

Hybrid-Gird: 遥感图像细粒度分类可解释方法

朱凯雯¹, 尤亚楠¹, 曹婧宜¹, 孟钢², 乔媛媛¹, 杨洁¹

1. 北京邮电大学 人工智能学院, 北京 100876;

2. 北京市遥感信息研究所, 北京 100192

摘要: 基于遥感图像的目标细粒度分类深度神经网络已技术日益成熟, 网络决策的可解释性研究是当前细粒度分类深度学习算法进一步提高决策可信度的关键问题。为精确表征对模型决策起决定性作用的本质特征, 本文基于博弈竞争理论对遥感图像目标细粒度分类任务进行建模, 分析了IG、SmoothGrad、Grad-CAM等可解释性方法在遥感图像目标细粒度分类网络上的适用性, 提出了一种尺度自适应的目标细粒度分类本质特征可解释性分析方法 Hybrid-Grid, 使用像素级与局部特征关系融合算法提高对支撑网络决策的目标本质特征的精确描述能力。结果表明: 本文提出的 Hybrid-Grid 对目标细粒度分类网络的解释效果在 ADCC 量化评估指标上达到 78.87, 相较 Score-CAM 有大幅提升; 与 SmoothGrad、Grad-CAM 的解释结果相对比, 本文方法在删除及精度损失实验上表现最好, 使 EFM-Net 的 Top-1 准确率、Top-5 准确率、F1 得分分别损失了 16.92%、1.61%、17.21%, 证明 Hybrid-Grid 准确解释了对细粒度分类网络决策贡献最大的目标本质特征。本文提出的可解释性分析方法能够更精准地揭示当前目标细粒度分类网络的决策特征依据。

关键词: 遥感图像, 可解释性分析方法, 目标细粒度分类网络, 可解释人工智能, 合作博弈理论, 本质特征

中图分类号: TP701/P2

引用格式: 朱凯雯, 尤亚楠, 曹婧宜, 孟钢, 乔媛媛, 杨洁. 2024. Hybrid-Gird: 遥感图像细粒度分类可解释方法. 遥感学报, 28(7): 1722-1734

Zhu K W, You Y N, Cao J Y, Meng G, Qiao Y Y and Yang J. 2024. Hybrid-Gird: An explainable method for fine-grained classification of remote sensing images. National Remote Sensing Bulletin, 28(7): 1722-1734 [DOI: 10.11834/jrs.20243252]

1 引言

遥感图像目标细粒度分类任务是遥感图像智能解译领域的一项重要任务, 基于光学遥感图像的目标细粒度分类技术已得到广泛应用, 例如空域管制、态势评估、重要港口目标监视等 (Zhang 等, 2023)。不同于对目标的大类进行判断 (Chen 等, 2016), 目标细粒度分类旨在对属于同一基础类别的对象进行子类划分, 如区分巡洋舰、集装箱船、油轮等不同用途的船舶目标细粒度识别 (Han 等, 2022)。目标细粒度分类任务的关键在于分类模型能否获得对目标类别具有决定性贡献的本质特征。Cheng 等 (2021) 研究证明占据目标较

大部分的结构对于分类的贡献度较低, 细粒度分类通常强迫神经网络去学习、记忆相似类别间的细微差异, 并去除一些共性特征。因此, 只有少量的、适应于特定任务的特征会被建模。

近年来, 随着深度学习的迅猛发展, 许多先进的目标细粒度分类网络被提出, 并取得了显著的性能提升。已有研究 (Nie 等, 2022; Xiong 等, 2022; Liang 等, 2020; Fu 等, 2019) 主要基于深度卷积神经网络, 即 DCNN (Deep Convolution Neural Network), 通过优化特征提取能力和增强特征来提升模型的性能。Yi 等 (2023) 建立了一种基于本质特征挖掘的目标细粒度分类网络 EFM-Net, 该网络通过局部特征提取、注意力机制、局部区域

收稿日期: 2023-07-11; 预印本: 2024-03-20

基金项目: 国家自然科学基金 (编号: 62101060)

第一作者简介: 朱凯雯, 研究方向为遥感图像智能解译、可解释人工智能。E-mail: zhukaiwen@bupt.edu.cn

通信作者简介: 尤亚楠, 研究方向为成像探测与智能感知, 人工智能遥感图像理解。E-mail: youyanan@bupt.edu.cn

相关代码链接: PMG-V2, <https://github.com/PRIS-CV/PMG-V2>

EFM-Net, <https://github.com/JACYI/EFM-Net-Pytorch>

特征融合以及迁移学习等策略, 获取对细粒度类别特定的本质特征表示, 并利用这些特征进行准确的分类, 在 FGSC-23、FGSCR-43、Aircraft-16 等具有挑战性的遥感图像目标细粒度分类数据集上取得了优异的结果。

尽管这些网络提升了遥感图像目标细粒度分类任务的精度, 但深度学习网络模型的“黑盒”性质使人们仍然难以理解网络在细粒度分类任务上的决策依据, 这不仅限制了深度神经网络通过反馈指导进行优化提升的可能, 更使其无法被人类充分信任并应用于军事、医疗等重要领域。

当前, 国内外针对深度神经网络的可解释性分析方法可分为激活值最大化分析法、代理模型分析法、归因分析法、扰动分析法、类激活图分析法及样例分析法等 6 类方法 (Gong 等, 2022)。而围绕遥感图像智能解译任务, 到目前为止, 仍主要采用面向自然图像的可解释性分析方法, 如 IG、GuidedBackPropagation、SmoothGrad、occlusion (Petsiuk 等, 2018) 等。然而, 上述方法缺乏统一理论框架, 在遥感图像目标细粒度分类任务中存在适用性有限、适用任务范围狭窄等问题。

图像分类任务主要使用包括归因分析法、扰动分析法和类激活图分析法在内的归因技术来对分类网络的决策过程进行解释。该技术通过生成显著图 (Saliency Map) 来显示不同像素对模型的影响, 又依据是否需要获取神经网络内部权重和参数分为内在和外在两种。内在归因方法通过神经网络的层将重要性得分从模型的输出反向传播到输入中的各个像素, 例如 Gradients、Grad-CAM (Selvaraju 等, 2017; Simonyan 等, 2014)。但是, 由于分类网络内部卷积层的尺度较小, 将其映射回原图时会导致显著图丢失细节信息, 难以生成较为清晰的特征边缘。而细粒度分类任务主要关注相似目标中的可分性特征, 因此, 这种归因方法不能精确显示对分类影响最大的本质特征区域, 无法对细粒度分类的决策依据给出清晰明确的解释。

外在归因方法也即基于遮挡的扰动方法, 通过对图像区域进行特定的扰动, 如遮挡、添加噪声、修复和模糊 (Petsiuk 等, 2018), 观察扰动对模型输出的影响以确定扰动区域的重要性。由于其实实现方法会赋予一些干扰像素较高的重要性, 在显著图中引入了遮挡偏差, 最终导致对细粒度分类网络的可解释工作出现误判。

总的来说, 现有可解释性分析方法对细粒度特征的描述能力有限, 缺乏面向可解释性的细粒度分类任务的数学框架, 以及在认知层面刻画本质特征、适用于遥感图像多尺度特性的可解释性分析方法。因此, 本文针对现有可解释性分析方法描述能力有限、缺乏统一数学框架等问题, 提出一种具有框架一致性、尺度自适应的可解释性分析方法, 探究了目标本质特征对网络决策过程的影响情况, 能够可视化地诠释细粒度目标本质特征, 以期为目标细粒度分类网络的可信应用提供参考。

