计算机科学
COMPUTER SCIENCE

基于ConvNeXt热图定位和对比学习的细粒度图像分类研究

郑世杰, 王高才

引用本文

郑世杰, 王高才. 基于ConvNeXt热图定位和对比学习的细粒度图像分类研究[J]. 计算机科学, 2023, 50(10): 119-125.

ZHENG Shijie, WANG Gaocai. Study on Fine-grained Image Classification Based on ConvNeXt Heatmap Localization and Contrastive Learning [J]. Computer Science, 2023, 50(10): 119-125.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220900196

基于 ConvNeXt 热图定位和对比学习的细粒度图像分类研究

郑世杰　王高才
广西大学计算机与电子信息学院　南宁 530004
(2371363651@qq.com)

摘　要　针对细粒度图像分类中高类内差异和低类间差异的挑战,提出一种以 ConvNeXt 网络为主干,使用 GradCAM 热图进行裁剪和注意力擦除的多分支细粒度图像分类方法。该方法利用 GradCAM 通过梯度回流得到网络的注意力热图,定位到具有判别性特征的区域,裁剪并放大该区域,使网络关注局部更深层次的特征。同时引入有监督的对比学习,扩大类间差异,减小类内差异。最后进行热图注意力擦除操作,使网络在关注最具判别性特征的前提下,也能关注其他对分类有用的区域。所提方法在 CUB-200-2011,Stanford Cars,FGVC Aircraft 和 Stanford Dogs 数据集上的分类准确率分别达到了 91.8%,94.9%,94.0%,94.4%,优于多种主流的细粒度图像分类方法,并且在 CUB-200-2011 和 Stanford Dogs 数据集上分别达到了 top-3 和 top-1 的分类准确率。

关键词:细粒度图像分类;注意力;有监督对比学习;热图;多分支
中图法分类号　TP391

Study on Fine-grained Image Classification Based on ConvNeXt Heatmap Localization and Contrastive Learning

ZHENG Shijie and WANG Gaocai
School of Computer and Electronic Information,Guangxi University,Nanning 530004,China

Abstract　Aiming at the challenges of high intra-class disparity and low inter-class disparity in fine-grained image classification, a multi-branch fine-grained image classification method based on ConvNeXt network and using GradCAM heatmap for cropping and attention erasure is proposed. This method uses GradCAM to obtain the attention heatmap of the network through gradient reflow, locates the region with discriminative features, crops and enlarges the region, and makes the network focus on local deeper features. At the same time, supervised contrastive learning is introduced to expand between-class differences and reduce intra-class differences. Finally, a heatmap attention erasure operation is performed to enable the network to focus on other regions useful for classification while focusing on the most discriminative features. The proposed method achieves 91.8%, 94.9%, 94.0%, and 94.4% classification accuracy on CUB-200-2011, Stanford Cars, FGVC Aircraft, and Stanford Dogs datasets, respectively, which is better than many mainstream fine-grained image classification methods. And this method achieves top-3 and top-1 classification accuracy on the CUB-200-2011 and Stanford Dogs datasets, respectively.

Keywords　Fine-grained image classification, Attention, Supervised contrastive learning, Heatmap, Multi-branch

1　引言

细粒度图像分类的目的是对某一类别的子类进行划分,如汽车的子类别[1]、鸟类[2]、飞机类[3]。细粒度图像分类的主要挑战是同一数据集内部存在高类内差异和低类间差异。如图 1 所示,高类内差异体现在:由于受到光照、视点、遮挡和背景杂乱等因素影响,图中同一类别的列呈现出不同的外观。低类间差异体现在:图中不同列的鸟类属于不同类别,但在第一行中具有非常相似的外观。目前主流的分类方法主要分为两个方面,即目标对象定位以及局部有判别性特征的捕获。近些年来,文献[4-7]提出了有关的方法,证明了该图像分类思想的有效性。本文则提出一种新的基于 GradCAM[8] 热图定位的多分支细粒度图像分类方法 GHOLM-Net(GradCAM Heatmap Object Location Multi-branch Network),同时融入有监督的对比学习,进一步提高网络的分类能力,方法总体结构如图 2 所示。首先将图像输入 ConvNeXt[9] 主干网络,通过梯度回流的方式使用 GradCAM 获得网络关注区域的热图。根据热图中响应值的大小进一步裁剪网络关注的区域,进行

到稿日期:2022-09-20　返修日期:2022-12-06
基金项目:国家自然科学基金(62062007)
This work was supported by the National Natural Science Foundation of China(62062007).
通信作者:王高才(wanggcgx@163.com)

