

基于策略梯度的黑盒模型显著性解释方法

果宇涵, 丁立中, 韦秋宁

(北京理工大学计算机学院, 北京 100081)

摘要: 作为可信人工智能的核心支撑技术, 机器学习模型的可解释性在自动化决策系统中扮演重要角色, 其中显著性解释能够以热力图的形式直观揭示模型决策依据, 已成为可解释人工智能的关键研究方向。然而, 现有方法存在根本性缺陷: 基于扰动的方法受限于人工设计的分布假设, 计算效率与泛化能力难以兼顾; 基于梯度的方法需要预设模型结构, 无法解释黑盒模型; 基于线性近似的方法因局部过度简化引入模型偏差。本文提出一种基于策略梯度的黑盒模型显著性解释方法 Polixplain, 将 Deletion Score 作为核心奖励信号, 通过可导松弛技术嵌入策略梯度优化流程, 实现忠实性目标的端到端优化。采用图像分类任务作为验证载体, 其高维特征空间与可视化直观性为解释方法的多维度评估提供了理想测试环境。Polixplain 方法在解释忠实性以及热力图的人类可理解性方面均超越了多种现有解释方法。

关键词: 可解释性; 黑盒解释; 显著性图; 强化学习; 策略梯度

中图分类号: TP391

文献标志码: A

引用格式: 果宇涵, 丁立中, 韦秋宁. 基于策略梯度的黑盒模型显著性解释方法[J]. 数据采集与处理, 2026, 41(5): 1366-1377. GUO Yuhan, DING Lizhong, WEI Qiuning. Policy gradient-based saliency explanation method for black-box models[J]. Journal of Data Acquisition and Processing, 2026, 41(5): 1366-1377.

引言

随着深度学习模型在医疗诊断、金融风控等高敏感性领域的广泛应用, 其黑箱特性引发的信任赤字与伦理风险已成为制约人工智能规模化应用的关键瓶颈^[1-3]。可解释性研究通过揭示模型决策逻辑与归因依据, 成为弥合技术能力与社会接受度之间鸿沟的核心技术路径^[4-7]。其中, 显著性解释因其直观可读备受关^[8-10]注: 通过生成与输入特征同维度的重要性图谱, 定位模型预测依赖的关键区域, 为用户提供模型关注局部特征的直接证据, 如图像中的关键区域^[11-13]或文本中的关键词^[14-16]。相较于全局特征重要性排序等抽象解释形式, 显著性解释通过空间语义一致性与可视化直观性, 更符合人类认知逻辑, 其在临床辅助诊断、自动驾驶决策等高风险场景中展现出独特的应用价值^[17-19]。

当前主流显著性解释方法可分为3类。SHAP^[20]、RISE^[11]等基于扰动的方法通过蒙特卡洛采样评估特征重要性, 但人工设计的扰动分布导致计算复杂度随特征维度呈指数级增长, 且高频噪声易引发伪影干扰; GradCAM^[21]、Gradient Input^[22]等基于梯度的方法依赖反向传播计算梯度幅值, 其本质受限于白盒假设, 且梯度饱和现象可能导致非关键区域的虚假激活; LIME^[23]、C-LIME^[24]等基于线性近似的方法通过线性代理模型拟合局部决策边界, 但模型简化假设在非线性特征交互场景下引入偏差, 且对超参数的高度敏感性导致解释结果不稳定。上述方法的共性缺陷在于无法直接优化忠实性指标, 导

致解释结果与模型真实决策逻辑存在语义鸿沟,同时缺乏对可读性约束的显式建模,显著性解释易出现空间碎片化或过度平滑现象。

为量化解释方法的忠实性,Petsiuk等^[11]提出显著性解释的忠实性评估指标 Deletion Score,通过降序遮蔽显著性解释高重要性区域并观测模型预测稳定性,直接反映特征对模型决策的实际贡献。然而,现有方法难以直接以 Deletion Score 为优化目标生成解释,其核心挑战源于高维特征空间的组合爆炸、Deletion Score 自身的不可微性质以及多目标动态平衡需求^[5]。针对热力图生成提出基于策略梯度的黑盒模型显著性解释方法 Polixplain,核心贡献包括:(1)提出 Deletion Score 的形式化表达,通过可导松弛技术设计了一种可微近似掩码,并提出 Deletion Score 的可微近似以解决 Deletion Score 离散性障碍,为显著性解释提供了针对忠实性的优化目标;(2)首次将显著性解释优化任务建模为奖励函数可扩展的强化学习问题,并使用策略梯度方法求解,在最小化 Deletion Score 的同时平衡可读性约束,为显著性解释提供了兼具忠实性与可读性的技术路径;(3)设计解耦的策略网络架构,支持多种任务与待解释模型,其自监督学习机制消除了对人工标注与先验假设的依赖。

在图像分类任务^[25]上的对比实验报告了 Polixplain 和多种对比基线方法的平均 Deletion Score,体现了 Polixplain 的解释更具忠实性。在手写数字识别任务^[26]和图像分类任务上的可视化分析实验表明,Polixplain 生成的显著性解释具有良好的可读性,高亮区域能够对应完整的语义特征。此外,对运行时间和超参数鲁棒性的分析实验展示了提出方法的可靠性。Polixplain 生成的解释可通过高精度归因为模型开发者提供决策归因依据,并为医疗和金融从业者验证高风险场景下模型的可靠性提供支持。

