

FastScore-CAM: 一种高效的视觉解释方法

袁月璨, 郑 熠, 张 琅

(成都信息工程大学 计算机学院, 四川 成都 610225)

摘要: 在深度学习模型应用中, 可解释性对于建立用户对模型的信任至关重要。视觉解释方法 Score-CAM 利用特征图扰动输入样本, 并与特征重要性权重进行线性组合, 表现出较高的准确率和跨模型的稳定性。然而, Score-CAM 中的扰动样本需要反复调用模型而导致算法效率降低、响应缓慢。为了解决该问题, 提出一种高效的基于扰动的视觉解释方法——FastScore-CAM, 该算法在保持 Score-CAM 算法性能的同时可缓解运行效率低下的问题。进一步通过研究分析 Score-CAM 的基本工作原理, 引入了非负矩阵分解(NMF)针对性地减少扰动样本并过滤冗余特征。实验结果表明, 该方法不仅保持了较高的准确率和跨模型的稳定性能, 还显著提高了原方法的运算速度。

关键词: Score-CAM; 可解释性; 深度学习; 视觉解释方法; 显著性图; 非负矩阵分解

DOI: 10.11907/rjdk.251056

开放科学(资源服务)标识码(OSID):

中图分类号: TP391.41; TP18

文献标识码: A

文章编号: 1672-7800(2025)004-0108-07



FastScore-CAM: An Efficient Saliency Visual Interpretation Method

YUAN Yuecan, ZHENG Yi, ZHANG Lang

(School of Computer Science, Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: In the application of deep learning models, interpretability is crucial for establishing users' trust in the model. The visual explanation method Score-CAM perturbs the input samples using feature maps and linearly combines them with feature importance weights, demonstrating high accuracy and cross-model stability. However, the perturbed samples in Score-CAM require repeated calls to the model, resulting in reduced algorithm efficiency and slow response. To address this issue, an efficient perturbation-based visual explanation method, FastScore-CAM, is proposed. This method alleviates the problem of low operational efficiency while maintaining the performance of the Score-CAM. Specifically, by studying and analyzing the basic working principle of the Score-CAM, Nonnegative Matrix Factorization (NMF) is introduced to specifically reduce perturbed samples and filter redundant features. Experimental results show that this method not only maintains high accuracy and cross-model stable performance but also significantly improves the computational speed of the original method.

Key Words: Score-CAM; interpretability; deep learning; visual explanation method; saliency map; non-negative matrix decomposition

0 引言

深度学习模型在人工智能领域的快速发展中扮演着关键角色, 相较于传统的机器学习模型, 神经网络等深度学习学习方法已展现出更高的准确性。然而, 由于深度学习模型通常被视为“黑盒”, 如何使用户理解这些模型的决策过程成为一个亟需解决的问题。尤其在医疗、气象、智能驾驶等关键领域, 具备可解释性(Interpretability)的模型能够

增强用户的信任, 从而促进其广泛应用^[1]。对于工业界而言, 可解释性甚至是人工智能技术能否成功落地的决定性因素。

近年来, 深度学习模型在视觉任务中的可解释性已受到广泛关注, 研究者提出了多种解释方法, 这些方法通过显著性图可视化, 突显图像中驱动模型决策的关键特征, 以提高模型可解释性^[2-4]。目前, 视觉解释方法大致可以分为两类: ①基于梯度的方法; ②基于扰动的方法。基于梯度的方法, 如层相关性传播(Layer-wise Relevance Propa-

收稿日期: 2025-01-22

扫描二维码阅读全文:

基金项目: 四川省科技计划资助项目(2024ZDZX0007)

作者简介: 袁月璨(1995-), 男, 成都信息工程大学计算机学院硕士研究生, 研究方向为机器学习与可解释性; 郑熠(1999-), 男, 成都信息工程大学计算机学院硕士研究生, 研究方向为机器学习与图像处理; 张琅(2000-), 男, 成都信息工程大学计算机学院硕士研究生, 研究方向为图像处理与目标检测。本文通讯作者: 袁月璨。



gation, LRP)^[5]、Grad-CAM^[6]和 Grad-CAM++^[7],通过计算输出层到中间层或输入层的梯度以获取特征权重,并结合特征进行线性组合以实现可视化解释。这些方法计算效率高,但对图像噪声较为敏感。与之相比,基于扰动的方法,如 RISE^[8]、Group-CAM^[9]和 Abs-CAM^[10],通过扰动输入特征并观察输出置信度的变化以计算特征权重。这些方法不依赖梯度,因而能在跨模型应用中提供更强的鲁棒性。然而,基于扰动的方法需要反复调用网络模型,导致算法资源开销增加,并使得计算时间显著增长。

基于扰动的解释方法认为,通过改变输入特征某些特定部分,可以有效地衡量这些特征对输出结果的贡献^[11]。Wang等^[12]提出了 Score-CAM,该方法通过目标类别在前向传播中的得分为每个特征图分配权重,从而摆脱了对梯度的依赖。Score-CAM通过权重与特征图的线性组合生成显著性图,实验表明其在视觉解释任务中具有较高的准确率和鲁棒性。然而,Score-CAM对输入图像进行的扰动数目与通道数相等,导致该方法需要使用大量扰动样本重新调用网络模型,从而降低了算法整体运行效率,甚至影响其实际应用的可行性。

为了解决这一问题,本文提出了一种高效的视觉解释方法——FastScore-CAM。通过对 Score-CAM 的运行机制进行分析发现,使用原始扰动样本的子集可以近似保留原有的解释质量。为此,本文引入非负矩阵分解(Nonnegative Matrix Factorization, NMF)对特征图进行降维,从而减少了冗余特征样本数量,并大幅降低了算法对 DNN 的访问频次,从而显著提升了计算效率,同时保持了 Score-CAM 的优良性能^[13]。

