计算机科学
COMPUTER SCIENCE

**基于多分支注意力增强的细粒度图像分类**

张文轩，吴秦

引用本文

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 Wen-xuan, WU Qin. Fine-grained Image Classification Based on Multi-branch Attention-augmentation[J]. Computer Science, 2022, 49(5): 105-112.

---

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210100108

# 基于多分支注意力增强的细粒度图像分类

张文轩　吴　秦
江南大学人工智能与计算机学院　江苏 无锡 214122
江南大学江苏省模式识别与计算智能工程实验室　江苏 无锡 214122
(6181914045@stu.jiangnan.edu.cn)

**摘　要**　针对细粒度图像类间差距小、类内差距大的问题,文中提出以弱监督学习的方式使用多分支注意力增强卷积网络,从而实现细粒度图像分类。文中采用 Inception-V3 网络提取图像的基础特征,从中获取多个局部响应区域并进行特征融合,在此基础上采用注意力机制对图像关键区域进行自约束的局部裁剪和局部擦除,避免仅提取目标单个部位的特征,促使网络更加关注目标物体不同部位的细节特征,同时也提升了目标区域的定位精度。此外,文中提出中心正则化损失函数来约束训练过程中获取的注意力区域,以进一步提升目标定位精度和扩大图像特征的类间差距。在 3 个公开数据集上进行了实验,结果表明,所提方法取得了比当前最优方法更好的结果。

**关键词**:细粒度图像分类;弱监督学习;多分支注意力增强;卷积神经网络;中心正则化损失
**中图法分类号**　TP391

## Fine-grained Image Classification Based on Multi-branch Attention-augmentation

ZHANG Wen-xuan and WU Qin
School of Artificial Intelligence and Computer Science, Jiangnan University, Jiangsu, Wuxi 214122, China
Jiangsu Provincial Engineering Laboratory for Pattern Recognition and Computational Intelligence, Jiangnan University, Jiangsu, Wuxi 214122, China

**Abstract**　In order to address the challenges of high intra-class variances and low inter-class variances in fine-grained image classification, a multi-branch attention-augmented convolution neural network is proposed to solve the problem. The pre-trained Inception-V3 network is used to extract basic feature. In order to solve the problem that features are extracted from one part of an object and encourage the network to pay more attention to the discriminative features of different parts, we apply self-constrained attention-wised cropping and self-constrained attention-wised erasing on the central parts of the original images. It also improves the detection accuracy of object locations. Meanwhile, a central regularization loss function is proposed to constrain attention-augmented training process to obtain better attention regions and expand the gap between different classes of images. Comprehensive experiments on three benchmark datasets show that our approach surpasses the state-of-art works.

**Keywords**　Fine-grained image classification, Weakly supervised learning, Multi-branch attention-augmentation, Convolutional neural network, Central regularization loss

## 1　引言

细粒度图像分类,即识别不同物体的子类,是计算机视觉、模式识别等领域新兴的研究课题。其研究内容主要为识别同一基础类别(如鸟[1]、狗[2]、花[3]、车[4]、飞机[5]等)下细化程度更高的子类。如,生态保护中针对生物不同种、科、属的识别是必需的研究基础;城市管理领域中对不同型号车辆的细化分类可以作为交通检测与跟踪参考的依据。与其他计算机视觉任务类似,细粒度图像分类方法存在着很多普遍性问题,例如图像的光照不均、场景差异大、尺度及视角多变。但是细粒度图像分类最大的挑战来自于其类间差距小、类内差距大。如图 1 所示,图像实例为取自 Stanford Dogs 数据集和 CUB-200-2011 数据集的图像。图中黑嘴杜鹃和黄嘴美洲鹃的区别仅仅在于鸟喙的颜色,这就要求细粒度分类模型对不同类别间的细微差异具有高度敏感性。在同一个子类中,由于目标个体的差异,甚至不同位置、不同角度的变化,不同个体会呈示出差异性(如图 1 所示,由于姿态差异,不同图像中的巴塞犬有明显差异)。

到稿日期:2021-01-14　返修日期:2021-04-21
基金项目:国家自然科学基金(61972180)
This work was supported by the National Natural Science Foundation of China(61972180).
通信作者:吴秦(qinwu@jiangnan.edu.cn)


图 1 细粒度图像实例

Fig. 1 Examples of fine-grained images

传统的机器学习方法往往很难解决上述问题，在早期发布的细粒度图像分类数据集 CUB-200-2011 的技术报告中，Wah 等[1]采用词包(bag of Words)和支持向量机模型对该数据集中的图像进行分类，分类正确率仅为 17.3 %。随着深度卷积神经网络在 ImageNet 上的成功，深度卷积特征[6-7]逐渐代替传统特征提取方法[8-9]成为主流。Donahue 等[10]通过对在 ImageNet 数据集上训练的卷积网络模型进行分析发现，卷积网络提取的特征具有更强大的语义特性和更优良的迁移能力。与此同时，越来越多的卷积神经网络模型被提出，如 ResNet101[11]、VGG19[12]、Inception 系列[13-14]，这些网络为解决问题提供了广泛的思路。

近年来，一些基于卷积神经网络的方法在细粒度图像分类任务上达到了较高的精度[15-18]。这些模型一般通过提议区域候选框(region Proposal)的方法在图像上提取上千个候选框，进而筛选出目标可能存在的关键区域，在一定程度上区分出了前景目标和背景，但是目标局部特征容易在繁多的候选区域中被混淆，并且效果因目标而异。此外，由于需要对每张训练图片人工标注目标位置框，且深度学习需要的训练图像数量巨大，因此造成这些方法的训练成本高昂。

