MGRW-Transformer:多粒度随机游走可解释性Transformer模型

耿 宇,丁卫平*,黄嘉爽,鞠恒荣,孙 颖,王海鹏
(南通大学信息科学技术学院,江苏南通 226019)

摘 要: 深度学习模型凭借特征学习能力应用于图像识别任务,但由于缺乏对工作机制的语义解释,因此难以识别复杂医学图像.Vision Transformer模型的自注意力机制具备可解释性.然而,医学图像中的病灶区域往往存在位置多变且大小不定等现象,这使得单纯依靠自注意力模块的深度学习模型难以提供有效的语义解释.为此,本文提出基于多粒度随机游走的可解释性Transformer模型(Multi-Granularity Random Walk Transformer Model For Interpretable Learning,MGRW-Transformer)寻找对识别任务重要的区域.具体来说,首先将图像划分多个子图像块,输入到Vision Transformer中的多头注意力层输出注意力矩阵,然后将图像块作为结点构建无向图,将注意力指引结点作为游走起点进行粗粒度随机游走,接着将每个粗信息粒划分为更细的图像块进行细粒度随机游走,最后根据信息重要度选取最优粗、细信息粒集合并约简融合.综上便可获取输入图像的可视化语义解释效果.本文在自然图像和医学图像两类数据集上对MGRW-Transformer模型进行了验证,在ImageNet-segmentation数据集上比现有方法的pixel accuracy提高了8.09%,mIoU提高了13.82%,在医学图像数据集上能得到合理语义解释.

关键词: 多粒度分析方法;可解释性方法;Vision Transformer;自注意力机制;扰动遮挡法;图随机游走

基金项目: 国家自然科学基金(No.61976120,No.62006128,No.62102199);江苏省自然科学基金(No.BK20191445);江苏省双创博士计划(No.(2020)30986);江苏省高等学校自然科学研究重大项目(No.21KJA510004);南通市科技局基础科学研究项目(No.JC2021122)

中图分类号: TP18 **文献标识码:** A **文章编号:** 0372-2112(XXXX)XX-0001-15

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20221181

MGRW-Transformer: Multi-Granularity Random Walk Transformer Model for Interpretable Learning

GENG Yu, DING Wei-ping*, HUANG Jia-shuang, JU Heng-rong, SUN Ying, WANG Hai-peng
(*School of Information Science and Technology*, *Nantong University*, *Nantong*, *Jiangsu* 226019,*China*)

Abstract: Deep learning model is applied to image recognition task with feature learning ability, but it is difficult to recognize complex medical images due to lack of semantic interpretation of working mechanism. The vision transformer model with a self-attention mechanism offers great interpretability. However, medical images often contain lesions of variable size in different locations, which makes it difficult for a deep learning model with a self-attention module to reach correct and explainable conclusions. We propose a multi-granularity random walk transformer model (MGRW-Transformer) guided by an attention mechanism to find the regions that influence the recognition task. Our method divides the image into multiple sub-image blocks and transfers them to the vision transformer module for classification. The segmented image blocks are used as nodes to construct an undirected graph using the attention node as a starting node and guiding the coarse-grained random walk. We appropriately divide the coarse blocks into finer ones to manage the computational cost and combine the results based on the importance of the discovered features. The result is that the model offers a semantic interpretation of the input image, a visualization of the interpretation. In this paper, the MGRW-Transformer model is verified on natural image and medical image data sets, and the pixel accuracy and mIoU of the ImageNet-segmentation data sets are improved by 8.09% and 13.82%, respectively. Reasonable semantic interpretation can be obtained in medical image data set.

收稿日期:2022-10-20;修回日期:2023-03-13;责任编辑:梅志强
*通讯作者:丁卫平

Key words: multi-granularity formal analysis; Interpretable method; vision transformer; self-attention mechanism; disturbance occlusion method; graph random walk

Foundation Item(s): National Natural Science Foundation of China (No.61976120, No.62006128, No.62102199); Natural Science Foundation of Jiangsu Province (No.BK20191445); Double-Creation Doctoral Program of Jiangsu Province (No.(2020)30986); Natural Science Key Foundation of Higher Education of Jiangsu Province (No.21KJA510004); Basic Science Research Program of Nantong Science and Technology Bureau (No.JC2021122)

1 引言

近年来,深度学习网络凭借强大的特征学习能力、优越的性能,在计算机视觉、自然语言处理、推荐系统等领域得到了广泛应用,但由于深度学习网络高维度的特征,深度学习网络中特征、权重等含义模糊不清,决策边界如何划定不为人知,使得深度学习网络缺乏其工作机制的可解释性,难以运用于一些风险高、容错率低的任务中. 例如医学图像的辅助诊断中,出于病人安全的考虑,可解释性的缺乏限制了医生对深度学习网络结果的信任程度,故在深度学习领域,医学图像辅助诊断已成为众多学者研究的课题之一.

医学图像是临床医学辅助诊断的一种重要工具,如乳腺癌检测[1]、肺癌检测[2]、视网膜检测[3]等,深入挖掘医学图像特征信息有助于医生临床诊断. 模型的准确率是衡量模型是否学习到医学图像的特征信息以及能否有助于医生临床诊断的一个重要指标. Rustam等人[4]提出将卷积神经网络(Convolutional Neural Network,CNN)与内核k均值算法(k-means clustering algorithm)相结合的方法提高了肺癌图像分类的准确率;Shi等人[5]提出了一种深度卷积神经网络的迁移学习方法提高肺癌分类准确率的同时降低了假阳性(FP)的概率;Zhao等人[6]提出一种弱监督多实例学习方法,通过引入多分辨率期望最大化卷积神经网络来定位感兴趣病变区域(Region Of Interest,ROI),提高了肺癌分类的准确率. 以上众多医学图像诊断方法大多基于卷积神经网络,近期,视觉转换器(Vision Transformer,ViT)[7]模型已在自然图像数据集中拥有与卷积神经网络相当甚至更好的性能;ViT模型已成功应用于医学图像诊断中[8~11],然而ViT模型由于缺少归纳偏置的能力,所以在中小型数据集上预训练的结果要略低于传统的卷积神经网络,Matsoukas等人[12]提出ViT模型使用迁移学习预训练参数,在医学图像分类中拥有与卷积神经网络相当的表现,而当使用自监督方法进行预训练时,ViT模型拥有更好的分类性能.

医疗图像辅助诊断需要拥有较好的可解释性才能获得医生、患者以及相关机构的认可,才能安全有保障的运用到医学诊断中. 面向医学图像辅助诊断任务的深度学习模型不仅需要拥有较高的性能,还需要额外增加可解释性算法,使得深度学习模型具备一定的可解释性,Papanastasopoulos等人[13]采用可解释性人工智能(Explainable Artificial Intelligence,EAI)来可视化深度卷积神经网络(Deep-learning Convolutional Neural Nnetworks,DCNN)中的重要特征用于乳房雌激素受体状态分类;Zhang等人提出医学影像诊断网络(Medical Image Diagnosis Network,MDNet)[14]在医学图像和诊断报告之间建立一个直接的多模态映射,通过症状描述检索图像得到可视化注意力结果并生成诊断报告,为网络诊断过程提供依据;Lin等人提出了生物可见神经网络(Biological Visible Neural Network,BioVNN)[15],利用路径知识来预测癌症依赖,可解释的VNN有助于了解癌症依赖和开发靶向治疗.

然而目前面向医学图像辅助诊断的可解释性方法存在计算成本高,病灶区域位置多变且大小不定等问题使得诸多可解释性算法例如梯度反向传播法、显著性映射法、扰动遮挡法以及注意力法等难以运用;在传统的医学图像任务中,可解释性任务往往依靠医生的主观观察和手动ROI医学图像中可能存在的病变的区域,不仅耗时长,而且手动ROI病变的区域往往存在主观性差异甚至存在错标漏标等问题,从而影响医学图像分析结果,耽误患者治疗.

因此基于现有研究存在的上述问题,本文提出MGRW-Transformer模型,针对医学图像病灶区域位置多变等问题,本文基于注意力机制改进了随机游走算法,通过注意力热图有效定位重要像素特征位置,减少无效游走的次数,多种停止条件进一步降低了随机游走算法的复杂程度,能够在一定规模下得到对模型分类重要的像素特征;针对扰动遮挡法对原始图像进行逐像素遮挡计算成本高等问题,本文采用多粒度分析方法,将原始图像切分为多个图像块,通过多粒度随机游走算法得到重要度较高的像素特征,同时为了减少拆分图像块而导致医学图像病灶信息不完整对特征重要度的影响,本文采用了融合约简方法,将图像块中的部分可解释性结果拼接成原始图像的可解释性图;针对粗粒度可解释性模型以及传统可解释性算法得到的可解释性结果图包含大量噪声且难以有效定位病灶信息等问题,细粒度可解释性模型选用了较小的约简阈值,使得标注的结果能够集中在病灶区域,有效的解释分类结果. 多粒度分析方法可以通过整体、轮廓、边缘

等特征信息多粒度分析图像,Feng等人[16]提出多粒度卷积神经网络用于患者预后检测;Wang等人[17]开发了多粒度尺度感知网络实现肺结节的分割;本文通过粗粒度随机游走方法得到重要度较大的信息粒,再通过细粒度随机游走方法来获取并标注粗粒度中对模型分类重要的像素特征,最后通过约简融合的方法得到原始图像的可解释性结果图.

本文的主要工作如下:

(1)提出一种基于自注意力机制的图随机游走算法,该算法从随机游走的游走开始结点、停止条件以及评价函数三个方面进行改进,解决了注意力方法指引不准确,扰动遮挡法计算成本高等问题.

(2)采用多粒度分析方法,根据信息粒的重要度从粗到细选取并可视化重要的特征信息,解决了选用过小信息粒使得网络规模大、计算成本高,而选用过大信息粒无法捕捉大小不定的特征等问题.

