计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(3):126−133 [doi: 10.15888/j.cnki.csa.009419]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

跨层协同注意和通道分组注意的细粒度图像分类①

何志祥, 齐　琦, 何　伟, 郭龙源

(湖南理工学院 信息科学与工程学院, 岳阳 414006)
通信作者: 齐　琦, E-mail: qiqi@hnist.edu.cn

摘　要: 细粒度图像分类的主要挑战在于类间的高度相似性和类内的差异性. 现有的研究多数基于深层的特征而忽略了浅层细节信息, 然而深层的语义特征由于多次卷积和池化操作往往会丢失大量的细节信息. 为了更好地整合浅层和深层的信息, 提出了基于跨层协同注意和通道分组注意的细粒度图像分类方法. 首先, 通过 ResNet50 加载预训练模型作为骨干网络提取特征, 由最后 3 个阶段提取的特征以 3 个分支的形式输出, 每一个分支的特征通过跨层的方式与其余两个分支的特征计算协同注意并交互融合, 其中最后一个阶段的特征经过通道分组注意模块以增强语义特征的学习能力. 模型训练可以高效地以端到端的方式在没有边界框和注释的情况下进行训练, 实验结果表明, 该算法在 3 个常用细粒度图像数据集 CUB-200-2011、Stanford Cars 和 FGVC-Aircraft 上的准确率分别达到了 89.5%、94.8% 和 94.7%.

关键词: 细粒度图像分类; 注意力机制; 深度学习; 特征提取; 语义特征

引用格式: 何志祥,齐琦,何伟,郭龙源.跨层协同注意和通道分组注意的细粒度图像分类.计算机系统应用,2024,33(3):126–133. http://www.c-s-a.org.cn/1003-3254/9419.html

Cross Layer Collaborative Attention and Channel Group Attention for Fine-grained Image Classification

HE Zhi-Xiang, QI Qi, HE Wei, GUO Long-Yuan

(School of Information Science and Engineering, Hunan Institute of Science and Technology, Yueyang 414006, China)

Abstract: The main challenge of fine-grained image classification lies in the high similarity between classes and differences within classes. Most of the existing research is based on deep features and ignores shallow details. However, deep semantic features often lose a lot of details due to multiple convolution and pooling operations. To better integrate shallow and deep information, this study proposes a fine-grained image classification method based on cross-layer collaborative attention and channel grouping attention. First, the pre-trained model loaded by ResNet50 is taken as the backbone network to extract features, and the features extracted by the last three stages are output in the form of three branches. The features of each branch are calculated and coordinated with the features of the other two branches in a cross-layer manner and interactive fusion. Specifically, the features of the last stage pass through the channel grouping attention module to enhance the learning ability of semantic features. Model training can be efficiently trained in an end-to-end manner without bounding boxes and annotations. Experimental results show that the algorithm performs well on three common fine-grained image datasets CUB-200-2011, Stanford Cars, and FGVC-Aircraft. The accuracy rates reach 89.5%, 94.8%, and 94.7%, respectively.

Key words: fine-grained image classification; attention mechanism; deep learning (DL); feature extraction; semantic features

① 基金项目: 国家自然科学基金 (62271200); 湖南省高校创新平台开放基金项目 (20K062); 湖南省教育厅优秀青年项目 (21B0590)
收稿时间: 2023-09-12; 修改时间: 2023-10-09; 采用时间: 2023-10-16; csa 在线出版时间: 2024-01-09
CNKI 网络首发时间: 2024-01-10

1 引言

细粒度图像分类 (fine-grained visual classification, FGVC) 是计算机视觉领域内的一个任务, 其目的是对具有相似外观的图像进行子类别划分[1], 例如区分鸟的种类、汽车的型号以及飞机的型号等. 相较于一般的分类任务, 细粒度图像分类由于细微的类间差异和较大的类内差异, 使其更具有挑战性. 如图 1 所示, 其中每一行为同一种鸟类, 但受到飞行姿态、拍摄视角等差异的影响, 即使是同一个种类也表现出了较大的类内差异, 而每一列不同的鸟类在相似的飞行姿态、背景情况下, 则展现出细微的类间差异.

