DOI:10.19651/j.cnki.emt.2416153

基于改进 MMAL 的细粒度图像分类研究*

李冰锋[1,2] 冀得魁[1,2] 杨 艺[1,2]
(1.河南理工大学电气工程与自动化学院 焦作 454000; 2.河南省煤矿装备智能检测与控制重点实验室 焦作 454003)

摘 要: 针对细粒度图像分类中目标区域难以精准定位及其内部细粒度特征难以识别的问题,提出了一种基于改进 MMAL 的细粒度图像分类方法。首先,利用形变卷积的感知区域可变性原理,动态地感知样本图像中不同尺度和形状的目标区域特征,从而增强网络对目标区域位置的感知能力。随后,采用 GradCAM 梯度回流的方法生成网络注意力热图,以减小特征背景噪声的干扰,实现对图像目标区域的精准定位。最后,提出位置感知空间注意力模块,通过融合坐标位置和双尺度空间信息,显著提升了网络对目标区域细粒度特征的提取能力。实验结果表明,与基线算法相比,该方法在 CUB-200-2011、Stanford Car 和 FGVC-Aircraft 三个公共数据集上分类精度分别提升了 1.4%、1.5%、1.9%,该结果验证了所提方法的有效性。

关键词: 细粒度图像分类;多尺度形变分组;位置感知空间注意力;GradCAM 热图定位;多分支
中图分类号: TP391.4;TN791 **文献标识码**: A **国家标准学科分类代码**: 520.60

Analysis of fine-grained image classification through improved MMAL

Li Bingfeng[1,2] Ji Dekui[1,2] Yang Yi[1,2]
(1. School of Electrical Engineering and Automation, Henan University of Technology, Jiaozuo 454000, China;
2. Henan Province Key Laboratory for Intelligent Detection and Control of Coal Mine Equipment, Jiaozuo 454003, China)

Abstract: To address the challenges of accurately locating target regions and identifying fine-grained features in fine-grained image classification, we propose a fine-grained image classification method based on an improved multi-scale deformable convolution (MMAL). Firstly, by leveraging the variable receptive field principle of deformable convolution, our method dynamically adapts to different scales and shapes of target regions in sample images, enhancing the network's ability to perceive the position of these regions. Subsequently, we utilize the Grad-CAM gradient backpropagation technique to generate network attention heatmaps, which reduces the interference from background noise and achieves precise localization of the image target regions. Finally, we introduce a position-aware spatial attention module that integrates coordinate positions and dual-scale spatial information, significantly improving the network's capability to extract fine-grained features of the target regions. Experimental results demonstrate that, compared to baseline methods, our approach achieves improvements of 1.4%, 1.5%, and 1.9% in classification accuracy on the CUB-200-2011, Stanford Car, and FGVC-Aircraft datasets, respectively, validating the effectiveness of the proposed method.

Keywords: fine-grained image classification; multi-scale deformable grouping; position-aware spatial attention; GradCAM heatmap localization; multi-branch architecture

0 引 言

细粒度图像辨识是指对属于同一大类别下的不同子类的图像进行更为精细的识别和分类[1]。不同于传统的图像识别任务,细粒度图像辨识可以关注图像中更加细微的差异,其在日常生活中也有广泛的应用。例如,不同类型汽车的识别、不同类型飞机的识别等。因此,该任务已成为目前计算机视觉研究领域的热点问题。

对于细粒度图像辨识任务而言,由于不同子类之间的差异仅存在于关键细微之处,因此,充分提取图像中目标的可判别区域特征具有极大挑战性[2]。鉴于卷积神经网络[3-4](convolution neural network, CNN)在提取图像局部

收稿日期:2024-05-31
* 基金项目:河南省科技攻关项目(222102210230)、河南理工大学博士基金(B2018-33)项目资助

特征方面的先天优势,因此其与细粒度图像辨识任务场景高度契合。随着近年来 CNN 的快速发展,细粒度图像辨识研究也取得了突破性的进展,逐渐成为细粒度图像辨识领域最为主流的方法。

其主要分为基于目标区域定位和注意力机制识别的方法。如朱丽等[5]提出双路通道注意力并融合残差网络实现特征提取,完成对图像的分类,但该方法缺少对空间信息的关注。基于此,张文轩等[6]提出多分支注意力获取局部关键区域特征并进行自约束的局部裁剪和擦除,提高了对关键区域的辨识能力,但该方法仍无法避免背景信息对关键区域识别造成的影响。随后,刘万军等[7]提出前景特征增强和区域掩码自注意力的识别方法,有效突出前景区域并通过掩码注意力网络学习丰富的特征信息,取得了较好的分类效果,但单线性的网络结构无法完成区域定位和关键特征提取的协同工作。因此,为进一步提升细粒度图像识别精度,Yang 等[8]提出了一种更为复杂的多子网络协同的细粒度分类方法,包括导航(navigator)、指导(advise)、审查(scrutinizer)子网络。具体而言,导航子网络和指导子网络共同完成图像中目标区域的定位工作,而审查子网络负责细粒度特征的提取并完成类别预测。

