第48卷 第2期
Vol.48 No.2
计 算 机 工 程
Computer Engineering
2022年2月
February 2022
·图形图像处理·
文章编号：1000-3428(2022)02-0237-06 文献标志码：A 中图分类号：TP391.41

结合双语义数据增强与目标定位的细粒度图像分类

谭 润，叶武剑，刘怡俊
（广东工业大学 信息工程学院，广州 514000）

摘 要：细粒度图像分类旨在对属于同一基础类别的图像进行更细致的子类划分，其较大的类内差异和较小的类间差异使得提取局部关键特征成为关键所在。提出一种结合双语义数据增强与目标定位的细粒度图像分类算法。为充分提取具有区分度的局部关键特征，在训练阶段基于双线性注意力池化和卷积块注意模块构建注意力学习模块和信息增益模块，分别获取目标局部细节信息和目标重要轮廓这2类不同语义层次的数据，以双语义数据增强的方式提高模型准确率。同时，在测试阶段构建目标定位模块，使模型聚焦于分类目标整体，从而进一步提高分类准确率。实验结果表明，该算法在CUB-200-2011、FGVC Aircraft和Stanford Cars数据集中分别达到89.5%、93.6%和94.7%的分类准确率，较基准网络Inception-V3、双线性注意力池化特征聚合方式以及B-CNN、RA-CNN、MA-CNN等算法具有更好的分类性能。

关键词：细粒度图像分类；数据增强；双线性网络；注意力学习；目标定位

开放科学（资源服务）标志码（OSID）：

中文引用格式：谭润，叶武剑，刘怡俊.结合双语义数据增强与目标定位的细粒度图像分类[J].计算机工程，2022，48(2)：237-242，249.

英文引用格式：TAN R，YE W J，LIU Y J.Fine-grained image classification combining dual semantic data augmentation and target location[J].Computer Engineering，2022，48(2)：237-242，249.

Fine-Grained Image Classification Combining Dual Semantic Data Augmentation and Target Location

TAN Run，YE Wujian，LIU Yijun
（School of Information Engineering，Guangdong University of Technology，Guangzhou 514000，China）

【Abstract】Fine-grained image classification aims to classify images of the same basic category into more specific subcategories. These images are characterized by large intra-class differences and minor inter-class differences，so the extraction of local key features is crucial to fine-grained image classification. A fine-grained image classification algorithm combining dual semantic data augmentation and target location is proposed.To extract discriminative local key features，two modules are constructed in the training phase to obtain two types of data at different semantic levels.The attention learning module is constructed based on Bilinear Attention Pooling（BAP）to obtain local detail information of the target，and the information gain module is constructed based on Convolutional Block Attention Module（CBAM）to obtain the important contour of the target.Then the accuracy of the model can be improved in the way of dual semantic data augmentation.At the same time，a target location module is built in the testing phase to make the model focus on the overall classification target and further improve the classification accuracy. The experimental results show that the proposed model displays a classification accuracy of 89.5% on CUB-200-2011 dataset，93.6% on FGVC Aircraft dataset and 94.7% on Stanford Cars dataset，delivering higher performance than benchmark network Inception-V3，Bilinear Attention Pooling（BAP）feature aggregation method，B-CNN，RA-CNN，MA-CNN and other algorithms.

【Key words】fine-grained image classification；data augmentation；bilinear network；attention learning；target location
DOI：10. 19678/j. issn. 1000-3428. 0060111

0 概述

细粒度图像分类是计算机视觉领域一项极具挑战和应用价值的研究课题，其在传统图像分类基础上进行更精细的图像类别子类划分，如区分鸟的种类等。区别于传统图像，细粒度图像均来自同一基本类别，不同子类图像间的差异较小，只通过目标整体轮廓往往无法取得良好的分类效果；而同一子类不同图像中又存在姿态、光照、背景遮挡等诸多影响因素，类内差异较大，因此，细粒度图像分类往往只

基金项目：广东省重点区域研究开发计划项目(2018B030338001)；广东工业大学青年百人项目(220413548)。
作者简介：谭 润(1996—)，男，硕士研究生，主研方向为细粒度图像分类；叶武剑(通信作者)，讲师、博士；刘怡俊，教授、博士。
收稿日期：2020-11-26 **修回日期：**2021-01-24 **E-mail：**1802934809@qq.com

