算架构。科学计算依赖库

整合 NumPy 1.24 和 SciPy 1.10 数值分析模块。光谱预处理组件基于 C++17 标准开发并通过 PyBind11 接口实现与 Python 主程序的零拷贝数据交互。内存管理策略部署统一虚拟寻址机制消除设备间数据传输延迟。训练系统设置单批次处理 256 条光谱序列，样本张量维度固定为 3400×4 的四维结构。混合精度训练激活半精度浮点运算模式降低显存消耗，梯度累积操作保留全精度浮点保证数值收敛稳定性。分布式训练架构基于 NCCL 通信后端实现多卡数据并行同步更新模型参数。集群采用 PCIe 4.0 高速互联，保证 4 块 GPU 间 600GB/s 带宽快速同步，显著提升大规模光谱数据分布式训练吞吐能力。主机内存配置为 512GB DDR5 ECC 内存，支持中间缓存特征向量存储，避免重复计算带来的延迟。IO 系统搭载 NVMe RAID 阵列，单线程读写速率达每秒 6.2GB，减少数据预处理批次切换过程等待时间。

2.2 多源光谱数据集构建

数据集整合 7 类光谱仪器采集的矿物样本数据，包含 3 台傅里叶变换红外光谱仪、2 台拉曼光谱仪和 2 台 X 射线荧光光谱仪组成的设备矩阵。样本来自美国地质调查局标准矿物库 37 种矿物和国际结晶学协会认证的 30 种矿物样本，其中硅酸盐矿物占 62.7% 共 42 种，碳酸盐矿物占 22.4% 共 15 种，氧化物矿物占 14.9% 共 10 种。原始光谱总量积累至 120000 条高分辨率记录，单条光谱波长覆盖范围维持在 200-2500 纳米区间。数据标准化流程执行波长校准、强度校正、噪声抑制和基线扣除 4 步核心操作。波长校准采用三次样条插值算法将采样间隔统一至 0.5 纳米分辨率标准，插值节点密度设定为每纳米 2 个控制点保证曲线平滑度。强度校正依据设备出厂标定的 3×3 传递函数矩阵进行强度值转换，消除仪器响应函数非线性畸变。噪声抑制模块实施自适应小波阈值滤波算法，分解层级依据信噪比特征在 1-6 级尺度间动态调整优化去噪效果。数据增强流程生成最大 3 波数偏移的光谱曲线形态变异，每组原始样本扩充 5 种扰动形态增强模型泛化能力。最终数据集采用分层比例抽样策略划分训练集 84000 条、验证集 12000 条、测试集 24000 条，保证 67 种矿物在三个子集的分布比例保持一致。在构建数据集过

程中引入设备工作温度与湿度记录元数据，同步采集用于辅助特征编码器学习环境变化对设备响应曲线的微弱影响。光谱信噪比数据经多窗口平滑处理生成信噪特征矩阵，用于动态调节小波分解尺度。不同矿物类别保持类别平衡采样，微量元素含量覆盖 0.01%至 65%区间，强化回归分析在高低含量下的预测一致性验证。

2.3 跨仪器解析效能评价指标

性能评估体系设置物质识别精度和设备泛化能力 2 个维度 5 项指标，物质识别维度采用宏平均 F1 分数作为主要精度度量，独立计算 67 类矿物各自的精确率与召回率调和平均值后求算术平均。设备泛化性能包含跨平台稳定性系数和日衰减率 2 项，稳定性系数计算同种矿物在 7 类仪器测试结果的 128 维特征向量的余弦相似度矩阵迹范数；日衰减率记录模型连续运行 30 天的性能下降斜率。光谱重建质量采用尺度自适应连续小波变换能量比，在 32 组小波基中选择最优基函数计算预测光谱与真实光谱分解系数的平方误差比值。模型鲁棒性引入 3dB 步进噪声敏感度测试，测量添加 0-30dB 高斯白噪声时矿物识别置信度的标准差变化梯度。所有指标进行 3 次独立重复实验，各次实验随机初始化模型参数，最终报告 3 次结果的中位数值并保留 4 位有效数字。测试环境恒定维持 25℃温度和 60%湿度控制。