图 1 细粒度图像分类任务两大挑战

由于卷积神经网络 (convolutional neural network, CNN) 的快速发展, 细粒度图像分类方法的性能也获得了飞速的提升. 目前细粒度图像分类方法主要分为两大类: 一是基于强监督的方法, 二是基于弱监督的方法. 其中基于强监督的方法最具代表性的是 Zhang 等人[2]提出的 Part R-CNN, 该方法在 Faster R-CNN 目标检测框架的基础上进行改进, Part R-CNN 通过将对象分解为局部部分并将其特征输入分类模型中以更好地捕获细粒度特征. 在细粒度图像中, 对象的姿态变化导致类内方差变大, 为了克服这一变化, Branson 等人[3]提出了姿态归一化 CNN, 该方法通过姿态归一化算法对图像进行裁剪、对齐使得对象在图像中的位置和朝向保持一致, 再通过 CNN 对姿态归一化后的图像进行特征提取, 最后通过分类器实现分类.

虽然基于强监督的方法取得了不错的效果, 但由于这类方法严重依赖额外的标注信息, 而获取这些标注信息需要耗费大量的人力物力, 从而极大地限制了这类方法的应用. 因此, 仅使用类别标签的弱监督方法逐渐成为细粒度图像分类领域的主流方法. Lin 等人[4]提出了双线性 CNN, 该方法通过将两个独立的 CNN 之间的特征图进行双线性池化, 从而捕捉到特征之间的高阶交互关系. 但是双线性池化操作导致了在训练和推理的过程中计算开销的增加. 因此, 为减少计算复杂度、提高分类效率, 一些改进的方法相继被提出, 例如紧凑型双线性池化[5]、低秩双线性池化[6]. Fu 等人[7]提出循环注意力卷积神经网络 (recurrent attention convolutional neural network, RA-CNN), 该方法通过迭代地使用注意力机制来增强模型对重要细节区域的关注, 经过一系列的迭代操作之后得到最终的特征图进行分类和预测. Zheng 等人[8]提出了多注意力卷积神经网络 (multi-attention convolutional neural network, MA-CNN), 通过学习多个局部和全局注意力区域和特征融合, MA-CNN 能够更好地捕捉图像中具有判别性的细粒度特征, 以提高模型的分类性能.

近年来, 许多研究人员提出了大量新颖的方法解决细粒度图像分类问题, 取得了一系列丰硕的成果. 但 FGVC 仍然是一个非常具有挑战性的任务, 还没有得到有效的解决. 鉴于深度神经网络强大的特征提取能力, 可以有效地捕获具有判别性的特征, 常被用来处理细粒度分类任务. 但这类方法通常采用深层特征而忽略了浅层特征, 不利于图像的细粒度分类. 因此, 本文提出采用跨层协同注意策略来融合浅层和深层特征, 通过浅层和深层特征之间的协同注意并融合不同层之间的特征来发掘具有判别力的特征. 同时, 由于深层特征具有丰富的语义信息, 本文进一步提出采用通道分组注意模块来改善模型对深层语义信息的学习能力, 从而提高模型的性能和泛化能力.

2 相关工作

2.1 细粒度特征学习

细粒度图像分类任务的首要任务是提取具有判别性的细粒度特征. Chen 等人[9]提出了破坏与重建学习 (destruction and construction learning, DCL) 来增强细粒度图像识别任务的性能. 核心思想是通过破坏与重建操作来强化模型对于细粒度特征之间的细微差异的感知能力. Chang 等人[10]提出了互通道损失 (mutual channel loss, MCL) 用于特征表示和类别区分能力. 首先通过 CNN 来提取图像特征, 通过计算特征图之间不同通道间的交互信息来表示通道间的相关性. 实验证明, MCL 通过增强深度特征之间的交互和通道之间的关联, 可以有效地捕获细粒度类别之间的细微差异. Huang 等人[11]提出了随机部分交换 (stochastic partial swap, SPS) 的方法来改善模型的泛化能力, 使用 SPS

交换图像中的局部特征来生成新的训练样本, SPS 既可以增加训练数据的多样性又可以促使模型学习细粒度类别之间的细微差异. SPS 的优势有两个方面, 一是通过随机部分交换增强了训练数据的多样性, 在减少过拟合风险的同时可以兼顾对模型泛化能力的提升. 二是随机部分交换操作可以帮助模型更聚焦于图像中更具有判别性的局部特征区域, 增强了模型的可解释性.

