自动化学报
ACTA AUTOMATICA SINICA

基于视觉属性的多模态可解释图像分类方法

王辉 黄宇廷 夏玉婷 范自柱 罗国亮 杨辉

Multimodal Interpretable Image Classification Method Based on Visual Attributes

WANG Hui, HUANG Yu-Ting, XIA Yu-Ting, FAN Zi-Zhu, LUO Guo-Liang, YANG Hui

在线阅读 View online: https://doi.org/10.16383/j.aas.c240218

基于视觉属性的多模态可解释图像分类方法

王 辉[1,2] 黄宇廷[2,3] 夏玉婷[1,2] 范自柱[4] 罗国亮[1,2] 杨 辉[2]

摘 要 基于深度神经网络 (Deep neutral networks, DNN) 的分类方法因缺乏可解释性, 导致在金融、医疗、法律等关键领域难以获得完全信任, 极大限制了其应用. 现有多数研究主要关注单模态数据的可解释性, 多模态数据的可解释性方面仍存在挑战. 为解决这一问题, 提出一种基于视觉属性的多模态可解释图像分类方法, 该方法将可见光和深度图等不同视觉模态提取的属性融入模型的训练过程, 不仅能通过视觉属性和决策树对已有的神经网络黑盒模型进行解释, 而且能在训练过程中进一步提升模型解释信息的能力. 引入可解释性通常会造成模型精度的降低, 该方法在保持模型具有良好可解释性的同时, 仍具有较高的分类精度, 在 NYUDv2、SUN RGB-D 和 RGB-NIR 三个数据集上, 相比于单模态可解释方法, 该模型准确率明显提升, 并达到与多模态不可解释模型相媲美的性能.

关键词 可解释性, 视觉属性, 多模态融合, 决策树, 图像分类

引用格式 王辉, 黄宇廷, 夏玉婷, 范自柱, 罗国亮, 杨辉. 基于视觉属性的多模态可解释图像分类方法. 自动化学报, 2025, **51**(2): 1−12

DOI 10.16383/j.aas.c240218 **CSTR** 32138.14.j.aas.c240218

Multimodal Interpretable Image Classification Method Based on Visual Attributes

WANG Hui[1,2] HUANG Yu-Ting[2,3] XIA Yu-Ting[1,2] FAN Zi-Zhu[4] LUO Guo-Liang[1,2] YANG Hui[2]

Abstract The classification methods based on deep neutral networks (DNN) lack interpretability, which makes it difficult to gain complete trust in key fields such as finance, medical treatment, and law, greatly limiting their applications. Most existing research mainly focuses on the interpretability of uni-modal data, while there are still challenges in the interpretability of multimodal data. To address this issue, a multimodal interpretable image classification method based on visual attributes is proposed. This method incorporates attributes extracted from different visual modalities such as visible light and depth maps into the training process of the model. It not only interpret the existing black box model of neural networks through visual attributes and decision trees, but also further enhances the model′s ability to interpret information during the training process. Introducing interpretability often leads to a decrease in model accuracy. This method maintains good interpretability while still maintaining high classification accuracy. Compared to uni-modal interpretable methods, the accuracy of this model is significantly improved on the NYUDv2, SUN RGB-D, and RGB-NIR datasets, and it achieves performance comparable to multimodal uninterpretable models.

Key words Interpretability, visual attributes, multimodal fusion, decision tree, image classification

Citation Wang Hui, Huang Yu-Ting, Xia Yu-Ting, Fan Zi-Zhu, Luo Guo-Liang, Yang Hui. Multimodal interpretable image classification method based on visual attributes. *Acta Automatica Sinica*, 2025, **51**(2): 1−12

深度神经网络 (Deep neutral networks, DNN) 由于其出色性能而在计算机视觉分类等任务中得到广泛运用[1−2], 然而, 其多层网络架构拟合复杂、非线性特征空间的特点致使其模型通常包含数亿个权重参数, 导致人们难以追踪、理解网络的决策过程[3]. 这种黑盒模型在需要高透明和高可信的领域 (如医学、法律和金融等) 难以应用, 因为它可能会以错误的结果误导用户, 甚至可能危害生命安全[4]. 相比之下, 一些经典的白盒模型如决策树[5]、规则学习[6] 和贝叶斯网络[7] 分别使用层次结构、推理规则和概率

收稿日期 2024-04-22 录用日期 2024-09-04
Manuscript received April 22, 2024; accepted September 4, 2024
国家自然科学基金 (61991401, U2034211, 61991404), 江西省自然科学基金 (20224BAB212014, 20232ABC03A04), CAD&CG 国家重点实验室开放课题 (A2334) 资助
Supported by National Natural Science Foundation of China (61991401, U2034211, 61991404), Natural Science Foundation of Jiangxi, China (20224BAB212014, 20232ABC03A04), and Open Project Program of the State Key Laboratory of CAD&CG (A2334)
本文责任编委 鲁继文
Recommended by Associate Editor LU Ji-Wen
1. 华东交通大学信息与软件工程学院 南昌 330013 2. 轨道交通基础设施性能监测与保障国家重点实验室 南昌 330013 3. 浙江大学软件学院 宁波 315048 4. 上海电力大学计算机科学与技术学院 上海 201306
1. School of Information and Software Engineering, East China Jiaotong University, Nanchang 330013 2. State Key Laboratory of Performance Monitoring and Protecting of Rail Transit Infrastructure, Nanchang 330013 3. School of Software Technology, Zhejiang University, Ningbo 315048 4. College of Computer Science and Technology, Shanghai University of Electric Power, Shanghai 201306

分布来展示推理过程, 然而这些白盒模型的预测精度远远低于深度神经网络. 研究人员尝试使用多种方法来解释黑盒模型, 例如基于梯度的可视化方法[8−9]、类激活图[10−11]、通过白盒模型创建代理局部拟合黑盒模型[12−13] 或原型学习[14−15] 等. 然而, 这些解释黑盒模型的方法并不直接参与到模型训练中, 因此不能提高模型提供解释信息的能力. 随着人工智能模型对可解释性的要求不断提高, 模型需要在保持高精度的同时实现良好的可解释性.

神经网络最初模仿大脑的神经元进行设计, 而在分类任务中, 即使面对未知的物体, 人类也可以做出预测并说明理由. 人类在认知物体过程中通过其所具有的视觉属性进行识别[16], 并借助触觉、听觉和味觉等多种模态信息帮助做出更精确的判断. 医学研究结果证实, 多感官的整体感知效果大于各感官的独立感知效果[17], 多模态数据的综合信息对于准确理解和认知物体至关重要. 具体而言, 人类通过多个感官感知物体所具有的属性, 并从已有知识中找到具有相同属性的类别原型. 例如, 铁门通常看上去很坚固, 摸起来冰凉光滑, 开关门时会发出低沉的声音, 这些属性就是铁门的原型. 当人类不知道一种门的材质时, 仍然可以根据它与铁门相同的属性来做出判断.

同时, 类别原型中天然存在一定的层次结构, 类似于图 1 所示的生物分类学. 每个类别都继承其父类的所有属性, 但又有其独特的部分. 例如食肉目和兔形目同属哺乳纲, 都继承哺乳纲有四肢、有耳廓的生理特征, 又存在食肉目眼睛前置, 兔形目眼睛在头部两侧的区别. 这种人类认知物体的方式启发我们可提取多模态数据的属性作为解释信息,

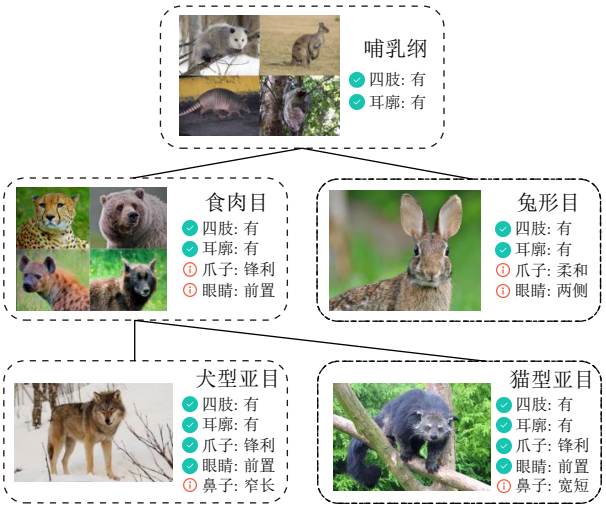

图 1 以生理特征为依据的生物分类学

Fig. 1 Biological taxonomy based on physiological characteristics

并使用决策树推理将这种层次结构加以利用. 通过提取和利用这些属性, 可以更好地理解和解释模型的决策过程, 使得模型在可解释性和精度方面取得更好的平衡.