基于此思想,Zhang 等[9]提出了一种多尺度多分支细粒度图像识别算法(multi-branch and multi-scale attention learning for fine-grained visual categorization, MMAL)。该算法任务分为初级图像分类、目标区域定位、最具判别力区域识别 3 个阶段。第一阶段用于实现对输入图像进行特征提取并完成初步分类;然后基于目标区域定位模块(attention object location module, AOLM)对特征图进行目标区域定位,并裁剪出目标区域;最后,利用关键部件区域生成模块(attention part proposal module, APPM)将目标区域划分成不同尺寸的子区域,并识别出子区域中目标最具判别力区域的特征。然而,由于初步分类网络的特征提取部分采用了传统卷积,导致难以提取不规则的目标区域特征。同时,AOLM 采用通道的全局平均池化来获取特征图中的目标区域,易受背景噪声干扰,这些因素均影响了目标区域定位的准确性。此外,视觉变化、遮挡等因素进一步影响了细粒度特征的获取,进而降低了分类的精度。

为了进一步提高细粒度图像辨识的性能,本文在 MMAL 网络为基础上引入多尺度形变思想,提出了一种基于改进 MMAL 的细粒度图像分类研究方法(analysis of fine-grained image classification through improved MMAL, IMAFC)。通过多尺度形变分组残差(multi-scale deformable convolutional residual,MDCR)模块,在提取的特征图上动态调整感受野的尺寸,使模型能够在各个尺度上自适应感知图像中的目标特征,从而提高目标区域的定位精度。同时,多尺度特征信息的分组融合不仅能够捕获目标区域的细节特征和全局上下文信息,还能在一定程度上减少计算参数量。其次,为减少定位过程中背景噪声的影响,受文献[10]的启发,提出利用 GradCAM[11]通过梯度回流的方法得到相应注意力热图,相对于沿通道全局平均的方法,梯度回流可有效减少背景噪声干扰。最后,为准确获取图像中目标的细粒度特征,提出了一种位置感知空间注意力(location aware spatial attention, LASA)方法。借助融合坐标位置信息和双尺度空间信息,进一步增强了网络对目标区域细粒度特征的提取能力。通过在 3 个公共细粒度数据集上展开实验,实验结果表明本文提出的方法具有良好的分类结果和性能。

1 本文方法

1.1 IMAFC 网络架构

IMAFC 网络架构如图 1 所示,由 3 个子网络构成。首先,目标定位子网络采用 ResNet50 网络的 Conv1 到 Conv4 层作为特征提取骨干网络,通过融合 MDCR 模块动态调整卷积来操作感受野的尺寸和形状,自适应感知输入图像中不同大小的目标区域特征。然后,将目标定位子网络的分类预测值进行反向传播,计算 Conv4 输出特征图的梯度信息,生成 GradCAM 特征图,并对其进行二值化处理,以获取图像中的目标区域并进行裁剪。接着,部件生成子网络以裁剪后的目标图像作为输入,在 LASA 模块的作用下,通过融合空间位置坐标和双尺度空间信息,增强目标区域的细粒度特征。最后,将增强后的特征图送入 APPM,获得不同大小窗口的特征图,并在通道维度上激活各窗口特征图,依据激活值大小挑选包含最具判别区域的部件图像,进行最终分类。

