ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 40,No. 2,Mar. 2025,pp. 384－400
DOI:10. 16337/j. 1004-9037. 2025. 02. 009

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于注意力机制和多尺度集成学习的细粒度图像识别方法

季晟宇，江志康，马　翔，杨绿溪

（东南大学信息科学与工程学院，南京 211102）

摘　要：细粒度图像识别是计算机视觉领域中一项重要的研究课题，其主要目标是分辨同属一大类下外观具有高度相似性的子类。以弱监督的细粒度图像识别为研究内容，针对现有研究中存在的图像细粒度特征利用不充分以及判别性区域难以挖掘的问题，提出了基于注意力机制和多尺度集成学习策略的细粒度图像识别方法。该方法引入渐进式学习网络，利用集成学习的策略，基于深度神经网络3个层级的输出特征并行构建多尺度基分类器，并使用标签平滑的方法对分类器进行渐进式训练，从而大幅度提高低层特征的利用率；同时采用高效双通道注意力机制对特征施加通道权重，使得网络能够在通道层面自主筛选特征，从而提升高信息相关度通道的利用率。该方法还引入了自注意力区域建议网络，通过构建循环反馈机制促使模型逐步定位到更加具有判别性的区域，并在最后的分类模块中将完整图像与判别性区域的特征信息进行融合。实验结果表明，该方法在CUB-200-2011、FGVC Aircraft和Stanford Cars细粒度图像数据集上的识别准确率达到行业先进水平。

关键词：深度学习；细粒度图像识别；弱监督；注意力机制；集成学习

中图分类号：TP391　　**文献标志码：**A

Fine-Grained Image Recognition Method Based on Attention and Multi-scale Ensemble Learning

JI Shengyu, JIANG Zhikang, MA Xiang, YANG Lvxi

(School of Information Science and Engineering, Southeast University, Nanjing 211102, China)

Abstract: Fine-grained image recognition (FGIR) is an important research topic in the field of computer vision. Its main goal is to distinguish subclasses with high similarity in appearance under the same category. This paper focuses on the research of weakly-supervised fine-grained image recognition technology. Given the problems of insufficient use of feature of fine-grained images and difficulty in digging discriminative regions existing in the research of FGIR, the attention and multi-scale ensemble-learning based network (AMEN) is proposed. This method introduces a progressive learning network, which uses the strategy of ensemble learning to construct multi-scale base-classifiers based on three levels of output features of deep neural network in parallel, and uses the label smoothing method to carry out progressive training for multi-scale base-classifiers, so as to greatly improve the utilization of low-level features. At the same time, the efficient dual channel attention is used to impose channel weights on features, so that the network can independently select features at the channel level, so as to improve the utilization of high information

基金项目：国家自然科学基金(61971128)。
收稿日期：2024-04-29；**修订日期：**2024-09-17

correlation channels. This method also introduces a self-attention region proposal network, which promotes the model to gradually locate the more discriminative region by constructing a circular feedback mechanism, and fuses the feature information of the complete image and the discriminative region in the final classification module. Experimental results show that the recognition accuracy of AMEN on three fine-grained image datasets of CUB-200-2011, FGVC Aircraft and Stanford Cars has reached the advanced level of the field.

Key words: deep learning; fine-grained image recognition; weakly supervised annotation; attention mechanism; ensemble learning

引　　言

细粒度图像识别技术经历了从依赖人工设计特征到利用深度神经网络自主学习特征的转变,其中,基于深度神经网络的方法依据训练过程中使用监督信息的强弱,可分为基于强监督的方法和基于弱监督的方法[1]。基于强监督的细粒度图像识别方法一般会用到目标边界框标注和精细部位特征点标注,Zhang等[2]基于R-CNN[3]提出的Part-based R-CNN就是一种典型的强监督细粒度图像识别算法。Branson等[4]对Part-based R-CNN算法进行改进提出了姿态归一化网络PN-CNN。虽然强监督细粒度图像识别算法在细粒度识别任务上取得了令人满意的识别精度,但其借助人工标注的额外监督信息涉及大量的人力和财力资源的投入,这也限制了该技术在实际场景中的应用。

得益于深度学习技术[5]的不断发展,仅依赖于图像级标签的弱监督细粒度图像识别技术的研究已经取得了巨大的进步,并逐渐成为了细粒度图像识别领域的主流研究方向。Lin等[6]提出一种模仿大脑工作方式的双线性卷积神经网络Bilinear CNN,其使用两路相互独立的卷积神经网络对图像的全局视觉特征进行提取,之后通过矩阵外积运算得到双线性特征描述矩阵,从而捕获细粒度图像的二阶统计信息。Fu等[7]提出一种循环注意力卷积神经网络RA-CNN,由3个不同尺度且参数相互独立的子网络组成,每个子网络的结构一样,都包含注意力区域建议网络(Attention proposal network,APN)和分类网络。通过APN可得到局部推荐区域的空间位置信息,接着将其从原始图像中裁剪下来,再使用双线性差值进行放大后输入到下一尺度的子网络。Zheng等[8]提出一种三线性注意力采样网络(Trilinear attention sampling network, TASN),包含三线性注意力模块、基于注意力的采样器和特征蒸馏器。三线性注意力模块通过建模特征图通道间的关系来生成注意力图;基于注意力的采样器通过随机选取一个注意力图生成细节保留的图像,通过平均化注意力图生成结构保留的图像;特征蒸馏器通过权值共享和特征保存策略,将部分特征蒸馏到一个全局的特征。除此之外,Ji等[9]提出了一种注意力卷积二元神经树架构,应对因变形、遮挡等因素导致的高类内方差和低类间方差的问题。Ding等[10]提出了一种注意力金字塔卷积神经网络,通过捕捉图像的低级信息,如颜色、边缘和纹理等,来增强特征表示和判别区域的精确定位,从而提升网络的整体性能。Shu等[11]提出了一种自增强注意力机制,通过将网络正则化去关注跨样本和类所共享的关键区域。

在细粒度图像识别问题中,多数方法只使用深度神经网络的顶层特征,这样的做法忽略了低层局部细节特征,导致细粒度特征利用不充分。此外,多数弱监督方法借助卷积神经网络提取特征中所包含的空间位置信息,来帮助模型定位到图像中具有判别性的区域,然而这种方法存在关注区域单一、不准确的问题,从而导致有效的判别性区域难以被挖掘。本文以弱监督的细粒度图像识别为研究内容,针对现有研究中存在的图像细粒度特征利用不充分的问题,以及弱监督条件下判别性区域难以挖掘的问题,提出了一种基于注意力机制和多尺度集成学习策略的细粒度图像识别方法。

1 本文方法

提出了一种基于注意力机制和多尺度集成学习策略的细粒度图像识别方法(Attention and multi-scale ensemble-learning based network, AMEN),其结构包含渐进式学习网络(Progressive learning network, PLN)、自注意力区域建议网络(Self-attention region proposal network, SRPN)和分类网络3个部分。其中,PLN利用集成学习的策略,基于深度神经网络不同层的输出特征构建多尺度基分类器,并使用标签平滑的方法根据不同尺度基分类器的学习能力设置相应的软标签进行监督。SRPN利用主干网络输出的顶层特征和锚框机制生成局部候选区域,并通过构建循环反馈机制促使网络逐步定位到更加具有判别性的区域。分类网络将完整图像与判别性区域的特征信息进行对应尺度的融合,并基于融合特征构建分类器进行渐进式训练,最后采取相应的集成策略汇聚多个分类器的结果进行类别的预测。此外,还提出了一种高效双通道注意力机制(Efficient dual channel attention, EDCA),该模块通过对特征施加通道注意力权重,从而帮助模型自主选择相关性高的特征。将多尺度细粒度特征学习和判别性区域定位进行了有效结合,使得模型性能得到显著提升。

1.1 渐进式学习网络

在细粒度特征提取过程中,仅使用深度神经网络顶层特征会导致低层局部细节特征被忽略,进而降低识别模型对小尺度差异信息的敏感度。PLN利用集成学习[12]的策略,基于深度神经网络不同层的输出特征构建多尺度基分类器,并使用标签平滑[13]的方法根据不同尺度基分类器的学习能力设置相应的软标签进行渐进式训练,从而大幅度提高对低层特征的利用率。

1.1.1 多尺度基分类器的构建与集成

PLN基于深度残差ResNet-50网络[14],利用其第三、第四和第五个卷积块的输出分别作为低层、中层和高层特征构建多尺度基分类器,其中的高层特征即为ResNet-50网络输出的顶层特征,网络结构如图1所示。