多模态融合方法已经广泛应用于各领域中[18−19], 然而这些方法通常缺乏可解释性. 研究表明卷积神经网络 (Convolutional neural networks, CNN) 在分类任务中具有学习视觉属性的能力, 因此不再需要人工定义属性以建模物体[20]. 受这些工作的启发, 本文提出一种基于可见光、深度图等多模态图像输入的视觉属性可解释分类方法, 通过骨干网络在原型约束下学习输入数据的属性, 在预先设定的层次结构中进行树推理. 每个属性用一组具有相同特征的图像集合表示, 利用梯度加权类激活映射 (Gradient-weighted class activation mapping, Grad-CAM)[21] 可视化后归纳集合具有的特征.

该方法流程如图 2 所示, 提供属性集合和决策树推理过程对模型决策进行解释. 以一组浴室场景图像为例, 骨干网络计算可见光、深度图中具有的属性强度, 将该浴室场景表示为马桶、橱柜、桌面、布料等视觉属性的集合. 同时, 决策树依据属性强度进行推理, 认为这些属性首先符合住宅的特点, 进一步符合住宅中浴室的特点, 最终完成分类.

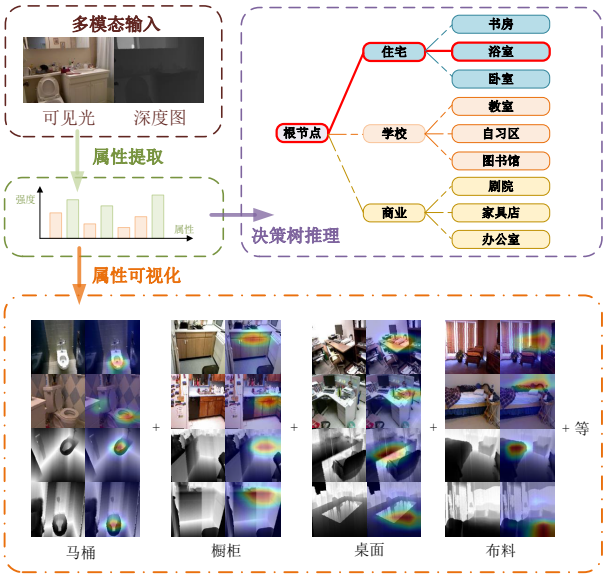

图 2 本文提出模型的推理流程

Fig. 2 The inference process of the proposed model

本文的主要贡献是:

1) 提出一种基于属性的多模态可解释分类方法, 该方法能够使用输入数据所具有的属性信息和决策树推理来解释分类结果, 并通过消融实验研究模型每个部分对整体性能的影响.

2) 提出一种面向多模态数据的决策树逐层融合方法, 实现在决策树推理阶段的证据融合, 在保证模型可解释性的同时提升准确性.

3) 本文的方法在准确性和可解释性方面都表现出色, 在三个多模态场景分类数据集 NYUDv2、SUN RGB-D 和 RGB-NIR 中达到最先进的性能, 并通过插入/删除测试验证了方法的可解释性.

1　相关工作

相关的前期工作包括可解释方法和多模态学习两部分.

1.1　可解释方法

近年来, 研究人员更加关注复杂、非线性的黑盒模型在做出决策时的内部机制, 以解决在医疗、法律和金融等特定应用场景中使用黑盒模型可能带来的风险[3].

在分类任务中, 一类解释黑盒模型的方法是展示模型关注输入数据的哪些部分. 通过观察模型的梯度[8]、积分梯度[9]、神经元激活情况[22] 或引入反卷积[23] 以突出输入数据的特定部分. 其中最具代表性的方法是类激活图 (Class activation mapping, CAM)[10] 及其改进[21], 通过特征图和输出层权重或梯度计算得到原始输入图像中各区域的重要性, 并生成热力图叠加到原始图像. 另一类方法与黑盒模型的架构无关. 例如局部可解释性模型诊断解释 (Local interpretable model-agnostic explanations, LIME)[12] 关注局部的输入输出之间的关系, 通过创建一个局部的可解释模型来拟合整个黑盒模型. Shapley 加性解释 (Shapley additive explanations, SHAP)[13] 将 LIME 与可解释理论 Shapley 值相结合, 通过博弈论计算最优的 Shapley 值. 原型学习[14−15] 通过比较输入数据的属性和类别原型之间的差异提供分类的解释信息, 原型通过人工编码[24] 或机器学习方法提取[25−26]. 这两类方法都是对已有黑盒模型进行解释, 不直接参与到模型训练过程中.

使用透明的白盒模型具有良好的可解释性. 常用的白盒模型包括线性模型、决策树[27−28] 和基于规则的模型[29] 等. 白盒模型可以提供比黑盒模型更好的解释性, 但通常准确率更低. 因此, 本文提出的方法优势在于能够在训练过程中利用神经网络学习输入图像的属性和类别原型, 并根据属性使用决策树进行分类推理, 从而实现良好的准确性和可解释性.

1.2　多模态学习

在现实生活中, 人类通过视觉、听觉和触觉等多种感官来感知世界, 医学研究表明多感官的整体感知效果大于各感官独立感知效果的总和[17]. 因此, 多模态学习旨在从多个不同数据源的丰富信息中进行学习.

多模态学习早期的代表性方法是典型相关分析 (Canonical correlation analysis, CCA)[30], 尝试找到不同模态间相关性最大的线性变换. 之后, 各种基于 CCA 的模型[31−32] 被广泛用于多模态学习中. 然而, 线性的嵌入函数并不能很好地拟合复杂非线性的多模态数据. 深度学习方法可以对复杂的非线性关系进行建模, 因此多模态融合可以使用深度学习方法[33]. 根据融合方法的不同可以分为聚合[34−35]、对齐[36−37] 和混合方法; 根据融合时间可以分为前期、中期和后期融合[38−40].

虽然多模态学习在分类准确性上取得显著效果, 但是针对多模态的可解释性问题研究较少. 本文提出的方法是针对来自可见光、深度图、远红外等不同图像数据源的多模态数据提供解释信息的一次尝试, 旨在增强模型的可解释性, 使得模型的决策过程更加透明和可理解.

2　模型架构

模型框架如图 3 所示, 每个模态都使用卷积神经网络作为骨干网络来学习属性的表示向量, 属性

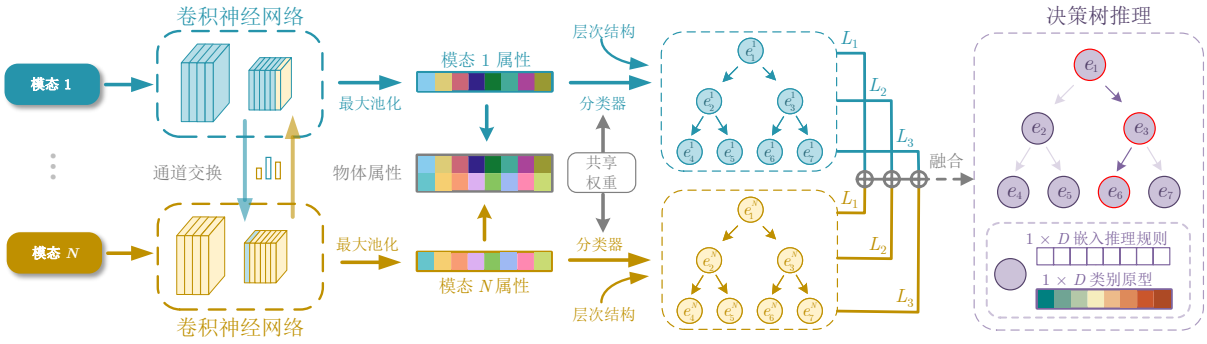

图 3　可解释图像分类框架

Fig. 3　The interpretable image classification framework

向量与引入的层次结构、学习得到的决策规则共同构建决策树. 各模态的决策树根据 Dempster-Shafer 证据理论[41] 进行逐层融合, 最后, 在融合后的决策树中进行软推理. 在骨干网络提取属性的过程中, 引入通道交换去除某一模态的低质量信息, 并保证不同模态之间学习得到的属性一致性. 模型框架主要包含视觉属性提取、决策树构建与融合、决策树推理和原型限制四个模块.