2.2 注意力机制

注意力机制用于改善传统方法中模型均匀对待输入信息而导致的对输入数据可能具有不同的相关性和重要性的忽略. 挤压和激励网络 (squeeze and excitation network, SENet)[12]的主要思想是通过学习通道之间的相关性来调整每个通道的权重. 挤压操作通过全局平均池化 (global average pooling, GAP) 将特征降维压缩成一个通道维度的特征向量以捕获特征图中每个通道的全局信息. 激励操作通过学习生成用于表示每个通道重要性的权重, 并调整特征图中每个通道的相应值. 挤压和激励操作使得模型更加关注于有利于细粒度分类的通道信息. Zhang 等人[13]提出了渐进式协同注意网络, 其中的协同注意模块通过比较输入图像对之间的模型所共同关注的区域, 注意力擦除模块则通过删除协同注意模块所建议的区域来以此关注图像的其他区域来发现潜在的细粒度特征.

3 本文方法

3.1 模型整体架构

本文提出的模型整体架构如图 2 所示. 骨干网络基于 ResNet50, 包含了 5 个主要阶段. 本文选取了 ResNet50 最后 3 个阶段的卷积层的输出进行处理, 3 个阶段所提取的特征可表示为$\{x_1, x_2, x_3\}$, $x_n \in \mathbb{R}^{H\times W\times C}$, 其中 HWC 分别表示当前阶段所提取深度特征的高度、宽度和通道数. x_n作为卷积块的输入得到F_n, 将不同层级得到的特征进行融合以计算不同层级之间的相关性, 引入协同注意机制, 使得不同层级之间的特征可以实现有效的交互. 将最后阶段提取的特征x_3输入通道分组注意模块 (channel group attention module, CGAM) 以增强语义特征学习. 通过跨层的协同注意学习可以充分利用 ResNet50 不同层级所提取的特征信息, 并在不同层级之间实现信息交互. 深层卷积所提取的特征具有丰富的语义信息, 引入通道分组注意模块可以进一步增强语义特征的学习. 有助于提取更丰富和更准确的判别性特征, 从而提高细粒度图像分类的有效性和泛化能力.

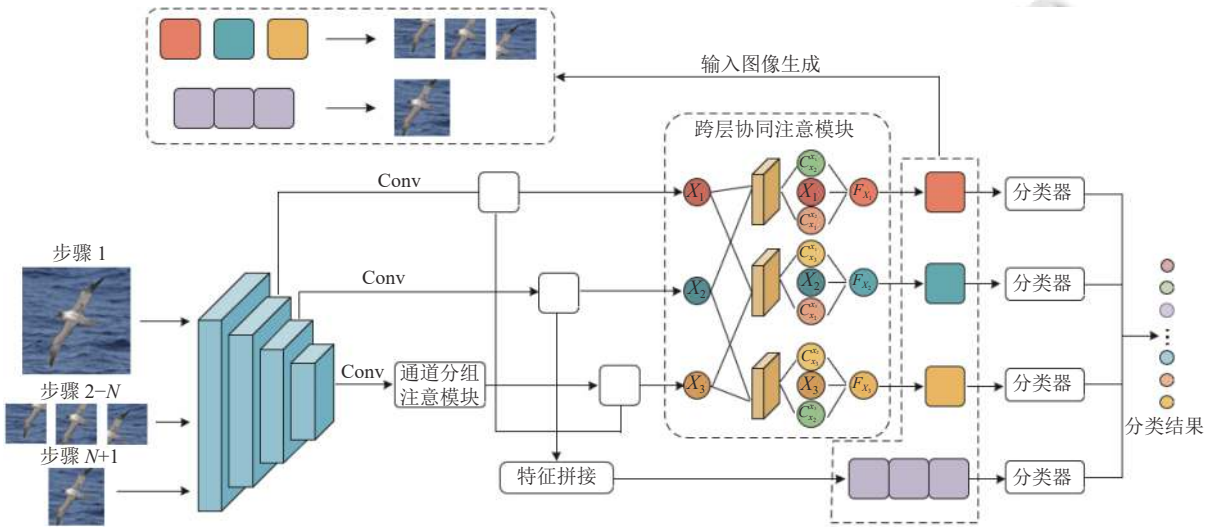


图 2 本文方法结构图

3.2 通道分组注意模块

传统的 CNN 中, 不同的卷积核负责提取不同的特征, 但这些特征之间的相关性较弱. 为了改善这一现象, 使用通道分组注意模块来增强这些特征之间的相关性, 通过对特征图中的通道进行分组并进行增强来提高特征的语义表达能力. 通道分组注意模块如图 3 所示. 具体而言, 将网络最后一个阶段输出的特征图$F \in \mathbb{R}^{H\times W\times C}$沿着通道维度均匀分成$n$组. 每组的特征可以表示为

