系统工程理论与实践
Systems Engineering-Theory & Practice
ISSN 1000-6788,CN 11-2267/N

《系统工程理论与实践》网络首发论文

题目：基于多模态可解释模型的文化艺术品价值评估研究
作者：倪渊，许愿清，张健，潘小宇
收稿日期：2024-12-23
网络首发日期：2025-09-15
引用格式：倪渊，许愿清，张健，潘小宇．基于多模态可解释模型的文化艺术品价值评估研究[J/OL]．系统工程理论与实践.
https://link.cnki.net/urlid/11.2267.N.20250912.1430.049

doi: 10.12011/SETP2024-3174　　中图法分类号: J124　　文献标志码: A

基于多模态可解释模型的文化艺术品价值评估研究

倪渊[1], 许愿清[2], 张健[2,3], 潘小宇[4]

(1. 北京信息科技大学 商学院, 北京 100192; 2. 北京信息科技大学 管理科学与工程学院, 北京 100190; 3. 北京知识管理研究基地, 北京 100190; 3. 北京信息科技大学 计算机学院, 北京 100192)

摘　要　随着文化艺术品市场的快速发展, 艺术品交易平台亟需透明高效的智能价值评估服务, 以提升市场活跃度. 本文提出了一种基于多模态可解释模型的艺术品价值评估方法, 通过 Vision Transformer(ViT)、基于全词掩码的语言模型和双向门控循环单元网络分别提取图像、文本和数值数据特征, 并采用瓶颈注意力机制实现多模态特征融合. 同时, 引入注意力展开和可解释人工智能等方法提升模型透明度. 实证结果表明, 该模型在预测准确性和稳定性方面显著优于传统单模态和其他融合模型, 能全面解析艺术品的复杂价值特征. 不仅为艺术品交易提供智能化、可解释的估值工具, 助力市场流通与规范化, 还为政策制定者和金融机构提供科学依据, 以优化定价机制与监管策略, 推动文化艺术品产业高质量发展.

关键词　文化艺术品; 艺术品价值评估; 多模态融合; 可解释性

Research on cultural artworks value assessment based on a multimodal explainable model

NI Yuan[1], XU Yuanqing[2], ZHANG Jian[2,3], PAN Xiaoyu[4]

(1. School of Business, Beijing Information Science & Technology University, Beijing 100192, China;
2. School of Management Science and Engineering, Beijing Information Science & Technology University, Beijing 100190 China;
3. Research Center for Knowledge Management, Beijing 100190 China;
4. School of Computer Science, Beijing Information Science & Technology University, Beijing 100192, China)

Abstract　In light of the accelerated growth of the cultural art market, it is imperative that art trading platforms adopt transparent and efficient intelligent value assessment services to stimulate activity. This paper puts forward a multimodal, interpretable methodology for the assessment of artwork value. This methodology combines a Vision Transformer (ViT), an all-word mask-based language model, and a bidirectional gated recurrent network in order to extract features from image, text, and numerical data. A bottleneck attention mechanism enables multimodal feature fusion, while attention expansion and interpretable AI enhance transparency. The empirical results demonstrate that the model outperforms traditional unimodal and fusion models in terms of accuracy and stability, offering a comprehensive artwork value analysis. It provides an intelligent valuation tool and scientific support for policymakers and financial institutions to refine pricing and regulatory strategies, thereby promoting high-quality growth in the cultural art industry.

收稿日期: 2024-12-23

作者简介: 通信作者: 倪渊 (1984–), 汉, 男, 教授, 博士, 研究方向: 文化产品产权价值评估与数智服务管理,E-mail:niyuan230@163.com; 许愿清 (2001–), 汉, 女, 硕士研究生, 研究方向: 文化产品产权价值评估与数智服务管理,E-mail:13269269599@163.com; 张健 (1974–), 汉, 男, 教授, 博士, 研究方向: 大数据与智能决策,E-mail:zhangjian@bistu.edu.cn; 潘小宇 (1998–), 汉, 女, 硕士, 研究方向: 大数据与智能决策,E-mail:3240357935@qq.com.

基金项目: 国家重点研发计划青年科学家项目: 文化产品产权价值评估与确权标识应用技术研究 (2021YFF0900200);
Foundation item: National Key Research and Development Program for Young Scientists Project: Research on the Application Technology of Cultural Product Property Rights Value Assessment and Confirmation Identification (2021YFF0900200)
中文引用格式: 倪渊, 许愿清, 张健, 等. 基于多模态可解释模型的文化艺术品价值评估研究 [J]. 系统工程理论与实践. doi: 10.12011/SETP2024-3174.
英文引用格式: NI Yuan, XU Yuanqing, ZHANG Jian, et al. Research on cultural artworks value assessment based on a multimodal explainable model[J]. Systems Engineering — Theory & Practice. doi: 10.12011/SETP2024-3174.

Keywords Cultural artworks; Artwork value assessment; Multimodal fusion; Explainability

1 引言

艺术品作为人类文明的瑰宝, 兼具文化与经济双重属性, 其价值的体现不仅在于文化艺术本身, 更在于市场的认可与流通[1]. 近年国家对文化产业日益重视, 出台了《关于深入推进文化金融合作的意见》、《艺术品发展规划》等一系列政策促进文化艺术品产业发展, 助推文化经济新动能的形成[2]. 在政策引导下全国各地陆续设立约一千家艺术品交易中心, 为买卖双方提供了集艺术品展示、交易、鉴定、评估、金融服务于一体的线上综合性平台, 旨在突破时空限制, 推动文化艺术品从收藏走向流通. 北京工艺美术品交易平台、北京宋庄艺术品交易中心等示范平台投入运营短短几年来交易额就达到了数十亿元, 发展潜力巨大. 这些新兴平台依托大数据、人工智能等新质生产力, 开创了艺术品交易变现新局面, 其中智能化价值评估服务是不可或缺的关键环节.

艺术品流通变现的核心在于交易定价, 而定价的起点是价值评估. 文化艺术品智能化价值评估就是通过大数据和人工智能技术实现高效、低成本的艺术品估值. 这种数据驱动的理性分析补了传统主观评估的偏差[3], 不仅能快速提供市场价值参考, 反映市场主流偏好, 还能有效减少买卖双方在艺术品价值认知上的分歧, 降低交易中的不确定性和潜在风险. 同时, 智能技术的支持使平台的价值评估服务更具专业性和公信力, 从而扩大市场参与度, 提升交易规模和频率, 为未来艺术品市场金融化开发奠定基础. 然而, 文化艺术品智能化价值评估在实际应用中仍面临一系列问题. 一方面, 大多数平台主要依靠交易数据来辅助定价, 但艺术品的市场价值还取决于其视觉特征、艺术风格、文化背景等多重因素, 单一模态的数据难以全面体现艺术品的复杂价值属性, 导致估值结果波动性大. 另一方面, 现有艺术品估值过程信息披露不足且缺乏统一评估标准, 用户难以理解估值模型背后的价格形成逻辑[4]. 这种信息不对称的现象会削弱用户对评估结果的信任, 阻碍交易的顺利达成. 据中央美术学院发布的《2025 艺术品市场趋势报告》显示, 近年艺术品高端市场表现低迷, 许多高价值艺术品因定价依据不明而流通受阻, 形成挂牌频繁但成交率低的现象. 因此, 为了破解艺术品估值复杂和透明度低的难题, 具有多模态且可解释性的文化艺术品价值评估方法已成为研究的核心议题.

近年来, 大数据分析、自然语言处理、图像识别等技术的出现推动文化艺术品价值评估逐渐从专家经验驱动到数据智能驱动过渡. 早期研究者主要利用统计学相关理论构建价格指数来反映艺术品的一般价格水平变化, 然后进行价值评估. 艺术品价格指数的构建方法主要包括代表作法、重复销售法、特征价格法以及混合模型法等[5]. 这些方法主要针对同行业、同水平作者的相似作品, 评估会受到历史数据的局限, 难以全面预测艺术品的市场价值. 在新兴技术的推动下, 学者们开始利用机器学习和深度学习方法去挖掘艺术品的价值信息. 如 Ugail 等[6] 通过卷积神经网络 (CNN) 和长短期记忆网络 (LSTM) 等方法利用图像识别艺术品物理特征进行价值分析, Iigaya 等[7] 通过深度 CNN 识别艺术品图像的层级视觉特征的来预测其美学价值, Li 等[8] 运用文本分析的 BERT 模型和用于图像分析的 CNN 方法对艺术品价格进行分类估算, Liu 等[9] 考虑艺术品价格数据的时间序列特征, 构建融合单向和双向的 LSTM 模型预测价格. Indrawan 等[10] 结合 CNN、BERT 和常规神经网络, 拼接不同模型预测结果, 根据图像、文本和类别数据来评估画作价值. 这些研究展现了数据驱动的方法在深度挖掘艺术品多维价值上的巨大潜力.

综上, 已有研究逐步探索基于大数据和智能技术的价值评估方法, 但仍存在一些不足: 1) 多数研究仅依赖单一模态数据的输入进行文化艺术品的价值评估, 忽略了文化艺术品的多维度价值来源, 导致评估结果存在片面性; 2) 部分研究开始尝试引入多模态数据 (如融合图像和文本信息) 进行评估, 但多模态数据的融合仍然是一个难点. 现有方法通常只是简单地拼接不同模态的特征, 未能充分考虑各模态之间的相关性和互补性, 无法有效融合多模态数据的潜在信息, 影响了模型的准确性; 3) 目前大多数艺术品价值评估模型都属于 “黑箱” 模型, 虽然能够给出较为准确的预测结果, 但很难解释模型的预测依据. 这种缺乏透明度的模型难以让用户信服, 尤其在文化艺术品价值评估这样一个复杂且主观性强的领域, 评估结果的可解释性对用户信任至关重要. 针对现有研究的不足, 本文提出了基于多模态融合与可解释性的文化艺术品价值评估模型, 旨在全面整合文化艺术品的多模态信息, 提高模型的透明度和适应性, 提升艺

术品交易平台的估值服务能力, 激发艺术品市场潜力.

本文的研究创新与贡献主要体现在三个方面: 1) 针对文化艺术品价值评估的复杂性和异质性, 本文结合图像、文本、数值三种模态数据, 通过 ViT、MacBERT、Bi-GRU 模型分别提取数据特征, 并基于注意力机制的 MBT 进行多模态信息的深度融合, 构建了多维信息的价值评估模型, 充分利用不同模态之间的互补性和相关性, 全面捕捉文化艺术品的多维度价值; 2) 为了提高模型的透明度和可信度, 本文引入多种可解释性技术, 分别利用 Attention Rollout 算法、多头注意力、SHAP 算法可视化模型内部决策过程, 增强了价值评估结果的可信度, 为用户提供了直观的解释依据; 3) 针对艺术品市场行情的变化和潜在波动, 本文引入时序分析模块, 实现了动态更新机制, 使得模型能够更精确地反映市场对艺术品价值的最新评估, 提升了模型在实际应用中的灵活性和适用性.