2.1 视觉属性提取

已有研究表明 CNN 在分类任务中具有学习视觉属性的能力. 在本文的方法中, CNN 被用作骨干网络来学习每个模态的属性表示向量. 对于第 i 个视图 v^i, 通过一个卷积神经网络的映射 f^i 来提取一维的属性强度向量 $\boldsymbol{A}^i$, 即 $f^i: v^i \mapsto \boldsymbol{A}^i \in [0,\ +\infty)^D$, 其中 D 表示网络提取的属性数量. 通过这个映射, 能够从输入的视图数据中提取出一组属性强度值, 这些属性描述该视图具有的特征. 这些属性向量 $\boldsymbol{A}^i$ 在后续的步骤中将用于构建决策树, 从而实现对模型决策过程的解释性分析.

通过卷积神经网络处理后, 将得到大小为 $D\times w\times h$ 的特征图, 其中 w 和 h 分别对应特征图的宽度和高度. 特征图中的每一项代表着在某一区域内特定属性的强度. 考虑到模型更关注某一属性在整个图像中是否存在, 因此使用全局最大池化将特征图转换为属性表示向量. 在进行全局最大池化之前, 需要确保属性的强度是非负的, 可以使用 ReLU 函数实现. 通过激活函数和全局最大池化操作, 可以将特征图中每个属性的强度汇总成一个属性表示向量, 该向量能更有效地表示图像中不同属性的存在情况, 并为后续的决策树推理提供更有意义的解释信息.

为提高骨干网络面对不同模态数据时提取属性的一致性, 在批归一化 BatchNorm 层用其他模态的数据替换特定模态中的低质量信息, 即 BatchNorm 通道交换[18]. 公式可表示为:

$$\boldsymbol{y}^i=\begin{cases}\gamma^i\dfrac{\boldsymbol{x}^i-\mu(\boldsymbol{x}^i)}{\sqrt{\delta^2(\boldsymbol{x}^i)}}+\beta^i, & \gamma^i>\theta\\ \dfrac{1}{N-1}\displaystyle\sum_{j\neq i}\gamma^j\dfrac{\boldsymbol{x}^j-\mu(\boldsymbol{x}^j)}{\sqrt{\delta^2(\boldsymbol{x}^j)}}+\beta^j, & \text{其他}\end{cases}\tag{1}$$

其中 $\boldsymbol{x}^i$ 表示第 i 个模态的 BatchNorm 层输入; $\boldsymbol{y}^i$ 表示相应的输出; γ^i 和 β^i 是 BatchNorm 层可训练的参数, 分别表示缩放因子和偏移量. 当缩放因子小于一定阈值时 $\theta\approx 0^+$, 输入 $\boldsymbol{x}^i$ 对输出 $\boldsymbol{y}^i$ 的影响很小, 可以认为输入 $\boldsymbol{x}^i$ 不重要, 能够用其他模态数据的平均值对齐进行替代. $\mu(\boldsymbol{x}^i)$ 和 $\delta^2(\boldsymbol{x}^i)$ 分别表示 $\boldsymbol{x}^i$ 的平均值和方差.

为保证这种交换能够进行, 本文对一半的 BatchNorm 层进行 ℓ_1 正则化, 使权重参数更加稀疏. 正则化损失可以表示为:

$$\mathcal{L}_{\ell_1}=\eta\sum_{i=1}^{N}\sum_{l=1}^{\lceil\frac{L}{2}\rceil}|\widehat{\gamma_l^i}|\tag{2}$$

其中, η 是超参数, L 表示单一模态 BatchNorm 层的数量, $\widehat{\gamma_l^i}$ 表示第 i 个模态第 l 个 BatchNorm 层的缩放因子参数.

2.2 决策树的构建与融合

类别之间天然具有一定的可用树形表示的层次结构, 利用这种层次结构可以构建决策树进行推理, 由于层次结构与模态无关, 因此不同模态的决策树遵循相同的类别层次结构.

在树形层次结构中, 第 k 个节点的高度 h_k 反映第 k 类的所有上级类别的数量. $\boldsymbol{H}\in\{0,\ 1\}^{M\times M}$ 表示引入的类别层次结构, 其中 $\boldsymbol{H}_{k\cdot}$ 表示第 k 类的所有上级类别, M 表示类型的个数. 类别层次结构 $\boldsymbol{H}$ 可表示为:

$$\boldsymbol{H}_{ij}=\begin{cases}1, & \text{节点 } j \text{ 是 } i \text{ 的祖先}\\ 0, & \text{其他}\end{cases}\tag{3}$$

决策树 T 由属性 $\boldsymbol{A}^i\in[0,\ +\infty)^D$, 推理规则 $\boldsymbol{W}\in\mathbf{R}^{D\times M}$ 和层次结构 $\boldsymbol{H}$ 组成. 推理规则 $\boldsymbol{W}$ 将属性 $\boldsymbol{A}$ 映射为属于该节点类型的证据 $\boldsymbol{e}$, 可由全连接层实现. 决策树的每一个节点对应一个类别, $\boldsymbol{W}_{\cdot k}$ 表示节点 k 的推理规则[27], 那么, 第 i 个模态中节点 k 的证据可以通过 $e_k^i=\boldsymbol{A}^i\boldsymbol{W}_{\cdot k}$ 计算.

决策树按层融合过程如图 4 所示. 受根据每个模态的数据质量动态调整融合权重的可信多模态融合方法[19] 的启发, Dempster-Shafer 证据理论和主观逻辑理论[42] 可以用于决策树的融合, 提升融合准确率. 但是在评估模态质量时, 主观逻辑理论要求样本只能拥有一个标签 k. 然而, 在决策树中, 子分类同时属于其父级分类, 但是在同一层级仅属于一个分类, 因此需要按层进行融合. 对于高度为 q 的节点, 满足以下关系:

$$\sum_{h_k=q}p_k=1\tag{4}$$

对于第 i 个模态具有相同高度的节点, 在主观逻辑理论中, 狄利克雷分布 $\boldsymbol{\alpha}^i$ 与证据 $\boldsymbol{E}^i=\{e_{k_1}^i,\ e_{k_2}^i,\ \cdots,\ e_{k_r}^i\}$ 的关系是 $\boldsymbol{\alpha}^i=\boldsymbol{E}^i+1$, 其中满足 $h_{k_1}=h_{k_2}=\cdots=h_{k_r}$, r 代表该高度的节点数, 狄利克雷强度定义为 $S^i=\sum_{k=1}^{r}\alpha_k^i$. 因此, 本文可以得到分类 k 的置信度 b_k^i 和第 i 个模态不确定度 u^i 如下:

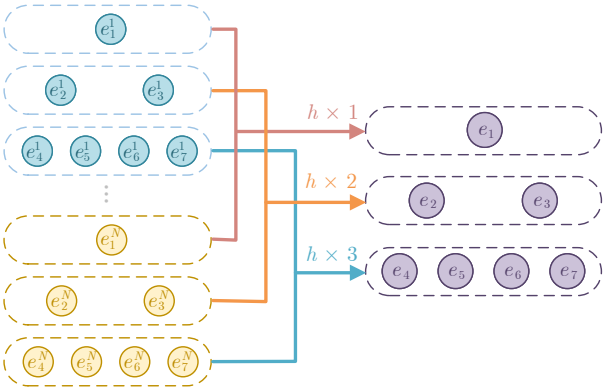

图 4　按层融合各模态决策树

Fig. 4　Fuse the decision tree of each modal layer-by-layer

$$b_k^i = \frac{E_k^i}{S^i} = \frac{\alpha_k^i - 1}{S^i} \quad 和 \quad u^i = \frac{r}{S^i} \tag{5}$$

将各模态依次两两融合, 第 i 个模态和第 $i+1$ 个模态融合公式表示为:

$$C = \sum_{k_i \neq k_{i+1}} b_{k_i}^i b_{k_{i+1}}^{i+1} \tag{6}$$

$$b_k = \frac{1}{1-C}(b_k^i b_k^{i+1} + b_k^i u^{i+1} + b_k^{i+1} u^i) \tag{7}$$

$$u = \frac{1}{1-C} u^i u^{i+1} \tag{8}$$

$$E_k = b_k \times S = b_k \times \frac{r}{u} \tag{9}$$

其中 $\boldsymbol{b}^i$ 和 u^i 表示第 i 个模态的分类置信度和不确定度, 相应的 $\boldsymbol{b}^{i+1}$ 和 u^{i+1} 表示第 $i+1$ 个模态的分类置信度和不确定度, C 反映两个模态之间的冲突程度, $\boldsymbol{E}$ 表示融合后的证据.