$x=\{x_1,x_2,\cdots,x_m\},x_i\in\mathbb{R}^{\frac{C}{n}},m=H\times W$, 然后将$x$进行全局平均池化 (global average pooling, *GAP*) 和全局最大池化 (global max pooling, *GMP*) 后相加得到这一组的特征语义向量, 计算公式如下:

$$g=GAP(x)+GMP(x) \tag{1}$$

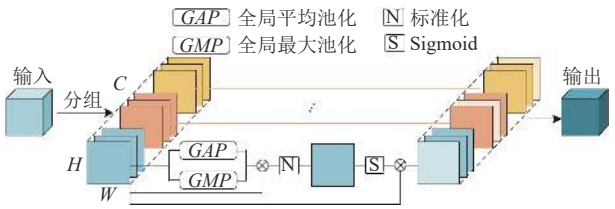


图 3 通道分组注意模块

使用全局特征g与每组特征做点积操作计算相似性, 生成初始注意力得分A. 对A进行标准化处理, 然后经过 Sigmoid 激活函数得到最终的注意力得分$\overline{A}$并对原始特征组进行点积得到最后的通道分组注意力.

3.3 跨层协同注意模块

合理利用网络不同层级所提取的深度特征可以更有效地提高 FGVC 任务的性能, 因此如何有效地整合深层网络和浅层网络所提取的特征是关键. 在一次迭代训练中, 第 1 步训练输入网络的是原始图像, 在这一步生成所有分支所生成的注意区域, 记为$\{a_1,a_2,a_3\}$. 在后续的步骤 2–N 中, 首先以跨层的方式对 3 个分支的注意区域图计算协同注意力并进行融合, 具体来说, 从a_1,a_2,a_3中选取一个作为主要特征, 其余两个特征则组成一对特征图$F_1,F_2\in\mathbb{R}^{C\times H\times W}$, 将特征图$F_1,F_2$重塑为$F_1',F_2'\in\mathbb{R}^{c\times l},l=h\times w$, 然后对$F_1'$和$F_2'^{\mathrm{T}}$进行乘积得到矩阵$M$. 然后将得到矩阵$M$输入到 Softmax 得到权重矩阵$W$, 将原始特征$F_1,F_2$与权重矩阵$W$相乘得到:

$$\begin{cases}F_a=F_1+(F_1\times W)\\F_b=F_2+(F_2\times W)\end{cases} \tag{2}$$

以a_1作为主要特征为例, a_2和a_3组成一对特征图计算得出F_a,F_b, 构建模型的后续输入图像$I_1=a_1+(F_a+F_b)$.

3.4 分类模型

本文将原始图像、跨层协同注意模块所构建的关注区域图像分别使用本文网络模型进行训练, 网络的不同分类器关注不同的区域, 输入跨层协同注意模块输出特征的分类器 1、2、3 主要关注细节特征, 输入经过 3 个卷积块提取并拼接特征的分类器 4 主要关注整体特征. 通过分类模型结合不同分支的不同层次特征的优点, 增加判别性信息的获取能力, 得到最终的分类结果. 本文使用分类器的详细结构如表 1 所示.

表 1 分类器结构详细信息

网络层	类型	输入	输出
分类器1	全连接层+BN激活函数 (ELU)	1 024	200
分类器2	全连接层+BN激活函数 (ELU)	1 024	200
分类器3	全连接层+BN激活函数 (ELU)	1 024	200
分类器4	全连接层+BN激活函数 (ELU)	1024×3	200

本文提出方法的具体步骤如下.

步骤 1. 将原始图像输入到主干网络 ResNet50 模型进行训练, 将模型最后 3 个阶段的特征图提取, 其中最后 1 个阶段的特征图先经过通道分组注意模块处理后与前两个阶段的特征进行拼接.

步骤 2. 使用跨层协同注意模块对网络步骤 1 所输出的 3 个阶段的特征进行融合, 以有效地整合深层网络和浅层网络所提取的特征, 并依据提取的特征与步骤 1 所拼接的特征生成具有不同关注区域和整体关注区域的图像.