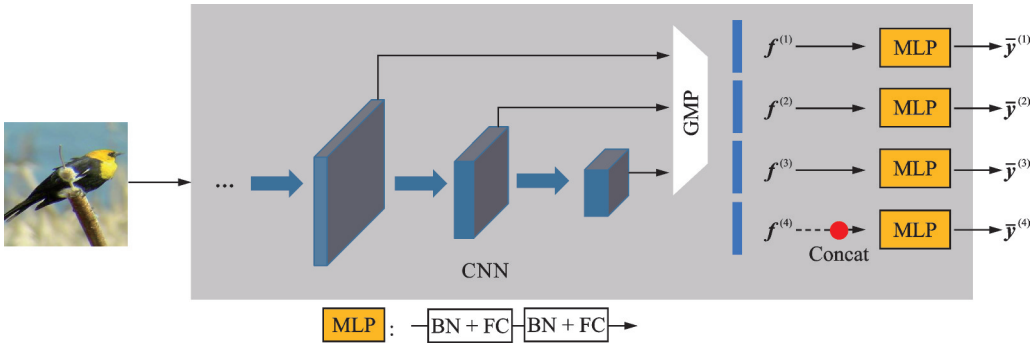

图1 渐进式学习网络结构示意图

Fig.1 Diagram of PLN structure

从图1可看出,渐进式学习网络首先将ResNet-50的低、中、高层特征输入到全局最大池化(Global maximum pooling, GMP)层,全局最大池化是指对输入特征每个通道的全部空间位置取最大值。之后,将全局最大池化层的输出结果展开得到低、中和高层特征描述向量$f^{(1)}$、$f^{(2)}$和$f^{(3)}$,其维度分别为512、1 024和2 048。最后,将低、中、高层特征描述向量$f^{(1)}$、$f^{(2)}$、$f^{(3)}$分别输入对应的多层感知机(Multi-layer perceptron,MLP)从而构建多尺度基分类器。因此,本文构建的分类器是一种多层感知机结构,其将批量标准化层(Batch normalization,BN)和全连接层(Fully connected layers,FC)依序进行组合。此外,本

文还将特征描述向量$\boldsymbol{f}^{(1)}$、$\boldsymbol{f}^{(2)}$、$\boldsymbol{f}^{(3)}$依序进行级联获得多尺度联合特征描述向量$\boldsymbol{f}^{(4)}$,并将$\boldsymbol{f}^{(4)}$也输入对应的多层感知机。因此,本文渐进式学习网络利用多尺度的特征信息一共构建了4个基分类器。

在构建好多尺度基分类器后,使用软投票的集成策略汇聚多个分类器的结果进行类别的预测。具体而言,对4个基分类器输出的预测向量$\{\bar{\boldsymbol{y}}^{(1)},\bar{\boldsymbol{y}}^{(2)},\bar{\boldsymbol{y}}^{(3)},\bar{\boldsymbol{y}}^{(4)}\}$进行相加获得最终的预测向量,选择其中预测值最大的类别作为集成模型的预测结果,其数学表达式如下

$$c=\underset{c_j}{\operatorname{argmax}}\sum_{i=1}^{4}h_i^j \tag{1}$$

式中:h_i^j表示基分类器h_i在类别标记c_j上的预测值,c为集成模型的预测类别。此时,基分类器h_i输出的预测向量可以表示为$\{h_i^j|j=1,2,\cdots,K\}$,其中K为类别总数。

1.1.2 渐进式训练策略

考虑到随着网络的加深,提取到的更高层特征往往具有更强的语义信息,从而对应尺度的基分类器会拥有更强的学习能力,因此提出了一种渐进式策略对多尺度基分类器进行训练。具体而言,使用标签平滑的方法,根据不同尺度基分类器的学习能力设置了相应的软标签,从而调整其学习的难度。考虑一个one-hot编码的图像类别标签向量$\boldsymbol{y}\in\mathbf{R}^K$,其中$K$为类别总数。假设该标签指示的真实类别的索引$m$是一个整数,取值范围为$[0,K-1]$,则标签向量$\boldsymbol{y}$可表示为

$$y[t]=\begin{cases}1 & t=m\\ 0 & t\neq m\end{cases} \tag{2}$$

式中:t表示标签向量的位置索引,取值范围为$[0,K-1]$。使用标签平滑的方式将原始标签向量$\boldsymbol{y}$修改为一个软标签,即

$$y_\alpha[t]=\begin{cases}\alpha & t=m\\ \dfrac{1-\alpha}{K} & t\neq m\end{cases} \tag{3}$$

式中:α为引入的平滑因子,取值范围为$[0,1]$。当α取1时,软标签$\boldsymbol{y}_\alpha$就是原始标签向量$\boldsymbol{y}$。依据式(3),可通过改变平滑因子α来调节软标签$\boldsymbol{y}_\alpha$中真实类别处值的大小,从而设置难易度不同的学习目标。因此,在训练期间可定义不同的软标签来监督不同尺度的基分类器,其损失函数可写为

$$L_{\mathrm{sce}}(\bar{\boldsymbol{y}}^{(i)},\boldsymbol{y},\alpha^{(i)})=L_{\mathrm{ce}}(\bar{\boldsymbol{y}}^{(i)},\boldsymbol{y}_{\alpha^{(i)}})=\sum_{t=0}^{K-1}-\boldsymbol{y}_{\alpha^{(i)}}[t]\log\bar{\boldsymbol{y}}^{(i)}[t] \tag{4}$$

式中:$L_{\mathrm{ce}}(\cdot)$表示交叉熵损失函数,i表示基分类器的索引。将上述损失函数定义为平滑交叉熵损失函数。渐进式学习网络构建了4个基分类器,$i=1,2,3,4$。将多尺度基分类器进行联合训练,因此整个渐进式学习网络整体的损失函数为

$$L_{\mathrm{pln}}=\sum_{i=1}^{4}L_{\mathrm{sce}}(\bar{\boldsymbol{y}}^{(i)},\boldsymbol{y},\alpha^{(i)}) \tag{5}$$

综上所述,考虑本文构建的多尺度基分类器$\mathrm{MLP}^{(i)}$,其对应输入特征为$\boldsymbol{f}^{(i)}$。随着i的增大,$\boldsymbol{f}^{(i)}$所包含的语义信息逐步增加,基分类器$\mathrm{MLP}^{(i)}$的学习能力也随之增强。因此,采取渐进式训练(Progressive training,PT)策略,将平滑因子$\alpha^{(i)}$从0.7逐步增加到1,从而设置由易到难的学习目标匹配基分类器$\mathrm{MLP}^{(i)}$的学习能力,最后将各个基分类器的平滑交叉熵损失求和从而对整个PLN网络进行训练。

1.2 自注意力区域建议网络

针对细粒度图像中判别性区域难以挖掘的问题,设计了一种自注意力区域建议网络(SRPN)。该网络首先利用主干网络输出的顶层特征产生大量可能包含细粒度目标的锚框,并计算这些区域的得

分；然后利用非极大值抑制(Non-maximum suppression, NMS)算法[15]对生成的候选区域进行初步筛选，得到得分最高的N个区域；之后，将这 N 个区域输入共享参数的主干网络计算相应的分类损失；最后，依据分类损失与得分的排序构建循环反馈机制，促使网络逐步定位到更具有判别性的区域。

1.2.1 候选区域生成

为了便于阐述，假设输入图像的尺寸为448×448，使用共享参数的主干网络是ResNet-50网络，那么输入自注意力区域建议网络的顶层特征尺寸即为2 048×14×14。首先使用卷积核尺寸为3×3，步长为1，输出通道数为128的卷积层对输入特征进行降维，生成尺寸为128×14×14 的特征图。之后，再使用两个卷积核尺寸为3×3，步长为2，输出通道数为128的卷积层降低特征的分辨率，分别得到尺寸为128×7×7和128×4×4 的特征图。经过上述操作总计得到了3张不同分辨率的降维特征图。对于这3张特征图中的每一像素点，都将其映射回原始图像产生不同尺度和比例的的锚框(Anchor)。同时在特征图的通道维度上提取像素点对应的128维特征向量，将其输入1×1卷积层获得该像素点所映射锚框的得分(Score)，这一数值代表了锚框所包含的区域属于前景的概率。对3张降维特征图分别使用3个相互独立的1×1 卷积层进行上述操作。在获得了大量可能包含细粒度目标的锚框之后，还需要使用NMS算法对这些区域进行初步筛选。记包含所有生成锚框的集合为S，通过筛选的集合为D，初始集合D为空集。NMS算法对集合S按照锚框的得分进行降序排列，选取其中得分最高的锚框记为 M，将其放入集合D并从集合S中删除。之后，遍历集合S中的所有锚框，计算其与 M 的交并比(Intersection over union, IoU)，如果交并比大于阈值 N_t，则将该锚框从集合S中删除。重复上述步骤，直至集合D中锚框的数量满足设定的要求。经过NMS算法的处理，去除了冗余的锚框，并在集合D中筛选出得分最高的N个区域。候选区域生成模块的结构如图2所示。