2 研究方法

2.1 细粒度分类网络

随着高分辨率遥感图像数据的持续增加, 遥感图像的有效判读变得越来越重要, 也使遥感图像的目标细粒度分类任务成为可能, 然而, 目标样本长尾分布限制了目标细粒度识别精度的提升, 准确获取图像中目标的可分性特征是实现高精度目标细粒度分类的关键。当前领域内主要通过特征挖掘与样本增强来提高分类精度: 一方面, 通过使用先进的特征提取与融合网络, 深度挖掘本质特征; 另一方面, 通过对目标样本进行数据增强, 提高样本质量以优化模型性能。

具体来说, PMG-V2 是目标细粒度分类领域目前性能最优的模型之一 (Du 等, 2022), 该网络由两个相互作用的模块组成。通过一个创新的类别一致块卷积来鼓励网络学习特定粒度中类别一致的特征。在训练阶段, 块状卷积使特征图在输入每个卷积层之前被分成一些块, 每个块在卷积时不能获取临近部分的信息, 这一操作并不会引入人工边界, 同时延续了拼图补丁的益处。通过应用一致性约束, 进一步在特征块进行卷积时从各个粒度获取具有意义的区域。

EFM-Net 是遥感图像目标细粒度分类领域目前性能最优的模型之一, 由本质特征挖掘模块和数据增强模块组成。EFM-Net 通过图像金字塔网络来提取更具代表性的特征, 使用 ConvNext 骨干网络提取目标的本征特征, 采用自注意力机制进行特征增强, 提出双线性特征聚合池化方法, 用以融合本征特征和注意力特征, 最终得到最具可分性的本质特征, 帮助网络更准确地定位判别区

域，抓住对细粒度分类任务最关键的本质特征。在训练阶段，数据增强模块可以自适应地增强训练样本。本质特征挖掘模块和数据增强模块以无监督的方式相互加强以提取准确的特征。

除了上述具有代表性的PMG-V2、EFM-Net网络之外，一些经典的分类网络，如ResNet18（He等，2016）也可用于目标细粒度分类任务。因此，本文以PMG-V2、EFM-Net和ResNet18为研究对象，试图建立一种可行的面向遥感图像目标细粒

度分类任务的可解释性分析方法。

2.2 内在与外在可解释性分析方法

现有可适用于目标细粒度分类任务的可解释性分析方法主要专注于模型在测试样本点附近局部空间上的决策行为，也即目标级可解释（Yang等，2023）。根据可解释性分析方法构建过程是否基于网络权重及内部结构，本文将归因方法分为内在、外在2种，并总结了常见的内在、外在归因方法，详见表1。

表 1 本研究内在与外在归因方法
Table 1 Subjective and objective attribution methods in this study

类别	名称	公式	实现方法
内在归因方法	Score-CAM	$a_k = f(X \circ H_{mn}^k) - f(X_b)$ $a_{mn} = \text{RELU}\left(\sum_{k=1}^K a_k A_{mn}^k\right)$	通过将卷积层特征图与目标类别的分数相乘,然后进行反向传播,生成类别激活图,强调模型决策的关键区域(Wang等,2020)
	Grad-CAM	$a_k^c = \frac{1}{W \times H} \sum_{m=1}^W \sum_{n=1}^H \frac{\partial y^c}{\partial A_{mn}^k}$ $a_{mn} = \text{RELU}\left(\sum_{k=1}^K a_k A_{mn}^k\right)$	计算模型预测类别与特征图之间的梯度信息生成一个热力图来可视化模型对不同特征的注意力分布(Selvaraju等,2020)
	DeepSHAP	$[h_x(x')]_d = \begin{cases} x_d & \text{if } x'_d = 1 \\ \mathbb{E}(x_d) & \text{if } x'_d = 0 \end{cases}$ $\mathcal{F}(h_x(x')) \approx \phi_0 + \sum_{d=1}^D \phi_0 x'_d$	结合了DeepLIFT和SHAP的原则,通过生成样本、构建参考分布和计算预测差异,最终计算每个特征的Shapley值(Lundberg和Lee,2017)
	LRP	$\forall_k: z_k = \epsilon + \sum_{0,j} a_j \cdot \rho(w_{j,k})$ $\forall_k: s_k = R_k / z_k$ $\forall_j: c_j = \sum_k \rho(w_{j,k}) \cdot s_k$ $\forall_j: R_j = a_j c_j$	将网络的输出得分 $f(x)$ 作为相关性得分逐层反向传播,最终将 $f(x)$ 分配到输入样本 x 的各个维度上,通过定义每个像素点的贡献来衡量像素级特征的重要性(Bach等,2015)
	SmoothGrad	$S_C(x) = \frac{\partial F_C(x)}{\partial x}$	通过引入高斯噪声,在局部邻域内多次采样求平均值来消除噪点,得到平滑锐化的效果(Smilkov等,2017)
	Occlusion	$g(z h_v) = \mathbb{E}_{p(h_i h_v)}[z(h)] = \sum_{h_i} p(h_i h_v) z(h_v, h_i)$	通过在输入数据中逐个遮挡不同区域,然后观察模型对遮挡后数据的响应来分析模型对不同特征的敏感性
外在归因方法	LIME	$\xi(x) = \sum_{z', z \in Z} \pi_x(z) (f(z) - g(z'))^2$	在输入空间中生成随机样本,对这些样本进行模型预测,然后拟合一个解释性好的本地模型,最终解释模型对于特定实例的决策原因(Ribeiro等,2016)
	RISE	$S_{l,f}(\lambda) \approx \frac{1}{\mathbb{E}[\mathbf{M}] \cdot N} \sum_{i=1}^N f(I \odot \mathbf{M}_i) \cdot \mathbf{M}_i(\lambda)$	随机生成图像子集,将其输入模型并记录输出,通过对比不同子集的输出,确定图像中哪些区域对于模型的预测最为关键(Petsiuk等,2018)

在内在归因方法中，由于Sigmoid函数的饱和问题或ReLU函数的零梯度区域，深度神经网络的梯度可能会有噪声，并存在梯度消失的问题，导致输出关于输入的梯度或内部层激活在视觉上存在噪音（Wang等，2020）。CAM/Grad-CAM将全局平均池化或最后一个卷积层的梯度信息映射回原图时会导致显著图丢失细节信息，难以生成较为清晰的特征边缘。即当卷积层最后一层得到7×7像素的特征图时，若映射回224×224像素的输入

图像，其显示区域将以32个像素为单位。由于细粒度分类任务主要关注于相似目标中的微小细节，这种归因方法不能精确显示对分类影响最大的本质特征区域，无法对细粒度分类的结果给出清晰明确的解释。

而外在归因方法，也即基于遮挡的扰动方法，普遍具有噪声和尺度问题。该方法随机生成遮挡图并利用其重要性加权以生成显著图，由于未对遮挡图中有效遮挡和无效噪声进行区分，因此，

会错误地赋予一些无关像素较高的归因得分, 导致生成的显著图中存在噪声。部分方法设计小尺度的掩膜图像(或遮挡窗口)并将其缩放成原始图像的大小进行遮挡, 通过遮挡输入图像的不同部分, 观察模型对遮挡的响应可以推断出模型对不同特征的重要性。在尺度缩放的过程中有模糊细粒度特征边缘的可能。

2.3 目标细粒度分类任务的博弈竞争数学模型

目前仍缺乏统一的数学框架对目标细粒度分类任务进行建模。为了在数学理论上定义目标细粒度分类任务所面对的问题, 指导可解释性分析方法的理论设计和优化, 本文从目标内部的特征博弈这一角度对目标细粒度分类任务进行数学建模。

