计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(11): 2737−2751, 2025

基于序贯三支掩码和注意力融合的 Transformer 解释方法

成晓天　丁卫平　耿　宇　黄嘉爽　鞠恒荣　郭　静
（南通大学人工智能与计算机学院　江苏南通　226019）
（1005335413@qq.com）

Transformer Interpretation Method Based on Sequential Three-Way Mask and Attention Fusion

Cheng Xiaotian, Ding Weiping, Geng Yu, Huang Jiashuang, Ju Hengrong, and Guo Jing
（*School of Artificial Intelligence and Computer Science*, *Nantong University*, *Nantong*, *Jiangsu* 226019）

Abstract　Transformer has gradually become the preferred solution for computer vision tasks, which has promoted the development of its interpretability methods. Traditional interpretation methods mostly use the perturbation mask generated by the Transformer encoder's final layer to generate an interpretable map. However, these methods ignore uncertain information on the mask and the information loss in the upsampling and downsampling processes, which can result in rough and incomplete positioning of the object area. To overcome the mentioned problems, a Transformer explanation method based on sequential three-way and attention fusion (SAF-Explainer) is proposed. SAF-Explainer mainly includes the sequential three-way mask (S3WM) module and attention fusion (AF) module. The S3WM module processes the mask by applying strict threshold conditions to avoid the uncertainty information in the mask from damaging the interpretation results, so as to effectively locate the object position. Subsequently, AF module uses attention matrix aggregation to generate a relationship matrix for cross-layer information interaction, which is used to optimize the detailed information in the interpretation results and generates clear and complete interpretation results. To verify the effectiveness of the proposed SAF-Explainer, comparative experiments are conducted on three natural image datasets and one medical image dataset. The results show that SAF-Explainer has better explainability. This work advances visual explanation techniques by providing more accurate and clinically relevant interpretability for Transformer-based vision systems, particularly in medical diagnostic applications where precise region identification is crucial.

Key words　interpretable method; Transformer; self-attention mechanism; sequential three-way decision; attention fusion; perturbation mask

摘　要　Transformer 逐渐成为计算机视觉任务的首选方案，这推动了其可解释性方法的发展. 传统解释方

收稿日期：2024-05-31；**修回日期**：2025-03-10
基金项目：国家重点研发计划项目 (2024YFE0202700)；国家自然科学基金项目（U2433216, 61976120, 62006128, 62102199）；江苏省自然科学基金项目（BK20231337）；江苏省双创博士计划项目；江苏省高等学校自然科学研究重大项目（21KJA510004）；中国博士后科学基金项目（2022M711716）；江苏省研究生科研与实践创新计划项目（SJCX24_2021）
This work was supported by the National Key Research and Development Program of China (2024YFE0202700), the National Natural Science Foundation of China (U2433216, 61976120, 62006128, 62102199), the Natural Science Foundation of Jiangsu Province (BK20231337), the Double-Creation Doctoral Program of Jiangsu Province, the Natural Science Key Foundation of Higher Education of Jiangsu Province (21KJA510004), the China Postdoctoral Science Foundation (2022M711716), and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (SJCX24_2021).
通信作者：丁卫平（dwp9988@163.com）

法大多采用 Transformer 编码器的最终层生成的扰动掩码生成可解释图,而忽略了掩码的不确定信息和上下采样中的信息丢失,从而造成物体区域的定位粗糙且不完整. 为克服上述问题,提出基于序贯三支掩码和注意力融合的 Transformer 解释方法(SAF-Explainer),SAF-Explainer 主要包含序贯三支掩码(sequential three-way mask, S3WM)模块和注意力融合(attention fusion, AF)模块. S3WM 通过应用严格的阈值条件处理掩码,避免掩码中的不确定信息对解释结果产生损害,以此有效定位到物体位置. 随后,AF 利用注意力矩阵聚合生成跨层信息交互的关系矩阵,用来优化解释结果中的细节信息,生成边缘清晰且完整的解释结果. 为验证所提出 SAF-Explainer 的有效性,在 3 个自然图像与 1 个医学图像数据集上进行比较实验,结果表明 SAF-Explainer 具有更好的可解释性效果.

关键词 可解释性方法;Transformer;自注意力机制;序贯三支决策;注意力融合;扰动掩码

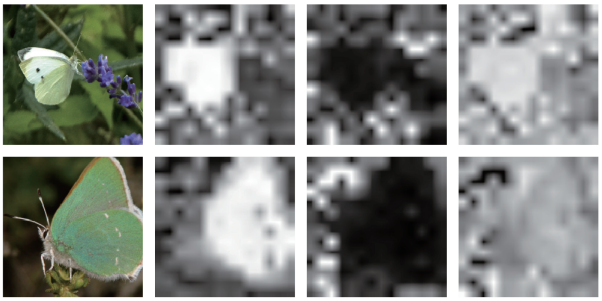
中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202440382 **CSTR:** 32373.14.issn1000-1239.202440382

随着人工智能技术的迅猛发展,出现了许多优秀的深度学习模型,如转换器(Transformer)[1]. 最初,Transformer 被广泛应用于自然语言处理领域并取得了巨大成功. 随后,视觉转换器(vision Transformer, ViT)[2]在计算机视觉领域也逐渐取代卷积神经网络(convolutional neural network, CNN),在图像分类[3]、语义分割[4-5]、目标检测[6-7]、图像检索[8]任务中,ViT 展现出了优异的效果. 然而,由于这些模型具有高度复杂的线性关系以及超大规模的参数量,人类无法理解其内部决策过程,因此通常被视为黑盒模型,这成为许多需要安全性和社会认可的现实应用中的主要缺陷[9-10]. 尽管在现实世界已经部署了一些人工智能系统,但在医疗诊断领域,由于其固有且不可否认的风险,Transformer 衍生模型难以真正投入使用[11].

传统的深度学习模型解释方法通常是对输入进行遮挡扰动,当遮挡区域为模型认为的重要区域时,往往会导致模型的输出得分较低[12]. 但这种基于遮挡的解释方法通常对遮挡策略有着较高的要求,随机遮挡往往难以达到较好的解释效果. Ding 等人[13]提出一种基于多粒度随机游走算法的遮挡策略,从粗粒度到细粒度对图像进行遮挡,并以注意力作为指引,大大提升了扰动的效率,并取得了较好的解释效果. Petsiuk 等人[14]提出利用蒙特卡洛采样生成一定数量的掩码,通过掩码与输入相乘对输入进行扰动,利用模型输出的类别分数与掩码加权求和得到最终的解释结果. 这一过程无需访问模型内部,因此这是一种黑盒解释方法. 不同的是,Xie 等人[15]提出利用 ViT 模型的最终层输出生成掩码,而无需利用蒙特卡洛采样,解决了解释效率低、伪影导致的分数误差以及掩码像素覆盖偏差的问题.

然而,简单利用 ViT 编码器最终层输出生成掩码忽略了 2 个关键方面:一是掩码中存在大量的不确定信息,如果不采取策略将会导致不完整、不可靠的物体定位信息;二是掩码在上下采样过程中造成的信息丢失问题,会导致解释结果中存在大量的空间噪声及细节缺失. 正如文献 [16–17] 所指出的,ViT 存在过度平滑问题,即随着堆叠的层数的增加,各令牌的语义信息趋于一致,导致最终层输出的掩码高度相似,我们遵循了文献 [15] 中的层次聚类方法来减少冗余掩码数量. 图 1 展示了积极掩码、消极掩码以及一些不确定性掩码. 为获得更优的显著性解释结果,我们应尽可能充分利用积极掩码. 图 1(b)显示了积极掩码,它能够基本完整地反映出原图对预测做出重大贡献的物体位置. 相反,图 1(c)展示了消极掩码,这些掩码无意义或与原图完全相反,对可视化结果产生消极影响. 此外,图 1(d)中的不确定掩码通常会对整体产生一定程度的掩盖,无法确定其对可视化结果的影响,需要进一步的细粒度信息来确定其是否为积极掩码或消极掩码.



(a) 输入图像 (b) 积极掩码 (c) 消极掩码 (d) 不确定性掩码

Fig. 1 Visualization results of three masks

图 1 3 种掩码可视化结果

Yao[18]提出的三支决策方法能够有效应对上述不确定性问题. 与传统两支决策不同,当目前证据不能

做出接受或拒绝的决策时，三支决策采用延迟决策的方式处理信息，将信息置于边界域，以降低错误决策带来的损失. 基于多粒度结构，序贯三支决策应运而生，序贯三支决策对进入边界域的信息通过可靠的细粒度的信息再次进行决策，形成了从粗粒度到细粒度的顺序过程. Savchenko[19] 将序贯三支决策应用于图像识别，提高了图像识别的速度，Ju 等人[20] 开发的具有合理子空间的序贯三支分类器有效提升了分类性能，Li 等人[21] 提出了代价敏感的序贯三支决策用于人脸识别. 通过序贯三支策略，我们能够从多粒度的角度获取积极掩码，生成更有益的可视化结果[22-23].

自注意力机制是 Transformer 架构的核心模块，可视化类别令牌的注意力矩阵是 ViT 的一种天然解释方法[24]. 然而，需要注意的是，可视化注意力矩阵的结果与类别无关，即无论输入何种类别，模型的解释结果均相同，这对于多类别图像显然是不合理的. 因此，我们不使用类别令牌，而是通过注意力矩阵分析各令牌之间的信息交互，获取所有令牌包含的前景、背景等细节语义信息. 我们提出一个注意力融合（attention fusion, AF）模块，旨在生成更为准确的可视化解释图.

针对 ViT 架构的独有特性，本文提出了一种基于序贯三支掩码和注意力融合的 Transformer 解释方法（SAF-Explainer）. 首先，将图像输入 Transformer 编码器，并上采样最终层输出作为掩码. 然后，通过序贯三支掩码模块，将掩码划分为积极掩码、消极掩码和不确定掩码，并通过多粒度层级顺序分析选出所有的积极掩码. 接着，利用 Transformer 计算置信度分数，并与积极掩码加权求和，得到初步解释结果. 接下来，使用提出的注意力融合模块对结果进行细节修正. 首先，保存每一层 Transformer 编码器中的注意力矩阵，并跨层聚合生成关系矩阵. 关系矩阵能够反映图像块之间的关系密切程度，例如前景图像块会与其他前景图像块具有较高的关系程度，与背景图像块的关系程度则较低. 因此，通过计算初步解释结果与关系矩阵之间的余弦相似度，得出每一个图像块的块重要性分数，利用重要性分数对初步解释结果进行融合，即可得到最终解释结果. 通过热图的方式，可以清晰地呈现最终的可视化结果，这些结果能够有效反映 ViT 模型在决策时的依据，从而体现模型的可解释性.