更深层次特征的捕获。同时引入有监督的对比学习和热图注意力擦除方案，最后使用有监督的对比损失和交叉熵损失共同更新网络参数。本文方法的主要优势为：

1)以 ConvNeXt 为主干，提出了基于 GradCAM 热图定位对象区域的细粒度图像分类网络 GHOLM-Net，通过裁剪后的对象进一步捕获更深层次有判别性的特征。

2)利用对象数据增强的样本，引入有监督的对比学习，在原有分类网络基础之上，进一步增大类间差异，减小类内差异。

3)提出热图注意力擦除方案，在网络关注局部深层特征表示的同时，使网络关注其他对分类有用的区域，提高分类网络的鲁棒性。

4)在 4 个公共的数据集上进行对比实验，证明了所提方法的有效性和优秀的分类效果。

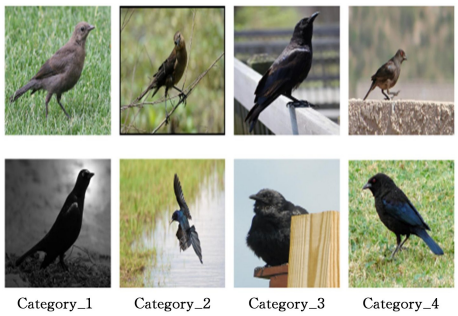

图 1 高类内差异和低类间差异

Fig. 1 High within-class and low between-class variance

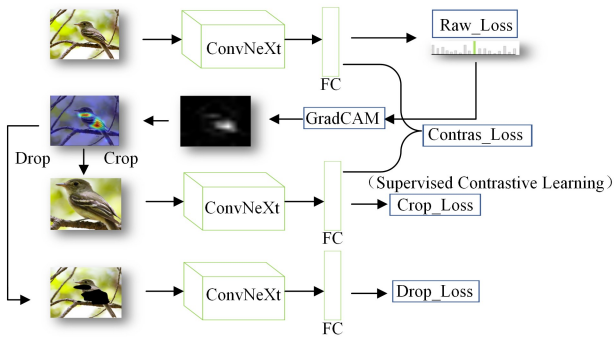

图 2 网络总体结构图

Fig. 2 Overall network structure

2 相关研究工作

2.1 细粒度图像分类

细粒度图像分类方法主要分为两大类，一是基于强监督的细粒度图像分类，二是基于弱监督的细粒度图像分类。强监督方法最大的特点是它使用了除图像级标签之外的标注，比如对象边界框、部位框图标注、部位标注点等，使分类网络能够根据标注信息捕获特定的图像特征。Branson 等[10]通过标注点和描框来训练网络进行对象姿态估计，利用图的聚类算法来紧密学习姿态归一化空间，进一步计算局部对象特征以进行分类。Huang 等[11]提出了一种基于标注部位堆叠的 CNN 结构，并利用全卷积网络(FCN)[12]定位到多个对象部位，再借助两流分类网络，同时捕获标注好的对象级和部位级特征。Lam 等[13]通过生成输入图像部位的提议框，使得网络捕获到更具有判别性的细粒度特征。以上这些强监督方法能够充分利用图像的标注信息，也为应对细粒度图像分类的挑战打开了局面。但过度依赖对象标注框或者部位标注点成为了其发展的最大障碍，这些强监督方法需要人工进行繁琐的部位标注，难以在实际当中进行应用。近年来，基于弱监督的细粒度图像分类方法发展迅速，其研究主要集中在目标对象定位以及捕获局部判别性的特征。文献[14-16]分别提出了 RA-CNN，MA-CNN，B-CNN，它们仅使用标签作为监督信息进行细粒度图像的识别，且都取得了较好的分类效果，证明了弱监督分类方法的有效性。B-CNN 利用双线性网络对细粒度图像的特征进行捕获，再进行丰富的特征融合。RA-CNN 利用深层滤波器对关键部位进行多尺度的定位裁剪，从而捕获关键部位的更深层次特征。MA-CNN 利用特征映射反应网络对不同通道的关注程度，聚集响应值相近的通道，使得相同部位的值更接近，不同部位的差值尽可能大，从而提高网络辨别不同部位特征的能力。

2.2 对象数据增强与 GradCAM

对象数据增强是近年来细粒度图像分类领域常用的方法之一，并取得了较好的效果。Hu 等[5]利用主干网络特征图的掩码对原图进行裁剪和擦除，实现对象数据增强。Zhang 等[4]通过计算主干网络最后一层特征图的均值，得到大于均值的区域掩码后，裁剪掩码的极大连通区域，得到对象定位图，利用对象定位图进行对象数据增强，使网络捕获对象更深层次的特征。Hanselmann 等[17]提出了一种新的对象定位模块，该模块基于反向传播的梯度以及额外的自监督损失函数进行训练并更新参数。Zhang 等[7]提出了一种新的视觉 Transformer 模型，该模型可进行自适应注意力的多尺度融合，由自适应的注意力定位到对象区域并裁剪，从而进行对象数据增强。Hu 等[6]提出了循环注意多尺度视觉 Transformer 模型，模型使用动态 Patch 建议模块(DPPM)引导区域放大进行对象数据增强。