(3)综合ImageNet自然图像数据集和肺癌医学图像数据集对本文提出的可解释性模型进行验证,结合多种指标与可视化结果得出本文提出的可解释性模型具有更好的可解释性.

2　相关研究

面向机器学习中分类模型的可解释性算法主要分为事先可解释性算法和事后可解释性算法,事先可解释性算法主要体现模型本身具备一定的可解释性例如决策树等机器学习算法;而事后可解释性算法是指模型本身不具备可解释性,额外增加可解释性算法后使得模型具备一定的可解释性,深度学习网络往往采用后者,故本文在这一小节将主要介绍事后可解释性算法中的梯度反向传播法、显著性映射法、扰动遮挡法以及注意力法等可解释性方法的研究进展及其不足之处.

2.1　梯度反向传播法

梯度反向传播可解释性方法主要通过深度学习网络中输入信息的变化对于输出的影响,来计算输入特征对于输出决策的重要性,最终输出分类显著性图作为可视化可解释性结果.

Binder等人提出的相关性分数逐层传播(Layer-wise Relevance Propagation,LRP)[18]算法将一阶或更高阶泰勒展开到非线性神经网络中,得到了层与层之间相关性计算准则,然后再逐层计算前一层的相关性,最终得到输入与输出之间的相关性来解释图像中哪些像素对于分类决策起到重要的作用;Elena等人提出的局部相关性分数逐层传播(Partial Layer-wise Relevance Propagation,Partial LRP)[19]算法将可解释性网络结构拓展到Transformer模型中,通过逐层关联传播验证多头注意力机制中各部分的相对贡献不同,为确定少部分对模型分类重要的可解释性头,Partial LRP引入新的方法来修剪注意力头使得可解释性结果更加准确;Chefer等人[20]通过对Transformer自注意力层深度泰勒分解,将LRP算法运用到Vision Transformer模型中,使用非正激活函数、频繁跳跃连接以及在自注意力模块中使用矩阵乘法运算等方法解决了Vision Transformer模型中自注意力可解释性结果分散、难以信服等问题,最终用热力图的形式解释了自然图像数据分类结果;Lee等人提出了相关性加权类激活映射(Relevance-weighted Class Activation Mapping, Relevance-CAM)[21]算法,针对传统的类激活映射法梯度溢出、置信度低等问题,结合了梯度反向传播法以及类激活映射法,将LRP的相关性分数作为类激活映射权重成分,同样通过类激活映射图解释模型的分类结果.

基于梯度反向传播的可解释性算法由于输入梯度反应了损失函数变化最快的方向,故该类型算法的优点在于简单高效,然而存在着可解释性结果噪声多的现象,甚至当输入图像存在多个类别的对象时,该算法会可视化标注出所有的对象,这对于可解释性问题产生了极大的干扰.

2.2　显著性映射法

显著性映射可解释性方法往往将特征图作为初始信息输入,计算特征之间的权重信息后得到类别热力图,最后再与原始特征图像叠加后得到可视化解释结果图像.

Zhou等人最先提出类激活映射(Class Activation Mapping,CAM)[22]模型,通过将全连接层替换为全局平均池化层,输出最后一层卷积层中特征图的均值,最后再加权求和得到类别热力图,用于解释分类的结果.然而CAM模型修改了深度学习网络结构,存在着重新训练深度学习网络成本高、可移植性较差等问题,因此Selvaraju等人提出了基于梯度的类激活映射(Gradient-weighted Class Activation Mapping,GradCAM)[23]模型,该模型将卷积神经网络输出的分类结果的梯度输入到最后一层卷积层,与CAM不同,GradCAM模型对特征图的梯度求均值作为权重,得到最终与CAM模型等效的类别热力图,用于解释分类结果.GradCAM模型无需更改网络结构,可移植性好且适用于众多卷积神经网络框架.针对GradCAM模型多目标定位不够准确等问题,Chattopadhyay等人提出了GradCAM++[24]模型,考虑在梯度图中每个像素贡献度不同,该模型增加额外的输出梯度对部分像素进行加权,在保证计算量不变的情况下,得到了更为精准的定位和可解释性结果;Wang等人提出了基于类激活映射的卷积神经网络的分数加权视觉解释(Score-Weighted Visual Explanations for Con-

volutional Neural Networks based on Class Activation Mapping, Score-CAM)[25]算法,针对传统类激活映射法包含大量噪声是梯度溢出导致的,Score-CAM算法摆脱了对梯度的依赖,主要通过特征图的全局置信分数来衡量线性权重解释分类结果.

显著性映射法优点在于计算速度较快,但也存在着容易标记大量背景信息,对于细微细节难以突出等问题,例如在位置多变且大小不定的医学图像任务中解释性方法缺乏可靠性.

2.3 扰动遮挡法

扰动遮挡法通过随机生成的掩膜(Mask)来扰动遮挡原始图像的特征,测量比较模型扰动前与扰动后的差异,进而判断特征的重要程度,作为图像分类的特征证据.

Zeiler等人可视化反卷积网络(Deconvolution Network, DeConvNet)[26]中各隐藏层,并在模型深浅层分别提取到图像不同粗细粒度信息,通过扰动遮挡输入图像的不同像素,提取出对模型分类结果影响较大的特征;Petsiuk等人提出的随机输入采样解释性模型(Randomized Input Sampling for Explanation, RISE)[27]使用蒙特卡洛采样将多个掩码与输入图像逐元相乘,得到遮挡图像的预测重要度,将重要度得分与掩膜图进行加权取均值后得到最终可解释性图像.

扰动遮挡法相较于其他可解释性算法定位更准,部分模型中可解释性结果更细致,然而该可解释性方法也存在解释性结果不直观、计算复杂度较高、模型整体效率不高等问题.

2.4 注意力法

由于注意力层可以计算层的表征权重,基于注意力的可解释性算法通过模型训练得到的注意力系数绘制注意力热图将重要的特征表现出来.

Dosovitskiy等人最初提出了Vision Transformer模型,通过Vision Transformer模型中多头自注意力模块生成的注意力热图解释分类结果即传统注意力方法(Raw-attention算法);张等人提出了一种通道注意机制实现前后端融合网络(Frontend-backend Fusion network through Channel-Attention Mechanism, FF-CAM)[28]模型基于通道注意力机制计算并可视化图像中人群数量;Bamba等人[29]采用将解码器中的注意力门突出重要特征进行脑图像分割.

Serrano等人[30]提出基于梯度的注意力权重等级能更好地预测其影响,但在许多方面这并不成立,注意力法绝不是安全的方法,所以注意力法与显著性映射法相同不需要进行大量的计算各个特征的重要度,但对于例如数据量小,输入图像位置不定、大小不一的医学图像等任务中难以获得较好的效果.

3 多粒度随机游走可解释性Transformer模型

针对上述梯度反向传播法噪声多、扰动遮挡法计算成本高、显著性映射法和注意力法指引不准确等问题,本文提出了MGRW-Transformer模型如图1所示. 该模型的基本思路如下:首先将输入图片中心裁剪并修改图像尺寸为224×224像素,其次在粗粒度下,将整张图片切割成多个32×32像素的图像块,共计49张图像块(粗信息粒),即$G_{\text{img}}=\{G_1,G_2,\cdots,G_{49}\}$,然后将49张图像块输入到预训练好的Vision Transformer(32×32)模型中,与此同时,每一张图像块也将作为多粒度随机游走模块中的游走结点$V_{\text{img}}=\{v_1,v_2,\cdots,v_{49}\}$,相邻的结点存在边信息$E=\{e_1,e_2,\cdots,e_n\}$,$n$为图中无向边个数,构建无向连通图$\boldsymbol{G}=(V_{\text{img}},E)$,再然后在Vision Transformer模块中,将每个图像块向量$\boldsymbol{E}_{\text{patch}}$与位置编码$\boldsymbol{E}_{\text{position}}$相加后得到最终的输入编码$\boldsymbol{I}$输入到Encoder编码器中,接着通过多个Encoder编码器后经过最后一个全连接层(Multilayer Perceptron Head, MLP Head)输出最终的分类结果,与此同时,多头注意力层能够可视化注意力热图并将注意力矩阵$\boldsymbol{h}_{\text{multi}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})$输入到多粒度随机游走模块,注意力指引结点$V_{\text{attention}}$将作为粗粒度下随机游走的初始结点,再接着根据本文提出的多粒度随机游走的三种游走停止条件最终得到游走路径$V_{\text{randomwalk}}=\{G_{\text{attention}},\cdots,G_{\text{final}}\}$,根据本文提出的粗信息粒重要度$I(E)$这一指标选取最佳游走路径,对粗信息粒约简后得到最终粗信息粒集合G_{reduct},最后将每个粗信息粒分割为16×16或8×8(更精确的任务中使用)像素的图像块,构建结点信息集合$V_{G_i}=\{V_1^i,V_2^i,\cdots,V_{16}^i\}$(采用$8\times8$像素的图像块时)和边信息$E_{G_i}$,根据本文提出的三个停止条件与细信息粒重要度$I(e)$进行细粒度下的随机游走来选取最佳游走路径(细信息粒集合),对细信息粒集合进行约简得到每个粗信息粒下细信息粒约简集合g^i_{reduct},最终得到输入图像的可解释性信息粒集合$G_{\text{img}}=\{g^i_{\text{reduct}},\cdots,g^j_{\text{reduct}}\}$并可视化结果.

基于上述基本思路,本文提出的MGRW-Transformer模型主要包含两个核心内容:(1)Transformer分类热图可视化;(2)多粒度随机游走. 模型总体设计如图1所示,具体的实现方法将在下面小节详细分析.

3.1 Transformer分类热图可视化

Vision Transformer是完全基于自注意力机制实现图像分类任务的模型,包含基本型、大型、巨型三种Vision Transformer模型见表1所示,例如ViT-L/16表示采用16×16图像块大小的大型Vision Transformer模型,故在保持输入图像大小不变的情况下采用较小的图像块大小模型的计算成本越高.