2　文献综述

2.1　多模态数据驱动的价值评估

传统的价值评估方法多以单模态数据构建线性模型, 如基于回归分析、统计学习或简单机器学习模型进行估值预测, 这些方法内部逻辑清晰, 易于解释, 但预测效果不佳[11]. 面对复杂价值要素构成的文化艺术作品, 多模态的数据可以从不同角度展现艺术品的多层次价值, 能有效避免模态信息不足、关联挖掘不充分、语义表示建模范式单一[12] 等问题. 近年来多模态数据驱动的机器学习、深度学习模型已被广泛运用到各领域中. 张大斌等[13] 使用双向长短期记忆神经网络 (BiLSTM)、文本卷积神经网络 (textCNN) 和 SnowNLP, 融合价格特征、文本特征和新闻情感特征对玉米期货收盘价进行预测. 刘洋等[14] 通过 LSTM 和 CNN 模型抽取旅游评论的图文信息特征, 进行拼接融合用以预测酒店股票价格. 这些研究展示了多模态融合在提升预测性能方面的潜力.

在文化艺术品的多模态价值评估中, 各个模态特征的提取是关键. 在图像模态方面, 基于 CNN 的 YOLO[15]、mask R-CNN[16] 等模型擅长提取局部信息, 但艺术作品在表现上注重整体性和细节之间的协调, 如作品的构图、色彩搭配、空间分布等, 由于 CNN 依赖固定大小的卷积核和池化层, 其全局特征提取能力有限. 基于 Transformer 的 Vision Transformer 模型 (ViT)[17] 则能利用自注意力机制捕捉全局信息, 更适合文化艺术品的价值评估. 在文本模态方面, 静态词嵌入方法 (如 Word2Vec[18]) 在文本表示上存在局限性, 无法捕捉上下文信息, 难以处理艺术品描述中具有歧义的词汇. 动态词嵌入模型 (如 BERT[19]、WoBERT[20]) 则能适应不同上下文, 提升对艺术品描述的理解能力. 然而, 由于描述文本包含大量专有名词 (如作者姓名、艺术流派、历史背景), 这些模型难以捕捉与艺术品价值评估密切相关的特征. 在 RoBERTA 模型的[21] 动态掩码机制的基础上,MacBERT[22] 通过引入全词 MASK 机制, 优化预训练目标, 提高了对专有名词的表征能力, 更适用于文化艺术品的文本分析任务. 在结构化数据方面, 艺术品的市场价值因时间和需求变化而波动显著, 需要动态建模. RNN 对动态变化敏感, 能提取复杂的数值变化模式, 但在处理长时间序列任务易存在梯度消失的问题[23]. 门控循环单元 (GRU)[24] 在 LSTM 模型架构的基础上进行了简化, 将遗忘门和输入门合并为更新门, 并利用记忆单元和隐藏层构成重置门, 解决了训练过程中面临的梯度问题[25]. 双向 GRU 网络模型 (Bi-GRU) 进一步通过正向和反向传递特征, 可以同时捕捉时间序列中的历史和未来信息[26], 更适合于数值模态数据的特征提取. 基于 Transformer 的模型通过引入稀疏自注意力机制、自适应分解机制和频域变换等创新方法, 显著提升时间序列建模的效率与精度[27], 但该模型的复杂性和对数据量的高需求限制了其在艺术品交易数据中的应用.

在模态融合方面, 拼接、相加、相乘等简单融合方法难以有效学习模态间的复杂交互, 并导致高维度问题, 影响模型收敛[28]. 为此, 有研究采用基于图卷积网络 (GCN) 的方法, 以图结构表示模态间的关系[29], 但此方法需要对多模态特征的关联性有较为清晰的理解, 在文化艺术品多样化特征下实现较为复杂. 基于生成对抗网络 (GAN) 的方法在处理模态不均衡和生成高质量模态特征方面表现出色, 但 GAN 的训练过程复杂, 容易出现模式崩溃等不稳定性的问题[30]. 基于注意力机制的融合方法可以在特征层进行深度交互, 显著提升评估精度和模型鲁棒性[31].Nagrani 等[32] 提出了一种新的模态融合方式 Multimodal Bottleneck Transformer(MBT), 在 Transformer 架构的基础上,MBT 使用融合瓶颈 (FSN) 整合各模态的关

键信息, 能实现多层次的有效信息共享, 提高价值评估模型的性能.Tang 等[33] 提出了基于跨注意力机制的多模态融合网络, 能够有效捕获模态之间的互补特性, 但是在数据规模不足时容易出现过拟合问题. MBT 相比于传统的拼接、GCN、GAN 以及跨注意力机制等方法, 在计算效率、信息筛选能力和泛化能力方面均具有优势, 所以本文最终选择 MBT 作为核心的模态融合框架.

2.2 可解释性方法

模型的高精度不应以牺牲可解释性为代价[34], 可解释的人工智能旨在使复杂模型的内部决策过程透明化[35]. 在传统机器学习模型中, 线性回归、决策树等方法因其内嵌可解释性而便于理解[36], 但模型的拟合能力受限, 预测精度不高; 复杂的机器学习、深度学习算法虽然可以提供高预测结果的准确率, 但人们很难理解模型决策逻辑, 因而无法信任其在价值评估、信用评分、医疗保健以及军事领域等实际工作场景中的应用, 大大降低了模型的实际效益[37].

于是, 学者们针对特定的复杂模型设计了特定的可解释方法. 根据输入数据, 激活最大化[38] 可以最大程度地激活特定层的神经元, 可视化每个神经元的输入偏好, 帮助人们理解模型内部工作逻辑, 但可视化的图像辨识度低, 不能很好地满足人们的预期. 基于类激活映射方法 (CAM)[39] 可以利用神经网络的特征图得到原图各个部分的重要性, 帮助人们理解模型关注图像的重要部分. 但 CAM 仅适用于 CNN 结构, 无法直接应用于 Transformer 模型. 梯度解释方法[40] 通过计算模型输出对输入特征的梯度, 衡量各特征对最终决策的贡献, 但在艺术品价值评估中, 特征对决策的影响是有限度的, 超过某一阈值后, 梯度值会变为零, 可能导致模型忽略一些细微但重要的特征. 由于 Transformer 在自然语言处理、计算机视觉等领域的广泛应用, 学者们针对该架构的可解释性进行了相关研究.Attention Rollout 方法[41] 通过层层展开注意力矩阵, 追踪注意力的流动路径, 能够解释 Transformer 在不同层次上关注的信息. 多头注意力分析可以直接分析 Transformer 模型的不同注意力头, 理解其在不同特征上的关注度[42], 但在多层结构下注意力模式较为复杂, 不易直观呈现. 研究者还提出了适用于不同数据类型的通用可解释性框架:LIME[43] 不关心模型的内部结构, 通过局部线性近似解释模型决策, 但该方法的稳定性较差, 解释结果容易受采样数据影响. SHAP[44] 基于 Shapley 值衡量特征的重要性, 为模型提供全局和局部的解释. 知识蒸馏[45] 通过压缩模型大小来提取关键决策信息进行解释, 但其适用条件苛刻且过程不可控, 无法有效应用于需要深度理解的艺术品价值评估任务.

基于以上分析, 本文基于图像模态、文本模态、数值模态进行融合建模, 在图像模态上, 采用 ViT 模型提取图像中的文学和艺术价值特征; 在文本模态上, 通过微调 MacBERT 模型, 获得统一的文本表示; 在数值模态上, 利用 Bi-GRU 模型挖掘市场需求特征. 针对各模态之间的异构性和语义不一致性问题, 通过基于注意力机制的 MBT 进行特征融合. 在模型的解释分析中, 根据不同模态特征提取模型的特点, 分别应用 Attention Rollout 算法、多头注意力、SHAP 算法可视化其内部决策过程.

3 多模态可解释的文化艺术品价值评估模型构建

3.1 模型框架

在艺术品多模态数据的复杂性与模型可解释困难的双重挑战下, 本文提出了一种基于多模态可解释的文化艺术品价值评估模型. 针对艺术品价值跨度大、作者信息敏感、多模态信息丰富等特点, 本文首先通过混合模型对艺术品的图像、文本、数值模态信息进行特征提取与融合, 实现对艺术品价值的全面评估. 然后, 针对评估模型的复杂性和可解释性需求, 设计了多维度的解释方式, 确保在保证模型准确性的基础上, 提升其可解释性与透明度. 模型的整体框架如下图 1 所示, 包括价值评估模型和多维可解释模型两部分, 具体步骤如下:

步骤 1: 数据处理. 针对不同艺术品图片的分辨率差异, 对数据进行填充、归一化及尺寸缩放处理, 保证模型在特征提取时的稳定性. 文本数据来源于拍卖记录、艺术评论、作者信息等, 需进行数据清洗、去重、分词及语义标准化, 确保不同来源的文本数据可对齐. 市场价格、交易时间、拍卖行信息等数值模态数据通常存在缺失值及量纲不一的问题, 因此采用均值填补、标准化归一化等方法进行预处理.

步骤 2: 单模态特征提取. 单一模态的特征稳定性表征是多模态融合的基础, 面对不同模态选取最匹配的模型进行深度特征挖掘. 其中, 图像模态采用 ViT 进行艺术价值的抽取; 文本模态采用全词掩码的 MacBert 模型进行文化价值和社会价值的提取; 结构化数据采用 Bi-GRU 模型进行数值特征的关联.

步骤 3: 多模态融合预测. 多模态的有效融合是艺术品多维价值要素集成的核心, 本文采用基于瓶颈注意力的 MBT 进行特征融合, 用少量的 FSN 进行多个模态信息的特征压缩和交互, 融合后的多模态价值向量综合艺术品图像的艺术风格、文本的文化解读以及数值的市场趋势, 该融合特征用于全连接层进行最终价值预测, 进而实现文化艺术品价值的评估.

步骤 4: 图像模态的解释. 对于图像特征抽取模型, 利用 Attention Rollout 解释模型中 Transformer 结构的注意力矩阵, 可视化模型关注的关键图像区域.