为了解决细粒度图像类间差距小、类内差距大导致的识别率不高以及人工标注成本高的问题，本文提出了一种基于多分支注意力增强机制的卷积神经网络，该网络能够精准定位目标关键区域并获得较为理想的分类能力，训练中仅需类别标签而无需标注目标位置框。该方法的优势如下：

(1)利用生成的局部响应图来反映模型所关注的目标局部位置，为之后的特征融合提供不同部位重要性的权重，采取网络局部优先的特征选择策略，使网络对目标局部特征更为敏感。

(2)设计了两大注意力增强分支，用于引导注意力训练过程。对目标局部轮廓、位置进行针对性强化，去除背景的噪音信息，提供准确的目标位置，相比提议区域候选框的方法，本文方法能更轻松地生成目标局部区域。

(3)在 3 个公开的数据集上进行一系列的对比实验，结果表明，本文方法的分类准确率均高于其他方法。

## 2 相关工作

按照网络训练过程中对监督信息依赖程度的强弱，细粒度分类网络可以分为“基于强监督信息的分类网络”和“基于弱监督信息的分类网络”两类。

基于强监督信息的分类模型除了需要图像类别标签监督样本训练以外，还需要提供目标边界框和部位标注点等额外的人工标注信息进行补充监督。Zhang 等[19]提出的基于部位的 R-CNN 用区域提议框同时检测目标整体位置和各个局部区域，并通过整体和局部区域的空间几何关系来训练网络。Mask R-CNN[20]使用全卷积网络[21]对生成的每一个候选区域进行分类回归，筛选出可能的目标局部区域。Ge 等[22]提出的姿态正则化卷积神经网络在图像不同的卷积层提取各个层次的局部区域，并进行姿态对齐。尽管基于强监督信息的分类方法能够充分地利用目标信息，但过于依赖目标边界框和部位标注点这类额外监督信息，提高了实现算法的门槛，限制了其实际应用。

基于弱监督信息的分类模型仅仅依赖于类别标签来完成分类。由于没有额外的人工标注信息，因此早期的弱监督图像分类方法使用全局平均池化生成特定于类的类激活区域，借此反映目标可能的位置。如何提升网络对图像目标的关注程度，成为了研究人员意图突破的难题。Xiao 等[23]提出的两级注意力网络旨在提取出两个不同层次的特征，即目标级(object-level)特征和局部级(part-level)特征，它们分别对应强监督学习中所使用的目标边界框和局部位标注点。Lin 等[24]提出了双线性卷积神经网络，设计了目标定位和目标识别两个子网络对细粒度图像进行分类。LIO 网络[25]通过目标内在结构信息作为额外监督的方式来预测目标内部的相对位置，以提升分类效果。FDL 网络[26]利用知识蒸馏筛选出图像中最具区分度的区域候选框，提升系统对重要区域的关注程度。SnapMix[27]利用类激活图减少细粒度图像数据扩充时的标签噪音，从而获得质量较好的增强数据集。基于弱监督信息的分类有训练成本低、算法速度快以及目标定位策略灵活的优点。但是，目标定位的正确性和局部特征提取能力的优劣程度极大地影响了这类方法的性能。与上述工作相比，本文方法结合了自约束的注意力增强分支和中心正则化损失函数来进一步提高图像中目标的定位准确度，以提取更有代表性的局部特征，从而提升细粒度图像的分类性能。

## 3 基于多分支注意力增强的卷积网络框架

本文提出的基于多分支注意力增强的卷积网络框架如图 2 所示，其由主干网络、局部响应特征融合模块、自约束注意力增强分支 3 个模块组成。作为主干网络的 Inception-V3 主要用于提取图像的基础特征；局部响应特征融合则在保留原图的整体特征的同时，增加了目标各个局部区域的特征权重，使网络更关注于目标物体的局部特征；注意力增强分支通过类似对抗的训练方式来凸显具有区分度的特征，以强化局部特征的捕捉能力。此外，我们还提出在损失函数中增加中心正则化损失来校正弱监督学习中注意力区域的位置。