在细粒度任务的目标级博弈中, 需要寻找少量的、适应于特定目标细粒度分类任务的专用特征, 这些可分性特征在博弈中一定被赋予最大贡献。因此, 细粒度分类任务 $f_{\text{fine-grained}}$ 的决策过程可定义如下:

$$f_{\text{fine-grained}}(\Psi) = \sum_{i=1}^m S_i + \prod_{j=1}^n P_j = \sum_{i=1}^m S_i^{\text{ess}} + \delta + \prod_{j=1}^n P_j^{\text{ess}} + \gamma \quad (1)$$

式中, Ψ 为目标级的内部特征集合, S 为 Ψ 内部的独立特征子集, P 为 Ψ 内部的交互特征子集, δ 为独立特征余项, γ 为交互特征余项。针对目标细粒度分类任务, 需要关注独立本质特征集 S_i^{ess} 和交互本质特征集 P_j^{ess} , 当其被分配了最大贡献值时使得该目标细粒度分类问题有解。

2.4 本质特征的特征博弈作用

为研究目标细粒度分类问题中本质特征的特征博弈作用, 首先, 基于像素级的特征关系, 可以将细粒度分类网络的映射过程按照泰勒级数展开如下:

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \left(\frac{\partial f}{\partial \mathbf{x}_0} \right)^T (\mathbf{x} - \mathbf{x}_0) + o((\mathbf{x} - \mathbf{x}_0)^T (\mathbf{x} - \mathbf{x}_0)) \quad (2)$$

式中, f 为网络模型, $\mathbf{x}$ 为输入图像, $\mathbf{x}_0$ 为 $\mathbf{x}$ 局部邻域内的参考点。

其次, 基于局部特征关系, 可以将细粒度分类网络的映射过程如下式展开:

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \frac{1}{i} \left(\frac{\partial^i f}{\partial^i \mathbf{x}_0} \right)^T (\mathbf{x} - \mathbf{x}_0)^i \quad (3)$$

根据式(2)和(3)的2种展开方式, 对于输入网络的任意特征, 可以把它对网络输出产生的影响归纳为独立效应和相互作用效应(Deng等, 2023)。由于独立效应和相互作用效应都可以量化为特定的泰勒相互作用, 所以网络的输出 $f(\mathbf{x})$ 可以分解为不同输入特征 p 的一般独立效应 $\mathcal{I}(p)$ 和输入特征集 Ψ 某一子集 Ω 的一般相互作用效应 $\mathcal{T}(\Omega)$ 之和。其中, $\Omega = \{\Omega_1, \Omega_2, \dots, \Omega_n\}$ 。因此, a_p 表示特征 p 的归因得分为

$$a_p = \sum_{q \in \Omega} w_{p,q} \mathcal{I}(q) + \sum_{\Omega_i \subseteq \Omega_p \in \Omega_i} w_{p,\Omega_i} \mathcal{T}(\Omega_i) \quad (4)$$

式中, $w_{p,q}$ 为 q 的一般独立效应 $\mathcal{I}(q)$ 分配给特征 p 的比率, w_{p,Ω_i} 为特征集 Ω 中特征之间的一般相互作用效应 $\mathcal{T}(\Omega)$ 分配给特征 p 的比率。

2.5 Hybrid-Grid 算法流程

基于目标细粒度分类的目标级博弈竞争数学模型(式(4)), 本文提出了一种满足泰勒展开框架的可解释性分析方法 Hybrid-Grid, 该方法融合了像素级特征关系和图像局部特征关系, 可以有效地提高目标细粒度分类任务的解释精度。将 Hybrid-Grid 扩展为特征博弈作用的重新分配, 从交互的角度将混合归因重写为独立效应和交互效应的加权和, 计算公式如下:

$$a_p = \sum_{\theta \in \Omega_p, Z_i \subseteq Z_n} \sum_{\vartheta \in \Psi/Z_i} \text{Ind}(\vartheta) + \sum_{\theta \in \Omega_p, Z_i \subseteq Z_n} \sum_{\Phi \in \Psi/Z_i} \text{Int}(\Phi) \quad (5)$$

式中, $\text{Ind}(\cdot)$ 为将权重比率整合入独立效应的计算函数, $\text{Int}(\cdot)$ 为将权重比率整合入相互作用效应的计算函数, θ 表示 p 所在特征集 Ω_p 中的某一像素单元, ϑ 代表泰勒展开项中的度向量, 遮挡窗口 $M = \{\Omega_1, \Omega_2, \dots, \Omega_m\}$, Z_i 代表经第 i 个窗口遮挡后参与交互的特征度向量集。其中, $\vartheta \in \mathbb{R}^{1 \times n}$, 对于 n 个类别的细粒度判断问题, 当目标归属于第 i 个类别时, $\vartheta_i=1$; 否则, $\vartheta_i=0$ 。

Hybrid-Grid 方法改进了现有外在归因方法中将特征集中所有的特征都分配独立效应的问题, 避免引入不涉及分类器判断的无关特征, 从理论上可以减少噪声; 同时, 还解决了现有内在归因方法中仅将特定神经元的独立效应分配给归因的模式问题, 理论上可以提高对特征间相互作用的精准刻画, 提高归因结果的精确度和准确性。

本文提出的 Hybrid-Grid 可解释性分析方法的结构如图1所示, 具体流程如下: 首先将类激活函数 $F_c(x)$ 对输入图像 x 求偏导得到灵敏图 $S_c(x)$, 外

在像素级特征提取算法对灵敏图进行遮挡后采样, 对每个采样图像的结果灵敏度图取平均值, 而后通过阈值 φ_1 进行限定, 其公式为

$$T_1(x) = \frac{1}{n} \sum_{c=1}^n S_c(x + \mathcal{N}(0, \sigma^2)) - \varphi_1 \quad (6)$$

然后使用内在局部特征提取算法对图像分类结果预测类别进行反向传播, 利用特征层上的梯度信息计算特征图每个通道信息 A^k 的重要程度 α_k^c , 根据重要程度对特征层每个通道进行加权求和, 而后通过阈值 φ_2 进行限定:

$$T_2(x) = \text{RELU}\left(\sum_k \alpha_k^c A^k\right) - \varphi_2 \quad (7)$$

$$T'(i, j) = \begin{cases} 1, & \text{if } \sum_i \sum_j \left(T\left(i - \frac{\text{grid}}{2} : i + \frac{\text{grid}}{2}\right), \left(j - \frac{\text{grid}}{2} : j + \frac{\text{grid}}{2}\right) \right) \geq \text{grid}^2 \cdot 0.8; \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

$$E_h(x) = \sum_i \sum_j T'(i, j) \quad (10)$$

基于网格的特征融合解释算法通过指定区域尺度控制了生成显著图的细度, 可以设置特征融合解释算法的网格尺度为输入图像尺度的任意因子, 以达到自适应遥感图像尺幅的目的。以输入图像尺度为 224×224 为例, 网格尺度可以设置为 2×2 ,