步骤 5: 文本模态的解释. 对于文本特征抽取模型, 利用 MacBERT 模型的多头注意力机制, 对不同词语的影响权重进行可视化, 分析模型对艺术品描述、作者背景等信息的关注程度.

步骤 6: 数值模型的解释. 对于数值特征抽取模型, 由于数值信息由结构化数据构成, 模型的输入特征含义明确, 直接采用 SHAP 进行特征贡献度分析, 从全局和局部两个方面进行解释.

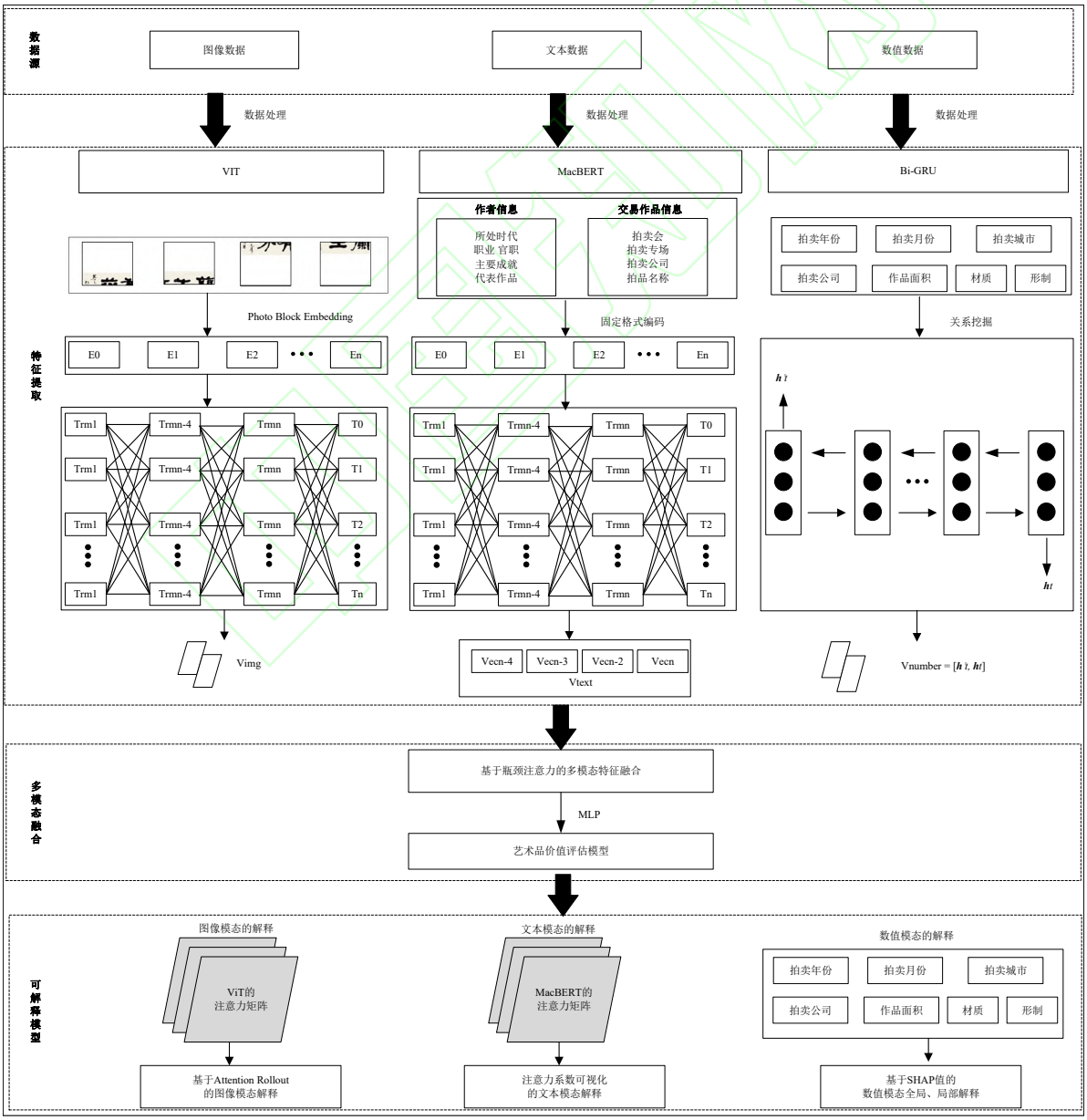


图 1　模型整体框架图

3.2 基于多模态融合的艺术品价值评估模型

3.2.1 图像特征提取

文化艺术品的图像模态中, 包含丰富的艺术风格、构图等特征, 反映了作品的艺术价值. 为了充分提取这些复杂多样的视觉特征, 需要采用能够捕捉到全局信息的图像特征提取方法. 相比常见的图像处理方法, 基于 ViT 能通过自注意力机制捕捉长距离依赖关系, 能够更好地建模图像中的全局特征, 尤其在处理大规模数据时, 其性能超越了传统 CNN, 弥补了全局信息提取的不足, 所以本文采用 ViT 模型实现图像中文学价值和艺术价值的挖掘.

首先, 将输入图像划分为固定大小的块, 并通过线性变换将每个图像块转化为特征向量, 同时引入位置编码以记录各图像块的位置信息. 随后, 利用 Transformer 编码器对艺术品图像进行全局特征提取. 其中,Transformer 网络通过不同 Transformer Encoder 层中多头注意力机制, 实现图像信息的处理和传播. 经过 12 层的编码,ViT 模型输出的 [CLS] 标记中包含图像的全局语义特征, 以便后续进行多模态的融合.ViT 计算过程可用公式 (1)-(4) 表示:

$$z_0 = \left[X_s; X_p^1 E; X_p^2 E; \ldots, X_p^n E\right] + E_p, \quad E \in \mathbb{R}^{p^2 \times C \times D}, \quad E_p \in \mathbb{R}^{(N+1) \times D}, \tag{1}$$

$$z_l' = \mathrm{MSA}\left(\mathrm{LN}(z_{l-1})\right) + z_{l-1}, \quad l = 1, \ldots, L, \tag{2}$$

$$z_l = \mathrm{MLP}\left(\mathrm{LN}(z_l')\right) + z_l', \quad l = 1, \ldots, L, \tag{3}$$

$$y = \mathrm{LN}\left(z_L^0\right). \tag{4}$$

其中, 图像划分为 N 个块, 每个图像块的大小为 $p \times p$, 图像的通道数为 C, 图像块的嵌入向量为 $\boldsymbol{E}$,D 表示线性变换后的特征大小, 图像块的位置编码为 $\boldsymbol{E}_p$. 第 l 层经过多头注意力 (MSA) 和残差连接获得的特征向量为 $\boldsymbol{z}_l'$, 第 l 层经过多层感知机 (MLP) 和残差连接获得的向量特征序列为 $\boldsymbol{z}_l$. 通过 L 个 Transformer 编码器层得到的特征向量为 $\boldsymbol{z}_L^0$, 最后 Transformer 编码器处理后的特征向量为 $\boldsymbol{y}$.

3.2.2 文本特征提取

目前艺术品的文本描述普遍存在信息不完整、用词不统一等现象, 增加了模型提取有效文本特征的难度. 在特征提取时, 不仅需要关注作品的描述性文本, 还需结合创作者的背景、社会关系及市场影响, 全面挖掘文本中隐含的价值信息. 为解决这一问题, 本文采用基于全词掩码的预训练语言模型 MacBERT 模型进行文本特征提取. 相比传统的 BERT,MacBERT 通过引入全词动态掩码机制, 使专有名词向量化表示更加精准, 保证语义一致性, 尤其适用于处理艺术品中大量出现的专有名词和短语.

在文本内容选择上, 本文重点提取反映艺术品选题立意、创作者背景及市场运作的信息, 包括作品标题、拍卖公司、拍卖会和拍卖专场. 又由于作者的社会关系、地位与艺术品的收藏投资价值密切相关, 本文引入思知知识图谱作为外部知识库, 整合作者所处时代、职业、官职、主要成就及代表作品等信息, 将其作为文本数据的概念集, 增强艺术作品的艺术投资价值表达. 为实现离散文本信息的语义关联, 本文设计了固定格式的句子模板, 将各类信息整合. 例如, 拍卖信息被格式化为: “在 auction 的 auction_session 专场上 auction_company 公司拍卖了 title”; 创作者信息被格式化为: “author, 所处时代:time, 职业 occupation, 官职:government_post, 主要成就:major_achievements, 代表作品:representative_works.” 通过这种方式, 文本信息以结构化形式输入模型, 提升其可理解性.

在 MacBERT 的训练任务中, 模型会在原始文本前插入 [CLS] 标记, 并在每个分句后插入 [SEP] 标记, 经过多层 Transformer 编码器提取文本特征, 模型将 [CLS] 标记对应的输出向量作为整段文本的语义表示. 同时通过多头注意力机制, 将每个单词与上下文关联, 丰富语义表达. 浅层的 Transformer 块提取基本的语法、句法信息, 高层的 Transformer 捕捉更抽象的语义概念和复杂的语义结构. 为更全面地提取潜在价值信息, 本文将 MacBERT 模型最后四层的编码结果组合拼接为句子向量, 以支持后续多模态融合的任务.

3.2.3 数值特征提取

在艺术品价值评估过程中, 国内经济环境、政府政策和市场供需也是重要影响因素. 这些因素随着市场环境不断波动, 其内在规律难以直接捕捉. 但是, 随着数据挖掘技术的应用, 可以根据近十几年的艺术品交易记录, 通过深度学习模型自动揭示价格随市场环境变化的潜在规律. 因此, 数值模态的数据, 主要包括拍卖时间、拍卖公司、作品面积等, 是衡量艺术品市场价值的重要依据.

为了同时捕捉时间序列中的未来信息和历史信息, 并提升模型对复杂模式的识别能力, 本文采用 Bi-GRU 结构进行数值特征挖掘. 通过拼接正向 GRU 和反向 GRU 的输出, 生成数值关系特征向量, 以便后续多模态融合.Bi-GRU 的核心部件是重置门和更新门, 负责实现记忆的存储与更新, 其内部处理逻辑用数学形式表示为公式 (5)-(10).

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]), \tag{5}$$

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]), \tag{6}$$

$$\tilde{h}_t = \phi(W_{\tilde{h}} \cdot [r_t \cdot h_{t-1}, x_t]), \tag{7}$$

$$h_t = (I - z_t) \cdot h_{t-1} + z_t \cdot \tilde{h}_t, \tag{8}$$

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \tag{9}$$

$$\phi(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}. \tag{10}$$

其中,h_{t-1} 为上一时刻的状态变量,r_t 为更新门状态变量,z_t 为重置门更新状态,$\tilde{h}_t$ 为当前时刻的候选状态变量,h_t 为当前时刻的状态变量.W_r、W_z、$W_{\tilde{h}}$ 为输出向量与 x_t、h_{t-1} 矩阵相乘的权重系数.σ 表示 *sigmoid* 激活函数,ϕ 表示 *tanh* 激活函数.