本文的主要贡献有 3 点：

1）提出序贯三支掩码（sequential three-way mask, S3WM）模块，以解决 ViT 最终层特征图生成的掩码质量不确定性问题. 通过设置特定的阈值及条件，可以有效地将掩码划分为积极掩码、消极掩码与不确定掩码 3 种类型，使用积极掩码可以明显提高模型解释效果.

2）提出注意力融合（attention fusion, AF）模块，它通过聚合 ViT 中每一层的注意力矩阵，生成可以反映图像块前景、背景与边缘区域关系的关系矩阵，通过余弦相似度度量获取图像块重要性分数，对解释结果进行加权融合，这一过程有效地解决了初步解释结果中存在的噪声问题，提高模型解释效果.

3）在 3 个自然图像数据集以及 1 个脑部肿瘤医学图像数据集对 SAF-Explainer 进行定量定性分析以及可视化结果验证.

1 相关工作

ViT 由于其黑盒特性，本身不具备可解释性，通常需要应用事后解释方法来为训练后的模型生成解释. ViT 的可解释性方法可以分为两大类：一类是最初为 CNN 设计的解释方法，尽管这些方法最初是为 CNN 设计的，但它们可以迁移并应用于 ViT；另一类是专门针对 ViT 特性设计的解释方法.

1.1 基于反向传播的方法

早期基于反向传播的方法通过梯度近似像素扰动前后模型输出的差值来衡量像素的重要性，梯度的绝对值越高，该像素对预测的影响越显著. 然而，基于梯度的灵敏度图通常包含大量噪声，Smilkov 等人[25] 进行多次采样并添加高斯噪声，取多次反向传播结果的平均值来获取平滑的灵敏性图. Sundararajan 等人[26] 也是计算梯度的平均值，其样本通过输入图像与基线图片之间的插值获得，将梯度积分累计以获取灵敏性图. 相关性逐层传播（layer-wise relevance propagation, LRP）方法从模型输出开始，反向传播计算各层间的相关性系数[27]，像素的相关性反映了该像素对模型决策的贡献. Voita 等人[28] 通过 LRP 挑选出相关性较高的 Transformer 注意力头. 这些注意力头被视为对预测贡献较大的部分. 然后，剪枝绝大多数相关性较低的注意力头. 即使在剪枝后，模型仍能保持较好的性能. Chefer 等人[29] 提出了一个 Transformer 归因（Transformer attribution, T-Attribution）方法，它采用 LRP 计算 Transformer 模型中每一层、每一个注意力头的相关性，然后通过梯度加权以聚合生成新的注意力矩阵，最终生成特定于类的可视化解释.

1.2 基于类激活图的方法

类激活图（class activation map, CAM）的概念最早

出现在文献 [30] 中. CAM 将网络中的全连接层替换为全局平均池化层, 每个激活图与最终预测输出之间的权重, 用于衡量该激活图对预测的贡献程度. 接着, 通过权重与激活图之间的线性加权求和, 聚合生成重要性分数. 最后, 通过热图方式可视化重要区域. 作为 CAM 的推广, 梯度加权的类激活图(gradient-weighted class activation map, Grad-CAM[31])及其变体方法[32]使用反向传播获取的梯度均值作为每个激活图对最终预测的贡献程度, 对激活图进行线性加权求和, 使其不依赖于特定架构. Wang 等人[33]指出由于梯度饱和或梯度消失问题, 往往会导致产生大量视觉噪声和虚假置信度问题, 即权重较高的激活图输入到网络中获取的分数很低, 从而提出一种分数加权的类激活图(score-weighted class activation map, Score-CAM)方法. 尽管以上方法可以有效定位物体, 但由于使用的激活图均来自最终层, 生成的可视化结果通常较为粗糙. Jiang 等人[34]利用梯度信息为激活图中的每个位置生成单独的权重, 来自浅层的激活图可以捕获细粒度的对象定位信息, 将不同层的类激活图进行融合后, 即可生成精确完整的类激活图. 这种方法生成的解释图可作为伪标签, 有效应用于弱监督对象定位和分割任务[35-36].

1.3 基于扰动的方法

Zeiler 等人[12]通过滑动一个灰度正方形块来遮挡图像扰乱输入. 然后, 将被遮挡的图像输入分类模型中, 观察深层特征图和输出结果. 当遮挡区域为图像重要区域时, 通常会导致模型输出的分类分数降低. Ding 等人[13]使用随机游走算法, 从粗粒度到细粒度对图像进行扰动, 观察对模型预测的影响进而解释模型. 随机输入采样解释(randomized input sampling for explanation, RISE)通过蒙特卡洛采样生成一定数量的掩码. 然后, 将被掩盖的图像输入模型, 以获取特定于类的置信度得分作为权重. 接着, 将权重与掩码进行线性加权求和, 即可得到图像的重要性分数[14]. 由于掩码是随机生成且与模型无关, RISE 是一种无需访问模型内部结果的通用方法. 与 RISE 类似, 因果解释视觉 Transformer(causal explanation of vision Transformer, ViT-CX)通过将掩码输入到模型中获取权重以生成可视化解释[15]. 但与 RISE 不同的是, ViT-CX 的掩码不是随机生成的, 而是通过最终层补丁嵌入上采样获取, 并通过层次聚类算法减少掩码的数量.

1.4 基于注意力的方法

ViT 通过自注意力机制对图像上下文进行全局建模, 因此可视化注意力分数是了解其工作机理的直觉想法. 然而, 简单地使用最后一层编码器的原始注意力矩阵, 通常在定位预测目标物体上效果较差. 因为 ViT 中还涉及前馈网络及大量的跳跃连接, 且中间层信息也被忽略. Abnar 等人[24]假设自注意力是线性组合的, 通过使用有向无环图对网络中的信息流进行建模, 将上下编码器层中的令牌视作结点, 注意力分数为边的权重, 递归相乘注意力矩阵以获取更为准确的注意力分数. 文献 [24] 的工作均使用注意力矩阵中的类别令牌进行可视化. 然而, 我们的工作不依赖于注意力矩阵中的类别令牌, 而是分析了其他令牌之间的交互作用, 通过聚合跨层注意力矩阵生成令牌关系矩阵, 融合生成准确合理的可视化解释结果.

在针对 ViT 架构的方法中, 基于反向传播的方法仅考虑梯度信息, 导致解释结果比较保守, 通常只能识别部分前景物体信息, 且重要性得分不高; 由于缺乏针对 ViT 特性定制, 基于扰动的方法通常会产生包含大量背景区域的解释结果, 难以令人信服; 而基于注意力的方法通常只能提供类别无关的解释结果, 并且由于专注于注意力组件, 缺乏对 ViT 架构的整体探索, 解释结果往往包含大量空间噪声. 为了解决上述缺点, 我们的工作充分利用了 ViT 架构的特性, 使用 S3WM 模块来选择积极掩码并生成初步的解释结果, 准确定位到前景物体; 接着, 本研究深入探讨了 ViT 架构的自注意力机制, 通过所提出的 AF 模块, 消除空间中的噪声信息, 优化前景物体边缘细节, 最终生成定位精准、噪声低的可视化解释结果.

2 Transformer 解释方法

本文提出了基于序贯三支掩码和注意力融合的 Transformer 解释方法, 具体流程如图 2 所示.

图 2 左图为标准 ViT 模块, 图 2 右上为 S3WM 模块, 图 2 右下为 AF 模块. 首先, 将图像输入 ViT 模块, 获取最后一层编码器的输出, 通过重塑和上采样生成掩码以扰动图像输入. 由于最后一层编码器的输出具有高度的语义相似性, 我们采用了层次凝聚聚类算法来聚合相似的掩码, 以减少冗余性. 由于掩码质量的不确定性, 将掩码集放入 S3WM 模块, 通过多粒度层级顺序分析选出积极掩码集 M_{pos}, 并生成初步解释结果 $\boldsymbol{S}$, $\boldsymbol{S}$ 可以准确定位物体位置, 但往往包含背景噪声信息. 接着, 本文分析了自注意力机制, 探索了图像块之间的信息交互, 并提出了 AF 模块. 该模块利用跨层的注意力信息生成关系矩阵 $\boldsymbol{R}$, 用于进一步优化解释结果中的细节信息, 最终生成最优的

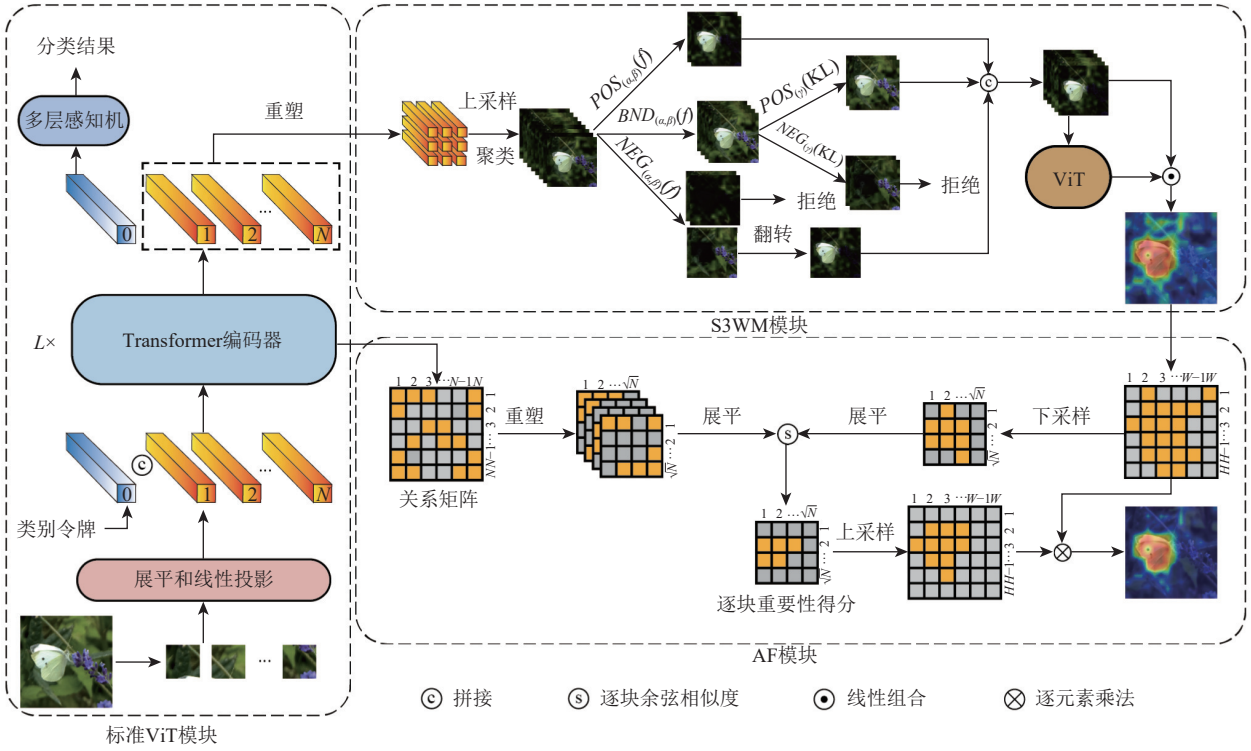


Fig. 2 SAF-Explainer model architecture

图 2 SAF-Explainer 模型结构

解释结果 $\boldsymbol{V}$, 有效提高模型可解释性.