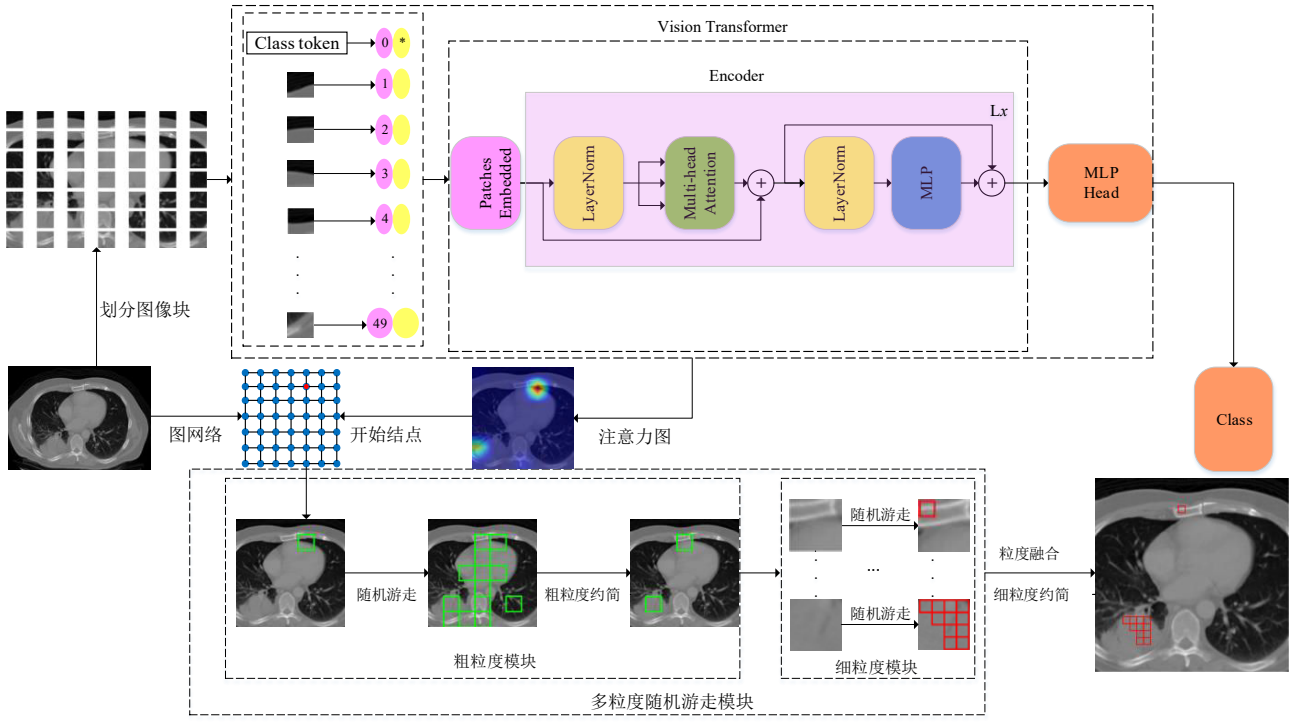


图1　MGRW-Transformer:多粒度随机游走分类解释性模型

表1　三种Vision Transformer模型框架

Model	Layers	Hidden size D	MLP size	Heads	Params/Mbit
ViT-Base	12	768	3 072	12	86
ViT-Large	24	1 024	4 096	16	307
ViT- Huge	32	1 280	5 120	16	632

本文采用Vision Transformer作为模型的分类器不仅可以提高模型的整体精度,与此同时Vision Transformer中自注意力模块可以很好的解释模型分类结果,虽然在医学图像分类中,显著性映射法、注意力方法往往难以奏效,但却对模型的分类结果有一定的指导作用,故本文将在这一小节以医学图像分类任务为例,详细介绍Vision Transformer模块分类与注意力热图可视化的过程,用于指导下一小节中的多粒度随机游走.

针对医学图像分类任务,由于医学图像的边缘信息对模型分类与可解释性结果影响较小,故将输入的医学图像进行中心裁剪并修改图像尺寸为224×224像素,为了简化矩阵运算以及处理序列化数据,首先需要将图像划分为多个Patch(图像块),然后在粗信息粒度下再将该图像切割成32×32像素的图像块,共计49块输入到Vision Transformer(32×32)模型中,最后再将每个图像块根据宽、高、通道数展开共计32×32×3=3 072个一维向量,最后再进入线性映射层进行线性投影变换.

为获取医学图像的全局信息用于分类,模型引入分类标志位CLS TokenClass Token用于表示图像的整体信息,位置编码$\boldsymbol{E}_{\text{position}}=\{\boldsymbol{P}_{\text{cls}},\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_{3\,072}\}$来获取每个图像块在原始图像所在的位置信息,图像块向量$\boldsymbol{E}_{\text{patch}}=\{\boldsymbol{p}_{\text{cls}},\boldsymbol{p}_1,\boldsymbol{p}_2,\cdots,\boldsymbol{p}_{3\,072}\}$与位置编码相加后,将编码$\boldsymbol{I}=\{\boldsymbol{P}_{\text{cls}}+\boldsymbol{p}_{\text{cls}},\boldsymbol{P}_1+\boldsymbol{p}_1,\boldsymbol{P}_2+\boldsymbol{p}_2,\cdots,\boldsymbol{P}_{3072}+\boldsymbol{p}_{3\,072}\}$输入到Vision Transformer模块中的Encoder编码器中.

Vision Transformer模块延用了Transformer[31]模型中Encoder编码器,Encoder编码器由多头注意力机制、LayerNorm归一化、残差连接以及全连接层组成,其核心内容多头注意力机制可以将查询与键值对信息映射到输出;将输入编码$\boldsymbol{I}$输入到Encoder编码器中记为$\boldsymbol{I}=\boldsymbol{X}=\{\boldsymbol{x}_0,\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{3072}\}$,通过注意力初始化权重矩阵$\boldsymbol{W}_Q$,$\boldsymbol{W}_K$,$\boldsymbol{W}_V$,并动态调整权重矩阵使得权重能够反映样本的重要性,最终得到$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$矩阵如下:

$$\boldsymbol{Q}=\boldsymbol{W}_Q\boldsymbol{X} \tag{1}$$

$$\boldsymbol{K}=\boldsymbol{W}_K\boldsymbol{X} \tag{2}$$

$$\boldsymbol{V}=\boldsymbol{W}_V\boldsymbol{X} \tag{3}$$

其中,$\boldsymbol{Q}$表示查询矩阵,$\boldsymbol{K}$和$\boldsymbol{V}$分别表示键、值矩阵,由于都是由自身样本X线性变化得到,最终得到注意力矩阵:

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{4}$$

其中d_k表示键矩阵的维度. 自注意力中$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$矩阵多次进行线性变化,每次变换称为单头:

$$\boldsymbol{h}_i=\text{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^{\boldsymbol{Q}},\boldsymbol{K}\boldsymbol{W}_i^{\boldsymbol{K}},\boldsymbol{V}\boldsymbol{W}_i^{\boldsymbol{V}}) \tag{5}$$

其中，$\boldsymbol{W}_i^{Q}$、$\boldsymbol{W}_i^{K}$、$\boldsymbol{W}_i^{V}$是三个注意力初始化的权重矩阵. 将多个头进行拼接后再进行线性变化即可得到多头注意力：

$$\boldsymbol{h}_{\text{multi}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Concat}(\boldsymbol{h}_1,\cdots,\boldsymbol{h}_n)\boldsymbol{W}^{O} \tag{6}$$

其中，Concat为拼接函数，n为head的个数，$\boldsymbol{W}^{O}$为权重矩阵.

模型将输入编码$\boldsymbol{I}$输入到Transformer模块，再经过多个Encoder（编码器）模块中训练后，全连接层会将多个Transformer模块输出的结果作为输入，输出每个类别对应的概率，同时多头注意力模块输出注意力矩阵$\boldsymbol{h}_{\text{multi}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})$，由于每个图像块大小为$32\times32$像素，故对注意力矩阵上采样得到注意力掩码矩阵$\boldsymbol{M}$：

$$\boldsymbol{M}=\text{Upsample}(\boldsymbol{M}_0,B) \tag{7}$$

其中，$\boldsymbol{M}_0$为需要上采样的矩阵，B为矩阵需要扩充的维数，在粗粒度下多头注意力层输出大小为7×7的矩阵，每个图像块为32×32像素，故此时需要对矩阵的维数扩充32倍得到注意力掩码矩阵$\boldsymbol{M}$，将注意力掩码矩阵$\boldsymbol{M}$归一化后覆盖在原始图像上，修改RGB三个通道，最终得到注意力热力图CAM，与此同时将注意力掩码矩阵$\boldsymbol{M}$输入到多粒度随机游走模块.

3.2 多粒度随机游走

多粒度分析方法在深度学习领域的应用往往与数据特征相关，通过对数据的处理获取局部特征、主要特征以及全局特征等；本文在粗粒度下将输入图像切分为32×32像素的图像块作为粗信息粒集合$G_{\text{img}}=\{G_1,G_2,\cdots,G_{49}\}$，给定大小的粗粒度中选取部分细图像块：

$$g_i=\Phi(G_i,S^t),i\in\{1,2,\cdots,d\} \tag{8}$$

其中，S^t是细信息粒的大小，g_i为该粗信息粒提取到特征的细图像块（细信息粒），Φ为具体的选取方法如局部二值特征裁剪[32]，本文将采用图随机游走算法来选取.