3.2.4 多模态融合

对单模态数据进行特征挖掘后, 获得图像特征向量、 文本特征向量和数值特征向量. 由于每个单模态特征的维度较高, 若直接将三个模态的特征拼接后通过自注意力机制进行重新编码, 不仅会显著增加计算复杂度, 还可能导致特征冗余问题, 影响模型的收敛效果. 为此, 本文采用多模态瓶颈 Transformer(MBT) 方法, 构建包含输入层、多模态融合层和回归预测层的多模态融合框架.MBT 通过引入融合瓶颈 (FSN), 在多模态融合过程中对不同模态的信息进行筛选、压缩和跨模态交互, 以降低计算复杂度, 并提高特征融合的有效性. 其核心思想是通过瓶颈注意力机制在特征层级进行跨模态信息共享, 从而增强不同模态间的协同作用, 避免简单特征拼接带来的信息损失. 具体融合步骤如下:

1) 单模态特征筛选: 首先通过 FSN 模块对图像模态特征进行筛选, 剔除冗余信息, 仅保留与艺术品价值评估高度相关的内容, 减小计算复杂度并提高信息质量. 2) 模态特征融合: 将浓缩了图像信息的 FSN 模块添加到文本模态特征中, 利用自注意力机制进行模态信息的压缩. 随后, 将融合了图像和文本内容的 FSN 模块添加到数值模态特征中, 再利用自注意力机制进行三个模态的关键内容融合. 3) 融合后的多模态价值向量同时整合了图像模态的文化艺术价值、文本模态的投资收藏价值以及数值模态的市场收益价值, 之后该多模态向量被输入到全连接层, 用于对艺术品的综合价值进行回归预测, 评估框架如图 2 所示.

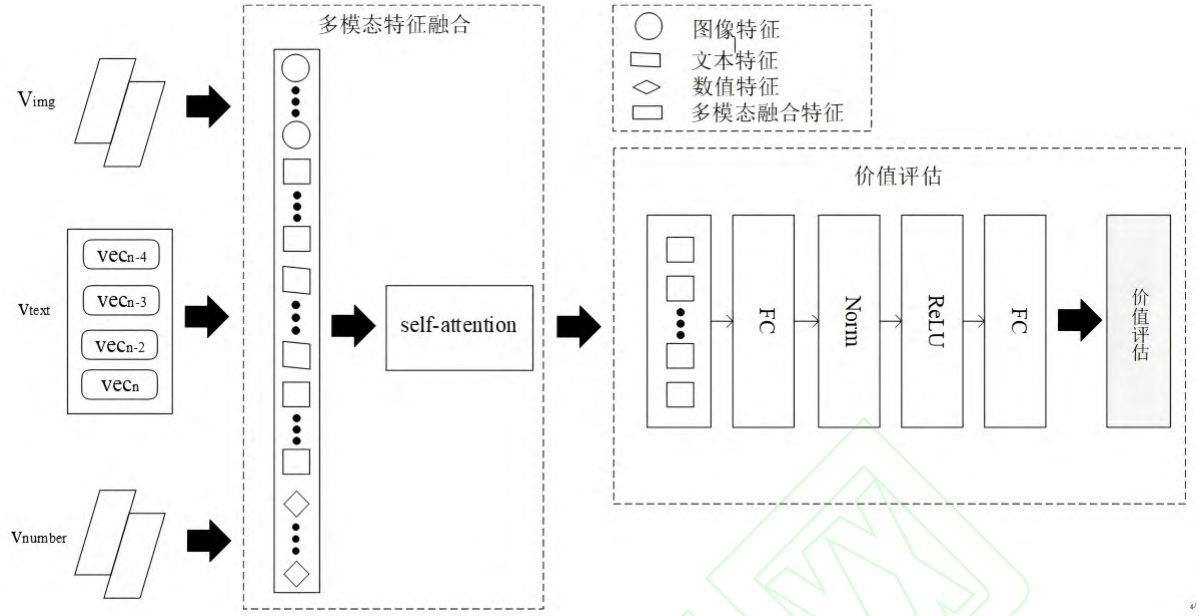


图 2 多模态融合的艺术品价值评估框架

3.3 艺术品价值评估模型的可解释性设计

3.3.1 图像模态的解释

每件文化艺术品在色彩浓淡、构图起伏及其他细节处理方面都能够展现出独特的视觉效果, 相比于全局可解释方法, 局部可解释算法能够更加精准地识别艺术品的关键视觉要素, 更适合用于突出文化艺术品的个性化与艺术性. 本文采用 Attention Rollout 方法对 ViT 模型的艺术品图像模态进行可解释性分析. 该方法通过计算不同 Transformer 层的注意力分布, 追踪模型在图像输入上的注意力流向, 观察图像中哪些区域与作品的价值相关. 由于 ViT 网络采用多层自注意力机制, 每一层的注意力信息会在层间逐步叠加, 最终的注意力矩阵能够汇聚整个图像的特征信息, 累积注意力权重的计算方式如下:

$$\tilde{A}_i = \begin{cases} A_i \tilde{A}_{i-1}, & \text{if } i > j \\ A_i, & \text{if } i = j \end{cases} \tag{11}$$

其中,$\tilde{A}_{li}$ 表示第 l 层的注意力输出结果,A_{li} 是第 l 层的原始注意力,$j = 0$, 表示从第一层 transformer 模块开始计算, 通过层间的递归计算, 浅层的局部特征在深层结构中得以累积, 从而形成最终的全局特征表达. 本文利用热力图可视化加权特征的注意力分布, 理解模型在图像输入中的关注重点, 从而揭示其在评估艺术品价值时的内在逻辑.

3.3.2 文本模态的解释

文本特征的提取通过 Transformer 编码器完成, 而 Transformer 的内置注意力机制能够自动评估文本中不同词语对艺术品价值的贡献, 本文采用多层注意力可视化方法对文本模态进行详细的可解释性分析.

1) 多层 Transformer 注意力可视化. Transformer 不同层级的编码器对文本特征的关注点有所不同, 浅层主要捕捉局部词语的细粒度语义信息, 如基本的词法结构和词义组合关系, 中层编码器进一步学习词汇间的长距离依赖性, 随着层次加深, 高层次编码器逐步提取更复杂的语义特征, 聚焦于整体语境. 作为 Transformer 预训练模型的全局语义标记,[CLS] 标记在 Transformer 结构中承担了句子级别的特征聚合作用, 其注意力分布可以反映模型对不同词语的重要性评估. 通过可视化每一层中 [CLS] 标记与其他词语之间的注意力权重, 可以观察到 Transformer 编码器如何逐步从文本输入中提取语义信息, 以及不同层级的模型对文本内容的关注变化.

2) 最后四层 Transformer 多头注意力矩阵可视化. 在高层编码器中, 提取的特征已高度语义化. 本文基于最后四层 Transformer 的多头注意力矩阵, 对艺术品描述文本、作者信息等内容中的关键词进行可视化, 分析这些词语对艺术品价值评估的影响. 通过词语颜色深浅的变化 (颜色越深代表权重越高), 直观展示关键词的影响程度, 揭示模型如何综合不同文本特征来支持艺术品价值评估.

3.3.3 数值模态的解释

结构化信息包括艺术品的拍卖年份、拍卖月份、拍卖公司、拍卖城市、尺寸面积、形制、材质等市场信息和作品自身属性. 这些特征隐含了不同时间节点上艺术品交易市场对作品价格的动态影响. 为深入解析这些特征在价值评估模型中的作用, 本文引入基于博弈论的 SHAP 解释框架, 从全局和局部两个视角分析 Bi-GRU 模型输入端各特征对预测结果的贡献程度及影响方向. SHAP 框架基于合作博弈理论, 通过计算特征变量在所有可能特征组合下的边际贡献, 量化各变量对最终估值结果的影响, 从而提供透明的解释机制全局解释通过计算所有样本的平均 Shapley 值, 量化每个特征在整个数据集上的平均贡献, 揭示特征的重要性排序及其对评估结果的整体影响. 局部解释则针对单个样本输入, 分析各特征变量对该样本预测结果的具体影响方向和影响力, 提供个性化的解释视角.

4 实验设计

4.1 实验数据与评价指标

本实验所用数据来源于北京保利、中国嘉德、中贸圣佳、北京翰海等中国艺术品拍卖网站, 涵盖近五年的艺术品交易记录, 共计 30841 条. 数据集中的模态信息包括图像模态: 作品的图像信息, 主要用于提取艺术品的视觉风格、色彩搭配等视觉特征; 文本模态: 包括艺术品的描述信息、拍卖公司、拍卖会、作者简介等文本信息, 作为艺术品的文化、社会价值的表达; 数值模态: 艺术品的拍卖年份、拍卖月份、拍卖公司、拍卖城市、尺寸面积、形制、材质等结构化数据, 用于分析作品在市场上的表现及其市场价值. 目前的数据集主要基于书法艺术品, 书法作为中国传统文化的重要组成部分, 其价值评估涉及艺术风格、书写技法、历史文化背景和市场认可度等多维因素, 涉及视觉、文本和市场特征, 能够较为全面地体现文化艺术品的多维价值. 并且不同艺术品类别 (书法、绘画、瓷器等) 在各数据模态上具有较强的共性, 因此模型具备在不同类别的艺术品数据上进行验证的潜力.

由于艺术品的市场估值呈现长尾分布, 即高端艺术品数量较少, 低价艺术品占比较大, 直接随机划分可能会导致训练集中高端艺术品比例较低, 影响模型学习的稳定性. 为确保模型的训练稳定性和泛化能力, 本文采用随机划分策略, 按照 7:3 的比例将数据集划分为: 训练集 (占比 70%, 约 21588 条样本), 用于模型训练; 测试集 (占比 30%, 约 9253 条样本), 用于模型在未见数据上的评估. 在数据划分过程中, 采用分层抽样方法, 确保训练集和测试集在价格区间、拍卖时间、艺术风格等维度上的数据分布保持一致, 以防止模型在测试时因数据分布不均而出现偏差.

