E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2025)07-2468-16

论文引用格式:Zhang J, Zhang Y R and Wang Z C. 2025. Spatial attention-based multilayer fusion method for high-quality class activation map. Journal of Image and Graphics, 30(7):2468-2483(张剑, 张一然, 王梓聪. 2025. 多层混合注意力机制的类激活图可解释性方法. 中国图象图形学报, 30(7):2468-2483)[DOI:10.11834/jig.240216]

多层混合注意力机制的类激活图可解释性方法

张剑,张一然*,王梓聪

武汉数字工程研究所,武汉　430205

摘　要:**目的**　深度卷积神经网络在视觉任务中的广泛应用,使得其作为黑盒模型的复杂性和不透明性引发了对决策机制的关注。类激活图已被证明能有效提升图像分类的可解释性从而提高决策机制的理解程度,但现有方法在高亮目标区域时,常存在边界模糊、范围过大和细粒度不足的问题。为此,提出一种多层混合注意力机制的类激活图方法(spatial attention-based multi-layer fusion for high-quality class activation maps,SAMLCAM),以优化这些局限性。**方法**　在以往的类激活图方法忽略了空间位置信息只关注通道级权重,降低了目标物体的定位性能,SAMLCAM方法提出一种结合了通道注意力机制和空间注意力机制的混合注意力机制,实现增强目标物体定位减少无效位置信息的效果。在得到有效物体定位结果后,根据神经网络多层卷积层的特点,改进多层特征图融合的方式提出多层加权融合机制,改善类激活图的边界效果范围过大和细粒度不足的问题,从而增强类激活图的视觉解释性。**结果**　引用广泛用于计算机视觉模型的基准测试ILSVRC 2012(ImageNet Large-scale Visual Recognition Challenge 2012)数据集和MS COCO2017(Microsoft common objects in context 2017)数据集,对提出方法在多种待解释卷积网络模型下进行评估,包括消融实验、定性评估和定量评估。消融实验证明了各模块的有效性;定性评估对其可解释性效果进行视觉直观展示,验证效果的提升;定量评估中数据表明,SAMLCAM在Loc1和Loc5指标性能相较于最低数据均有大于8%的提升,在能量定位决策指标相较于最低数据均有大于7%的提升。由于改进方法减少了目标样本区域的上下文背景区域,使其对结果置信度存在负影响,但在可信度指标中,与其他方法比较仍可以保持比较小的差距并维持较高性能。**结论**　本文方法在多种卷积神经网络架构上均展现出优异的解释性能,通过扩大目标样本区域的响应覆盖度并有效抑制背景或无关区域的响应,提升了可解释性结果的精确性与可靠性。

关键词:类激活图(CAM);可解释性;注意力机制;图像分类;卷积神经网络(CNN)

Spatial attention-based multilayer fusion method for high-quality class activation map

Zhang Jian, Zhang Yiran*, Wang Zicong

Wuhan Digital Engineering Institute, Wuhan　430205, China

Abstract: **Objective** The success of deep convolutional neural networks (DCNNs) in image classification, object detection, and semantic segmentation has revolutionized the field of artificial intelligence (AI). Models based on DCNNs have demonstrated exceptional accuracy and have been deployed in various real-world applications. However, a major drawback of DCNNs is their lack of interpretability, which is frequently referred to as the "black-box" problem. When a DCNN

收稿日期:2024-04-17;**修回日期**:2024-12-12;**预印本日期**:2024-12-19

* **通信作者**:张一然　y1random@163.com

makes a prediction, understanding how and why it arrived at that decision is a challenging task. This lack of transparency hinders our ability to trust and rely on these models' output, particularly in critical domains, such as healthcare, autonomous driving, and finance. In medical diagnosis, for example, comprehending the reasoning behind a model's diagnosis is crucial for healthcare professionals to make informed decisions about patient care. Explainable AI (XAI) aims to address this issue by providing human-interpretable explanations for decisions made by complex machine learning models. XAI seeks to bridge the gap between model performance and interpretability, allowing users to understand the inner workings of a model and have confidence in its output. Researchers have been actively developing techniques and methods for enhancing the interpretability of deep learning models. One approach is to generate visual explanations through techniques, such as class activation map (CAM), gradient-weighted CAM (Grad-CAM), and smooth Grad-CAM. These methods provide heat maps or attention maps that highlight the areas of an input image that influence the model's decision the most. By visualizing this information, users can gain insights into the features and patterns that the model focuses on when making predictions. Experimental evidence shows that CAM methods can effectively enhance the interpretability of image classification. However, existing methods can only provide rough range explanations and suffer from the issues of excessively large boundary effects and insufficient granularity. To address these problems, spatial attention-based multilayer fusion for high-quality CAM (SAMLCAM) is proposed. SAMLCAM combines channel attention and spatial attention mechanisms based on Gra'd-CAM. SAMLCAM achieves more effective object localization and enhances visual interpretability by addressing the issues of excessively large activation map boundaries and lack of fine granularity through multilayer fusion. **Method** In existing CAM methods, only the channel weights are considered, while beneficial information from spatial position, which contributes to target localization, is frequently overlooked. In our study, a hybrid attention mechanism that combines channel attention and spatial attention is proposed to enhance the interpretability of target localization. The spatial attention mechanism focuses on the spatial relationship among different regions in feature maps. By assigning higher weights to regions that are more likely to contain the target object, SAMLCAM can enhance the precision of object localization while reducing false positives. This attention mechanism allows the model to allocate more attention to discriminative features, improving object localization. One key improvement of SAMLCAM lies in its multilayer attention mechanism. Previous methods frequently suffer from boundary effects, wherein activation maps tend to have excessively large boundaries that may include irrelevant regions. SAMLCAM addresses this issue by refining attention maps at multiple layers of a network. It not only relies on the results from the final convolutional layer but also considers multiple aspects, including attention to shallow layers. This feature enriches the reference information, resulting in a more comprehensive understanding of the semantic information of the target object while reducing unnecessary background information. This multilayer attention mechanism helps gradually refine boundaries and improve localization accuracy by reducing the influence of irrelevant regions. Moreover, SAMLCAM deals with the problem of insufficient granularity in CAM. In some cases, activation maps generated using previously available methods lack fine details, making precisely identifying the object of interest a challenging task. SAMLCAM overcomes this limitation by leveraging the multilayer attention mechanism to capture more detailed information in activation maps, resulting in high-quality CAM with enhanced visual interpretability. The ImageNet Large-scale Visual Recognition Challenge (ILSVRC) 2012 dataset is a large-scale image classification dataset that consists of over a million labeled images from 1 000 different categories. It is widely used in benchmarking computer vision models. The evaluation results on the ILSVRC 2012 validation dataset demonstrate the effectiveness of SAMLCAM in improving object localization and energy localization decision metrics. The proposed method contributes to the field by offering a more comprehensive understanding of how deep models make decisions in visual tasks and provides insights into improving the interpretability of these models. The proposed SAMLCAM method is evaluated on five backbone convolutional network models by using the ILSVRC 2012 validation dataset and compared with five state-of-the-art saliency models, namely, Grad-CAM, Grad-CAM++, XGradCAM, ScoreCAM, and LayerCAM. The results demonstrate the performance improvement of SAMLCAM compared with the lowest-performing methods in the Loc1 and Loc5 metrics, with an increase of over 8%. In addition, when comparing energy localization decision metrics, SAMLCAM exhibits an improvement of more than 7% over the lowest-performing methods. Notably, the improved method reduces the contextual background areas that surround the target sample region, negatively affecting the confidence metric. However, in terms of the credibility metric,

SAMLCAM maintains relatively high performance with only a small gap compared with the other methods. In addition, we conduct a series of comparative experiments to clearly demonstrate the effectiveness of the fusion algorithm in the form of images. **Result** In conclusion, the SAMLCAM method presents a novel approach for enhancing the interpretability of DCNN models. By incorporating channel attention and spatial attention mechanisms, the proposed method improves object localization and overcomes the limitations of previous methods, such as excessive boundary effects and lack of fine granularity, in CAM. The evaluation results on the ILSVRC 2012 dataset highlight the performance improvement of SAMLCAM compared with other methods in terms of localization and energy localization decision metrics. The proposed method contributes to advancing the field of visual deep learning and offers valuable insights into understanding and improving the interpretability of black-box models. **Conclusion** The proposed method demonstrates superior explanatory performance across various convolutional neural network architectures by expanding the response coverage of target sample regions while effectively suppressing responses in background or irrelevant areas, thereby enhancing the precision and reliability of the interpretability results.