图随机游走算法是一种最优化寻解方法，图随机游走算法不仅成功运用于统计学和经济学等领域，在医学图像分割等领域也有广泛的运用[33]. 多粒度分析方法将原始图像切分为多个图像块有利于简化模型降低计算成本，但切分图像块会损失图像块之间信息即破坏原始图像的完整性，故如图1中构建的图网络所示，本文将保留粗粒度下切分后图像块的位置信息，将每个图像块作为结点得到图网络结点信息$V_{\text{img}}=\{v_1,v_2,\cdots,v_{49}\}$，四面相邻的结点之间构建边信息$E=\{e_1,e_2,\cdots,e_n\}$，其中$n$为图中无向边个数，相邻图像块结点之间的相似关联度作为边信息的权重，对于任意两个相邻结点v_i,v_j之间若存在边e_{ij}，那么边e_{ij}的权重w_{ij}为：

$$w_{ij}=\exp(-\beta(g_i-g_j)^2) \tag{9}$$

其中，g_i,g_j表示对应像素的强度，β表示加权参数，最终得到无向连通图$\boldsymbol{G}=(V_{\text{img}},E)$.

图随机游走算法需要在一定规模下才能够得到全局最优解，本文针对图随机游走算法计算成本高以及规模小难以得到最优解等问题，在游走的初始位置，游走的停止条件以及游走的评价函数三个方面进行了改进.

3.2.1 注意力机制引导的随机游走

在上一小节中，Transformer分类热图可视化模块将注意力掩码矩阵$\boldsymbol{M}$输入到了多粒度随机游走模块中，由于注意力掩码矩阵$\boldsymbol{M}$对应着输入图像中每个像素点对于决策的重要程度，在多粒度随机游走模块中，将注意力掩码矩阵$\boldsymbol{M}$按照掩盖图像$G_{\text{img}}=\{G_1,G_2,\cdots,G_{49}\}$进行划分，即$\boldsymbol{M}=\{\boldsymbol{M}_1,\boldsymbol{M}_2,\cdots,\boldsymbol{M}_{49}\}$，矩阵$\boldsymbol{M}_i$中所有元素之和的大小反应了对应图像块$G_i$对分类决策的重要程度，那么元素之和最大的矩阵$\boldsymbol{M}_{\max}$掩盖区域即为注意力热图CAM中的热力区域，矩阵$\boldsymbol{M}_{\max}$对应的图像块$G_{\max}$所处结点即为注意力指引结点$V_{\text{attention}}$，本文将充分利用注意力信息用于选取随机游走的起点$V_{\text{attention}}$，将随机游走的起点设置为注意力掩码矩阵$\boldsymbol{M}$指引的图像块结点$V_{\text{attention}}$，可以将游走起点定位到包含重要信息的结点位置，有效减少无效游走次数，降低计算成本. 本文提出的MGRW-Transformer模型将根据选取信息粒的重要度来寻求最佳路径，具体基于注意力机制的图随机游走的实现如算法1所示.

算法1是具体的图随机游走策略，在初始位置选取方面，采用Transformer模块中多头注意力层输出的注意力热图作为指引，将热力区域所在的位置结点$V_{\text{attention}}$即算法1中的Attention_Node作为初始起点Start_Node，用注意力指引结点替代随机初始化位置结点可以有效融合分类模型的注意力信息，减少因随机初始化而导致的部分无效游走，使得随机游走算法能在较小的规模下得到全局最优解.

基于注意力机制的图随机游走算法与传统的图随机游走算法相比共有三处创新点，除了游走的初始位置选取外，基于注意力机制的图随机游走算法在游走的停止条件以及游走的评价函数也进行了一定的改进.

3.2.2 随机游走评价函数和停止条件

在游走的停止条件方面，除了延用了最大游走路径长度终止游走的方法之外，还额外增加了两种停止条件，一是每次游走后删除游走前Start_Node和游走后Neighbor_Node两个结点之间的边信息，避免出现在部分结点之间往复无效游走的现象，若当前结点不存在邻居结点即Neighbors为空时游走停止，删除游走后两个游走结点的边信息再进行图随机游走可以有效避免

算法1　基于注意力机制的图随机游走算法

输入:注意力指引结点Attention_Node,最大游走路径长度Max_len,输入图像Img.

输出:游走路径Node_trave,选取信息粒的重要度*G*_importance.

1:Start_Node=Attention_Node;//注意力指引结点作为随机游走初始结点

2:Node_trave=np.array(Start_Node);//将初始结点添加到游走路径

3:While len(Node_trave)< Max_len://开始游走

4:　Neighbors=self.network.neighbors(StartNode);//获取当前结点邻居结点

5:　If　Neighbor_Node==None://无邻居结点

　　Break;//终止游走

6:　Neighbor_Node=choose_neighbor(Neighbors);//选取邻居结点

7:　If　Neighbor_Node==AttentionNode://回到初始化结点

　　Break;//终止游走

8:Self.network.delete_edge(StartNode, Neighbor_Node);//删除当前结点与选择结点之间的边信息

9:StartNode=Neighbor_Node;//选择结点作为当前结点

10:Node_trave=np.append(Node_trave, StartNode);//将结点添加到游走路径

11:*G*_importance=Get_significant(Node_trave,Img);//获取游走路径遮挡下粒度重要度

12:Return Node_trave, *G*_importance.//返回游走路径和对应信息粒下重要度

大量无效游走,快速找到最佳的游走路径;二是注意力方法往往对于游走具有积极的指导作用,简而言之,本文采用图随机游走算法旨在捕获注意力指引结点的周边结点信息,所以当部分边信息被删除后,游走结点Neighbor_Node返回到注意力指引结点Attention_Node时,游走路径必然包含了注意力周边结点,故此时停止游走;在算法1中,若随机游走不满足上述三种停止条件,则将游走到达的结点Neighbor_Node添加到游走路径Node_trave中,若满足停止条件则计算当前的游走路径(信息粒集合)的重要度,最后返回游走路径Node_trave和信息粒重要度I(Node_trave).

在游走的评价函数方面,由于不同的粒度大小对于模型有不同的影响,若直接选取在较小的图像块之间游走,大量的图像节点信息与边信息使得网络规模较大,不可避免出现计算量过大而在一定规模下无法得到全局最优解的现象,另一方面,选用较大的粒度使得模型网络规模较小,难以得到较好的可解释性结果;本文采用多粒度分析方法可以有效解决这一问题,通过计算选取信息粒的重要度I(Node_trave)来作为图随机游走评价标准,在粗、细粒度下将使用不同的评价函数$I(E)$和$I(e)$分别获取对应信息粒的重要度,最终选取最佳粗、细粒度. 粗、细粒度重要度评价函数如算法2所示.

算法2　粗、细粒度下重要度评价函数(Get_significant)

输入:图随机游走选取游走粒度Node_trave,输入图像Img.

输出:游走粒度重要度Imp.

1:S_1=Get_Classes(Img);//获取图像被预测为每个类别的概率

2:Img.permute(1,2,0);//将图像(通道,宽,高)转换为(宽,高,通道)

3:Mask(Img,Node_trave);//对图像中游走粒度进行掩码

4:Img.permute(2,0,1);//将图像(宽,高,通道)转换为(通道,宽,高)

5:Normalize (Img);//图像归一化

6:S_2=Get_Classes(Img);//获取图像被预测为每个类别的概率

7:Imp=Get_importance(S_1,S_2);//获取信息粒重要度

8:Return Imp.// 返回信息粒重要度

算法2主要针对图像的变换处理获取游走粒度的重要度,对于输入图像,在粗粒度下,可以将其切分为多个图像块构成粒度集合$G_{\text{img}}=\{G_1,G_2,\cdots,G_{49}\}$,该图像可能的决策信息满足$D=\{d_1,d_2,\cdots,d_m\}$,其中$m$为决策类别个数,在粗粒度下图随机游走最终选取信息粒的集合$E=\text{Node_trave}=\{G_{\text{attention}},\cdots,G_{\text{final}}\}$,$G_{\text{attention}}$为游走初始位置,$G_{\text{final}}$为游走终止结点,满足$E\subseteq G$,那么粗粒度下信息粒的重要度定义如下:

$$I(E)=\frac{1}{|D|}\sum_{d_j\in D}|S(G,d_j)-S(G,d_j,E)| \tag{10}$$

其中$S(G,d_j,E)$表示在G粗信息粒集合(整个图像)下将图随机游走信息粒集合E掩码后预测为类别j的概率,E默认为空.

与粗粒度相比,细粒度在深度学习中更具优势因其包含详细的主要特征也包含了粗略的整体信息以及整体信息与局部信息之间的联系信息. 本文在粗粒度下随机游走后对每个图像块再进行切分,构成更小的细粒度,即$G_i=\{g_1,g_2,\cdots,g_k\}$,其中$k$为每个大图像块包含小图像块的数量,令$e$为细粒度下随机游走的粒度集合,满足$e\subseteq G_i$,那么细粒度下粒度重要度定义如下:

$$I(e)=\frac{1}{|D|}\sum_{d_j\in D}|S(G_i,d_j)-S(G_i,d_j,e)| \tag{11}$$

其中$S(G_i,d_j,e)$表示在G_i细粒度集合(整个编号为i的图像块)下将e信息粒集合掩码后预测为类别j的概率,e默认为空.

粗、细粒度重要度越大则表明选取的信息粒对于模型的分类决策越关键,所以粗、细粒度重要度对于随机游走最佳粒度的选取起到决定性作用具体见算法3所示,在一定规模下选取信息粒重要度I(Node_trave)最大值对应的游走路径Node_trave作为最佳游走路径,此外粗、细粒度重要度对于粗、细粒度下的粒度约简和粒度融合都起到决定性作用.

在信息粒约简方面,为了进一步降低模型网络的规模,本文提出的多粒度随机游走的可解释性Transformer模型在分解粒度的基础上增加了粗、细信息粒的

算法3 多粒度随机游走算法

输入:注意力矩阵 Attention,游走规模 Num_P,最大游走路径长度 Max_len,输入图像 Img.

输出:粒度集合 G.