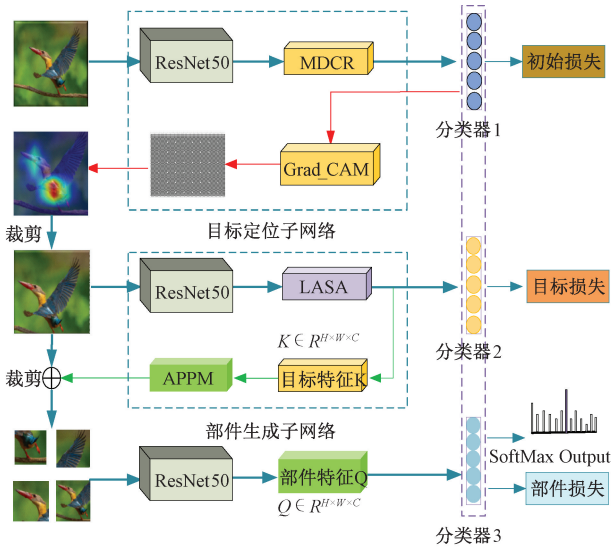

图 1 IMAFC 网络架构图

Fig. 1 IMAFC network architecture

对于 IMAFC 网络而言,MDCR 模块、GradCAM 热力图定位模块及 LASA 模块是提升其性能的关键,下文将详细描述其具体实现。

1.2　MDCR 模块

MDCR 模块由分组卷积[12]和多尺度的形变卷积[13]组成，其网络结构如图 2 所示。

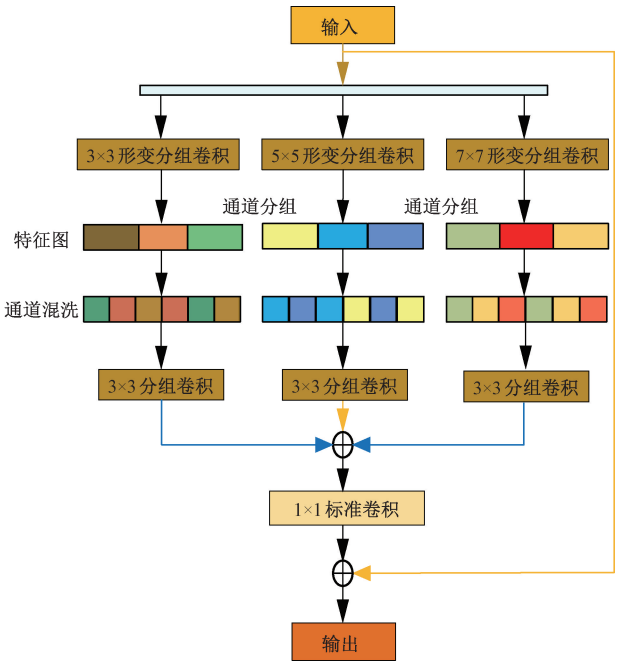

图 2　多尺度形变分组残差模块架构

Fig. 2　Multi-scale deformable grouped residual module architecture

首先，该模块首先使用不同尺度卷积对输入 X 进行特征提取，然后通过引入形变卷积来学习每个卷积核的偏移量，并将该偏移量 ΔP_k 应用到固定的采样位置，使卷积核能够动态调整采样区域，从而更好地捕捉和表征复杂的形变特征。为解决采样位置非整数像素坐标的问题，采用双线性插值的方法获取该采样位置的像素值。如式(1)所示，其中 ΔP_k 和 ΔM_k 分别代表卷积核中第 k 个位置处的可学习偏移量与调制标量，W_k 和 P_k 分别表示卷积核在第 K 个位置所代表的权重值和预设偏移量。

$$X(p)=\sum_{k=1}^{K} w_k \times x(p+p_k+\Delta p_k)\times \Delta m_k \tag{1}$$

为实现输出特征间通道信息的交互，首先利用分组卷积的思想对各尺度下获得的特征进行分组处理，然后采用通道混洗的方法增强特征的跨通道交流。具体地，将式(1)所得的输出特征图按通道重新划分成 G 组并做 Reshape 处理，使其通道维数扩展为两维，即 $V\in(N,C,c/g,H,W)$。然后对各子组内的通道按照交替排列重新规划通道顺序，再将重新规划后的各子组依照原始顺序排列，实现各组之间的通道特征交互。详细实现步骤如式(2)所示。

$$\begin{cases} Y_1=Gconv(Transpose(\mathrm{ReShape}(Dconv_1(\mathrm{X1})))) \\ Y_2=Gconv(Transpose(\mathrm{ReShape}(Dconv_2(\mathrm{X2})))) \\ Y_3=Gconv(Transpose(\mathrm{ReShape}(Dconv_3(\mathrm{X3})))) \end{cases} \tag{2}$$

最终，在通道维度上聚合各尺度的输出特征 Y_1，Y_2，Y_3，并将所得深层特征与初始输入特征融合，以增强模型的学习能力和特征表示的丰富性。公式化描述如式(3)所示。

$$F_{out}=Conv_1(Concat(Y1,Y2,Y3))+X \tag{3}$$

MDCR 模块采用多尺度的动态感受野，使模型自适应地感知图像中不同大小的目标区域特征，极大地改善了传统卷积存在的窗口局限性问题，同时引入分组处理的方法在一定程度上减小了多尺度卷积操作带来的计算量。