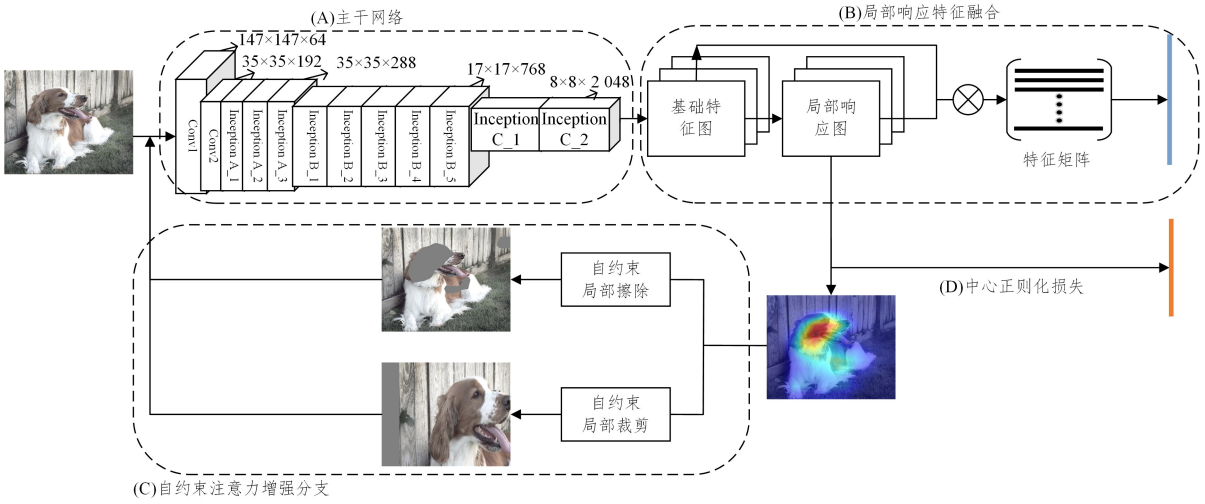



图 2　多分支注意力增强卷积网络的框架

Fig. 2　Overview of multi-branch attention-augmented convolution network

### 3.1　主干网络

细粒度图像目标位置、姿态、角度的差异，导致固定大小的卷积核很难适应细粒度图像分类中目标的多变性，因此特征的提取至关重要。相比 ResNet 和 VGG 网络，Inception-V3[13]使用了多分支的卷积结构，不同大小的卷积核能够兼顾图像全局语义特征与局部细节特征的提取。此外，Inception-V3 将 $N\times N$ 的卷积分解为 $1\times N$ 和 $N\times 1$ 的非对称卷积，在增加特征多样性的同时有效减少计算量。因此，我们选择 Inception-V3 网络作为主干网络，其包含 3 个 Inception A 模块、5 个 Inception B 模块和 2 个 Inception C 模块，具体结构如图 3—图 5 所示。

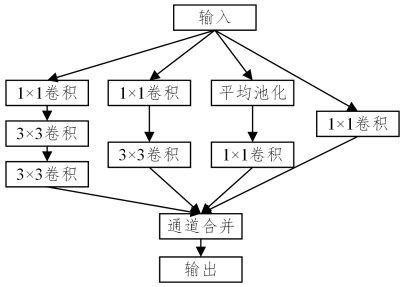



图 3　Inception A 模块的结构

Fig. 3　Structure of Inception A

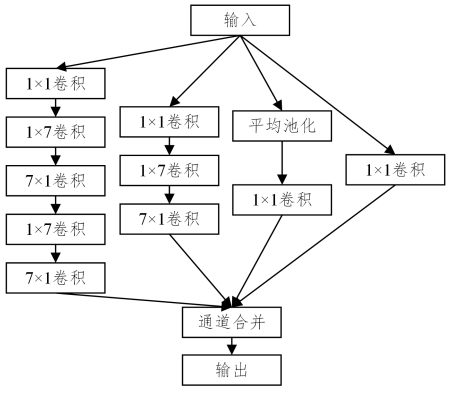



图 4　Inception B 模块的结构

Fig. 4　Structure of Inception B

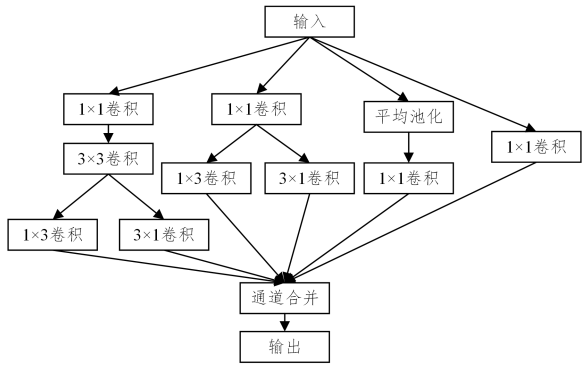



图 5　Inception C 模块的结构

Fig. 5　Structure of Inception C

3 个 Inception A 模块通过 1×1，3×3，5×5（分解为两个 3×3）大小的卷积运算对应不同的特征图区域，在降低参数计算量的同时保留了图像的浅层特征；Inception B 模块数量更多，引入的卷积核尺寸更大，避免了特征图严重压缩导致的特征表示瓶颈；2 个 Inception C 模块既起到稳定网络深层特征的作用，又为下文的自约束策略的方法提供了可行性。

### 3.2　局部响应特征融合

细粒度图像不同类别之间的差异往往体现在目标的各个局部区域。如果网络能够对这些不同的部位进行特征提取并对图像分类影响较大的部位所对应的局部区域给予较大的特征权重，则可以提升网络的分类能力[28]。基于此，我们提出局部响应特征融合方法。首先，用 1×1 的卷积操作来生成对应不同部位信息的局部响应图。如图 6 所示，假设使用主干网络提取出的基础特征图为 $F\in R^{H\times W\times C}$（如图 2 中的 Inception C_2），使用 1×1 卷积操作来生成各个通道的局部响应图，如式(1)所示：

$$P_k=Conv_{1\times 1}^{k}(F),k=1,2,\cdots,K \tag{1}$$

其中，$F\in R^{H\times W\times C}$ 代表基础特征图，$P_k\in R^{H\times W}$ 代表物体的第 $k$ 个特征图，该特征图主要提取了目标第 $k$ 部分的位置信息和权重，$K$ 为局部响应图的通道数量，本文实验将其设置为 32。

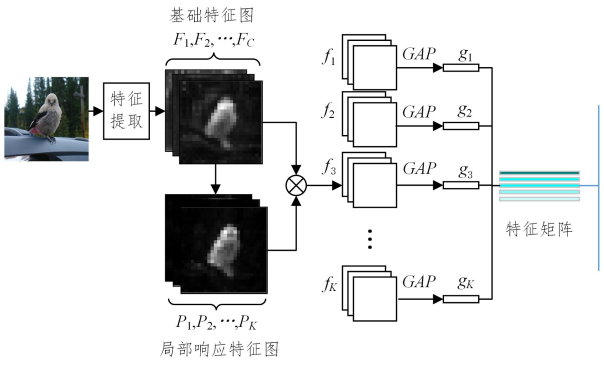



图 6 局部响应特征的融合流程

Fig. 6 Process of parts-activated feature fusion

首先将基础特征图 $F$ 与目标第 $k$ 个部分的局部响应图 $P_k$ 进行按元素位相乘操作，获得融合后的特征 $f_k=P_k\otimes F$，然后全局池化自适应特征融合的特征图 $f_1,f_2,\cdots,f_K$，获得相应的局部响应特征 $g_1,g_2,\cdots,g_K$，最后将这些特征经过拼接操作获得特征矩阵。其过程如式(2)所示：

$$\boldsymbol{\Phi}=\begin{bmatrix}GAP[P_1\otimes F]\\GAP[P_2\otimes F]\\\cdots\\GAP[P_K\otimes F]\end{bmatrix}=\begin{bmatrix}GAP[f_1]\\GAP[f_2]\\\cdots\\GAP[f_K]\end{bmatrix}=\begin{bmatrix}g_1\\g_2\\\cdots\\g_K\end{bmatrix}\tag{2}$$