1 相关工作

1.1 基于扰动的解释

基于扰动的方法通过系统性地遮蔽输入特征并观测模型输出的变化,量化特征重要性。典型方法如 SHAP^[20]基于博弈论中的 Shapley 值概念,通过计算特征在所有特征子集组合中的边际贡献,生成全局或局部解释。RISE^[11]则通过随机生成输入掩码并评估对应预测概率,以蒙特卡洛积分逼近特征重要性分布。此类方法的优势在于无需访问模型参数,且理论基础扎实,如 SHAP^[20]满足对称性、可加性等公理。然而,基于扰动的方法需生成大量扰动样本,计算复杂度高,如 RISE 需数千次前向传播;扰动分布设计主观性强,如 SHAP^[20]假设特征独立性,RISE^[11]依赖高斯噪声,易引入分布偏移偏差;对高频噪声敏感,遮蔽非关键高频特征可能误判其重要性。

1.2 基于梯度的解释

基于梯度的方法需预先了解模型结构,通过反向传播计算输入特征的梯度幅值或方向,直接关联特征变化与模型输出。Grad-CAM^[21]结合卷积层梯度与全局平均池化,生成类激活热力图;Gradient input^[22]则将梯度与原始输入相乘以增强重要性定位。此类方法仅需一次反向传播,计算高效,且与深度网络无缝兼容。但其依赖模型可微性假设,无法直接应用于树模型、图神经网络等非可微架构;梯度饱和现象导致显著性解释失真,在饱和激活区域梯度趋近于零,引发虚假高亮或低亮;对初始化与超参数敏感,Grad-CAM^[21]的卷积层选择直接影响解释结果。

1.3 基于线性近似的解释

基于线性近似的方法假设复杂模型在局部区域可由线性模型近似,拟合一个低复杂度的内在可解释代理模型。LIME^[23]通过在输入数据邻域内采样扰动样本,训练稀疏线性模型拟合预测结果,以线性系数作为显著性解释。C-LIME^[24]进一步引入因果约束,增强解释稳定性。此类方法的优势在于直观性强,可以解释非线性模型的局部决策逻辑。但局部近似精度受限于采样策略,欠采样会导致无法准

确拟合模型局部动作,过采样则会引入噪声;线性模型无法捕提高阶交互规则,对复杂非线性关系解释能力不足;解释结果对文本替换策略或图像遮蔽形状等样本扰动方式高度敏感,导致结果不稳定。

上述方法均未直接以解释忠实性为优化目标,而是通过梯度幅值、Shapley值和局部近似误差等代理指标间接生成解释,导致优化方向与忠实性需求存在偏差。Polixplain通过策略梯度优化机制,创新性地实现了以 Deletion Score最小化为直接优化目标的显著性解释方法,突破了上述方法的局限性。

2 基于策略梯度的黑盒模型显著性解释

本节介绍一种基于强化学习策略梯度算法直接优化忠实性目标函数的解释方法 Polixplain。如图1所示,Polixplain通过智能体与待解释黑盒模型交互学习模型激活区域,生成输入特征的显著性解释,作为预测结果的特征归因。

本节首先给出显著性解释及 Deletion Score的形式化定义,然后设计可解释性学习目标奖励函数,最后给出基于策略梯度的优化算法。

2.1 显著性解释移除分数

在本文中,假设待解释黑盒模型从数据集 $\mathcal{D}=\{(x_i, y_i)\}_{i=1}^n$ 学习模型 $f: \mathbf{R}^d \rightarrow \mathbf{R}$, 其中 $x \in \mathbf{R}^d, y \in \mathbf{R}$ 。输入特征 $x=(x_1, x_2, \dots, x_d), x_i$ 可为像素颜色等图像特征。显著性解释 $s \in \mathbf{R}^d$ 与 x 的形状相同,每个位置的值表示 x 对应位置特征的重要性程度,可以体现1次预测过程中模型关注的特征区域以及对不同维度特征的关注程度。

尽管 Deletion Score 的概念已经得到了广泛的应用^[27-29],但其形式化表达首次在本文中给出。Deletion Score 用降序移除的方法评价1个显著性解释 s 的忠实性。给定1个输入特征 $\bar{x}$,模型预测结果为 $f(\bar{x})$,通过可解释性算法 $A(f, \bar{x})$ 得到对此次预测的显著性解释 $\bar{s}$ 。Deletion Score 首先根据 $\bar{s}$ 的 top-K 元素生成重要性掩码 $m(k; \bar{x}, \bar{s})$,有

$$m_i(k; \bar{x}, \bar{s}) = \begin{cases} 1 & \bar{x}_i \text{ in Top}_k(\bar{s}_i) \\ 0 & \text{其他} \end{cases} \quad (1)$$

移除 top-K 重要特征后的输入为 $\bar{x} \otimes m(k; \bar{x}, \bar{s})$,其中运算符 $\otimes$ 表示输入特征与掩码逐元素相乘。令 $\phi(x; f, \bar{x}) = P[f(x) = f(\bar{x})]$ 表示1个输入特征的预测结果与待解释特征的预测结果相同的概率,Deletion Score 定义为

$$D(\bar{s}; f, \bar{x}) = \sum_{i=0}^d \phi[\bar{x} \otimes m(i; \bar{x}, \bar{s}); f, \bar{x}] \quad (2)$$

式中 i 为向量的第 i 个位置。

更低的 Deletion Score 表示认为重要性程度高的更少量特征对模型预测结果的贡献更大,即能够忠实地反映模型预测过程所依赖的特征。Deletion Score 的计算过程中涉及离散的掩码生成操作,使得通过梯度优化 Deletion Score 存在困难。令 $(\bar{s}, \bar{c}) = \text{sort}(\bar{s})$ 为降序排列的显著性解释及对照索引,设定正整数 $T > 0$,构造向量 $u, v \in \mathbf{R}^d$,其中

$$u_i = \begin{cases} 0 & i \text{ in } \bar{s}[:k] \\ -T & \text{其他} \end{cases}, \quad v_i = \begin{cases} 1 & i \text{ in } \bar{s}[:k] \\ 0 & \text{其他} \end{cases} \quad (3)$$