步骤 3. 将步骤 2 中生成的图像依次输入到网络模型进行训练, 该图像由原始图像和该分支以外的分支提取出的注意力区域组成.

步骤 4. 一次性训练所有分支与整体注意力区域的图像. 整体注意力由所有分支共同提出, 包含了所有分支认为重要且有助于分类的特征, 对他们共同获取的注意力信息进行放大和研究以提取具有判别性的特征.

步骤 5. 本文方法共有 N+1 个分类器. 在推理阶段, 给定一幅输入图像, 网络模型可以产生 N+1 个预测分数. 在执行推理时, 依次将原始输入和整体注意力区域输入到网络模型得到预测分数, 对这些预测分数取平均, 计算出最终预测分数.

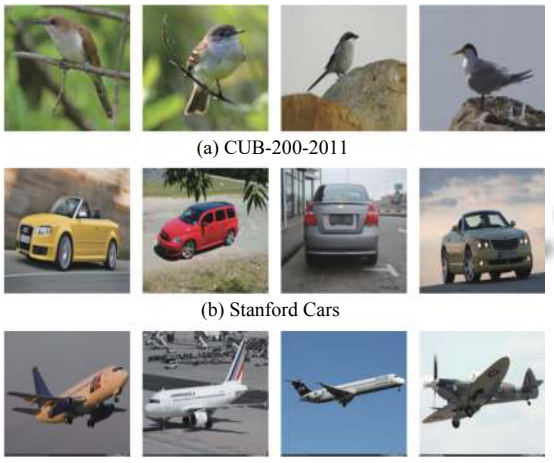
4 实验说明及分析

4.1 实验数据集

本文使用细粒度图像分类领域内常用的 3 个数据集来评估本文方法的有效性, 分别是 CUB-200-2011、Stanford Cars、FGVC-Aircraft.

CUB-200-2011 数据集是细粒度图像分类领域内最具挑战性的数据集之一, 该数据集有 200 种不同鸟类的图像共 11 788 张. 由于鸟类飞行姿态、光照条件等差异增加了识别的难度. Stanford Cars 数据集有

196 种不同品牌和型号汽车的图像共 16 185 张. 每个类别包含约 80 张不同角度和视角的汽车图像. FGVC-Aircraft 数据集有 100 种不同型号的飞机图像共 10 200 张. 每个类别包含约 100 张不同视角和姿态的飞机图像. 3 个常用细粒度图像分类数据集的示例图像如图 4 所示, 每个数据集的训练集和测试集的划分如表 2 所示.



(a) CUB-200-2011

(b) Stanford Cars

(c) FGVC-Aircraft

图 4 细粒度图像数据集示例

表 2 本文使用数据集信息

数据集	类别	训练集数量	测试集数量
CUB-200-2011	200	5 994	5 794
Stanford Cars	196	8 144	8 041
FGVC-Aircraft	100	6 667	3 333

4.2 评价指标

本文采用分类准确率作为最终分类精度的评价指标. 分类准确率定义如下:

$$Accuracy = \frac{R_a}{R} \tag{3}$$

其中, R_a 表示测试集中模型预测正确的数量, R 表示测试集中所有图像的数量.

使用分类准确率作为本文的实验评价指标在 3 个常用细粒度图像分类数据集上进行比较. 在实验中, 本文方法采用细粒度图像分类领域常用数据集训练集和测试集划分方法、数据预处理方法. 其他方法的实验结果均引自原论文所给出的实验结果.

4.3 实验环境及参数设置

本文提出的模型采用 ResNet50 作为骨干网络, 并加载预训练模型权重. 实验所采用的服务器硬件配置为 i9 12900K 的 CPU 和 Nvidia RTX 3090 的 GPU. 软件配置为 Windows 10 操作系统, 并基于 Python 3.7, PyTorch 1.11.0 和 TorchVision 0.12.0 搭建深度学习框架.

训练参数: 将输入图像统一调整为 550 像素×550 像素的大小, 然后将图像随机裁剪为 448 像素×448 像素的大小. 本文使用随机梯度下降 (SGD) 来优化网络模型, 以批大小为 16 训练 200 个最大迭代次数. 为使用预训练权重的卷积层学习率设置为 0.000 2, 新添加的卷积层和全连接层的学习率设置为 0.002. 在训练的过程中使用余弦退火的策略来优化学习率. SGD 优化器设置动量和权重衰减为 0.9 和 0.000 5.