多模态融合后的交叉熵损失根据狄利克雷分布进行调整, 以模拟交叉熵损失的最大似然估计. 对于一个狄利克雷分布 $\boldsymbol{\beta}$, 其交叉熵损失的计算公式为[43]:

$$\mathcal{L}_{\mathrm{CE}}(\boldsymbol{\beta}_i) = \sum_{j=1}^{M} y_{ij} \left(\psi\left(\sum_{j=1}^{M} \beta_{ij} \right) - \psi(\beta_{ij}) \right) \tag{10}$$

其中类别的个数和 β_{ij} 分别表示类别数和第 i 个样本在第 j 个类别概率的狄利克雷分布, $\psi(\cdot)$ 是二伽马函数.

此外, 相比于保证正确标签的贡献大于其他标签, 引入 KL 散度[44] 减少错误标签的共享. 最终, 狄利克雷分布 $\boldsymbol{\beta}$ 的融合损失表示为:

$$\mathcal{L}_{\mathrm{fusion}}(\boldsymbol{\beta}_i) = \mathcal{L}_{\mathrm{CE}}(\boldsymbol{\beta}_i) + \lambda_t KL\left[D(\boldsymbol{p}_i \mid \widetilde{\boldsymbol{\beta}}_i) \| D(\boldsymbol{p}_i \mid \mathbf{1}) \right] \tag{11}$$

其中, $D(\boldsymbol{p}_i \mid \boldsymbol{\beta}_i)$ 是针对狄利克雷分布 $\boldsymbol{\beta}_i$ 形成的多项式意见, $\boldsymbol{p}_i$ 表示单纯形上的类别概率. $\widetilde{\boldsymbol{\beta}}_i$ 由公式 $\widetilde{\boldsymbol{\beta}}_i = \boldsymbol{y}_i + (1 - \boldsymbol{y}_i) \odot \boldsymbol{\beta}_i$ 计算得到, $\odot$ 是逐元素乘法, 表示移除错误证据后的狄利克雷分布. $\lambda_t = \min(1, t/\lambda)$ 从 0 逐渐增加到 1, 以减少在训练阶段早期 KL 散度的影响, t 是当前训练的周期数, λ 是限制增长率的超参数.

2.3　决策树推理

各模态的决策树 T^i 经过融合后, 得到可用于推理的决策树 $\widehat{T}$. 研究表明, 软推理比硬推理更准确[45], 软推理过程如图 5 所示. 基于软推理规则, 对于决策树 T, 计算节点 v 处所有子节点的归一化指数 Softmax, 公式表示为:

$$e_{k_i} = \frac{\exp s_{k_i}}{\sum_{k_i \in \mathrm{child}_v} \exp s_{k_i}} \tag{12}$$

其中 $k_i \in \mathrm{child}_v$ 表示 v 的子节点, e_{k_1} 是节点 k_1 的证据, s_{k_1} 表示从节点 v 向其子节点 k_1 转移的单级转移概率. 每个节点的分类预测概率 p_k 是从节点 k 到根节点之间所有节点 $\mathcal{V}$ 的转移概率乘积, 表示为:

$$p_k = \prod_{v \in \mathcal{V}} s_v \tag{13}$$

由叶子节点概率 $\boldsymbol{p}'$ 和证据 $\boldsymbol{E}'$ 计算决策树 T 的推理损失 $\mathcal{L}_{\mathrm{infer}}$, 即:

$$\mathcal{L}_{\mathrm{infer}}(T) = \mathcal{L}_{\mathrm{CE}}(\boldsymbol{p}' + \mathbf{1}) + \mathcal{L}_{\mathrm{fusion}}(S(\boldsymbol{E}') + \mathbf{1}) \tag{14}$$

其中 $S(\cdot)$ 是 Softplus 函数, $\boldsymbol{p}' + \mathbf{1}$ 和 $S(\boldsymbol{E}') + \mathbf{1}$ 分别为 $\boldsymbol{p}'$ 和 $S(\boldsymbol{E}')$ 的狄利克雷分布.

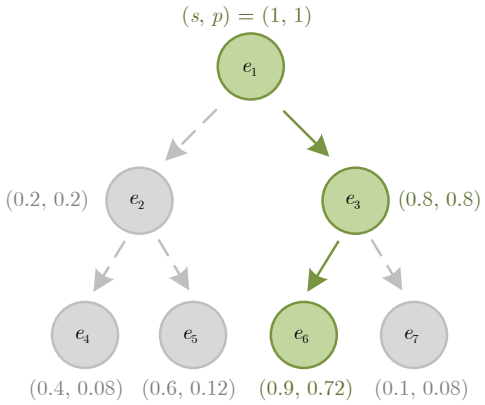

图 5　决策树进行软推理

Fig. 5　Apply soft inference for the decision tree

决策树共有 $N+1$ 棵, 包含 N 棵各模态的决策树 T^i 和一棵融合后的决策树 $\widehat{T}$. 最终的属性推理损失定义为正则化损失 $\mathcal{L}_{\ell_1}$ 与各个决策树推理损失平均值的和, 即:

$$\mathcal{L}_{\text{attribute}} = \mathcal{L}_{\ell_1} + \frac{1}{N+1}\left[\mathcal{L}_{\text{infer}}(\widehat{T}) + \sum_{i=1}^{N}\mathcal{L}_{\text{infer}}(T^i)\right] \tag{15}$$

2.4 原型限制

每个类别都继承其直接上级类别所有的属性, 并有自身独特的属性, 即该类别的原型. 因此, 类别原型是其直接上级的类别原型和其独特属性的和, 而其直接上级的类别原型也为继承部分与独特部分的和, 故类别原型最终可表示为其所有上级独特属性与该类别独特属性的和.

假设 $\boldsymbol{U} \in [0, +\infty)^{M\times D}$ 和 $\boldsymbol{P} \in [0, +\infty)^{M\times D}$ 分别表示独特属性和原型, 关系如下:

$$\boldsymbol{P} = \boldsymbol{H}\boldsymbol{U} \tag{16}$$

其中 $\boldsymbol{U}$ 是可训练的参数. 原型限制帮助模型提高提取属性的能力, 将第 i 个类别的原型作为属性向量构建原型决策树 $T_{p_i} = \{\boldsymbol{P}_i, \boldsymbol{W}, \boldsymbol{H}\}$, 原型损失类似于式 (14), 表示为:

$$\mathcal{L}_{\text{prototype}}(T_{p_i}) = \mathcal{L}_{\text{CE}}(\boldsymbol{p}' + \boldsymbol{1}) - \log\left(\frac{\exp \boldsymbol{P}_i\boldsymbol{W}_{\cdot i}}{\sum\limits_{k\in\text{leaf}}\exp \boldsymbol{P}_i\boldsymbol{W}_{\cdot k}}\right) \tag{17}$$

不同之处在于, 多模态融合损失 $\mathcal{L}_{\text{fusion}}$ 被替换为 Softmax 分类损失.

最后, 整个模型的损失函数为属性推理损失和原型损失的和, 公式为:

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{attribute}} + \mathcal{L}_{\text{prototype}}(T_{p_y}) \tag{18}$$

其中 y 表示样本的类别标签.

3 实验

本节将通过多个实验来评估模型在 3 个多模态图像数据集 NYUDv2[46]、SUN RGB-D[47] 和 RGB-NIR[48] 上的性能. 介绍超参数的选择方法, 验证模型每个部分的功能, 并与之前的方法进行比较, 以验证本文的方法具有出色的准确性和可解释性.

3.1 实验设置

3.1.1 数据集

本文在 NYUDv2、SUN RGB-D 和 RGB-NIR 数据集上进行 RGB-D 和 RGB-NIR 的多模态场景分类. NYUDv2 数据集包括 1 449 组对齐的可见光和深度图像, 分为 27 个场景. 实验挑选出包含超过 50 组图像的 7 个类别. 同样, 包含 10 335 组 RGB-D 图像的 SUN RGB-D 数据集被重组为包含 20 个类别的 9 585 组图像. RGB-NIR 数据集包含 476 组对齐的 RGB 和远红外图像, 分为 9 个场景.

根据场景的类型, 为每个数据集人工指定 3 层的层次结构. 为减少随机性, 本文将所有数据集随机分成 10 份并采用 10 折交叉验证, 并且设置固定的随机种子.