Key words: class activation map(CAM); explainability artificial intelligence; attention mechanism; image classification; convolutional neural network(CNN)

0 引 言

深度神经网络(deep neural networks, DNN)已经在计算机视觉(何伟和潘晨, 2022)和自然语言处理(Ali等, 2021)等领域广泛流传,深度神经网络相比传统机器学习算法不仅能够更加有效地处理分类和回归等问题,而且可以减少人工干预,增强泛化能力,实现物理世界的实际应用。然而,这些模型逐渐从最初人类可以理解其内部逻辑的智能算法转变为人类难以理解的黑盒模型(Guidotti等, 2019)。这导致在提升模型准确性的过程中失去了对计算过程的逻辑解释基础,降低了模型的透明度(Zhang等, 2021)。尽管这些深度学习算法在诸多分类、回归等任务中得到了很高的反馈结果,但由于其潜在内部机制可能存在与人类理解不同的情况,导致使用者对算法的结果存在质疑,降低算法的落地部署率(董胤蓬 等, 2022),因此提高智能模型可信度迫在眉睫(国家互联网信息办公室 等, 2021)。

在图像分类任务中,类激活图(class activation map, CAM)(Zhou等, 2016)是一种常用的可视化可解释性方法(Arrieta等, 2020; Linardatos等, 2020)。类激活映射方法以热力图的形式展示卷积神经网络(convolutional neural network, CNN)感兴趣的目标区域(窦慧 等, 2024),提供决策依据,辅助决策者理解模型(Li等, 2022)。但该方法需要修改原网络模型结构,为解决该问题, Selvaraju等人(2017)基于反向传播通过结合梯度信息和类激活图提出一种对CAM的改进方法Grad-CAM(gradient-weighted class activation map)。反向传播方法(化盈盈 等, 2020)是一种特定于模型的方法,其计算特定输出相对于输入的梯度,并使用反向传播来推导特征在模型推断过程中的贡献。同时Selvaraju等人(2017)还将导向反向传播(guided-backpropagation, GBP)算法与类激活映射相结合,提出Guided Grad-CAM算法,模型向后传播,通过特定任务的计算获得类别的原始分数,然后将该信号反向传播到卷积特征图以定位物体位置,使用ReLU(rectified linear unit)函数进行激活提取正向信息,得到高分辨率具有特定概念的可视化结果。

Grad-CAM++(Chattopadhay等, 2018)认为梯度矩阵上的每一个元素的贡献不同,为了获得更好的效果,舍弃了Grad-CAM中梯度的全局平均值,引入了二阶梯度和三阶梯度。Omeiza等人(2019)提出Smooth Grad-CAM++方法,多次输入带有随机噪声的图像,并对结果求平均,用以消除输出显著图的噪声,随着噪声的减少类激活图会逐渐聚焦到目标物体区域。Fu等人(2020)引入了两个公理:敏感性(Dong等, 2019)和一致性(Fong和Vedaldi, 2017),通过归一化的激活对梯度进行加权,提出XGradCAM(axiom-based Grad-CAM)方法。ScoreCAM(Wang等, 2020)基于扰动的方法和基于CAM通过每个激活映射在目标类上的前向传递分数来获得每个激活映射的权重,从而摆脱了对梯度的依赖,最终结果由权重和激活映射的线性组合得到,其时间复杂度很高。使用浅层的特征图产生类别激活图时,效果并不理想,因此Jiang等人(2021)提出LayerCAM,该方法收集所有层可靠的类激活图,为特征图的每个空间位置生成单独的权重,这个元素级别权重可以考

虑特征图中每个空间位置的作用。LPCAM（local prototype CAM）（Chen 和 Sun，2023）对对象类的所有局部特征进行聚类，然后进行类激活图显示。一般来说，在图像分类任务中，决策者倾向于使用类激活图的方法去进行结果展示，不仅可以反映结果，还可以使人们更容易理解，对深度网络有更好的解释性。

上述方法已经可以为决策者提供直观的可视化结果，使得能够直观地理解在决策时所关注的区域和特征，从而解释模型的决策。但仍存在以下 3 个关键问题：1）权衡解释性与性能之间的平衡，不能以失去过多的算法性能为代价提高解释性，例如算法计算效率；2）类激活图的完整性问题，在视觉图像分类问题中，需要其解释结果完全覆盖所有目标对象，不能存在缺失；3）类激活图的细节问题，在完整表达的同时需要抓取尽可能多的目标对象特征细节，而不是大致区域覆盖。

因此本文提出一种基于混合注意力机制的多层加权融合的类激活图算法 SAMLCAM（spatial attention-based multi-layer fusion for high-quality class activation maps）。首先进行消融实验，证明了模块设计的合理性，同时在两个数据集上分别进行定性和定量评估，定性评估展示了单目标可视化分析结果和多目标可视化分析结果；接着进行了定量的评估，为定性评估结果增加数据证明，其中包括弱监督定位、能量定位决策指标、可信度评估和延迟基准评估，从不同角度证明了本实验提出的改进方法可以带来更好的性能提升。

1 SAMLCAM 方法描述

1.1 混合注意力机制权重提取

传统类激活图方法大多只计算通道级权重，即通道注意力模块（Hu 等，2020）。该模块获取对各卷积层中的特征图通道上的全局权重，确定各个通道的重要性程度，分配通道角度的注意力资源，以达到对目标对象特征的注意力调整。但仍存在一个重要限制，即缺乏对空间注意力的关注。这意味着这些方法主要关注图像中的某些区域，而忽略了不同区域的空间分布。在图像中，不同区域的空间布局对于理解模型的决策过程和对目标的解释至关重要。设模型 F 是一个图像 n 分类模型，对输入图像 $\boldsymbol{x}$ 进行分类，获得分类结果对应置信度分数 $y^1, y^2, \cdots, y^c, \cdots, y^n$，感兴趣类别为 c，计算 $\boldsymbol{x}$ 对于类别 c 的通道级权重矩阵，具体为

$$\boldsymbol{w}_{lk}^{c} = \frac{1}{I \times J} \sum_{i} \sum_{j} ReLU\left(\frac{\partial y^{c}}{\partial \boldsymbol{A}_{ij}^{lk}} \right) \tag{1}$$

式中，$ReLU$ 为线性整流函数，$I \times J$ 表示特征图大小。i 和 j 表示特征图位置坐标 (i,j)，后续变量中的下标 i 和 j 与这里的代表 (i,j) 位置处意思相同。$\boldsymbol{A}_{ij}^{lk}$ 则表示在第 l 层第 k 个特征图中 (i,j) 位置处的特征值，$\boldsymbol{w}_{lk}^{c}$ 表示目标类别为 c 时的第 l 层第 k 个通道的权重值。

对通道维上（特征图的各个位置）进行数据处理，计算空间权重矩阵。具体为

$$\boldsymbol{w}_{lij}^{c} = \frac{1}{k} \sum_{k} ReLU\left(\frac{\partial y^{c}}{\partial \boldsymbol{A}_{ij}^{lk}} \right) \tag{2}$$

式中，$\boldsymbol{w}_{lij}^{c}$ 表示目标类别为 c 时第 l 层 (i,j) 位置处的权重值。

如图 1 所示，采取并行的方式对空间权重矩阵和通道权重矩阵进行点积计算，获得最终的综合权重矩阵，具体为

$$\boldsymbol{w}_{l}^{c} = \boldsymbol{w}_{lij}^{c} \odot \boldsymbol{w}_{lk}^{c} \tag{3}$$

式中，$\boldsymbol{w}_{l}^{c}$ 表示目标类别 c 时第 l 层的综合权重值，$\odot$ 表示哈达玛积。

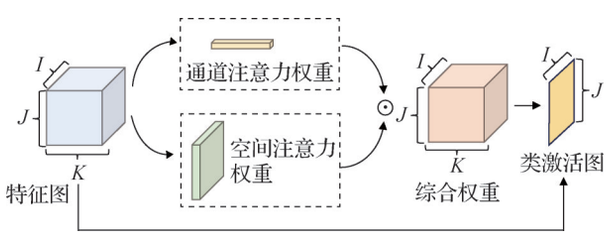

图 1 混合注意力机制

Fig. 1 Hybrid attention mechanism

1.2 多层加权融合机制

目前，已有一些研究针对目标物体覆盖区域不完整的问题展开，例如 LayerCAM。其方法如图 2 所示，通过将注意力聚焦于融合所有层的解释图上，尝试获取更精细的类激活图。具体而言，LayerCAM 通过正梯度权重与特征图中各位置的激活值相乘来生成细粒度激活图。

然而，LayerCAM 最终对所有层的解释图取极限最大值，这导致浅层非目标物体的特征也被引入，进而生成大量噪声，具体为

$$\boldsymbol{M}^{c\prime} = \max_{l} \left(\hat{\boldsymbol{M}}_{l}^{c} \right) \tag{4}$$

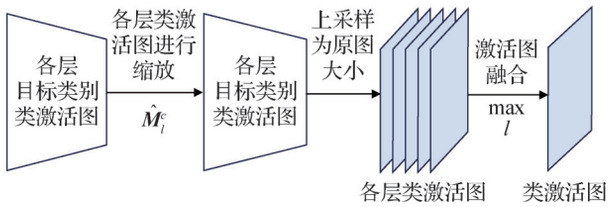

图2　LayerCAM流程图

Fig. 2　Flow diagram of LayerCAM

式中，$\hat{\boldsymbol{M}}_l^c$表示第l层卷积层目标类别c的类激活值，$\boldsymbol{M}^{c\prime}$表示取各层每个位置的最大值作为最后的类激活值。