1:Igraph.Create_Network();//创建图网络

2:Create_Nodes(Attention.size());//根据图像块数量(Attention矩阵规模)构建结点信息

3:Create_Edges(len(nodes));//相邻的 Patch 即结点存在边信息,构建边信息

4:StartNode=Max_index(Attention);//根据 Attention 矩阵索引图随机游走初始位置

5:For person in Num_P://随机游走规模

Nodes,imp=Randomwalk(StartNode, Max_len, Img);

//获取随机游走路径结点和对应游走路径

6:G_Nodes=Max_imp(Nodes,imp);//获取最大遮挡粒度重要度对应的路径结点

7:G=Reduct(G_Nodes).//对遮挡路径根据重要性进行约简得到最终粒度集合

约简方法 Reduct;对于不同的数据集,多粒度随机游走的可解释性 Transformer 模型在信息粒约简中允许不同的误差.

在自然图像等简单解释性数据集下,首先由于该类数据集中图像的主体在图像中占比较大且往往分布在图像中心位置,采用较大的误差降低模型网络的规模,模型可解释性结果仍分布在图像的中心位置即模型仍具有较好的可解释性,此外由于该类数据集训练样本量大,适当增加训练规模往往也能够获得增大模型网络的规模同样的可解释性结果,最后从实验分析部分展现的实验结果中可以得出采用较大误差粗粒度网络相比于细粒度网络虽然在细节上表现稍差,但对比传统的可解释性算法仍具有较好的可解释性,因此在自然图像等简单解释性数据集下可采用较大的误差降低模型网络的规模的同时对模型可解释性的效果影响较小.

但在医学图像等精度要求较高数据集下,首先由于医学图像数据集中病灶特征在图像中占比较小且分布在图像位置不定,采用较大的误差将使得模型无法得到较好的可解释性结果,此外由于该类数据集训练样本量小,适当增加训练规模难以提升模型可解释性性能,最后从实验分析部分展现的实验结果中可以得出采用较大误差粗粒度网络相比于传统的可解释性算法具备一定的可解释性,但相较于细粒度网络存在大量噪声像素,难于运用于医学图像可解释性任务中,因此在医学图像等需要极为精确的可解释性任务中则建议采用较小误差增大网络规模来保证可解释性的结果;那么粗、细信息粒下最简信息粒集合 Red 为:

$$I(\mathrm{Red})-I(\mathrm{Red}-\{i\})>k, i\in \mathrm{Red} \tag{12}$$

其中 i 为 Red 集合中任意一个信息粒,满足约简其中任意一个信息粒都会降低信息粒集合的重要度使得误差大于阈值 k,那么集合 Red 即为最简信息粒集合.

在粒度融合方面,主要针对细信息粒约简后进行粒度融合这一过程,通过启发式算法得到粗粒度下的约简集合 $G_{\mathrm{reduct}}=\{G_i,\cdots,G_j\}$,对于每个粗信息粒通过随机游走算法能够都得细粒度下的细信息粒集合 g^i_{reduct},那么粒度融合后最终得到输入图像可解释性信息粒集合 $G_{\mathrm{img}}=\{g^i_{\mathrm{reduct}},\cdots,g^j_{\mathrm{reduct}},g^k_{\mathrm{reduct}}\}$,由于粗信息粒下得到的部分细粒度约简集合满足粗粒度约简阈值却不满足细粒度约简阈值,这些细粒度约简集合标注的特征在局部图像块中具有一定重要性,但与其他细粒度融合后对模型的分类精度影响较小,故要去除细粒度融合后存在的部分冗余特征,以每个细信息粒约简集合 g^k_{reduct} 为单位,对 g^k_{reduct} 中标注的特征进行扰动遮挡,若不满足细粒度约简阈值则从可解释性图中去除该冗余特征,最后得到最终的可解释性图 $G_{\mathrm{img}}=\{g^i_{\mathrm{reduct}},\cdots,g^j_{\mathrm{reduct}}\}$.

4 实验分析

本文通过 ImageNet 自然图像数据集和肺癌医学图像数据集对本文提出的 MGRW-Transformer 模型进行验证. 本文采用的实验平台为 PC(Intel(R) Core(TM) i9-12900K CPU@3.19 GHz, RAM 32 GB, 显卡为 NVIDIA GeForce RTX3090,内存容量 64 GB),Windows10 专业版操作系统,开发工具为 JetBrains PyCharm,使用 Python 语言实现实验中相关算法.

4.1 实验数据集

本文使用经典的 ImageNet[34](ILSVRC)2012 自然图像数据集,由 1 000 个类别的 5 万张图像组成,以及 ImageNet-Segmentation[35] 数据集,包含来自 445 个类别的 4 276 张图像.

本文在医学图像领域采用的是 Kaggle 中肺部 CT 扫描图像数据集[36],包含四个类别分别是腺癌、大细胞癌、鳞状细胞癌以及正常,其中训练集占比 70%,测试集占比 20%,验证集占比 10%,共收录 1 000 张图像信息.

4.2 实验结果评估指标

本文提出一种可解释性模型,通过标注特征信息来解释模型的分类结果,故本文采用图像分割中的部分标准来评价模型精度. 假设在图像语义分割中共有 k+1 个类别,p_{ij} 表示标签为类别 i 却被预测为类别的像素数量,那么和分别对应假正和假负的像素数量,和对应真正和真负的像素数量.

像素准确率(Pixel Accuracy,PA)表示预测正确的像素数量占像素总量的比例,计算公式定义为:

$$\mathrm{PA}=\frac{\sum_{i=0}^{k}p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k}p_{ij}} \tag{13}$$

平均交并比(Mean Intersection over Union,MIoU)计算真实值和预测值两个集合的交并比,计算公式定义为:

$$\mathrm{MIOU}=\frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ij}-p_{ii}} \tag{14}$$

平均精度均值(mean Average Precision,mAP)衡量图像多个类别的识别精度,求解平均精度均值首先需要绘制PR曲线(Precision-Recall Curve):

$$\mathrm{Precision}=\frac{p_{ii}}{p_{ii}+p_{ij}} \tag{15}$$

$$\mathrm{Recall}=\frac{p_{ii}}{p_{ii}+p_{jj}} \tag{16}$$

将召回率(Recall)作为横轴,精准度(Precision)作为纵轴构建直角坐标系得到PR曲线,以0.1为步长选取横轴上11个点对应的精准度取平均得到平均精准度:

$$\mathrm{AP}=\frac{1}{11}\sum_{r\in\{0,0.1,0.2,\cdots,1.0\}}P(r) \tag{17}$$

其中r表示横轴即召回率的数值,表示召回率为r时对应的精准度. AP对应一个类别下平均精准度,那么mAP表示在所有类别下的平均精度均值,公式定义如下:

$$\mathrm{mAP}=\frac{\sum_{q=1}^{Q}\mathrm{AP}(q)}{Q} \tag{18}$$

其中Q为预测图像的类别总数,AP(q)表示在第q个类别下的平均精准度AP.

4.3　实验结果与分析

分类模型的性能决定我们将如何进行决策,而模型的可解释性从侧面反应了分类模型决策的原因,两者相辅相成,这一小节我们将从模型分类性能以及模型的可解释性两个方面对自然图像数据集和医学图像数据集分别进行分析.

由于反向传播法和类激活映射法主要依赖于卷积神经网络模型中层与层之间的梯度信息,故可以兼容所有的卷积神经网络分类模型,而注意力法以及本文提出的MGRW-Transformer模型主要依赖于Transformer模块中多头自注意力机制,故可以兼容所有Vision Transformer模型以及包含Transformer模块的其他模型等;本文侧重于在通过优化可解释性算法来得到比传统可解释性算法更好的可解释性结果,故本文选用了经典深度学习分类模型中两种VGG[37]网络结构、五种Resnet[38]网络结构与常用的五种Vision Transformer网络结构进行比较,选择合适的网络作为本文提出的MGRW-Transformer模型以及传统的可解释性算法的基分类器,在ImageNet数据集和肺部CT扫描图像数据集上均使用对应网络结构在ImageNet数据集上的预训练参数,在保证学习率、优化函数、训练规模等参数一致的情况下,从模型的Top-1准确率和模型的复杂度等评估指标来选取最佳分类模型.

本文采用不同预训练后的网络结构在ImageNet数据集上进行测试,表2为本次测试中各个网络结构的Top-1准确率、模型的总参数大小、模型的计算速度,用每秒十亿次的浮点运算量(Giga Floating-point Operations Per Second,G-Flops)表示,从表中可以看出ViT-H/14模型即采用图像块大小的巨型Vision Transformer模型在ImageNet上得到最高88.08%的Top-1准确率,与此同时该模型参数量最大为660.39 Mbit,计算速度也相应为最高的161.96 G-Flops,而Resnet-34包含的参数总量最小仅为21.8 Mbit,相应的计算速度也最低为3.68 G-Flops. 在众多网络结构中,Top-1准确率越高越好,模型的复杂度与模型总参数量和计算速度有关,故总参数大小越低越好,相应的计算速度也应降低. 本文将综合考虑准确率与复杂度两个方面选取最优网络结构作为本文中的分类模型.