3.1.2 卷积神经网络搭建

在提取属性的过程中, 本文所提出的模型不关注骨干网络的具体架构, 任何卷积神经网络通过简单的修改都可以作为骨干网络. 在本文研究过程中, 采用残差神经网络 ResNet[49] 提取特征, 如 ResNet-18 特征图大小为 $512\times 7\times 7$, 这意味着 ResNet-18 可以提取 $D = 512$ 个属性. 为避免 ReLU 函数可能导致神经元“死亡”问题, 采用负斜率为 0.01 的 leaky ReLU[50] 作为激活函数.

通过复制将深度图像和远红外图像从单通道转换为三通道, 并将所有图像大小调整为 256×256 像素, 经过一系列图像增强处理后, 输入卷积神经网络, 得到表示属性强度的长度为 512 或 2 048 的一维向量. 在 10 个 Tesla P40 GPU 上进行 10 折交叉验证的并行训练, 使用 3×10^{-4} 学习率和权重衰减为 10^{-5} 的 Adam 优化器训练网络. BatchNorm 层的初始权重 γ 服从 0 到 1 之间的均匀分布.

3.2 超参数设置

如式 (1) 和 (2) 中所讨论的, BatchNorm 交换由两个超参数组成: 阈值 θ 和正则化损失权重 η. 一般来说, 随着阈值 θ 和正则化损失权重 η 的增加, 正则化损失和发生交换的概率也随之增加. 为找到合适的超参数设置, 在样本规模最大的 SUN RGB-D 上进行实验来寻找超参数与准确率之间的关系.

通过实验枚举 θ 在 10^{-5} 到 10^{-1} 与 η 在 5×10^{-6} 到 10^{-2} 之间的组合, 得到该参数组合下模型的准确率, 利用高斯拟合消除抖动后绘制热力图, 如图 6 所示, 图中颜色越浅, 准确率越高. 观察热力图发现, 在 $\theta \in [10^{-2}, 10^{-1}]$, $\eta \in [10^{-5}, 10^{-4}]$ 区域准确率较高, 因此在后续实验中, 本文设置阈值 $\theta = 10^{-2}$ 和正则化损失权重 $\eta = 2\times 10^{-5}$.

3.3 模型准确性

3.3.1 消融实验

除通过骨干网络提取属性外, 模型主要包括三个模块: BatchNorm 通道交换、决策树构建与融合以及决策树推理. 为验证这三个模块的有效性而进行消融实验, 逐步应用这些模块以便分析它们对模

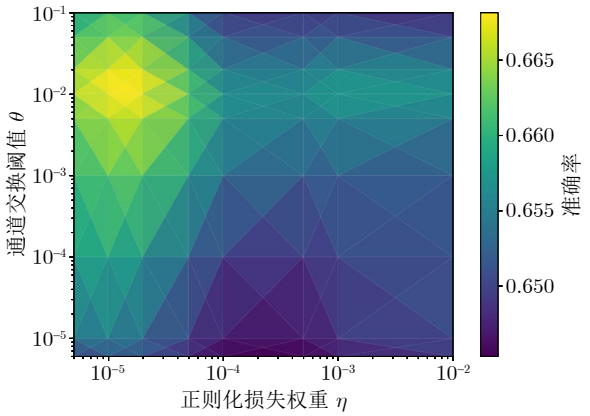

图 6　通道交换阈值 θ、正则化损失权重 η 与 Top-1 准确率在 SUN RGB-D 数据集上的关系

Fig. 6　The relationship between channel exchange threshold θ, regularization loss parameters η, Top-1 accuracy on SUN RGB-D

型性能的影响, 结果如表 1 所示. 在不应用树推理、树融合和通道交换的基础模型中, 我们将各模态的分类概率取平均值作为融合结果并计算损失.

为增强不同模态数据之间的一致性, 通道交换通过使用其他模态数据来替换特定模态中的低质量信息, 从而提高模型对多模态数据的表示能力. 相较于基础模型, 仅应用通道交换模块后, 三个数据集准确率均得到提升, 在 SUN RGB-D 数据集上提升超过 4%. 而在进行决策树推理、融合的模型上, 应用该模块后, 所有数据集上融合准确率都进一步得到提升, 在 SUN RGB-D 数据集上提升约 2%, 这一结果验证了通道交换模块的有效性.

决策树推理为模型提供更好的解释性, 利用决策树的逐层推理路径, 能够清楚地看到模型是如何从根节点沿着父类推理到最终类型, 使得模型的决策过程更加透明可理解. 模型的可解释性和准确性一般是互斥的, 决策树的引入增加了模型的学习和推理难度, 当决策树推理的层次结构变得复杂时, 准确率可能会降低. 例如在 SUN RGB-D 数据集中仅应用决策树推理时, 准确率下降了 3.2%, 但是通过决策树提供的层次信息, 模型才能够进行准确的属性和原型学习.

决策树融合模块为有效地结合多个模态的决策树, 使用 Dempster-Shafer 证据理论进行逐层融合, 动态适应各模态的数据质量, 最终得到一个融合决策树, 进一步提升模型的分类准确率, 并显著提高融合精度, 针对 3 个数据集, 准确率分别提高了 1.99%、7.17% 和 5.54%. 这表明决策树融合在将多个模态的决策树结合起来时, 能够更好地捕捉到不同模态之间的相关性, 从而提高整体模型的分类精度.

值得注意的是, 由于使用通道交换的方法, 每个模态会引入其他模态的数据, 因此无法直接比较单模态精度, 表 1 和表 2 中相关的数据用星号标记, 以示区别.

3.3.2　对比实验

与残差神经网络 ResNet-18 、针对小数据集优化的 Transformer 骨干网络 ViT-S-16、三种单模态可解释模型 (dNDF、NBDT、HCN) 以及三种多模态融合方法 (CBCL、TMC、TMNR) 进行比较, 结果如表 2 所示. 在小规模数据集场景下, ResNet 架构优于 Transformer ViT 架构, 即使使用针对小规模数据集优化的方法, 骨干网络 ViT-S-16 在三个数据集中准确率均低于 ResNet-18.

多模态数据能够提供比单模态数据更加丰富的信息, 实现数据互补, 从而提高分类准确率. 其中, CBCL[52] 是一种 RGB-D 聚类方法, TMC[19] 和 TMNR[53] 属于多模态证据融合方法, 相较于单模态方法, 准确率均有提升. 同时, 证据融合方法相较聚类方法不依赖所有模态都提供高质量数据, 有更高的准确率. 我们的方法与骨干网络 ResNet-18 在 NYUDv2、SUN RGB-D 和 RGB-NIR 数据集上相比, 准确率分别提高了 8.81%、1.97% 和 6.71%.

当前研究主要关注单模态可解释性, dNDF[54] 和 NBDT[27] 是基于决策树的解释方法, NBDT 引入

表 1　不同模块在 NYUDv2、SUN RGB-D 和 RGB-NIR 数据集上的 Top-1 准确率 (%)

Table 1　Top-1 accuracies with different components on NYUDv2, SUN RGB-D and RGB-NIR (%)

树推理	树融合	通道交换	NYUDv2			SUN RGB-D			RGB-NIR		
			RGB	Deep	Fusion	RGB	Deep	Fusion	RGB	NIR	Fusion
×	×	×	43.08	59.26	71.98	52.10	38.49	62.19	58.33	52.08	77.78
×	×	√	47.74*	59.47*	72.07	54.29*	47.05*	66.28	62.23*	53.76*	80.43
√	×	×	46.28	57.68	72.41	50.98	36.00	58.99	58.68	53.47	79.17
√	√	×	**61.43**	**61.00**	74.40	**59.96**	**51.62**	66.16	**71.08**	**66.45**	84.71
√	√	√	71.14*	70.99*	**74.74**	66.76*	66.37*	**68.01**	78.85*	77.37*	**85.54**

注: * 表示使用通道交换为单个模态引入其他模态数据后的准确率, 加粗表示单模态或融合后最高准确率.