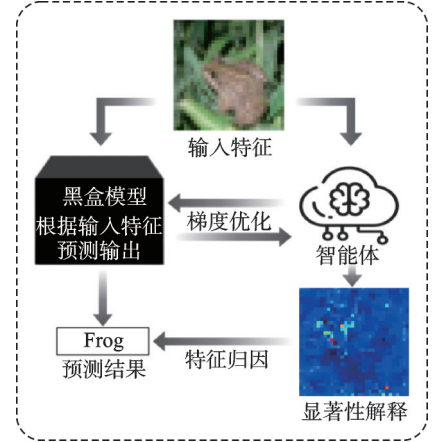


图1 Polixplain解释流程

Fig.1 Explaining process of Polixplain

令掩码 $\tilde{m} = k \text{Softmax}[\text{sort}^{-1}(\tilde{s} \otimes v + u, \tilde{c})/\tau]$, 其中 sort^{-1} 用于按原始顺序将元素顺序还原。当 $T \rightarrow +\infty$ 时, 由于 $\text{sort}^{-1}(\tilde{s} \otimes v + u, \tilde{c})$ 中 top-K 位置保留原值, 而其他位置的值趋于 $-\infty$; 当 Softmax 函数的平滑参数 $\tau \rightarrow +\infty$ 时, $\tilde{m}$ 中 top-K 位置趋近于 1, 其余位置趋近于 0, 因此 $\tilde{m}$ 是重要性掩码 m 的可微近似。定义可微 Deletion Score 为

$$\tilde{D}(\bar{s}; f, \bar{x}) = \sum_{i=0}^d \phi[\bar{x} \otimes \tilde{m}(i; \bar{x}, \bar{s}); f, \bar{x}] \quad (4)$$

式(4)为 $D(\bar{s}; f, \bar{x})$ 的近似, 且由于 $\tilde{m}$ 仅使用对 $\bar{s}$ 的连续函数, 可以获取对显著性解释的梯度 $\frac{\partial \tilde{D}}{\partial \bar{s}}$, 使得可以通过梯度优化方法提高显著性解释的忠实性。

2.2 可解释性学习目标

直接在全空间中求解最优显著性解释的复杂度为 $\mathcal{O}([0, 1]^d)$, 指数级的搜索空间导致优化方法困难。为解决此挑战, 将求解显著性解释问题建模为 1 个基于策略梯度^[30-31]的强化学习问题。定义 1 个 MDP 四元组 $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$, 其中状态 $s_t \in \mathcal{S}$ 表示 t 时刻的显著性解释; 动作 $a_t \in \mathcal{A}$ 表示 t 时刻策略网络 $\pi_\theta(s_t, x)$ 依概率采取的生成下一时刻显著性解释的动作, 被隐式地编码在网络参数中; 状态转移函数 $P(s_t, a_t) \in \mathcal{P}$ 是一个确定性转移; $r_t = R(s_t)$ 表示 t 时刻的奖励值, 奖励函数 $R(s_t) \in \mathcal{R}$ 衡量 s_t 的忠实性和可读性并指导策略网络的优化方向。

作为一种强化学习方法, 策略梯度需要通过最大化奖励函数指导智能体行为。奖励函数定义为

$$R(s_t) = R_D(s_t) - \sum_{i=1}^m \lambda_i \cdot \Gamma_i(s_t) \quad (5)$$

式中 $R_D(s_t) = 1 - \tilde{D}(s_t; f, x)$ 为可优化忠实性奖励, 直接关联 Deletion Score 最小化目标。 $\Gamma_i(s_t)$ 为正则化项, 通过约束显著性解释的重要性值提升可读性, 其权重 λ_i 可人为设置或通过交叉验证动态调整。

由于 Deletion Score 关注显著性解释元素相对大小关系而不关注其绝对值, 以下正则化约束可以在保持忠实性的同时增加解释的可读性。

(1) 稀疏性约束定义为

$$\Gamma^{\text{sparse}}(s_t) = \|s_t\|_1 \quad (6)$$

通过 L_1 正则化抑制冗余特征的高亮, 避免全区域激活导致的解释模糊。

(2) 对比度约束定义为

$$\Gamma^{\text{contrast}}(s_t) = 1 - \frac{\|s_t - \mu_t\|_2^2}{\mu_t^2} \quad (7)$$

式中 $\mu_t = \frac{1}{d} \sum_{i=1}^d s_t(i)$ 为显著性解释均值。通过最大化重要性值的动态范围, 增强重要特征与非重要特征的区分度。此约束惩罚低方差分布, 鼓励显著性解释中高值特征更突出、低值特征更抑制, 同时保持原始排序不变。

(3) 分位数约束定义为

$$\Gamma^{\text{quantile}}(s_t) = \sum_{k=1}^K \left(\frac{\sum_{i=1}^{\lfloor kd \rfloor} s_t^{\text{sorted}}(i)}{\sum_{i=1}^d s_t(i)} - q_k \right)^2 \quad (8)$$

式中 s_t^{sorted} 为降序排列的显著性值, $q_k \in [0, 1]$ 为预设的分位数目标, K 为预设的分位数数量。通过约束分位数分布, 避免极少数特征占据极高权重或权重过度平均化的极端情况, 同时保持特征间相对重要