由于内在局部特征是基于梯度构建, 而深度模型的复杂度和高度非线性导致梯度存在梯度噪声、梯度饱和、连续性较差等问题, 本文提出了基于网格的特征融合解释算法, 在特征融合过程中经阈值 φ 过滤保留、利用网格法聚合本质特征, 最终输出特征显著图 $E_h(x)$, 该过程如式(8)—(10)所示。在其遍历过程中, 有部分像素点被丢弃, 部分像素点被补充, 像素点总数基本保持不变。背景中的分散像素点被忽略, 突出了前景的目标特征区域, 保证了显著图的简洁性和有效性。

$$T(i, j) = T_1(i, j) + T_2(i, j) - \varphi \quad (8)$$

$4 \times 4, 7 \times 7, 8 \times 8, 14 \times 14, 16 \times 16, \dots$, 当网格尺度为 2×2 时, 本文提出的遥感图像细粒度分类可解释方法可以在显著图中指示细度为输入图像尺度0.9%的特征信息, 因此, 该方法生成的显著图具有清晰的特征边缘, 能够保证对目标细粒度分类任务的高精度解释。

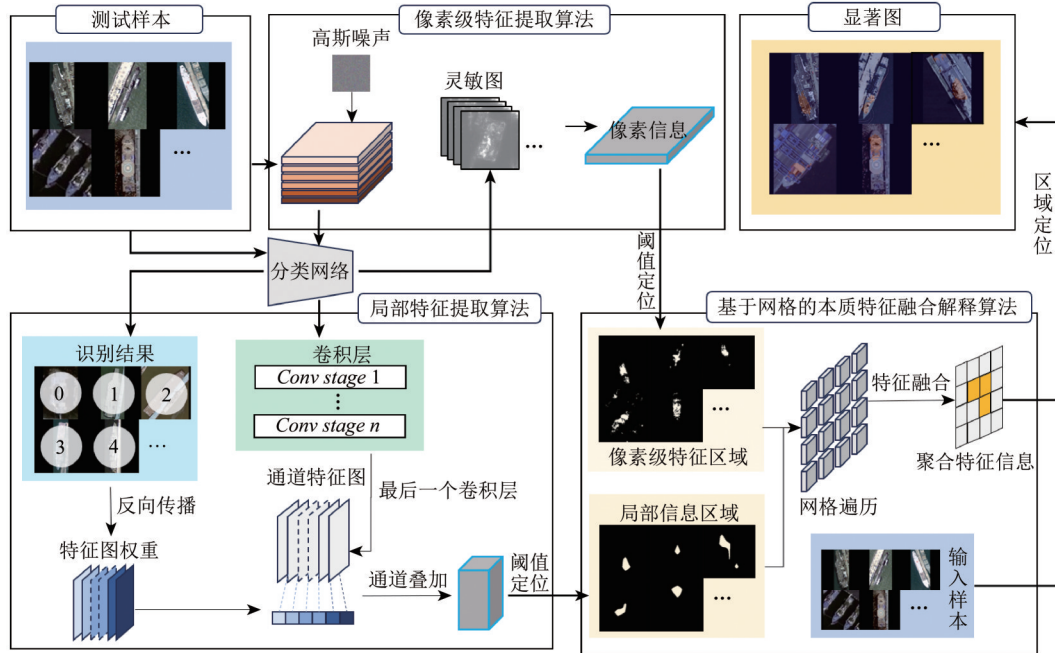


图1 遥感图像目标细粒度分类可解释方法 Hybrid-Grid 结构示意图

Fig. 1 Structure diagram of grid based explainable analysis method for essential features

3 实验设计

3.1 实验数据集

FGSC-23 数据集共 23 类细粒度船舶类的样本,

包括塔川类两栖攻击船舶、两栖攻击船舶等船舶类别共 22 类, 以及非船舶类别 1 类。Aircraft-16 数据集共 16 类细粒度飞机样本, 包括云遮挡、小目标和不平衡样本等困难情况。在 FGSC-23、Aircraft-16 训练集上分别训练了 EFM-Net、PMG-

V2、Resnet18,并在测试集上验证分类网络的解释情况,为展示测试集样本分布均衡度,FGSC-23、Aircraft-16训练集的类别和样本数如表2所示。

表2 数据测试集样本分布数量

Table 2 Number of sample distributions in the data testing set

FGSC-23				Aircraft-16					
类型	数量	类型	数量	类型	数量	类型	数量	类型	数量
非船舶	97	两栖攻击舰	32	气垫船	24	类0	40	类8	2
航空母舰	32	指挥船	10	散货船	69	类1	196	类9	8
驱逐舰	107	潜艇	48	油轮	33	类2	130	类10	54
登陆舰	22	医疗船	10	渔船	20	类3	67	类11	748
护卫舰	59	战斗艇	29	客船	18	类4	155	类12	100
两栖船坞运输舰	18	辅助船	45	液化气船	20	类5	87	类13	136
巡洋舰	59	集装箱船	20	驳船	11	类6	141	类14	115
两栖攻击舰-T	18	汽车运输船	14			类7	110	类15	44

3.2 实验平台

本文实验操作平台为 Ubuntu 18.04 操作系统,使用 CUDA11.3 和 cuDNN8 加速训练,处理器为 Intel® Xeon® Silver 4214R CPU@2.40 GHz, GPU 为 NVIDIA GeForce RTX 3090 (24 G 显存)。使用的编程语言为 Python,开发框架为 Pytorch。其中,实验代码为:PMG-V2, <https://github.com/PRIS-CV/PMG-V2>[2023-06-30]; EFM-Net, <https://github.com/JACYI/EFM-Net-Pytorch>[2023-06-30]。

3.3 评价指标

为定量评价可解释性分析方法的性能,采用平均下降 (Average Drop)、一致性 (Coherency)、复杂度 (Complexity)、ADCC (Average DCC)、删除和精度损失 (Deletion and precision loss) 作为可解释性分析方法的评价指标 (Poppi 等, 2021)。

(1) 平均下降 (Average Drop)。该评价指标衡量了当模型仅看到解释图而不是完整图像时,目标类别 c 的置信度平均下降的百分比。其计算公式为

$$\text{AverageDrop}(x) = \left(\frac{\max(0, y_c(x) - o_c(x))}{y_c(x)} \right) \times 100\% \quad (11)$$

(2) 一致性 (Coherency)。该评价指标用于衡量可解释性分析方法的显著图在解释预测时是否包含所有相关特征并以连贯的方式去除无用特征。因此,对于给定的输入图像 x 和感兴趣的类别 c , x

的显著图在以显著图本身为条件时不应发生变化。其计算公式为

$$E_c(x \odot E_c(x)) = E_c(x) \quad (12)$$

$$\text{Coherency}(x) = \frac{\text{Cov}(E_c(x \odot E_c(x)), E_c(x))}{\sigma E_c(x \odot E_c(x)) \sigma E_c(x)} \quad (13)$$

(3) 复杂度 (Complexity), 该评价指标要求显著图尽可能简单,即它必须包含最少的像素来解释预测结果。当归因方法仅突出显示少量像素时,复杂度最小化。这个评价指标有助于评估可解释性分析方法的简洁性,即使用最少的像素来解释预测结果。其计算公式为

$$\text{Complexity}(x) = \|E_c(x)\|_1 \quad (14)$$

(4) ADCC (Average DCC), 该评价指标提出于论文《重新评估用于解释性的类激活映射:一种新的评估指标和实验分析》(Poppi 等, 2021),用于对可解释性分析方法进行解释性评估。ADCC 综合考虑了平均减少、一致性和复杂度等多个方面,从而提供了对可解释性分析方法可解释性的全面和公正的评估。其计算公式为

$$\text{ADCC}(x) = 3 \left(\frac{1}{\text{Coherency}(x)} + \frac{1}{1 - \text{Complexity}(x)} + \frac{1}{1 - \text{AverageDrop}(x)} \right)^{-1} \quad (15)$$

(5) 删除和精度损失 (Deletion and precision loss), 该评价指标衡量在从图像中移除重要特征 (由可解释性分析方法提供) 后,分类器精度的下