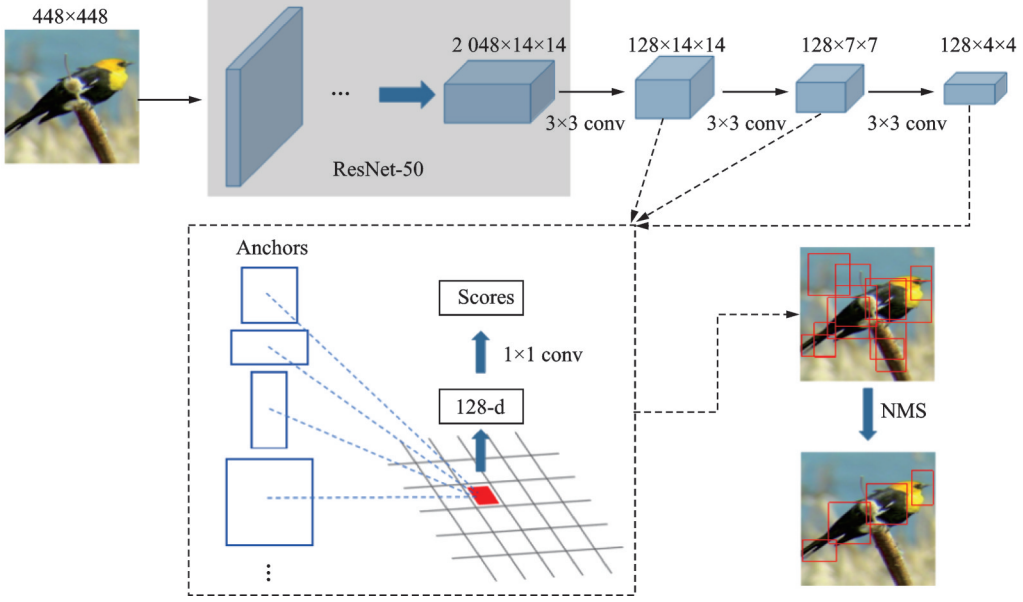

图2 候选区域生成模块结构示意图

Fig.2 Diagram of structure of candidate region generation module

1.2.2 循环反馈机制

经过候选区域生成模块获取了得分最高的N个区域的集合$\{R'_1, R'_2, \cdots, R'_N\}$，以及它们对应分数的集合$\{s_1, s_2, \cdots, s_N\}$，其中 $s_1 \geqslant s_2 \geqslant \cdots \geqslant s_N$。将这些候选区域在原图上进行裁剪后，调整至统一的尺寸

输入共享参数的主干网络。设定这些候选区域的类别标签与完整图像的类别标签相同,那么经主干网络进行特征提取和类别预测后,可得到这些区域分类损失(Loss)的集合$\{l_1, l_2, \cdots, l_N\}$。候选区域的得分代表该区域属于前景的概率,能够反映区域内所蕴含的语义信息丰富程度,而候选区域的分类损失则代表了该区域输入主干网络后得到的预测结果与真实标签的接近程度,能够反映局部区域的判别性。换句话说,分类损失值越低,说明网络对这片候选区域的预测越接近于真实标签,意味着这片区域具有更高的判别性。考虑到蕴含语义信息更为丰富的局部区域应该具有更强的判别性,基于这一判断构建了循环反馈机制,其结构如图3所示。自注意力区域建议网络将得到的分类损失集合$\{l_1, l_2, \cdots, l_N\}$反馈给候选区域生成模块,候选区域生成模块根据反馈结果优化卷积层的参数,进而生成具有更高信息量的候选区域。如此循环往复,直至候选区域得分的排序与分类损失的排序完全相反,即得分越高的区域其分类损失值越低。

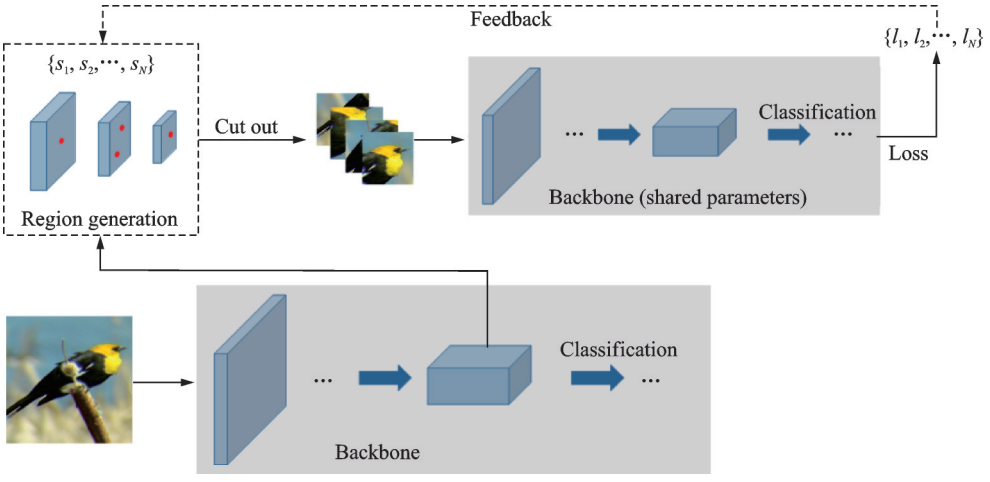

图3 循环反馈机制示意图

Fig.3 Diagram of circular feedback mechanism

1.2.3 训练策略

为了更好地训练所提出的自注意力区域建议网络,需要设计适当的损失函数。在循环反馈机制中,通过将候选区域的分类损失集合$\{l_1, l_2, \cdots, l_N\}$进行反馈,让候选区域生成模块比较损失集合$\{l_1, l_2, \cdots, l_N\}$和得分集合$\{s_1, s_2, \cdots, s_N\}$的关系,进而优化卷积层的参数。考虑给定候选区域的索引i和j,根据1.2.2节循环反馈机制的分析,若损失$l_i < l_j$,则得分s_i应当大于s_j,即$s_i > s_j$。因此,对损失和得分的上述关系进行建模,定义了排序损失函数,其表达式如下

$$L^{\text{rank}} = \sum_{i=1}^{N}\sum_{j=1}^{N} \max(0, l_i - l_j + \delta) * k_{ij} \tag{6}$$

式中:N为候选区域的数目,δ为裕度,是一个超参数,在本文中设置为1,k_{ij}定义如下

$$k_{ij} = \begin{cases} 1 & s_i > s_j \\ 0 & s_i \leqslant s_j \end{cases} \tag{7}$$

结合式(6)和式(7)可看出,当$s_i > s_j$时,排序损失函数鼓励$l_i < l_j$。换句话说,L^{rank}惩罚损失l和得分s之间正向配对,鼓励l和s的顺序相反。理想情况下,通过最小化排序损失L^{rank},候选区域的得分

$\{s_1, s_2, \cdots, s_N\}$和分类损失$\{l_1, l_2, \cdots, l_N\}$应该朝相反的方向变化，最终实现两者排序在反方向达成一致性。

为了充分考虑判别性区域定位与细粒度特征学习之间的联系，在自注意力区域建议网络的训练策略中还引入了分类损失$L_{\text{srpn}}^{\text{cls}}$，其表达式如下

$$L_{\text{srpn}}^{\text{cls}} = \sum_{i=1}^{N} l_i \tag{8}$$

从式(8)可看出，自注意力区域建议网络的分类损失即为所有候选区域的分类损失之和。而$L_{\text{srpn}}^{\text{cls}}$的具体表达形式将在1.5节进行阐述。综上所述，本文联合排序损失和分类损失对自注意力区域建议网络进行训练，其整体损失函数的表达式为

$$L_{\text{srpn}} = L^{\text{rank}} + L_{\text{srpn}}^{\text{cls}} \tag{9}$$

1.3 高效双通道注意力机制

在细粒度特征提取的过程中，多数方法没有充分考虑到对特征的通道信息进行筛选，针对这一问题，受SENet[16]启发，设计了一种高效双通道注意力机制(EDCA)，通过对特征施加通道权重，使得网络能够在通道层面自主筛选特征，从而提升高信息相关度通道的利用率，其结构如图4所示。