通过上述分析可见，现有的类激活图方法虽然能够高亮目标区域并提取目标样本的特征，但仍存在两个主要问题：1）目标物体的特征覆盖不完整，存在特征丢失；2）目标区域外的噪声较多，干扰物体识别。具体的实验分析将在第2节展示。

如图3所示，通过特征图实验验证在第1层的类激活图中，可以清晰地看到大象的轮廓，而随着层次的加深，轮廓逐渐变得模糊，更多的信息集中在局部特征上。这表明，浅层类激活图能够提供更精确的位置信息，而深层类激活图则包含更丰富的语义信息。不同层次的类激活图融合有助于提升解释的完整性。此外，实验还发现，浅层类激活图的激活值明显低于深层类激活图，因此，单纯对各层类激活图进行线性组合并不能提升性能，必须采用加权组合策略。

在本实验中，当组合来自不同层的类激活图时，首先通过缩放函数缩放浅层的类激活图，其中缩放的激活图记做$\hat{\boldsymbol{M}}_l^c$。具体为

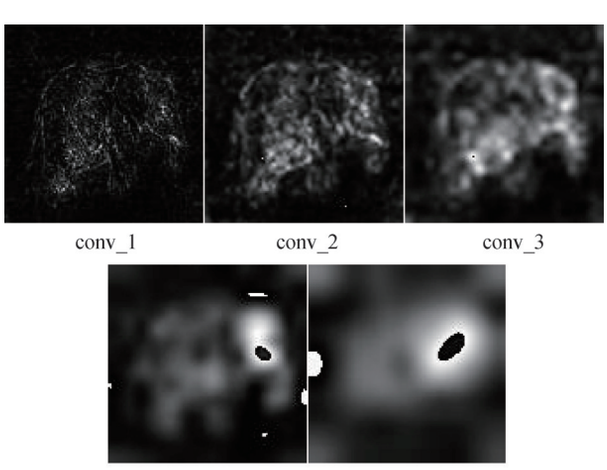

图3　各卷积层的特征图

Fig. 3　Feature maps of each convolutional layer

$$\hat{\boldsymbol{M}}_l^c = \tanh\left(\frac{\gamma \boldsymbol{M}_l^c}{\max\left(\boldsymbol{M}_l^c\right)}\right) \tag{5}$$

式中，γ为缩放因子，$\boldsymbol{M}_l^c$和$\hat{\boldsymbol{M}}_l^c$表示缩放前后第l层目标类别c的类激活值。

将浅层的类激活值放大至与深层的类激活值同一级别，然后采用多层类激活图均值化处理，获取激活图中的全局重要性，浅层激活图与深层激活图之间相互补充，输出神经元激活的视觉特征，计算权重矩阵$\boldsymbol{M}^{c\prime}$，具体为

$$\boldsymbol{M}^{c\prime} = \underset{l}{\text{mean}}\left(\hat{\boldsymbol{M}}_l^c\right) \tag{6}$$

然而，由于浅层类激活图保留了非目标类别的干扰信息，包含大量噪声点，这些噪声在缩放过程中未被有效抑制，反而在接近1时被放大，进而影响最终融合结果，导致激活图中出现噪声点。因此通过类激活图内激活值的关系，分别对各(i,j)位置进行约束，计算出约束矩阵$\boldsymbol{S}_{ij}^c$，具体为

$$\boldsymbol{S}_{ij}^c = \text{sign}\left(\boldsymbol{M}_{ij}^{c\prime} - \lambda \max_{i,j}\left(\boldsymbol{M}_{ij}^{c\prime}\right)\right) \tag{7}$$

式中，λ为范围因子，用于抑制不相关神经元的激活，减少非目标区域对感兴趣目标类别的噪声干扰，并增强目标区域的响应，特别是对浅层特征的约束效果。

$\boldsymbol{M}_{\text{final}}^c$为本文方法的最终目标类别$c$的类激活值，具体为

$$\boldsymbol{M}_{\text{final}}^c = \eta \boldsymbol{S}_{ij}^c \odot \boldsymbol{M}_{ij}^{c\prime} \tag{8}$$

式中，η为放大参数（通常设为1.2）。

1.3　整体流程

结合卷积网络的特点，本文提出一种基于多层注意力机制的多层加权融合的可解释类激活图方法（SAMLCAM）算法，整体设计思路如图4所示，具体流程如以下算法所示。

算法1　SAMLCAM类激活图生成算法

输入：待解释图像n分类模型F，对输入图像$\boldsymbol{x}$进行分类，输入图像$\boldsymbol{x} \in \mathbf{R}^{h \times w}$，目标类别$c$。

输出：对目标类别c生成类激活图解释结果。

1）将样本$\boldsymbol{x}$输入模型F，返回类别c的分类得分y^c，获得各层的特征$\boldsymbol{A}^l$；

2）确定模型F的目标层$target_layers=\{layer1, layer2, layer3\}$；

3）for l in target_layers do

4）$\boldsymbol{g}_{ij}^{lkc}=\left(\frac{\partial y^c}{\partial \boldsymbol{A}_{ij}^{lk}}\right)$　#计算目标层的梯度；

5）end for

6）for g in $\boldsymbol{g}_{ij}^{lkc}$ do　#计算综合权重

7）$\boldsymbol{w}_{lk}^c=\frac{1}{I\times J}\sum_i\sum_j ReLU\left(\boldsymbol{g}_{ij}^{lkc}\right)$　#计算通道权重矩阵；

8）$\boldsymbol{w}_{lij}^c=\frac{1}{k}\sum_k ReLU\left(\frac{\partial y^c}{\partial \boldsymbol{A}_{ij}^{lk}}\right)$　#计算空间权重矩阵；

9）$\boldsymbol{w}_l^c=\boldsymbol{w}_{lij}^c\odot\boldsymbol{w}_{lk}^c$　#计算综合权重矩阵；

10）end for

11）for w in $\boldsymbol{w}_l^c$ do

12）$\boldsymbol{L}^{lc}=ReLU\left(\sum \boldsymbol{w}_{lkij}^c\boldsymbol{A}_{ij}^{lk}\right)$　#计算各层激活图；

13）$\boldsymbol{L}^{lc\prime}=\frac{\boldsymbol{L}^{lc}-\boldsymbol{L}^{lc}{}_{\min}}{\boldsymbol{L}^{lc}{}_{\max}-\boldsymbol{L}^{lc}{}_{\min}}$　#min-max 归一化；

14）end for；

15）for $\boldsymbol{M}_l^c$ in $\boldsymbol{L}^{lc\prime}$ do

16）$\hat{\boldsymbol{M}}_l^c=\tanh\left(\frac{\gamma\boldsymbol{M}_l^c}{\max\left(\boldsymbol{M}_l^c\right)}\right)$　#将激活值进行缩放；

17）将激活图上采样至 $h\times w$；

18）end for；

19）$\boldsymbol{M}^{c\prime}=\underset{l}{\text{mean}}\left(\hat{\boldsymbol{M}}_l^c\right)$　#计算权重矩阵；

20）$\boldsymbol{S}_{ij}^c=\text{sign}\left(\boldsymbol{M}_{ij}^{c\prime}-\lambda\max_{i,j}\left(\boldsymbol{M}_{ij}^{c\prime}\right)\right)$　#计算约束矩阵；

21）$\boldsymbol{M}_{\text{final}}^c=\eta\boldsymbol{S}_{ij}^c\odot\boldsymbol{M}_{ij}^{c\prime}$　#计算最终激活图；