为了建立对深度学习模型的信任，研究人员在过去几年中提出了各种方法来解释为什么网络模型会做出这样或那样的决定。Oquab 等[18]提出了 CAM，其对卷积神经网络的结构进行了修改，将原来的全连接层替换为全局平均池化层，对得到的特征图进行加权组合，得到视化网络焦点的热图，通过可视化的方式说明了网络做出最终分类决策的原因。CAM 方法虽然简单，但是需要对网络结构进行修改，导致需要重新训练网络模型，从而限制了 CAM 的使用场景。为了弥补 CAM 方法需要修改网络结构的不足，Selvaraju 等[8]提出了 GradCAM。相对于 CAM 而言，GradCAM 不需要修改网络结构，它能根据逻辑输出的最终对应类别反映出网络的关注区域。首先通过网络正向传播得到最后一层的特征图和网络预测值；然后根据预测类别 c 进行梯度回流，计算特征图梯度信息，根据梯度信息得到针对特征层中每个通道的重要程度 w；最后经过加权求和以及 ReLU 函数激活得到最终 GradCAM 的热图，该过程如图 3 所示。本文借助 GradCAM 生成

的热图，定位到网络所关注的区域，通过注意力裁剪和擦除的方式进行相应的对象数据增强。

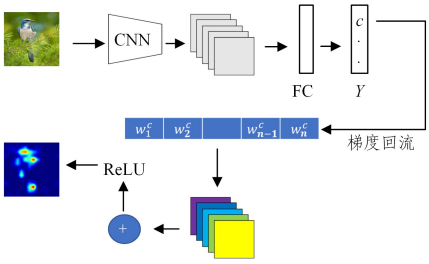

图 3 GradCAM 流程图

Fig. 3 GradCAM flow chart

2.3 ConvNeXt

近年来，Vit[19] 和 Swin-Transformer[20] 在图像分类、图像分割以及目标检测领域展现出了强大的性能。正当许多的视觉领域的工作逐渐被视觉 Transformer 取代时，Liu 等[9] 结合 MobileNet[21]，ResNet[22]，Vit[19]，Swin-Transformer[20] 等网络的特点和理念提出了 ConvNeXt 主干网络，其在图像分类和图像分割以及目标检测等领域可以达到和 Vit 以及 Swin-Transformer 同等或更好的效果和性能。ConvNeXt 是基于 ResNet 网络进行改进的，改进内容主要包括：增加单一模块的堆叠次数；用深度可分离卷积（Depthwise Convolution）代替普通卷积、引入 MobileNet 当中的倒置瓶颈结构（Inverted Bottleneck）；在模块中使用大量的 7 * 7 卷积核代替 3 * 3 卷积核；使用 GELU 激活函数和 LN（Layer Normalization）代替原来 ResNet 使用的 RELU 和 BN（Batch Normalization）。

ConvNeXt 模块结构如图 4 所示，首先将输入模块的特征经过 7 * 7 深度分离卷积，接着通过 LN 以及 1 * 1 的普通卷积进行升维，再使用 GELU 进行激活操作，利用 1 * 1 的普通卷积恢复通道数，最后与输入前的特征相加。本文使用 ConvNeXt 作为主干网络，结合 GradCAM 生成对应热图，根据热图进行注意力区域的裁剪和擦除，并验证 ConvNeXt 在细粒度图像分类领域的有效性以及优秀的分类效果。

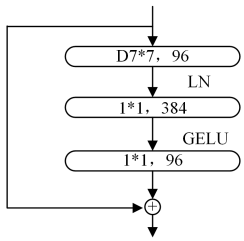

图 4 ConvNeXt 模块结构图

Fig. 4 Diagram of ConvNeXt module structure

2.4 有监督对比学习

近年来，对比学习的工作，如 Simclr[23]，Moco[24]，MoCoV2[25] 等，在自监督领域取得了较大的成果。这些方法有一些共同之处，即它们都区分了一定的正样本和负样本，正样本通常由同源数据增强的数据组成，负样本为批量中的其他样本，通过增大正负样本之间的差异以及减小正样本之间的差异来更新网络。而在无监督的对比学习的基础之上，Khosla 等[26] 提出了一种有监督的对比学习方法，利用标签信息指导对比学习，正样本在同一类别中提取，而不仅是数据增强后的同源样本。本文受到有监督对比学习的启发，对裁剪分支与原始分支的特征输出引入有监督的对比学习，使用对比损失辅助网络进行参数更新。