为量化模型在艺术品价值评估任务中的预测准确性, 本文采用均方误差 (MSE)、均方根误差 (RMSE) 和平均绝对误差 (MAE) 作为模型的评价指标[46], 定义如下:

均方误差 (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2, \tag{12}$$

均方根误差 (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}, \tag{13}$$

平均绝对误差 (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|, \tag{14}$$

其中,$\hat{y}_i$ 表示第 i 个样本的预测值,y_i 表示第 i 个样本的真实值,n 表示样本的数量.MSE、RMSE 和 MAE 的数值越小说明模型具有更低的误差水平和更好的预测性能.

4.2 实验过程

1) 数据预处理. 将图像数据大小调整统一, 归一化处理, 使图像数据符合预训练的 ViT 模型输入要求. 将文本数据通过 MacBERT 模型进行预处理, 包括包括分词和全词 MASK 操作. 为保持词汇的语义完整性, 采用固定格式的句子模板, 将拍卖描述和创作者简介转化为序列输入. 将结构化数值数据中拍卖年份、拍卖月份、拍卖公司等进行编码处理, 并对缺失数据进行插值填补. 所有数值数据均归一化处理, 以确保数值特征在模型输入时具有一致的尺度范围.

2) 多模态价值评估模型构建. 结合图像、文本和数值三种模态信息全面评估艺术品的价值. 图像模态利用 ViT-large 进行全局特征抽取, 同时引入 GoogleNet、Resnet50 和 ViT-base 进行对比分析, 实验结果验证了 ViT-large 在提取视觉信息中的优越性. 文本模态利用 MacBERT 模型对特征进行编码, 主要通过字符 MASK 的 Bert、ALBERT 和全词 MASK 的 BERT-WWM、RoBERTa-WWM、MacBERT 进行对比分析, 实验结果表明 MacBERT 的全词 MASK 机制在文本语义表达上效果最佳. 数值模态采用 Bi-GRU 模型挖掘潜在的市场价值信息, 并利用简化门控机制解决长距离依赖问题, 有效提升了训练效率. 在多模态融合中, 通过与拼接、相加、投票策略的对比, 证明了基于注意力融合的 MBT 可以整合各模态特征, 实现信息的深度互补和高效融合, 显著提升了模型的预测准确性.

3) 可解释性分析. 本文对艺术品图像、文本、数值三个模态的特征抽取模型进行可解释分析. 图像模态解释采用 Attention Rollout 方法生成多层注意力热力图, 结合多头注意力权重矩阵的 max、min、mean 分析, 揭示模型对图像关键区域的关注. 文本模态解释中通过 MacBERT 模型的 Tramsformer 注意力权重进行逐层分析, 重点分析最后四层注意力分布的变化, 展示模型在文本特征提取中的语义聚焦能力. 数值模态的解释利用 SHAP 框架, 从全局和局部两个视角解释各特征对预测结果的贡献和影响方向. 全局层面量化特征重要性, 局部层面揭示单个样本中的决策依据.

5 实验结果与分析

5.1 单模态价值评估模型性能分析

为验证不同卷积神经网络和 transformer 网络在提取图像特征上的表现、不同预训练语言模型的语义表征能力, 以及不同的序列模型与艺术品结构化数据的匹配程度, 本文对三个单模态价值评估模型进行性能分析, 并与已有研究中主流方法进行对比, 实验结果如表 1 所示,

表 1　单模态价值评估的性能对比

模态类型	模型	MSE	RMSE	MAE
图像模态	GoogleNet	0.0394	0.1984	0.1063
	Resnet50	0.0278	0.1667	0.0941
	ViT-base	0.0385	0.1962	0.1038
	ViT-large	**0.0169**	**0.1300**	**0.0758**
文本模态	BERT	0.0288	0.1697	0.0934
	ALBERT	0.0206	0.1435	0.0899
	BERT-WWM	0.0184	0.1356	0.0851
	RoBERTa-WWM	0.0167	0.1292	0.0776
	MacBERT	**0.0129**	**0.1135**	**0.0709**
数值模态	RNN	0.0401	0.2002	0.1133
	Bi-RNN	0.0395	0.1987	0.1087
	LSTM	0.0393	0.1982	0.0972
	Bi-LSTM	0.0312	0.1766	0.0918
	GRU	0.0297	0.1723	0.0871
	Bi-GRU	**0.0189**	**0.1374**	**0.0891**

在图像模态的性能分析中, 传统卷积神经网络模型 GoogleNet 和 ResNet50 主要依赖于局部特征提取, 在艺术品图像的复杂背景、纹理特征提取方面具有一定局限性[47]. 相比之下, 基于 Transformer 架构的 ViT 采用自注意力机制, 能够捕捉长距离依赖关系, 适用于解析艺术品的整体布局和艺术风格[48]. 实验结果表明,ViT-base 在预训练数据不足的情况下, 其 MSE、RMSE 和 MAE 均较高, 这可能是由于 Transformer 结构对大规模训练数据依赖性较强, 导致在数据较少的情况下学习能力受限. CNN 通过卷积操作提取局部信息, 而 ViT 依赖自注意力机制, 全局信息建模的能力需要大规模数据来优化权重参数, 在小数据集上容易出现欠拟合或高方差问题[49]. 在充足数据训练后,ViT-large 的性能显著提升, 明显优于其他模型,MSE 降至 0.0169,MAE 仅为 0.0758, 表明 ViT-large 在捕获整体构图、空间布局和色彩和谐等艺术特征方面具有显著优势.

在文本模态中, 本文对比了 BERT、ALBERT、BERT-WWM、RoBERTa-WWM 和 MacBERT 五种预训练语言模型的文本表征能力. 传统的 BERT 采用子词级别的 Masking 机制, 在处理文化艺术品描述、作者信息等文本数据时, 容易破坏专有名词的语义完整性, 影响模型的语义表达能力.ALBERT 优化了 BERT 的计算效率, 但由于模型轻量化, 其语义表达能力略逊于 BERT, 在高精度文本分析任务上的表现有限[50]. BERT-WWM 采用全词掩码机制, 避免了语义割裂问题[51].RoBERTa-WWM 在 BERT-WWM 的基础上通过更大规模数据和更长训练时间优化了预训练过程, 提高了模型的泛化能力[52]MacBERT 进一步优化了全词掩码机制, 并引入动态 Masking 和拼写纠正任务, 增强了模型对艺术品名称、拍卖行、作者背景等文本信息的理解能力[22]. 实验结果表明, 基于完整词语 MASK 的 BERT-WWM、RoBERTa-WWM 和 MacBERT 均显著优于 BERT 和 ALBERT. 尤其是 MacBERT, 其 MSERMSE、MAE 均达到最佳. 同时,MacBERT 因减轻了预训练和微调阶段之间的差距, 使得模型在文本价值评估微调过程中, 句子语义表示更完整、准确, 评估效果显著提升.

在数值模态中, 艺术品的市场数据呈现明显的时间序列特征, 其中包括拍卖价格、成交时间、市场活跃度等. 传统的 RNN 在时间序列建模方面存在梯度消失和长期依赖性不足的问题. 为此,LSTM 和 GRU 通过引入门控机制, 提高了模型对长时间依赖的建模能力[53], Bi-GRU 在此基础上进一步通过双向特征提取, 增强了时间序列信息的有效性. 实验结果显示, 与其他模型相比,Bi-GRU 的 MSE、RMSE 最低, 验证了其在结构化数据的时间依赖关系挖掘中具有显著优势. 结果表明双向编码网络在挖掘艺术品市场表现的潜在规律方面尤为有效.

5.2 多模态融合价值评估模型性能分析

在图像、文本和数值模态的融合过程中, 为了最大程度挖掘各模态之间的互补信息并增强融合效果, 本文采用特征层的融合策略, 包括拼接 (Concat)、相加 (Add) 和基于注意力融合的 MBT. 同时, 为了验证三个模态在不同层级在融合方式对艺术品价值评估的影响, 本文引入了基于投票机制的决策层融合策略, 并绘制各模态信息的价值评估结果以进行全面分析, 实验结果如表 2 所示.

表 2 多模型混合的艺术品价值评估的性能对比

模型	MSE	RMSE	MAE
ViT + MacBERT + Voting	0.0321	0.1791	0.0900
ViT + Bi-GRU + Voting	0.0416	0.2039	0.1254
MacBERT + Bi-GRU + Voting	0.0317	0.1780	0.0914
ViT + MacBERT + Bi-GRU + Voting	**0.0223**	**0.1493**	**0.0824**
ViT + MacBERT + Concat	0.0228	0.1509	0.0888
ViT + Bi-GRU + Concat	0.0273	0.1652	0.0992
MacBERT + Bi-GRU + Concat	0.0211	0.1452	0.0873
ViT + MacBERT + Bi-GRU + Concat	**0.0193**	**0.1389**	**0.0798**
ViT + MacBERT + Add	0.0146	0.1208	0.0811
ViT + Bi-GRU + Add	0.0253	0.1590	0.0994
MacBERT + Bi-GRU + Add	0.0160	0.1264	0.0848
ViT + MacBERT + Bi-GRU + Add	**0.0105**	**0.1024**	**0.0742**
ViT + MacBERT + MBT	0.0092	0.0959	0.0992
ViT + Bi-GRU + MBT	0.0158	0.1256	0.0826
MacBERT + Bi-GRU + MBT	0.0087	0.0932	0.0731
ViT + MacBERT + Bi-GRU + MBT	**0.0079**	**0.0888**	**0.0680**

从表 2 的实验结果可以看出, 与仅融合两个模态的信息相比, 三个模态融合的评估结果更优, 表明多模态信息的联合建模能更全面地刻画文化艺术品的多维价值特征. 基于特征层的融合方式总体优于决策层的投票机制, 其 MSE、RMSE、MAE 得分显著降低, 表明直接交互融合特征能更充分利用模态间的互补性, 提高数据的全面性和准确性, 增强模型的鲁棒性.

在特征层融合方法中,MBT 表现最佳,MSE、RMSE、MAE 分别达到 0.0079、0.0888 和 0.0680, 显著优于 Concat 和 Add 方式.Concat 因直接拼接特征向量导致高维稀疏性问题, 性能较低; 而 Add 通过特征相加保留了低维结构信息, 效果略优于 Concat, 但对模态特征的权重分配过于平均, 无法突出重要特征的贡献, 导致性能提升有限. 相比之下,MBT 通过注意力机制精准捕捉不同模态特征之间的深层关联, 同时动态调整特征权重分配, 能够有效整合关键信息并抑制无关噪声, 显著提升了评估精度.