测试参数: 将输入图像调整为 550 像素×550 像素后中心裁剪为 448 像素×448 像素的大小. 其余未提及参数设置与训练参数设置一致.

4.4 实验结果与分析

为了充分验证本文所提出方法的有效性, 在 CUB-200-2011、Stanford Cars、FGVC-Aircraft 等 3 个常用细粒度图像分类数据集上完成了实验, 如图 5 所示.

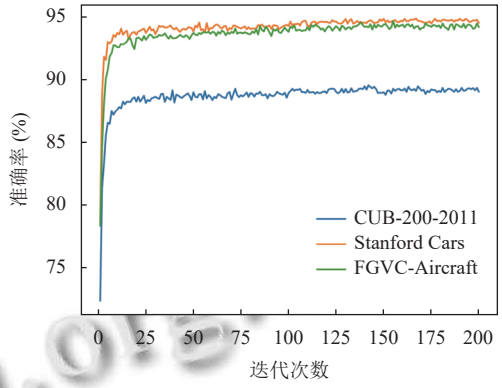


图 5 本文方法准确率变化趋势

CUB-200-2011: 表 3 展现了不同模型在 CUB-200-2011 数据集上的性能的对比. DCL 通过打破图像的全局结构并打乱局部区域来迫使网络发现潜在的细微特征取得了 87.8% 的准确率. SSSNet 通过分离和平滑采样操作改善了分类模型的性能, 能够更好地处理类内差异和类间相似, 获得了 89.0% 的准确率. 相比他们的方法, 本文提出的方法通过融合不同网络层之间的协同关注特征, 并使用特征分组注意获得了 89.5% 的准确率.

Stanford Cars: 如表 4 所示, 本文提出的方法优于大多数方法. MCL-Loss 通过将不同通道的特征进行区分, 并通过损失函数来制约它们的分布, 使得属于同一类别下的特征具有判别力, 获得了 93.7% 的准确率.

Cross-X 方法利用不同图像之间和不同网络层来学习多尺度特征获得了 94.6% 的准确率.

表 3 不同模型在 CUB-200-2011 数据集上的性能比较

方法	骨干网络	准确率 (%)
PA-CNN[14]	VGG19	87.8
S3N[15]	ResNet50	88.5
DF-GMM[16]	ResNet50	88.8
FDL[17]	ResNet50	88.6
ACNet[18]	ResNet50	88.1
CMN[19]	ResNet50	88.2
SSSNet[20]	ResNet50	89.0
PMA[21]	ResNet50	88.7
DCL[9]	ResNet50	87.8
MGE-CNN[22]	ResNet50	88.5
本文方法	ResNet50	89.5

表 4 不同模型在 Stanford Cars 数据集上的性能比较

方法	骨干网络	准确率 (%)
PA-CNN	VGG19	93.3
DCL	ResNet50	94.5
ACNet	ResNet50	94.6
MCL-Loss[10]	ResNet50	93.7
FDL	ResNet50	94.3
S3N	ResNet50	94.7
Cross-X[23]	ResNet50	94.6
MGE-CNN	ResNet50	93.9
PMA	ResNet50	93.1
本文方法	ResNet50	94.8

FGVC-Aircraft: 表 5 展示了本文提出方法在 FGVC-Aircraft 数据集上取得了最好的准确率为 94.7%. API-Net 利用对比的方法, 通过不同图像区域之间的成对交互, 捕捉细粒度细节和区分信息, 以关注对象部分之间的局部关系和微小差异, 获得了 93.4% 的准确率. GHORD 通过构建对象之间的关系图来解释对象之间的复杂关系和细微差异从而提高分类的准确率并获得了 94.3% 的准确率.

4.5 消融实验

为了验证各个模块是否能有效地提升模型性能, 对本文提出的算法模型在 3 个数据集上进行消融实验, 实验结果如表 6 所示. 其中 A 代表跨层协同注意模块, B 代表通道分组注意模块. 基准模型在 CUB-200-2011、Stanford Cars 和 FGVC-Aircraft 数据集上的准确率分别为 89.1%、94.5% 和 94.2%, 通过引入跨层协同注意模块分别获得了 0.3%、0.1% 和 0.2% 的提升, 这得益于模型有效地整合了浅层和深层的特征, 有效地发现并关注到了有助于分类的信息. 通过引入通道分组注意模块, 分别提升了 0.1%、0.2% 和 0.3% 的准确率, 这说明对通道进行分组能有效地提升模型的语义特征学习能力, 通过通道分组注意和跨层协同注意的共同作用, 本文方法相比基准模型分别获得了 0.4%、0.3% 和 0.5% 的准确率提升, 表明了本文方法在细粒度图像分类任务上的有效性, 表明通过整合浅层和深层的特征以及通道分组注意的有效性.