2.1 预备知识

2.1.1 标准 ViT 模块

本文主要考虑用于分类任务的标准 ViT 模块. 令 $f(\cdot)$为标准 ViT 模块, $\boldsymbol{I}\in\mathbb{R}^{H\times W\times C}$为输入图像, 其中 H, W, C 分别是图像的高、宽和通道, 将其切割分成大小$P\times P\times C$的共 N 块小图像块, 接着展平后线性嵌入成 D 维, 在头部拼接类别令牌并加入位置信息编码嵌入, 最终产生 $N+1=HW/P^2+1$个令牌作为输入 $\boldsymbol{X}_0\in\mathbb{R}^{(N+1)\times D}$馈送到 Transformer 编码器重复 L 遍, 每一层的输出为$\boldsymbol{X}_i\in\mathbb{R}^{(N+1)\times D}$, i=1, 2, …, L. 在每一个 Transformer 编码器中, $\boldsymbol{X}_i$经过多头自注意力模块和多层感知机模块处理, 并在每个模块后应用跳跃连接, 以对全局进行建模. 最后, 从$\boldsymbol{X}_L$中取出类别令牌, 输送到多层感知机中获取分类结果.

2.1.2 多头自注意力机制

受人类在复杂场景中能够有效注意到重点区域的启发, 注意力机制被引入计算机视觉领域. 以第 l 层 Transformer 编码器中自注意力模块为例, 将序列 $\boldsymbol{X}$作为输入, 通过投影变换为查询矩阵 $\boldsymbol{Q}$、键矩阵 $\boldsymbol{K}$ 和值矩阵 $\boldsymbol{V}$. 接着, 将 $\boldsymbol{Q}$ 与 $\boldsymbol{K}^{\mathrm{T}}$进行矩阵乘法, 并整体除以 $\sqrt{d_{\boldsymbol{K}}}$, 以防止经过 softmax 后出现梯度消失的问题, 然后将矩阵放入 softmax 函数进行归一化, 即可得到原始注意力矩阵$\boldsymbol{A}^{(l)}$, 其中每一行表示当前令牌对其他图像块的关注程度, 最后将$\boldsymbol{A}^{(l)}$与 $\boldsymbol{V}$进行矩阵乘法即可得到一次自注意力层的输出结果:

$$\boldsymbol{A}^{(l)}=softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_{\boldsymbol{K}}}}\right), \tag{1}$$

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\boldsymbol{A}^{(l)}\boldsymbol{V}, \tag{2}$$

其中$d_{\boldsymbol{K}}$是 $\boldsymbol{K}$ 的通道维度. 以上为单个自注意力头的工作过程, 此时$\boldsymbol{A}^{(l)}\in\mathbb{R}^{(N+1)\times(N+1)}$.

受卷积神经网络中使用多个卷积核提取不同特征的启发, 多头自注意力使用多个自注意头学习多种不同类型的上下文交互信息, 增强模型的表达能力, 然后, 将每个头学习到的特征进行拼接, 获得多头自注意力的输出结果:

$$\boldsymbol{head}_i=Attention(\boldsymbol{Q}\boldsymbol{W}_i^{\boldsymbol{Q}},\boldsymbol{K}\boldsymbol{W}_i^{\boldsymbol{K}},\boldsymbol{V}\boldsymbol{W}_i^{\boldsymbol{V}}), \tag{3}$$

$$MultiHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Concat(\boldsymbol{head}_1,\cdots,\boldsymbol{head}_h)\boldsymbol{W}, \tag{4}$$

其中$\boldsymbol{W}_i^{\boldsymbol{Q}}$, $\boldsymbol{W}_i^{\boldsymbol{K}}$, $\boldsymbol{W}_i^{\boldsymbol{V}}$, $\boldsymbol{W}$是投影变换参数矩阵, $\boldsymbol{head}_i$为每个头的计算输出结果, h为头的个数, $Concat$表示拼接操作, 经过多头自注意力操作后的注意力矩阵为$\boldsymbol{A}^l\in\mathbb{R}^{h\times(N+1)\times(N+1)}$.

2.2 S3WM 模块

2.2.1 掩码生成

与 Score-CAM 通过上采样 CNN 的最后一层激活图获取掩码类似, 我们使用 ViT 最后一层编码器输

出重塑并上采样生成掩码. 设初始输入为 $\boldsymbol{X}_0 \in \mathbb{R}^{(N+1)\times D}$, 经过 L 层编码器后的输出为 $\boldsymbol{X}_L \in \mathbb{R}^{(N+1)\times D}$, 舍去类别令牌后重塑为 $\boldsymbol{m} \in \mathbb{R}^{\sqrt{N}\times\sqrt{N}\times D}$, 其中每个位置的值代表图像块在空间位置的信息, 上采样后获取 D 个掩码 $\boldsymbol{M}_i \in \mathbb{R}^{H\times W}$, 构成掩码集 $M=\{\boldsymbol{M}_1,\boldsymbol{M}_2,\cdots,\boldsymbol{M}_D\}$. 在 Transformer 架构中, 图像被分割并转换为补丁令牌, 然后与类令牌拼接后, 馈送到 Transformer 编码器中进行 L 次重复操作, 每次输出的补丁令牌的形状保持不变. 由于多次执行自注意力模块, 补丁令牌逐渐趋于一致. 随着编码器层的加深, 补丁令牌之间的成对余弦相似度逐步升高, 这意味着令牌的语义信息逐渐丢失[17]. 因此, 掩码之间存在冗余, 并包含高质量的积极掩码、低质量的消极掩码以及大量不确定性掩码. 我们采用层次凝聚聚类算法减少掩码冗余. 首先将每个掩码视作 1 个单独的簇, 使用余弦相似度计算相似度度量矩阵, 并通过对相似性度量矩阵求反得到距离矩阵. 假设掩码 $\boldsymbol{M}_i$ 与掩码 $\boldsymbol{M}_j$ 已经被展平为 1 维张量, 则它们之间的距离可以表示为

$$Distance(\boldsymbol{M}_i,\boldsymbol{M}_j)=1-Cosine\_sim(\boldsymbol{M}_i,\boldsymbol{M}_j), \tag{5}$$

$$Cosine\_sim(\boldsymbol{M}_i,\boldsymbol{M}_j)=\frac{\boldsymbol{M}_i\cdot\boldsymbol{M}_j}{\|\boldsymbol{M}_i\|\,\|\boldsymbol{M}_j\|}, \tag{6}$$

其中“$\cdot$”表示点乘操作, $\|\cdot\|$ 表示欧几里得范数.

通过式(5)计算可以得到距离矩阵, 距离矩阵中各元素值表示掩码之间的成对距离, 通过设定阈值 d 控制生成的聚类数量. 计算簇间距离时采用完全连接策略, 即使用 2 簇中掩码之间的最大距离作为 2 簇的距离, 确保聚合生成的掩码的差异性. 假设经过聚类后共产生 c 个簇, 每个簇包含不同数量的掩码, 我们通过对簇内掩码取平均, 生成最终经过层次聚类后的掩码集 $M_{\mathrm{C}}=\{\boldsymbol{M}_1,\boldsymbol{M}_2,\cdots,\boldsymbol{M}_c\}$.

2.2.2 序贯三支决策

层次聚类在一定程度上减少了掩码的冗余性, 但掩码质量仍具有不确定性. 信息量不足是造成不确定性问题的主要原因, 这里, 我们采用序贯三支决策处理不确定性掩码. 首先, 将掩码放入 ViT 模型 $f(\cdot)$, 输出的概率可以提供粗粒度的掩码质量信息. 利用这些信息, 我们可以将掩码分为积极掩码、消极掩码和不确定性掩码.

对 $\boldsymbol{M}_i \in M_{\mathrm{C}}$, 如果放入模型后输出与输入图像 $\boldsymbol{I}$ 相当的概率, 即体现了图像核心区域信息, 而翻转掩码 $\boldsymbol{1}-\boldsymbol{M}_i$ 输出较低概率, 即仅捕捉到一些次要信息时, 则该掩码属于积极掩码, 放入接受域 *POS*.

$$POS_{(\alpha,\beta)}(f)=\{\boldsymbol{M}_i\in M_{\mathrm{C}}|(\beta<f((\boldsymbol{1}-\boldsymbol{M}_i)\odot\boldsymbol{I})<\alpha)\wedge(f(\boldsymbol{M}_i\odot\boldsymbol{I})>\alpha)\}, \tag{7}$$

其中 $\boldsymbol{M}_i \in M_{\mathrm{C}}$ 表示当前处理的是属于候选掩码集 M_{C} 中的掩码, $\boldsymbol{M}_i\odot\boldsymbol{I}$ 可以得到经过掩码 $\boldsymbol{M}_i$ 覆盖后的图像, 将其放入 ViT 模型 $f(\cdot)$ 中即可得到输出概率, “$\wedge$”为逻辑且运算符, “$\odot$”为逐元素乘法, α 与 β 为控制阈值.

在某些情况下, 掩码 $\boldsymbol{1}-\boldsymbol{M}_i$ 可能才是我们所需要的积极掩码, 此时应将翻转后的掩码归入接受域 *POS*. 如图 3 所示, 第 1 行为掩码 $\boldsymbol{M}_i$ 及 $\boldsymbol{M}_i\odot\boldsymbol{I}$ 的可视化结果, $\boldsymbol{M}_i$ 没有捕捉到图像中的核心信息, 但将其取反后, 第 2 行显示的翻转掩码 $\boldsymbol{1}-\boldsymbol{M}_i$ 是我们需要的积极掩码, 因此我们将其放入 *POS*.