22）return $\boldsymbol{M}_{\text{final}}^c$

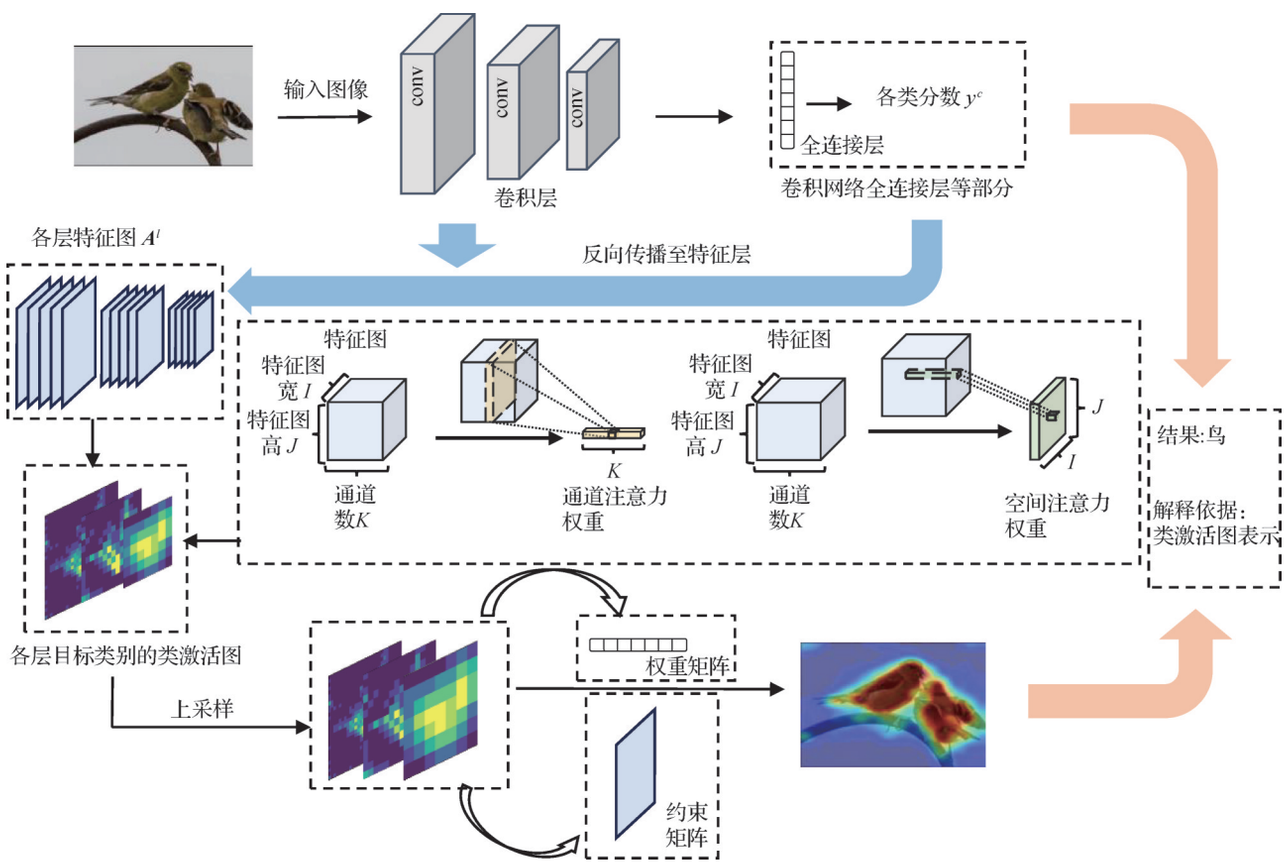

图 4　SAMLCAM 整体设计

Fig. 4　The overall design of SAMLCAM

2　验　证

2.1　数据集

为了评估类激活图方法的性能，选择计算机视觉领域具有权威度的两个图像数据集：ILSVRC2012（ImageNet Large-scale Visual Recognition Challenge 2012）（Krizhevsky 等，2012）数据集和 MS COCO2017（Microsoft common objects in context 2017）（Lin 等，2014）数据集进行实验。ILSVRC2012 数据集可以用

于图像分类、目标定位和目标检测等视觉任务，其中包括1 000个类别，约120万幅训练图像、5万幅验证图像和10万幅测试图像。在本实验中，选取ILSVRC2012验证集的2 000幅图像进行定量指标的评估实验。MS COCO2017数据集可以用于目标检测、物体分割等视觉任务，主要从复杂的日常场景中截取，包括80个对象类别，118 287幅训练集图像、5 000幅验证集图像和40 670幅测试图像。在本实验中，随机选取MS COCO2017验证集图像进行定性指标的评估实验。

2.2 实验环境配置

实验的环境配置如表1所示。

表1 环境配置

Table 1 Environment configuration

实验环境	环境配置
CPU	Intel(R) Xeon(R) Gold 6230 CPU@2.10 GHz
GPU	NVIDIA GeForce GTX 3090Ti
操作系统	Ubuntu 18.04.6 LTSx86_64
Python版本	Python3.7.5
深度学习框架	PyTorch1.13.1
CUDA版本	CUDA11.1

2.3 实现细节

本实验选用VGG19（Visual Geometry Group19）（Simonyan 和 Zisserman，2014）、ResNet34（residual network34）（He等，2016）和DenseNet201（Huang等，2017）3种经典卷积网络作为待解释的主干网络，这些预训练模型均由PyTorch中的torchvision库提供。理论上，图像分类模型的准确率越高，意味着模型能够更有效地捕捉与分类结果相关的关键特征，并更准确地定位感兴趣类别的重要特征。因此，通过可视化解释方法生成的类激活图能够更直观有效地测试模型在特征定位方面的能力。

在将测试图像输入分类网络前，需要进行数据预处理。将图像尺寸统一调整为3 × 224 × 224，并对图像进行归一化处理，将像素值映射到[0，1]区间内，并按通道进行标准化，以加速模型收敛。

在对比实验中，将图像分类模型输出中最大置信度对应的类别视为感兴趣目标类别，基于该类别进行目标物体识别的解释分析。其中，Grad-CAM、Grad-CAM++、XGradCAM、SmoothGradCAM++、ScoreCAM、ShapCAM和LPCAM选用最后一层的最后一个卷积层作为目标层生成激活图，其余模型的目标层都选用最后3个阶段的最后一个卷积层作为目标层生成激活图，双曲正切函数作为浅层激活图的缩放函数。

2.4 消融实验

为研究混合注意力机制模块在本文算法中的影响，在ILSVRC2012数据集上进行消融实验，分别在不同的待解释卷积网络上进行了Loc1（location 1）、Loc5和EBPG（energy-based pointing game）指标的数据对比。其中分别对GradCAM进行混合注意力机制和多层加权融合机制的加强版实验，SACAM为GradCAM的混合注意力机制模块加强实验，MLCAM进行了加多层加权融合机制模块的加强版实验，SAMLCAM是混合注意力机制和多层加权融合机制两者的加强版最终实验。

消融实验的结果如表2所示。数据表明，混合注意力机制模块和多层加权融合机制模块均可以对算法效果进行增强。具体来说，如表2中的加粗数据显示，在同一待解释卷积网络下，SACAM可以对GradCAM实现在Loc1、Loc5和EBPG指标不同程度的提升，具体地，在ResNet34网络下Loc1指标提升至49.96%，Loc5指标提升至62.75%，EBPG提升至45.65%，同样在VGG19和DenseNet201网络下也有不同程度的提升；相比MLCAM，SAMLCAM也分别在不同指标上对MLCAM有一定的效果提升，其中在ResNet34网络下，Loc1指标提升至55.65%，Loc5指标提升至69.25%，EBPG提升至51.36%，同样在VGG19和DenseNet201网络下也有不同程度的提升。多层加权融合机制更多是增加对目标范围更加准确的定位和轮廓区分，在数据上会有显著提高，而混合注意力机制更多是对细节方面的提升，在数据上会有小幅度提升，同时在可视化方面会看到细节上效果的提升，提高用户使用体验。实验结果可以说明，混合注意模块和多层加权融合机制模块能够提高目标对象的可解释性。

2.5 定性评估

在定性评估中，选用卷积框架ResNet34作为主干框架进行测试，并从MS COCO2017验证集中随机抽取图像进行可视化，将SAMLCAM分别与基于梯度和分数的类激活图可视化方法进行对比，其中包括基于梯度的类激活图方法Grad-CAM、Grad-

表 2　SAMLCAM 的消融实验对比结果

Table 2　Comparison results of ablation experiments for SAMLCAM

/%

卷积网络	混合注意力机制	多层加权融合机制	可解释方法	Loc1	Loc5	EBPG
ResNet34	–	–	GradCAM(Selvaraju 等,2017)	49.87	62.13	45.62
	√	–	SACAM	49.96	62.75	45.65
	–	√	MLCAM	55.01	68.36	50.99
	√	√	SAMLCAM	**55.65**	**69.25**	**51.36**
VGG19	–	–	GradCAM(Selvaraju 等,2017)	34.24	43.52	43.45
	√	–	SACAM	46.25	59.72	47.75
	–	√	MLCAM	45.62	59.47	56.21
	√	√	SAMLCAM	**49.62**	**64.04**	**56.70**
DenseNet201	–	–	GradCAM(Selvaraju 等,2017)	54.96	66.07	47.44
	√	–	SACAM	55.15	66.39	47.74
	–	√	MLCAM	56.54	68.03	50.43
	√	√	SAMLCAM	**56.67**	**68.17**	**50.58**

注:加粗字体数据为相应的对比实验中的最优结果。“√”和“–”分别表示使用和未使用对应机制。

CAM++、XGradCAM、SmoothGradCAM++、LayerCAM 和基于分数的类激活图方法 ScoreCAM,直观地比较不同可视化方法生成的目标物体类激活图效果。