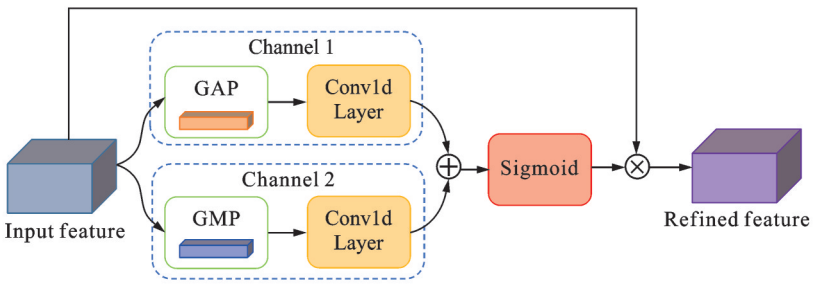

图4 EDCA结构示意图

Fig.4 Diagram of EDCA structure

EDCA利用两条路径分别计算特征的通道权重，然后进行组合，使用组合后的权重与输入特征进行乘法运算，从而实现通道权重的赋予。EDCA使用一维卷积层代替了SENet中的全连接层来捕捉特征通道间的关系，有效降低了参数量，并且避免了使用两层全连接层先降维后升维而对通道与权重间直接对应关系的破坏。其中，一维卷积层中卷积核的尺寸是根据输入的通道数自适应调整的，计算公式如下

$$k = \frac{\log_2 C + b}{\gamma} \tag{10}$$

式中：C为输入的通道数，b和γ为固定的超参数，本文将其分别设置为1和2。在实际操作中，对于计算得到的卷积核尺寸k，需要进行向下取整，如若得到的结果为偶数还需要加一将其变为奇数。此外，卷积步长设置为1，填充设置为卷积核尺寸的一半，向下取整。

在EDCA的通道1中，首先使用全局平均池化(Global average pooling, GAP)对输入特征进行挤压操作，然后使用一维卷积层计算通道权重。在通道2中，使用全局最大池化(GMP)对输入特征进行挤压操作，然后同样使用一维卷积层计算通道权重。值得注意的是，两个通道所使用的卷积层相互独立。上述操作过程可用公式描述为

$$s^1 = \mathrm{Conv}^1(\mathrm{GAP}(\boldsymbol{U})) \tag{11}$$

$$s^2 = \mathrm{Conv}^2(\mathrm{GMP}(\boldsymbol{U})) \tag{12}$$

式中:$\boldsymbol{U}$为输入特征。之后,将两个通道计算得到的权重s^1和s^2逐元素相加(Add)进行组合,然后采用Sigmoid激活函数将权重映射到0至1之间。将最后得到的权重值与输入特征做乘法运算。上述操作过程可用公式描述为

$$s = \sigma(s^1 \oplus s^2) \tag{13}$$

$$\boldsymbol{U}' = s * \boldsymbol{U} \tag{14}$$

式中:$\sigma(\cdot)$表示Sigmoid激活函数,"$\oplus$"表示逐元素相加运算,"$*$"表示对输入特征$\boldsymbol{U}$中每个通道全部空间位置的像素值乘以对应的通道权重,$\boldsymbol{U}'$为被赋予通道权重后整个模块的输出特征。EDCA是即插即用的模块,使用起来非常方便。此外,由于EDCA整个模块的参数量仅来源于两个一维卷积层,因此将该模块添加到深度神经网络中,增加的参数量可以忽略。

1.4 整体结构

基于渐进式学习网络、自注意力区域建议网络和高效双通道注意力机制的设计方案,提出了一种基于注意力机制和多尺度集成学习策略的细粒度图像识别方法AMEN。整个网络可进行端到端的训练,其结构如图5所示。主干网络使用基于ResNet-50的渐进式学习网络(PLN),图5中的CNN即为ResNet-50网络,去除其末端的全连接层和最后一层池化层,仅用作渐进式学习网络的特征提取器。将高效双通道注意力机制(EDCA)插入在了主干网络的特征提取部分,具体而言插入在ResNet-50第3个卷积块的最后一个Bottleneck和第4个卷积块的最后一个Bottleneck中,插入EDCA模块的Bottleneck结构如图6所示。

对于输入图片,先送入到主干网络,经特征提取和全局最大池化得到对应不同尺度的特征描述向量$\{f^{(1)}, f^{(2)}, f^{(3)}, f^{(4)}\}$,维度分别为512、1 024、2 048和3 584。与此同时,将主干网络的顶层特征,也就

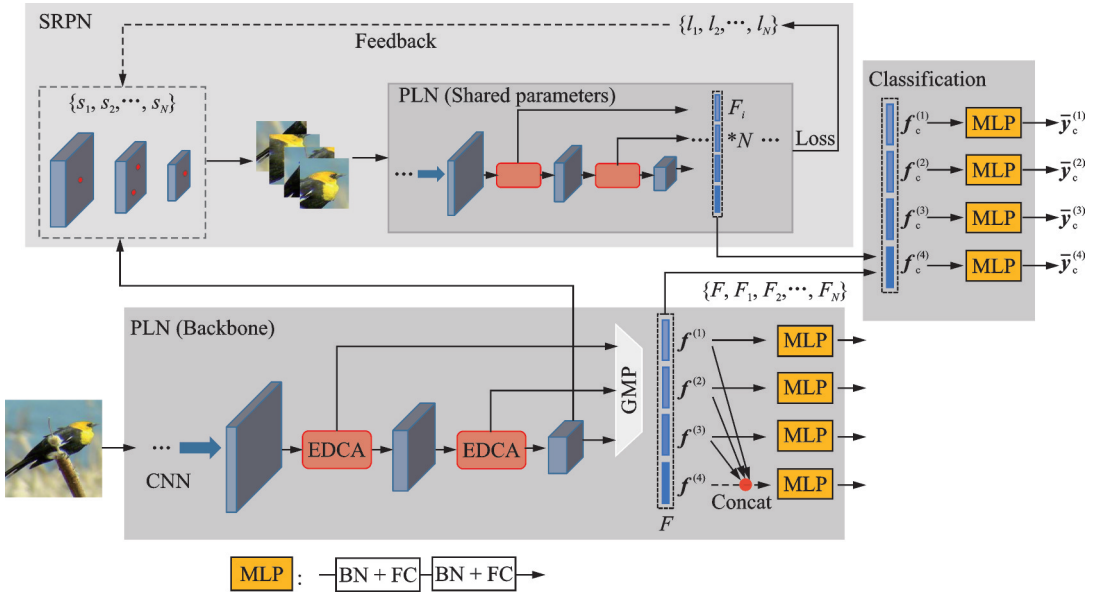

图5 基于注意力机制和多尺度集成学习策略的细粒度图像识别方法整体结构示意图

Fig.5 Diagram of overall structure of AMEN for fine-grained image recognition

是ResNet-50提取的顶层特征送入自注意力区域建议网络(SRPN)。SRPN筛选出最具有判别性的N个局部区域$\{R_1, R_2, \cdots, R_N\}$,将其调整至统一的尺寸送入共享参数的主干网络,也就是PLN,此处PLN的特征提取器和分类器的参数均共享给SRPN。对于候选区域R_i,送入主干网络后,经特征提取和全局最大池化同样能够获得一组特征描述向量$\{f_i^{(1)}, f_i^{(2)}, f_i^{(3)}, f_i^{(4)}\}$,维度也分别是512、1 024、2 048和3 584,这里$i=1, 2, \cdots, N$。

如图5所示,记完整图像输入主干网络得到的特征集合为$F=\{f^{(1)}, f^{(2)}, f^{(3)}, f^{(4)}\}$,判别性区域$R_i$输入主干网络得到的特征集合为$F_i=\{f_i^{(1)}, f_i^{(2)}, f_i^{(3)}, f_i^{(4)}\}$,那么得到的完整图像和局部区域的全部特征描述向量的集合为$\{F, F_1, F_2, \cdots, F_N\}$。最后,将上述特征集合输入到模型的最后一个部分——分类网络(Classification network, CN)进行图像类别的预测,分类网络结构如图7所示。