表 2 不同方法在 NYUDv2、SUN RGB-D 和 RGB-NIR 数据集上的 Top-1 准确率 (%)
Table 2 Top-1 accuracies with different methods on NYUDv2, SUN RGB-D and RGB-NIR (%)

方法	解释性	NYUDv2			SUN RGB-D			RGB-NIR		
		RGB	Deep	Fusion	RGB	Deep	Fusion	RGB	NIR	Fusion
ViT-S-16[51]	×	54.95	62.56	—	59.23	49.43	—	74.44	66.32	—
ResNet-18[49]	×	65.28	65.93	—	66.04	57.85	—	**78.83**	**75.70**	—
CBCL[52]	×	56.87	63.20	73.85	50.74	43.59	65.78	74.23	62.91	81.72
TMC[19]	×	60.14	62.19	74.57	60.89	52.95	66.69	72.76	68.77	84.29
TMNR[53]	×	56.61	64.50	74.10	60.60	53.53	66.30	69.50	65.26	82.20
dNDF[54]	√	61.86	**65.76**	—	64.78	57.30	—	78.61	72.11	—
NBDT[27]	√	**65.28**	62.85	—	**66.20**	**57.93**	—	74.24	74.22	—
HCN[20]	√	62.20	63.18	—	61.91	53.03	—	72.92	68.75	—
Ours	√	71.14*	70.99*	**74.74**	66.76*	66.37*	**68.01**	78.85*	77.37*	**85.54**

注: * 表示使用通道交换为单个模态引入其他模态数据后的准确率, 加粗表示单模态或融合后最高准确率.

神经网络作为分类器, 因此准确率更高. 而 HCN[20] 是一类原型学习方法, 虽然准确率较低, 但可解释性更强. 可解释性的引入通常会造成模型精度降低, 得益于我们的方法能够利用互补的多模态信息, 因此相较于单模态方法更加精确, 与准确率最高的单模态可解释模型相比, 准确率分别提高了 8.98%、1.81% 和 6.93%. 层次信息和原型限制的引入, 使该方法在保持模型良好可解释性的同时, 仍具有较好的分类精度. 与不可解释的多模态融合方法相比, 准确率分别提高了 0.17%、1.32% 和 1.25%, 达到与不可解释的模型相近的准确率.

3.3.3 预训练骨干网络

任何 CNN 骨干网络都可被用于本文提出的模型之中, 并且骨干网络也可以在其他数据集上进行预训练. 如表 3 所示, 使用在 ImageNet 上预训练的 ResNet-18 骨干模型较表 2 中未预训练的骨干模型在 NYUDv2、SUN RGB-D 和 RGB-NIR 数据集上的准确率分别提升了 7.19%、6.95% 和 5.25%. 这表明预训练的骨干网络能够更准确地提取图像中具有的视觉属性.

复杂的骨干网络能够学习更丰富的特征表示, 从而提高模型的分类准确率. 例如预训练的 ResNet-101 可以提取 2 048 个属性, 与只能提取 512 个属性的 ResNet-18 相比, 在 NYUDv2、SUN RGB-D 和 RGB-NIR 数据集上的准确率分别提高了 1.03%、1.46% 和 0.64%.

表 3 不同预训练骨干网络在 NYUDv2、SUN RGB-D 和 RGB-NIR 数据集中的 Top-1 准确率 (%)
Table 3 Top-1 accuracies with different pretrained backbones on NYUDv2, SUN RGB-D and RGB-NIR (%)

骨干网络	NYUDv2	SUN RGB-D	RGB-NIR
ResNet-18	80.90	73.50	90.15
ResNet-34	81.58	73.87	90.15
ResNet-50	81.92	73.88	90.58
ResNet-101	**81.93**	**74.96**	**90.79**

3.4 模型可解释性

本文方法提供的解释性, 核心是将输入图像转化为一系列可视化的属性集合, 并展示决策树推理路径, 流程如图 2 所示.

3.4.1 可视化

为可视化模型学习到的属性, 本文用一组图像及其热力图来表示, 如图 7 所示, 同组图像具有一个共同的视觉属性特征. 具体而言, 要获得第 i 个模态中具有第 k 个属性的图像集合, 根据第 k 个属性强度 $\boldsymbol{A}_k^i$ 从大到小将所有图像进行排序, 并选择前若干个图像, 图像之间的共同部分就反映这一属性. 使用 Grad-CAM 可以生成该组图像对特点属性的热力图, 更加直观地表示图像集合具有的属性.

对于一组图像或类别原型, 通过其对应的属性强度 $\boldsymbol{A}$ 或类别原型 $\boldsymbol{P}_i$ 可以找到具有代表性的属性. 与可视化属性的过程类似, 将每个属性 $\boldsymbol{A}_k$ 或原型 $\boldsymbol{p}_{ik}$ 的属性强度从大到小排序, 选择强度最大的属性为代表, 属性由上一步提取的一组图像及热力图表示.

使用层次结构构建决策树后, 可以展示决策树在各单一模态和融合模态的推理过程, 可视化推理过程有助于验证模态的数据质量. 以图 8 为例, 输入数据首先被分类至校园场景, 随后进一步归属于自习室. 此时输入的深度图像不含任何可用信息, 难以进行场景分类, 得到概率较高且错误的分类结果. 通过 Dempster-Shafer 证据理论和主观逻辑理论计算深度图模态不确定度 $u = 1.62$, 自动降低该

模态在融合后的权重, 融合时更加采信不确定度较低 $u = 0.34$ 的可见光模态, 最终得到正确的分类结果.

3.4.2　插入/删除测试

为验证提取属性的正确性, 如图 9 所示, 本文对属性进行插入/删除测试[55]. 插入测试是向一张空白的图像中插入一些像素, 如果这些像素能够反映图像的分类类别独特且重要, 那么预测准确率将迅速增加. 同理, 删除测试是从一张完整的图像中删除一些像素, 如果这些像素是重要的, 那么预测准确率将迅速下降.

Grad-CAM 可以为每个属性生成热力图, 在插入测试中, 根据热力图的强度按热力值从大到小向空白图像中插入一定比率像素的图像. 在删除测试中, 根据热力图的强度从大到小删除一定比率像素的图像, 并比较准确率曲线 (Area under curve, AUC), 结果如表 4 所示. 结果表明, 删除最强属性