3 基于 ConvNeXt 和热图定位的多分支方法——GHOLM-Net

3.1 GHOLM-Net 网络总体结构

如图 2 所示，整体网络框架由 4 个部分组成：1）Raw 分支，原图像经过主干网络 ConvNeXt 得到最终分类输出；2）Crop 分支，结合 Raw 分支最终的分类输出，通过 GradCAM 生成的热图进行注意力区域的裁剪，并将裁剪图输入主干网络中；3）Drop 分支，使用 GradCAM 生成的热图进行注意力区域的擦除，并将注意力擦除后的图像输入主干网络中；4）特征输出的对比学习，本文只选择对 Raw 分支和 Crop 分支进行对比学习。最终测试时，逻辑输出结果由 Crop 分支和 Raw 分支相加组成。

3.2 GradCAM 生成热图

我们定义 $A \in R^{C\times h\times w}$，表示输入图像 X 的最后一个卷积层的输出为 A，其代表具有 C 个通道和空间大小为 $h\times w$ 的特征图。

利用 GradCAM 生成热图的具体步骤如下：首先给定一张输入图像 X，通过主干网络 ConvNeXt 提取特征，再通过全连接层得到图像所属类别 c 的预测值 y^c。根据 y^c 进行反向传播，得到指定的最后一层特征图 A 的梯度信息，根据梯度信息计算出特征图 A 中每个通道特征的重要程度 a_k^c，即权重。a_k^c 中 c 表示类别，A^k 表示 A 中的第 k 个特征图，A_{ij}^k 表示在 A^k 中坐标为 (i,j) 位置处的数据。a_k^c 的计算式如式（1）所示：

$$a_k^c=\frac{1}{h\times w}\sum_i\sum_j\frac{\partial y^c}{\partial A_{ij}^k} \tag{1}$$

求得 y^c 对 A 中每个通道特征的权重后，对 A 中每个通道特征进行加权求和，使用 ReLU 函数激活得到输出的 GradCAM 特征图，即 $gradcam$。最后上采样得到原图大小的热图 heat。该过程如式（2）、式（3）所示：

$$gradcam=\mathrm{ReLU}\sum_k a_k^c A^k \tag{2}$$

$$heat=upSample(gradcam) \tag{3}$$

3.3 注意力裁剪与擦除

在注意力裁剪和注意力擦除部分，首先计算热图 $heat$ 的特征平均值 $\bar{a}$，如式（4）所示。其中 $heat(x,y)$ 代表热图 $heat$ 中特征点的值，h 和 w 代表热图的高和宽。

$$\bar{a}=\frac{\sum_{x=0}^{w-1}\sum_{y=0}^{h-1}heat(x,y)}{h\times w} \tag{4}$$

裁剪分支的目的是放大原图中的关键区域。本文使用 $heat$ 热图中的特征平均值 $\bar{a}$ 来界定网络对区域的关注程度，对 $\bar{a}$ 取系数 μ，设置一个阈值 $\mu\bar{a}$，最佳的 μ 值通过 4.3.2 节中的消融实验获得。当热图 $heat$ 中的值大于阈值时，将掩码值设置为 1，否则设置为 0，如式（5）所示：

$$Mask^{crop}_{(x,y)}=\begin{cases}1, & heat(x,y)>\mu\bar{a}\\ 0, & otherwise\end{cases} \tag{5}$$

GradCAM 热图反映了 ConvNeXt 网络最关注的区域，即热图当中响应值较大的区域，这些响应值较大的区域均为识别对象的某些部位或区域，根据这些区域生成对应的掩码图，进一步对值为 1 的掩码位置求极大联通区域。根据联通区域在原图上进行裁剪，从而裁剪出网络较为关注的局部区域，同时也保证了对象裁剪的有效性和准确性，得到对象数据增强后的裁剪图。

注意力擦除部分的目的是根据热图进一步擦除网络最关注的区域，鼓励网络关注其他有辨别性的区域。注意力擦除分支和裁剪分支类似，同样使用 $heat$ 的特征平均值 $\bar{a}$ 来界定网络对区域的关注程度，对 $\bar{a}$ 取系数 θ，设置一个阈值 $\theta\bar{a}$，本文的 θ 值参考文献[5]，将其设置为 1～2 的随机数。同样地，当 $heat$ 中的值大于阈值时，将掩码值设置为 1，否则设置为 0，如式(6)所示：

$$Mask^{drop}_{(x,y)}=\begin{cases}1, & heat(x,y)>\theta\bar{a}\\ 0, & otherwise\end{cases} \tag{6}$$