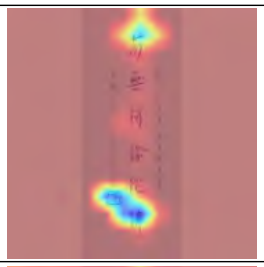
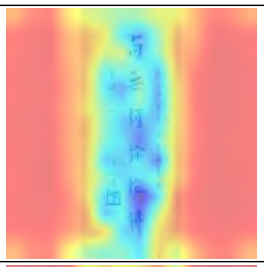
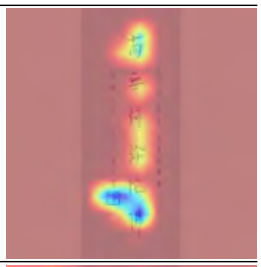
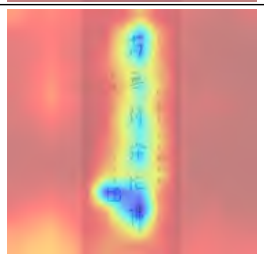
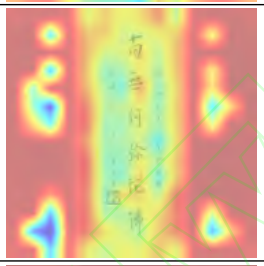
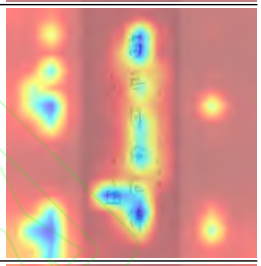
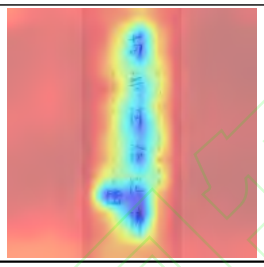
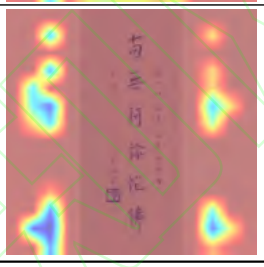
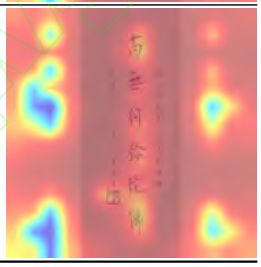
ViT 模型的浅层编码主要关注艺术品的局部特征, 例如色彩过渡、边缘形状等. 随着编码层数的加深, 模型的注意力从局部扩展到整体, 逐步捕捉作品的全局特征, 如构图布局、空间分布和整体视觉平衡. 在专家评估中, 艺术品的纹理细腻度、色彩层次感、元素排列的均衡性及整体美感是价值评估的重要参考依据. 因此, 在图像特征提取过程中, 通过对每层 Transformer 注意力矩阵求最大值, 不仅能过滤掉冗余信息, 还能更加准确地聚焦于符合专家分析视角的关键内容, 证明了图像特征模型在文化艺术品价值评估中的适用性和有效性.

5.3 艺术品价值评估可解释性分析

5.3.1 图像模态的价值评估解释

在 Attention Rollout 方法实现过程中, 考虑到不同 Transformer 层对图像关注区域的差异性, 本文对 1-4 层、1-8 层、1-12 层的注意力矩阵进行可视化分析. 为处理每个 Transformer 层中多个注意力头的权重, 分别采用了 max,min,mean 三种方式来聚合多头注意力的权重. 实验结果如表 3 所示, 展示了不同计算方式在不同层级的注意力分布效果.

表 3　基于 Attention Rollout 的艺术品图像模态解释

Transformer 层	max	min	mean
1–4 层			
1–8 层			
1–12 层			

从实验结果可以看出, 每层 Transformer 中, 对注意力头权重求 max 的方式生成的解释图像效果最好, 模型的关注区域集中在艺术作品中书写文字、印章、题识等核心内容, 有效过滤掉背景和无关信息. 这表明 max 策略可以显著减少噪声干扰, 确保关键特征进入多模态融合的环节.

ViT 模型的浅层编码主要关注艺术品的局部特征, 例如色彩过渡、边缘形状等. 随着编码层数的加深, 模型的注意力从局部扩展到整体, 逐步捕捉作品的全局特征, 如构图布局、空间分布和整体视觉平衡. 在专家评估中, 艺术品的纹理细腻度、色彩层次感、元素排列的均衡性及整体美感是价值评估的重要参考依据. 因此, 在图像特征提取过程中, 通过对每层 Transformer 注意力矩阵求最大值, 不仅能过滤掉冗余信息, 还能更加准确地聚焦于符合专家分析视角的关键内容, 证明了图像特征模型在文化艺术品价值评估中的适用性和有效性.

5.3.2 文本模态的价值评估解释

为验证注意力机制解释 MaBERT 算法的特征抽取能力, 探究文本特征模型对艺术品价值评估中关注的重点词语, 本文基于艺术品的拍卖描述文本和作者简介信息, 选取 MacBERT-large 模型的第 8 层、第 16 层和第 24 层的 [CLS] 注意力权重进行可视化分析, 揭示模型在不同层次对文本特征的关注变化. 结果表明, 随着编码层数的增加,[CLS] 语义表示逐渐从关注全体词语转向聚焦于少数关键词语, 说明高层 Transformer 编码器能够有效抽取与作品价值密切相关的信息. 实验结果如表 4, 表 5 所示.

表 4 基于多层注意力的艺术品拍卖文本解释

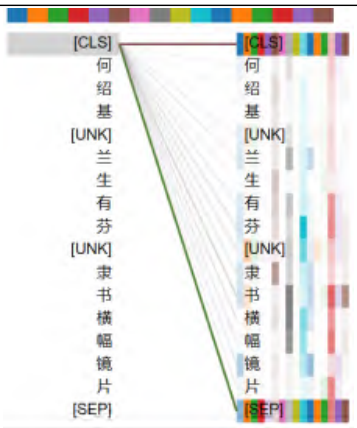
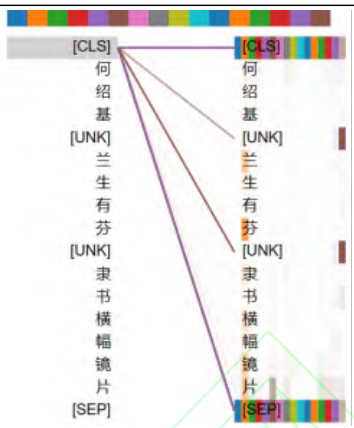
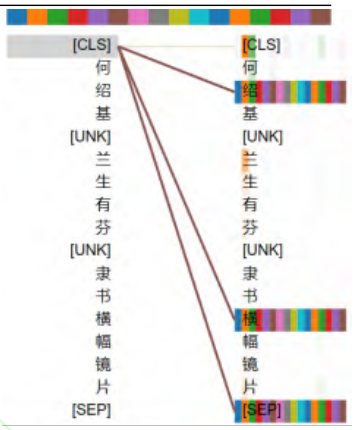
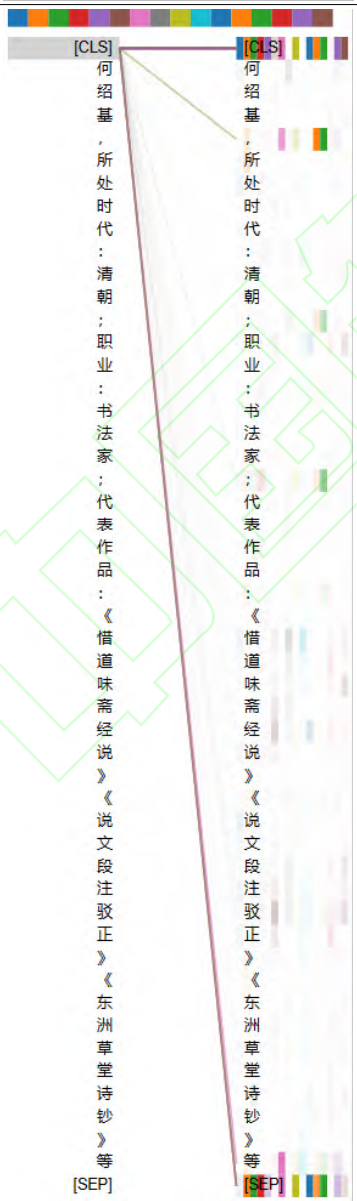
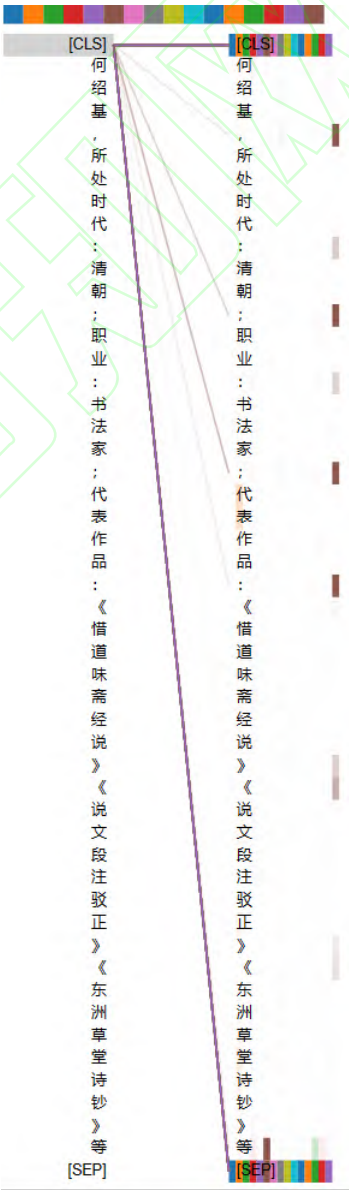
文本示例	第 8 层	第 16 层	第 24 层
何绍基 《兰生有芬》 隶书横幅镜片			
何绍基 所处时代: 清朝 职业: 书法家 代表作品: 《惜道味斋经说》 《说文段注驳正》 《东洲草堂诗钞》等			

表 5　最后四层注意力的艺术品文本模态解释

示例编号	拍卖文本	作者简介文本
1	何绍基兰生有芬隶书横幅镜片	弘一，职业：美术教育家,书法家；主要成就：新文化运动和中日文化交流的先驱；代表作品：《送别》《南京大学校歌》
2	何绍基，所处时代：清朝；职业：书法家；代表作品：《惜道味斋经说》《说文段注驳正》《东洲草堂诗钞》等	赵朴初书法对联镜片
3	弘一楷书南无阿弥陀佛镜心	赵朴初，职业：社会活动家；主要成就：中国民主促进会的创始人之一；代表作品：《佛教常识答问》,《滴水集》

从实验结果可以看出, 拍卖描述文本中诸如 “作者”“形制”“内容” 等词语的注意力权重较高, 颜色最深; 在作者简介文本中,“所处时代”“主要成就”“代表作品” 等词语的注意力权重最突出. 这表明, 模型能够有效捕捉艺术品拍卖文本和作者背景中与作品价值相关的关键信息. 在实际艺术品交易市场中, 艺术品的市场价值受多方面因素的影响, 其中作者的历史地位、代表作品的市场表现、作品的创作时间及艺术风格等因素是收藏家和投资者最为关注的价值评估核心指标[54]. 研究表明, 在艺术品估值过程中, 作品所处时代及其艺术风格的独特性会直接影响市场认可度, 而具有较高市场流通性的作品, 其成交价往往更稳定, 并能获得更高估值[55]. 此外, 拍卖行的估价通常参考相似作品的历史成交数据, 重点考量作品的文化价值、稀缺性及市场需求. 由此可见, 多头注意力的可视化分析结果不仅揭示了模型对重点信息的提取能力, 还与实际艺术品交易市场的评估标准高度一致, 进一步验证了文本特征模型在艺术品价值评估中的实用性和可靠性. 模型能够关注与市场价值高度相关的关键特征, 表明基于 Transformer 的文本特征提取方法在艺术品价值评估任务中的有效性.