1.3　GradCAM 热图定位及目标区域裁剪

图像中目标区域的准确定位对细粒度分类至关重要。虽然基于通道全局平均池化的方法可以轻松获取特征图中的目标区域，但其易受噪声干扰，尤其是背景复杂时，影响更为明显。相比之下，基于梯度信息的 CAM 方法可以有效减少定位过程中噪声的影响。因此，IMAFC 网络采用 GradCAM 梯度回流方法来定位图像中的目标区域。

如图 1 所示，基于定位子网络的分类输出 y^c，通过反向传播预测值获取最终特征图的梯度信息，再依据梯度信息为各个通道分配权重，如式(4)所示。

$$a_k^c=\frac{1}{H\times W}\sum_i\sum_j\frac{\partial y^c}{\partial T_{i,j}^k} \tag{4}$$

式中：T^k 表示输出通道中第 k 个特征图，T_{ij}^k 表示在第 k 个特征图中坐标(i,j)位置处的数据信息。

求得各通道特征权重后，对各通道特征进行加权求和并引入 ReLU 函数激活，生成 GradCAM 特征图，即类激活图。最后通过上采样得到与输入图像尺寸相同的热力图。该过程如式(5)和(6)所示。

$$gradcam=\mathrm{ReLU}\sum_k a_k^c A^k \tag{5}$$

$$heat=upSample(gradcam) \tag{6}$$

最后通过对热力图进行二值化处理来获取目标定位模版，其过程如式(7)和(8)所示。

$$\tilde{a}=\frac{\sum_{x=0}^{W-1}\sum_{y=0}^{H-1}heat(x,y)}{H\times W} \tag{7}$$

$$Mask^{crop}=\begin{cases}1, & heat(x,y)\geqslant\beta\tilde{a}\\ 0, & \text{其他}\end{cases} \tag{8}$$

式中：H 和 W 代表热力图的高度和宽度。

不同于常规的二值化方法，本文引入了一个可调系数 β，通过调整该系数，更精确地获取图像中的目标区域。

GradCAM 热力图定位方法通过引入梯度信息，减小全局平均池化在获取特征图的目标区域时的背景噪声干扰，并将梯度信息与特征图进行加权平均，突出显示对分类决策贡献最大的区域，从而提高网络的目标定位性能。

1.4　LASA 机制

细粒度图像的类别间差异细微，导致其难以区分，影响最终的分类结果。为了增强网络对目标细微特征的表达能力，引入注意力机制是一种有效的解决方案。常用的视觉注意力机制包括通道注意力和空间注意力两种。通道注意力 SE-Net[14]使网络能够关注显著特征通道并抑制无用特

征通道,但忽略了对空间特征的关注,从而限制了其性能。基于此,CBAM[15]通过自适应地学习特征图中通道和空间注意力权重,在通道和空间两个方面提升了网络的特征表示能力。然而,CBAM 未能有效捕捉图像中远距离像素点间的位置关系,导致模型无法准确理解图像中目标的形状和空间位置,从而难以识别出目标的关键判别区域。

基于此,本文提出一种新型 LASA 机制,该机制在通道维度引入位置坐标信息,有效地捕获像素间的相对位置关系。同时,在空间维度提取双尺度的空间上下文信息,通过结合坐标位置信息和空间特征信息,使网络模型能够更准确地关注图像中目标区域的细粒度特征。LASA 机制的网络结构如图 3 所示。

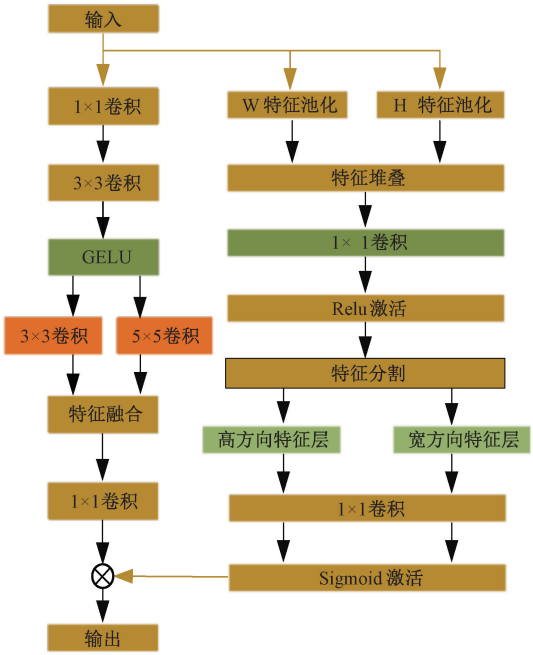

图 3 位置感知空间注意力机制

Fig. 3 Position-aware spatial attention mechanism

该注意力机制分为两个关键部分:通道位置信息嵌入和空间特征提取。传统的全局平均池化会导致输入特征图的空间位置信息丢失,因此在通道位置信息嵌入部分,对输入特征图 $X \in (N,C,H,W)$ 在 W、H 两个方向分别进行特征池化,得到 W 和 H 方向的池化特征,分别表示为 $X_1 \in (N,C,H,1)$ 和 $X_2 \in (N,C,1,W)$。随后,对 W 方向的池化特征图 X_1 进行转置操作,并在最后的通道维度对 W 和 H 方向特征进行整合处理,增强其对空间信息的感知能力,表示为 $X_3 \in (N,C,1,W+H)$。再采用 1×1 卷积层对其进行通道降维,并引入 ReLU 激活函数,使网络学习更复杂的函数映射。