最后将掩码图中特征值为 1 的区域在原图对应位置进行擦除，将得到的图像再次输入网络中。

3.4 损失函数

损失函数主要由交叉熵损失和有监督的对比损失组成。首先，在训练阶段，我们使用原始分支、裁剪分支以及注意力擦除分支组成一个三分支网络结构，共享一个主干网络进行特征提取和分类，三分支均使用交叉熵损失作为分类损失，如式(7)－式(10)所示：

$$L_{raw}=-\log(P_r(c)) \tag{7}$$

$$L_{crop}=-\log(P_c(c)) \tag{8}$$

$$L_{drop}=-\log(P_d(c)) \tag{9}$$

$$L_{total\_ce}=L_{raw}+L_{crop}+\gamma L_{drop} \tag{10}$$

其中，P_r，P_c 以及 P_d 分别表示原始分支、裁剪分支以及擦除分支的最后一个 softmax 层的类别输出概率；c 代表输出类别；L_{total_ce} 代表总的交叉熵损失；γ 为损失权重超参数，由于注意力擦除后的图像包含较多的背景区域，为了避免网络过度关注背景和其他区域，本文对 γ 的值取为 0.2。

将在有监督的对比学习部分，本文将同一批量中同一类别的数据作为正样本，其他类别作为负样本。即当同一批次中出现同类不同源的样本时，仍视其为正样本，根据是否同类来区分该批次数据的正负样本。正负样本划分如图 5 所示，对比损失的计算式如式(11)所示。

$$L_c=\sum_{i\in I}\frac{-1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp(z_i\times z_p/\tau)}{\sum_{a\in A(i)}\exp(z_i\times z_a/\tau)} \tag{11}$$

其中，I 代表一个批量中的样本；i 代表 I 中的任意一个样本；z_i 和 z_p 分别代表原样本和正样本经过主干网络的全连接层之后输出的一维特征；z_a 代表所有不同标签样本的输出特征；$P(i)$ 代表每一个批量中与 i 互为正样本的集合，$A(i)$ 代表与 i 互为负样本的集合；exp 是计算样本相似度的函数；τ 为对比损失的温度系数，它控制了模型对负样本的区分度，我们参考文献[26]，设 τ 值为 0.7。

(a)Positives (b)Negatives

图 5 正负样本分类图

Fig. 5 Diagram of positive and negative sample classification

总的损失函数为交叉熵损失和对比损失之和，两者共同指导网络参数的更新，如式(12)所示：

$$L_{total}=L_{total\_ce}+L_c \tag{12}$$

4 实验设计与结果分析

4.1 数据集与评价指标

本文使用的 4 个数据集为鸟类 CUB-200-2011、飞机类 FGVC-Aircraft、车辆类 Stanford Cars 和狗类 Stanford Dogs。数据集具体数据如表 1 所列。

表 1 数据集介绍

Table 1 Dataset introduction

数据集	类别数	训练集/张	测试集/张
CUB-200-2011	200	5994	5794
FGVC-Aircraft	100	6667	3333
Stanford Cars	196	8144	8041
Stanford Dogs	120	12000	8580

同时，使用模型在测试集上的分类准确率来评判所提方法的分类效果，计算式如式(13)所示。

$$Acc=\frac{T\_num}{A\_num} \tag{13}$$

其中，T_num 和 A_num 分别代表预测正确样本的数量和测试集样本总数量。

4.2 实验细节与数据预处理

本文的网络架构基于 Pytorch 实现，ConvNeXt 作为主干网络，使用 ImageNet21k 预训练模型进行参数初始化；对于 Stanford Dogs 数据集，则使用 ImageNet1k 预训练模型进行参数初始化。在训练过程中，仅使用图像级别的标签。使用 SGD 作为优化器，动量设置为 0.9，权重衰减设置为 0.0001，Batchsize 为 16。初始学习率为 0.001，学习率在 15 和 30 个 epoch 后均乘以 0.1。训练阶段将图像尺寸调整为 510 * 510 大小，接着随机裁剪至 448 * 448，水平随机翻转所有数据集使用并调整鲜艳度和亮度。测试阶段，将图像尺寸放大到 510 * 510 大小，以中心裁剪至 448 * 448。对 Stanford Dogs 数据集则不进行裁剪操作，在训练和测试过程中直接调整大小为 448 * 448，所有的实验均在一台 GPU 型号为 RTXA5000、CPU 核数为 14 核，以及内存为 30GB 的服务器上进行。

4.3 消融实验

4.3.1 验证模块的有效性

为了验证模块的有效性，在 CUB-200-2011 数据集和

Stanford Cars 数据集上进行了 4 组消融实验，实验结果如表 2 和表 3 所列。

从表 2 和表 3 中可以得出，本文提出的模块均可以提升分类准确率。两个表中的实验 1 仅使用主干网络 ConvNeXt 进行分类，在两个数据集上的分类准确率分别达到了 90.6% 和 93.2%。