2.5.1　单目标图像可视化分析

图 5 为单目标图像不同类激活图方法的可视化结果。可视化结果中,第 1 行为输入图像和对应设置的目标类别,左侧显示的是对应行使用的类激活图方法。GradCAM、Grad-CAM++、XGradCAM 和 ScoreCAM 这 4 种方法未能综合多个卷积层的特征来提取感兴趣目标的类激活图。因此,其生成的类激活图中,热力图所覆盖的区域较为分散,且目标类别的特征存在覆盖不完整的问题,例如,在第 3 列的“garbage truck”示例中,车头部分未被覆盖,且对目标对象的覆盖多呈块状分布,仅显示了目标的大致位置区域。SmoothGradCAM++虽然通过平滑操作增强了对目标物体的覆盖率,但仍未能突出显示目标的清晰轮廓。此外,平滑处理对边缘的影响导致额外的背景信息被引入,如第 1 列的“airliner”示例中,目标物体的轮廓边缘包含了更多无关背景信息。LayerCAM 方法融合了多个卷积层的类激活图,效果有所提升但目标物体的轮廓区分仍不清晰,覆盖区域不连续呈现点状分布,并且目标区域外的噪声点增多,影响目标物体的直观展示。相比之下,SAMLCAM 能够准确突出感兴趣类别的所有相关区域,并清晰展现目标物体与背景信息之间的边界。

2.5.2　多目标图像可视化分析

多目标可视化旨在处理目标类别在图像中多次出现的情况,不仅要求准确定位目标物体,还需同时突出显示所有出现的目标。然而,现有的类激活映射可视化方法在同时高亮多个目标对象时面临较大挑战,往往无法全面展示支撑网络分类结果的所有关键特征。为验证本文提出方法在多目标可视化方面的优势,进行了多种方法的对比实验并进行分析。图 6 展示的是多目标图像多种类激活图方法的可视化结果。GradCAM、Grad-CAM++和 XGradCAM 在目标区域定位上表现相似,整体存在目标缺失的问题,无法全面突出所有目标物体。ScoreCAM 在高亮目标物体时,常出现仅标注单个目标的情况,例如,第 4 列仅标注了右上角的“zebra”,第 5 列仅标注了左侧的“sorrel”。虽然 SmoothGradCAM++和 LayerCAM 在多目标覆盖方面性能有所提升,但仍存在不足。例如,第 2 列对“bagel”的覆盖不够充分,第 3 列和第 4 列则出现多个目标物体之间粘连现象,影响定位精度。对于多物体目标识别的可视化,SAMLCAM 能够有效覆盖多个目标物体的区域,并显著减少目标物体之间的粘连现象,从而清晰地分辨出每个目标物体的独立区域,显著提高对不同目标物体的分辨能力。