降程度。该指标的数值越大,表示删除的特征对分类器的预测精度影响越大,反映了模型对于关键特征的依赖程度。其计算公式为

$$\text{Precision}(x) = \frac{\text{Acc}_{\text{full}} - \text{Acc}_{\text{excl}}}{\text{Acc}_{\text{full}}} \times 100\% \quad (16)$$

4 结果与分析

4.1 目标细粒度分类网络的可解释性分析效果评估

为了对本文提出的遥感图像细粒度分类可解释方法进行效果评估,使用现有的主流归因分析法和本文所提方法 Hybrid-Grid 对 EFM-Net 的全局决策进行解释,包括 Smooth-Grad、Deep-Lift 和 Integrated-Gradients,其解释效果如图 2(b)~(d) 所示。可见: Deep-Lift 和 Integrated-Gradients 生成的显著图存在噪声和解释效果不佳的问题。其中:由于 Integrated-Gradients 通过引入与输入相似的多个样本来计算梯度,深度模型内部对这些尺度缩放图像的处理可能存在很大的差异,因此会引入虚假相关性 (Yang 等, 2023),导致生成的显著图

错误地强调或忽视一些特征,不能够抓住细粒度分类网络判断依据的本质特征;由于 Deep-Lift 方法涉及对参考激活的选择,参考激活的不准确或不合适会导致无法捕捉到神经元间复杂的相互作用和依赖关系,此外,梯度传播受限也会导致信息传递不完整等问题,因此解释效果欠佳。与上述两种方法相比相比,Smooth-Grad 和 Grad-CAM 得出较为符合目视解译判断依据的显著图。其中: Smooth-Grad 可以生成更清晰、细节较多的显著性图,但会错误地指示飞机的影子等并不属于目标细粒度分类本质特征的干扰信息; Grad-CAM 可以关注到对全局决策更重要的块状区域,然而由于全局池化操作的使用,Grad-CAM 会丢失一些细节区域的信息,同时,受限目标细粒度分类网络最后一层卷积层的尺度问题,将热力图映射回原图时导致显著图难以生成较为清晰的特征边缘。综上,目前现有的可解释性分析方法都不能对目标细粒度分类网络的决策过程进行可信赖、稳定的解释,并给出符合人类感知的本质特征表示。

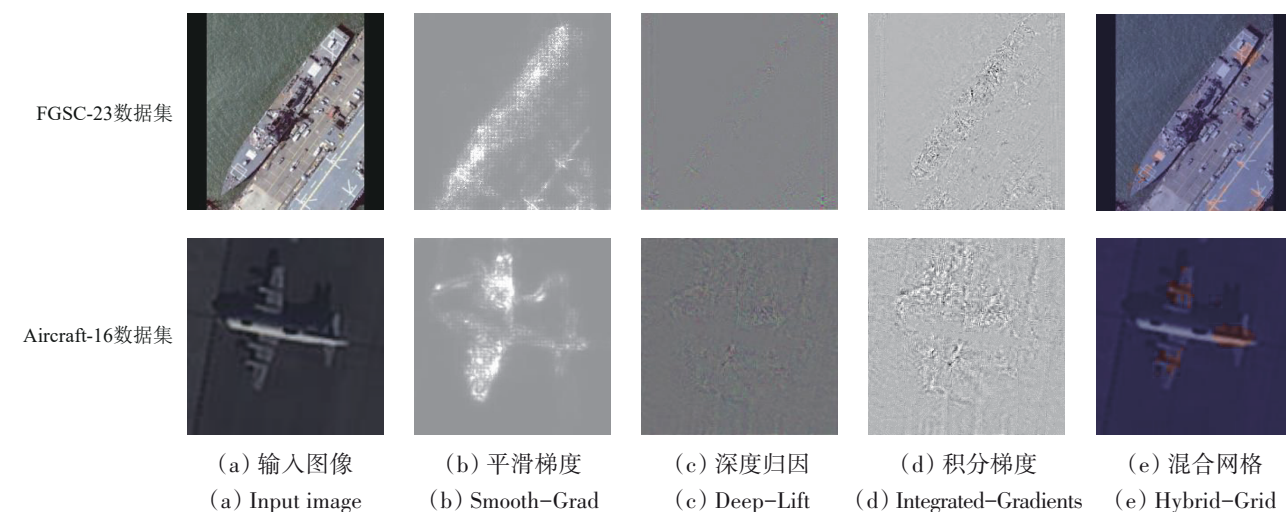


图2 不同解释方法对 EFM-Net 的解释效果对比(红色响应为网络在分类任务中抓取到的关键特征)

Fig. 2 Comparison of the explanation effect of different explanation methods on EFM-Net (the red area represents the key features captured by the classification network)

本文提出的遥感图像细粒度分类可解释方法——Hybrid-Grid 对于以 EFM-Net 为例的遥感图像目标细粒度分类网络具有更好的解释效果,详见图 2 (e)。红色响应为网络在分类任务中抓取到的关键特征。可见目标细粒度分类网络在决策过程中对目标的关注区域具有一致性。具体而言,

对于船舶细粒度分类任务,以 FGSC-23 数据集为例,网络主要依赖于船舶的停机坪、舰岛、相控阵雷达等区域进行判断。类似地,在飞机目标细粒度分类任务中,以 Aircraft-16 数据集为例,网络主要依赖于飞机的机翼、发动机等区域进行判断。Hybrid-Grid 可以确定目标细粒度分类网络在分类

过程中关注的特定区域包含了目标的本质特征,从而深入理解网络的决策过程。

从视觉连贯性和判别性的角度来看 (Simonyan 等, 2014; Zeiler 和 Fergus, 2014; Springenberg 等, 2014; Selvaraju 等, 2016; Sundararajan 等, 2017), Hybrid-Grid 在解释区域的定位上表现更准确, 其指示区域在目标细粒度分类网络感兴趣的细粒度特征上, 同时不易受到影子等干扰因素的影响。从视觉感知角度来看, Hybrid-Grid 给出的显著图具有较小的噪声, 在解释具体目标的分类时保持了归因分析法所具有的像素级信息和细节性, 相对于类激活图映射系列方法具有更清晰的特征边缘。另外, Hybrid-Grid 的显著图具有较好的像素连贯性, 能够实现以区域响应的形式定位目标物体, 显示分类器决策依赖的本质特征的区域。综上所述, 本文提出的遥感图像细粒度分类可解释方法同时具有先进的原理和较好的视觉解释效果, 能给出对目标细粒度分类网络更准确和直观的解释结果。

4.2 基于网格的特征融合解释算法参数

Hybrid-Grid 的先进性之一体现于它对像素级特征和高层特征的良好平衡, 通过调节基于网格的特征融合解释算法的参数, 一方面可以达到减少噪声、平滑波动、聚合特征的效果, 另一方面可以控制生成显著图的细度, 使其在保持对目标细粒度分类网络高精度解释效果的情况下适应不同尺度的遥感图像。以 224×224 的输入图像尺度为例, 可以设置特征融合解释算法的网格尺度为输入图像尺度的任意因子。图 3 为当网格尺度参数设置为不同值时, 所提方法 Hybrid-grid 对参考分类网络在 FGSC-23 数据集上进行决策的可解释分析示例。可见: 从左到右特征融合解释算法的网格尺度逐渐减小, 并且随着该参数的变化, Hybrid-Grid 生成的显著图细节性越来越好; 对目标细粒度分类网络决策依赖的本质特征的边界描述得更加清晰, 而特征像素间的聚合性越来越差, 对目标决策重要区域响应的定位变得分散。