能借助于极其细微的局部差异才能较好地完成分类。同时,细粒度图像数据库的获取和标注依赖于专家级别的知识,制作成本和时间成本昂贵。上述这些问题都给细粒度图像分类造成了极大的困难,使现有算法难以很好地完成分类任务。

细粒度图像分类的研究工作主要分为基于强监督信息和基于弱监督信息2个方向[1-2]。两者区别在于,基于强监督信息的算法需要引入额外的人工标注信息,如局部区域位置、标注框等,用于定位图像局部关键区域,而基于弱监督信息的算法仅依靠图像标签完成图像局部关键部位的定位和特征提取。目前研究思路主要分为2种:一是通过构建更具判决力的特征表征,适配于复杂的细粒度图像分类任务;二是在算法中引入注意力机制,通过注意力机制弱监督式地聚焦于部分局部区域,进一步提取特征,但仍存在定位不准确的问题。同时,细粒度图像中存在较多的遮挡,只通过提取少部分的局部关键特征,往往无法在所有同一类别图像上得到对应,不能达到良好的分类效果。

本文提出一种基于双语义增强和目标定位的细粒度图像分类算法。以双线性注意力池化(Bilinear Attention Pooling,BAP)方式构建注意力学习模块和信息增益模块提取双语义数据,并结合原图通过双语义数据增强的方式提高模型分类准确率。该算法一方面通过模块相互增益可控地学习图像中多个局部关键特征,另一方面分别获取2种语义层次数据,用于丰富模型训练数据,以双语义数据增强的方式辅助模型训练,同时在测试阶段构建目标定位模块,实现目标整体定位。

1 相关工作

目前,单独使用传统的卷积神经网络(如VGG[3]、ResNet[4]和Inception[5-6])无法很好地完成细粒度图像分类任务,因此,研究者通常在传统卷积神经网络的基础上进行基于强监督信息或基于弱监督信息方向的算法研究。

1.1 基于强监督信息的细粒度图像分类

基于强监督信息的细粒度图像分类算法需要利用训练数据集中已有的人工标注信息定位局部关键部位,再进一步提取特征。ZHANG等提出Part R-CNN算法[7],利用选择性搜索形成关键部位的候选框,通过目标检测R-CNN[8]算法对候选区域进行检测,挑选出评分值高的区域提取卷积特征用于训练SVM分类器。BRANSON等从分类目标姿态入手,提出姿态归一化CNN[9]。LIN等提出Deep LAC[10],在同一个网络中进行部件定位、对齐及分类,设计VLF函数用于Deep LAC中的反向传播。但基于强监督信息的算法所依赖的人工标注信息获取耗时且代价昂贵,导致该类算法实用性较差。

1.2 基于弱监督信息的细粒度图像分类

仅依靠图像标签信息完成细粒度图像分类任务成为近年来主要的研究方向。JADERBERG等提出时空卷积神经网络(ST-CNN)[11],在目标合适的区域进行适当的几何变换校正图像姿态。FU等提出循环注意力神经网络(RA-CNN)[12],在多尺度下递归式地预测注意区域的位置并提取相应的特征,由粗到细迭代地得到最终的预测结果。但该模型在同一时间只能关注于一个注意力区域,存在时间效率问题。ZHENG等提出多注意力卷积神经网络(MA-CNN)[13],通过构建一个部位分类子网络学习多个特征部位。但该模型在同一时间只能定位2~4个关键局部区域,这对于复杂的细粒度图像仍是不够的。

1.3 双线性网络

双线性网络也是弱监督算法的一种,与同类算法不同,其从高阶特征表达的角度出发,以外积汇合的方式聚合2个特征块。这种高阶特征间的交互作用适配于细粒度图像分类任务,如LIN等提出双线性CNN[14]和改进的双线性CNN[15]用于细粒度图像分类,LI等利用矩阵平方进一步改进双线性CNN[16]。但该类算法往往受限于较高的计算复杂度。HU等在双线性CNN的基础上提出双线性注意力池化方法[17],同时对原图采取注意力式剪切、注意力式丢弃,得到可以随着模型迭代更新变动的增强数据,这些数据和原图一起以数据增强的方式提高模型分类准确率。但该算法只利用了单一语义的数据增强方式,对于更复杂的细粒度图像任务仍存在缺少有效分类信息的问题。