其中，$\boldsymbol{\Phi}$ 为得到的特征矩阵，$GAP[*]$ 为全局平均池化操作，$\otimes$ 为按位相乘的操作。

特征图和局部响应图的特征融合模块能够计算不同空间位置的特征，而全局平均池化则可得到全局特征。这样的融合操作捕获了特征通道之间不同位置细节特征之间的关系，提供了比普通线性模型更强的特征。

### 3.3 自约束多分支注意力增强

现有的多分支结构网络大多通过对图像进行裁剪的方式来获取更好的目标特征。鉴于局部响应图可以反映目标局部注意力特征的位置，我们在原始图像上裁剪出对应局部响应图响应值较大的区域，从而挖掘细节特征，实现对目标部位的增强。然而，单纯裁剪的增强操作会使网络倾向于对目标的同一部位进行循环训练，可能导致网络对次要细节特征的不敏感或者是出现参数过拟合的现象。因此，我们使用图像擦除的增强方法来强迫网络学习其他部位的特征。一般的随机图像擦除策略擦除的区域可能是不相关的背景，增强图像质量低且效果差；对于小目标则会擦除整个目标，生成噪声数据。为了解决这些问题，本文在注意力局部裁剪分支之外，并列设计了注意力局部擦除分支，将局部响应图中响应值最高的部分擦除，避免网络过于关注目标某一局部特征而忽略其他细节特征，强化了网络的泛化能力。

此外，单纯以局部响应图上的阈值区域作为注意力增强的依据会导致部分样本在训练时裁剪、擦除的区域过小或过大，导致注意力增强效果降低。针对这一问题，我们提出在注意力增强分支中加入自约束策略，约束每一次裁剪、擦除的区域，通过对不同网络层输出的裁剪、擦除区域进行训练来提取更加有效的空间特征，以获得更加准确的局部增强区域。

结合上述两个方面，本文提出了两大自约束注意力增强分支——注意力局部裁剪分支和注意力局部擦除分支。

对于训练数据，我们在局部响应图 $P\in R^{H\times W\times K}$ 中随机挑选一张 $P_k$ 来指导注意力增强分支的训练过程。首先对 $P_k$ 作归一化处理，其具体过程如式(3)所示：

$$P_k^*(u,v)=\frac{P_k(u,v)-\min(P_k(u,v))}{\max(P_k(u,v))-\min(P_k(u,v))}\tag{3}$$

其中，$P_k(u,v)$ 表示局部响应图 $P_k$ 在第 $u$ 行第 $v$ 列位置的响应值，$P_k^*(u,v)$ 作为局部注意力裁剪和局部注意力擦除区域的判断依据。

在局部注意力裁剪分支上，当 $P_k^*(u,v)$ 大于裁剪阈值 $\delta_{\text{crop}}=random(0.4,0.6)$ 时，$M_k^{\text{crop}}(u,v)=1$；否则 $M_k^{\text{erase}}(u,v)=0$。然后寻找一组最小的边界框来覆盖 $M_k^{crop}(u,v)=1$ 的所有位置，并将其放大到原图像大小，以呈现更多的细节信息。

在局部注意力擦除分支上，当 $P_k^*(u,v)$ 小于裁剪阈值 $\delta_{\text{erase}}=\text{random}(0.2,0.5)$ 时，$M_k^{\text{erase}}(u,v)=1$；否则 $M_k^{\text{crop}}(u,v)=0$。即将输入图像中高于阈值的相应像素区域值置为 0，得到擦除图像，以解决网络的多张特征图关注目标同一部位的问题。

在局部注意力裁剪和擦除操作之后，结合不同尺寸网络层特征的自约束机制可以使网络模型逐渐聚焦到目标的局部区域。如图 7 所示，若将 Inception $C_2$ 作为输出的特征图，则将产生的相应的局部响应图记为 $P_k|_{\text{Inception C}_2}$，将对应的裁剪区域记为 $M_k^{\text{crop}}|_{\text{Inception C}_2}$，将擦除区域记为 $M_k^{\text{erase}}|_{\text{Inception C}_2}$。同时，将利用 Inception $C_1$ 提取的新的局部响应图记为 $P_k|_{\text{Inception C}_1}$，将对应的裁剪区域记为 $M_k^{\text{crop}}|_{\text{Inception C}_1}$，将擦除区域记为 $M_k^{\text{erase}}|_{\text{Inception C}_1}$，使用式(4)、式(5)得到新的更加精确的注意力增强区域 $\widetilde{M_k^{\text{crop}}}$ 和 $\widetilde{M_k^{\text{erase}}}$。

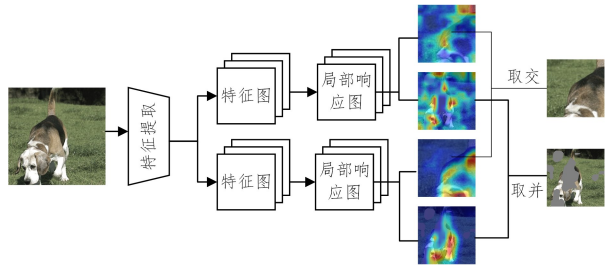



图 7 自约束的注意力增强分支

Fig. 7 Self-constrained attention-augmented branches

$$\widetilde{M_k^{\text{crop}}}=M_k^{\text{crop}}|_{\text{Inception C}_1}\cap M_k^{\text{crop}}|_{\text{Inception C}_2}\tag{4}$$

$$\widetilde{M_k^{\text{erase}}}=M_k^{\text{erase}}|_{\text{Inception C}_1}\cup M_k^{\text{erase}}|_{\text{Inception C}_2}\tag{5}$$