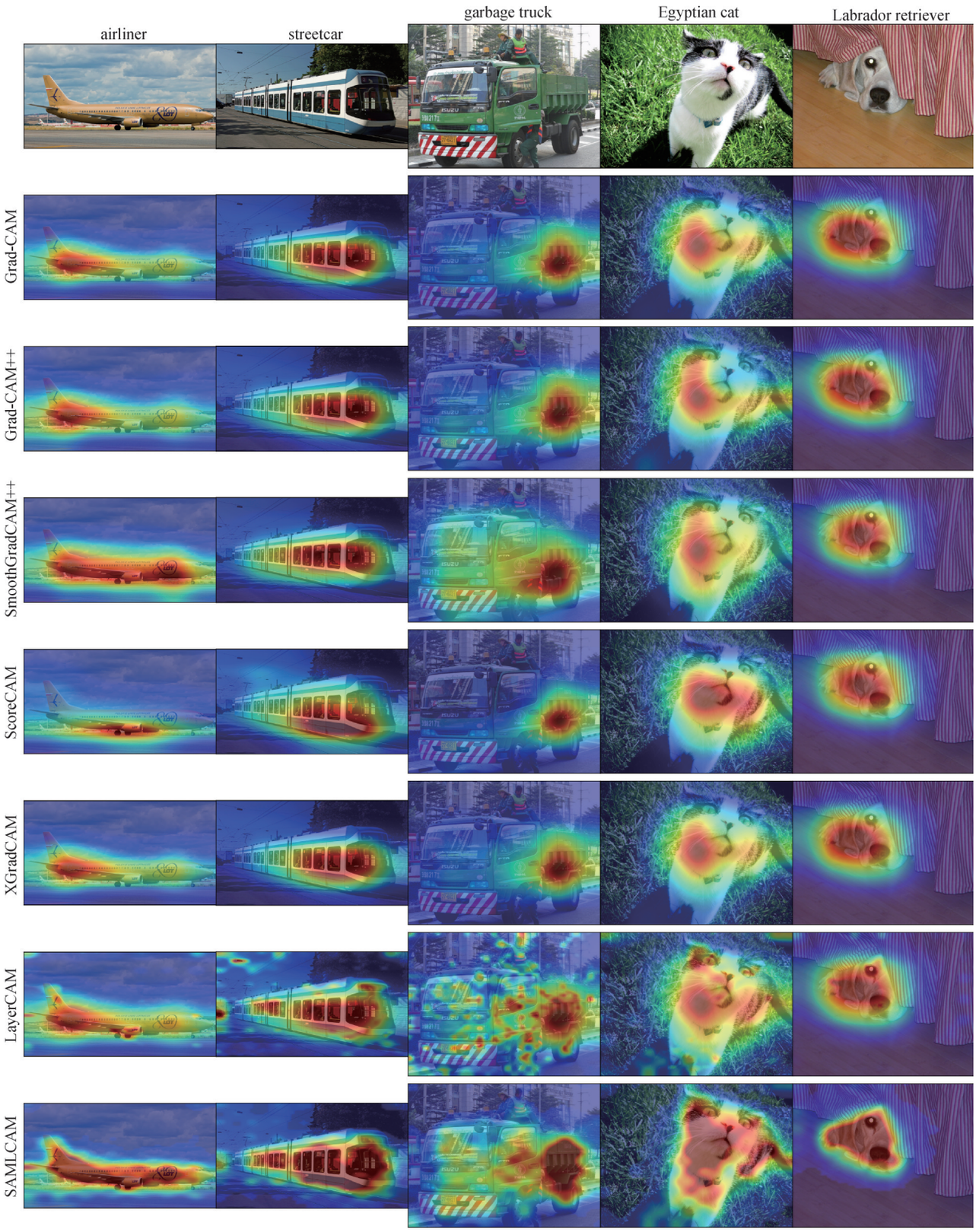

图5　单目标图像可视化分析

Fig. 5　Visual analysis of single target images

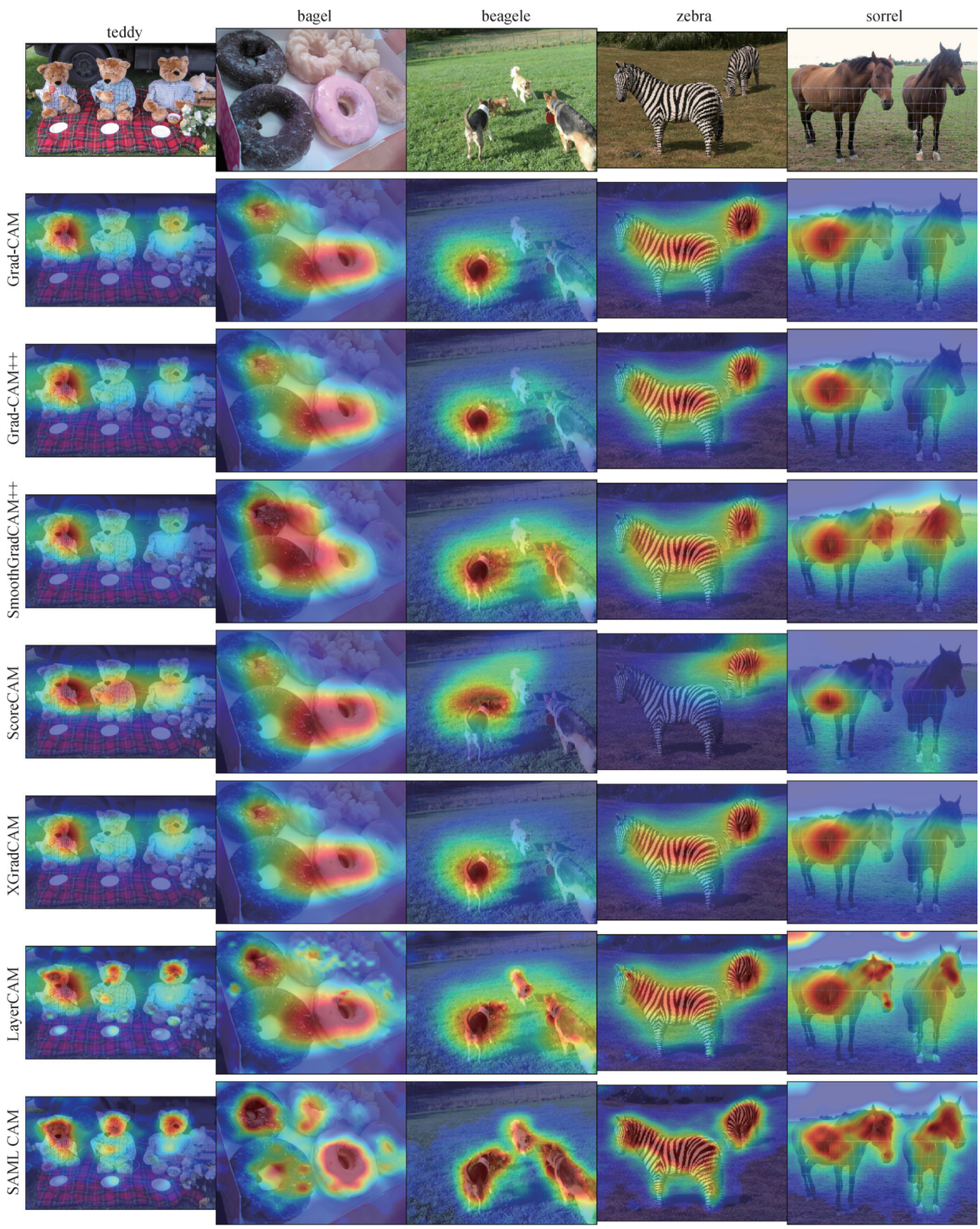

图 6　多目标图像可视化分析

Fig. 6　Visual analysis of multi-objective images

2.6 定量评估

2.6.1 弱监督定位能力

本实验使用图像定位任务的Loc-K指标来评估生成的类激活图是否可以准确地突出显示感兴趣的目标对象。计算与图像的真实标注框之间的重叠程度IoU(intersection over union),根据分类预测结果,选取置信度最高的1个类别,当预测的边界框与选取的目标类别真实边界框大于0.5时,认为该类激活图可视化的解释结果成功地进行了弱监督目标定位,计算预测解释结果成功的占比,即Loc1指标(Loc5指标为取置信度最高的5个类别)。

根据表3中的数据显示,本文算法比其他算法有更可靠的细粒度目标定位信息。对可解释性方法分别在不同的卷积网络上进行了测试,在ResNet34网络中,SAMLCAM的Loc1和Loc5指标可以提升至55.65%和69.25%,相较于其他方法的Loc1有3.84%～10.8%的提升,Loc5有4.97%～12.98%的提升;在VGG19网络中,SAMLCAM的Loc1和Loc5指标可以提升至49.62%和64.04%,相较于其他方法Loc1有2.07%～15.38%的提升,Loc5有3.82%～20.52%的提升;在DenseNet201网络中,SAMLCAM的Loc1和Loc5指标可以提升至56.67%和68.17%,相较于其他方法Loc1有1.52%～9.02%的提升,Loc5有1.78%～9.85%的提升。总体可以看出,SAMLCAM在对不同卷积网络中的目标对象定位方面均有所提升,可以通过对目标物体定位效果的提升来表示对目标物体区域的有效解释,帮助使用者找到目标物体,促进理解决策结果。

表3 SAMLCAM的弱监督定位对比结果

Table 3 Comparison results of weakly supervised localization of SAMLCAM

/%

可解释性方法	待解释卷积网络					
	ResNet34		VGG19		DenseNet201	
	Loc1	Loc5	Loc1	Loc5	Loc1	Loc5
GradCAM(Selvaraju等,2017)	49.87	62.13	34.24	43.52	54.96	66.07
GradCAM++(Chattopadhay等,2018)	50.51	62.90	43.84	56.16	53.30	64.42
SmoothGradCAM++(Omeiza等,2019)	51.31	64.28	45.24	53.23	54.12	65.16
XGradCAM(Fu等,2020)	49.87	62.13	36.59	46.63	53.75	64.93
ScoreCAM(Wang等,2020)	44.85	56.27	41.11	52.48	47.65	58.32
LayerCAM(Jiang等,2021)	49.56	61.75	46.25	59.72	55.15	66.39
ShapCAM(Zheng等,2022)	50.11	62.33	41.31	53.42	54.81	65.23
LPCAM(Chen和Sun,2023)	50.28	63.71	47.55	60.22	55.05	62.39
PolyCAM(Englebert等,2024)	51.81	62.29	42.73	53.48	54.29	65.31
SAMLCAM(本文)	**55.65**	**69.25**	**49.62**	**64.04**	**56.67**	**68.17**

注:加粗字体表示各列最优结果。

2.6.2 能量定位决策能力

能量定位决策指标EBPG是一种体现对目标物体是否具有分辨能力的指标。该指标主要体现在不仅要对目标对象进行正确识别,还需要体现出对不同区域的重要性程度(能量)区别,其主要思想为类激活图内的目标对象的能量与类激活图内的所有能量值之比,具体计算为

$$f_{\mathrm{EBGP}}=\frac{\sum \boldsymbol{L}^{c}_{(i,j)\in \boldsymbol{B}_g}}{\sum \boldsymbol{L}^{c}_{(i,j)\in \boldsymbol{B}_g}+\sum \boldsymbol{L}^{c}_{(i,j)\notin \boldsymbol{B}_g}} \tag{9}$$

式中,$\sum \boldsymbol{L}^{c}_{(i,j)\in \boldsymbol{B}_g}$和$\sum \boldsymbol{L}^{c}_{(i,j)\notin \boldsymbol{B}_g}$分别表示位于目标类别真实标注框内和框外的能量总和。

表4展示了分别对待解释卷积网络ResNet34、VGG19和DenseNet201做出的类激活图可解释性方法的EBPG指标结果。对可解释性方法分别在不同的卷积网络上进行了测试,在ResNet34网络中,SAMLCAM可以将EBPG指标提升至51.36%;在VGG19网络中,可以将其提升至56.70%;在DenseNet201网络中,可以将其提升至50.58%。数据表明,SAMLCAM对不同的卷积网络都有提升的

表 4　SAMLCAM 的能量定位决策指标对比结果

Table 4　Comparison results of EBPG for SAMLCAM

/%

可解释性方法	待解释卷积网络		
	ResNet34	VGG19	DenseNet201
GradCAM(Selvaraju 等,2017)	45.62	43.45	47.44
GradCAM++(Chattopadhay 等,2018)	45.42	46.64	47.52
SmoothGradCAM++(Omeiza 等,2019)	46.23	48.56	48.63
XGradCAM(Fu 等,2020)	45.62	44.10	48.10
ScoreCAM(Wang 等,2020)	43.56	52.01	46.03
LayerCAM(Jiang 等,2021)	42.75	46.43	43.57
ShapCAM(Zheng 等,2022)	44.25	52.11	46.38
LPCAM(Chen 和 Sun,2023)	48.71	50.22	49.26
PolyCAM(Englebert 等,2024)	47.28	51.36	47.64
SAMLCAM(本文)	**51.36**	**56.70**	**50.58**

注:加粗字体表示各列最优结果。

效果,说明了 SAMLCAM 具有补充解释细节信息的能力,并且在卷积网络中具有一定的普适性,而且提升效果也说明 SAMLCAM 目标对象区域的能量值占总能量值较高,可以对目标对象区域有效地定位并且高能量突出显示,明显区分重要特征与无效特征区域。

2.6.3　可信度的评估

采用 Average Drop 和 Average Increase 作为可信度的判断指标,这两个指标的主要思想在于将类激活图的激活值与原始输入图像进行点乘得到解释图,达到对非目标区域的遮盖效果,并对图像进行了裁剪,使输入图像更加集中在目标对象区域,如图 7 所示,将处理后的解释图输入卷积网络进行分类,并观察目标物体的置信度结果。通过监测置信度的变化,评估解释结果对卷积网络决策的影响,从而验证类激活图的高亮区域是否能够作为网络分类决策的依据。其中,Average Drop 是对置信度下降情况的关注指标,Average Increase 针对的是置信度上升的情况,计算置信度上升的个数占比。

表 5 显示,使用 DenseNet201 作为待解释卷积网络时,SAMLCAM 的 Average Drop 降至 6.756%,是对比方法中最优的情况。在其他情况中,Average Drop 和 Average Increase 两种指标处于总体偏上水平并非最优,这是由于本实验方法可以有效减少非目标范围的背景语义信息,但当目标物体区域较小时,可能导致语义上下文的缺失,从而降低置信度。然而,整体而言,在提升定位能力和视觉直观效果的同时,这些指标相较于最优结果的差比较小,能够在保证算法性能提升的基础上维持较高的可信度。此部分是未来研究中需要进一步改进和优化的方向。

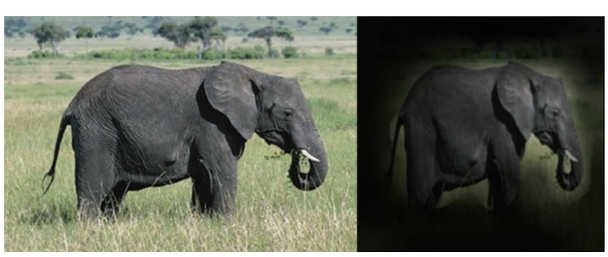

图 7　原图(左)与解释图(右)

Fig. 7　Original image (left) and interpretation image (right)

2.6.4　延迟基准

在某些特定的应用场景下,不仅要求可视化方法具有优异的准确定位效果,更需要满足不同任务的延时需求。本节测试依旧使用上述定量评估方法中的 ResNet34、VGG-16 和 DenseNet201 网络作为待解释卷积模型,从公开数据集 ImageNet ILSVRC 2012 的验证集中随机挑选了 2 000 幅样本图像,进行延迟基准评估。FPS(frames per second)用于评估模型在给定硬件上的处理速度,即每秒可以处理的