随后,将经过非线性变换后的特征图 X_3 重新分割成 W 和 H 方向上的特征图,并调整其通道维度。同时引入 Sigmoid 激活函数,生成两个方向的注意力权重,其向量特征分别表示为 $X_4 \in (N,C,1,W)$ 和 $X_5 \in (N,C,H,1)$。最后,沿着 W 和 H 方向对特征向量进行扩展,并将扩展后的特征向量逐元素相乘,以获得在两个方向上的各个像素点的注意力权重。

在空间特征提取部分,采用 1×1 卷积层对输入特征图 X 进行特征维度的调整,将原通道 C 转变为 C/R,其输出特征图表示为 $D \in (N,C/R,H,W)$。然后使用 3×3 大小的卷积核结合 GELU 非线性激活函数对特征图进行处理,增强模型对特征图中局部信息的提取能力和非线性表达能力。丰富的特征表示是提高模型性能的关键,而单一尺度下的特征提取限制了特征的多样性,因此采用双尺度感受野获取不同尺度的特征,增强特征的多样性和表达能力,并通过加权求和的方式进行特征融合,实现对全局上下文信息和局部细节的关注。最后恢复原始的通道维度,并将其与拓展所得的特征向量进行交互融合,获得结合位置嵌入信息的空间特征。

与传统的通道和空间注意力机制相比,LASA 不仅在捕获图像中远距离像素点之间的相互关系方面表现出明显优势,而且结合图像目标的位置信息与空间信息,使得网络能够更准确地聚焦于图像中关键区域的具体位置特征,从而提升了网络对目标区域特征的识别性能。

2 实验设计与结果分析

2.1 实验数据集

为了公平地验证本文方法的性能,所有实验均在 CUB-200-2011[16]、FGVC-Aircraft[17] 和 Stanford Cars[18] 三个知名的细粒度图像数据集上进行。其中,FGVC-Aircraft 数据集包含来自 100 种不同飞机类的 9 980 张图像,CUB-200-2011 数据集包含来自 200 种不同鸟类的 11 788 张图像,而 Stanford Cars 数据集由 196 种不同汽车类型的 16 185 张图像组成。这些数据集的类别及训练集、测试集的详细划分情况如表 1 所示。

表 1 细粒度图像数据集详细信息

Table 1 Detailed information of fine-grained image datasets

数据集	类别数	训练集/张	测试集/张
Aircraft	100	6 647	3 333
CUB	200	5 994	5 794
Cars	196	8 144	8 041

同时,本文选择通用的准确率(Accuracy)来评判所提方法的分类效果,计算公式如下:

$$Accuracy = \frac{T\_num}{A\_num} \tag{9}$$

式中:T_num 和 A_num 分别代表测试准确的样本数量和测试集总样本数量。

2.2 实验细节与数据预处理

本文的网络架构基于 Pytorch 实现,所有实验均在配

置有 Ubuntu20.04.6LTS 操作系统的计算机上进行。该机器搭载了一块拥有 16 GB 显存的 NVIDIA GeForce RTX 4080 显卡和一款 Intel Core i7-12700 处理器。软件环境配置包括：Cuda12.0 版本，Cudnn8.8.0 版本，以及 Python3.9.13 版本。

IMAFC 网络架构采用 ResNet50 作为特征提取的骨干网络，以提取输入图像的特征信息。为了实现更有效的特征提取，将原始图像分类子网络和目标区域定位子网络的输入尺寸大小为 480×480，并将关键部位特征提取子网络的输入尺寸调整为 224×224；采用随机梯度下降(SGD)算法，将批次数量(batchsize)设定为 4 个样本以提高计算效率。动量参数(momentum paramete-r)设置为 0.9，学习率(learning rate)设定为 0.002，训练轮数为 200 轮，并将权重衰减参数(weight decayparameter)设置为 0.000 1。为了更精细地控制学习率的下降过程，本文采用了余弦退火(cosine annealing)策略来调整学习率。此外，通过水平翻转、颜色变换及图像裁剪等方式增加样本数据，以获取更丰富的样本信息，从而提升模型的泛化性和鲁棒性。样本数据展示如图 4 所示。