5.3.3 数值模态的价值评估解释

在数值模态的解释中, 本文利用 SHAP 生成了基于所有样本的全局解释图, 以分析各特征对模型输出值的影响. 由图 3 所示, 拍卖公司、拍卖年份、拍卖月份、尺寸面积是对模型输出值影响较大的关键特征. 其中, 拍卖公司成交额对数值价值评估模型的影响最大. 按照成交额从高到低编码, 成交额越大的通常具备更强的作品质量审核能力和媒体宣传能力, 其对应的 SHAP 值也随之增大. 拍卖年份的值越大, 其对应的 SHAP 值却越小. 可能是受近年来疫情等外部因素的影响, 导致市场波动和作品成交表现不佳所致. 拍卖月份按照从低到高编码, 月份越大, 作品的市场表现和交易热度越高, 其 SHAP 值也随之增大. 尺寸面积也表现出类似的趋势, 尺寸越大,SHAP 值越高, 反映出大尺寸作品更受市场青睐. 拍卖城市人均 GDP 依据人均国内生产总值进行编码, 结果显示一、二线城市因其较高的市场需求和文化影响力, 对作品价值评估产生正向影响, 其对应的 SHAP 值也较大. 作品的材质和形制按照做工的复杂度进行编码, 对价值评估的结果影响最小.

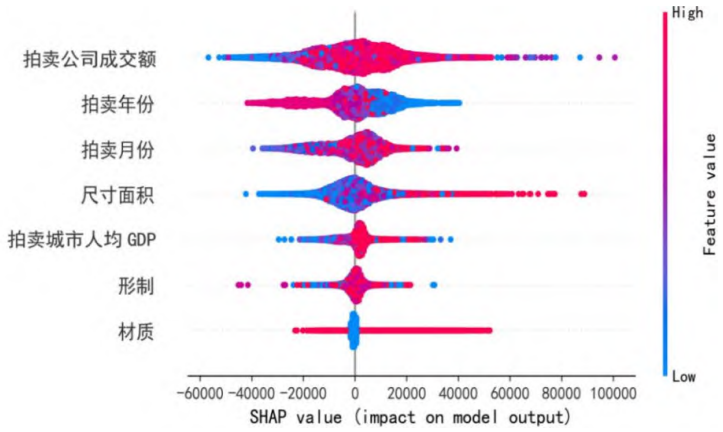
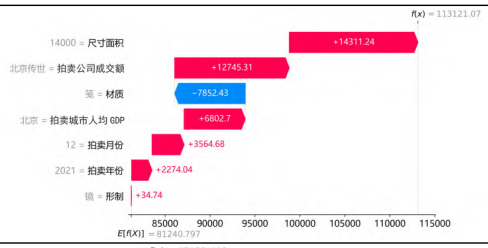
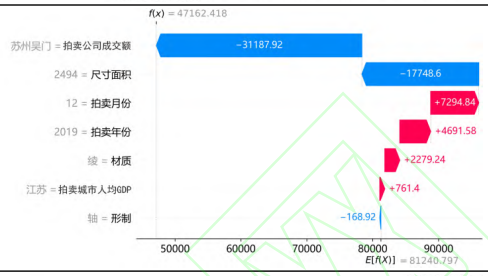
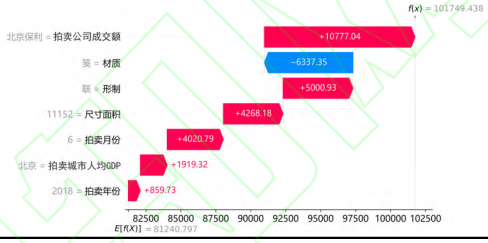


图 3　基于 SHAP 的艺术品数值模态全局解释

表 6 基于 SHAP 的艺术品数值模态局部解释

示例	可视化结果
局部解释示例 1	
局部解释示例 2	
局部解释示例 3	

在局部解释中,SHAP 为单个样本生成了特定特征的近似影响值, 如表 6 所示. 从三个样本的局部解释结果可见, 材质为纸、笺以及形制简单的作品, 对价值评估呈现负向影响; 而拍卖公司成交额较高、拍卖地点位于高 GDP 城市的一二线城市, 则对价值评估具有显著的正向影响. 此外, 尺寸面积过小也显著降低了作品的市场价值. 这些结果表明, 在艺术品交易中, 材质、形制和尺寸面积等特征会影响市场接受度, 而选择具有良好市场表现的拍卖公司和成交地点则有助于提升作品的价值评估结果.

综上, 基于 SHAP 算法的全局和局部解释结果, 与艺术品交易的现实规则高度吻合. 这一解释框架不仅揭示了不同特征在价值评估中的影响机制, 也为艺术品市场参与者优化交易策略、选择拍卖平台和目标市场提供了科学依据.

6 结论

本文基于图像、文本、数值三类模态数据, 提出一种多模态融合与可解释性的文化艺术品价值评估模型. 实验结果显示, 融合模型的 MSE、RMSE、MAE 指标显著优于单一模型和其他模型, 且可解释性强, 为文化艺术品的智能化评估提供了具有应用价值的方法和实践工具. 本文的研究结论主要包括:

1) 多模态融合的文化艺术品价值评估模型构建. 本文创新型地提出了 ViT- MacBERT-Bi-GRU 的多模态融合模型, 克服了单一模态难以全面反映艺术品多维价值的局限, 通过视觉、文本和结构化数据的深度融合, 从文化艺术品的艺术价值、历史价值、市场价值等不同维度进行全面分析, 为艺术品价值评估提供了系统化的理论框架. 2) 提升模型的可解释性, 解决 “黑箱” 问题. 本研究引入 Attention Rollout、多头注意力和 SHAP 的可解释方法, 分别从三个模态解释模型的决策依据, 揭示了不同模态特征对价值评估结果的贡献, 帮助用户理解模型的决策过程. 为人工智能在复杂评估场景中的可解释应用提供了理论支持, 推动了智能技术在文化艺术品价值评估中的可信应用. 3) 多模态融合模型的跨领域应用潜力. 本文提出的多模态融合框架不仅适用于艺术品的价值评估, 还可以推广至医疗、金融等需要多维数据融合的复杂领域, 为这些领域的多维度数据整合提供了新的理论参考. 4) 推动文化艺术品平台的智能化发展. 本文提出的多模态评估模型为文化艺术品交易平台提供了智能化可解释的估值工具, 有助于提升平台的

价值评估服务能力, 增强公信力, 提升市场活跃度. 5) 为政策制定者和金融机构提供科学依据. 文化艺术品的多模态数据融合分析能够帮助市场监管者更好地理解艺术品的价值构成, 制定更科学合理的市场定价机制, 促进艺术品市场的长期规范化发展, 降低 “黑箱操作” 带来的投机风险. 此外, 还能为金融机构提供可靠的艺术品价值参考, 辅助金融机构在艺术品投资、融资等资产管理中作出更加科学的决策.

目前本文的研究仍然存在一些不足, 未来的工作可以在以下几个方面进行改进和拓展: 第一, 数据源优化. 当前的实验数据主要使用的是书法艺术品数据, 限制了模型的泛化能力, 但所采用的多模态融合方法具有可扩展性, 可适用于不同类别的文化艺术品, 未来可以加入绘画、陶瓷等多种类的艺术品数据, 丰富模型的应用场景. 第二, 模态特征提取优化. 在面对低像素图像、复杂市场信息和古代术语, 可能会导致多模态信息提取不完全. 在未来的研究中, 可以通过优化模型网络结构并结合最新的技术, 进一步提升各个模态的特征表示能力.

参考文献

[1] 刘小娟, 魏农建, 傅晓红. 价值因素与非价值因素的互动: 艺术品定价机制的一种新可能 [J]. 上海大学学报 (社会科学版), 2022, 39(06): 133–146.
Liu X J, Wei N J, Fu X H. The interaction of value factors and non-value factors: A new possibility of artwork pricing pattern[J]. Journal of Shanghai University(Social Sciences Edition), 2022, 39(06): 133–146.

[2] 周建新, 朱学平. 中国文化产业研究 2024 年度学术报告 [J]. 深圳大学学报 (人文社会科学版), 2025, 42(01): 46–58.
Zhou J X, Xie J M. 2024 Annual academic report on China’s cultural industry[J]. Journal of Shenzhen University(Humanities & Social Sciences), 2024, 41(01): 55–70.

[3] 洪永淼, 汪寿阳. 大数据如何改变经济学研究范式?[J]. 管理世界,2021,37(10):40-55+72+56.
Hong Y M, Wang S Y. How is big data changing economic research paradigms?[J]. Management World, 2021, 37(10): 40–55+72+56.

[4] 倪渊, 华君鹏, 张健, 等. 融合情感特征和可解释性的弹幕视频传播效果预测模型 [J]. 数据分析与知识发现, 2025, 9(02): 146–158.
Ni Y, Hua J P, Zhang J, et al. A prediction model for danmaku video’s propagationEffects with sentiment features and interpretability[J]. Data Analysis and Knowledge Discovery , 2025, 9(02): 146–158.

[5] Renneboog L, Spaenjers C. Buying beauty: On prices and returns in the art market[J]. Management Science, 2013, 59(1): 36–53.