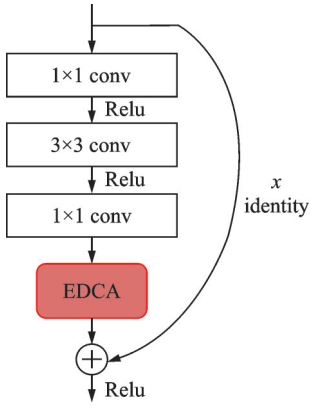

图6 包含EDCA模块的Bottleneck结构示意图

Fig.6 Diagram of the structure of Bottleneck including EDCA module

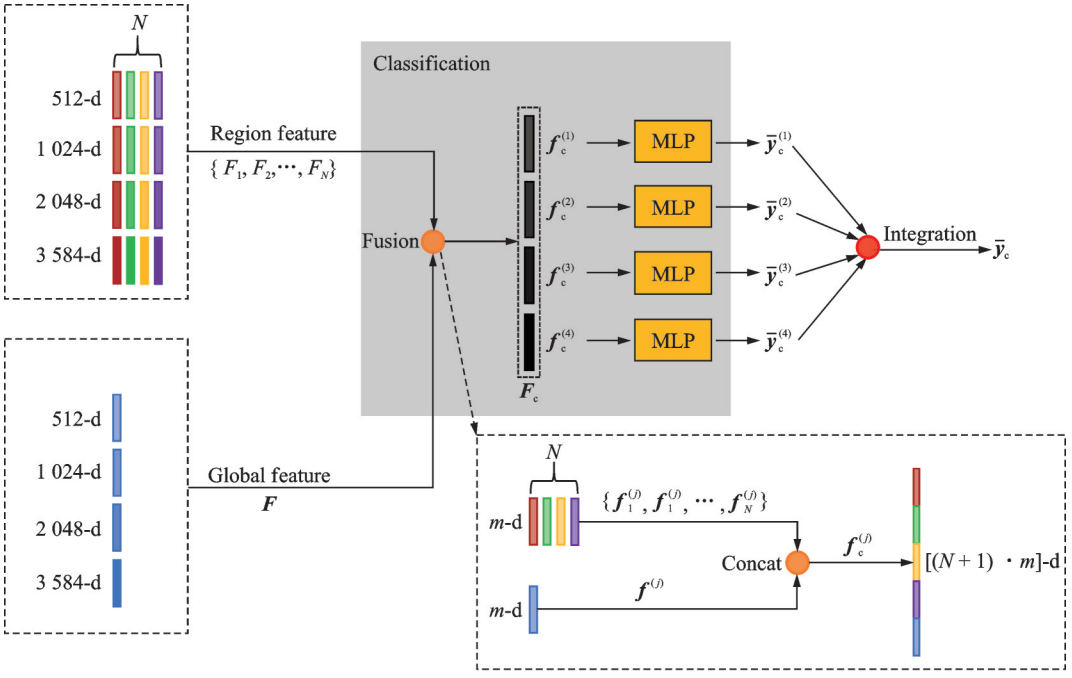

图7 分类网络结构示意图

Fig.7 Diagram of structure of classification network

分类网络首先将完整图像与判别性区域的特征使用级联(Concat)的方式进行对应尺度的融合,得到全局-区域联合特征$F_c=\{f_c^{(1)}, f_c^{(2)}, f_c^{(3)}, f_c^{(4)}\}$,其中

$$f_c^{(j)}=[f_1^{(j)}, f_2^{(j)}, \cdots, f_N^{(j)}, f^{(j)}]\in \mathbf{R}^{(N+1)\cdot D(f^{(j)})} \tag{15}$$

式中:$D(f^{(j)})$为特征描述向量$f^{(j)}$的维度。之后,分类网络基于全局-区域联合特征F_c构建新的一组多尺度基分类器,其中包含的4个基分类器与$\{f_c^{(1)}, f_c^{(2)}, f_c^{(3)}, f_c^{(4)}\}$相对应。最后,将$f_c^{(j)}$输入对应的分类

器,得到一组预测结果$\{\bar{\boldsymbol{y}}_{\mathrm{c}}^{(1)},\bar{\boldsymbol{y}}_{\mathrm{c}}^{(2)},\bar{\boldsymbol{y}}_{\mathrm{c}}^{(3)},\bar{\boldsymbol{y}}_{\mathrm{c}}^{(4)}\}$,并使用软投票的集成策略汇聚多个分类器的结果。综上所述,分类网络最终输出的预测结果为

$$\bar{\boldsymbol{y}}_{\mathrm{c}}=\sum_{j=1}^{4}\bar{\boldsymbol{y}}_{\mathrm{c}}^{(j)}\in\mathbf{R}^{K} \tag{16}$$

式中:K表示类别总数。选取$\bar{\boldsymbol{y}}_{\mathrm{c}}$中预测值最大的类别作为模型的判定类别。值得注意的是,本文自注意力区域建议网络的循环反馈机制仅在训练阶段被激活。

1.5 损失函数

模型的总体损失函数包括3个部分,分别是渐进式学习网络损失L_{pln}、自注意力区域建议网络损失L_{srpn}和分类网络损失L_{cn}。对于渐进式学习网络,即本文的主干网络,其损失函数L_{pln}由式(5)定义。对于自注意力区域建议网络,考虑候选区域R_i送入共享参数的主干网络后得到的特征描述向量$\{\boldsymbol{f}_i^{(1)},\boldsymbol{f}_i^{(2)},\boldsymbol{f}_i^{(3)},\boldsymbol{f}_i^{(4)}\}$,将其继续送入主干网络的分类器中,能够得到一组预测结果$\{\bar{\boldsymbol{y}}_i^{(1)},\bar{\boldsymbol{y}}_i^{(2)},\bar{\boldsymbol{y}}_i^{(3)},\bar{\boldsymbol{y}}_i^{(4)}\}$,从而整个SRPN网络的分类损失可表示为

$$L_{\mathrm{srpn}}^{\mathrm{cls}}=\sum_{i=1}^{N}\sum_{j=1}^{4}L_{\mathrm{sce}}(\bar{\boldsymbol{y}}_i^{(j)},\boldsymbol{y},\alpha^{(j)}) \tag{17}$$

式中:$L_{\mathrm{sce}}(\cdot)$表示平滑交叉熵损失函数,由式(4)定义,N为候选区域的总数,$\alpha^{(j)}$为平滑因子,其设置与L_{pln}中相同。此外,为了训练的稳定性,对于候选区域R_i,仅使用$L_{\mathrm{sce}}(\bar{\boldsymbol{y}}_i^{(4)},\boldsymbol{y},\alpha^{(4)})$作为该区域的分类损失进行循环反馈。此时,SRPN网络的排序损失函数可表示为

$$L^{\mathrm{rank}}=\sum_{i_1=1}^{N}\sum_{i_2=1}^{N}\max(0,L_{\mathrm{sce}}(\bar{\boldsymbol{y}}_{i_1}^{(4)},\boldsymbol{y},\alpha^{(4)})-L_{\mathrm{sce}}(\bar{\boldsymbol{y}}_{i_2}^{(4)},\boldsymbol{y},\alpha^{(4)})+\delta)*k_{i_1i_2} \tag{18}$$

式中:δ为超参数,本文设置为1,$k_{i_1i_2}$由式(7)定义。SRPN网络的总体损失函数L_{srpn}为分类损失和排序损失之和,由式(9)定义。

对于分类网络,考虑将全局-区域联合特征F_{c}输入分类器中,此处的分类器与PLN中的结构相同但参数相互独立,可得到一组预测结果$\{\bar{\boldsymbol{y}}_{\mathrm{c}}^{(1)},\bar{\boldsymbol{y}}_{\mathrm{c}}^{(2)},\bar{\boldsymbol{y}}_{\mathrm{c}}^{(3)},\bar{\boldsymbol{y}}_{\mathrm{c}}^{(4)}\}$,从而分类网络的损失函数可表示为

$$L_{\mathrm{cn}}=\sum_{j=1}^{4}L_{\mathrm{sce}}(\bar{\boldsymbol{y}}_{\mathrm{c}}^{(j)},\boldsymbol{y},\alpha^{(j)}) \tag{19}$$

式中:$\alpha^{(j)}$为平滑因子,其设置与L_{pln}中相同。综上所述,本文模型的总体损失函数可表示为

$$L=\lambda\cdot L_{\mathrm{pln}}+\mu\cdot L_{\mathrm{srpn}}+\eta\cdot L_{\mathrm{cn}} \tag{20}$$

式中:λ、μ和η为超参数,本文均设置为1。

2 实验结果与分析

2.1 实验环境与参数设置

本文基于CUB-200-2011[17]、FGVC Aircraft[18]和Stanford Car[19]3个公开数据集进行实验,验证所提出的基于注意力机制和多尺度集成学习策略的细粒度图像识别方法AMEN的有效性。上述细粒度图像识别数据集的基本信息如表1所示,其中包括:数据集的名称(Dataset name)、所属的粗粒度类别(Object)、总类别数(Catego-

表1 实验数据集基本信息

Table 1 Basic information of experimental datasets

Dataset name	Object	Categories	Train	Test
CUB-200-2011[17]	Bird	200	5 994	5 794
FGVC Aircraft[18]	Aircraft	100	6 667	3 333
Stanford Car[19]	Car	196	8 144	8 041

ries)、训练集的图片样本数目(Train)以及测试集的图片样本数目(Test)。本文所有实验均基于Pytorch深度学习框架并在服务器端完成,所使用的GPU型号为NVIDIA Tesla V100。