为提取足够多的有区分度的局部关键特征,本文在训练阶段以双线性注意力池化的方式在网络不同深度构建注意力学习模块和信息增益模块,同时为提高模型中期特征表达能力,并行地在注意力学习模块和信息增益模块中分别引入卷积块注意模块(Convolutional Block Attention Mmodule,CBAM)。而在测试阶段,通过注意力学习模块和信息增益模块分别得到特征图,并以此构建目标定位模块用于聚焦图像中的目标整体,从而进一步提高分类准确率。

2 本文分类算法

本文算法的训练流程及网络模型如图1和图2所示,测试流程如图3所示。训练流程模块分别提

取两种语义层次的数据,以2种语义数据增强的方式辅助模型训练。

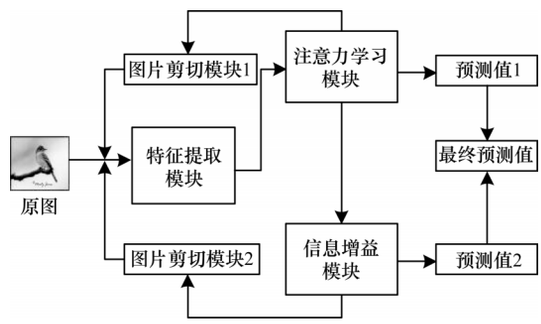

图1　训练流程

Fig.1　Training procedure

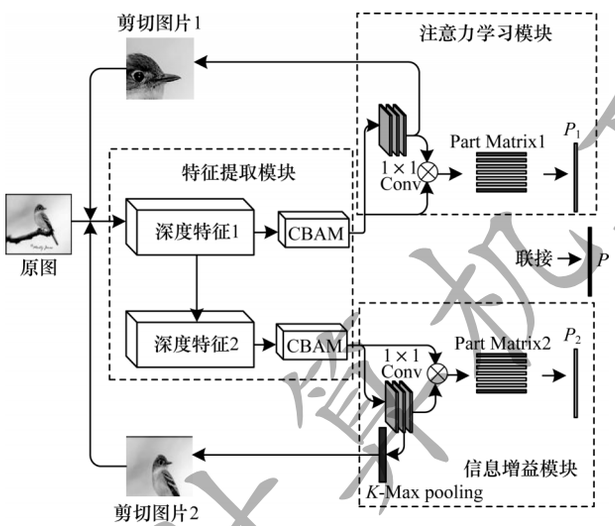

图2　网络模型框架

Fig.2　Framework of network model

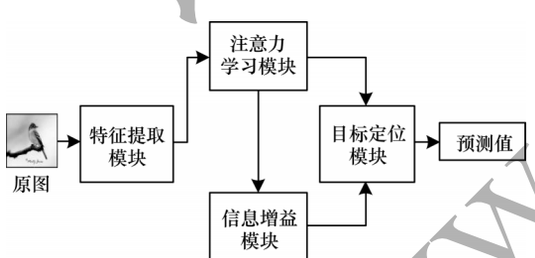

图3　测试流程

Fig.3　Testing procedure

2.1　第1类语义数据增强

图1中的注意力学习模块和图片剪切模块1用于第1类语义数据增强,其中注意力学习模块负责分类特征的学习,图片剪切模块1从分类特征中得到第1种语义类型的剪切图片辅助模型训练,该语义类型图片更关注于分类目标的局部细节信息。

1)注意力学习模块

如图1所示,模型首先从特征提取模块得到深度特征$f_1 \in \mathbb{R}^{C\times H\times W}$。对于细粒度图像分类任务,为使特征图有足够的特征表达能力同时增强特定区域的表征,从特征图本身出发,模型加入卷积块注意力模块(CBAM)[18],从通道维度和空间维度引入关注权重,提升特征图对关键局部区域的关注度。

(1)从通道维度引入关注权重。

对得到的初始深度特征f_1分别进行全局平均池化和全局最大池化,得到2个C维的池化特征,这2个池化特征均经过一个共享参数的多层感知器(Multi-Layer Perceptron,MLP),分别得到2个$1\times1\times C$维的通道关注权重,最后将其分别对应元素相加,经sigmoid激活函数激活得到最终的通道关注权重$M_c(f_1)$,如式(1)所示:

$$M_c(f_1)=\sigma(\mathrm{MLP}(\mathrm{avgpool}(f_1))+\mathrm{MLP}(\mathrm{maxpool}(f_1))) \tag{1}$$