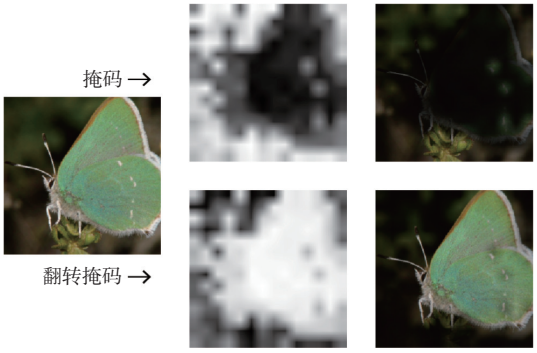


Fig. 3 Visualization of mask and reverse mask

图 3 掩码及翻转掩码可视化

当掩码 $\boldsymbol{M}_i$ 或其翻转掩码 $\boldsymbol{1}-\boldsymbol{M}_i$ 放入模型时输出极低的概率, 说明掩码 $\boldsymbol{M}_i$ 捕捉到的是杂乱无用的信息, 该掩码属于消极掩码, 放入拒绝域 *NEG*:

$$NEG_{(\gamma)}(f)=\{\boldsymbol{M}_i\in M_{\mathrm{C}}|(f((\boldsymbol{1}-\boldsymbol{M}_i)\odot\boldsymbol{I})<\gamma)\vee(f(\boldsymbol{M}_i\odot\boldsymbol{I})<\gamma)\}, \tag{8}$$

其中“$\vee$”为逻辑或运算符, γ 为控制阈值.

剩余掩码在未有进一步的细粒度信息时, 无法判断其为积极掩码还是消极掩码, 属于不确定性掩码, 不确定性掩码放入边界域 *BND* 等待延迟决策:

$$BND_{(\alpha,\beta,\gamma)}(f)=M_{\mathrm{C}}-POS_{(\alpha,\beta)}(f)-NEG_{(\gamma)}(f), \tag{9}$$

其中“–”表示差集运算符, 表示集合 A 中的元素不在集合 B 中. 在这里, 式(9)表示取出候选掩码集 M_{C} 中所有的剩余掩码, 将其放入 *BND*.

接着, 我们建议使用 KL 散度(Kullback-Leibler divergence)来进一步处理 *BND* 中的不确定性掩码. 具体来说, 我们的模型通过衡量原始图像 $\boldsymbol{I}$ 与 *BND* 中掩码的概率分布差异来筛选积极掩码. 其中, $f(\boldsymbol{I})$ 为目标分布 $\boldsymbol{P}$, $f(\boldsymbol{M}_i\odot\boldsymbol{I})$ 为近似分布 $\boldsymbol{Q}_i$, 我们定义目标分布 $\boldsymbol{P}$ 和 $\boldsymbol{M}_i$ 对应的近似分布 $\boldsymbol{Q}_i$ 之间的 KL 散度:

$$KL(\boldsymbol{P}\|\boldsymbol{Q}_i)=\sum_i \boldsymbol{P}\mathrm{lb}\frac{\boldsymbol{P}}{\boldsymbol{Q}_i}. \tag{10}$$

通过式(10)计算出所有 $\boldsymbol{M}_i \in BND$ 与 $\boldsymbol{I}$ 的 KL 散

度值，并将其按从小到大的顺序排序. 取其中较大的 KL 散度值与较小的掩码 $\boldsymbol{M}_i$ 放入 POS. KL 散度大的掩码与输入图像 $\boldsymbol{I}$ 的概率分布差异大，说明该掩码具备特异性，有利于生成有意义的解释结果，而 KL 散度小的掩码则可以较好体现原始图像 $\boldsymbol{I}$ 的特征. 对于 KL 散度处于中间范围的掩码，它们的表现较为模糊和不确定. 为了避免这种不确定性对解释结果产生损害，我们将其放入 NEG 中：

$$POS_{(\delta)}(KL)=\{\boldsymbol{M}_i\in BND|(KL(f(\boldsymbol{I})\|f(\boldsymbol{M}_i\odot\boldsymbol{I}))<\delta)\vee(KL(f(\boldsymbol{I})\|f(\boldsymbol{M}_i\odot\boldsymbol{I}))>1-\delta)\},\tag{11}$$

$$NEG_{(\delta)}(KL)=\{\boldsymbol{M}_i\in BND|\delta<KL(f(\boldsymbol{I})\|f(\boldsymbol{M}_i\odot\boldsymbol{I}))<1-\delta\},\tag{12}$$

其中 δ 为控制阈值，$\boldsymbol{M}_i\in BND$ 指在第 1 轮三支决策中放入边界域中的掩码.

通过序贯三支决策筛选后，获得最终的积极掩码集 $M_{\text{pos}}=\{\boldsymbol{M}_1,\boldsymbol{M}_2,\cdots,\boldsymbol{M}_p\}$，其中 p 为积极掩码集中掩码的个数. 对于输入图像 $\boldsymbol{I}$ 中每个像素 x 的重要性值计算通过对掩码 $\boldsymbol{M}_i$ 与其放入模型后对应类别 c 的输出 $f_c(\boldsymbol{M}_i\odot\boldsymbol{I})$，加权求和，并除以掩码集的期望值 $E(M_{\text{pos}})$ 获取：

$$\boldsymbol{S}(x)=\frac{\sum_{i=1}^{p}f_c(\boldsymbol{M}_i\odot\boldsymbol{I})\cdot\boldsymbol{M}_i(x)}{E(M_{\text{pos}})},\tag{13}$$

其中 $f_c(\cdot)$ 表示特定于类别 c 的模型，$E(\cdot)$ 为取均值函数.

接着，为了便于利用 AF 模块对解释结果进行优化，对 $\boldsymbol{S}$ 进行归一化处理：

$$\boldsymbol{S}=\frac{\boldsymbol{S}-\boldsymbol{S}_{\min}}{\boldsymbol{S}_{\max}-\boldsymbol{S}_{\min}}.\tag{14}$$

S3WM 模块算法如算法 1 所示.

算法 1. 序贯三支掩码模块算法（S3WM）.

输入：掩码集 M_{C}，输入图像 $\boldsymbol{I}$，标准 ViT 模型 $f(\cdot)$，阈值集 $\{\alpha,\beta,\gamma,\delta\}$；

输出：初步解释结果 $\boldsymbol{S}$.

① $POS=NEG=BND=\varnothing$; /*初始化 3 个域*/

② $POS_{(\alpha,\beta)}(f)=\{\boldsymbol{M}_i\in M_{\text{C}}|(\beta<f((\boldsymbol{1}-\boldsymbol{M}_i)\odot\boldsymbol{I})<\alpha)\wedge(f(\boldsymbol{M}_i\odot\boldsymbol{I})>\alpha)\}$;

③ $NEG_{(\gamma)}(f)=\{\boldsymbol{M}_i\in\boldsymbol{M}_{\text{C}}|(f((1-\boldsymbol{M}_i)\odot\boldsymbol{I})<\gamma)\vee(f(\boldsymbol{M}_i\odot\boldsymbol{I})<\gamma)\}$;

④ $BND_{(\alpha,\beta,\gamma)}(f)=M_{\text{C}}-POS_{(\alpha,\beta)}(f)-NEG_{(\gamma)}(f)$;

⑤ $POS=POS\cup POS_{(\alpha,\beta)}(f)$;

⑥ $BND=BND\cup BND_{=(\alpha,\beta,\gamma)}(f)$;

⑦ $POS_{(\delta)}(KL)=\{\boldsymbol{M}_i\in BND|(KL(f(\boldsymbol{I})\|f(\boldsymbol{M}_i\odot\boldsymbol{I}))<\delta)\vee(KL(f(\boldsymbol{I})\|f(\boldsymbol{M}_i\odot\boldsymbol{I}))>1-\delta)\}$;

⑧ $NEG_{(\delta)}(KL)=\{\boldsymbol{M}_i\in BND|\delta<KL(f(\boldsymbol{I})\|f(\boldsymbol{M}_i\odot\boldsymbol{I}))<1-\delta\}$;

⑨ $POS=POS\cup POS_{(\delta)}(KL)$;

⑩ $\boldsymbol{S}$=**0**;

⑪ for $\boldsymbol{M}_i$ in POS do/*利用积极掩码加权求和*/

⑫ $\quad\boldsymbol{S}=\boldsymbol{S}+f_c(\boldsymbol{M}_i\odot\boldsymbol{I})\cdot\boldsymbol{M}_i$;

⑬ end for

⑭ $\boldsymbol{S}=\boldsymbol{S}/E(POS)$;

⑮ $\boldsymbol{S}=(\boldsymbol{S}-\boldsymbol{S}_{\min})/(\boldsymbol{S}_{\max}-\boldsymbol{S}_{\min})$; /*归一化*/

⑯ return $\boldsymbol{S}$. /*返回初步解释结果*/

2.3 AF 模块

2.3.1 自注意力信息交互

自注意力机制的核心功能是实现全局信息的交互. 然而，在可解释人工智能领域，以往对自注意力机制的研究大多仅限于类别令牌[24,29]，忽略了占据多数的其余补丁令牌，从而缺乏对自注意力机制整体效益的探索. 我们通过实验分析了自注意力机制中其他补丁令牌的关注区域，进一步探索了跨补丁之间的信息交互[37]. 如图 4 所示，我们将图像划分为 N 个部分，并将它们标记为前景块、背景块和边缘块，以便进行详细分析.

由 2.1.2 节分析可知，第 l 层 Transformer 编码器层的注意力矩阵为 $\boldsymbol{A}^{(l)}\in\mathbb{R}^{h\times(N+1)\times(N+1)}$，为了便于分析，我们将注意力矩阵在头部方向上取平均，并舍去类别令牌后，其维度为 $N\times N$，每行表示当前图像块对其余图像块的关注程度. 如图 4 所示，我们分别使用猫与蝴蝶作为前景块对自注意力信息进行分析. 图 4（a）（b）中第 1 列为原始图像，图像被切块处理并按号划分为前景图像块、背景图像块和边缘图像块，图 4（a）（b）中第 2~4 列分别可视化了每一层 Transformer 编码器所有前景图像块、背景图像块、边缘图像块对其余图像块的平均注意力分数，其中第 2 列表示对前景图像块的注意力分数，第 3 列表示对背景图像块的注意力分数，第 4 列表示对边缘图像块的注意力分数. 观察结果表明，前景区域的图像块始终对前景区域保持较高的关注度. 随着编码器层数的加深，对边缘区域的关注有所提升，但对背景区域的关注始终较低；背景区域的图像块对背景区域也保持较高关注度，但随着层数的加深，对前景区域和边缘区域的关注度也有所提升；边缘区域的图像块对边缘区域始终保持很高的关注度，对背景区域关注度低，呈明显的区分性.