性顺序,使显著性解释更符合人类对关键区域集中分布的认知习惯。

2.3 策略梯度优化算法

本节介绍 Polixplain 的策略梯度优化算法,根据奖励函数构建策略网络评价函数,并使用梯度提升方法对策略网络参数进行优化以生成高质量解释。如图 2 所示,策略网络参数化为 $\pi_{\theta}(a|s_t, \mathbf{x})$,输入为当前显著性解释 s_t 与原始样本 $\mathbf{x}$,输出下一时刻的隐式动作和显著性解释 s_{t+1} 。策略评价函数为最大化期望累积奖励。

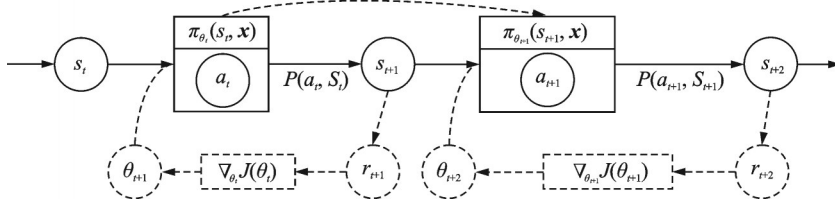


图2 策略网络优化过程

Fig.2 Optimizing process of policy network

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\sum_{t=0}^T R(s_t) \right] \quad (9)$$

根据 REINFORCE 定理^[12],策略梯度为

$$\nabla_{\theta} J(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_t | s_t, \mathbf{x}) \cdot G_t \right] \quad (10)$$

式中 $G_t = \sum_{k=t}^T R(s_k)$ 为从时刻 t 到终止时刻 T 的累积回报。策略梯度 $\nabla_{\theta_t} J(\theta_t)$ 通过梯度提升方法引导网络参数优化,得到下一时刻的策略网络 $\pi_{\theta_{t+1}}(a|s_{t+1}, \mathbf{x})$ 。策略网络具体优化步骤如算法 1 所示。

算法 1 策略梯度解释优化

输入:待解释模型 f , 样本 x_0 , 迭代轮数 I ;

输出:显著性解释 $s^* = \pi_{\theta_I}(s_I, \mathbf{x})$ 。

(1) 初始化显著性解释 $s_0 \sim \text{Uniform}([0, 1]^d)$, 随机初始化策略网络参数 θ_0

(2) for $T = 0, 1, \dots, I - 1$ do

(3) for $t = 0, 1, \dots, T$ do

(4) $a_t \sim \pi_{\theta_t}(\cdot | s_t, \mathbf{x})$

(5) $s_{t+1} \leftarrow P(s_t, a_t)$

(6) end for

(7) $G_T \leftarrow \sum_{k=t}^T R(s_k)$

(8) $\nabla_{\theta_T} J(\theta_T) \approx \sum_{i,t} \nabla_{\theta_t} \log \pi_{\theta_t}(a_t^{(i)} | s_t^{(i)}, \mathbf{x}) G_t^{(i)}$

(9) $\theta_{T+1} \leftarrow \theta_T + \frac{\alpha}{N} \cdot \nabla_{\theta_T} J(\theta_T)$

(10) end for

算法 1 中, I 为用户指定的迭代轮数, 根据被解释模型的复杂度和策略网络的优化深度动态选择; α 为策略网络优化器的学习率。

3 实验验证

本节将Polixplain应用于图像数据集的分类任务上的实验分析。在单个NVIDIA RTX A6000 48 GB显卡上基于编码器-解码器结构的卷积神经网络实现了Polixplain的策略网络(表1中 H 、 W 分别表示卷积层固定的图像高度与图像宽度,取决于数据集中图像大小),超参数 $\lambda_{\text{sparse}} = 0.1, \lambda_{\text{contrast}} = 0.01, \lambda_{\text{quantile}} = 0.001, \tau = 10^5, q_k = 0.1$,迭代轮数 $I_{\text{MNIST}} = 30, I_{\text{CIFAR}} = 80$,学习率 $\alpha = 0.01$,并在两个公开图像数据集上验证了其性能。

表1 Polixplain策略网络架构
Table 1 Architecture of Polixplain's policy network

Module	Layer	Input	Output	Kernel size	Padding	Activation
Encoder	1	$3, H, W$	$32, H, W$	3	1	ReLU
	2	$32, H, W$	$64, H, W$	3	1	ReLU
	3	$64, H, W$	$128, H, W$	3	1	ReLU
	4	$128, H, W$	$128, H, W$	3	1	ReLU
	5	$128, H, W$	$64, H, W$	3	1	ReLU
Decoder	6	$64, H, W$	$64, H, W$	3	1	ReLU
	7	$64, H, W$	$32, H, W$	3	1	ReLU
	8	$32, H, W$	$32, H, W$	3	1	ReLU
	9	$32, H, W$	$16, H, W$	3	1	ReLU
	10	$16, H, W$	$1, H, W$	1	0	Sigmoid

本节首先对Polixplain生成的显著性解释与经典可解释性方法进行忠实性定量评估;随后进行可视化分析,以直观展示Polixplain如何捕捉影响模型决策的关键特征;最后分析Polixplain的运行时间开销与超参数敏感性。

3.1 忠实性评估

为全面评估Polixplain的解释性能,在CIFAR-10数据集上设计了系统的对比实验,将其与7种经典解释方法进行忠实性定量比较。实验以Deletion Score为核心评价指标,旨在验证Polixplain在精确揭示模型决策依据方面的优势。参与对比的解释方法涵盖不同技术路线,具有良好的代表性:(1)基于扰动的方法:Occlusion^[32]、RISE^[11];(2)基于梯度/反向传播的方法:Saliency^[33]、Input $\times$ Gradient^[34]、Grad-CAM^[21];(3)基于线性近似的方法:LIME^[23]、Kernel SHAP^[20]。