采用细粒度图像识别领域常用的随机翻折和随机剪裁的基础数据增强手段对训练集进行处理。在模型训练阶段,使用ImageNet大规模数据集[20]上的预训练权重对用于特征提取的基础网络ResNet-50进行权重参数的初始化。将模型训练的初始学习率设置为0.002,特别地,将主干网络特征提取部分的学习率设置为其他层的1/10,这其中包括了基础网络ResNet-50和插入的高效双通道注意力机制(EDCA),上述设置是为了使训练更加稳定。采取余弦退火策略对学习率进行调整,表达式如下

$$l_r=\frac{1}{2}l_{r_0}\cdot\left[\cos\left(\pi\cdot\frac{t}{\text{epoch}}\right)+1\right] \tag{21}$$

式中:t表示当前训练轮次,epoch表示总训练轮次,l_{r_0}表示初始学习率。提出的基于注意力机制和多尺度集成学习策略的细粒度图像识别方法AMEN的详细训练参数如表2所示。

使用测试集上的总体分类准确率(Overall accuracy, OA)作为评价指标,其计算公式如下

$$\text{acc}(D)=\frac{1}{N_0}\sum_{i=1}^{N_0}\mathbb{I}(y_i=c_i) \tag{22}$$

式中:D为数据集,N_0为样本总数,y_i为模型的预测类别,c_i为样本对应的类别标签,$\mathbb{I}(\cdot)$为指示函数,其值根据输入的真假而定,当输入为真时,函数值为1;反之,函数值则为0。下文将总体分类准确率简称为准确率(Accuracy)。

表2 训练参数设置

Table 2 Training parameter settings

参数名称	参数值	参数含义
Batch Size	16	批量大小
LR	0.002	初始学习率
Optimizer	SGD	优化器
Weight decay	5×10^{-4}	权值衰减系数
Momentum	0.9	动量
Epoch	300	总训练轮次
Image size	448×448	完整图像输入尺寸
Part size	224×224	局部区域输入尺寸
$\{\alpha^{(j)}\}$	{0.7, 0.8, 0.9, 1.0}	平滑因子
N	4	候选区域的数目

2.2 渐进式学习网络对模型性能的影响

本节首先评估了渐进式学习网络(PLN)中多尺度基分类器的性能表现,以及采取集成策略汇聚多个分类器预测结果后的性能表现,并与基础网络ResNet-50在同等实验条件下进行对比。表3展示了在CUB-200-2011、FGVC Aircraft和Stanford Cars三个数据集上的实验结果,其中,FC表示分类全连接层,其输入特征为ResNet-50的顶层特征;$\text{MLP}^{(i)}$表示渐进式学习网络中第i个基分类器,是一种多层感知机的结构,其对应的输入特征为$\boldsymbol{f}^{(i)}$;CUB、AIR和CAR分别为3个数据集的简称。从表3可看出PLN中多尺度基分类器的性能是逐级提升的,其中$\text{MLP}^{(2)}$相比$\text{MLP}^{(1)}$的提升最为明显,这是由于低层特征$\boldsymbol{f}^{(1)}$虽然包含较多的位置和细节信息,但其噪声多、语义性低,从而导致$\text{MLP}^{(1)}$分类性能较差。此外,$\text{MLP}^{(4)}$的性能是最好的,这是由于其输入的特征$\boldsymbol{f}^{(4)}$是由$\boldsymbol{f}^{(1)}$、$\boldsymbol{f}^{(2)}$、$\boldsymbol{f}^{(3)}$依序

表3 渐进式学习网络对模型性能的影响

Table 3 Impact of PLN on model performance

模型	基础网络	分类器	Accuracy/%		
			CUB	AIR	CAR
ResNet-50	ResNet-50	FC	84.4	89.1	91.1
渐进式学习网络(PLN)	ResNet-50	$\text{MLP}^{(1)}$	72.5	83.1	82.3
		$\text{MLP}^{(2)}$	84.4	90.2	92.3
		$\text{MLP}^{(3)}$	86.7	91.3	93.1
		$\text{MLP}^{(4)}$	87.7	92.3	93.8
		集成模型	**88.3**	**92.3**	**94.0**

级联(Concat)得到的,同时包含了低、中和高层特征的信息。值得关注的是,集成模型汇聚了多个分类器的预测结果,其识别性能达到最优,相较于基础ResNet-50网络,其在3个数据集上的识别准确率分别提升了3.9%、3.2%和2.9%,性能提升极为显著,验证了本文所提出的渐进式学习网络的有效性。

其次,本节对基于标签平滑方法的渐进式训练(PT)策略进行了评估,与仅使用one-hot编码的标签对多尺度基分类器进行训练的方式进行对比。表4展示了在CUB-200-2011数据集上的实验结果,其中,"No PT"表示仅使用one-hot编码的标签进行训练,相当于将平滑因子$\{\alpha^{(j)}\}$设置为$\{1.0,1.0,1.0,1.0\}$,"With PT"表示采用渐进式训练策略;"Gain"表示使用渐进式训练策略在各个分类器上获得的性能增益。从表4可看出,在使用渐进式训练策略之后,除了在$MLP^{(1)}$上分类性能有所下降外,在其余的分类器包括集成模型上均有不同程度的性能增益,这说明渐进式训练是一种系统性的优化策略,能够提升模型整体的识别性能,从而体现在集成预测的结果上。

表4 渐进式训练策略对模型性能的影响

Table 4 Impact of PT strategy on model performance

分类器	Accuracy/%		Gain/%
	No PT	With PT	
$MLP^{(1)}$	**73.7**	72.5	−1.2
$MLP^{(2)}$	84.1	**84.4**	+0.3
$MLP^{(3)}$	86.4	**86.7**	+0.3
$MLP^{(4)}$	87.1	**87.7**	+0.6
集成模型	88.0	**88.3**	+0.3

2.3 引入注意力机制对模型性能的影响

本节评估了本文引入的注意力机制对模型性能的影响,主要包括自注意力区域建议网络(SRPN)和高效双通道注意力机制(EDCA)。表5展示了在CUB-200-2011、FGVC Aircraft和Stanford Cars三个数据集上的实验结果,其中,PLN表示渐进式学习网络,其用于特征提取的基础网络为ResNet-50;PLN+EDCA表示基于渐进式学习网络添加EDCA模块,插入在了其基础网络的特定Bottleneck中。此外,表5中SRPN模型均使用完整图像与判别性区域的级联(Concat)特征进行预测,其中主干网络为PLN或PLN+EDCA的模型使用如图7所示的分类网络的集成预测结果。从表中可看出,在主干网络使用ResNet-50的情况下,引入SRPN使得模型在3个数据集上的识别准确率分别提升了3.0%、1.5%和1.6%。将模型3与表3中仅使用PLN的结果进行对比,引入SRPN使得模型在3个数据集上的识别准确率分别提升了0.7%、0.9%和0.6%。通过上述两组数据的对比,可验证本文自注意力区域建议网络的有效性。模型4在模型3的基础上向主干网络中插入了EDCA模块,使得模型在3个数据集上的识别准确率又分别进一步提升了0.6%、0.2%和0.2%,这验证了EDCA模块设计的有效性。此外,EDCA模块是轻量化的,将其添加到已有模型中额外增加的参数量是可以忽略的。表5中性能表现最好的模型4即为本文AMEN模型。

表5 自注意力区域建议网络和高效双通道注意力机制对模型性能的影响

Table 5 Impact of SRPN and EDCA on model performance

模型	主干网络	区域建议	Accuracy/%		
			CUB	AIR	CAR
1	ResNet-50		84.4	89.1	91.1
2	ResNet-50	SRPN	87.4	90.6	92.7
3	PLN	SRPN	89.0	93.2	94.6
4	PLN + EDCA	SRPN	**89.6**	**93.4**	**94.8**

本节还分析了自注意力区域建议网络中候选区域数目 N 的选取对模型性能的影响,并基于表5中的模型4在CUB-200-2011数据集上进行了实验,表6展示了 N 的不同取值对模型性能的影响。从表中可看出,当候选区域数目从2增加到4时,模型性能逐步提升,说明定位更多的判别性区域可帮助模型更好地学习细粒度特征。当 N 的取值进一步增加时,模型识别的准确率并没有继续提升,这是因为排序靠后的区域所包含的判别性信息有限,无法帮助模型更好识别目标。