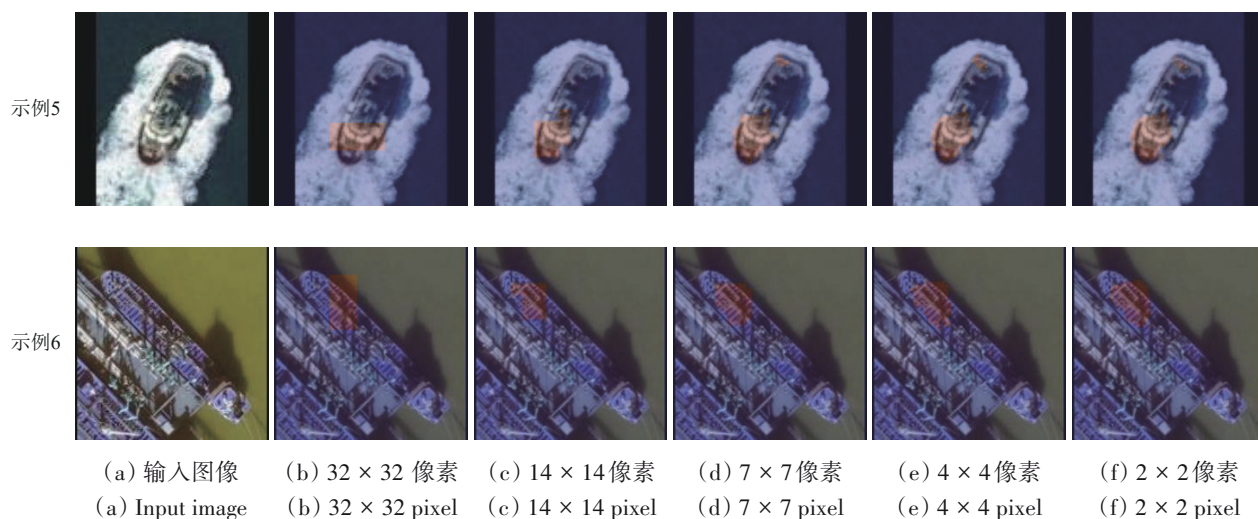


图3 不同网格尺寸参数设置下的Hybrid-Grid可解释性分析方法对EFM-Net的解释效果

Fig. 3 The explanatory effect of different size grids on EFM-Net

为了进一步论证上述结论,在表3中,通过本文提出的Hybrid-Grid所指示的本质特征区域,根据之前所提的可解释性分析方法量化评价指标,以EFM-Net在FGSC-23数据集上的分类任务为例,计算遮挡Hybrid-Grid指示区域造成的删除和精度损失,考察Hybrid-Grid可解释性分析方法能否抓住目标细粒度分类决策依据的本质特征,探究特征融合解释算法的网格参数对其产生的影响。可见:遮挡Hybrid-Grid指示的本质特征所在区域可以使目标细粒度分类网络的Top-1准确率、Top-5准确率和F1指标产生大幅下降;同时,随着掩膜尺寸参数的减小,删除和精度损失指标呈现出一种特殊的变化模式。

表3 掩膜尺寸对目标细粒度分类网络精度造成的影响
Table 3 Influence of different mask sizes on the accuracy of target fine-grained classful network

掩膜尺寸/像素	Top-1准确率 损失/%	Top-5准确率 损失/%	F1损失/%
32×32	4.40	0.14	4.59
14×14	4.40	0.01	4.43
7×7	8.45	0.87	8.65
4×4	11.88	1.23	12.16
2×2	16.92	1.61	17.21

具体来说,在调整掩膜尺寸从 32×32 到 2×2 的过程中,删除和精度损失指标,即Hybrid-Grid对目标细粒度分类网络的解释性能,在掩膜尺寸减小的初期阶段有所下降。这是因为较大的网格

尺寸使Hybrid-Grid的特征融合解释算法的定位区域较广,过大的掩膜遮挡面积使得目标细粒度分类网络的准确率大幅下降。但与图3所展示的可视化效果结合来看,较大的网格尺寸可能会导致特征融合解释算法不能良好地融合像素级特征信息和高层特征信息,使Hybrid-Grid无法准确地定位细粒度的本质特征信息,导致可解释性分析方法的性能下降。因此,应进一步减小特征融合解释算法的网格参数。

随着掩膜尺寸进一步减小,发现删除和精度损失指标持续上升,这是因为较小的网格尺寸能够使特征融合解释算法更好地捕捉到目标细粒度分类网络决策关注的细微特征,从而提高Hybrid-Grid的解释性能。当掩膜尺寸达到 2×2 时,Hybrid-Grid的删除和精度损失指标达到了最大值,这意味着在输入图像为 224×224 的情况下,当特征融合解释算法的网格尺寸为 2×2 时,Hybrid-Grid能够以最佳的方式解释目标细粒度分类网络的分类决策依据,并提供最准确、详细的本质特征定位。

4.3 量化指标评价可解释性分析方法

使用平均下降、连贯性、复杂度和ADCC指标对Hybrid-Grid及其他现有可解释性分析方法在EFM-Net、PMG-V2、Resnet18上进行量化指标评估,以定量衡量不同可解释性分析方法在目标细粒度分类网络上的性能。

对于普通分类网络来说,包括ResNet18、VGG16

在内, Samuele Poppi 等 (2021) 发现 CAM 系列的 Score-CAM 在 ADCC 指标上取得了最好的性能指标, 这表明 Score-CAM 方法能够有效地解释传统分类网络的决策过程, 因此在本节中, 使用 Score-CAM 作为分类网络的基准可解释性分析方法。

使用 Hybrid-Grid 对经典分类网络应用于细粒度分类网络任务进行解释, 表 4 为 Score-CAM 在解

释 ResNet-18 应用于 FGSC-23 测试集的评价指标情况。可见与在传统分类任务中相比, 这 4 个指标都产生了显著的下降。其中, ADCC 指标与 Samuele Poppi 提出的最优数值 77.30% 相比下降了 16.27%。这是因为 ResNet-18 在目标细粒度分类任务上的适应性不足, 无法有效地捕捉到重要的目标区域进行分类。

表 4 不同可解释性分析方法量化指标评估 Hybrid-Gird

Table 4 Evaluating Hybrid-Gird by different quantitative indicators for explainability

指标	ResNet18		PMG-V2		EFM-Net	
	Score-CAM	Hybrid-Grid	Score-CAM	Hybrid-Grid	Score-CAM	Hybrid-Grid
平均下降 ↓	28.12	34.03	82.77	47.87	82.94	38.24
一致性 ↑	54.71	93.68	12.76	94.61	74.10	86.45
复杂度 ↓	41.07	3.28	44.15	2.17	42.83	2.71
ADCC ↑	61.03	82.93	19.44	75.05	33.47	78.87

注: 加粗数值表示不同方法的最优数值, “↓”表示越低越好, “↑”表示越高越好。

为了评估 Hybrid-Grid 在解释目标细粒度分类网络方面的优势, 本文选择 EFM-Net 作为目标模型, 并使用上述量化指标来比较本文提出的可解释性分析方法 Hybrid-Grid 和基准方法 Score-CAM 对 EFM-Net 应用于 FGSC-23 测试集的解释效果 (表 4)。可见 Hybrid-Grid 在所有指标上都取得了最好的结果, 该指标证明其能够对提供一致且准确的解释; 去除冗余特征、突出本质特征; 其解释具有更高的可靠性和稳定性。与之相比, Score-CAM 虽然在解释传统分类任务时表现良好, 但在解释目标细粒度分类网络时存在一定的局限性, 这也证明了现有可解释性分析方法对目标细粒度分类网络关注的本质特征的解释能力有限。