表 2 CUB-200-2011 上的分支消融实验

Table 2 Branch ablation experiments on CUB-200-2011

	Raw	Crop	Sup	Drop	*Acc*/%
实验 1	√				90.6
实验 2	√	√			91.4(+0.8)
实验 3	√	√	√		91.6(+1.0)
实验 4	√	√	√	√	91.8(+1.2)

表 3 Stanford Cars 上的分支消融实验

Table 3 Branch ablation experiments on Stanford Cars

	Raw	Crop	Sup	Drop	*Acc*/%
实验 1	√				93.2
实验 2	√	√			94.3(+1.1)
实验 3	√	√	√		94.6(+1.4)
实验 4	√	√	√	√	94.9(+1.7)

实验 2 验证 Crop 分支的有效性，表 2 和表 3 的实验 2 的准确率分别在实验 1 的基础上提升了 0.8% 和 1.1%，说明热图注意力裁剪方案能够使网络捕获更深层次的特征，从而提高分类准确率。

实验 3 的验证有监督对比学习模块 Sup 的有效性，表 2 和表 3 的实验 3 的准确率分别在实验 2 的基础上提升了 0.2% 和 0.3%，说明对比学习模块能够在一定程度上使网络模型进一步区分各类别之间的差异，提升分类效果。

实验 4 验证注意力擦除 Drop 模块的有效性，表 2 和表 3 的实验 4 的准确率分别在实验 3 的基础上有 0.2% 和 0.3% 的提升，说明该模块能够使网络进一步关注其他区域，提升分类网络的鲁棒性。

4.3.2 裁剪 μ 参数消融实验

为了探究超参数 μ 对分类准确率的影响，我们在 CUB-200-2011 数据集上进行了 6 组对比实验，实验结果如表 4 所列。

表 4 裁剪参数 μ 消融实验

Table 4 Crop parameter μ ablation experiment

μ	1.6	1.7	1.8	1.9	2.0	2.1
Acc/%	91.5	91.8	91.7	91.6	91.5	91.4

通过观察表 4 可知，当 μ 为 1.7 时，分类准确率最高。根据实验结果，我们分析：相对于某一数据集而言，当 μ 较小时，裁剪模块在原图上裁剪的区域较大，导致不能很好地定位到关键区域；当 μ 较大时，裁剪模块在原图上裁剪的区域较小，一定程度上丢失了许多关键区域，导致分类准确率下降。

4.3.3 逻辑输出选择消融实验

为了探究逻辑输出类别概率的最优结果，我们对 Raw 和 Crop 分支及其组合的逻辑输出在 CUB-200-2011 数据集上进行了分类准确率的对比实验，实验结果如表 5 所列。从表中可以看出，实验 1 和实验 2 分别单独使用 Raw Logit 和 Crop Logit 的分类结果相比基准方法都有着较大的提升。实验 3 将 Raw Logit 和 Crop Logit 相加，其结果优于实验 1 和实验 2。我们分析：Crop 分支中对于局部判别性区域的关注较多，其逻辑输出与 Raw 分支的逻辑输出相加，弥补了 Raw 分支在原始图像上难以捕获局部重要特征的缺陷。而仅使用 Raw Logit 的好处是，在测试时不需要进行梯度的计算，在一定程度上减少了计算量。

表 5 逻辑输出选择消融实验

Table 5 Logical output selection ablation experiment

	Raw Logit	Crop Logit	*Acc*/%
实验 1	√		91.6(+1.0)
实验 2		√	91.5(+0.9)
实验 3	√	√	91.8(+1.2)

4.4 算法比较

为了验证所提方法的优越性，我们将其和现有先进的细粒度图像分类方法进行比较。表 6－表 9 分别展示了本文方法与其他现有方法在 Stanford Cars，CUB-200-2011，FGVC-Aircraft，Stanford Dogs 这 4 个公开数据集上的分类准确率的对比。

表 6 Stanford Cars 上的对比实验

Table 6 Comparative experiment on Stanford Cars

Method	Backbone	*Acc*/%
WS-DAN[5]	RestNet-50	94.5
TransFG[27]	Vit-B_16	94.8
SEB+Eff-Net-B5[28]	EfficientNet	94.6
Vit[19]	Vit-B_16	93.7
ConvNeXt[9]	ConvNeXt-S	93.2
(ours)	ConvNeXt-S	94.9

表 7 CUB-200-2011 上的对比实验

Table 7 Comparative experiment on CUB-200-2011

Method	Backbone	*Acc*/%
MMAL[4]	ResNet-50	89.6
TransFG[27]	Vit-B_16	91.7
AFTrans[7]	Vit-B_16	91.5
RAMS[6]	Vit-B_16	91.3
Vit[19]	Vit-B_16	90.2
ConvNeXt[9]	ConvNeXt_S	90.6
(ours)	ConvNeXt_S	91.8