自约束策略使注意力增强分支在对训练数据提取细节特征的同时，能筛选出更精确的局部信息，并尽可能地摒弃样本噪声，而且对于内部变化大的样本能够降低类内差异，提升裁剪与擦除分支的性能。

### 3.4 损失函数

分类损失函数是决定卷积神经网络模型性能是否优越的关键，当前主流的细粒度分类模型损失函数大多使用 softmax 损失函数，如式(6)所示：

$$L_{\text{softmax}}=-\sum_{i=1}^{m}\log\frac{e^{W_{y_i}^{T}x_i+b_{y_i}}}{\sum_{j=1}^{s}e^{W_j^{T}x_i+b_j}} \tag{6}$$

其中，$m$ 表示一个批次送入模型训练的图像数量，$y_i$ 表示第 $i$ 张图像的真实类别标签，$\boldsymbol{x}_i$ 表示第 $i$ 张图像放入最后全连接层前的特征向量，$W$ 代表网络的最后一个全连接层，$b$ 代表网络偏置，$s$ 代表目标类别的数量。

损失函数的重要作用是修正卷积神经网络中的权重，使得提取的图像特征可以对图像中的目标进行正确分类。对于细粒度图像分类而言，损失函数对图像类间差异与类内差异的关注程度起到了关键作用。softmax 损失函数反映了类间差异但未能很好地反映类内相似性。因此，我们有必要在 softmax 函数的基础上加入反映表达类内相似性的函数。受 centerloss 缩短类内距离方法的启发，我们提出中心正则化损失函数。中心正则化损失函数的定义如式(7)所示：

$$L_{\text{center}}=\frac{1}{2}\sum_{i=1}^{m}\sum_{k=1}^{K}\left\|g_k(I_i)-c_k(y_i)\right\|_2^2 \tag{7}$$

其中，$K$ 表示图像中每类目标的部位数量，$g_k(I_i)$ 表示第 $i$ 张图像 $I_i$ 第 $k$ 个部位的局部响应特征，$c_k(y_i)\in R^{1\times N}$ 表示第 $i$ 张图像 $I_i$ 所属类别 $y_i$ 在第 $k$ 个部位的特征中心。中心正则化损失函数计算了每个样本的每个部位到对应部位的类别中心的距离。$L_{\text{center}}$ 的值越小，说明同一类别图像在同一部位的特征差异越小，即类内距离越小。综合分类损失函数 $L_{\text{softmax}}$ 和中心正则化损失 $L_{\text{center}}$，我们提出的网络模型最终的损失函数如式(8)所示：

$$L_{\text{total}}=L_{\text{softmax}}+\alpha * L_{\text{center}} \tag{8}$$

其中，$\alpha$ 是两种损失函数的线性组合系数。

整个网络中的参数计算采用小批量随机梯度下降法，更新每个类别的每个部位的特征中心和利用损失函数迭代计算网络参数的过程如算法 1 所示。算法 1 中，$\alpha$ 为损失函数线性组合系数，$\varepsilon$ 为学习率，$\beta$ 为中心正则化损失训练速度控制参数，本文实验中将其设置为 0.95。

**算法 1** 利用损失函数计算网络参数的基本过程

输入：一个批次 m 张图像$\{I_i\}$及对应的标签$\{\text{label}(I_i)\}$；每张图像对应的第 k 个部位的局部响应特征$g_k(I_i)$；网络参数 $\alpha,\varepsilon,\beta$；MAXITERATION

输出：网络参数 $\theta$，W

1. n←0
2. 初始化的网络参数 $\theta$，W
3. 初始化每 l 个类别第 k 个部位的特征中心
4. For k=1∶K
5. For l=1∶L
6. $c_k^{(n)}(l)\leftarrow 0$
7. While n<MAXITERATION
8. n←n+1
9. 计算损失函数值 $L_{\text{total}}^{(n)}=L_{\text{softmax}}^{(n)}+\alpha * L_{\text{center}}^{(n)}$
10. 计算反向传播误差 $\frac{\partial L_{\text{softmax}}^{(n)}}{\partial x_i^{(n)}}$ 和 $\frac{\partial L_{\text{center}}^{(n)}}{\partial g_k^{(n)}(I_i)}$
11. 更新权重 $W^{(n+1)}=W^{(n)}-\varepsilon^{(n)}\cdot\frac{\partial L_{\text{total}}^{(n)}}{\partial x_i^{(n)}}=W^{(n)}-\varepsilon^{(n)}\cdot\frac{\partial L_{\text{softmax}}^{(n)}}{\partial x_i^{(n)}}$
12. 更新各个类别各个部位的中心特征
13. For k=1∶K
14. For l=1∶L
15. $\Delta c_k(l)=\frac{\sum_{i\in\{1,\cdots,m\}\cap\{y_i=l\}}\left[c_k^{(n)}(l)-x_i^{(n)}\right]}{1+\sum_{i\in\{1,\cdots,m\}\cap\{y_i=l\}}1}$
16. $c_k^{(n+1)}(l)=c_k^{(n)}(l)-\beta\cdot\Delta c_k(l)$，更新卷积层参数
17. $\theta^{(n+1)}=\theta^{(n)}-\varepsilon^{(n)}\cdot\left(\sum_{i}^{m}\frac{\partial L_{\text{softmax}}^{(n)}}{\partial x_i^{(n)}}\cdot\frac{\partial x_i^{(n)}}{\partial\theta^{(n)}}+\alpha\sum_{i}^{m}\sum_{k=1}^{K}\frac{\partial L_{\text{center}}^{(n)}}{\partial g_k^{(n)}(I_i)}\cdot\frac{\partial g_k^{(n)}(I_i)}{\partial\theta^{(n)}}\right)$
18. End while