实验使用预训练的ResNet-18模型在CIFAR-10测试集上进行评估。从每个类别中随机抽取50个样本,共500个测试样本,计算平均Deletion Score以确保统计显著性。表2中的实验数据表明,本文所提方法在解释忠实性上具有显著优势。具体而言,Polixplain的Deletion Score为0.065,远低于对比方法。其中,基于扰动的方法(如LIME^[23]的0.112和Kernel SHAP^[20]的0.111)与基于梯度/反向传播的方法(如Saliency^[33]的0.152、Input $\times$ Gradient^[34]的0.154)的忠实性均显著弱于Polixplain。Grad-CAM^[21]方法的平均Deletion Score高达0.668,这表明其生成的类激活图虽能定位与语义相关的宽泛区域,但在像素级的精细归因上存在严重不足,难以精确反映模型的决策逻辑。

3.2 可视化分析

为直观验证Polixplain的显著性解释能力,在MNIST手写数字数据集上对数字0~9的分类归因进

表 2 CIFAR-10数据集上的忠实性评估对比实验

Table 2 Comparative experiment on faithfulness evaluation on CIFAR-10 dataset

Method	Polixplain	Saliency	RISE	Grad-CAM	LIME	Kernel SHAP	Occlusion	Input $\times$ Gradient
Deletion Score	0.065	0.152	0.154	0.668	0.112	0.111	0.321	0.154

行了可视化分析。图3展示了原始图像与Polixplain生成的显著性解释及其Deletion Score,其中高亮区域(紫色)表示模型决策时关注的关键特征。

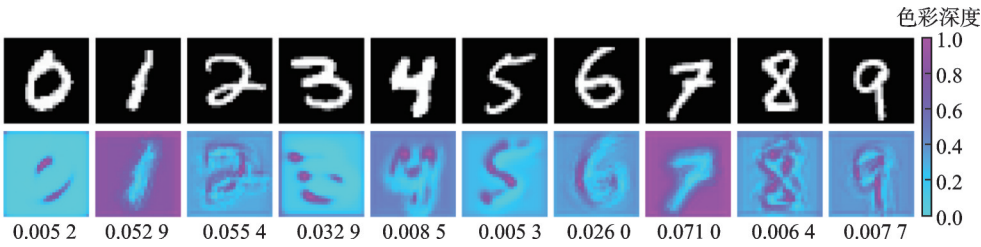


图3 Mnist数字识别解释可视化

Fig.3 Visualization of explaining digit recognition on Mnist

数字0的显著性解释高亮区域集中于左上和右下弧线,与数字0的闭合环形结构高度吻合;数字3左侧3个连续点状区域被显著高亮,对应数字3的垂直笔画特征,验证了模型对局部细节的敏感性;数字4上方两个顶部开放区域的激活,说明模型依赖此类几何特征进行分类。然而,数字1和7的显著性解释高亮区域集中于背景区域,而非数字本身的笔画。这一现象表明模型可能通过排除法识别1和7,如通过背景空白区域推断数字形状,且当前分类模型未充分学习1和7的细微差异特征,这一发现可为模型改进提供方向。

为评估Polixplain的解释忠实性与人类偏好的一致性,在CIFAR-10数据集图像分类任务中,对比了Polixplain与SHAP^[20]、Grad-CAM^[21]和LIME^[23]这3种经典方法在忠实性指标与可视化效果上的差异。实验基于ResNet-18分类模型,针对标签为“Ship”的示例图像生成解释结果,如图4与表3所示。

SHAP^[20]与LIME^[23]生成离散的高亮斑块,呈现不规则超像素组合,二者重要性值分布缺乏空间连续性,难以关联具体船舶特征,可读性较差;Grad-CAM^[21]高亮区域覆盖整个船体,未能区分关键特征与

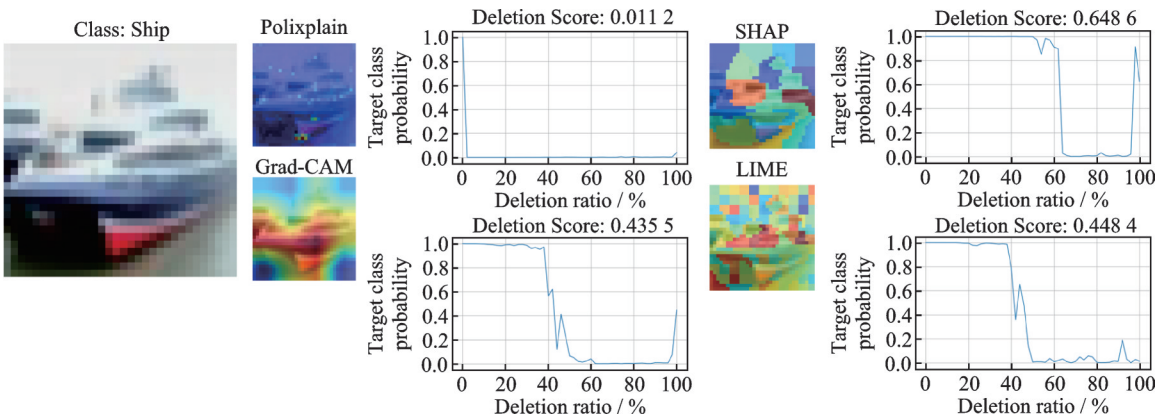


图4 CIFAR-10分类任务解释忠实性对比

Fig.4 Comparison of explanation faithfulness on CIFAR-10 classification

冗余背景,解释过于宽泛;Polixplain利用可读性约束,高亮区域集中于船体底部吃水线特征,与船舶类图像的浮力特征高度吻合,具有明确的语义。

如图4中曲线所示,SHAP^[20]、Grad-CAM^[21]和LIME^[23]需遮蔽约50%高重要性区域像素才能使准确率下降至随机水平,而Polixplain仅需移除1.2%的重要像素即可触发准确率骤降,表示Polixplain高亮区域关注到了模型决策的关键依据;Polixplain的Deletion Score显著低于SHAP^[20]、Grad-CAM^[21]和LIME^[23]这3种经典方法,表明Polixplain解释更具忠实性。