这些观察驱使我们利用这些注意力信息生成关系矩阵 $\boldsymbol{R}$，并与初步解释结果 $\boldsymbol{S}$ 融合，从而生成

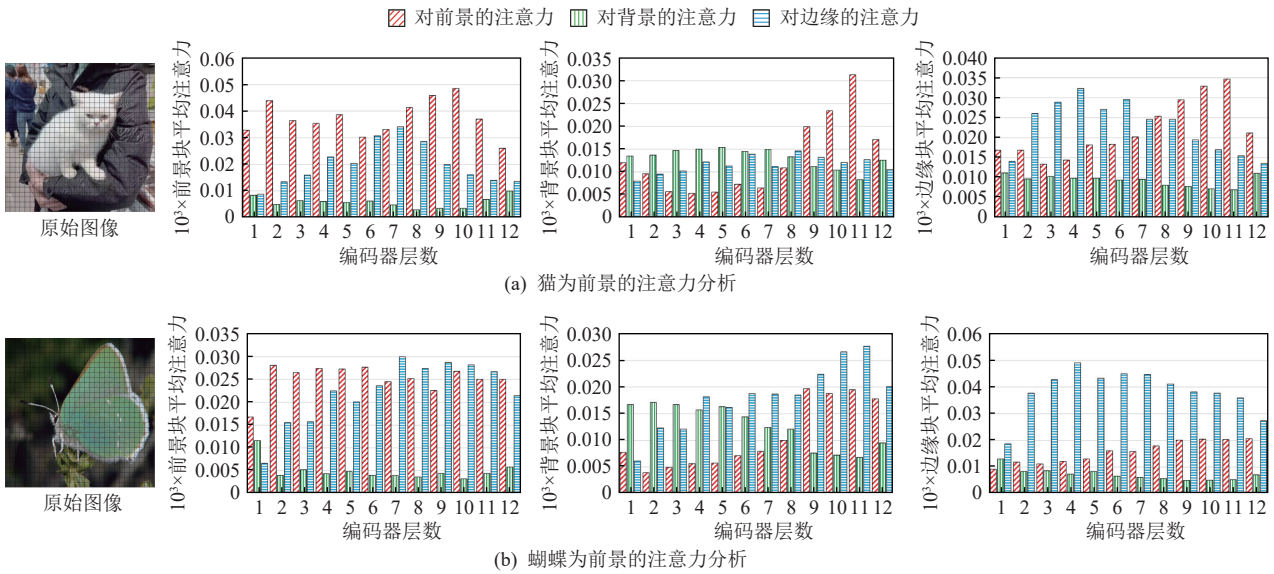


Fig. 4 Results of information interaction analysis of self-attention mechanism

图 4 自注意力机制信息交互分析结果

逐块重要性分数. 通过对图像块逐块加权融合, 进一步细化可视化解释结果, 生成最终可视化解释结果 $\boldsymbol{V}$.

2.3.2 关系矩阵

考虑一个标准 ViT 模块, 其包括 l 层编码器, 每层包含 h 个自注意力头. 我们提取所有 l 层中的 h 个头的注意力矩阵, 并舍弃类别令牌得到 $\boldsymbol{A}\in\mathbb{R}^{l\times h\times N\times N}$. 为了考虑不同编码器层中语义信息, 我们首先从层的维度上进行聚合. 由于第 1 层编码器层尚未获取明确的语义信息, 而最后一层编码器层的语义信息已经趋于平滑, 我们舍弃这 2 层, 聚合剩下的 $l-2$ 层:

$$\boldsymbol{r}=\frac{\sum_{i=2}^{l-1}(\boldsymbol{A}_i-E(\boldsymbol{A}))^2}{l-2}, \tag{15}$$

$$E(\boldsymbol{A})=\frac{\sum_{i=2}^{l-1}\boldsymbol{A}_i}{l-2}, \tag{16}$$

其中 $E(\boldsymbol{A})$ 可以理解为考虑了不同编码器层的一般性关系的注意力矩阵.

利用 $\boldsymbol{A}_i-E(\boldsymbol{A})$ 计算可以得到前景、背景、边缘图像区域的关系差异, $\boldsymbol{r}\in\mathbb{R}^{h\times N\times N}$ 表示考虑了跨编码器层的关系矩阵. 根据文献 [28], 在多头自注意力中, 不同的头部关注的重点不同, 我们采取类似策略对 $\boldsymbol{r}$ 进行聚合得到最终的关系矩阵 $\boldsymbol{R}$:

$$\boldsymbol{R}=\frac{\sum_{i=1}^{h}(\boldsymbol{r}_i-E(\boldsymbol{r}))^2}{h}, \tag{17}$$

$$E(\boldsymbol{r})=\frac{\sum_{i=1}^{h}\boldsymbol{r}_i}{h}, \tag{18}$$

其中 $E(\boldsymbol{r})$ 可以理解为考虑了不同注意力头部的一般性关系的注意力矩阵.

利用 $\boldsymbol{r}_i-E(\boldsymbol{r})$ 计算可以得到前景、背景、边缘图像区域的关系差异, $\boldsymbol{R}\in\mathbb{R}^{N\times N}$ 表示图像全局的关系矩阵, 每行表示当前图像块与其他图像块之间的关系, 列中的值越大, 表明 2 个图像块之间的关系越密切.

我们随机采样了部分前景、背景和边缘图像块, 并选取图像块序号在关系矩阵 $\boldsymbol{R}$ 中对应的行进行重塑操作, 如图 5 第 1 列所示, 正方形块为采样的前景图像块, 圆形块为采样的边缘图像块, 三角形块为采样的背景图像块. 图 5 中第 2~4 列分别是前景、边缘、背景图像块的可视化结果. 观察结果显示, 3 种类型的图像块与自身周围的图像块关系更为密切, 同时也与具有相同属性的远处图像块存在一定的联系,

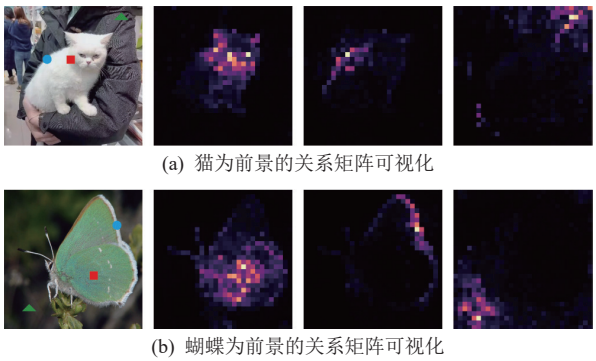


Fig. 5 Visualization result of relation matrix

图 5 关系矩阵可视化结果

这进一步验证了我们在 2.3.1 节中的结论.

2.3.3 注意力融合

由于初步解释结果 $\boldsymbol{S}$ 能够有效定位物体位置，我们将关系矩阵 $\boldsymbol{R}$ 的注意力信息与 $\boldsymbol{S}$ 融合. 具体而言，我们首先将 $\boldsymbol{S}\in\mathbb{R}^{H\times W}$ 下采样至 $\sqrt{N}\times\sqrt{N}$ 后展平为 $1\times N$，接着计算 $\boldsymbol{R}_i\in\mathbb{R}^{i\times N}, i=1,2,\cdots,N$ 与 $\boldsymbol{S}$ 的余弦相似度作为第 i 个图像块的重要性分数：

$$\boldsymbol{P}_i=\frac{\boldsymbol{S}\cdot\boldsymbol{R}_i}{\|\boldsymbol{S}\|\|\boldsymbol{R}_i\|}, i=1,2,\cdots,N, \tag{19}$$

直觉上 $\boldsymbol{P}_i$ 可以理解为第 i 个图像块关联区域与物体位置信息的相关程度. 当第 i 个图像块为前景图像块时，会获取较高的重要性得分；当第 i 个图像块为边缘图像块时，得分会处于中等水平；当第 i 个图像块为背景图像块时，会获取较低的重要性得分.

通过上述方法计算得到重要性得分 $\boldsymbol{P}\in\mathbb{R}^{1\times N}$，我们将其重塑为 $\sqrt{N}\times\sqrt{N}$ 并上采样至 $H\times W$ 的尺寸，接着与 $\boldsymbol{S}$ 进行逐元素相乘，最终得到重要性图 $\boldsymbol{V}$：

$$\boldsymbol{V}(x)=\boldsymbol{P}(x)\odot\boldsymbol{S}(x). \tag{20}$$

可视化实验及消融实验结果表明，通过注意力融合产生的重要性分数可以有效提升可视化解释结果的定位能力.

SAF-Explainer 的完整过程如算法 2 所示.

算法 2. 基于序贯三支掩码和注意力融合的 Transformer 解释方法.

输入：输入图像 $\boldsymbol{I}$；

输出：解释结果 $\boldsymbol{V}$.

① $\boldsymbol{X}$=*Get_Output*($\boldsymbol{I}$)；

② $\boldsymbol{Mask}$=*Upsample*(*Reshape*($\boldsymbol{X}$))；

③ M_C=*Agglomerative_Clustering*($\boldsymbol{Mask}$)；

④ $\boldsymbol{S}$=*S3WM*(M_C)；/*获取初步解释结果*/

⑤ $\boldsymbol{A}$=*Get_Attention*($\boldsymbol{I}$)；

⑥ $\boldsymbol{r}$=**0**；

⑦ for $\boldsymbol{A}_i$ in $\boldsymbol{A}$ do /*聚合注意力矩阵*/

⑧ $\boldsymbol{r}$=$\boldsymbol{r}$+*square*($\boldsymbol{A}_i-E(\boldsymbol{A})$)；

⑨ end for

⑩ $\boldsymbol{r}$=$\boldsymbol{r}$/($l-2$)；/*除以层数取均值*/

⑪ $\boldsymbol{R}$=**0**；

⑫ for $\boldsymbol{head}$ in $\boldsymbol{r}$ do /*聚合生成关系矩阵*/

⑬ $\boldsymbol{R}$=$\boldsymbol{R}$+*square*($\boldsymbol{head}-E(\boldsymbol{r})$)；

⑭ end for

⑮ $\boldsymbol{R}$=$\boldsymbol{R}$/h；

⑯ $\boldsymbol{P}$=*dot*($\boldsymbol{S}$,$\boldsymbol{R}$)/(*norm_2*($\boldsymbol{S}$)·*norm_2*($\boldsymbol{R}$))；

⑰ $\boldsymbol{V}$=$\boldsymbol{P}\odot\boldsymbol{S}$；

⑱ return $\boldsymbol{V}$. /*获取最终解释结果*/

3 实验设置及结果分析

我们首先将所提出的 SAF-Explainer 架构与其他基线架构在定性评估和分割任务上进行比较，以验证其优越性；再去掉 SAF-Explainer 中重要模块进行消融实验，验证架构中所提出模块的有效性.