## 4 实验

### 4.1 数据集

本文在 3 个公开的数据集上验证本文方法的有效性，下面简单介绍这 3 个数据集。

(1)CUB-200-2011。该数据集包含 200 种不同的鸟类，共计 11 788 张图像，每张图像标注了 15 个部位标注点、1 个目标边界框。其中，训练数据集有 5 994 张图像，测试集有 5 794 张图像。

(2)Stanford Dogs。该数据集提供了 120 种狗类图像，每类约 150 张图像，共计有 20 580 张图像，其中 12 000 张用于训练，8 580 张用于测试。

(3)Stanford Cars。该数据集基于汽车出产时间、品牌和车型将汽车图像分为 196 类，共有 16 185 张图像，其中训练图像和测试图像分别为 8 144 张和 8 041 张。

### 4.2 实验细节

#### 4.2.1 数据增强

针对每一张训练图像，多分支注意力增强网络通过注意力增强的两个分支产生局部裁剪、局部擦除两张增强图像，相当于将训练数据集扩充至 3 倍，从而满足深度学习对训练图像的数量需求。

#### 4.2.2 实验设置

本文代码基于 Tensorflow 深度学习框架来实现，主干网络部分的参数采用 ImageNet 上预训练的结果来进行初始化，该网络使用训练动量为 0.9 的随机梯度下降优化器。初始学习率设置为 0.0001，在训练过程中学习速率设置为 $1\times10^{-5}$。在测试过程中，保持测试图像的长宽比，通过调整图像的高度来调整输入图像的尺寸。

#### 4.2.3 评价指标

本文使用分类准确度 Accuracy 来检验细粒度图像分类模型的训练效果，能直观地反映模型的分类性能，其计算方式如式(9)所示：

$$Accuracy=\frac{\text{分类预测正确的图像数量}}{\text{总测试样本数量}} \tag{9}$$

### 4.3 消融实验

为了验证本文提出的各个模块是否能有效地提升网络性能，我们在 Stanford Dogs 数据集上做了 4 组对比实验，实验设置和结果如表 1 所列。

表 1 Stanford Dogs 数据集上的消融实验结果

Table 1 Results of ablation experiment on Stanford Dogs

| | 实验 1 | 实验 2 | 实验 3 | 实验 4 |
|---|---|---|---|---|
| Inception-V3 主干网络 | √ | √ | √ | √ |
| 局部响应特征融合模块 | | √ | √ | √ |
| 注意力增强分支（无约束） | | | √ | |
| 注意力增强分支（自约束） | | | | √ |
| 准确度/% | 88.9 | 91.82 | 92.21 | 92.63 |

实验 1 中使用 Inception-V3 网络作为主干网络提取特征，并使用 Inception-V3 网络最后一层全连接层进行分类。实验 2 在实验 1 的基础上加入局部响应特征融合模块，实验 2 的结果比实验 1 提升了 2.92%，说明局部响应特征融合通过提取目标不同部位的特征，可以更精准地提取不同类别目标的细节特征，从而提升细粒度图像的分类正确率。

实验 3 在实验 2 的基础上采用无约束的注意力增强分支（即没有引入 Inception C_1 约束注意力增强区域）。与实验 2 相比，实验 3 的结果有小幅度提升，说明注意力增强分支可以在一定程度上扩张训练数据，并引导网络关注各部位的细节区域。

实验 4 在实验 2 的基础上采用了自约束的注意力增强分支，其结果比完全不采用注意力增强分支提升了 0.81%，与采用无约束的注意力增强分支相比，提升了 0.42%。我们认为，其主要原因在于自约束方法可有效避免无约束数据增强中增强效果差的情况或噪声的产生。

### 4.4 参数优化

本文在传统的损失函数的基础上增加了中心正则化损失来优化训练过程。为了确定式(8)中 $\alpha$ 的值，我们在 Stanford Dogs 数据集上对不同的 $\alpha$ 值进行了实验，实验结果如图 8 所示。

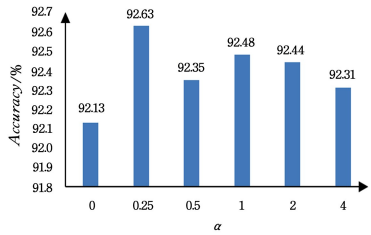



图 8 Stanford Qogs 上不同 $\alpha$ 取值的对比结果

Fig. 8 Comparison of results with different values of $\alpha$ on Stanford Dogs

由图 8 可知，当多分支注意力增强网络训练添加中心正则化损失时（$\alpha>0$），结果比仅使用 softmax 损失函数（对应 $\alpha=0$）时更好，原因在于本文提出的中心正则化损失将目标各个部位的位置误差计入损失函数，通过纠正各部位的位置信息使得目标定位更精准，提取的特征也就更精确，从而提升了模型的分类性能。当 $\alpha=0.25$ 时分类效果最好，在下面的实验中，$\alpha$ 的取值均为 0.25。

### 4.5 算法比较

为了验证本文方法的优越性，我们在 4.1 节描述的 3 个数据集上进行了实验，并与最新的一些方法进行了比较。实验结果如表 2－表 4 所列，相比当前最优的结果，本文方法在 3 个数据集上的分类准确率均有所提升。在 Stanford Dogs 数据集上，本文方法的正确率提升最为明显，比最新方法（SEF）提升了 3.83%。