3.3 运行时间分析

策略梯度方法由于需要与优化目标进行多次交互,通常与高训练成本相关。Polixplain进行解释时,需要用户指定迭代轮数,轮数不足则解释忠实性未能充分收敛,而迭代过度则时间开销过大。表4给出了CIFAR-10数据集上Polixplain在迭代过程中的平均Deletion Score与运行时间。在迭代5轮后,运行时间稳定在平均每5次迭代耗时0.6 s左右,而Deletion Score呈逐渐收敛趋势。

表4 Polixplain在迭代中的性能
Table 4 Performance of Polixplain among iterations

Iteration	5	10	15	20	25	30
Deletion Score	0.264	0.147	0.098	0.045	0.023	0.012
Runtime/s	3.255	3.831	4.425	5.021	5.594	6.170

在实际应用中,Polixplain的Deletion Score收敛到一定程度即可获取兼具忠实性与可读性的解释,因此建议用户设定阈值(如0.05),当Deletion Score降低至阈值以下时停止策略网络的迭代。

3.4 超参数敏感性分析

为了全面评估Polixplain的鲁棒性与可复现性,本节对策略网络优化过程中的关键超参数进行了敏感性分析。实验在CIFAR-10数据集上进行,以ResNet-18对单一样本(标签为“Ship”)的解释任务为基准,主要考察了平滑参数 τ 、可读性约束权重 λ_{sparse} 、 $\lambda_{\text{contrast}}$ 和 $\lambda_{\text{quantile}}$ 对解释结果忠实性的影响。

可微掩码 $\tilde{m}$ 的生成依赖于Softmax函数的平滑参数 τ ,其值直接影响对原始离散掩码的近似程度。当 τ 趋近于0时, $\tilde{m}$ 趋近于one-hot向量,但梯度易饱和;当 τ 趋近于无穷大时, $\tilde{m}$ 则趋近于均匀分布,导致近似误差增大。本文测试了 τ 在区间 $[10^3, 10^7]$ 内解释忠实性与one-hot相似性(计算为 $\tilde{m}$ 和one-hot向量的Pearson相关性)的变化,结果如表5所示。结果表明, $\tilde{m}$ 和one-hot向量的相似性随 τ 的增加逐渐增大。当 $\tau < 10^4$ 时,由于梯度饱和,策略网络优化不稳定,Deletion Score波动较大;当 $\tau > 10^6$ 时,近似掩码过于平滑,无法有效区分特征重要性,导致最终Deletion Score升高。在 $\tau = 10^5$ 时,Deletion Score达到最低且收敛过程最稳定,这与第3节设定的实验参数一致,验证了该取值的合理性。

表3 CIFAR-10可视化实验中的Deletion Score
Table 3 Deletion Score in the CIFAR-10 visualization experiments

Deletion ratio	0%	20%	40%	60%	80%	100%
Polixplain	1.000	0.011	0.011	0.011	0.011	0.031
SHAP	1.000	0.998	0.998	0.907	0.016	0.622
Grad-CAM	1.000	0.992	0.578	0.052	0.012	0.436
LIME	1.000	0.997	0.803	0.018	0.007	0.190

表5 超参数 τ 的敏感性分析
Table 5 Sensitivity analysis of hyperparameter τ

τ	10^3	10^4	10^5	10^6	10^7
Deletion Score	0.098	0.076	0.011	0.045	0.063
Similarity	0.912	0.925	0.964	0.973	0.985

奖励函数中的正则化权重 λ 用于在优化忠实性的同时引入可读性约束。然而,过强的可读性约束可能会与忠实性目标产生冲突。本文通过固定其他参数,系统调整某一 λ 值,观察其对最终 Deletion Score 的影响,结果如表 6 所示。

稀疏性权重 λ_{sparse} :该参数通过 L_1 正则化惩罚非零重要性特征的数量。实验发现,当 $\lambda_{\text{sparse}} < 0.05$ 时,解释图中包含过多冗余激活,Deletion Score 因噪声而略有升高;当 $\lambda_{\text{sparse}} > 0.15$ 时,过度稀疏化会强制抹去模型实际依赖的次要特征,反而损害了忠实性,导致 Deletion Score 快速上升。

对比度权重 $\lambda_{\text{contrast}}$:该参数用于增强显著性与非显著性区域的差异。分析显示, $\lambda_{\text{contrast}} = 0.01$ 是一个理想平衡点。过小的权重(< 0.001)使热力图对比度不足,重要性分布模糊,不利于优化;过大的权重(> 0.05)则可能导致优化过程过于关注极少数特征,忽略了特征间的协同效应,从而损害忠实性。

分位数权重 $\lambda_{\text{quantile}}$:该约束用于控制重要性值的分布形态。相较于其他参数, $\lambda_{\text{quantile}}$ 在较大范围(0.000 1~0.005)内对 Deletion Score 的影响均较小,这表明分位数约束是一种相对温和的优化引导方式,其鲁棒性较强。

4 结束语

本文提出了一种基于策略梯度的黑盒模型显著性解释方法 Polixplain,通过设计可微 Deletion Score 并使用策略梯度方法对其优化过程建模,首次实现了以归因忠实性为直接优化目标的动态解释生成。相较于 SHAP^[20]、Grad-CAM^[21]和 LIME^[23]等方法,Polixplain 不依赖模型结构和其他先验信息,适用于任意黑盒模型。实验表明其解释结果在保持高可读性的同时,忠实性显著高于基线方法。未来研究可从 3 方面展开:探索更多模态数据如文本、视频的统—解释框架,验证方法的跨模态泛化能力;开发生态正则化策略,根据任务需求自适应平衡忠实性与可读性;将 Polixplain 扩展至自动驾驶决策系统等实时解释场景是重要的应用方向。

参考文献:

[1] LONGO L, BRCIC M, CABITZA F, et al. Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions[J]. *Information Fusion*, 2024, 106: 102301.