3.1 实验设置

本文实验使用 4 个数据集进行定性定量评估.

1）ImageNet 2012[38] 的验证集，由来自 1 000 个类别的 5 万张图像组成. 主要用于验证 ViT 模型在自然图像上的可解释性能力.

2）COCO 2017（Microsoft common objects in context 2017）[39] 验证集，包含来自 80 个不同类别的 5 000 个带注释的分割图像.

3）BraTS 2023（brain tumor segmentation 2023）[40-42] 数据集，BraTS 2023 训练数据集包含 1 251 名受试者，每个受试者均有来自 4 种不同磁共振成像模态的 3D 体积：①原生 T1；②对比后 T1 加权；③ T2 加权；④ T2 流体衰减反转恢复，这些模态严格对齐，并以 1 mm × 1 mm × 1 mm 的各向同性分辨率重新采样，输入 3D 体积的大小为 240 × 240 × 155. 我们遵循文献 [43] 中的处理方法，通过沿 z 轴切片每个模态体积以形成总共 177 175 个 2D 图像来预处理数据. 由于官方提供的验证数据集中没有真实分割标签，我们将 BraTS 训练数据集按 9∶1 划分为训练集与验证集，训练集用于 ViT 分类模型的训练，验证集用于验证 SAF-Explainer 模型在医学图像上的可解释性.

4）VOC 2012（pascal visual object classes 2012[44]）数据集，包含来自 20 个类别的 1 449 个图像的带注释分割的验证集.

分割实验将每个解释结果通过阈值进行二值化形成分割结果，通过图像分割指标来衡量模型性能：1）像素准确率（pixel accuracy, *PA*）；2）平均交并比（mean intersection over union, *mIoU*）；3）戴斯相似性系数（dice similarity coefficient, *DSC*）.

PA 为所有预测正确的像素个数与总预测像素个数之比，它易于理解和计算，可以简单直观地反映分割结果的效果，*PA* 的计算公式为：

$$PA=\frac{\sum_i n_{ii}}{\sum_i\sum_j n_{ij}}, \tag{21}$$

其中，n_{ij}为类别 i 被预测成类别 j 的像素个数，n_{cls}为目标类别个数(包括背景).

预测区域和真实区域交集除以预测区域和真实区域的并集可以得到单个类别下的 *IoU*，通过计算出每个类别下的 *IoU* 取平均即可得到 *mIoU*：

$$mIoU=\frac{1}{n_{\text{cls}}}\frac{n_{ii}}{\sum_j n_{ij}+\sum_j n_{ji}-n_{ii}}. \tag{22}$$

DSC 是衡量 2 个集合相似度的指标，它通过计算模型预测结果与真实标签之间的相似程度来度量模型的准确性，取值范围是 [0,1]，值越大说明模型预测结果与真实标签相似度越高，分割效果越好. *DSC* 的计算公式为：

$$DSC=\frac{2\times n_{ii}}{2\times n_{ii}+\sum_{j\neq i} n_{ji}+\sum_{j\neq i} n_{ij}}. \tag{23}$$

本文实验选用超参数$\alpha=0.98$，$\beta=0.1$，$\gamma=0.01$，$\delta=0.15$，$d=0.1$，本文采用的实验平台为 PC(13th Gen Intel® Core™ i9-13900K @ 3.00 GHz)，显卡为 NVIDIA GeForce RTX 4090，内存容量为 64 GB，Windows10 专业版操作系统，开发工具为 JetBrains PyCharm 2022.2.1 专业版，使用 Python 语言实现实验中相关算法.

在 3.2 节和 3.3 节中，我们将 SAF-Explainer 与 Raw Attention[2]，T-Attribution[29]，Grad-CAM[31]，ViT-CX[15] 四个基线进行比较.

3.2 定性评估

本文使用 ImageNet 2012 自然图像数据集和 BraTS 医学图像数据集上进行定性评估. 图 6 展示了 SAF-Explainer 与其他基线方法在单类别自然图像上的解释结果. Grad-CAM 和 Raw Attention 几乎无法定位到图像中的主体部分，T-Attribution 虽然可以定位到主体，但其热图覆盖区域往往只能占到主体的 50%，例如，它仅关注熊与狼狗的脸部或鸽子的头部，而对它们的身体部分关注较少，无法全面展现模型的解释能力. SAF-Explainer 能够实现最为准确的定位效果，这得益于 S3WM 模块中积极掩码的关键作用. 尽管 ViT-CX 能够定位大部分主体对象，但背景区域往往包含大量噪声信息. 在描绘主体边缘的细节时，其效果也不够精确，无法准确描述物体轮廓. 由于 AF 模块对边缘图像块的独特设计，SAF-Explainer 在一些细节处理方面表现更好，如蝴蝶的边缘更为平滑以及熊和狼狗的腿部轮廓描绘得更为准确. 同时，注意力融合也对背景图像块信息进行处理，大大减少了背景中包含的噪声信息.

SAF-Explainer 和基线方法对在多类别自然图像的解释结果如图 7 所示. Raw Attention 获得的结果与类别无关，1 张图片上 2 个不同的类别会生成相同的解释，无法体现模型工作机理. Grad-CAM 虽然能生

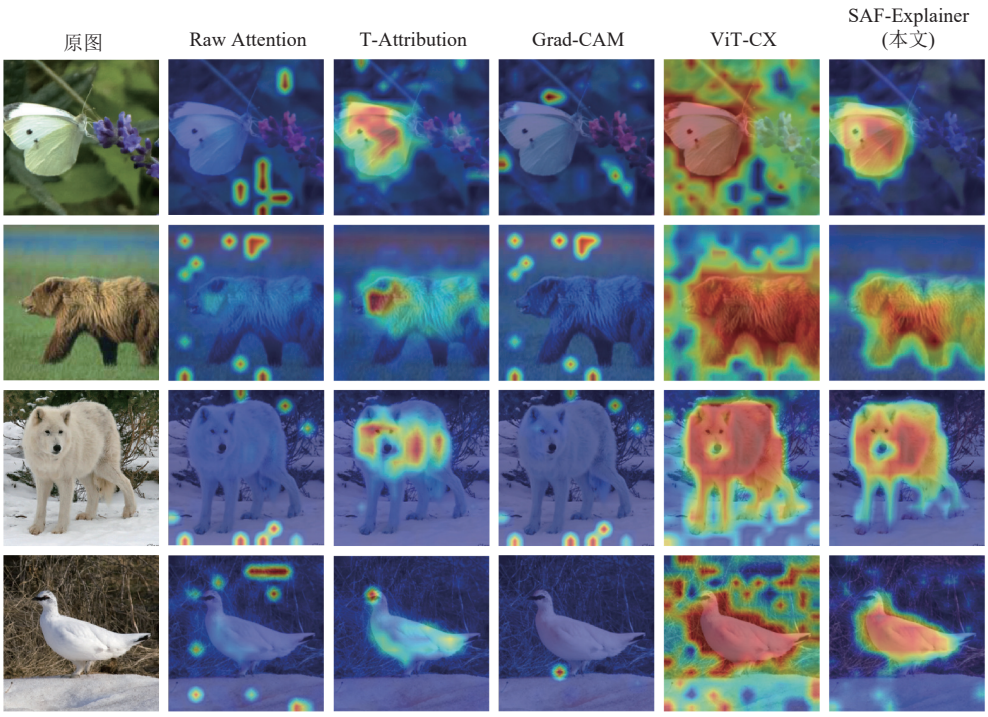


Fig. 6 Interpretation results of single-category natural images

图 6 单类别自然图像解释结果

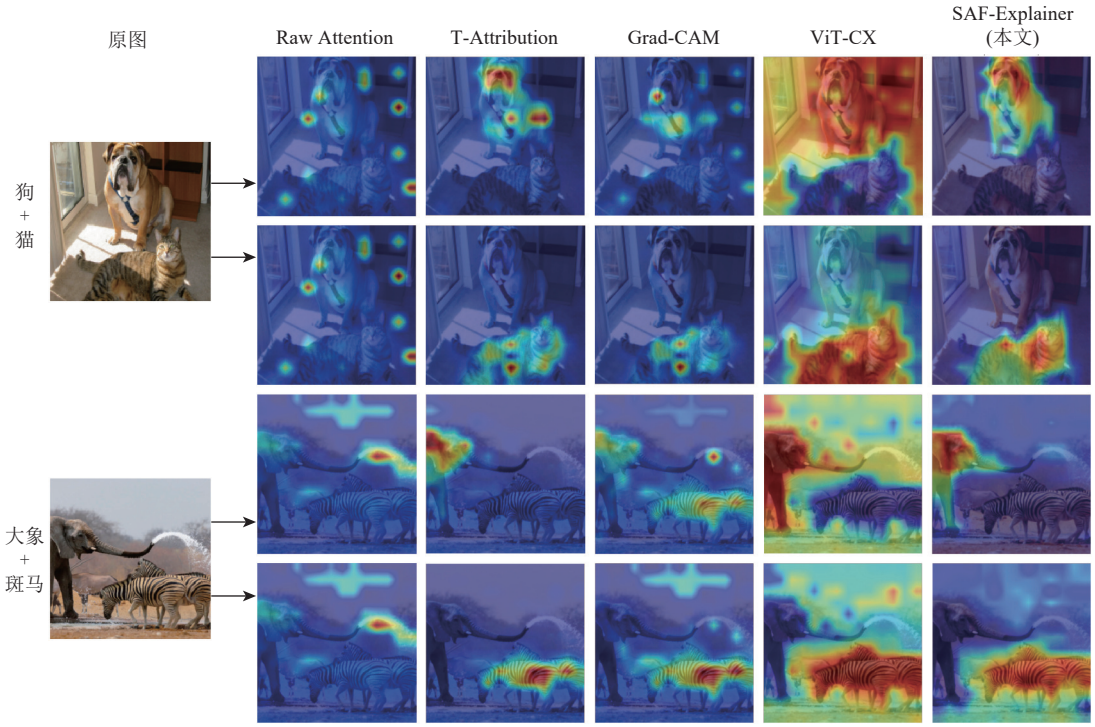


Fig. 7 Interpretation results of multi-category natural images

图 7 多类别自然图像解释结果

成不同类别的解释，但结果通常只包含零散的点，并不令人信服. ViT-CX 对 2 个类别的结果是互补的，不能反映类别特定的特征. 尽管 T-Attribution 方法可以为 2 个类别生成不同的解释图，但它仍然缺乏信心. 相反，SAF-Explainer 可以为 2 个类别提供准确的解释结果.