本文主要贡献如下:①分析了扰动样本数量及其分布对线性决策边界构建的影响,并引入 NMF 减少冗余扰动,提出了 FastScore-CAM 解释方法;②实验结果表明, FastScore-CAM 不仅保留了 Score-CAM 的高解释准确性和跨模型稳定性,还显著提升了计算效率;③通过支持性实验,探讨了降维算法的选择及目标维度的确定,验证了 NMF 在避免随机扰动方面的有效性。

1 相关工作

根据生成视觉解释方式的不同,目前的解释方法可分为基于梯度的方法和基于扰动的方法。

1.1 基于梯度的解释方法

早期的梯度解释方法,如 LRP 通过反向传播,从输出层逐层计算相关性,直到输入层。但在非线性网络结构中,LRP 无法得到完美的相关性分解。Zhou 等^[14]提出类激活图(Class Activation Map, CAM),该方法通过特征层与输出层之间的全局平均池化(Global Average Pooling, GAP)为每个通道的特征图赋予权重。然而,CAM 方法非常依赖于 GAP。随后, Selvaraju 等^[6]将反向传播与 CAM 相结合,提

出了 Grad-CAM 方法,该方法通过计算从输出层到任意目标中间层的梯度以实现优化,而不需要关注网络具体结构,因而表现出更为鲁棒的性能。Chattopadhyay 等^[15]在 Grad-CAM 的基础上引入了额外权重,这些权重根据神经网络中的局部结构进行调整,旨在更准确地标记图像中同一类别的多个目标。Fu 等^[16]通过对梯度进行归一化处理,使得梯度信息的利用更加合理,从而在一些复杂场景下能够更准确地定位重要特征区域。然而,基于梯度的解释方法对图像噪声高度敏感,并且其性能会受到梯度消失和梯度爆炸问题的影响。

1.2 基于扰动的解释方法

通过扰动输入特征并观察输出结果的变化幅度,可以衡量该扰动对模型的影响程度,这一思路出自因果解释模型^[17]。Petsiuk 等^[8]采用蒙特卡罗采样生成随机掩码扰动输入图像,再将扰动后的图像输入模型,并计算置信度的下降幅度,作为该扰动图像的权重。研究者认为,这种基于扰动的方法可以广泛应用于不同的机器学习模型。

随后,因果关系启发对输入进行扰动的研究,开始探讨影响深度学习模型输出的各类特定因素,包括但不限于特征图^[18-19]、神经元^[20]、图像网格^[21-22]以及输入图像^[10,12]的变化。其中,Score-CAM 利用中间层特征图扰动输入图像。相较于基于梯度的方法,该方法不依赖于梯度,因而在跨模型解释时展现出更强的鲁棒性。然而,该方法因扰动样本过多而非非常耗时。此外,基于合作博弈的 Shapley 算法也被引入到解释方法中^[23-24],其目的是为特征权重分配提供更高的公平性。然而,由于 Shapley 值的计算具有指数级的时间复杂度,其实际应用面临着显著的可行性挑战。

Zhang 等^[9]首次探讨了扰动数量对效率的影响,提出了对特征进行分组融合的方法。然而,这一顺序编组的策略导致分组结果具有较强的随机性。为解决这一问题, Feng 等^[25]提出使用无监督聚类算法对特征图语义进行聚类,以降低扰动样本数量。然而,聚类方法本身也存在一些局限性,例如计算复杂度较高、对异常值敏感及无法自动选择最优聚类数。目前,以上扰动方法尚未明确探讨如何确定适当的扰动数量,以在保证模型性能的基础上有效降低计算成本。

2 研究方法

深度学习模型视觉解释方法的算法流程如下:首先,准备预训练的网络模型、输入图像及对应目标数据标签;接着,将输入图像输入网络模型,对于算法利用过程中产生的中间数据与输出结果,经特征提取、重要性计算及图像处理,最终生成显著性图。图中热区覆盖的特征与数据标签相关,是模型识别出的关键特征。

2.1 Score-CAM方法

Score-CAM 是一种基于扰动的解释方法,通过扰动样本在网络模型内获取前向传播得分,并将其作为特征权重。其算法分两阶段:①从网络提取特征图;②将特征图处理成掩码扰动输入图像,计算扰动图像在目标类别得分。最终,经特征图与得分线性组合生成显著性视觉解释结果。以下为形式化表示,其中加粗符号用于表示张量、矩阵或向量。

给定一个 RGB 格式的输入图像 $x \in \mathbb{R}^{C \times H \times W}$, 将其输入到 DNN 模型 $\mathcal{F}(\cdot)$ 中, 并在目标层 ℓ 中提取特征图 $\mathbf{A} \in \mathbb{R}^{c \times h \times w}$ 。设基线图像为 $\mathbf{x}_b$, 哈达玛积为 $\odot$, 目标类别标签为 y , Min-Max 归一化函数为 $MinMax(\cdot)$, 双线性插值为 $\mathcal{B}(\cdot)$ 。对于 $\mathbf{a} \in \mathbf{A} (\mathbf{a} \in \mathbb{R}^{h \times w})$, 扰动掩码定义为 $\mathcal{M}_\ell = \mathcal{B}(\mathbf{a})$, 其特征权重计算公式为 $w_a = \mathcal{F}_y(MinMax(\mathcal{M}_a) \odot \mathbf{x}) - \mathcal{F}_y(\mathbf{x}_b)$ 。最后线性组合 w_a 与 $\mathcal{M}_\ell$ 得到:

$$\text{Score-CAM} = MinMax(ReLU(\sum_{a \in \mathbf{A}} w_a \odot \mathcal{M}_a)) \quad (1)$$