表 8 FGVC-Aircraft 上的对比实验

Table 8 Comparative experiment on FGVC-Aircraft

Method	Backbone	*Acc*/%
M-Granularity[29]	ResNet-50	93.8
API-Net[30]	DenseNet-161	93.9
CAL[31]	ResNet-50	94.2
Vit-SAC[32]	Vit-B_16	93.1
Vit[19]	Vit-B_16	92.1
ConvNeXt[9]	ConvNeXt-S	91.9
(ours)	ConvNeXt-S	94.0

表 9 Stanford Dogs 上的对比实验

Table 9 Comparative experiment on Stanford Dogs

Method	Backbone	*Acc*/%
API-Net[30]	DenseNet-161	90.3
FFVT[33]	Vit-B_16	91.5
TransFG[27]	Vit-B_16	92.3
WS_DAN-SAC[32]	Vit-B_16	93.1
Vit[19]	Vit-B_16	91.7
ConvNeXt[9]	ConvNeXt-S	93.4
(ours)	ConvNeXt-S	94.4

根据对比实验的结果可知，本文方法的分类准确率相比原主干网络 ConvNeXt-S 有明显提升，展现了本文方法的有效性和优秀的分类效果。在 CUB-200-2011 数据集上，本文算法在准确率上优于所有以 CNN 和 Vit 为主干网络的细粒度图像分类方法，在主干 ConvNeXt 的基础上提升了 1.2%，并且达到了 top-3 的分类准确率。在 Stanford Cars 数据集和飞机类 Aircraft 数据集上，与基线网络相比，本文方法的准确率分别有 1.7%和 2.1%的提升，优于大多数现有的方法。在 Stanford Dogs 数据集上，与基线分类网络相比，本文方法准确率得到了 1.0%的提升，并达到了 top-1 的分类准确率。

4.5 可视化实验

本文在 4 个数据集上对各个阶段的图像进行可视化实验，如图 6 所示。

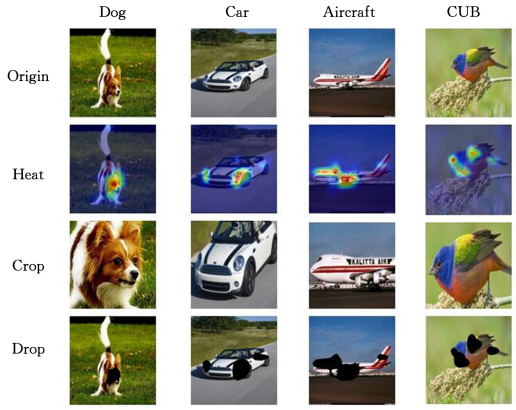

图 6 分支可视化图

Fig. 6 Branch visualization

其中，第一行为输入的原图；第二行为热图和原图叠加的可视化，该行反映了网络最关注的区域；第三行和第四行分别为裁剪和进行了注意力擦除后的图像。从图中可以看出，ConvNeXt 模型重点捕获具有判别性的特征，通过热图裁剪的方式，可以进行更深层次的特征捕获。又由于裁剪分支和原始分支更侧重于对最具判别性特征的捕获，导致网络忽略了一些对分类有用的其他区域，引入注意力擦除分支则能在一定程度上解决该问题。

结束语 本文通过 ConvNeXt 提取图像中的特征信息，利用 GradCAM 生成注意力热图，借助注意力热图对原图进行裁剪，使得网络能够挖掘局部的深层特征信息，同时引入有监督的对比学习，以更好地扩大类间差异，减小类内差异。最后引入注意力擦除方案，使网络关注其他有判别性的区域。在 4 个公开的数据集上验证了所提方法在细粒度图像分类领域的有效性和优秀的分类效果。在下一步工作中，我们将进一步提高注意力裁剪准确性，从而提升最终的分类准确率。

参 考 文 献

[1] KRAUSE J, STARK M, DENG J, et al. 3D Object Representations for Fine-Grained Categorization[C]// IEEE International Conference on Computer Vision Workshops. IEEE, 2014.

[2] ELINDER P, BRANSON S, MITA T, et al. The caltech-ucsd birds-200-2011 dataset [R]. California Institute of Technology, 2011.

[3] MAJI S, RAHTU E, KANNALA J, et al. Fine-grained visual classification of aircraft[J]. arXiv: 1306. 5151, 2013.

[4] ZHANG F, LI M, ZHAI G, et al. Multi-branch and multi-scale attention learning for fine-grained visual categorization[C]// International Conference on Multimedia Modeling. Cham: Springer, 2021: 136-147.