图 4　样本增强效果展示图

Fig. 4　Sample enhancement effect display

2.3　消融实验对比

1)裁剪参数 β 消融实验

为探究裁剪参数 β 对最终分类结果的影响，本文在 FGVC-Aircraft 数据集上进行了多组对比试验，通过设置不同的 β 值来观测最终实验结果。由表 2 中的数据分析可知，当 β=1.5 时，分类精度最高。当 β 值较小时，导致裁剪区域过大，无法精确定位关键区域；而当 β 值较大时，导致裁剪区域过小，导致某些关键区域丢失，从而使分类精度下降。

表 2　裁剪参数 β 消融实验

Table 2　Ablation study of the cropping parameter β

β	1.2	1.3	1.4	1.5	1.6	1.7
Acc/%	92.8	93.1	93.4	**93.7**	93.5	93.4

2)模块有效性验证

为验证所提出的模块对细粒度图像识别中的有效性，本文在 FGVC-Aircraft、CUB-200-2011 和 Standford Cars 三个知名数据集上进行了对比实验，实验结果如表 3 所示。表中共进行 7 组实验，各模块在独立和组合使用时对网络模型的提升效果。其实验结果证实了本文所提出的模块在细粒度图像识别中展现了良好的识别性能，有效提升了网络的分类准确度。

表 3　消融实验分析

Table 3　Ablation study analysis

实验	模型				准确率/%		
	MMAL	GradCAM	MDCR	LASA	Aircraft	CUB	Cars
1	√	—	—	—	93.0	88.3	93.7
2	√	√	—	—	93.7(+0.7)	88.9(+0.6)	94.2(+0.5)
3	√	—	√	—	93.5(+0.5)	88.5(+0.2)	94.0(+0.3)
4	√	—	—	√	93.9(+0.9)	88.7(+0.4)	94.5(+0.8)
5	√	√	√	—	94.2(+1.2)	89.0(+0.7)	94.8(+1.1)
6	√	√	—	√	93.9(+0.9)	89.4(+1.1)	94.5(+0.8)
7	√	√	√	√	94.9	89.7	95.2

2.4　现有算法对比实验

为证明本文方法的良好性能，将其与现有先进的细粒度图像分类方法在知名公开数据集上进行对比测试，以评估各算法的分类性能优劣。各算法的识别准确率如表 4 所示。

根据表中各数据集实验结果，本文方法相较于 MMAL 基线网络在识别准确率上有显著提升，充分展示了其有效性和优越的识别效果。此外，在同类算法中，IMAFC 在 3 个数据集上的识别准确率均高于以 CNN 为主干网络的其他分类方法，显示出良好的识别性能，优于大多数现有的方法。

2.5　定位准确度对比

在细粒度图像识别任务中，样本图像目标区域的准确定位至关重要，其定位准确度直接影响到分类的最终准确率。为验证本文所采用的 GradCAM 热图定位方法的有效性，采用多种定位方法与之进行比较，详细的实验结果如表 5 所示。

表 4 相关算法在公开数据集下 Top-1 准确率

Table 4 Top-1 accuracy of relevant algorithms on public datasets

算法	准确率/%		
	CUB	Aircraft	Cars
DFL-CNN[19]	87.4	92.0	93.8
DCL[20]	87.8	93.0	94.5
LIO[21]	88.0	92.7	94.5
FDL[22]	88.6	93.4	94.3
PART[23]	89.6	94.4	95.1
WS-DAN[24]	89.4	93.0	94.5
API-Net[25]	87.7	93.0	94.8
Vit[26]	90.2	92.1	93.7
MMAL-Net[9]	88.3	93.0	93.7
IMAFC	89.7	94.9	95.2

表 5 相关方法在 CUB-200-2011 数据集下定位精度

Table 5 Localization accuracy of related methods on the CUB-200-2011 dataset

方法	定位精度/%
	CUB-200-2011
ADL[27]	64.5
SCDA[28]	76.3
AOLM[9]	82.6
CAM[29]	83.1
GradCAM	85.4

根据表中各方法所得定位精度可知,相较于基线网络中采用的 AOLM 和其他定位方法,本文所采用的 GradCAM 热图定位方法取得了更高的定位精度,显示出更优的定位性能,这有助于更好的完成细粒度分类任务。