其中, Score-CAM 表示经特征与特征权重线性叠加后最后的显著性图可视化效果。该方法的基本原理是基于特征图与输入图像的空间相关性^[26], 通过掩码来扰动输入图像, 保留特征相关像素, 获取有特征表示能力的扰动样本。本文还提供了伪代码以总结 Score-CAM 算法(见算法 1)。

算法 1 Score-CAM

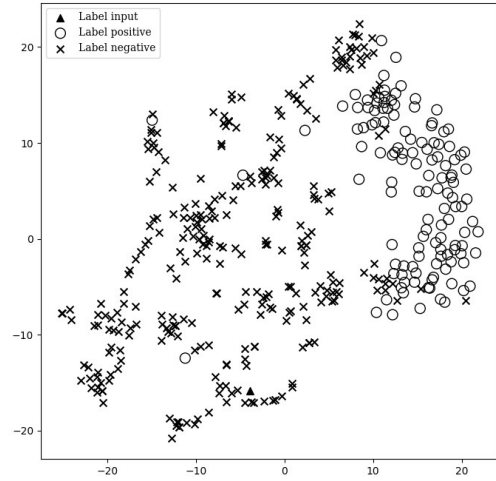
输入: 输入图像 $\mathbf{x}$; 基线图像 $\mathbf{x}_b$; 预训练的 DNN 模型 $\mathcal{F}(\cdot)$; 目标层 ℓ ; 目标类别 y 。

输出: Score-CAM。

1. 提取特征图: $\mathbf{A} \leftarrow \mathcal{F}_\ell(\mathbf{x})$
2. 初始化 CAM: $\text{CAM} \leftarrow 0 (\text{CAM} \in \mathbb{R}^{C \times H \times W})$
3. FOR $\mathbf{A}$ 的每个通道下特征图 $\mathbf{a}$
4. 上采样至输入图像尺寸已生成扰动样本: $\mathcal{M}_\ell \leftarrow \mathcal{B}(\mathbf{a})$
5. 计算得分权重: $w_a \leftarrow \mathcal{F}_y(MinMax(\mathcal{M}_\ell) \odot \mathbf{x}) - \mathcal{F}_y(\mathbf{x}_b)$
6. 将得分权重线性叠加至 CAM: $\text{CAM} += w_a \odot \mathcal{M}_\ell$
7. END
8. RETURN Score-CAM $\leftarrow MinMax(ReLU(\text{CAM}))$

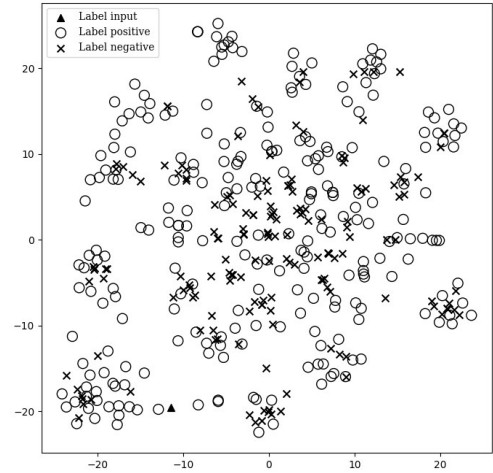
由算法 1 可知, Score-CAM 时间复杂度主要取决于特征图通道数 c 。由于网络模型依靠增加通道数提升特征表示能力^[27], 因而 Score-CAM 的执行速度随通道数增加呈线性增长。若以 T_{total} 表示为 Score-CAM 方法的总执行时间, 它满足 $T_{total} \approx c \cdot T_{single}$ 。其中, T_{single} 表示输入数据单次通过模型结构的执行时间, 受网络结构、模型参数和硬件环境等外部因素限制。因此, 优化策略多着眼于减少扰动样本数量, 以降低对网络模型的访问频率, 从而显著提高算法效率。

探讨减少特征样本数量的优化策略前, 需明确关键问题: Score-CAM 使用随机扰动能否实现有效解释? 若能, Score-CAM 可采用少量的随机扰动, 便可加快执行速度, 且无需从网络模型提取特征图。为此, 本文运用 t-分布



(a) Disturbance result of Score-CAM

(a) Score-CAM 扰动结果



(b) Random disturbance result

(b) 随机扰动结果

Fig. 1 The distribution of two kinds of disturbed sample masks

图 1 两种扰动样本掩码的分布情况

机领域嵌入 (t-Distributed Stochastic Neighbor Embedding, t-SNE)^[28] 的可视化技术进行分析, 图 1(a) 展示了扰动样本在 t-SNE 中的分布。结果显示, Score-CAM 扰动特征图生成的正样本 (o 符号) 与负样本 (x 符号) 之间存在相对的线性语义边界。同样地, 用相同数量随机扰动模拟上述过程, 公式 $\mathcal{M}_\ell = MinMax(\mathcal{B}(shuffle(\mathbf{a})))$ 表示随机掩码的生成方式, 其中 $shuffle(\cdot)$ 函数用于打乱矩阵元素值。随机扰动样本 $\mathcal{M}_\ell \odot \mathbf{x}$ 的数据分布结果如图 1(b) 所示, 对比发现: ①随机扰动样本的分布分散且无规律; ②随机扰动的正负样本数据点紧密粘连, 不具备构建有效线性边界的条件。因此, 扰动输入特征样本不应依赖随机掩码生成。本文将在下文以对比实验的方式讨论这种随机扰动的可行性。