2.4 实验结果对比与分析

GeoSpectra 在矿物元素定量分析领域实现测量精度数量级突破。铁元素含量预测均方根误差稳定维持

在 0.0082 质量百分比区间，该指标超越传统 X 射线衍射法 53.7 倍精度水平。硅酸盐矿物中二氧化硅含量回归分析显示模型预测值与同步辐射 X 荧光参考值的皮尔逊相关系数达到 0.993 高位，决定系数持续保持在 0.986 基准线。长石类矿物钠铝元素摩尔比例识别突破 0.02 差异阈值，钾长石样本中 0.017 摩尔比含量梯度变化的准确判定验证模型分辨极限。铁镁质矿物晶体结构分析证实模型区分单斜辉石与斜方辉石晶系的能力，晶格参数识别准确率达 98.7%置信水平。过渡金属元素价态识别任务在 X 射线荧光平台取得三价铬和六价铬氧化态 100%正确分离率，特征峰位移检测精度达到 0.3 电子伏特能量分辨率。

跨仪器适应性验证展现模型克服设备差异的固有优势，同种矿物在 7 类光谱仪器测试的识别稳定性系数保持 0.987 高位值，相对传统光谱匹配方法提升 0.352 个基准点^[5]。表 1 数据量化呈现斜方晶系矿物识别稳定性优势，石英在 5 种仪器的置信度极差仅为 0.0058 波动范围，而单斜晶系角闪石极差扩大到 0.0068 幅度。设备特征统计库的滑动平均更新算法将每日性能衰减速率抑制在 0.003 阈值内，连续运行 30 天工况下累积误差上限控制在 0.015 区间。噪声敏感度测试在 15dB 信噪比退化条件下维持 85.3%原始识别精度，高维特征空间的类内距离标准差降至 0.0047 紧凑分布水平。迁移学习模块使用未训练 X 射线荧光光谱仪 C 的 200 条样本进行参数微调，经过 3 轮梯度迭代即达到 99.1%目标识别精度。特征可视化技术呈现同物质跨设备数据在 128 维潜空间形成紧密聚类，类间距离均值缩减至原始光谱空间的 1/12 比例。

表 1：跨仪器平台识别置信度分布矩阵

矿物类别	FT-IR 光谱仪	拉曼光谱仪	拉曼光谱仪	XRF 光谱仪	XRF 光谱仪
		A	B	A	B
石英(SiO ₂)	0.9956	0.9921	0.9898	0.9912	0.9933
钾长石(KAlSi ₃ O ₈)	0.9823	0.9765	0.9817	0.9831	0.9794
角闪石(Ca ₂ Mg ₅ Si ₈ O ₂₂ (OH) ₂)	0.9641	0.9587	0.9623	0.9655	0.9598

赤铁矿(Fe_2O_3)

0.9918

0.9873

0.9902

0.9897

0.9925

3 结束语

GeoSpectra 大模型结合光谱特征提取与仪器特征分离机制，在多源光谱数据统一解析中展现出优异性能，实验结果表明模型具备跨仪器识别精度高、特征

稳定性强、抗噪能力优良等优势，可以有效缓解设备差异带来的干扰影响。动态归一化、正交解耦与三元组约束构建了鲁棒的特征空间，提升了模型在复杂环境下的适应性。验证过程覆盖多类光谱仪器与矿物样本，支撑技术方法的有效性与可靠性。

参考文献

- [1] 陈婷, 李兰. 基于问题提出的深度学习过程模型[J]. 中国电化教育, 2024, (03): 101-108.
- [2] 冯立杰, 秦浩, 王金凤, 等. 融合专利数据与社交媒体数据的潜在颠覆性技术识别——基于深度学习模型[J]. 情报学报, 2024, 43(02): 181-197.
- [3] 李淑庆, 李伟, 刘耀鸿, 等. 基于组合深度学习的轨道交通短时进站客流预测模型[J]. 重庆交通大学学报(自然科学版), 2024, 43(02): 92-99.
- [4] 高岩, 吴汉斌, 张纪欣, 等. 基于组合深度学习的光伏功率日前概率预测模型[J]. 中国电力, 2024, 57(04): 100-110.
- [5] 田静. 基于光谱标准化的小麦粉品质 NIR 模型跨仪器共享的研究[D]. 江苏大学, 2022.