表2　多种分类模型的复杂度以及在ImageNet数据集上的分类精度

Model	Top-1 Accuracy/%	Params/M	Operations/G-Flops
Vgg16	74.4	138.36	15.5
Vgg19	74.5	143.67	19.6
Resnet-34	74.97	**21.8**	**3.68**
Resnet-50	77.54	25.56	4.12
Resnet-101	80.67	44.55	7.85
Resnet-152	81.88	60.19	11.58
ViT-B/16	84.15	86.57	16.86
ViT-B/32	80.73	88.22	4.37
ViT-L/16	86.3	304.33	59.67
ViT-L/32	84.37	306.54	15.26
ViT-H/14	**88.08**	660.39	161.96

图2为不同的网络结构在不同数据集下的测试结果,综合了网络结构Top-1准确率、计算速度以及参数总量进行分析绘图,图2(a)表示不同网络结构在ImageNet数据集上测试结果,纵坐标表示不同网络结构在ImageNet数据集上Top-1准确率,横坐标对应不同网络结构的计算速度,每一个圆对应相应网络结构,圆的大小代表该网络结构的参数量的大小,Flops不仅可以用于衡量算法运行的快慢也能衡量模型的复杂度,Flops

越小、圆的面积越小则模型的复杂度越低，即在表2中Resnet-34网络结构复杂度最低，ViT-H/14网络结构复杂度最高，从图2中可以看出越靠近左上角的网络结构整体较好，综合ImageNet数据集和肺癌CT图像数据集可以看出ViT-L/32和ViT-B/16以及Resnet-34在自然图像和医学图像数据集上表现较好，Vision Transformer模型自身包含了注意力模块能够生成注意力热图，对多粒度随机游走模块具有指导作用，若选用传统卷积神经网络Resnet-34模型则无法得到注意力热图即只能通过传统可解释性方法例如显著性映射法得到显著性映射图，这无疑额外增加了计算成本，所以Resnet-34相较于ViT-L/32和ViT-B/16并没有减少计算成本反而降低了模型分类精度，故本文在粗粒度下采用ViT-L/32为基模型，在细粒度下采用ViT-B/16为基模型.

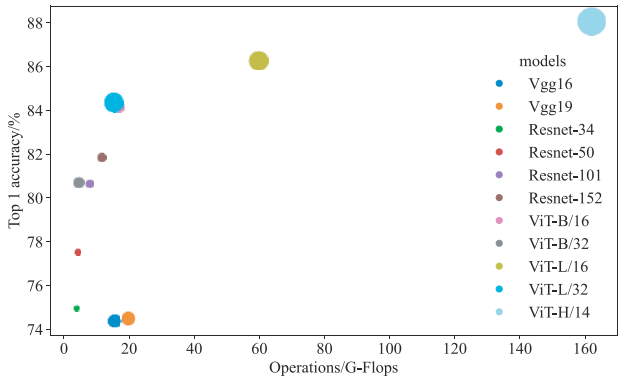
本文将MGRW-Transformer模型与注意力法、显著映射法、梯度反向传播法等方法进行比较，分别为Raw-attention算法、Score-CAM算法和Relevance-CAM算法.

本文提出的MGRW-Transformer模型在粗粒度下采用ViT-L/32为基分类器，在细粒度下采用ViT-B/16为基分类器，Raw-attention算法采用ViT-B/16为基分类器，Score-CAM算法、Relevance-CAM算法均采用ResNet-50为基分类器，以上可解释性算法采用的基分类器均使用ImageNet数据集预训练参数；本文提出的MGRW-Transformer模型在自然图像数据集中游走规模为500次，最大游走路径长度为36块，粗粒度下约简阈值为0.3，细粒度下约简阈值为0.2；而在精度要求更高的医学图像数据集下游走规模为5 000次，最大游走路径长度为20块，粗粒度下约简阈值为0.2，细粒度下约简阈值为0.1.

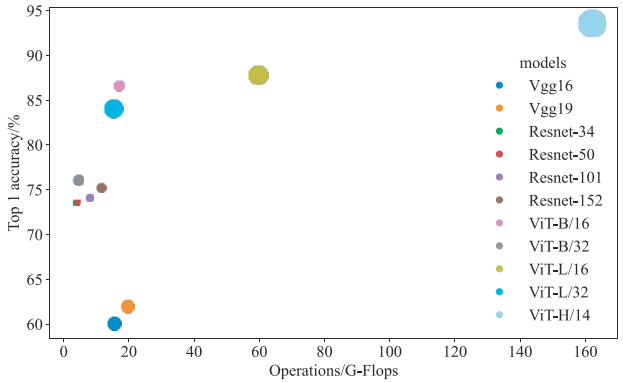
为了检验多粒度随机游走解释性模型在自然图像和医学图像领域分类性能及其可解释性，本文基于ImageNet(ILSVRC)2012数据集和肺部CT扫描图像数据集分别进行分类及其可解释性实验.

图3为自然图像单目标分类任务解释性效果图，在该图中，从左往右依此为输入图像、Raw-attention算法解释性效果图，Score-CAM算法解释性效果图和Relevance-CAM算法解释性效果图；本文提出的粗粒度下解释性效果图以及细粒度下解释性效果图；由于不同的可解释性方法展现形式不同，为了更好展示各种可解释性方法的性能好坏，本文将多种可解释性算法输出的最终结果如注意力热图CAM统一用红色边框进行展示，对于本文提出的多粒度分析方法，输出结果保留了每个图像块的完整性，用于展示粗、细信息粒的选取，与此同时图像块的外围边框也可用于与其他可解释性算法进行比较.

从图3中四张自然图像的可解释性结果可以看出



(a) ImageNet数据集



(b) 肺癌CT图像数据集

图2 多种分类模型在不同数据集上测试结果

Raw-attention算法能够关注到对于分类模型重要的像素特征，然而易受背景、噪声等干扰；Score-CAM算法几乎能够捕捉到主体的大部分特征，然而由于不依靠模型的梯度信息，在大部分图像中均出现了噪声像素且在噪声像素与主体像素接近时表现较差，如在海豚图像中由于海豚主体特征像素与海水特征像素较为接近，Score-CAM算法几乎没有捕捉到主体；由于Relevance-CAM算法主要贡献在于选用梯度更为平稳的LRP算法中相关性分数作为类激活映射权重成分，Relevance-CAM算法相较于Raw-attention算法和Score-CAM算法能够标记出主体绝大部分重要的特征，但是与Raw-attention算法和Score-CAM算法一样在主体特征像素与噪声特征像素较为接近时表现较差，如在海豚图像中标注的结果与正类完全相反；本文提出的粗粒度模型可以较好的标出自然图像中的生物特征用于解释分类结果，但由于粗粒度下选用的图像块较大，故最终选取到的特征包含了一小部分的背景噪声，于是本文继续在每个粗粒度中进行细化，进而粗粒度模型可以在不降低解释性模型精度的同时将粗粒度下标记好的粗图像块进一步的细化，得到最终切分出自然图像的总体特征.

为了进一步比较本文提出的多粒度随机游走的可解释性Transformer模型与多种传统可解释性模型的优

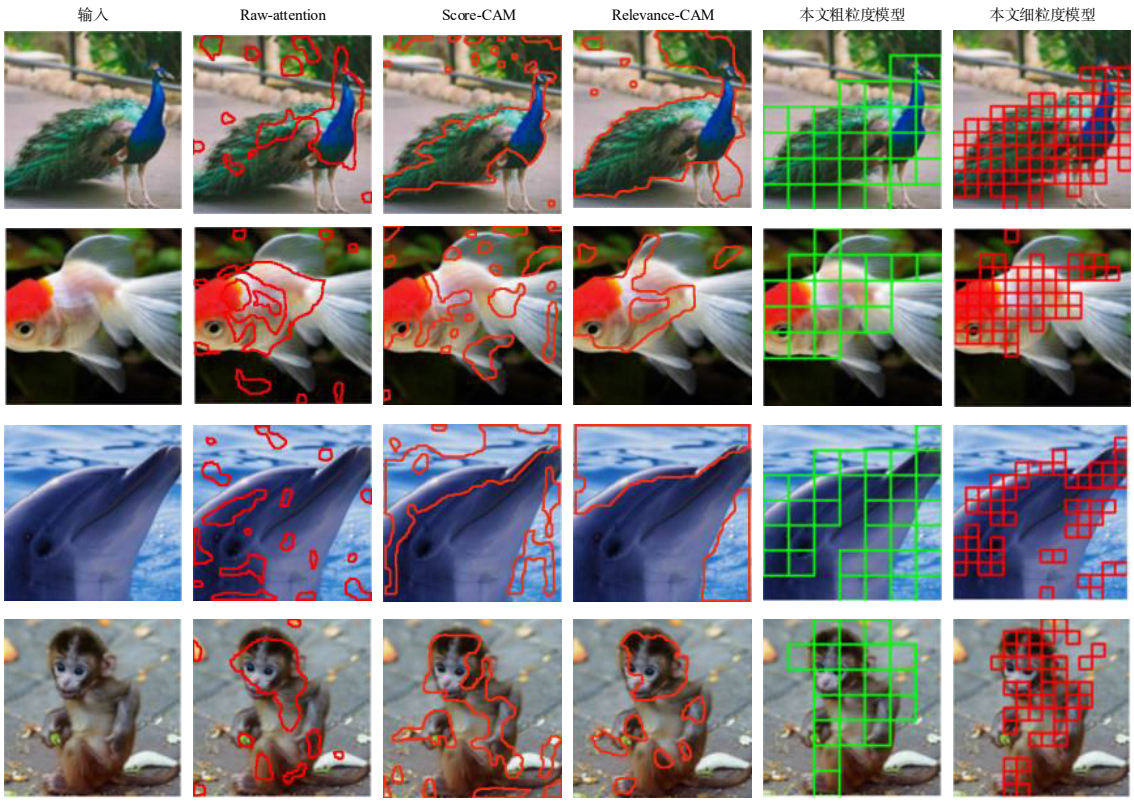


图3　自然图像单目标分类任务

劣,本文引入了像素准确率、平均交并比、平均精度均值三个指标进行衡量,多种模型最终在ImageNet-segmentation数据集上的表现见表3所示.

像素准确率(Pixel Accuracy)是图像分割中最常用的指标.像素准确率来衡量分类正确的像素占像素总数的比例,像素准确率值越高,代表图像分割模型的效果就越好;从表3可以看出本文提出的多粒度随机游走解释性Transformer模型在像素准确率下比另外三个解释性模型表现更好,细粒度模型达到了最高75.27%的像素准确率,细粒度解释性模型的像素准确率更是Score-CAM算法提高了8.09%,粗粒度模型达到了71.72%的次优值,从图3中四张自然图像也能看出本文提出的多粒度随机游走的可解释性Transformer模型相较于另三种算法能够捕捉到更多分类正确的像素.