2.2 Fast-ScoreCAM

基于上述分析, 优化 Score-CAM 的关键在于保留能形成分类边界的部分扰动样本, 有两种策略: ①特征图采样; ②特征图降维。由于随机采样结果的不确定性, 难以保持 Score-CAM 的性能, 本文不推荐采纳这种策略, 并通过实验对二者差异进行比较分析。为此, 本文引入一种基于低秩分解的降维技术——NMF, 用于对初始特征图进行降维处理, 获得拥有局部特征表示的特征掩码, 以生成少量扰动样本, 提高算法整体效率, 此方法称为 FastScore-CAM。

对于给定 DNN $\mathcal{F}(\cdot)$, 在 $\mathbf{x}$ 首次访问该网络时从其目标层 ℓ 中提取特征图 $\mathbf{A}$, 将 $\mathbf{A}$ 映射至非负空间。接着, 通过 NMF 对 $\mathbf{A}$ 降维 ($\mathbb{R}^{c \times (h \times w)} \rightarrow \mathbb{R}^{d \times (h \times w)}, d < c$) 获得低维特征图 $\mathbf{V}$, 该过程表示为 $\mathbf{V} \leftarrow \text{NMF}_d(\text{Translate}(\mathbf{A}))$, d 表示目标维度, $\text{Translate}(\cdot)$ 为平移函数。 $\mathbf{A}$ 与 $\mathbf{V}$ 的关系为:

$$\mathbf{A}^{c \times (h \times w)} \approx \mathbf{U}^{c \times d} \mathbf{V}^{d \times (h \times w)} \quad (2)$$

其中, $\mathbf{U}$ 为基矩阵, 由 d 个基向量张成, $\mathbf{V}$ 是系数矩阵, 表示特征图在特征子空间中的投影系数。 $\mathbf{V}$ 集合了特征图 $\mathbf{A}$ 中潜在的局部特征结构, 其值与 $\mathbf{A}$ 的空间位置相互对应, 并且系数值越高, 说明该位置的特征成分越显著。在更新后的算法中, 扰动样本的特征掩码由 $\mathcal{M}_v = \text{MinMax}(\mathcal{B}(\mathbf{v}))$ 生成, 其中 $\mathbf{v} \in \mathbf{V}$, 重要性权重由 $w_v = \mathcal{F}_y(\mathcal{M}_v \odot \mathbf{x})$ 计算得到, 则 FastScore-CAM 的生成公式如下:

$$\text{FastScore-CAM} = \text{MinMax}(\text{ReLU}(\sum_{v \in \mathbf{V}} w_v \odot \mathcal{M}_v)) \quad (3)$$

本文提供了伪代码算法逻辑 (见算法 2), 以便复现。与 Score-CAM 相比, FastScore-CAM 有如下显著不同之处: ①将 $\mathbf{A}$ 转换到非负空间, 使用 NMF 获得低维特征图 $\mathbf{V}$, 并替换 Score-CAM 中 $\mathbf{A}$ 的作用以加快速度; ②线性组合时, 特征表示用 $\mathcal{M}_v = \text{MinMax}(\mathcal{B}(\mathbf{v}))$ 替换 $\mathcal{M}_a = \mathcal{B}(\mathbf{a})$, 旨在减小不同特征的数值差异, 避免数值大的特征主导可视化效果以及权重失效; ③为简化步骤, 去除基线图像 $\mathbf{x}_b$ 计算的得分权重, 因为线性组合结果呈相对性, 更注重权重间的比例。

算法 2 FastScore-CAM

输入: 输入图像 $\mathbf{x}$; 预训练的 DNN 模型 $\mathcal{F}(\cdot)$; 目标层 ℓ ; 目标类别 y ; 目标维度 d ;

输出: FastScore-CAM。

1. 提取特征图: $\mathbf{A} \leftarrow \mathcal{F}_\ell(\mathbf{x})$
2. 将 $\mathbf{A}$ 转换到非负区间: $\mathbf{A} \leftarrow \mathbf{A} - \text{Min}(\mathbf{A})$ if $\text{Min}(\mathbf{A}) < 0$
3. 执行非负矩阵分解获得低维特征: $\mathbf{V} \leftarrow \text{NMF}_d(\mathbf{A})$
4. CAM 初始化: $\text{CAM} \leftarrow 0$ ($\text{CAM} \in \mathbb{R}^{c \times h \times w}$)
5. FOR $\mathbf{V}$ 的每个通道下特征图 $\mathbf{v}$
6. 归一化和上采样: $\mathcal{M}_v \leftarrow \text{MinMax}(\mathcal{B}(\mathbf{v}))$
7. 线性组合: $\text{CAM} += \mathcal{F}_y(\mathcal{M}_v \odot \mathbf{x}) \odot \mathcal{M}_v$
8. END
9. RETURN FastScore-CAM $\leftarrow \text{MinMax}(\text{ReLU}(\text{CAM}))$

已有研究表明, 基于独立成分分析 (Independent Component Analysis, ICA) 的非负矩阵分解能够有效地对实体

进行局部特征分解, 所得的特征表示为全体局部特征的集合^[29]。各局部特征表达滤除了内噪声, 并与特征图的部分特征语义相似且非随机。在 FastScore-CAM 的扰动样本的 t-SNE 可视化结果中 (见图 2), 正负样本在样本空间中显著分离, 表明 FastScore-CAM 与 Score-CAM 具有相同能力。