表7列出了本文AMEN模型与现有弱监督细粒度图像识别模型在CUB-200-2011、FGVC Aircraft和Stanford Cars三个公开数据集上的识别准确率。为了公平起见，只选取了基础网络为VGGNet[21]和ResNet-50[14]的方法。在表7列举的方法中，RA-CNN和MGE-CNN都是对细粒度图像中的注意力区域进行逐级放大，需要将每一级图像输入到参数独立的主干网络中，因此模型中包含了多个主干网络。相较而言，本文自注意力区域建议网络使用共享参数的主干网络，降低了模型的参数量，并且在CUB-200-2011和Stanford Cars数据集上的识别准确率比MGE-CNN分别高出1.1%和0.9%。HBP提出了一种分层双线性池化方法来捕获层间特征关系，并在此基础上集成多个跨层双线性特征以提高其表示能力。NTS-Net采用导航-教师-审查网络将细粒度特征学习和判别性区域定位进行了结合，其将注意力区域输入主干网络得到的置信度反馈给区域生成模块，本文模型与其不同的是使用平滑交叉熵损失进行反馈，并且引入了渐进式学习网络和ED-CA模块，在3个数据集上的识别准确率比NTS-Net分别提升了2.1%、2.0%和0.9%。DFL-CNN通过学习一组卷积滤波器，从而捕获类特定的判别性补丁。DCL提出了一种新的“破坏和构造学习”方法，仔细地“破坏”然后“重建”输入图像以学习判别区域和特征，在模型的识别性能上获得较大提升。LIO提出了一个对象范围学习模块用于定位目标，然后设计了一个空间上下文学习模块对目标的内部结构进行建模。FDL提出了一种“过滤和蒸馏学习”模型，以增强模型对局部区域的判别能力。SPS提出了一种新的随机部分交换策略来增强中层模型的泛化能力。综合来看，本文AMEN模型在CUB-200-2011数据集上与性能表现最好的SPS相比，识别准确率提升0.9%；在FGVC Aircraft数据集上与性能表现最好的FDL相比，识别准确率持平；而在Stanford Cars数据集上，识别准确率仅比性能表现最好的SPS低0.1%。上述的数据对比验证了本文AMEN算法设计的有效性，并且说明本文AMEN模型的细粒度图像识别性能处于行业先进水平。

表6 候选区域数目的选取对模型性能的影响

Table 6 Impact of the number of candidate regions on model performance

候选区域数目 N	Accuracy/%
2	88.4
3	89.4
4	**89.6**
5	89.5
6	89.5

表7 本文AMEN模型与现有弱监督细粒度图像识别方法的性能对比

Table 7 Performance comparison between the proposed AMEN and existing weakly-supervised fine-grained image recognition methods

模型	基础网络	Accuracy/%		
		CUB	AIR	CAR
RA-CNN[7]	VGG-19	85.3	88.2	92.5
HBP[22]	VGG-16	87.1	90.3	93.7
NTS-Net[23]	ResNet-50	87.5	91.4	93.9
DFL-CNN[24]	ResNet-50	87.4	91.7	93.1
TASN[8]	ResNet-50	87.9		93.8
MGE-CNN[25]	ResNet-50	88.5		93.9
DCL[26]	ResNet-50	87.8	93.0	94.5
BNT[9]	ResNet-50	88.1	92.4	94.6
LIO[27]	ResNet-50	88.0	92.7	94.5
FDL[28]	ResNet-50	88.6	**93.4**	94.3
SPS[29]	ResNet-50	88.7	92.7	**94.9**
AMEN(Ours)	ResNet-50	**89.6**	**93.4**	94.8

2.4 可视化结果

图8展示了AMEN模型检测到的判别性区域，图中依次使用橙、蓝、红、绿4种颜色的矩形框来表示得分递减的候选区域。为了方便观看，图中前3行仅显示得分最高的两个候选区域，在第4行中显示全

部的4个候选区域。从图中可看出,尽管是在弱监督的条件下,AMEN模型检测到的判别性区域仍然较为准确。对于鸟类而言,得分最高的前两个区域往往会关注鸟的头部和躯体;对于汽车而言,前两个区域则偏向于关注车的前脸和车身;对于飞机而言,前两个区域则会关注机身的前部和中后部。通过观察第4行可看出,AMEN模型检测的判别性区域主要分布在飞机的机头、机身、机翼、机尾等位置,鸟儿的头部、躯干、翅膀等位置以及汽车的前脸、车身、车灯等位置,与人类的视觉感知相一致。

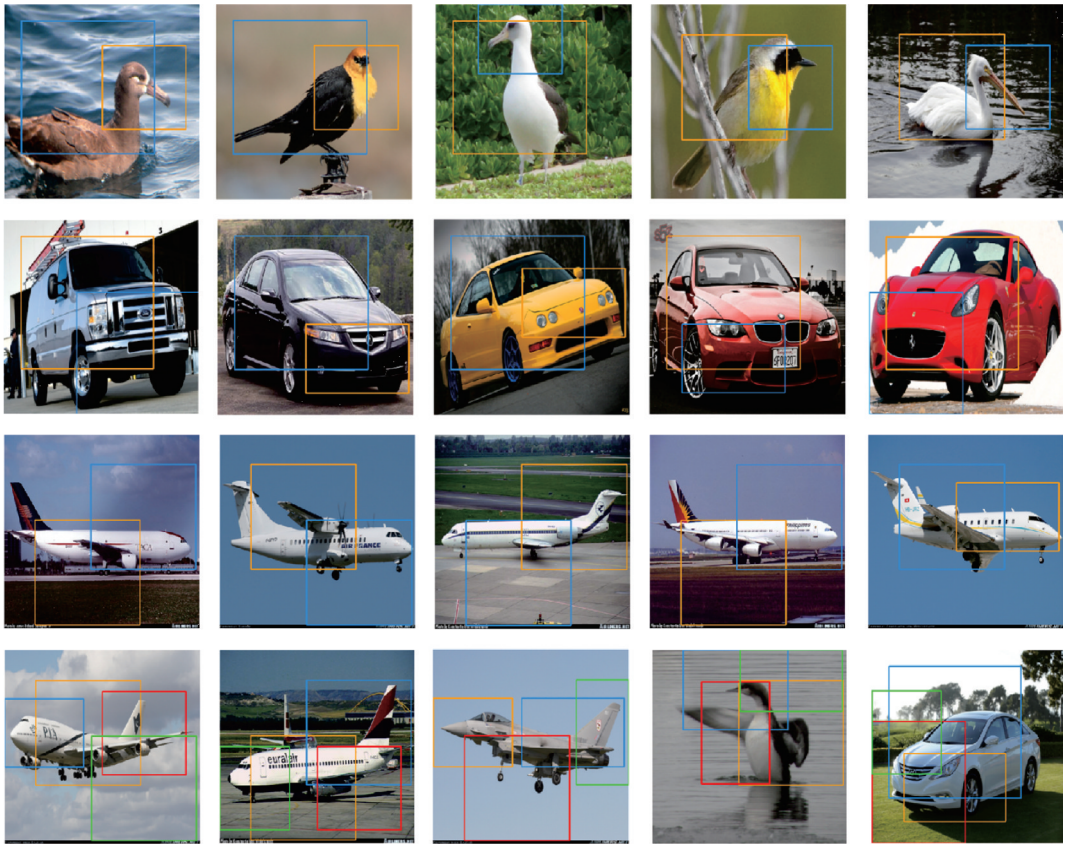

图8 细粒度图像识别模型AMEN检测的判别性区域

Fig.8 Discriminant regions detected by fine-grained image recognition model of AMEN

除此之外,还将Grad-CAM[30]应用于模型主干网络的最后一个卷积层,以生成图9所示的热力图,图中红色越深的区域表示其对模型预测结果的贡献越大,即模型对这部分区域的关注度越高。AMEN(no EDCA)表示没有在主干网络中插入EDCA模块,即为表5中的模型3。从图9可看出,ResNet-50的关注区域相对单一,主要集中在鸟的头部、车的前脸以及飞机机身的颈部。而本文通过引入渐进式学习网络和自注意力区域建议网络将多尺度细粒度特征学习和判别性区域定位有效结合,与ResNet-50相比热力图(第3列)上的关注区域更加多样,在鸟的翅膀、车身的侧面以及飞机的机尾处都额外出现了较为强烈的响应。添加EDCA模块之后,热力图(第4列)上的关注区域仍然保持了多样化,与添加模块之前的的差异在于削弱或增强了部分区域的响应。具体而言,削弱了冗余区域的响应,表现在第1行背景区域以及第2行背景、车前窗区域;增强了判别区域的的响应,表现在第3行机尾和机翼位置处。综上所述,本文AMEN模型能够从不同区域中提取有价值的信息,并削弱背景等冗余区域的干扰。