表 5 不同模型在 FGVC-Aircraft 数据集上的性能比较

方法	骨干网络	准确率 (%)
DCL	ResNet50	93.0
ACNet	ResNet50	92.4
MCL-Loss	ResNet50	92.6
FDL	ResNet50	93.4
Cross-X	ResNet50	92.6
DF-GMM	ResNet50	93.8
CMN	ResNet50	93.8
GHORD[24]	ResNet50	94.3
DTRG[25]	ResNet50	94.1
API-Net[26]	ResNet101	93.4
本文方法	ResNet50	94.7

表 6 本文方法在 3 个常用据集上的消融实验 (%)

方法	CUB-200-2011	Stanford Cars	FGVC-Aircraft
Baseline	89.1	94.5	94.2
Baseline+A	89.4	94.6	94.4
Baseline+A+B	89.5	94.8	94.7

4.6 可视化分析

为了直观地展示本文方法的有效性, 采用 Grad-CAM[27]的可视化方法来分析本文方法的性能. 随机选取各个数据集中测试集中的图像作为可视化分析的实验数据, 以热力图展示本文方法所预测出的具有判别力的区域. 图 6 展示了本文方法和基准模型 ResNet50 的可视化结果. 第 1 行图像表示从 3 个数据集上随机选取两张图像作为网络输入, 第 2 行图像表示基准模型所得出的热力图, 第 3 行图像表示本文方法所得出的热力图, 其中, 热力图中的红色区域表示网络模型所关注的区域. 根据热力图可以发现, 基准模型所关注的区域比较分散, 甚至更多的关注到了背景区域, 而本文方法能有效地找到并关注具有判别性的区域. 例如, 第 4 列的图像, 基准模型的注意区域在车轮下的草地, 而本文方法可以有效地将注意区域定位在车辆本身. 这表明本文方法可以有效地学习到具有判别性的特征信息, 有效地降低背景对预测模型的影响.

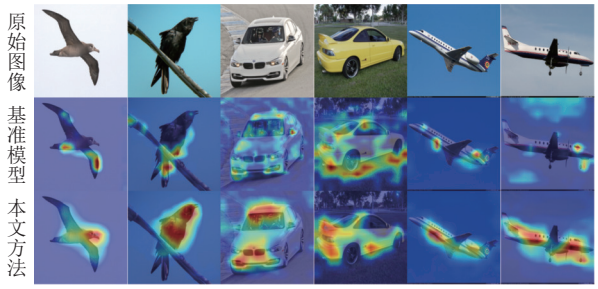


图 6 基准模型与本文方法在测试集上的热力图

5 结束语

本文提出了一种新颖的跨层协同注意和通道分组注意的细粒度图像分类网络. 本文提出的模型可以利用骨干网络 ResNet50 的浅层和深层信息来产生关注区域, 模型的最后一个阶段提取特征经过通道分组注意模块, 以挖掘通道之间的相关性和不同通道间具有判别力的特征信息. 引入跨层协同注意模块将骨干网络最后 3 个阶段所提取的特征通过协同信息交互并以跨层的方式进行多尺度融合, 使得网络学习更加全面且互补的特征信息. 实验表明, 本文提出的方法在多个常用细粒度图像分类数据集上都超过了大部分主流方法的精度. 在接下来的工作中, 将继续优化网络结构, 从而进一步提高模型的分类性能.

参考文献

1 申志军, 穆丽娜, 高静, 等. 细粒度图像分类综述. 计算机应用, 2023, 43(1): 51–60.

2 Zhang N, Donahue J, Girshick R, *et al*. Part-based R-CNNs for fine-grained category detection. Proceedings of the 13th European Conference on Computer Vision. Zurich: Springer, 2014. 834–849.

3 Branson S, van Horn G, Belongie S, *et al*. Bird species categorization using pose normalized deep convolutional nets. arXiv:1406.2952, 2014.