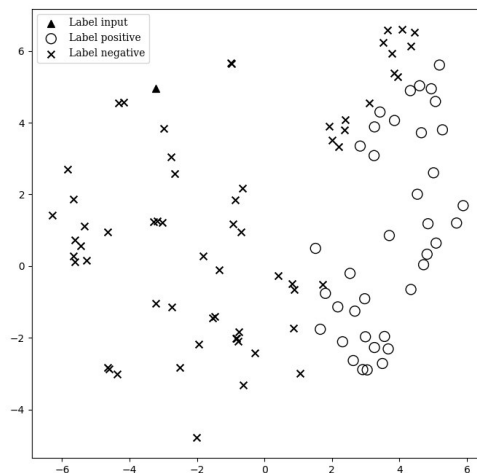


Fig. 2 Disturbance sample distribution of FastScore-CAM

图 2 FastScore-CAM 的扰动样本分布情况

3 实验与结果分析

3.1 数据集

本文使用公开数据集 Mini-ImageNet 的验证集 (val) 以评估所有方法的性能^[30]。Mini-ImageNet 是 ILSVRC 2012 的一个子集, 包含 1 000 个类别, 在每个类别中随机采样 1 张图像, 形成测试集。使用真实标签作为目标标签。数据预处理方法是使用无裁剪的方式将图像调整为 (3, 224, 224) 的 RGB 格式, 归一化参数为 $\text{mean} = (0.485, 0.456, 0.406)$, $\text{std} = (0.229, 0.224, 0.225)$ 。

3.2 实验设置

基线对比的解释方法包括 Grad-CAM、Group-CAM、FD-CAM、Score-CAM。

被解释的 DNN 模型包括 VGG-16^[31]、DenseNet-201^[32]、ResNet-152^[33]。①VGG-16 被广泛应用于大型网络的特征提取器、重要模块或骨干网络; ②DenseNet-201 的最后一层卷积块的特征图具有 1 920 个通道, 意味着 Score-CAM 所需扰动样本数量为 1 920, 这对其算法可行性提出了挑战; ③ResNet-152 引入残差结构以解决堆叠卷积块引发的梯度消失问题, 通常残差结构常见于深层网络模型并伴随着更多通道。对所有网络模型, 目标层 ℓ 设置为最后一个卷积层。

使用 NVIDIA RTX 1080Ti 的 GPU 以及 Pytorch 深度学习框架进行实验。设置 NMF 的目标维度 $d = 100$ 。

3.3 实验方法

本文使用两组实验指标综合评价待测试的所有解释方法:①删除(Deletion, Del)/插入(Insertion, Ins);②平均下降(Average Drop, AD)/平均上升(Average Increase, AI)。

(1)删除(Del)/插入(Ins)。将解释方法生成的显著性图对像素重要性进行从大到小排序,并从输出图像 $\mathbf{x}$ 中逐步删除对应位置像素(将像素值设置为0)。每次删除3.6%的像素,并将更新后的图像 $\mathbf{x}_i$ 输入DNN进行预测并记录 $\mathcal{F}_y(\mathbf{x}_i)$ 。接着,计算曲线 $\mathcal{F}_y(\mathbf{x}_i)$ 与 x 轴围成的面积而得到Del值。插入是删除的逆向操作,按照重要性排序逐步插入像素。最终计算整个测试集的平均Del和Ins,并乘以100,得到最终结果。

(2)平均下降(AD)/平均上升(AI)。首先,根据显著性图中值的大小,保留 $\mathbf{x}$ 中最重要的20%像素,并将其余像素值设置为0,更新修改后的输入表示为 $\mathbf{x}^*$;然后,将 $\mathbf{x}^*$ 输入到DNN中;最后,计算整个数据集的AD和AI。计算方式如下:

$$AD = \frac{1}{|\mathcal{X}|} \sum_{\mathbf{x} \in \mathcal{X}} \frac{\text{Max}(0, \mathcal{F}_y(\mathbf{x}) - \mathcal{F}_y(\mathbf{x}^*))}{\mathcal{F}_y(\mathbf{x})} \quad (4)$$

$$AI = \frac{1}{|\mathcal{X}|} \sum_{\mathbf{x} \in \mathcal{X}} \mathbf{1}_{\mathcal{F}_y(\mathbf{x}) < \mathcal{F}_y(\mathbf{x}^*)} \quad (5)$$

其中, $\mathcal{X}$ 为测试数据集,1为指示函数。

3.4 定性分析

解释方法生成的显著性图显示了DNN推理时所采纳的关键特征,其中热区表示重要特征。若热区覆盖较多无关特征,表示解释错误或噪声过多;若热区更聚焦,表明该方法能进行更细粒度的解释。如图3所示(第一行至第三行的标签分别为“半蹼鹬”“卷尾猴”“算盘”),随机抽取的3个标签的定性对比结果中,每行代表不同的标签图像,每列代表不同的解释方法。首先,FastScore-CAM和Score-CAM结果基本一致,验证了FastScore-CAM是Score-CAM的有效扩展;其次,FastScore-CAM在语义定位上更为精准,尤其在“卷尾猴”标签的例子中,准确捕捉到两个面部特征,而其他方法未能完全覆盖或精确定位关键特征;最后,FastScore-CAM的显著性图噪声较少,非重要特征没有热区覆盖。这些定性比较展示了FastScore-CAM在解释精度上的优势。

3.5 定量分析

对比实验的定量结果如表1所示,表中最优结果以黑色加粗表示,次优结果用下划线标出。可以看出,FastScore-CAM在解释不同DNN时大多表现最好或较好,显示了其优异的解释准确性和跨模型的稳定性。相反,Grad-CAM在VGG-16上的表现最差,Group-CAM在DenseNet-201上最差,在跨模型的鲁棒性测试方面表现稍逊。梯度方法在跨模型测试中的不稳定性源于与ReLU单元“就地”操作的冲突,在逐层回传梯度时,尽管处于推理模式,DNN的训练参数仍因“就地”设置不断更新,导致异常结果。此外,扰动方法通常与模型无关,而梯度方法由于预训练效