本文采用在 ImageNet 2012 预训练的 ViT 模块基础上，对磁共振成像脑肿瘤数据集进行 2 分类迁移学习，将图像分为正常图像与含脑肿瘤图像，并对迁移学习后的 ViT 模块进行可解释性评估. 图 8 展示

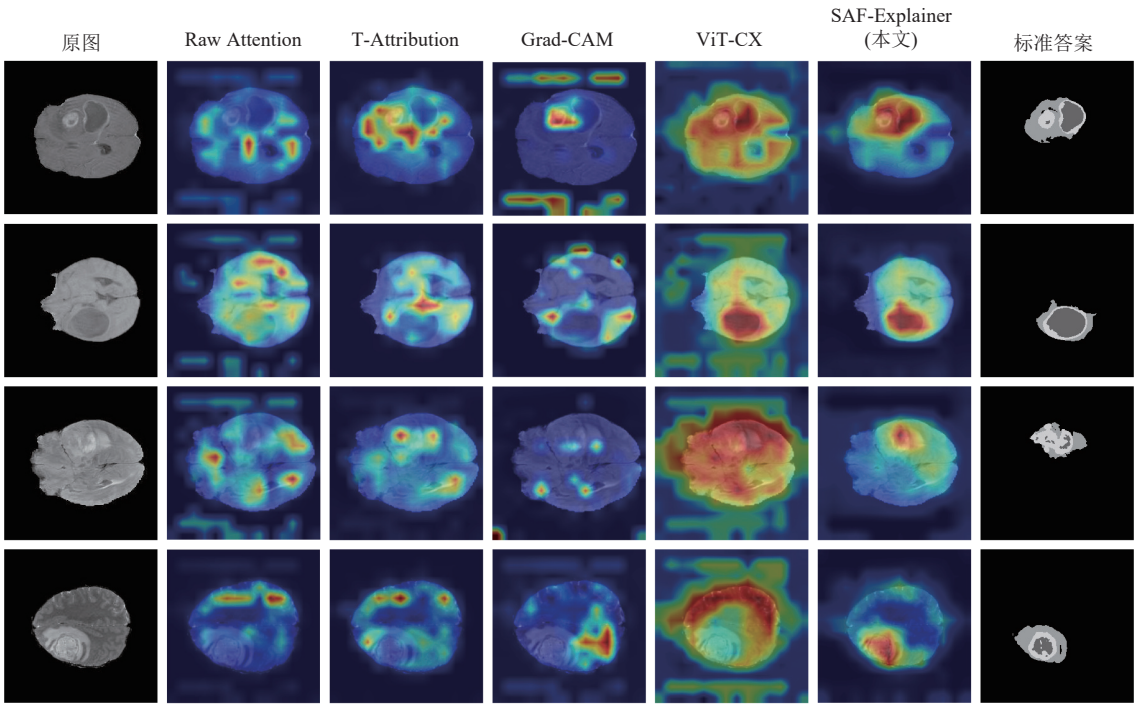


Fig. 8 Interpretation results of medical image

图 8 医学图像解释结果

了 SAF-Explainer 与其他基线方法在医学图像上的解释结果, 其中第 7 列为肿瘤分割标准答案. 由于医学图像数据集较小, 大多数方法难以产生良好的解释结果, Raw Attention 未能捕捉肿瘤位置, T-Attribution 和 ViT-CX 虽然在某些情况下可以定位肿瘤, 但其对肿瘤位置的信心不足, 通常只能定位少部分肿瘤位置, 且其噪声问题依旧存在. Grad-CAM 仅在少数情况下能够定位肿瘤位置. 相比之下, SAF-Explainer 不仅能够准确定位肿瘤位置, 还能够有效减少空间噪声, 表现出更好的解释效果.

3.3 分割实验

我们使用 VOC 2012 数据集和 COCO 2017 数据集进行分割结果评估. 由于数据集中部分图像包含多个标签注释, 我们在评估时将图像中所有带注释的对象均视为真实标签, 因此所有方法的整体性能指标较低. 各方法在 VOC 2012 数据集、COCO 2017 数据集上的分割指标如表 1 和表 2 所示.

Table 1 Comparison of Segmentation Performance of Each Method on VOC 2012 Dataset

表 1 各方法在 VOC 2012 数据集上分割性能对比 %

方法	*PA*	*mIoU*	*DSC*
Raw Attention	62.15	39.61	33.49
T-Attribution	74.01	50.64	46.89
Grad-CAM	65.32	41.25	30.59
ViT-CX	63.39	44.92	48.60
SAF-Explainer（本文）	**74.12**	**52.72**	**50.30**

注：黑体数值表示最优结果.

Table 2 Comparison of Segmentation Performance of Each Method on COCO 2017 Dataset

表 2 各方法在COCO 2017数据集上分割性能对比 %

方法	*PA*	*mIoU*	*DSC*
Raw Attention	62.02	40.10	33.38
T-Attribution	68.95	46.36	39.95
Grad-CAM	65.20	41.55	30.56
ViT-CX	56.30	37.73	38.98
SAF-Explainer（本文）	**69.73**	**46.59**	**40.25**

注：黑体数值表示最优结果.

从表 1 中可以看出, Raw Attention, Grad-CAM, ViT-CX 在 VOC 2012 数据集上的 *PA* 和 *mIoU* 表现相对较差. 原因在于 Raw Attention 只使用了模型注意力信息, Grad-CAM 仅依赖模型梯度信息, 而 ViT-CX 仅使用模型输出信息, 这些单一信息源不足以提供全面的解释. T-Attribution 结合了注意力信息和梯度信息, 取得了更好的效果, 但在图像细节处理方面仍然存在不足. SAF-Explainer 通过融合模型输出和注意力信息, *PA*, *mIoU*, *DSC* 指标均达到了最优值, 分别为 74.12%, 52.72%, 50.30%, 超过次优方法 0.11 个百分点、2.08 个百分点、1.7 个百分点.

由表 2 可知, COCO 2017 数据集由于类别更多, 导致各解释方法的分割性能相比 VOC 2012 数据集有所下降. 只使用单一模型信息的方法表现依旧较差, SAF-Explainer 在 *PA*, *mIoU*, *DSC* 指标上分别达到 69.73%, 46.59%, 40.25%, 超过次优值 0.78 个百分点、0.23 个百分点、0.3 个百分点. T-Attribution 在各指标上可以达到次优值, 这主要是由于它综合利用了梯度信息与注意力信息. 结合图 6 可以看出, 由于 T-Attribution 对解释细节的处理不足, 生成的解释结果过于保守, ViT-CX 生成的解释图虽然可以定位到主体对象, 但由于缺乏针对 ViT 特性的处理, 存在大量的空间噪声. SAF-Explainer 能够更精确地定位图像中的物体位置, 同时在主体边缘细节的描绘上表现出色, 且包含较少的空间噪声.

3.4 消融实验

为了验证所提模块的有效性, 我们设计了 3 种方法的变体, 并在 VOC 2012 和 COCO 2017 数据集进行分割实验, 各模块的消融实验结果如表 3 所示. 1)“w/o S3WM”, 即剔除序贯三支掩码模块; 2)“w/o AF”, 即剔除注意力融合模块; 3)“w/o S3WM, AF”, 即同时剔除序贯三支掩码模块和注意力融合模块.

Table 3 Results of Ablation Experiments on VOC 2012 and COCO 2017 Datasets

表 3 VOC 2012与COCO 2017数据集的消融实验结果 %

方法	*PA*		*mIoU*		*DSC*	
	VOC 2012	COCO 2017	VOC 2012	COCO 2017	VOC 2012	COCO 2017
w/o S3WM	72.84	67.64	51.02	45.54	47.89	39.93
w/o AF	71.09	69.06	50.91	45.78	49.51	39.01
w/o S3WM, AF	69.50	66.86	49.17	44.79	47.63	38.97
SAF-Explainer（本文）	**74.12**	**69.73**	**52.72**	**46.59**	**50.30**	**40.25**

注：黑体数值表示最优结果.

移除序贯三支掩码模块(即 w/o S3WM)后, VOC 2012 和 COCO 2017 数据集的 *PA* 分别下降 1.28 个百分点和 2.09 个百分点, *mIoU* 下降 1.7 个百分点和 1.05 个百分点, *DSC* 下降 2.41 个百分点和 0.32 个百分点. 由此可以看出, S3WM 模块有效提高了模型对物体

位置的定位能力. 通过 S3WM 模块处理掩码质量不确定性问题, 筛选积极掩码, 并利用这些掩码对图像进行扰动, 可以显著提高模型的解释效果.

此外, 我们也尝试同时剔除序贯三支掩码模块与注意力融合模块(即 w/o S3WM, AF). 结果显示, 3 项指标在 VOC 2012 和 COCO 2017 数据集上的分割性能相较于单独剔除 S3WM 或 AF 模块时进一步下降. 这验证了 2 个模块叠加的有效性, 并进一步说明了 S3WM 模块在处理掩码质量不确定性问题和筛选积极掩码方面的重要性. 同时, 跨层注意力信息聚合生成的关系矩阵能够有效优化解释结果的细节信息.

4 结论与展望

Transformer 模型在计算机视觉领域的重要性日益增加, 对模型完整与准确的可解释性需求也大大提高. 然而, 目前关于 Transformer 可解释性的研究有限. 本文提出 SAF-Explainer, 它基于掩码解释方法设计, 主要提出了 S3WM 模块, 通过特定的阈值条件设定解决了掩码质量不确定性问题, 提出了 AF 模块, 通过聚合注意力矩阵生成关系矩阵, 解决了解释结果包含大量噪声的问题, 以优化解释结果中的细节信息. 在 ImageNet 2012, VOC 2012, COCO 2017, BraTS 2023 数据集上的实验表明, 我们的方法 SAF-Explainer 无论在自然图像还是在医学图像上表现均优于其他基线方法. 由于 SAF-Explainer 是针对视觉领域的 Transformer 设计, 在未来的研究中, 我们会尝试延伸 SAF-Explainer 到其他领域如自然语言处理和多模态领域. 例如, 在自然语言处理领域, Transformer 模型被训练用于文本情感分类. 我们可以通过生成重要性热图, 展示文本中各个词语对模型预测的贡献程度. 这些热图有助于理解模型做出特定预测的原因.