表 2 CUB-200-2011 上的对比实验结果

Table 2 Comparison with state-of-the-art methods on CUB-200-2011

（单位：%）

| 方法 | 准确率 |
|---|---|
| MA-CNN[29] | 86.5 |
| MAMC[30] | 86.5 |
| NTS-Net[31] | 87.5 |
| SEF[32] | 87.3 |
| DCL[33] | 87.8 |
| DB[15] | 88.6 |
| WS-DAN[34] | 89.2 |
| LIO[25] | 88.0 |
| FDL[26] | 89.09 |
| SnapMix[27] | 89.32 |
| Ours | **89.67** |

表 3 Stanford Dogs 上的对比实验结果

Table 3 Comparison with state-of-the-art methods on Stanford Dogs

（单位：%）

| 方法 | 准确率 |
|---|---|
| DVAN[35] | 87.1 |
| PC[36] | 83.8 |
| MAMC[30] | 85.2 |
| DB[15] | 87.7 |
| FDL[26] | 85.5 |
| SEF[32] | 88.8 |
| Ours | **92.63** |

表 4 Stanford Cars 上的对比实验结果

Table 4 Comparison with state-of-the-art methods on Stanford Cars

（单位：%）

| 方法 | 准确率 |
|---|---|
| MA-CNN[29] | 92.8 |
| MAMC[30] | 93.0 |
| NTS-Net[31] | 93.9 |
| SEF[32] | 94.0 |
| DCL[33] | 94.5 |
| WS-DAN[34] | 94.5 |
| DB[15] | 94.9 |
| LIO[25] | 94.5 |
| FDL[26] | 94.5 |
| SnapMix[27] | 95.0 |
| Ours | **95.21** |

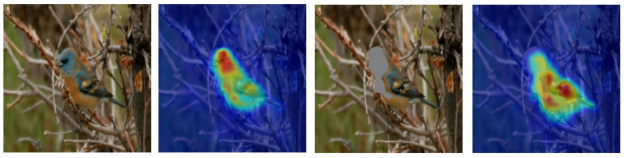
模型参数量是评价深度学习算法复杂度的重要指标。本文算法的参数量为 $26.2\times10^6$，在 3 个数据集上的结果与本文方法较为接近的算法（SEF 与 SnapMix 算法）的参数量如表 5 所列。本文方法不仅分类正确率优于 SEF 与 SnapMix，其参数量也更少。

表 5 模型参数量对比

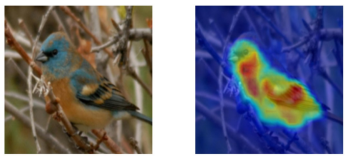
Table 5 Comparison on parameters of different models

| 模型 | 参数量 |
|---|---|
| SEF | $27.9\times10^6$ |
| SnapMix | $44.5\times10^6$ |
| Ours | $26.2\times10^6$ |

为了显示本文提出的注意力增强机制对细粒度图像的分类效果，我们将本文方法各分支的注意力图进行可视化。图 9 以 CUB-200-2011 数据集的一幅图像为例，给出了注意力增强的具体过程。图 9(a)为输入图像；图 9(b)为局部响应特征融合后的注意力图，该阶段的网络能够定位目标的整体位置，但是其注意力过于集中在头部，导致身体部分的权重较小，尾部没有呈现；图 9(c)是对注意力值较大的区域(鸟的头部)进行擦除后的图像；图 9(d)是对擦除鸟头后的图像进行注意力提取而获得的注意图，显而易见，擦除头部迫使网络关注鸟的头部以外的其他部位的特征；图 9(e)为图 9(a)裁剪放大后的增强图像；图 9(f)为图 9(e)的注意力图，裁减后的图像去除了原始图像中的复杂背景干扰，可以更好地提取鸟的整体特征和各个部位的细节特征。正是因为本文方法可以更好地定位和提取目标各个部位的细节特征，才使得本文方法取得了比其他方法更好的分类结果。



(a) (b) (c) (d)



(e) (f)

图 9 各分支的可视化结果

Fig. 9 Visualization results of each branch

**结束语** 本文通过 Inception-V3 网络提取图像中的高层语义特征，将其作为图像的基础特征，利用网络模型中生成的局部响应图来强化显著区域并对目标的不同部位赋予不同的权重，从而获取更加具有辨别性的目标局部区域。此外，多分支注意力增强通过自约束策略筛选出更准确、更丰富的细节特征，以提升网络的分类能力。在公开的细粒度图像数据集上的实验结果表明，本文方法作为端对端的弱监督学习模型，有较强的实用性和鲁棒性。虽然本文方法的正确率相比已有方法有所提升，但是我们发现，一些子类的误分类率仍旧较高。因此，在下一步的工作中，我们将重点研究如何进一步优化网络模型，以提取这些子类的有效特征，降低误分类率，从而提升整个数据集的分类正确率。

## 参 考 文 献


[1] WELINDER P, BRANSON S, MITA T, et al. The Caltech-UCSD Birds-200-2011 Dataset[R]. California Institute of Technology, 2011: 1-15.

[2] RABIEE H, HADDADNIA J, MOUSAVI H, et al. Novel dataset for fine-grained abnormal behavior understanding in crowd [C]// IEEE International Conference on Advanced Video & Signal Based Surveillance. 2016: 121-130.

[3] YANG W G, HUAI Y J. Flower Image Enhancement and Classification Based on Deep Convolution Generative Adversarial Network[J]. Computer Science, 2020, 47(6): 176-179.

[4] 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.

[5] MAJI S, RAHTU E, KANNALA J, et al. Fine-Grained Visual Classification of Aircraft[C]// IEEE International Conference on Advanced Video & Signal Based Surveillance. 2013: 1-6.