表3　ImageNet-segmentation数据集分割能力　　单位/%

	Raw-attention	Score-CAM	Relevance-CAM	本文粗粒度模型	本文细粒度模型
Pixel Accuracy	67.84	67.18	73.15	71.72	**75.27**
mAP	**80.24**	64.38	68.33	73.05	73.22
MIoU	46.37	45.46	53.40	55.72	**59.28**

平均交并比(Mean Intersection over Union,MIoU)是图像分割中最为简单有效的指标之一,是每个类别下像素预测值与标签的重叠区域占像素预测值和标签的联合区域的比例的均值,体现了预测值和标签的契合程度,平均交并比越高说明图像分割模型的效果就越好;从表3中可以看出粗、细粒度可解释性模型在平均交并比这一指标下分别取得了次优值55.72%和最优值59.28%,尤其是细粒度可解释性模型比Score-CAM算法提高了13.82%,可见本文提出的多粒度随机游走的可解释性Transformer模型相较于其他可解释性算法具有更好的分割能力.

平均精度均值(mean Average Precision,mAP)是图像分割、目标检测尤其是多类别目标检测中最常用的指标之一,平均精度(Average Precision,AP)代表由准确率以及召回率绘制出PR曲线下的面积,平均精度均值表示多个类别下平均精度的均值;平均精度均值越高,表示模型在图像分割以及多类别目标检测表现越好,从表3中可以看出Raw-attention算法在这一指标下表现最好达到80.24%,本文提出的多粒度随机游走的可解释性Transformer模型表现仍然较好但在ImageNet-segmentation数

据集未能达到最优,主要存在以下两点原因:

一是较好的图像分割、目标检测模型需要满足正确的类漏检少、错误的类误检少、目标类别分类准、目标检测框与标签贴合度高这四个要求,通过上述像素准确率以及平均交并比两个指标可以验证本文提出的多粒度随机游走的可解释性Transformer模型相较于Raw-attention算法在像素准确率提高了7.43%,在平均交并比提高了12.91%即在目标类别分类准和目标检测框与标签贴合度高这两个要求下表现更好;从图3可以看出,Raw-attention算法在错误的类误检少这一要求下相较于其他算法表现最差,该算法极易受背景、噪音等因素影响,所以综合以上可以得出Raw-attention算法相较于本文提出的算法仅仅只在正确的类漏检少这一要求下表现较好,即Raw-attention算法能够捕捉到更多正确类别的特征信息,比如图3中能够捕捉到猴子完整的脸部特征却是以忍受更多的噪声特征如猴子手中的果实、背景中的落叶等为代价. 这对于可解释性工作有一定的影响.

二是平均精度均值是针对特定数据集下计算的一种相对度量,代表多个类别的平均精度,所以往往在多类别目标检测模型中表现更好,且ImageNet-segmentation数据集中含有较多的多类别分割图像;从图3可以看出,基于注意力方法的Raw-attention算法对于多个类别,比如图3中正类猴子、负类猴子手中的果实、负类背景中树叶都较为敏感,适用于多类别图像分割;然而本文提出的粗、细粒度可解释性方法与多类别分割模型不同,主要用于标记图像中正类信息,即分类模型标签的特征信息.

综合图3和表3评价指标分析可以得出MGRW-Transformer模型在自然图像分类的可解释性方面表现较好.

本文采用不同的ImageNet预训练网络结构在肺癌CT图像数据集上进行测试见表4所示,ViT-H/14表现依旧最佳,Top-1准确率达到了93.66%,与此同时精准度以及召回率也分别达到了96.90%和69.62%,从整体而言Vision Transformer模型在肺癌CT数据集上要优于VGG模型和Resnet模型,由于模型更为侧重Top-1准确率,故排除VGG模型和Resnet模型后最终在粗细粒度模型选取时,参考图2(b)中分别选用ViT-L/32和ViT-B/16作为基模型.

在可解释性方面上述可解释性算法几乎都能够在自然图像数据集上展现一定的可解释性功能,是因为上述模型都通过了大规模自然图像数据的训练使得分类及其可解释性都具备了较好的泛化能力,然而由于医学图像可供的训练样本量较少,可解释性模型难以在医学图像领域拥有较好的表现,故本文提出的MGRW-Transformer模型主要面向癌症检测等医学图像分类任务,旨在提取医学图像中的主要特征用于解释医学图像的分类结果,与医学图像分割数据集不同,本文选用了医学图像分类数据集即包含一定数量的正类(类别为正常),本文将采用不同的方法提取负类中重要特征来解释分类结果如图4所示.

图4中第一列是四张被分类为肺癌且已被影像和临床医生标定疑似肺癌病灶区域的医学图像,依旧选用了Raw-attention算法、Score-CAM算法和Relevance-CAM算法三种可解释性模型与本文提出的MGRW-Transformer模型在医学图像分类可解释性进行比较,前三种对比算法依旧采用红色边框的统一形式展现,本文提出的MGRW-Transformer模型采用最终选取的粗、细粒度的形式展现.

从图4中可以清晰的看出每张图像存在着大小不一的肿块即在肺部CT图像中长度超过3CM的病灶区域,由于Raw-attention算法、Score-CAM算法和Relevance-CAM算法缺乏大规模的预训练,因而在医学图像数据集上表现较差,故本文通过多张肺癌图像分析可解释性模型的好坏. 从Raw-attention算法输出结果可以看出该可解释性模型主要标记心脏以及图像边缘部分,完全不能捕捉到病灶的特征,Score-CAM算法和Relevance-CAM算法由于均采用ResNet50模型进行训练,且训练样本有限,故两种算法更关注外部轮廓以及肺部枝叶部分,几乎没有捕捉到的任何病灶像素特征,难以用于解释医学图像分类结果,而本文提出的MGRW-Transformer模型相比之下拥有较好的性能,从粗粒度可解释性模型中可以看出整体模型能够找出用于解释分类结果的病灶特征,可解释性结果包含了大量的噪声像素,无法突出细微的病灶特征,这不利于医生或者患者去识别哪一部分才是病灶区域,因此本文根据粗粒度模型继续降低约简阈值,融合约简细粒度下的特征像素,将图像块中的部分可解释性结果拼接成原始图像的可解释性图,解决了较大医学图像病灶区域可能会被划分到多个不同的信息粒下,局部图像块中病灶信息缺乏完整性对可解释性结果的影响,由于细粒度下选用了较小的约简阈值来去除噪声像素的干扰,细粒度可解释性结果图中标注的细信息粒特征主要集中分布在病灶像素特征区域,能够突出细微的病灶像素特征,用于解释医学图像分类结果,极个别细粒度模块往往指引着噪声或者隐藏微小病灶. 所以不难看出常规的可解释性模型难以胜任训练样本较小的医学图像可解释性任务,而本文针对医学图像可解释性提出的MGRW-Transformer模型不仅仅在自然图像数据集上拥有较好的表现,在训练样本较小的医学图像分类及其可解释性任务中表现仍然较为突出.

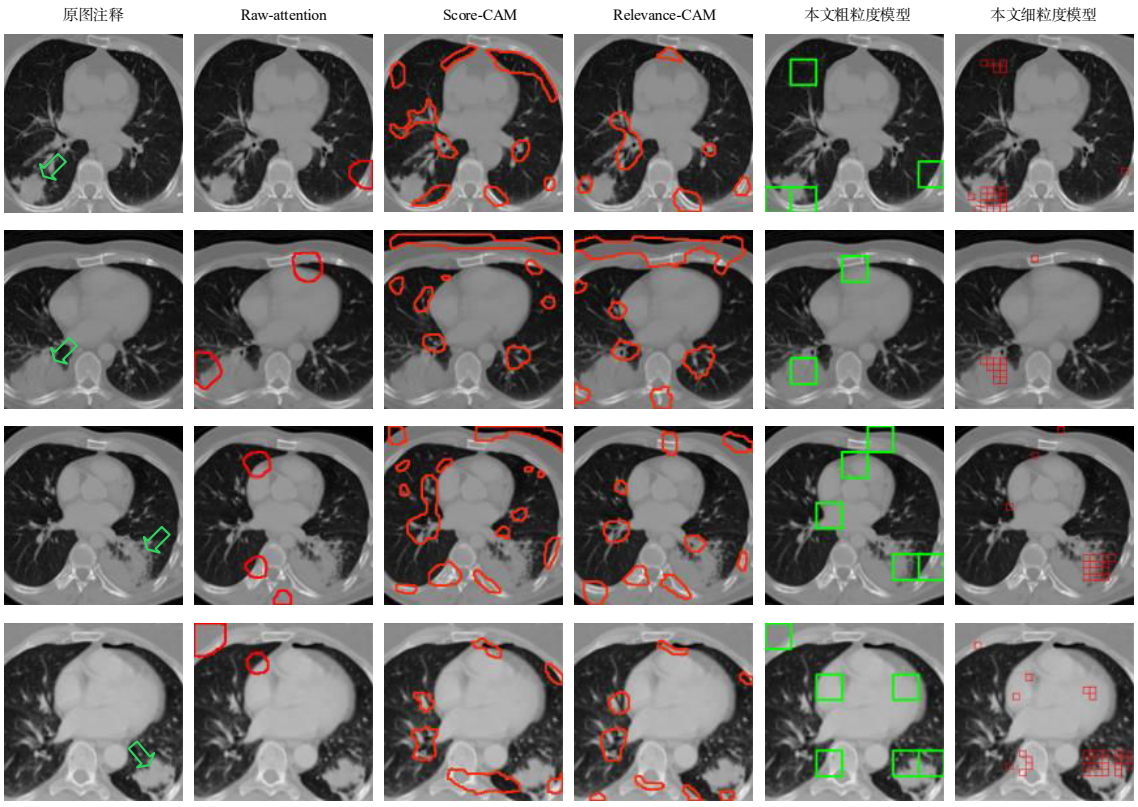


图4　肺癌图像病灶检测分类任务

表4　多种分类模型在肺癌CT数据集上测试结果

Model	Top-1 Accuracy/%	Precision/%	Recall/%
Vgg16	60.07	19.87	34.81
Vgg19	62.01	23.40	39.56
Resnet-34	73.59	34.98	41.85
Resnet-50	73.66	35.79	44.81
Resnet-101	74.17	38.73	56.67
Resnet-152	75.25	39.30	50.37
ViT-B/16	86.61	**100**	31.11
ViT-B/32	76.11	41.96	60.0
ViT-L/16	87.84	99.02	37.77
ViT-L/32	84.1	**100**	18.48
ViT-H/14	**93.66**	96.90	**69.62**