[5] HU T, QI H, HUANG Q, et al. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification[J]. arXiv: 1901. 09891, 2019.

[6] HU Y, JIN X, ZHANG Y, et al. Rams-trans: Recurrent attention multi-scale transformer for fine grained image recognition [C]// Proceedings of the 29th ACM International Conference on Multimedia. 2021: 4239-4248.

[7] ZHANG Y, CAO J, ZHANG L, et al. A free lunchfrom ViT: adaptive attention multi-scale fusion Transformer for fine-grained visual recognition[C]// 2022 IEEE International Conference on Acoustics, Speech and Signal Processing) ICASSP 2022). IEEE, 2022: 3234-3238.

[8] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization [J]. International Journal of Computer Vision, 2020, 128(2): 336-359.

[9] LIU Z, MAO H, WU C Y, et al. A convnet for the 2020s[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 11976-11986.

[10] BRANSON S, VAN HORN G, BELONGIE S, et al. Bird species categorization using pose normalizeddeep convolutional nets[J]. arXiv: 1406. 2952, 2014.

[11] HUANG S, XU Z, TAO D, et al. Part-Stacked CNN for Fine-Grained Visual Categorization[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016.

[12] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3431-3440.

[13] LAM M, MAHASSENI B, TODOROVIC S. Fine-Grained Recognition as HSnet Search for InformativeImage Parts[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2017.

[14] FU J, ZHENG H, TAO M. Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition[C]// IEEE Conference onComputer Vision & Pattern Recognition. IEEE, 2017.

[15] ZHENG H, FU J, TAO M, et al. Learning Multi-attention Convolutional Neural Network for Fine-Grained Image Recognition [C]// 2017 IEEE International Conference on Computer Vision. IEEE, 2017.

[16] LIN T Y, ROYCHOWDHURY A, MAJI S. Bilinear CNNs for fine-grained visual recognition[J]. arXiv: 1504. 07889, 2015.

[17] HANSELMANN H, NEY H. Fine-grained visual classification with efficient end-to-end localization [J]. arXiv: 2005. 05123, 2020.

[18] OQUAB M, BOTTOU L, LAPTEV I, et al. Is object localization for free? -weakly-supervised learning with convolutional neural

networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015：685-694.

[19] DOSOVITSKIY A，BEYER L，KOLESNIKOV A，et al. An image is worth 16x16 words：Transformers for image recognition at scale[J]. arXiv：2010. 11929，2020.

[20] LIU Z，LIN Y，CAO Y，et al. Swin transformer：Hierarchical vision transformer using shifted windows[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021：10012-10022.

[21] HOWARD A G，ZHU M，CHEN B，et al. Mobilenets：Efficient convolutional neural networks for mobile vision applications[J]. arXiv：1704. 04861，2017.

[22] 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. 2016：770-778.

[23] CHEN T，KORNBLITH S，NOROUZI M，et al. A simple framework for contrastive learning of visual representations [C]// International Conference on Machine Learning. PMLR，2020：1597-1607.

[24] HE K，FAN H，WU Y，et al. Momentum contrast for unsupervised visual representation learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：9729-9738.

[25] CHEN X，FAN H，GIRSHICK R，et al. Improved baselines with momentum contrastive learning[J]. arXiv：2003. 04297，2020.

[26] KHOSLA P，TETERWAK P，WANG C，et al. Super-vised contrastive learning[J]. Advances in Neural Information Processing Systems，2020，33：18661-18673.

[27] HE J，CHEN J N，LIU S，et al. TransFG：A transformer architecture for fine-grained recognition[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2022：852-860.

[28] SONG Y，SEBE N，WANG W. On the Eigenvalues of Global Covariance Pooling for Fine-grained Visual Recognition[J]. arXiv：2205. 13282，2022.

[29] CHANG D，PANG K，ZHENG Y，et al. Your "Flamingo" is My "Bird"：Fine-Grained，or Not[C]// Computer Vision and Pattern Recognition. IEEE，2021.

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

[31] RAO Y，CHEN G，LU J，et al. Counterfactual attention learning for fine-grained visual categorization and reidentification[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021：1025-1034.

[32] DO T，TRAN H，TJIPUTRA E，et al. Fine-Grained Visual Classification using Self Assessment Classifier[J]. arXiv：2205. 10529，2022.

[33] WANG J，YU X，GAO Y. Feature fusion vision transformer for fine-grained visual categorization[J]. arXiv：2107. 02341，2021.

ZHENG Shijie，born in 1999，postgraduate candidate. His main research interests include fine-grained image classification and image segmentation.

WANG Gaocai，born in 1976，Ph.D，professor，Ph. D supervisor，is a senior member of China Computer Federation. His main research interests include computer network，performance evaluation and network security.

（责任编辑：何杨）