2.6 准确率变化过程

为了直观的展示各模块在训练阶段对分类结果的影响,本文在 CUB-200-2011 数据集上进行了试验。根据实验数据,制作了折线图来展示 MMAL 网络和 IMAFC 网络中迭代轮数与分类准确率的变化关系,其中 B、C、D、E、F、G、H 分别对应表 3 消融实验中基线网络添加不同模块的折线图。详细的折线图变化如图 5 所示。

实验数据分析表明,在训练过程的前 90 轮中,与基线网络相比较,各模块分别显现出分类性能的提升,并达到较高水平。添加 3 个模块的 IMAFC 网络分类准确率相比基线网络分类准确率提升了 1.4%,此结果验证了本文方法的有效性。

2.7 网络的可视化分析

为验证所提方法在目标关键特征提取方面的显著效果,本文在 3 个数据集上进行了实验,并对特征提取网络的输出进行了可视化,展示了本文方法在细粒度图像中对显著目标区域的特征提取性能。图 6 展示了在不同数据集下的图像可视化输出结果。

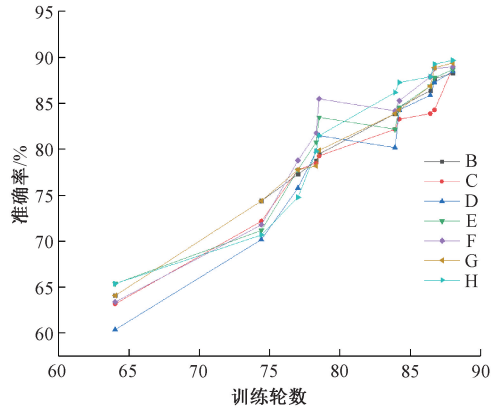

图 5 CUB-200-2011 数据集准确率变化曲线

Fig. 5 Accuracy variation curve on the CUB-200-2011 dataset

图 6 数据样本热力图示例

Fig. 6 Data sample heatmap example

根据图示结果,本文所提出的 IMAFC 网络相比与 MMAL 网络,在识别图像中关键区域方面表现出更全面和更精准的识别能力。热图显示,在鸟类图像中,IMAFC 网络主要关注眼睛、头部以及鸟爪区域;在飞机图像中,主要集中在机翼和涡扇等区域;而汽车图像中,则更加注重车标、车灯等细微区域。由此验证了 IMAFC 网络在不同类型的细粒度图像中对目标关键区域的识别性能表现更为出色,具有更强的识别能力。

3 结 论

本文提出了一种基于多尺度分组形变和位置感知的细粒度图像分类算法,该算法利用多尺度形变分组残差模块,通过动态调整卷积核的位置和形状,增强了模型对不同尺度和形状目标的特征表达能力。此外,采用 GradCAM 梯度回流方法定位目标区域,有效减轻了图像特征中背景噪声干扰,从而提升了网络对目标区域的识别和定位性能。最后,本文提出了位置感知空间注意力机制,通过结合坐标位置信息和双尺度空间特征信息,进一步提升了网络对图像目标区域细粒度特征的识别能力。实验结果表明,IMAFC 网络在 3 个细粒度图像数据集上

均表现出优异的识别性能。

参考文献

[1] 张志林,李玉鑑,刘兆英,等.深度学习在细粒度图像识别中的应用综述[J].北京工业大学学报,2021,47(8):942-953.
ZHANG ZH L, LI Y J, LIU ZH Y, et al. Deep learning for fine-grained image recognition: A survey[J]. Journal of Beijing University of Technology, 2021,47(8):942-953.

[2] 罗建豪,吴建鑫.基于深度卷积特征的细粒度图像分类研究综述[J].自动化学报,2017,43(8):1306-1318.
LUO J H, WU J X. A survey of fine-grained image categorization using deep convolutional features[J]. Acta Automatica Sinica, 2017,43(8):1306-1318.

[3] 朱阳光,刘瑞敏,黄琼桃.基于深度神经网络的弱监督信息细粒度图像识别[J].电子测量与仪器学报,2020,34(2):115-122.
ZHU Y G, LIU R M, HUANG Q T. Fine-grained image recognition of weak supervisory information based on deep neural network [J]. Journal of Electronic Measurement and Instrumentation, 2020, 34(2):115-122.

[4] 齐爱玲,王宣淋.融合通道与位置信息的 ResNet 细粒度图像识别[J].国外电子测量技术,2022,41(12):103-111.
QI AI L, WANG X L. ResNet fine-grained image identification with fused channel and location information [J]. Foreign Electronic Measurement Technology,2022,41(12):103-111.

[5] 朱丽,王新鹏,付海涛,等.基于注意力机制的细粒度图像分类[J].吉林大学学报:理学版,2023,61(2):371-376.
ZHU L, WANG X P, FU H T, et al. Fine-grained image classification based on attention mechanisms[J]. Journal of Jilin University: Science Edition, 2023, 61(2): 371-376.