[2] LI Y, LIU X, WANG X, et al. FakeBench: Uncover the achilles' heels of fake images with large multimodal models[EB/OL]. (2024-09-08). <https://arxiv.org/pdf/2404.13306>.

[3] YE J, ZHOU B, HUANG Z, et al. Loki: A comprehensive synthetic data detection benchmark using large multimodal models [EB/OL]. (2025-04-21). <https://arxiv.org/pdf/2410.09732>.

[4] ARRIETA A B, DÍAZ-RODRÍGUEZ N, DEL SER J, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI[J]. *Information Fusion*, 2020, 58: 82-115.

[5] 裴忠一, 刘璘, 王晨, 等. 面向机器学习应用的可解释性需求分析框架[J]. *计算机研究与发展*, 2024, 61(4): 983-1002.
PEI Zhongyi, LIU Lin, WANG Chen, et al. An explainability-centric requirements analysis framework for machine learning applications[J]. *Journal of Computer Research and Development*, 2024, 61(4): 983-1002.

[6] ZHANG S, HAN T, BHALLA U, et al. Building bridges, not walls-advancing interpretability by unifying feature, data, and

表 6 超参数 λ 的敏感性分析

Table 6 Sensitivity analysis of hyperparameter λ

参数	数值					
λ_{sparse}	0.01	0.05	0.1	0.15	0.2	0.5
Deletion Score	0.035	0.022	0.012	0.018	0.042	0.156
$\lambda_{\text{contrast}}$	0.000 1	0.001	0.005	0.01	0.02	0.05
Deletion Score	0.028	0.018	0.014	0.012	0.019	0.037
$\lambda_{\text{quantile}}$	0.000 01	0.000 1	0.000 5	0.001	0.002	0.005
Deletion Score	0.015	0.013	0.012	0.012	0.013	0.014

- model component attribution[EB/OL]. (2025-01-25). <https://arxiv.org/pdf/2501.18887v1>.
- [7] 李平, 宋舒寒, 张园, 等. HSEGRL:一种分层可自解释的图表示学习模型[J]. 计算机研究与发展, 2024, 61(8): 1993-2007.
LI Ping, SONG Shuhan, ZHANG Yuan, et al. HSEGRL: A hierarchical self-explainable graph representation learning model [J]. Journal of Computer Research and Development, 2024, 61(8): 1993-2007.
- [8] LUNDBERG S M, LEE S I. A unified approach to interpreting model predictions[J]. Advances in Neural Information Processing Systems, 2017, 30: 4768-4777.
- [9] BARKAN O, FUCHS Y, CACIULARU A, et al. Explainable recommendations via attentive multi-persona collaborative filtering[C]//Proceedings of the 14th ACM Conference on Recommender Systems. [S.l.]: ACM, 2020: 468-473.
- [10] GUYON I, ELISSEEFF A. An introduction to variable and feature selection[J]. Journal of Machine Learning Research, 2003, 3: 1157-1182.
- [11] PETSUK V, DAS A, SAENKO K. RISE: Randomized input sampling for explanation of black-box models[EB/OL]. (2018-09-25). <https://arxiv.org/pdf/1806.07421>.
- [12] BARKAN O, HAUON E, CACIULARU A, et al. Grad-SAM: Explaining transformers via gradient self-attention maps[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. [S.l.]: ACM, 2021: 2882-2887.
- [13] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks[C]//Proceedings of Computer Vision-ECCV 2014: the 13th European Conference. [S.l.]: Springer International Publishing, 2014: 818-833.
- [14] CHEN J, SONG L, WAINWRIGHT M, et al. Learning to explain: An information-theoretic perspective on model interpretation[C]//Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2018: 883-892.
- [15] ADEBAYO J, GILMER J, MUELLY M, et al. Sanity checks for saliency maps[J]. Advances in Neural Information Processing Systems, 2018, 31: 9505-9515.
- [16] SITU X, ZUKERMAN I, PARIS C, et al. Learning to explain: Generating stable explanations fast[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. [S.l.]: Association for Computational Linguistics, 2021: 5340-5355.
- [17] SCHWAB P, KARLEN W. CXPlain: Causal explanations for model interpretation under uncertainty[J]. Advances in Neural Information Processing Systems, 2019, 32: 10220-10230.
- [18] 孙莹, 章玉婷, 庄福振, 等. 基于集合效用边际贡献学习的可解释薪酬预测算法[J]. 计算机研究与发展, 2024, 61(5): 1276-1289.
SUN Ying, ZHANG Yuting, ZHUANG Fuzhen, et al. Interpretable salary prediction algorithm based on set utility marginal contribution learning[J]. Journal of Computer Research and Development, 2024, 61(5): 1276-1289.
- [19] RICH M L. Machine learning, automated suspicion algorithms, and the fourth amendment[J]. University of Pennsylvania Law Review, 2016, 164(4): 871-929.
- [20] LUNDBERG S M, LEE S I. A unified approach to interpreting model predictions[J]. Advances in Neural Information Processing Systems, 2017, 30: 4765-4774.
- [21] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE International Conference on Computer Vision. [S.l.]: IEEE, 2017: 618-626.
- [22] HECHTLINGER Y. Interpretation of prediction models using the input gradient[EB/OL]. (2016-11-23). <https://arxiv.org/pdf/1611.07634>.
- [23] RIBEIRO M T, SINGH S, GUESTRIN C. "Why should I trust you?" explaining the predictions of any classifier[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.]: ACM, 2016: 1135-1144.
- [24] ITO T, OCHIAI K, FUKAZAWA Y. C-LIME: A consistency-oriented lime for time-series health-risk predictions[C]//Proceedings of Knowledge Management and Acquisition for Intelligent Systems: the 17th Pacific Rim Knowledge Acquisition Workshop. [S.l.]: Springer International Publishing, 2021: 58-69.
- [25] KRIZHEVSKY A. Learning multiple layers of features from tiny images: Technical Report TR-2009[R]. Toronto: University of Toronto, 2009.