5　结论与展望

本文针对深度学习领域缺乏可解释性难以运用于医学图像分类等任务展开了深入研究,分析医学图像数据集因数据量小且病灶位置多变、大小不一,故无法采用常规的显著性映射法、注意力法等可解释性方法来解释分类结果,提出了基于注意力机制的随机游走算法,从游走起点、停止条件、评价函数三个方面进行图随机游走确定输入图像重要特征位置,并采用多粒度分析方法基于信息粒的重要度由粗到细选取并可视化对于分类结果重要的特征信息,建立了MGRW-Transformer模型,通过在ImageNet(ILSVRC)2012数据集、ImageNet-Segmentation数据集以及肺癌CT图像数据集下与Raw-attention算法、Score-CAM算法、Relevance-CAM算法三种可解释性算法进行比较验证得出本文提出的多粒度随机游走可解释性Transformer模型在自然图像和医学图像数据集下均表现较好,但本文提出的模型还存在着标注的特征不够全面等问题,我们将在未来的工作中考虑采用最新的深度学习分类方法来进一步提高模型的可解释性.

参考文献

[1] GOUR M, JAIN S, SUNIL KUMAR T. Residual learning based CNN for breast cancer histopathological image classification[J]. International Journal of Imaging Systems and Technology, 2020, 30(3): 621-635.

[2] PONNADA V T, SRINIVASU D S V N. Efficient CNN for lung cancer detection[J]. International Journal of Recent Technology and Engineering (IJRTE), 2019, 8(2): 3499-3503.

[3] 魏博文, 全红艳. 基于语义与形态特征融合的语义分割网络[J]. 电子学报, 2022, 50(11): 2688-2697.
WEI B W, QUAN H Y. Semantic segmentation network based on semantic and morphological feature fusion[J]. Ac-

ta Electronica Sinica, 2022, 50(11): 2688-2697. (in Chinese)
[4] RUSTAM Z, HARTINI S, PRATAMA R Y, et al. Analysis of architecture combining convolutional neural network (CNN) and kernel K-means clustering for lung cancer diagnosis[J]. International Journal on Advanced Science, Engineering and Information Technology, 2020, 10(3): 1200-1206.
[5] SHI Z H, HAO H, ZHAO M H, et al. A deep CNN based transfer learning method for false positive reduction[J]. Multimedia Tools and Applications, 2019, 78(1): 1017-1033.
[6] ZHAO L, XU X W, HOU R P, et al. Lung cancer subtype classification using histopathological images based on weakly supervised multi-instance learning[J]. Physics in Medicine and Biology, 2021, 66(23): 235013.
[7] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[EB/OL]. (2020-10-22)[2022-08-01]. https://arxiv.org/abs/2010.11929.
[8] CHEN J N, LU Y Y, YU Q H, et al. Transunet: Transformers make strong encoders for medical image segmentation [EB/OL]. (202102-08) [2022-08-01]. https://arxiv. org/abs/2102.04306.
[9] ZHANG Y D, LIU H Y, HU Q. TransFuse: fusing transformers and CNNs for medical image segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2021: 14-24.
[10] GAO X H, QIAN Y, GAO A. COVID-VIT: Classification of COVID-19 from CT chest images based on vision transformer models[EB/OL]. (2021-07-04) [2022-08-01]. https://arxiv.org/abs/2107.01682.
[11] VALANARASU J M J, OZA P, HACIHALILOGLU I, et al. Medical transformer: Gated axial-attention for medical image segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2021: 36-46.
[12] MATSOUKAS C, HASLUM J F, SÖDERBERG M, et al. Is it time to replace CNNs with transformers for medical images? [EB/OL]. (2021-08-24) [2022-08-01]. https://arxiv.org/abs/2108.09038.
[13] PAPANASTASOPOULOS Z, SAMALA R K, CHAN H P, et al. Explainable AI for medical imaging: Deep-learning CNN ensemble for classification of estrogen receptor status from breast MRI[C]//SPIE Medical Imaging. Proc SPIE 11314, Medical Imaging 2020: Computer-Aided Diagnosis. Houston: SPIE, 2020: 228-235.
[14] ZHANG Z Z, XIE Y P, XING F Y, et al. MDNet: A semantically and visually interpretable medical image diagnosis network[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 6428-6436.
[15] LIN C H, LICHTARGE O. Using interpretable deep learning to model cancer dependencies[J]. Bioinformatics, 2021, 37(17): 2675-2681.
[16] FENG Y J, MIN X, CHEN N, et al. Patient outcome prediction via convolutional neural networks based on multi-granularity medical concept embedding[C]//2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Piscataway: IEEE, 2017: 770-777.
[17] WANG K, ZHANG X B, ZHANG X H, et al. Multi-granularity scale-aware networks for hard pixels segmentation of pulmonary nodules[J]. Biomedical Signal Processing and Control, 2021, 69: 102890.
[18] BINDER A, MONTAVON G, LAPUSCHKIN S, et al. Layer-wise relevance propagation for neural networks with local renormalization layers[C]//Proceedings of the Artificial Neural Networks and Machine Learning. Barcelona: Springer, 2016: 63-71.
[19] 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]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 5797-5808.
[20] CHEFER H, GUR S, WOLF L. Transformer interpretability beyond attention visualization[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 782-791.
[21] LEE J R, KIM S, PARK I, et al. Relevance-CAM: Your model already knows where to look[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 14939-14948.
[22] ZHOU B L, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 2921-2929.
[23] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 618-626.
[24] CHATTOPADHAY 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). Piscataway: IEEE, 2018: 829-838.

[25] WANG H F, WANG Z F, DU M N, et al. Score-CAM: Score-weighted visual explanations for convolutional neural networks[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE, 2020: 24-25.

[26] ZEILER M D, FERGUS R. Visualizing and Understanding Convolutional Networks[M]//Computer Vision-ECCV 2014. Cham: Springer International Publishing, 2014: 818-833.

[27] Petsiuk V, Das A, Saenko K. RISE: Randomized input sampling for explanation of black-box models[EB/OL]. (2018-06-19)[2022-08-01]. http://arxiv.org/abs/1806.07421.

[28] 张宇倩, 李国辉, 雷军, 等. FF-CAM: 基于通道注意机制前后端融合的人群计数[J]. 计算机学报, 2021, 44(2): 304-317.
ZHANG Y Q, LI G H, LEI J, et al. FF-CAM: Crowd counting based on frontend-backend fusion through channel-attention mechanism[J]. Chinese Journal of Computers, 2021, 44(2): 304-317. (in Chinese)

[29] BAMBA U, PANDEY D, LAKSHMINARAYANAN V. Classification of brain lesions from MRI images using a novel neural network[C]//Proceedings Volume 11232, Multimodal Biomedical Imaging XV. San Francisco: SPIE, 2020: 23-31.

[30] Serrano S, Smith N A. Is attention interpretable?[EB/OL]. (2019-06-09)[2022-08-01]. https://arxiv.org/abs/1906.03731.

[31] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000-6010.

[32] REN S Q, CAO X D, WEI Y C, et al. Face alignment at 3000 FPS via regressing local binary features[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2014: 1685-1692.

[33] GRADY L. Random walks for image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(11): 1768-1783.

[34] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[35] GUILLAUMIN M, KÜTTEL D, FERRARI V. ImageNet auto-annotation with segmentation propagation[J]. International Journal of Computer Vision, 2014, 110(3): 328-348.

[36] Mohamed H. Chest CT-Scan images Dataset[EB/OL]. (2019-10-10)[2022-08-01]. https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images.

[37] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04)[2022-08-01]. https://arxiv.org/abs/1409.1556.

[38] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 770-778.

作者简介

耿　宇　男,1998年出生于江苏扬州. 南通大学信息科学技术学院硕士研究生. 主要研究领域为粒计算、深度学习.

丁卫平　男,1979年出生于江苏金坛,博士,教授,博士生导师. 2013年于南京航空航天大学获得工学博士学位. 主要研究方向为人工智能及计算智能、多模态机器学习及优化、深度神经网络、粒计算及其在不确定大数据中应用. E-maill: dwp9988@163.com

黄嘉爽　男,1988年出生于江苏南通,博士,讲师. 2015年取得南京工业大学硕士学位,2020年取得南京航空航天大学博士学位. 主要研究领域为脑网络分析,深度学习.

鞠恒荣　男,1989年出生于江苏泰州,博士,副教授. 2015年取得江苏科技大学硕士学位,2019年取得南京大学博士学位. 主要研究领域为粒计算、粗糙集、机器学习、知识发现.

孙　颖　女,1997年出生于江苏南通. 南通大学信息科学技术学院硕士研究生. 主要研究领域为粒计算、粗糙集、深度学习.

王海鹏　男,1998年出生于江苏泰州. 南通大学信息科学技术学院硕士研究生. 主要研究领域为模糊系统、医学图像分析、深度学习.