为了验证 Hybrid-Grid 解释方法在其他目标细粒度分类网络上的通用性和有效性, 选择了 2021 年提出的 PMG-V2 网络作为测试网络, 使用 FGSC-23 作为测试集, 将 Hybrid-Grid 对 PMG-V2 网络的分类过程进行可解释性分析, 并与基准方法 Score-CAM 进行比较, 结果如表 4 所示。可见: Hybrid-Grid 对 PMG-V2 的解释效果在所有指标 (包括平均下降、一致性、复杂度、ADCC) 上均优于 Score-CAM, 并展现出优秀的数值。这表明对于目标细粒度分类网络 PMG-V2 来说, Hybrid-Grid 具有更加准确和稳定的解释能力。该实验证明了 Hybrid-Grid 不仅适用于之前提出的 EFM-Net 网络,

还可以应用于其他目标细粒度分类网络, 且在这些网络上也具备良好的解释性能。

综上所述, 量化指标结果证明了 EFM-Net 在目标细粒度分类任务上的先进性, 以及 Hybrid-Grid 在解释目标细粒度分类模型的通用性和优越性。

4.4 删除及精度损失实验

本文使用删除及精度损失指标对 Hybrid-Grid 及其他现有可解释性分析方法进行评估, 通过在原测试集以及根据不同可解释性分析方法使用掩膜遮挡了测试图像特定区域的测试集进行测试, 考察目标细粒度分类网络的精度损失情况。考察的精度指标包括 Top-1 准确率、Top-5 准确率和 F1。Top-1 准确率是指模型在对单个样本进行分类时, 预测的最高置信度类别与实际类别相符的比例; Top-5 准确率是指模型在对单个样本进行分类时, 在前 5 个最高置信度的预测类别中, 至少有一个与实际类别相符的比例; F1 综合考虑了模型的精确率 (预测为正例的样本中, 真正为正例的比例) 和召回率 (实际为正例的样本中, 被正确预测为正例的比例) 的指标, 可以评估模型的分类性能。预测结果详见表 5。可见以在 FGSC-23 数据集上训练后的 EFM-Net 为例, 遮挡 Hybrid-Gird 指示的重要区域后, 目标细粒度分类网络的 Top-1 准确率下降了 16.92%, Top-5 准确率下降了 1.61%,

F1下降了17.21%；通过与根据其他可解释性分析方法的遮挡结果进行比较，Hybrid-Grid指示的重要区域对Top-1、Top-5准确率和F1指标产生的影响高于其他两个可解释性分析方法。以Aircraft-16数据集为例，所选网络在Aircraft-16测试集上的Top-1准确率下降了12.18%，Top-5准确率下降了

2.1%，F1下降了11.95%。通过与其他可解释性分析方法结果进行比较，Hybrid-Grid可解释性分析方法指示的重要区域对Top-1、F1指标产生的影响高于其他两个可解释性分析方法，对Top-5准确率的影响略低于Grad-CAM。

表5 不同遮挡情况对EFM-Net性能指标的影响
Table 5 The impact of occlusion experiments on accuracy

遮挡情况	FGSC-23			Aircraft-16		
	Top-1 准确率	Top-5 准确率	F1	Top-1 准确率	Top-5 准确率	F1
原FGSC-23测试集	91.39	99.03	91.35	92.35	99.29	92.27
遮挡Grad-CAM指示区域	85.97	98.53	86.03	81.17	97.13	81.24
遮挡Smooth-Grad指示区域	79.51	97.92	79.35	83.66	98.21	83.73
遮挡Hybrid-Gird指示区域	74.47	97.42	74.14	80.17	97.19	80.22

因此，Hybrid-Grid抓取的特征确实是对目标细粒度分类网络决策过程最关键的本质特征，证实了Hybrid-Grid对目标细粒度分类网络的解释性能优于现有可解释性分析方法。

4.5 可解释性分析方法的适应性和可行性

在本研究中，对Hybrid-Gird在不同数据集、不同分类网络上的适应性和可行性进行了实验。本文选择了一个常用的分类网络即传统的ResNet18，以及两个具有代表性的目标细粒度分类网络即PMG-V2和EFM-Net，并将其应用于2个具有代表性的FGSC-23和Aircraft-16数据集，获得在这些数据集上的分类结果。后使用本文提出的可解释性分析方法Hybrid-Gird对这些分类网络的决策过程进行解释，通过该可解释性分析方法，能够可视化输入图像中对网络模型决策具有关键作用的本质特征，这些解释结果如图3所示。可见，从左到右特征融合解释算法的网格尺度逐渐减小，并且随着该参数的变化，Hybrid-Grid生成的显著图细节性越来越好；对目标细粒度分类网络决策依赖的本质特征的边界描述得更加清晰，而特征像素间的聚合性越来越差，对目标决策重要区域响应的定位变得分散。

同时，本文还使用了多种评价指标来量化评价Hybrid-Gird的性能。这些评价指标包括平均下降（Average Drop）、一致性（Coherency）、复杂度（Complexity）、ADCC（Average DCC）、删除和精度损失（Deletion and precision loss）等，用于评估

Hybrid-Gird在不同数据集和不同目标细粒度分类网络上的表现，以判断可解释性分析方法的适应性和可行性。

上述实验结果证明，在不同的数据集和分类网络中，Hybrid-Gird可以准确捕捉到输入图像中与目标分类密切相关的关键特征，并通过可视化结果展示出来。这表明Hybrid-Gird能够应用于不同分类网络、不同数据集进行解释。

5 结 论

本文通过研究目标细粒度分类任务问题的数学模型，并提出一种基于本质特征博弈数学框架及语义可解释的归因方法Hybrid-grid，探究了目标本质特征对网络决策过程的影响情况。主要结论如下：（1）高分辨率遥感图像内容复杂、特征丰富，现有常用的可解释性分析方法仅考虑目标部件特征之间的竞争关系，不能实现高精度、可视化的可解释性分析；（2）为表明现有目标细粒度分类网络与数据集之间存在的决策偏见和路径依赖，需要对遥感图像目标细粒度分类决策过程建立数学框架，精细刻画目标本质特征，准确描述目标可信识别任务中决策特征归因过程；（3）通过融合像素级特征关系与图像局部特征关系，可以进一步提升目标本质特征的可视化效果和解释准确度，辅助提升细粒度分类网络的分类性能与决策可信性；（4）通过设置和调整特征融合的网格参数，可以实现可解释性分析方法对遥感图像

尺度的自适应。

未来的工作将进一步提升目标本质特征的可视化效果和解释准确度, 辅助提升细粒度分类网络的分类性能与决策可信性。

参考文献 (References)