- [26] DENG L. The MNIST database of handwritten digit images for machine learning research[J]. *IEEE Signal Processing Magazine*, 2012, 29(6): 141-142.
- [27] BARKAN O, ASHER Y, ESHEL A, et al. Learning to explain: A model-agnostic framework for explaining black box models [C]//*Proceedings of 2023 IEEE International Conference on Data Mining (ICDM)*. [S.l.]: IEEE, 2023: 944-949.
- [28] CHEN R, ZHANG H, LIANG S, et al. Less is more: Fewer interpretable region via submodular subset selection[J]. *International Conference on Learning Representations*, 2024, 2024: 18280-18302.
- [29] NOVELLO P, FEL T, VIGOUROUX D. Making sense of dependence: Efficient black-box explanations using dependence measure[J]. *Advances in Neural Information Processing Systems*, 2022, 35: 4344-4357.
- [30] SUTTON R S, MCALLESTER D, SINGH S, et al. Policy gradient methods for reinforcement learning with function approximation[J]. *Advances in Neural Information Processing Systems*, 1999, 12: 1057-1063.
- [31] DENKERT R, PHAM H, WARIN X. Control randomisation approach for policy gradient and application to reinforcement learning in optimal switching[J]. *Applied Mathematics & Optimization*, 2025, 91(1): 1-33.
- [32] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks[C]//*Proceedings of Computer Vision—ECCV 2014: the 13th European Conference*. Zurich, Switzerland: Springer, 2014: 818-833.
- [33] SIMONYAN K, VEDALDI A, ZISSERMAN A. Deep inside convolutional networks: Visualising image classification models and saliency maps[C]//*Proceedings of the 2nd International Conference on Learning Representations, ICLR 2014*. Banff, AB, Canada: ICLR, 2014: 1-8.
- [34] SHRIKUMAR A, GREENSIDE P, KUNDAJE A. Learning important features through propagating activation differences [C]//*Proceedings of the 34th International Conference on Machine Learning*. Sydney, Australia: PMLR, 2017: 3145-3153.

作者简介:



果宇涵(2001-),男,硕士研究生,研究方向:核方法与机器学习可解释性,E-mail: 3220231188@bit.edu.cn。



丁立中(1986-),通信作者,男,教授,研究方向:统计学习方法和理论、深度生成模型、核方法与深度核方法,E-mail: lizhong. ding@outlook.com。



韦秋宁(2002-),女,硕士研究生,研究方向:深度生成模型、流模型以及假设检验。

(编辑:刘彦东)

Policy Gradient-Based Saliency Explanation Method for Black-Box Models

GUO Yuhan, DING Lizhong*, WEI Qiuning

(School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China)

Abstract: As a core enabling technology for trustworthy artificial intelligence, research on the interpretability of machine learning models plays a vital role in automated decision-making systems. Among them, saliency-based explanations, which intuitively reveal the basis of model decisions in the form of heatmaps, have become a key research direction in explainable artificial intelligence. However, existing approaches suffer from fundamental limitations: Perturbation-based methods are constrained by manually designed distributional assumptions, making it difficult to balance computational efficiency and generalization ability; gradient-based methods require prior knowledge of the model's internal structure and thus fail to explain black-box models; and linear-approximation-based methods introduce model bias due to excessive local simplification. To address these limitations, we propose Polixplain, a policy-gradient-based saliency explanation method for black-box models. By using the Deletion Score as the core reward signal and incorporating a differentiable relaxation technique into the policy gradient optimization process, Polixplain enables end-to-end optimization toward the faithfulness objective. Image classification tasks are adopted as the experimental testbed, where the high-dimensional feature space and intuitive visual representation provide an ideal environment for multidimensional evaluation of explanation methods. Polixplain outperforms a range of existing explanation methods in terms of both explanation faithfulness and the human interpretability of the generated heatmaps.

Highlights:

1. A policy-gradient-based saliency framework, Polixplain, is developed for model-agnostic explanation of black-box models.
2. A differentiable Deletion Score enables direct end-to-end faithfulness optimization while balancing multiple readability constraints.
3. Polixplain outperforms seven representative baselines in explanation faithfulness and generates semantically coherent, human-interpretable heatmaps.

Key words: interpretability; black-box explanation; saliency map; reinforcement learning; policy gradient

Foundation item: National Key Research and Development Project (No.2022YFB2703100).

Received: 2025-06-12; **Revised:** 2025-08-25

***Corresponding author, E-mail:** lizhong.ding@outlook.com.