作者贡献声明: 成晓天提出了算法核心思想, 设计了实验方案, 完成实验并撰写论文; 丁卫平提出了整个算法的框架并对整个算法思想进行完善, 指导了论文修改; 耿宇、黄嘉爽、鞠恒荣完善了算法的思路, 指导了论文写作并修改论文; 郭静协助完成部分实验及完善论文内容.

参 考 文 献

[1] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C] //Proc of the 31st Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2017: 6000−6010

[2] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint, arXiv: 2010.11929, 2020

[3] Chen C F R, Fan Quanfu, Panda R. Crossvit: Cross-attention multi-scale vision Transformer for image classification[C] //Proc of the 18th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 357−366

[4] Wang Wenhai, Xie E, Li Xiang, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions[C] //Proc of the 18th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 568−578

[5] Ding Weiping, Wang Haipeng, Huang Jiashuang, et al. FTransCNN: Fusing Transformer and a CNN based on fuzzy logic for uncertain medical image segmentation[J]. Information Fusion, 2023, 99: 101880

[6] Liu Ze, Lin Yutong, Cao Yue, et al. Swin Transformer: Hierarchical vision Transformer using shifted windows[C] //Proc of the 18th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 10012−10022

[7] Zhou Qianyu, Li Xiangtai, He Lu, et al. TransVOD: End-to-end video object detection with spatial-temporal transformers[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(6): 7853−7869

[8] Ding Weiping, Liu Chuansheng, Huang Jiashuang, et al. ViTH-RFG: Vision Transformer hashing with residual fuzzy generation for targeted attack in medical image retrieval[J]. IEEE Transactions on Fuzzy Systems, 2023, 32(10): 5571−5584

[9] Cheng Keyang, Wang Ning, Shi Wenxi, et al. Research advances in the interpretability of deep learning[J] Journal of Computer Research and Development, 2020, 57(6): 1208−1217 (in Chinese)
(成科扬, 王宁, 师文喜, 等. 深度学习可解释性研究进展[J]. 计算机研究与发展, 2020, 57(6): 1208−1217)

[10] Ma Liantao, Zhang Chaohe, Jiao Xianfeng, et al. Dr. Deep: Interpretable evaluation of patient health status via clinical feature's context learning [J] Journal of Computer Research and Development, 2021, 58(12): 2645−2659 (in Chinese)
(马连韬, 张超贺, 焦贤锋, 等. Dr. Deep: 基于医疗特征上下文学习的患者健康状态可解释评估[J]. 计算机研究与发展, 2021, 58(12): 2645−2659)

[11] Zhou Tianyi, Ding Weiping, Huang Jiashuang, et al. Fuzzy logic guided deep neural network with multi-granularity[J] Pattern Recognition and Artificial Intelligence, 2023, 36(9): 778−792 (in Chinese)
(周天奕, 丁卫平, 黄嘉爽, 等. 模糊逻辑引导的多粒度深度神经网络[J]. 模式识别与人工智能, 2023, 36(9): 778−792)

[12] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C] //Proc of the 13th European Conf on Computer Vision. Berlin: Springer, 2014: 818−833

[13] Ding Weiping, Geng Yu, Huang Jiashuang, et al. MGRW-Transformer: Multigranularity random walk Transformer model for interpretable learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 36(1): 1104−1118

[14] Petsiuk V, Das A, Saenko K. Rise: Randomized input sampling for explanation of black-box models[J]. arXiv preprint, arXiv:

1806.07421, 2018

[15] Xie Weiyan, Li Xiaohui, Cao C C, et al. ViT-CX: Causal explanation of vision transformers[C] //Proc of the 32nd Int Joint Conf on Artificial Intelligence. San Francisco: Margan Kaufmann, 2023: 1569−1577

[16] Wang Peihao, Zheng Wenqing, Chen Tianlong, et al. Anti-Oversmoothing in deep vision Transformers via the Fourier domain analysis: From theory to practice[C/OL] //Proc of the 10th Int Conf on Learning Representations. La Jolla, CA: ICLR, 2022[2023-03-15]. https://openreview.net/forum?id=O476oWmiNNp

[17] Ru Lixiang, Zheng Heliang, Zhan Yibing, et al. Token contrast for weakly-supervised semantic segmentation[C] //Proc of the 36th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 3093−3102

[18] Yao Yiyu. Three-way decisions with probabilistic rough sets[J]. Information Sciences, 2010, 180(3): 341−353

[19] Savchenko A V. Sequential three-way decisions in multi-category image recognition with deep features based on distance factor[J]. Information Sciences, 2019, 489: 18−36

[20] Ju Hengrong, Pedrycz W, Li Huaxiong, et al. Sequential three-way classifier with justifiable granularity[J]. Knowledge-Based Systems, 2019, 163: 103−119

[21] Li Huaxiong, Zhang Libo, Huang Bing, et al. Sequential three-way decision and granulation for cost-sensitive face recognition[J]. Knowledge-Based Systems, 2016, 91: 241−251

[22] Li Jinhai, Li Yufei, Mi Yunlong, et al. Meso-Granularity labeled method for multi-granularity formal concept analysis[J]. Journal of Computer Research and Development, 2020, 57(2): 447−458 (in Chinese)
(李金海, 李玉斐, 米允龙, 等. 多粒度形式概念分析的介粒度标记方法[J]. 计算机研究与发展, 2020, 57(2): 447−458)

[23] Zhang Sulan, Guo Ping, Zhang Jifu, et al. Automatic semantic image annotation with granular analysis method[J]. Acta Automatica Sinica, 2012, 38(5): 688−697 (in Chinese)
(张素兰, 郭平, 张继福, 等. 图像语义自动标注及其粒度分析方法[J]. 自动化学报, 2012, 38(5): 688−697)

[24] Abnar S, Zuidema W. Quantifying attention flow in Transformers[C] //Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 4190−4197

[25] Smilkov D, Thorat N, Kim B, et al. Smoothgrad: Removing noise by adding noise[J]. arXiv preprint, arXiv: 1706.03825, 2017

[26] Sundararajan M, Taly A, Yan Qiqi. Axiomatic attribution for deep networks[C] // Proc of the 34th Int Conf on Machine Learning. New York: ACM, 2017: 3319−3328

[27] Binder A, Montavon G, Lapuschkin S, et al. Layer-wise relevance propagation for neural networks with local renormalization layers[C] //Proc of the 25th Int Conf on Artificial Neural Networks and Machine Learning. Berlin: Springer, 2016: 63−71

[28] Voita E, Talbot D, Moiseev F, et al. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned[C] //Proc of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019: 5797−5808

[29] Chefer H, Gur S, Wolf L. Transformer interpretability beyond attention visualization[C] //Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 782−791

[30] Zhou Bolei, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization[C] //Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 2921−2929

[31] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization[C] //Pro of the 14th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 618−626

[32] Chattopadhay A, Sarkar A, Howlader P, et al. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks[C] //Proc of the 6th IEEE Winter Conf on Applications of Computer Vision (WACV). Piscataway, NJ: IEEE, 2018: 839−847

[33] Wang Haofan, Wang Zifan, Du Mengnan, et al. Score-CAM: Score-weighted visual explanations for convolutional neural networks[C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2020: 24−25

[34] Jiang Pengtao, Zhang Changbin, Hou Qibin, et al. LayerCAM: Exploring hierarchical class activation maps for localization[J]. IEEE Transactions on Image Processing, 2021, 30: 5875−5888

[35] Chen Kaitao, Sun Shiliang, Du Youtian. Deconfounded multi-organ weakly-supervised semantic segmentation via causal intervention[J]. Information Fusion, 2024, 108: 102355

[36] Yang Jiaqi, Mehta N, Demirci G, et al. Anomaly-guided weakly supervised lesion segmentation on retinal OCT images[J]. Medical Image Analysis, 2024, 94: 103139

[37] Ma Jie, Bai Yalong, Zhong Bineng, et al. Visualizing and understanding patch interactions in vision transformer[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 35(10): 13671−13680

[38] Deng Jia, Dong Wei, Socher R, et al. ImageNet: A large-scale hierarchical image database[C] // Proc of the 22nd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2009: 248−255

[39] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context[C] // Proc of the 13th European Conf on Computer Vision. Berlin: Springer, 2014: 740−755

[40] Baid U, Ghodasara S, Mohan S, et al. The RSNA-ASNR-MICCAI brats 2021 benchmark on brain tumor segmentation and radiogenomic classification[J]. arXiv preprint, arXiv: 2107.02314, 2021

[41] Menze B H, Jakab A, Bauer S, et al. The multimodal brain tumor image segmentation benchmark (BRATS)[J]. IEEE Transactions on Medical Imaging, 2014, 34(10): 1993−2024

[42] Bakas S, Akbari H, Sotiras A, et al. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features[J]. Scientific Data, 2017, 4(1): 1−13

[43] Chen Y J, Hu Xinrong, Shi Yiyu, et al. AME-CAM: Attentive multiple-exit CAM for weakly supervised segmentation on MRI brain tumor[C] // Proc of the 26th Int Conf on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2023: 173−182

[44] Everingham M, Van Gool L, Williams C K I, et al. The pascal visual object classes (VOC) challenge[J]. International Journal of Computer Vision, 2010, 88: 303−338

Cheng Xiaotian, born in 2001. Master candidate. His main research interests include granular computing, deep learning, and computer vision.

成晓天, 2001 年生. 硕士研究生. 主要研究方向为粒计算、深度学习、计算机视觉.

Ding Weiping, born in 1979. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include data mining, machine learning, granular computing, evolutionary computing, and big data analytics.

丁卫平, 1979 年生. 博士, 教授, 博士生导师. CCF 高级会员. 主要研究方向为数据挖掘、机器学习、粒计算、演化计算、大数据分析.

Geng Yu, born in 1998. Master. His main research interests include granular computing, machine learning, and deep learning.

耿　宇, 1998 年生. 硕士. 主要研究方向为粒计算、机器学习、深度学习.

Huang Jiashuang, born in 1988. PhD, associate professor. His main research interests include brain network analysis and deep learning.

黄嘉爽, 1988 年生. 博士, 副教授. 主要研究方向为脑网络分析、深度学习.

Ju Hengrong, born in 1989. PhD, associate professor. His main research interests include granular computing, rough sets, machine learning, and knowledge discovery.

鞠恒荣, 1989 年生. 博士, 副教授. 主要研究方向为粒计算、粗糙集、机器学习、知识发现.

Guo Jing, born in 2000. Master candidate. Her main research interests include granular computing, machine learning, and deep learning.

郭　静, 2000 年生. 硕士研究生. 主要研究方向为粒计算、机器学习、深度学习.