[6] PERRONNIN F, DANCE C. Fisher Kernels on Visual Vocabularies for Image Categorization[C]// 2007 IEEE Conference on Computer Vision and Pattern Recognition. 2007: 1-8.

[7] SÁNCHEZ J, MENSINK T, VERBEEK J. Image Classification with the Fisher Vector: Theory and Practice[J]. International Journal of Computer Vision, 2013, 105(1): 222-245.

[8] LOWE D G. Object recognition from local scale-invariant features[C]// Proceedings of the Seventh IEEE International Conference on Computer Vision. 1999: 1150-1157.

[9] DALAL N, TRIGGS B. Histograms of Oriented Gradients for Human Detection[C]// IEEE Computer Society Conference on Computer Vision & Pattern Recognition. 2005.

[10] DONAHUE J, JIA Y Q, VINYALS O, et al. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition [C]// Proceedings of the 31st International Conference on Machine Learning. PMLR, 2014: 647-655.

[11] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 770-778.

[12] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]// ICLR. 2015: 1-14.

[13] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the Inception Architecture for Computer Vision[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 2818-2826.

[14] XIE L, HUANG C. A Residual Network of Water Scene Recognition Based on Optimized Inception Module and Convolutional Block Attention Module[C]// 2019 6th International Conference on Systems and Informatics (ICSAI). 2019: 1174-1178.

[15] SUN G, CHOLAKKAL H, KHAN S, et al. Fine-Grained Recognition: Accounting for Subtle Differences between Similar Classes[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(1): 12047-12054.

[16] TAN M, WANG G, ZHOU J, et al. Fine-Grained Classification via Hierarchical Bilinear Pooling With Aggregated Slack Mask [J]. IEEE Access, 2017, 7(1): 117944-117953.

[17] YAO B, BRADSKI G, LI F F. A codebook-free and annotation-free approach for fine-grained image categorization[C]// 2012 IEEE Conference on Computer Vision and Pattern Recognition. 2012: 3466-3473.

[18] CHERIYADAT A M. Unsupervised Feature Learning for Aerial Scene Classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2014, 52(1): 439-451.

[19] ZHANG N, DONAHUE J, GIRSHICK R, et al. Part-based R-CNNs for Fine-grained Category Detection[C]// European Conference on Computer Vision (ECCV). 2014: 834-849.

[20] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 386-397.

[21] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015: 3431-3440.

[22] GE W, LIN X, YU Y. Weakly Supervised Complementary Parts Models for Fine-Grained Image Classification From the Bottom Up[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 3029-3038.

[23] XIAO T J, XU Y C, YANG K Y, et al. The application of two-level attention models in deep convolutional neural network for fine-grained image classification[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015: 842-850.

[24] LIN T, ROYCHOWDHURY A, MAJI S. Bilinear CNN Models for Fine-Grained Visual Recognition[C]// 2015 IEEE International Conference on Computer Vision (ICCV). 2015: 1449-1457.

[25] ZHOU M, BAI Y, ZHANG W, et al. Look-Into-Object: Self-Supervised Structure Modeling for Object Recognition[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 11771-11780.

[26] LIU C, XIE H, ZHA Z J, et al. Filtration and Distillation: Enhancing Region Attention for Fine-Grained Visual Categorization [C]// AAAI Conference on Artificial Intelligence. 2020: 11555-11562.

[27] HUANG S, WANG X, DAO D. SnapMix: Semantically Proportional Mixing for Augmenting Fine-grained Data [C]// AAAI Conference on Artificial Intelligence. 2021: 1-8.

[28] WU J, XU J, DING T. Fine-grained Image Classification Algorithm Based on Ensemble Methods of Transfer Learning[J]. Journal of Chongqing University of Posts and Telecommunications(Natural Science Edition), 2020, 32(3): 452-458.

[29] ZHENG H, FU J, MEI T, et al. Learning Multi - attention Convolutional Neural Network for Fine-Grained Image Recognition [C]// 2017 IEEE International Conference on Computer Vision (ICCV). 2017: 5219-5227.

[30] SUN M, YUAN Y, ZHOU F, et al. Multi-Attention Multi-Class Constraint for Fine-grained Image Recognition[C]// European Conference on Computer Vision(ECCV). 2018: 834-850.

[31] YANG Z, LUO T, WANG D, et al. Springer International Publishing Learning to Navigate for Fine-Grained Classification [C]// European Conference on Computer Vision(ECCV). 2018: 438-454.

[32] LUO W, ZHANG H, LI J, et al. Learning Semantically Enhanced Feature for Fine-Grained Image Classification[J]. IEEE Signal Processing Letters, 2020, 27: 1545-1549.

[33] CHEN Y, BAI Y, ZHANG W, et al. Destruction and Construction Learning for Fine-Grained Image Recognition[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 5152-5161.

[34] HU T, QI H. See Better Before Looking Closer: Weakly Supervised Data Augmentation Network for Fine-Grained Visual Classification[J/OL]. https://arxiv.org/abs/1901.09891.

[35] ZHAO B, WU X, FENG J, et al. Diversified Visual Attention Networks for Fine-Grained Object Classification [J]. IEEE Transactions on Multimedia, 2017, 19(6): 1245-1256.

[36] DUBEY A, GUPTA O, GUO P, et al. Pairwise Confusion for Fine-Grained Visual Classification[C]// European Conference on Computer Vision(ECCV). 2018: 71-88.

**ZHANG Wen-xuan**, born in 1997, master candidate, is a member of China Computer Federation. His main research interests include computer vision and machine learning.

**WU Qin**, born in 1978, Ph.D, associate professor, is a member of China Computer Federation. Her main research interests include computer vision and pattern recognition.

(责任编辑:喻藜)