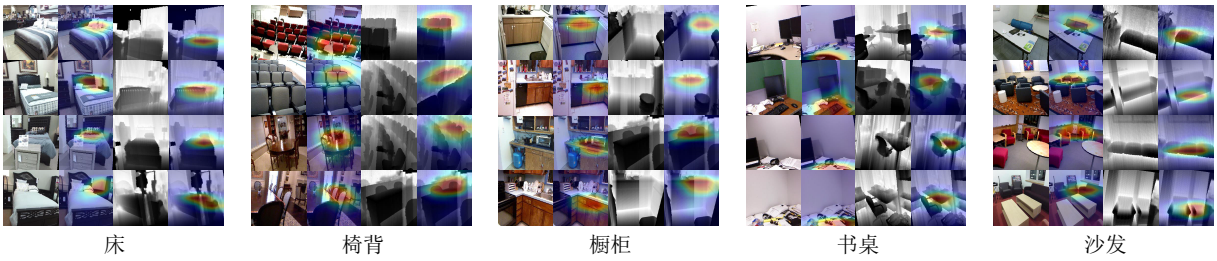

图 7　模型学习得到的属性

Fig. 7　Attributes learned by the model

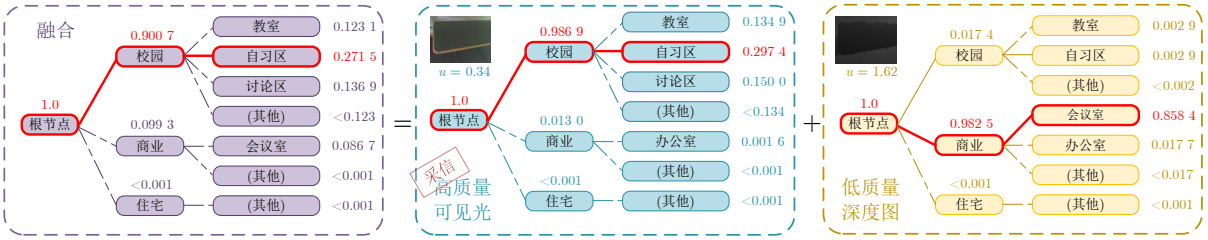

图 8　通过评估不确定度动态适应模态数据质量

Fig. 8　Dynamically adapt to modal data quality by evaluating uncertainty

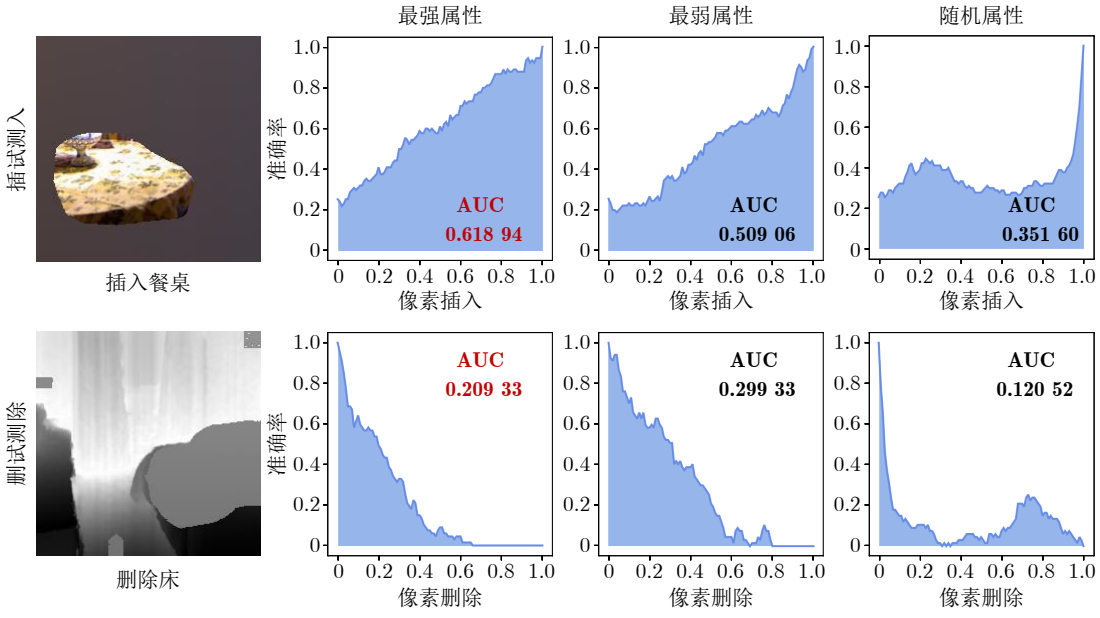

图 9　向原始图像中插入或删除属性

Fig. 9　Insert or delete attribute to the original image

表 4 插入或删除不同属性在 NYUDv2、SUN RGB-D 和 RGB-NIR 数据集中的 AUC

Table 4 AUC of different attribute inserted or deleted in NYUDv2, SUN RGB-D and RGB-NIR datasets

数据集	最强属性		最弱属性		随机	
	插入	删除	插入	删除	插入	删除
NYUDv2	**0.619**	0.209	0.509	0.299	0.351	**0.121**
SUN RGB-D	**0.601**	0.300	0.463	0.380	0.284	**0.168**
RGB-NIR	**0.636**	0.380	0.549	0.466	0.355	**0.207**

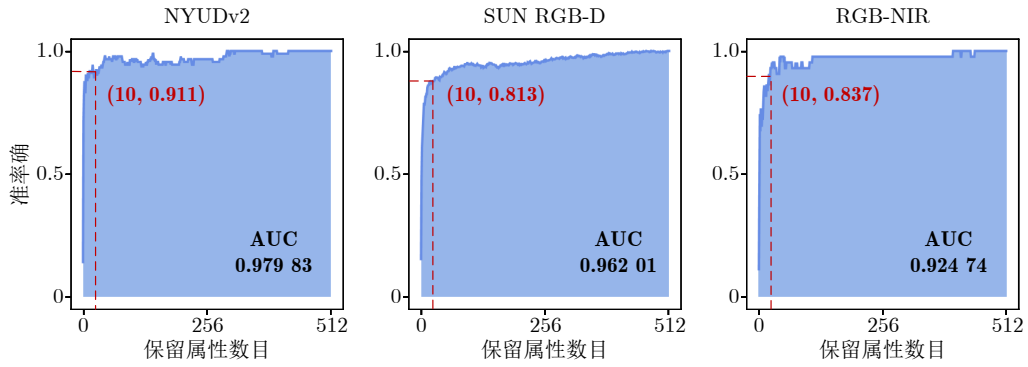

图 10 保留输入数据前 k 强的属性

Fig. 10 Preserve the first k strong attributes of the input data

比删除最弱的属性具有更低的 AUC, 在空白图像中插入最强的属性比插入最弱的属性有更高的 AUC, 这表明本文的模型能够正确评估属性强度, 进而找到图像所具有的属性. 由于每个样本都由多个属性组成, 因此在上述实验中, 随机删除的 AUC 比删除最强属性要低, 因为它可能会破坏多个属性.

同时, 为验证模型所提取属性的代表性, 本文尝试找出多少个属性可以代表一个样本. 我们将样本具有的属性按强度排序, 仅保留前 k 强的属性, 其余属性强度全部置为 0, 并绘制准确率曲线 AUC, 结果如图 10 所示. 实验结果表明, 仅保留 10 个属性即可在三个数据集中达到与原始模型 91.1%、81.3% 和 83.7% 的准确率, 这也表明本文的模型可以有效地提取具有代表性的属性.

4 总结

本文提出一种基于属性的多模态可解释分类方法. 通过使用骨干网络提取属性、构建和融合决策树并利用决策树进行推理, 该方法具有良好的可解释性, 能够可视化地展示输入数据所具有的属性以及决策树的推理过程. 同时, 与其他单模态可解释方法和多模态不可解释方法相比, 我们的方法均表现出优秀的性能. 本文方法是解决多模态融合可解释性问题的一个良好尝试, 在保持较高准确率的同时, 还能提供清晰的解释信息, 帮助人们理解模型的决策过程.

References

1 Zhao Jing, Pei Zi-Nan, Jiang Bin, Lu Ning-Yun, Zhao Fei, Chen Shu-Feng. Virtual tube visual obstacle avoidance for UAV based on deep reinforcement learning. *Acta Automatica Sinica*, 2024, **50**(11): 1−14
(赵静, 裴子楠, 姜斌, 陆宁云, 赵斐, 陈树峰. 基于深度强化学习的无人机虚拟管道视觉避障. 自动化学报, 2024, **50**(11): 1−14)

2 Miikkulainen R, Liang J, Meyerson E, Rawal A, Fink D, Francon O, et al. Evolving deep neural networks. *Artificial Intelligence in the Age of Neural Networks and Brain Computing (Second edition)*. Amsterdam: Academic Press, 2024. 269−287

3 Hassija V, Chamola V, Mahapatra A, Singal A, Goel D, Huang K Z, et al. Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation*, 2024, **16**(1): 45−74

4 Jung J, Lee H, Jung H, Kim H. Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. *Heliyon*, 2023, **9**(5): Article No. e16110

5 Costa V G, Pedreira C E. Recent advances in decision trees: An updated survey. *Artificial Intelligence Review*, 2023, **56**(5): 4765−4800

6 Aksjonov A, Kyrki V. A safety-critical decision-making and control framework combining machine-learning-based and rule-based algorithms. *SAE International Journal of Vehicle Dynamics, Stability, and NVH*, 2023, **7**(3): 287−299

7 Kitson N K, Constantinou A C, Guo Z G, Liu Y, Chobtham K. A survey of Bayesian Network structure learning. *Artificial Intelligence Review*, 2023, **56**(8): 8721−8814

8 Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In: Proceedings of the 2nd International Conference on Learning Representations (ICLR). Banff, Canada: ICLR, 2014. 1−8

9 Sundararajan M, Taly A, Yan Q Q. Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML). Sydney, Australia: JMLR,

2017. 3319−3328

10 Zhou B L, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 2921−2929

11 Chattopadhay A, Sarkar A, Howlader P, Balasubramanian V N. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV). Lake Tahoe, USA: IEEE, 2018. 839−847

12 Ribeiro M T, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA: Association for Computing Machinery, 2016. 1135−1144

13 Lundberg S M, Lee S I. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS). Long Beach, USA: Curran Associates Inc., 2017. 4768−4777

14 Chen C F, Li O, Tao C F, Barnett A J, Su J, Rudin C. This looks like that: Deep learning for interpretable image recognition. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS). Vancouver, Canada: 2019. Article No. 801

15 Nauta M, van Bree R, Seifert C. Neural prototype trees for interpretable fine-grained image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021. 14928−14938

16 Biederman I. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 1987, **94**(2): 115−147

17 Cohen L G, Celnik P, Pascual-Leone A, Corwell B, Faiz L, Dambrosia J, et al. Functional relevance of cross-modal plasticity in blind humans. *Nature*, 1997, **389**(6647): 180−183

18 Wang Y K, Huang W B, Sun F C, Xu T Y, Rong Y, Huang J Z. Deep multimodal fusion by channel exchanging. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS). Vancouver, Canada: Curran Associates Inc., 2020. Article No. 406

19 Han Z B, Zhang C Q, Fu H Z, Zhou J T. Trusted multi-view classification with dynamic evidential fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, **45**(2): 2551−2566

20 Liu H M, Wang R P, Shan S G, Chen X L. What is a tabby? Interpretable model decisions by learning attribute-based classification criteria. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, **43**(5): 1791−1807

21 Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 618−626

22 Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: Proceedings of the 34th International Conference on Machine Learning (ICML). Sydney, Australia: JMLR.org, 2017. 3145−3153

23 Zeiler M D, Fergus R. Visualizing and understanding convolutional networks. In: Proceedings of the 13th European Conference on Computer Vision (ECCV). Zurich, Switzerland: Springer, 2014. 818−833

24 Roberts L G. Machine Perception of Three-Dimensional Solids [Ph.D. dissertation], Massachusetts Institute of Technology, USA, 1963.