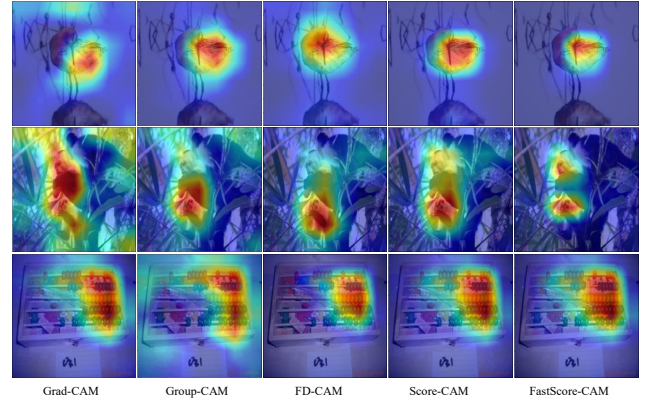


Fig. 3 Qualitative comparison

图3 定性对比

果差异,难以在不同DNN模型上保持稳定性。

FastScore-CAM与Score-CAM在指标上基本一致,表明其保留了大部分原有能力。执行时间方面,Score-CAM在解释DenseNet时耗时46.11 s,ResNet-152时为30.25 s,而FastScore-CAM分别为2.78 s和1.95 s,效率提升显著。对于Score-CAM,每次生成扰动样本都需重新访问DNN,产生固定的传输和推理损耗;而FastScore-CAM通过严格控制扰动样本数量,避免了计算成本呈陡峭线性增长,提升了算法可行性。

Table 1 Comparison experiment of ILSRVC 2012 validation set

表1 ILSRVC 2012验证集对比实验

DNN	可解释性方法	AD ↓	AI ↑	Del ↓	Ins ↑	Time/s
VGG-16	Grad-CAM	46.66	8.30	7.04	41.38	0.02
	Group-CAM	40.31	10.20	6.45	44.09	0.20
	FD-CAM	40.89	10.00	<u>6.49</u>	<u>43.77</u>	0.26
	Score-CAM	38.68	11.00	<u>6.49</u>	44.43	2.90
	FastScore-CAM	<u>39.25</u>	<u>10.80</u>	6.90	43.19	0.70
DenseNet-201	Grad-CAM	37.23	13.80	8.72	<u>51.00</u>	0.08
	Group-CAM	66.89	5.50	20.64	35.13	0.21
	FD-CAM	39.08	14.50	9.98	47.92	0.79
	Score-CAM	<u>37.31</u>	<u>13.20</u>	<u>9.13</u>	50.73	46.11
	FastScore-CAM	35.73	15.10	9.74	51.01	2.78
ResNet-152	Grad-CAM	37.85	13.50	9.91	55.91	0.05
	Group-CAM	41.66	12.90	<u>10.76</u>	55.12	0.20
	FD-CAM	45.72	11.70	12.26	51.40	0.80
	Score-CAM	38.94	13.10	11.00	<u>55.56</u>	30.25
	FastScore-CAM	<u>38.76</u>	<u>13.40</u>	10.97	55.02	1.85

3.6 定位能力评估

本文评估FastScore-CAM的显著性图的定位质量,并验证其在目标类别边界框内准确定位目标特征区域的能力。若显著性图的热区集中于目标特征区域,则可以增强用户对解释结果的信任。本文采用Score-CAM提出的基于能量的“指向游戏”量化定位能力,定义如下:

$$Proportion(x) = \frac{\sum_{L_{(i,j)}^y \in bbox} L_{(i,j)}^y}{\sum_{L_{(i,j)}^y \in bbox} L_{(i,j)}^y + \sum_{L_{(i,j)}^y \notin bbox} L_{(i,j)}^y} \quad (6)$$

其中, $L_{(i,j)}^y$ 表示在目标类别 y 的标签中位置 (i,j) 的显著性值。使用官方提供的标签,计算平均

$Proportion(\mathbf{x})(\mathbf{x} \in \mathcal{X})$ 。此外,加入一种弱监督语义分割方法 Layer-CAM^[34]进行对比评估。实验结果如表 2 所示, FastScore-CAM 的 $Proportion$ 为 61.82%, 优于 Score-CAM (60.18%) 和 Layer-CAM (55.94%), 在所有方法中表现更佳。这一优势源于 NMF 方法能够有效剔除与目标标签无关的冗余特征, 过滤特征图中的噪声, 使得在进行线性叠加时, 关键特征占据主导地位, 从而使显著性图中的高亮区域准确落入边界框内。该结果与定性分析结论一致, 进一步验证了 FastScore-CAM 在定位方面的先进性能。

Table 2 Comparative experiment of Energy-Based Pointing Game

表 2 Energy-Based Pointing Game 对比实验						
	Grad-CAM	Group-CAM	FD-CAM	Layer-CAM	Score-CAM	FastScore-CAM
Proportion/%	54.29	57.92	58.94	55.94	60.18	61.82

3.7 随机扰动与随机采样

上文讨论了一种对特征图的元素值进行打乱从而生成随机扰动特征掩码的方式, 在此称为 RandomDisturb-CAM (RD-CAM)。此外, 还提及一种对特征图进行随机采样的优化策略, 旨在以采样的方式直接降低通道维度, 称为 RandomSample-CAM (RS-CAM)。两者的主要区别在于, RD-CAM 是指扰动样本数据分布的随机性, 而 RS-CAM 则通过随机采样减少扰动样本数量。