[6] 张文轩,吴秦.基于多分支注意力增强的细粒度图像分类[J].计算机科学,2022,49(5):105-112.
ZHANG W X, WU Q. Fine-grained image classification based on multi-branch attention-augmentation[J]. Computer Science, 2022,49(5): 105-112.

[7] 刘万军,赵思琪,曲海成,等.结合前景特征增强与区域掩码自注意力的细粒度图像分类[J].智能系统学报,2022,17(6):1134-1144.
LIU W J, ZHAO S Q, QU H CH, et al. Combining foreground feature reinforcement and region mask self-attention for fine-grained image classification [J]. Journal of Intelligent Systems, 2022, 17 (6): 1134-1144.

[8] YANG Z, LUO T G, WANG D, et al. Learning to navigate for fine-grained classification[C]. European Conference on Computer Vision (ECCV), 2018: 420-435.

[9] ZHANG F, LI M, ZHAI G SH, et al. Multi-branch and multi-scale attention learning for fine-grained visual categorization[C]. Multi-Media Modeling: 27th International Conference, MMM 2021, 2021:136-147.

[10] XIE J H, LUO CH, ZHU X P, et al. Online refinement of low-level feature based activation map for weakly supervised object localization[C]. IEEE/CVF International Conference on Computer Vision, 2021: 132-141.

[11] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization[C]. IEEE International Conference on Computer Vision, 2017: 618-626.

[12] IOANNOU Y, ROBERTSON D, CIPOLLA R, et al. Deep roots: Improving cnn efficiency with hierarchical filter groups [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017:1231-1240.

[13] DAI J F, QI H ZH, XIONG Y W, et al. Deformable convolutional networks [C]. IEEE International Conference on Computer Vision, 2017:764-773.

[14] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2018:7132-7141.

[15] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]. European Conference on Computer Vision(ECCV), 2018:3-19.

[16] WAH C, BRANSON S, WELINDER P, et al. The caltech-UCSD birds-200-2011 dataset[J]. California Institute of Technology, 2011.

[17] MAJI S, RAHTU E, KANNALA J, et al. Fine-grained visual classification of aircraft [J]. ArXiv preprint arXiv:1306.5151, 2013.

[18] KRAUSE J, STARK M, DENG J, et al. 3D object representations for fine-grained categorization [C]. IEEE International Conference on Computer Vision Workshops, 2013: 554-561.

[19] WANG Y, MORARIU V I, DAVIS L S. Learning a discriminative filter bank within a CNN for fine-grained recognition [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2018:4148-4157.

[20] CHEN Y, BAI Y L, ZHANG W, et al.

Destructionand construction learning for fine-grained image recognition [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019:5157-5166.

[21] ZHOU M H, BAI Y L, ZHANG W, et al. Look-into-object: Self-supervised structure modeling for object recognition [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 11774-11783.

[22] LIU CH B, XIE H T, ZHA ZH J, et al. Filtration and distillation: Enhancing region attention for fine-grained visual categorization[C]. AAAI Conference on Artificial Intelligence, 2020, 34(7): 11555-11562.

[23] ZHAO Y F, LI J, CHEN X W, et al. Part-guided relational transformers for fine-grained visual recognition [J]. IEEE Transactions on Image Processing, 2021, 30: 9470-9481.

[24] HU T, QI H G, HUANG Q M, et al. See better beforelooking closer: Weakly supervised data augmentation network for fine-grained visual classification [J]. ArXiv preprint arXiv:1901.09891, 2019.

[25] ZHUANG P Q, WANG Y L, QIAO Y. Learning attentive pairwise interaction for fine-grained classification [C]. AAAI Conference on Artificial Intelligence, 2020, 34(7): 13130-13137.

[26] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale [J]. ArXiv preprint arXiv:2010.11929, 2020.

[27] CHOE J, SHIM H. Attention-based dropout layer for weakly supervised object localization[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 2219-2228.

[28] WEI X SH, LUO J H, WU J X, et al. Selective convolutional descriptor aggregation for fine-grained image retrieval [J]. IEEE Transactions on Image Processing, 2017, 26(6): 2868-2881.

[29] ZHOU B, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2921-2929.

作者简介

李冰锋,博士研究生,讲师,主要研究方向为迁移学习、计算机视觉与目标检测。

E-mail:libingfeng@hpu.edu.cn

龚得魁(通信作者),硕士研究生,主要研究方向为计算机视觉与目标检测。

E-mail:j1326553@163.com

杨艺,博士研究生,副教授,主要研究方向为深度学习与强化学习。