表5 SAMLCAM的可信度对比结果

Table 5 Results of the credibility comparison of SAMLCAM

/%

可解释性方法	待解释卷积网络					
	Average Drop			Average Increase		
	ResNet34	VGG19	DenseNet201	ResNet34	VGG19	DenseNet201
GradCAM(Selvaraju等,2017)	11.431	23.736	8.313	**35.769**	18.742	**42.058**
GradCAM++(Chattopadhay等,2018)	10.580	17.183	8.122	32.592	24.396	38.310
SmoothGradCAM++(Omeiza等,2019)	9.213	16.285	8.056	30.584	22.135	38.560
XGradCAM(Fu等,2020)	11.431	22.235	9.255	**35.769**	20.140	41.677
ScoreCAM(Wang等,2020)	10.430	17.460	8.663	32.719	21.220	37.675
LayerCAM(Jiang等,2021)	**8.645**	**13.160**	6.812	31.385	**24.778**	32.530
Shap-CAM(Zheng等,2022)	10.210	16.221	8.015	29.481	22.018	38.236
LPCAM(Chen和Sun,2023)	11.236	15.236	8.114	29.012	21.412	31.256
Poly-CAM(Englebert等,2024)	10.107	16.318	8.251	28.139	20.178	36.459
SAMLCAM(本文)	9.421	15.087	**6.756**	29.543	23.189	39.371

注:加粗字体表示各列最优结果。

图像数量。本实验采用FPS进行多算法的处理速度对比,在统计时间的起止范围取自图像处理后输入算法开始至输出结束。如表6所示,ScoreCAM、ShapCAM、LPCAM和PolyCAM处理速度明显较慢,本文算法仍可以达到与其他算法相近的较高处理速度,同时仍能够在不同程度上提升其他性能指标。

2.7 卷积网络错误分析

可解释性技术的重要作用之一是用来分析待解释网络的结果正确与否以及正误的原因,2.5节的定性评估和2.6节的定量评估都是对网络分类结果正确以及对正确结果提供决策依据的分析。而本节展示的则是卷积网络分类错误或卷积网络训练关注

表6 SAMLCAM的FPS比较结果

Table 6 Comparison results of FPS for SAMLCAM

/(帧/s)

可解释性方法	待解释卷积网络		
	ResNet34	VGG19	DenseNet201
GradCAM(Selvaraju等,2017)	**1.932**	**2.038**	**1.645**
GradCAM++(Chattopadhay等,2018)	1.844	1.954	1.556
SmoothGradCAM++(Omeiza等,2019)	1.851	1.902	1.561
XGradCAM(Fu等,2020)	1.866	1.919	1.555
ScoreCAM(Wang等,2020)	1.206	0.656	0.320
LayerCAM(Jiang等,2021)	1.845	1.892	1.537
ShapCAM(Zheng等,2022)	1.031	0.718	0.794
LPCAM(Chen和Sun,2023)	1.023	0.924	0.736
PolyCAM(Englebert等,2024)	0.958	0.618	0.481
SAMLCAM(本文)	1.856	1.923	1.550

注:加粗字体表示各列最优结果。

对象错误关注决策的情况。以 ResNet34 网络检测“crash helmet（防撞头盔）”和“parachute（降落伞）”为例，图 8 的第 1 列和第 2 列是 ResNet34 网络对目标对象错误关注的情况，图 8 的第 3 列和第 4 列是 ResNet34 网络错误分类及其错误关注原因。具体地，图 8 的第 1 列和第 2 列表明对于所设置的关注对象“crash helmet”，ResNet34 网络对其的决策关注部分为“摩托车身”部分，形成了错误关注，造成了关注目标对象与关注区域不匹配的情况；图 8 的第 3 列和第 4 列表明 ResNet 网络对图像分类结果是“parachute”并给出了错误决策原因，类激活图方法提供了错误结果的错误原因，错将突出显示部分分类为“parachute”。即该网络在训练过程中出现了部分偏差，也是分类网络在目标分类的正确率并非 100% 的原因。根据可解释方法提供的错误示例，可以针对性地对网络进行优化训练，提升正确分类结果，实现对待解释网络的正向优化作用。

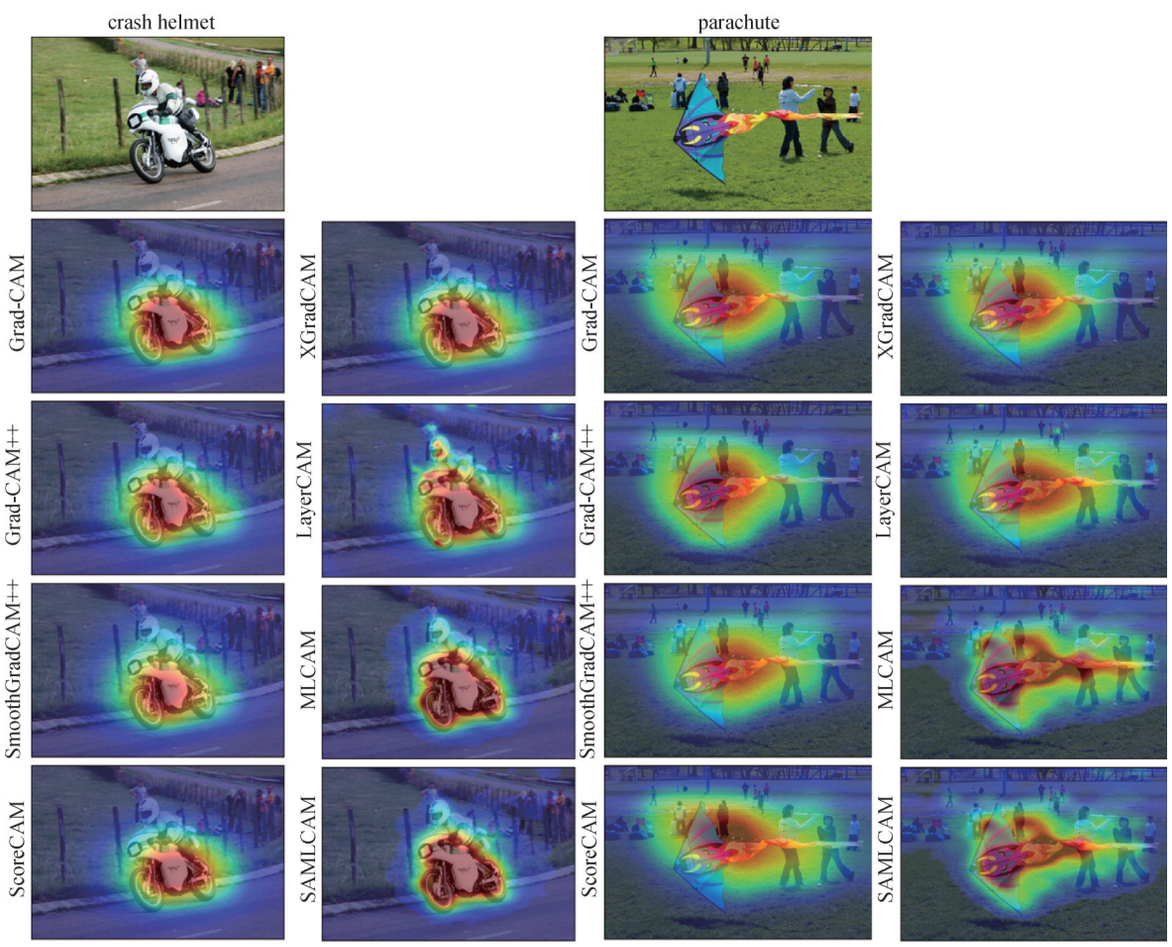

图 8　类激活图方法对卷积神经网络错误结果的分析展示

Fig. 8　Analysis of the error results of the CNN by the CAM

3　结　论

深度学习在图像领域的广泛应用推动了图像识别和分类技术的巨大进步，但由于深度学习模型的黑盒特性以及对抗样本的存在，其决策过程缺乏透明度，这在各个领域的实际应用中限制了其落地率。基于此，本文重点研究了类激活图可解释性方法，该方法通过梯度反向传播，直观地展示了模型在图像中关注的目标区域，从而帮助决策者更好地理解模型结果。然而，研究表明，目前的类激活图方法仍存在解释结果不完整和细粒度不足的问题，未能达到理想的解释效果。因此，本文针对深度学习在图像分类领域的应用，提出一种多层混合注意力机制的类激活图可解释性方法 SAMLCAM，旨在为决策过程提供更精细的解释，进一步提升卷积网络模型的