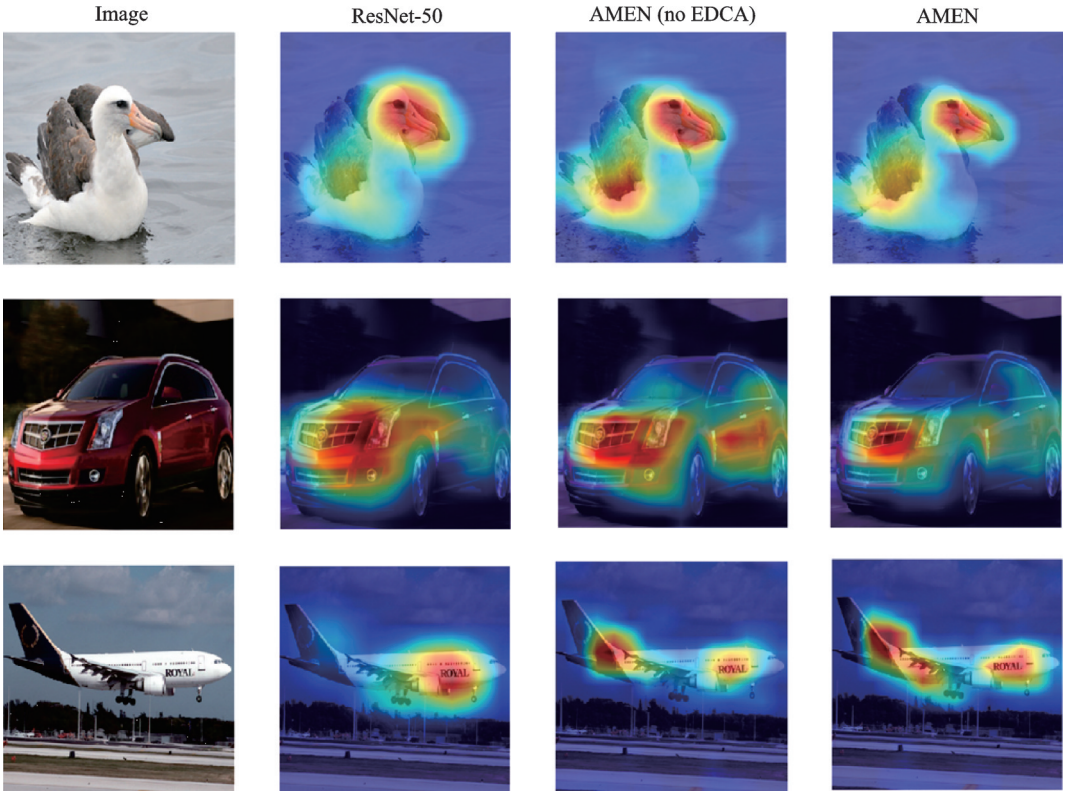

图9 部分样本的热力图

Fig.9 Heat maps of partial samples

3 结束语

现有细粒度图像识别研究中存在图像细粒度特征利用不充分以及判别性区域难以挖掘的问题，因此提出了一种基于注意力机制和多尺度集成学习策略的细粒度图像识别方法AMEN。针对图像细粒度特征利用不充分的问题，引入渐进式学习网络，利用集成学习的策略，基于深度神经网络3个层级的输出特征构建多尺度基分类器，并使用标签平滑的方法根据不同尺度基分类器的学习能力设置相应的软标签进行渐进式训练，从而大幅度提高低层特征的利用率；同时采用高效双通道注意力机制对特征施加通道权重，使得网络能够在通道层面自主筛选特征，从而提升高信息相关度通道的利用率。针对弱监督条件下判别性区域难以挖掘的问题，引入自注意力区域建议网络，通过构建循环反馈机制促使模型逐步定位到更加具有判别性的区域，并在最后的分类模块中将完整图像与判别性区域的特征信息进行融合。实验结果表明，AMEN在CUB-200-2011、FGVC Aircraft和Stanford Cars三个公开数据集上达到了先进的细粒度图像识别水平。本文模型的局限性在于，直接利用3个层级的输出特征构建分类器，而没有充分考虑不同层级输出特征间存在的相关性，容易导致存在于层级特征关系中有价值信息的丢失。因此，下一步的研究方向是需对层级间特征的相关性进行建模，促进跨层次特征信息的深度融合。

参考文献：

[1] ZHAO B, FENG J, WU X, et al. A survey on deep learning-based fine-grained object classification and semantic segmentation

[J]. International Journal of Automation and Computing, 2017, 14(2): 119-135.

[2] ZHANG N, DONAHUE J, GIRSHICK R, et al. Part-based R-CNNs for fine-grained category detection[C]//Proceedings of European Conference on Computer Vision (ECCV). [S.l.]: ECCV, 2014. 834-849.

[3] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2014: 580-587.

[4] BRANSON S, VAN HORN G, BELONGIE S, et al. Bird species categorization using pose normalized deep convolutional nets[EB/OL]. (2014-06-11). https://doi.org/10.48550/arxiv.1406.2952.

[5] 卢宏涛,罗沐昆.基于深度学习的计算机视觉研究新进展[J].数据采集与处理,2022, 37(2): 247-278.
LU Hongtao, LUO Mukun. Survey on new progresses of deep learning based computer vision[J]. Journal of Data Acquistion and Processing, 2022, 37(1): 247-278.

[6] LIN T Y, ROYCHOWDHURY A, MAJI S. Bilinear convolutional neural networks for fine-grained visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 40(6): 1309-1322.

[7] FU J, ZHENG H, MEI T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2017: 4438-4446.

[8] ZHENG H, FU J, ZHA Z J, et al. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2019: 5012-5021.

[9] JI R, WEN L, ZHANG L, et al. Attention convolutional binary neural tree for fine-grained visual categorization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2020: 10468-10477.

[10] DING Y, MA Z, WEN S, et al. AP-CNN: Weakly supervised attention pyramid convolutional neural network for fine-grained visual classification[J]. IEEE Transactions on Image Processing, 2021, 30: 2826-2836.

[11] SHU Y, YU B, XU H, et al. Improving fine-grained visual recognition in low data regimes via self-boosting attention mechanism[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2022: 449-465.

[12] GALAR M, FERNANDEZ A, BARRENECHEA E, et al. A review on ensembles for the class imbalance problem: Bagging- , boosting- , and hybrid-based approaches[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2011, 42(4): 463-484.

[13] MÜLLER R, KORNBLITH S, HINTON G E. When does label smoothing help?[C]//Proceedings of the 33rd International Conference on in Neural Information Processing Systems. [S.l.]: ACM, 2019: 4693-4703.

[14] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2016: 770-778.

[15] ROSENFELD A, THURSTON M, LEE Y H. Edge and curve detection for visual scene analysis[J]. IEEE Transactions on Computers, 1972, 20(5): 562-569.

[16] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018: 7132-7141.

[17] WAH C, BRANSON S, WELINDER P, et al. The caltech-UCSD birds200-2011 dataset[R]. [S.l.]: California Institute of Technology, 2011.

[18] MAJI S, RAHTU E, KANNALA J, et al. Fine-grained visual classification of aircraft[EB/OL]. (2013-06-21). https://doi.org/10.48550/arxiv.1306.5151.

[19] KRAUSE J, STARK M, DENG J, et al. 3d object representations for fine-grained categorization[C]//Proceedings of the IEEE International Conference on Computer Vision Workshops. [S.l.]: IEEE, 2013: 554-561.

[20] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database[C]//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2009: 248-255.

[21] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-12). https://doi.org/10.48550/arxiv.1409.1556.

[22] YU C, ZHAO X, ZHENG Q, et al. Hierarchical bilinear pooling for fine-grained visual recognition[C]//Proceedings of the European Conference on Computer Vision (ECCV). [S.l.]: ECCV, 2018: 574-589.

[23] YANG Z, LUO T, WANG D, et al. Learning to navigate for fine-grained classification[C]//Proceedings of the European Conference on Computer Vision (ECCV). [S.l.]: ECCV, 2018: 420-435.

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

[25] ZHANG L, HUANG S, LIU W, et al. Learning a mixture of granularity-specific experts for fine-grained categorization[C]//Proceedings of the IEEE/CVF international Conference on Computer Vision. [S.l.]: IEEE, 2019: 8331-8340.

[26] CHEN Y, BAI Y, ZHANG W, et al. Destruction and construction learning for fine-grained image recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2019: 5157-5166.

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

[28] LIU C, XIE H, ZHA Z J, et al. Filtration and distillation: Enhancing region attention for fine-grained visual categorization[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 11555-11562.

[29] HUANG S, WANG X, TAO D. Stochastic partial swap: Enhanced model generalization and interpretability for fine-grained recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE, 2021: 620-629.

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

作者简介:

季晟宇(1998-),男,硕士研究生,研究方向:深度学习、图像识别、目标检测,E-mail:220210936@seu.edu.cn。

江志康(1999-),男,硕士研究生,研究方向:深度学习、图像识别、小样本学习。

马翔(1998-),男,硕士研究生,研究方向:深度学习、图像异常检测。

杨绿溪(1964-),**通信作者**,男,教授,博士生导师,研究方向:通信信号处理、深度学习,E-mail:lxyang@seu.edu.cn。

(编辑:夏道家)