[6] Ugail H, Stork D G, Edwards H, et al. Deep transfer learning for visual analysis and attribution of paintings by Raphael[J]. Heritage Science, 2023, 11(1): 268.

[7] Iigaya K, Yi S, Wahle I A, et al. Aesthetic preference for art can be predicted from a mixture of low-and high-level visual features[J]. Nature Human Behaviour, 2021, 5(6): 743-755.

[8] Li B, Liu T. An analysis of multi-modal deep learning for art price appraisal[C]//2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom). IEEE, 2021: 1509–1513.

[9] Liu C. Prediction and analysis of artwork price based on deep neural network[J]. Scientific Programming, 2022, 2022(1): 7133910.

[10] Indrawan J O, Filia B J, Parmonangan I H. Multimodal Approach for Painting Price Prediction[C]//2023 5th International Conference on Cybernetics and Intelligent System (ICORIS). IEEE, 2023: 1–5.

[11] 曾春艳, 严康, 王志锋, 等. 深度学习模型可解释性研究综述 [J]. 计算机工程与应用, 2021, 57(08): 1–9.
Zeng C Y, Yan K, Wang Z F, et al. Survey of interpretability research on deep learning models[J]. Computer Engineering and Applications, 2021, 57(08): 1–9.

[12] Zhang Y, Zhou X, Yuan J, et al. Multimodal Named Entity Recognition Model Based on Cross-modal Feature Enhancement Mechanism[C]//2024 International Conference on Computational Linguistics and Natural Language Processing (CLNLP). IEEE, 2024: 36-40.

[13] 张大斌, 曾芷媚, 凌立文, 等. 基于多特征融合深度神经网络的玉米期货价格预测 [J/OL]. 中国管理科学, 1–11[2024-11-19].https://doi.org/10.16381/j.cnki.issn1003-207x.2022.1040.
Zhang D B, Zeng Z M, Ling L W, et al. Prediction of corn futures price based on multi-feature deep neuralnet-

work model[J/OL]. Chinese Journal of Management Science, 1–11[2024-11-19].https://doi.org/10.16381/j.cnki.issn1003-207x.2022.1040.

[14] 刘洋, 张雯, 胡毅, 等. 基于多模态深度学习的酒店股票预测 [J]. 数据分析与知识发现, 2023, 7(05): 21–32.
Liu Y, Zhang W, Hu Y, et al. Hotel stock prediction based on multimodal deep learning[J]. Data Analysis and Knowledge Discovery, 2023, 7(05): 21–32.

[15] Liu W, Ren G, Yu R, et al. Image-adaptive YOLO for object detection in adverse weather conditions[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(2): 1792–1800.

[16] Bi X, Hu J, Xiao B, et al. IEMask R-CNN: Information-enhanced mask R-CNN[J]. IEEE Transactions on Big Data, 2022, 9(2): 688–700.

[17] Zhang Y, Cao J, Zhang L, et al. A free lunch from ViT: Adaptive attention multi-scale fusion transformer for fine-grained visual recognition[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 3234–3238.

[18] Mikolov T. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013, 3781.

[19] Lee J, Toutanova K. Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018, 3(8).

[20] Su J. Speed up without losing points: Chinese WoBERT based on word granularity[R]. 2020.

[21] Liu Z, Lin W, Shi Y, et al. A robustly optimized BERT pre-training approach with post-training[C]//China National Conference on Chinese Computational Linguistics. Cham: Springer International Publishing, 2021: 471-484.

[22] Cui Y, Che W, Liu T, et al. Pre-training with whole word masking for Chinese BERT[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504–3514.

[23] Dhruv P, Naskar S. Image classification using convolutional neural network (CNN) and recurrent neural network (RNN): A review[J]. Machine learning and information processing: Proceedings of ICMLIP 2019, 2020: 367–381.

[24] Dey R, Salem F M. Gate-variants of gated recurrent unit (GRU) neural networks[C]. 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS). IEEE, 2017: 1597–1600.

[25] 谷丽琼, 吴运杰, 逄金辉. 基于 Attention 机制的 GRU 股票预测模型 [J]. 系统工程, 2020, 38(05): 134–140.
Gu L Q, Wu Y J, Pang J H. Stock prediction model based on attention mechanism and GRU[J]. Systems Engineering, 2020, 38(05): 134–140.

[26] Zeng L, Ren W, Shan L. Attention-based bidirectional gated recurrent unit neural networks for well logs prediction and lithology identification[J]. Neurocomputing, 2020, 414: 153–171.

[27] 李新, 张旭, 余乐安, 等. 基于改进 Transformer 模型的景区短时客流预测研究 [J/OL]. 中国管理科学, 1–15[2024-11-19].https://doi.org/10.16381/j.cnki.issn1003-207x.2023.1927.
Li X, Zhang X, Yu L A, et al. Enhancing short-term tourist flow forecasting and evaluation using an improved Transformer framework[J/OL]. Chinese Journal of Management Science, 1–15[2024-11-19].https://doi.org/10.16381/j.cnki.issn1003-207x.2023.1927.

[28] 王德鲁, 毛锦琦, 宋学锋, 等. 数据特征驱动的火电产能过剩分解集成预测模型 [J]. 系统工程理论与实践, 2021, 41(03): 727–743.
Wang D L, Mao J Q, Song X F, et al. A data-characteristic-driven decomposition ensembleforecasting model for thermal power overcapacity[J]. Systems Engineering-Theory & Practice, 2021, 41(03): 727–743.

[29] Wu Z, Pan S, Chen F, et al. A comprehensive survey on graph neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(1): 4–24.

[30] Wang T C, Liu M Y, Zhu J Y, et al. High-resolution image synthesis and semantic manipulation with conditional GANs[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8798–8807.

[31] 王刚, 陈红, 马敬玲, 等. 基于多尺度 1D-CNN 和注意力机制的汇率多步预测研究 [J]. 系统工程理论与实践, 2024, 44(06): 1934–1949.
Wang G, Chen H, Ma J L, et al. Multi-step forecasting of exchange rate based on multi-scale1D-CNN and attention mechanisms[J]. Systems Engineering-Theory & Practice, 2024, 44(06): 1934–1949.

[32] Nagrani A, Yang S, Arnab A, et al. Attention bottlenecks for multimodal fusion[J]. Advances in Neural Information Processing Systems, 2021, 34: 14200–14213.

[33] Tang L, Hu Q, Wang X, et al. A multimodal fusion network based on a cross-attention mechanism for the classification of Parkinsonian tremor and essential tremor[J]. Scientific Reports, 2024, 14(1): 28050.

[34] Zhou W, Lin M, Xiao M, et al. Higher Precision is Not Always Better: Search Algorithm and Consumer Engagement[J]. Management Science, 2024.

[35] Rudresh D, Devam D, Het N, et al. Explainable AI (XAI): Core ideas, techniques, and solutions[J]. ACM Computing Surveys, 2023, 55(9): 1–33.

[36] 张永, 黎嘉豪, 刘悦, 等. 基于 Transformer 和关键特征的可解释端到端投资组合策略 [J]. 计量经济学报, 2024, 4(05): 1381–1407.
Zhang Y, Li J H, Liu Y, et al. Interpretable end-to-end portfolio selection strategyBased on Transformer and key features[J]. China Journal of Econometrics, 2024, 4(05): 1381–1407.

[37] 孔祥维, 唐鑫泽, 王子明. 人工智能决策可解释性的研究综述 [J]. 系统工程理论与实践, 2021, 41(02): 524–536.
Kong X W, Tang X Z, Wang Z M. A survey of explainable artificial intelligence decision[J]. Systems Engineering-Theory & Practice, 2021, 41(02): 524–536.

[38] Nanfack G, Fulleringer A, Marty J, et al. Adversarial attacks on the interpretation of neuron activation maximization[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(5): 4315–4324.

[39] Jung H, Oh Y. Towards better explanations of class activation mapping[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 1336–1344.

[40] Zeng C, Yan K, Wang Z, et al. Abs-CAM: A gradient optimization interpretable approach for explanation of convolutional neural networks[J]. Signal, Image and Video Processing, 2023, 17(4): 1069–1076.

[41] Abnar S, Zuidema W. Quantifying attention flow in transformers[J]. arXiv preprint arXiv:2005.00928, 2020.

[42] Kumar S, Sumers T R, Yamakoshi T, et al. Shared functional specialization in transformer-based language models and the human brain[J]. Nature communications, 2024, 15(1): 5523.

[43] Garreau D, Luxburg U. Explaining the explainer: A first theoretical analysis of LIME[C]. International Conference on Artificial Intelligence and Statistics. PMLR, 2020: 1287–1296.

[44] Van den Broeck G, Lykov A, Schleich M, et al. On the tractability of SHAP explanations[J]. Journal of Artificial Intelligence Research, 2022, 74: 851–886.

[45] Cheng X, Rao Z, Chen Y, et al. Explaining knowledge distillation by quantifying the knowledge[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 12925–12935.

[46] Willmott C J, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance[J]. Climate Research, 2005, 30(1): 79-82.

[47] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 1-9.

[48] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[49] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention[C]//International Conference on Machine Learning. PMLR, 2021: 10347-10357.

[50] Lan Z, Chen M, Goodman S, et al. ALBERT: A lite BERT for self-supervised learning of language representations[J]. arXiv preprint arXiv:1909.11942, 2019.

[51] Cui Y, Che W, Liu T, et al. Pre-training with whole word masking for Chinese BERT[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514.

[52] Liu Y, Ott M, Goyal N, et al. RoBERTa: A robustly optimized BERT pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.

[53] Abumohsen M, Owda A Y, Owda M. Electrical load forecasting using LSTM, GRU, and RNN algorithms[J]. Energies, 2023, 16(5): 2283.

[54] 倪渊, 李晓娜, 张健, 等. 多源异构数据融合视角下文化 UGC 传播效果预测——基于 GRA-PSO-WRF 的组合建模 [J]. 管理评论, 2024, 36(11): 235-247.
Ni Y, Li X N, Zhang J, et al. Prediction of cultural UGC communication effectiveness from the perspective of multi-source heterogeneous dataFusion: A combination modeling of GRA-PSO-WRF method[J]. Management Review, 2024, 36(11): 235-247.

[55] 江凌. 我国书画类艺术品资产组合及投资收益优化 [J]. 深圳大学学报 (人文社会科学版), 2024, 41(05): 58-70.
Jiang L. Optimization of China’s calligraphy and painting artworks asset portfolio and investment income[J]. Journal of Shenzhen University(Humanities & Social Sciences), 2024, 41(05): 58-70.