- Bach S, Binder A, Montavon G, Klauschen F, Müller K R and Samek W. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One*, 10(7): e0130140 [DOI: 10.1371/journal.pone.0130140]
- Cheng X, Chu C, Zheng Y, Ren J and Zhang Q S. 2021. A game-theoretic taxonomy of visual concepts in DNNs. *arXiv preprint arXiv: 2106.10938*
- Deng H, Zou N, Du M, Chen W, Feng G, Yang Z, Li Z and Zhang Q. 2023. Understanding and unifying fourteen attribution methods with taylor interactions. *arXiv preprint arXiv:2303.01506*.
- Du R, Xie J, Ma Z, Chang D, Song Y Z and Guo J. 2021. Progressive learning of category-consistent multi-granularity features for fine-grained visual classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9521-9535.
- Fu K, Dai W, Zhang Y, Wang Z R, Yan M L and Sun X. 2019. Multi-CAM: multiple class activation mapping for aircraft recognition in remote sensing images. *Remote Sensing*, 11(5): 544 [DOI: 10.3390/rs11050544]
- Gong J Y, Huan L X and Zheng X W. 2022. Deep learning interpretability analysis methods in image interpretation. *Acta Geodaetica et Cartographica Sinica*, 51(6): 873-884 (龚健雅, 宦麟茜, 郑先伟. 2022. 影像解译中的深度学习可解释性分析方法. *测绘学报*, 51(6): 873-884) [DOI: 10.11947/j.AGCS.2022.20220106]
- Han Y Q, Yang X Y, Pu T and Peng Z M. 2022. Fine-grained recognition for oriented ship against complex scenes in optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 5612318 [DOI: 10.1109/TGRS.2021.3123666]
- He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//*Proceedings of the IEEE conference on computer vision and pattern recognition: 770-778* [DOI: 10.1109/CVPR.2016.90]
- Liang W, Li J H, Diao W H, Sun X, Fu K and Wu Y R. 2020. FGATR-Net: automatic network architecture design for fine-grained aircraft type recognition in remote sensing images. *Remote Sensing*, 12(24): 4187 [DOI: 10.3390/rs12244187]
- Lundberg S M and Lee S I. 2017. A unified approach to interpreting model predictions//*Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach: Curran Associates, Inc.: 4768-4777
- Nie Y, Bian C J and Li L G. 2022. Adap-EMD: adaptive EMD for aircraft fine-grained classification in remote sensing. *IEEE Geoscience and Remote Sensing Letters*, 19: 6509005 [DOI: 10.1109/LGRS.2022.3168581]
- Petsiuk V, Das A and Saenko K. 2018. RISE: randomized input sampling for explanation of black-box models//*Proceedings of the British Machine Vision Conference 2018*. Newcastle: BMVA Press: 151
- Poppi S, Cornia M, Baraldi L and Cucchiara R. 2021. Revisiting the evaluation of class activation mapping for explainability: A novel metric and experimental analysis//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2299-2304*
- Ribeiro M T, Singh S and Guestrin C. 2016. Why should I trust you?": explaining the predictions of any classifier//*Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco: ACM: 1135-1144 [DOI: 10.1145/2939672.2939778]
- Chen S, Wang H, Xu F and Jin Y Q. 2016. Target classification using the deep convolutional networks for sar images. *IEEE transactions on geoscience and remote sensing*, 54(8): 4806-4817
- Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D and Batra D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization//*Proceedings of the IEEE international conference on computer vision: 618-626*
- Selvaraju R R, Das A, Vedantam R, Cogswell M, Parikh D and Batra D. 2016. Grad-CAM: Why did you say that?. *arXiv preprint arXiv: 1611.07450*
- Simonyan K, Vedaldi A and Zisserman A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv: 1312.6034*
- Smilkov D, Thorat N, Kim B, Viégas F and Wattenberg M. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv: 1706.03825*
- Springenberg J T, Dosovitskiy A, Brox T and Riedmiller M. 2014. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv: 1412.6806*
- Springenberg J T, Dosovitskiy A, Brox T and Riedmiller M. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv: 1412.6806*, 2014.
- Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks//*International conference on machine learning*. PMLR, 2017: 3319-3328.
- Wang H F, Wang Z F, Du M N, Yang F, Zhang Z J, Ding S R, Mardziel P and Hu X. 2020. Score-CAM: Score-weighted visual explanations for convolutional neural networks//*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 2020: 24-25 [DOI: arXiv:1910.01279v2]
- Xiong W, Xiong Z Y and Cui Y Q. 2022. An explainable attention network for fine-grained ship classification using remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 5620314 [DOI: 10.1109/TGRS.2022.3162195]
- Yang P B, Sang J T, Zhang B, Feng Y G and Yu J. 2023. Survey on interpretability of deep models for image classification. *Journal of Software*, 34(1): 230-254 (杨朋波, 桑基韬, 张彪, 冯耀功, 于剑.

2023. 面向图像分类的深度模型可解释性研究综述. 软件学报, 34(1): 230-254 [DOI: 10.13328/j.cnki.jos.006415]
Yi Y H, You Y N, Li C and Zhou W L. 2023. EFM-net: An essential feature mining network for target fine-grained classification in optical remote sensing images. IEEE Transactions on Geoscience

and Remote Sensing. [DOI: 10.1109/TGRS.2023.3265669]
Zeiler M D and Fergus R. 2014. Visualizing and understanding convolutional networks//Computer Vision-ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13. Springer International Publishing, 2014: 818-833

Hybrid-Grid: An explainable method for fine-grained classification of remote sensing images

ZHU Kaiwen¹, YOU Yanan¹, CAO Jingyi¹, MENG Gang², QIAO Yuanyuan¹, YANG Jie¹

1. School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China;

2. Beijing Institute of Remote Sensing Information, Beijing 100192, China

Abstract: At present, the deep neural network for targets fine-grained classification based on remote sensing images has been widely applied in military and civilian fields. The “black box” problem of deep learning network models still makes it difficult for people to understand the decision-making basis of the network in fine-grained classification tasks. This not only limits the possibility of deep neural networks optimizing and improving through feedback guidance, but also makes them unable to be fully trusted by humans and applied in important fields such as military and medical fields. Thus, how to carry out explainability research on its internal decision-making mechanism is the key problem for the current fine-grained algorithm to further improve the credibility of decision-making basis. This article summarized the commonly used explainable methods in existing fine-grained classification networks for remote sensing image targets. On this basis, a mathematical model for the problem of target fine-grained classification is established. It also models the fine-grained classification task of remote sensing image targets based on game competition theory, analyzes the applicability of explainable methods such as IG, SmoothGrad, and Grad CAM on the fine-grained classification network of remote sensing image targets, and proposes a scale adaptive method for analyzing the explainability of essential features of fine-grained classification of targets, that is, Hybrid Grid. The fusion algorithm of pixel level and local feature relationships is used to improve the accurate description ability of essential features of targets that support network decision-making. The typical explaining methods involved in this article were evaluated using indicators such as Average Drop, Coherence, Complexity, ADCC, and Deletion and precision loss, confirming the accuracy of the proposed method. The experimental results show that the Hybrid-Grid proposed in this paper achieves a score of 78.87 on the quantitative evaluation index of ADCC for target fine-grained classification networks, which is significantly improved compared to Score-CAM. Compared with the explanation results of SmoothGrad and Grad-CAM, our method performed the best in deletion and accuracy loss experiments, resulting in a loss of 16.92%, 1.61%, and 17.21% in the Top-1 accuracy, Top-5 accuracy, and F1 score of EFM Net, respectively. This proves that the Hybrid Grid accurately explains the essential feature of the target that contributes the most to fine-grained classification network decision-making. Based on several typical fine-grained classification networks, the feasibility and universality of the explaining method proposed in this paper have been confirmed. The explaining method proposed in this article can more accurately reveal the decision feature basis of the current target fine-grained classification network. Future work will further enhance the visualization and interpretation accuracy of the essential features of the target, assisting in improving the classification performance and decision-making credibility of fine-grained classification networks.

Key words: remote sensing image, explainability analysis method, target fine-grained classification network, explainable artificial intelligence, cooperative game theory, essential feature

Supported by National Natural Science Foundation of China (No. 62101060)