可解释性。

该方法通过引入多层混合注意力机制，既确保了对通道级特征的关注，又补充了目标物体的位置信息，从而为类激活图提供更丰富的解释细节。首先，通过消融实验验证了混合注意力机制的有效性，并进一步与多层加权机制相结合，表明了SAMLCAM方法的优越性。在评估实验中，采用了多种待解释的卷积网络作为主干网络，实验结果表明，提出的方法在多种卷积模型中均有不同程度的性能提升，验证了其在提供更细粒度解释结果方面的优势，同时也表明了该方法在不同卷积网络中的适用性。此外，还进行了延时基准测试，确保在不牺牲处理效率的前提下实现了性能提升。综合来看，提出的SAMLCAM方法能够生成更完整、效果更佳的类激活图，为卷积网络决策过程提供更加全面的解释依据，提升深度学习智能模型的决策透明度，并有助于促进其实际应用落地。

尽管实验表明本文提出的类激活图可解释方法SAMLCAM在一定程度上提升了图像分类领域中卷积网络模型的可解释性，但仍存在一些不足。由于可解释性方法的多样性与复杂性，目前尚难以找到一种通用的评估标准适用于所有情况。不同应用领域和任务的需求差异使得评估指标和方法可能需要根据具体场景进行调整。因此，未来的研究应致力于开发更加全面、客观且切实可行的评估框架，考虑不同任务和领域的特点，并加强用户参与和反馈。这将有助于推动可解释性方法的不断完善，并提高其在实际应用中的有效性与可信度。通过解决当前问题，提升深度学习模型的可解释性，能够加速其在各类实际应用中的推广落地，并为决策者和用户提供更加可信、可靠的智能系统。

参考文献(References)

Ali N Y, Rahman L, Chaki J, Dey N and Santosh K C. 2021. Machine translation using deep learning for universal networking language based on their structure. International Journal of Machine Learning and Cybernetics, 12(8): 2365-2376 [DOI: 10.1007/s13042-021-01317-5]

Arrieta A B, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Garcia S, Gil-Lopez S, Molina D, Benjamins R, Chatila R and Herrera F. 2020. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58: 82-115 [DOI: 10.1016/j.inffus.2019.12.012]

Chattopadhay A, Sarkar A, Howlader P and Balasubramanian V N. 2018. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks//Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision. Lake Tahoe, USA: IEEE: 839-847 [DOI: 10.1109/WACV.2018.00097]

Chen Z Z and Sun Q R. 2023. Extracting class activation maps from non-discriminative features as well//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 3135-3144 [DOI: 10.1109/CVPR52729.2023.00306]

Cyberspace Administration of China, Publicity Department of the Communist Party of China, Ministry of Education of the People's Republic of China, Ministry of Science and Technology of the People's Republic of China, Ministry of Industry and Information Technology, The Ministry of Public Security of the People's Republic of China, Ministry of Culture and Tourism of the People's Republic of China, State Administration for Market Regulation and National Radio and Television Administration. 2021. Notice on Issuing the "Guiding Opinions on Strengthening the Overall Governance of Internet Information Service Algorithms" [EB/OL]. [2024-02-22]. https://www.gov.cn/zhengce/zhengceku/2021-09/30/content_5640398.htm (国家互联网信息办公室, 中央宣传部, 教育部, 科学技术部, 工业和信息化部, 公安部, 文化和旅游部, 国家市场监督管理总局, 国家广播电视总局. 2021. 关于印发《关于加强互联网信息服务算法综合治理的指导意见》的通知 [EB/OL]. [2024-04-17]. https://www.gov.cn/zhengce/zhengceku/2021-09/30/content_5640398.htm)

Dong Y P, Su H, Wu B Y, Li Z F, Liu W, Zhang T and Zhu J. 2019. Efficient decision-based black-box adversarial attacks on face recognition//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 7714-7722 [DOI: 10.1109/CVPR.2019.00790]

Dong Y P, Su H and Zhu J. 2022. Interpretability analysis of deep neural networks with adversarial examples. Acta Automatica Sinica, 48(1): 75-86 (董胤蓬, 苏航, 朱军. 2022. 面向对抗样本的深度神经网络可解释性分析. 自动化学报, 48(1): 75-86) [DOI: 10.16383/j.aas.c200317]

Dou H, Zhang L M, Han F, Shen F R and Zhao J. 2024. Survey on convolutional neural network interpretability. Journal of Software, 35(1): 159-184 (窦慧, 张凌茗, 韩峰, 申富饶, 赵健. 2024. 卷积神经网络的可解释性研究综述. 软件学报, 35(1): 159-184) [DOI: 10.13328/j.cnki.jos.006758]

Englebert A, Cornu O and De Vleeschouwer C. 2024. Poly-CAM: high resolution class activation map for convolutional neural networks. Machine Vision and Applications, 35(4): #89 [DOI: 10.1007/s00138-024-01567-7]

Fong R C and Vedaldi A. 2017. Interpretable explanations of black boxes by meaningful perturbation//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 3429-3437 [DOI: 10.1109/ICCV.2017.371]

Fu R G, Hu Q Y, Dong X H, Guo Y L, Gao Y H and Li B. 2020. Axiom-based grad-CAM: Towards accurate visualization and explanation of CNNs [EB/OL]. [2024-04-17]. https://arxiv.org/pdf/2008.02312.pdf

Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F and Pedreschi D. 2019. A survey of methods for explaining black box models. ACM Computing Surveys (CSUR), 51(5): #93 [DOI: 10.1145/3236009]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

He W and Pan C. 2022. The salient object detection based on attention-guided network. Journal of Image and Graphics, 27(4): 1176-1190 (何伟, 潘晨. 2022. 注意力引导网络的显着性目标检测. 中国图象图形学报, 27(4): 1176-1190) [DOI: 10.11834/jig.200658]

Hu J, Shen L, Albanie S, Sun G and Wu E H. 2020. Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(8): 2011-2023 [DOI: 10.1109/TPAMI.2019.2913372]

Hua Y Y, Zhang D C and Ge S M. 2020. Research progress in the interpretability of deep learning models. Journal of Cyber Security, 5(3): 1-12 (化盈盈, 张岱墀, 葛仕明. 2020. 深度学习模型可解释性的研究进展. 信息安全学报, 5(3): 1-12) [DOI: 10.19363/J.cnki.cn10-1380/tn.2020.05.01]

Huang G, Liu Z, Van Der Maaten L and Weinberger K Q. 2017. Densely connected convolutional networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 4700-4708 [DOI: 10.1109/CVPR.2017.243]

Jiang P T, Zhang C B, Hou Q B, Cheng M M and Wei Y C. 2021. LayerCAM: exploring hierarchical class activation maps for localization. IEEE Transactions on Image Processing, 30: 5875-5888 [DOI: 10.1109/TIP.2021.3089943]

Krizhevsky A, Sutskever I and Hinton G E. 2012. ImageNet classification with deep convolutional neural networks//Proceedings of the 26th International Conference on Neural Information Processing System. Lake Tahoe, Nevada: Curran Associates Inc.: 1097-1105

Li X H, Xiong H Y, Li X J, Wu X Y, Zhang X, Liu J, Bian J and Dou D J. 2022. Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond. Knowledge and Information Systems, 64(12): 3197-3234 [DOI: 10.1007/s10115-022-01756-8]

Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick C L. 2014. Microsoft COCO: common objects in context//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer: 740-755 [DOI: 10.1007/978-3-319-10602-1_48]

Linardatos P, Papastefanopoulos V and Kotsiantis S. 2020. Explainable AI: a review of machine learning interpretability methods. Entropy, 23(1): #18 [DOI: 10.3390/e23010018]

Omeiza D, Speakman S, Cintas C and Weldermariam K. 2019. Smooth grad-CAM++: an enhanced inference level visualization technique for deep convolutional neural network models [EB/OL]. [2024-04-17]. https://arxiv.org/pdf/1908.01224.pdf

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 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 618-626 [DOI: 10.1109/ICCV.2017.74]

Simonyan K and Zisserman A. 2014. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2024-04-17]. https://arxiv.org/pdf/1409.1556.pdf

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 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA: IEEE: 24-25 [DOI: 10.1109/CVPRW50498.2020.00020]

Zhang Y, Tiňo P, Leonardis A and Tang K. 2021. A survey on neural network interpretability. IEEE Transactions on Emerging Topics in Computational Intelligence, 5(5): 726-742 [DOI: 10.1109/TETCI.2021.3100641]

Zheng Q, Wang Z W, Zhou J and Lu J W. 2022. Shap-CAM: visual explanations for convolutional neural networks based on Shapley value//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 459-474 [DOI: 10.1007/978-3-031-19775-8_27]

Zhou B L, Khosla A, Lapedriza A, Oliva A and Torralba A. 2016. Learning deep features for discriminative localization//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 2921-2929 [DOI: 10.1109/CVPR.2016.319]

作者简介

张剑，男，研究员，主要研究方向为人工智能和作战指挥。
E-mail：1893664@qq.com
张一然，通信作者，女，硕士研究生，主要研究方向为人工智能可解释性。E-mail：y1random@163.com
王梓聪，男，硕士研究生，主要研究方向为人工智能。
E-mail：1421531677@qq.com