4 Lin TY, Roychowdhury A, Maji S. Bilinear convolutional neural networks for fine-grained visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(6): 1309–1322. [doi: 10.1109/TPAMI.2017.2723400]

5 Gao Y, Beijbom O, Zhang N, *et al*. Compact bilinear pooling. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016, 317–326.

6 Kong S, Fowlkes C. Low-rank bilinear pooling for fine-grained classification. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 7025–7034.

7 Fu JL, Zheng HL, Mei T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 4476–4484.

8 Zheng HL, Fu JL, Mei T, *et al*. Learning multi-attention convolutional neural network for fine-grained image recognition. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 5219–5227.

9 Chen Y, Bai YL, Zhang W, *et al*. Destruction and construction learning for fine-grained image recognition. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5152–5161.

10 Chang DL, Ding YF, Xie JY, *et al*. The devil is in the channels: Mutual-channel loss for fine-grained image classification. IEEE Transactions on Image Processing, 2020, 29: 4683–4695. [doi: 10.1109/TIP.2020.2973812]

11 Huang SL, Wang XC, Tao DC. Stochastic partial swap: Enhanced model generalization and interpretability for fine-grained recognition. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021, 600–609.

12 Hu J, Shen L, Albanie S, *et al*. Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011–2023. [doi: 10.1109/TPAMI.2019.2913372]

13 Zhang T, Chang DL, Ma ZY, *et al*. Progressive co-attention network for fine-grained visual classification. Proceedings of the 2021 International Conference on Visual Communications and Image Processing. Munich: IEEE, 2021. 1–5.

14 Zheng HL, Fu JL, Zha ZJ, *et al*. Learning rich part hierarchies with progressive attention networks for fine-grained image recognition. IEEE Transactions on Image Processing, 2020, 29: 476–488. [doi: 10.1109/TIP.2019.2921876]

15 Ding Y, Zhou YZ, Zhu Y, *et al*. Selective sparse sampling for fine-grained image recognition. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 6598–6607.

16 Weng ZH, Wang SJ, Yang SY, *et al*. Weakly supervised fine-grained image classification via Guassian mixture model

oriented discriminative learning. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 9746–9755.

17 Liu CB, Xie HT, Zha ZJ, *et al*. Filtration and distillation: Enhancing region attention for fine-grained visual categorization. Proceedings of the 27th AAAI Conference on Artificial Intelligence. Washington: AAAI Press, 2020. 11555–11562.

18 Ji RY, Wen LY, Zhang LB, *et al*. Attention convolutional binary neural tree for fine-grained visual categorization. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10465–10474.

19 Deng WJ, Marsh J, Gould S, *et al*. Fine-grained classification via categorical memory networks. IEEE Transactions on Image Processing, 2022, 31: 4186–4196. [doi: 10.1109/TIP.2022.3181492]

20 Rong SH, Wang ZL, Wang J. Separated smooth sampling for fine-grained image classification. Neurocomputing, 2021, 461: 350–359. [doi: 10.1016/j.neucom.2021.07.067]

21 Song KT, Wei XS, Shu XB, *et al*. Bi-modal progressive mask attention for fine-grained recognition. IEEE Transactions on Image Processing, 2020, 29: 7006–7018. [doi: 10.1109/TIP.2020.2996736]

22 Zhang LB, Huang SL, Liu W, *et al*. Learning a mixture of granularity-specific experts for fine-grained categorization. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 8330–8339.

23 Luo W, Yang XT, Mo XJ, *et al*. Cross-X learning for fine-grained visual categorization. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 8241–8250.

24 Zhao YF, Yan K, Huang FY, *et al*. Graph-based high-order relation discovery for fine-grained recognition. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 15074–15083.

25 Liu KJ, Chen K, Jia K. Convolutional fine-grained classification with self-supervised target relation regularization. IEEE Transactions on Image Processing, 2022, 31: 5570–5584. [doi: 10.1109/TIP.2022.3197931]

26 Zhuang PQ, Wang YL, Qiao Y. Learning attentive pairwise interaction for fine-grained classification. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 13130–13137. [doi: 10.1609/aaai.v34i07.7016]

27 Selvaraju RR, Cogswell M, Das A, *et al*. Grad-CAM: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision, 2020, 128(2): 336–359. [doi: 10.1007/s11263-019-01228-7]

(校对责编: 牛欣悦)