25 Farhadi A, Endres I, Hoiem D, Forsyth D. Describing objects by their attributes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Miami, USA: IEEE, 2009. 1778−1785

26 Yang H M, Zhang X Y, Yin F, Liu C L. Robust classification with convolutional prototype learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA: IEEE, 2018. 3474−3482

27 Wan A, Dunlap L, Ho D, Yin J H, Lee S, Petryk S, et al. NBDT: Neural-backed decision tree. In: Proceedings of the 9th International Conference on Learning Representations (ICLR). Austria: OpenReview.net, 2021.

28 Han X Y, Zhu X B, Pedrycz W, Li Z W. A three-way classification with fuzzy decision trees. *Applied Soft Computing*, 2023, **132**: Article No. 109788

29 Islam S, Haque M M, Karim A N M R. A rule-based machine learning model for financial fraud detection. *International Journal of Electrical and Computer Engineering (IJECE)*, 2024, **14**(1): 759−771

30 Hotelling H. Relations between two sets of variates. *Breakthroughs in Statistics: Methodology and Distribution*. New York: Springer, 1992. 162−190

31 Zhang J W, Yu Y, Tang S H, Wu J M, Li W. Variational autoencoder with CCA for audio——Visual cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023, **19**(3s): Article No. 130

32 Sapkota R, Thapaliya B, Suresh P, Ray B, Calhoun V D, Liu J Y. Multimodal imaging feature extraction with reference canonical correlation analysis underlying intelligence. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Korea: IEEE, 2024. 2071−2075

33 Tang Q, Liang J, Zhu F Q. A comparative review on multi-modal sensors fusion based on deep learning. *Signal Processing*, 2023, **213**: Article No. 109165

34 Li X J, Ma S Q, Xu J H, Tang J J, He S F, Guo F. TranSiam: Aggregating multi-modal visual features with locality for medical image segmentation. *Expert Systems With Applications*, 2024, **237**: Article No. 121574

35 Zheng X, Wang M H, Huang K, Zhu E. Global and cross-modal feature aggregation for multi-omics data classification and application on drug response prediction. *Information Fusion*, 2024, **102**: Article No. 102077

36 Hou M X, Zhang Z, Liu C, Lu G M. Semantic alignment network for multi-modal emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023, **33**(9): 5318−5329

37 Song Z Y, Wei H Y, Bai L, Yang L, Jia C Y. GraphAlign: Enhancing accurate feature alignment by graph matching for multi-modal 3D object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 3335−3346

38 Xue Z H, Marculescu R. Dynamic multimodal fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR). Vancouver, Canada: IEEE, 2023. 2575−2584

39 de Vries H, Strub F, Mary J, Larochelle H, Pietquin O, Courville A. Modulating early visual processing by language. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS). Long Beach, USA: Curran Associates Inc., 2017. 6597−6607

40 Du C Z, Teng J Y, Li T L, Liu Y C, Yuan T Y, Wang Y, et al. On uni-modal feature learning in supervised multi-modal learning. In: Proceedings of the 40th International Conference on Machine Learning (ICML). Honolulu, USA: JMLR.org, 2023. Article No. 345

41 Dempster A P. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 1967, **38**(2): 325−339

42 Jøsang A. *Subjective Logic: A Formalism for Reasoning Under Uncertainty*. Cham: Springer Publishing Company, 2016. 1−326

43 Sensoy M, Kaplan L, Kandemir M. Evidential deep learning to quantify classification uncertainty. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS). Montréal, Canada: Curran Associates Inc., 2018. 3183−3193

44 Higgins I, Matthey L, Pal A, Burgess C P, Glorot X, Botvinick M M, et al. Beta-VAE: Learning basic visual concepts with a constrained variational framework. In: Proceedings of the 5th International Conference on Learning Representations (ICLR). Toulon, France: OpenReview.net, 2017.

45 İrsoy O, Yıdız O T, Alpaydın E. Soft decision trees. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR). Tsukuba, Japan: IEEE, 2012. 1819−1822

46 Silberman N, Hoiem D, Kohli P, Fergus R. Indoor segmentation and support inference from RGBD images. In: Proceedings of the 12th European Conference on Computer Vision (ECCV). Florence, Italy: Springer, 2012. 746−760

47 Song S R, Lichtenberg S P, Xiao J X. SUN RGB-D: A RGB-D scene understanding benchmark suite. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015. 567−576

48 Brown M, Süsstrunk S. Multi-spectral SIFT for scene category recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Colorado Springs, USA: IEEE, 2011. 177−184

49 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 770−778

50 Maas A L, Hannun A Y, Ng A Y. Rectifier nonlinearities improve neural network acoustic models. In: Proceedings of the 30th International Conference on Machine Learning (ICML). Atlanta, USA: JMLR, 2013. 3−8

51 Lee S, Lee S, Song B C. Improving vision transformers to learn small-size dataset from scratch. *IEEE Access*, 2022, **10**: 123212−123224

52 Ayub A, Wagner A R. Centroid based concept learning for RGB-D indoor scene classification. In: Proceedings of the 31st British Machine Vision Conference (BMVC). Virtual Event: BMVA, 2020. 1−13

53 Xu C, Zhang Y L, Guan Z Y, Zhao W. Trusted multi-view learning with label noise. In: Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI). Jeju Island, Korea: IJCAI, 2024. 5263−5271

54 Kontschieder P, Fiterau M, Criminisi A, Bulò S R. Deep neural decision forests. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 1467−1475

55 Petsiuk V, Das A, Saenko K. Rise: Randomized input sampling for explanation of black-box models. In: Proceedings of the British Machine Vision Conference (BMVC). Newcastle, UK: BMVA, 2018. 151−163

王 辉 华东交通大学信息与软件工程学院副教授. 主要研究方向为人工智能, 计算机视觉.
E-mail: huiwangens@163.com
(**WANG Hui** Associate professor at the School of Information and Software Engineering, East China Jiaotong University. His research interest covers artificial intelligence and computer vision.)

黄宇廷 浙江大学软件学院硕士研究生. 主要研究方向为自然语言处理, 可解释人工智能.
E-mail: yutinghuang@zju.edu.cn
(**HUANG Yu-Ting** Master student at the School of Software Technology, Zhejiang University. His research interest covers natural language processing and explainable artificial intelligence.)

夏玉婷 华东交通大学信息与软件工程学院硕士研究生. 主要研究方向为计算机视觉, 多模态图像融合.
E-mail: xiayuting0403@126.com
(**XIA Yu-Ting** Master student at the School of Information and Software Engineering, East China Jiaotong University. Her research interest covers computer vision and multimodal image fusion.)

范自柱 上海电力大学计算机科学与技术学院教授. 主要研究方向为模式识别与机器学习. 本文通信作者.
E-mail: zzfan3@163.com
(**FAN Zi-Zhu** Professor at the College of Computer Science and Technology, Shanghai University of Electric Power. His research interest covers pattern recognition and machine learning. Corresponding author of this paper.)

罗国亮 华东交通大学信息与软件工程学院教授. 主要研究方向为计算机视觉, 人工智能.
E-mail: luoguoliang@ecjtu.edu.cn
(**LUO Guo-Liang** Professor at the School of Information and Software Engineering, East China Jiaotong University. His research interest covers computer vision and artificial intelligence.)

杨 辉 轨道交通基础设施性能监测与保障国家重点实验室教授. 主要研究方向为复杂系统建模, 控制与运行优化.
E-mail: yhshuo@263.com
(**YANG Hui** Professor at the State Key Laboratory of Performance Monitoring and Protecting of Rail Transit Infrastructure, East China Jiaotong University. His research interest covers modeling, control and operation optimization of complex systems.)