为了对比 FastScore-CAM 与 RS-CAM、RD-CAM 的性能差异, 本文进行了对比实验, 结果如表 3 所示。可以观察到: ①FD-CAM 在 3 种方法中表现最差, 这说明有效的解释结果需要避免扰动样本数据分布的随机性; ②FastScore-CAM 在各项评价指标上均显著优于 RS-CAM, 尽管 RS-CAM 在执行效率上相较于 NMF 更为高效, 但两者差距仅为 0.03 s, 优势并不明显。这一现象表明, 虽然随机采样能够有效减少扰动样本数量, 但其随机性无法确保每次采样都能获得理想的关键特征样本, 这种不可控性最终导致算法性能显著下降。

Table 3 Comparison experiment of FastScore-CAM, RS-CAM and RD-CAM

表 3 FastScore-CAM、RS-CAM、RD-CAM 的对比实验					
算法	AD ↓	AI ↑	Del ↓	Ins ↑	Time/s
FastScore-CAM	39.25	10.80	6.90	43.19	0.70
RS-CAM	42.41	9.30	7.14	42.66	0.67
RD-CAM	60.23	3.20	11.23	24.33	2.95

3.8 性能与效率平衡性

根据 FastScore-CAM 的算法流程, 执行 NMF 之前需预设目标维度 $d(d < c)$ 。已知 d 的增加将引起执行时间线性增长, 但增加更多的扰动样本有助于构建更准确的线性分类边界, 从而提高解释准确率的增益。因此, 为 FastScore-CAM 选择 d 值的问题涉及性能和效率的平衡性问题。为此, 本文通过一项支持性实验进行了探讨。以 FastScore-CAM 解释 VGG-16 为例, 预设 d 的范围为 [40, 200], 间隔为 20, 进行实验并记录评价指标, 实验结果如表 4 所示。可

以看出, 随着 d 的增加, 各项指标整体趋势朝着更好的结果发展, 但变化幅度逐渐减小, 表明解释准确性随着扰动样本的增加而递增, 但增加样本数量的边际收益逐渐递减。从当前实验环境看, 考虑 d 引起计算成本的增加, 设置 d 在 80~120 区间较为合适。

Table 4 The effect of setting the target dimension d on the result in FastScore-CAM

表 4 FastScore-CAM 方法中目标维度 d 设置对结果的影响										
目标维度 d	40	60	80	100	120	140	160	180	200	
评价指标										
AD ↓	41.30	40.61	39.47	39.65	39.91	39.26	38.78	38.56	39.11	
AI ↑	10.2	10.4	11.1	11.0	10.9	10.4	11.2	10.5	10.5	
Del ↓	7.26	7.14	7.18	7.03	6.88	6.99	6.95	6.88	6.88	
Ins ↑	42.69	42.88	43.19	43.29	43.31	43.67	43.56	43.76	43.50	

3.9 降维算法

主成分分析^[35] (Principal Component Analysis, PCA)、奇异值分解^[36] (Singular Value Decomposition, SVD) 等常见的降维算法也可用于将特征图映射到目标子空间。与 NMF 的区别在于, PCA 的原理是识别数据中最大方差的方向, 而 SVD 则通过保留较大的奇异值以实现图像压缩。为进一步验证 NMF 的优势, 本文对 3 种降维算法进行了严格控制变量的对照实验。实验中, 所有算法设置均相同, 由于 PCA 和 SVD 要求 d 必须严格小于 c 和 $h \times w$, 因此设置 d 为 30。实验结果如表 5 所示, PCA 和 SVD 均导致 AD 显著升高 (超过 76), 同时 AI 显著降低 (降至 2.20), Del 和 Ins 结果也较差, 表明 PCA 和 SVD 不适用于 FastScore-CAM。

Table 5 The effect of different dimensionality reduction algorithms on performance

表 5 不同降维算法对性能的影响				
降维算法	AD ↓	AI ↑	Del ↓	Ins ↑
NMF	41.47	9.70	7.58	42.66
PCA	76.12	2.20	12.70	18.13
SVD	76.32	2.20	12.78	17.78

4 结语

本文提出了一种改进 Score-CAM 的视觉解释方法——FastScore-CAM, 在保持 Score-CAM 的可解释性的同时获得显著的计算效率增益。具体地, 通过 NMF 降低通道维度和减少冗余特征以显著降低扰动样本数量, 使线性决策边界的构建更加简洁和高效。定性结果表明, FastScore-CAM 生成的显著性图, 定位精确, 噪声较少。定量评估表明, FastScore-CAM 提供了与 Score-CAM 相当质量的解释, 同时显著降低了计算成本。此外, 实验结果验证了如何平衡解释性能和运行效率, 以及降维算法对性能的影响。然而, FastScore-CAM 的整体算法逻辑上仍未完全突破 Score-CAM 的框架限制, 因此在解释准确率要求更高的应用场景中, 其性能提升空间有限。从当前设计看, 目标维度的选择仍依赖于经验性方法, 若能开发一种自适应

机制,根据输入特性动态确定最优目标维度,则有望进一步提升解释准确率。

参考文献:

- [1] KIM B, DOSHI-VELEZ F. Interpretable machine learning: the fuss, the concrete and the questions [C]//Sydney: The Thirty-Fourth International Conference on Machine Learning, 2017.
- [2] ITTI L, KOCH C, NIEBUR E. A model of saliency-based visual attention for rapid scene analysis [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(11): 1254-1259.
- [3] HSU C Y, LI W W. Explainable GeoAI: can saliency maps help interpret artificial intelligence's learning process? an empirical study on natural feature detection [J]. International Journal of Geographical Information Science, 2023, 37(5): 963-987.
- [4] SEYED MOJTABA M Z, DEVIN G, NILANJAN R, et al. Crown-CAM: interpretable visual explanations for tree crown detection in aerial images [J]. IEEE Geoscience and Remote Sensing Letters, 2023, 20: 1-5.
- [5] BACH S, BINDER A, MONTAVON G, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation [J]. PLoS One, 2015, 10(7): 0130140.
- [6] 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, 2017: 618-626.
- [7] CHATTOPADHYAY A, SARKAR A, HOWLADER P, et al. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks [C]//2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018: 839-847.
- [8] PETSUK V, DAS A, SAENKO K. Rise: randomized input sampling for explanation of black-box models [DB/OL]. <https://arxiv.org/abs/1806.07421>.
- [9] ZHANG Q L, RAO L, YANG Y B. Group-CAM: group score-weighted visual explanations for deep convolutional networks [DB/OL]. <https://arxiv.org/abs/2103.13859>.
- [10] ZENG C Y, YAN K, WANG Z F, et al. Abs-CAM: a gradient optimization interpretable approach for explanation of convolutional neural networks [J]. Signal, Image and Video Processing, 2023, 17: 1069-1076.
- [11] JOSHUA R L, LUCIUS E B, SAKINA H. Causal dependence plots for interpretable machine learning [DB/OL]. <https://arxiv.org/abs/2303.04209>.
- [12] WANG H F, WANG Z F, DU M N, et al. Score-CAM: score-weighted visual explanations for convolutional neural networks [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020: 24-25.
- [13] DANIEL D L, SEBASTIAN S H. Learning the parts of objects by non-negative matrix factorization [J]. Nature, 1999, 401: 788-791.
- [14] ZHOU B, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2921-2929.
- [15] CHATTOPADHYAY A, SARKAR A, HOWLADER P, et al. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks [C]//Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018: 839-847.
- [16] FU R G, HU Q Y, DONG X H, et al. Axiom-based grad-cam: towards accurate visualization and explanation of cnns [DB/OL]. <https://arxiv.org/abs/2008.02312>.
- [17] PATRICK S, WALTER K. Cxplain: causal explanations for model interpretation under uncertainty [DB/OL]. <https://arxiv.org/abs/1910.12336>.
- [18] DESAI S, RAMASWAMY H G. Ablation-CAM: visual explanations for deep convolutional network via gradient-free localization [C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020: 983-991.
- [19] LI H, LI Z H, MA R, et al. FD-CAM: improving faithfulness and discriminability of visual explanation for CNNs [C]//2022 26th International Conference on Pattern Recognition (ICPR), 2022: 1300-1306.
- [20] OLIVER N, PÉREZ-CRUZ F, KRAMER S, et al. Causal explanation of convolutional neural networks [C]//Proceedings of European Conference on Machine Learning and Knowledge Discovery in Databases, 2021: 633-649.
- [21] SEOK Y B, LEE W. Recipro-CAM: gradient-free reciprocal class activation map [DB/OL]. <https://arxiv.org/abs/2209.14074>.
- [22] ENGLEBERT A, CORNU O, DE VLEESCHOUWER C. Poly-CAM: high resolution class activation map for convolutional neural networks [J]. Machine Vision and Applications, 2024, 35(4): 89.
- [23] BEECHY D, SMITH T M S, ŞİMŞEK Ö, et al. Explaining reinforcement learning with shapley values [C]//International Conference on Machine Learning, 2023: 2003-2014.
- [24] ZHENG Q, WANG Z, ZHOU J, et al. Shap-CAM: visual explanations for convolutional neural networks based on Shapley value [C]//Proceedings of European Conference on Computer Vision, 2022: 459-474.
- [25] FENG Z P, JI H B, MILOS D, et al. Cluster-CAM: cluster-weighted visual interpretation of CNNs' decision in image classification [DB/OL]. <https://arxiv.org/abs/2302.01642>.
- [26] RUDIN C, CHEN C F, CHEN Z, et al. Interpretable machine learning: fundamental principles and 10 grand challenges [DB/OL]. <https://arxiv.org/abs/2103.11251>.
- [27] JONATHAN L, EVAN S, TREVOR D. Fully convolutional networks for semantic segmentation [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 3431-3440.
- [28] LAURENS V M, GEOFFREY H. Visualizing data using t-SNE [J]. Journal of Machine Learning Research, 2008, 9: 2579-2605.
- [29] JAMES V S. Independent component analysis: a tutorial introduction [M]. Cambridge: MIT Press, 2004.
- [30] DENG J, DONG W, RICHARD S, et al. ImageNET: a large-scale hierarchical image database [C]//2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009: 248-255.
- [31] KAREN S, ANDREW Z. Very deep convolutional networks for large-scale image recognition [DB/OL]. <https://arxiv.org/abs/1409.1556>.
- [32] HUANG G, LIU Z, LAURENS V D M. Densely connected convolutional networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4700-4708.
- [33] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.
- [34] JIANG P T, ZHANG C B, HOU Q B, et al. LayerCAM: exploring hierarchical class activation maps for localization [J]. IEEE Transactions on Image Processing, 2021, 30: 5875-5888.
- [35] MARKUS R. What is principal component analysis? [J]. Nature Biotechnology, 2008, 26: 303-304.
- [36] GILBERT W S. On the early history of the singular value decomposition [J]. SIAM Review, 1993, 35: 551-566.

(责任编辑:孙娟)