将该权重与初始特征f_1相乘,得到通道关注特征$f_{1c}\in\mathbb{R}^{C\times H\times W}$,如式(2)所示:

$$f_{1c}=M_c(f_1)\otimes f_1 \tag{2}$$

(2)从空间维度引入关注权重。

对上一步得到的通道关注特征$f_{1c}\in\mathbb{R}^{C\times H\times W}$,沿着通道方向进行取平均(mean)和最大(max),得到2个维度为$1\times H\times W$的特征图,将这2个特征图进行维度拼接得到维度为$2\times H\times W$的特征图,最后本模型选择用一个卷积核大小为7×7的卷积层对其进行卷积操作,经sigmoid激活函数激活得到最终的空间关注权重$M_s(f_{1c})$,如式(3)所示:

$$M_s(f_{1c})=\sigma(f^{7\times7}([\mathrm{mean}(f_{1c});\ \max(f_{1c})])) \tag{3}$$

将该权重与特征f_2相乘,得到最终的聚焦特征$F_1\in\mathbb{R}^{C\times H\times W}$,如式(4)所示:

$$F_1=M_s(f_{1c})\otimes f_{2c} \tag{4}$$

通过上述过程,得到经过CBAM模块的聚焦特征F_1,之后模型采用双线性注意力汇合的思想,将聚焦特征F_1与其经过k个1×1卷积核后得到的注意力图A_{1k}以外积即点乘的形式汇合,从注意力图出发,使得到的分类特征的每一维代表分类目标中的一部分关键部位,最终得到注意力学习模块的分类特征P_1,如式(5)所示:

$$P_1=\begin{Bmatrix} g(A_{11}\cdot F_1)\\ g(A_{12}\cdot F_1)\\ g(A_{13}\cdot F_1)\\ \vdots\\ g(A_{1k}\cdot F_1)\end{Bmatrix} \tag{5}$$

其中:g为特征聚合函数。本文在该模块中采用全局平均池化方式为特征聚合函数聚合特征。

2)图片剪切模块1

从注意力学习模块得到注意力图$A_{1k}\in\mathbb{R}^{K\times H\times W}$,其中每一通道的注意力图代表分类目标的一关键部位。模型在每个迭代过程中随机挑选一个通道的注意力图,这样随着网络训练,每个通道的注意力图都有可能被挑选到。然后由挑选到的注意力图$A_{1k1}\in\mathbb{R}^{1\times H\times W}$按是否大于阈值$\theta$可以生成剪切的掩模

图M_1^{Mask}，如式(6)所示：

$$M_1^{\text{Mask}}=\begin{cases}1, A_{1k1}>\theta\\0, \text{其他}\end{cases} \tag{6}$$

最后将其放大至原图，对应原图采样可得剪切图片$C_1^{\text{CropImage}}$，如式(7)所示：

$$C_1^{\text{CropImage}}=S(I, M_1^{\text{Mask}}) \tag{7}$$

其中：I为原图；S为采样函数。

2.2 第2类语义数据增强

图1中的信息增益模块和图片剪切模块2用于第2类语义增强，其中信息增益模块负责更深层次分类特征的学习，图片剪切模块2从深度分类特征得到第2种语义类型的剪切图片辅助模型训练，该语义类型图片更关注于分类目标的重要轮廓。

1)信息增益模块

对比注意力学习模块，模型从特征提取模块更深层次的卷积层中得到深度特征$f_2\in\mathbb{R}^{C\times H\times W}$，一方面，更深网络层次的卷积特征可以更关注于分类目标整体的重要信息；另一方面，随着训练迭代，模型分类逐渐满足于注意力学习模块的分类特征映射，通过构建一个结构相似但关注点区别于注意力学习模块的新的信息学习模块，可以形成相对的信息差，共同作用于最后的模型分类。因此，区别于以往的CBAM模块以单个或残差的形式出现，模型并行地引入一个额外的CBAM模块得到特征$F_2\in\mathbb{R}^{C\times H\times W}$，同理，最后运用双线性注意力汇合的思想，将特征$F_2$与生成的注意力图$A_{2k}$汇合得到最后的分类特征$P_2$。

2)图片剪切模块2

与图片剪切模块1同理，模型从信息增益模块生成的注意力图$A_{2k}\in\mathbb{R}^{K\times H\times W}$中得到剪切图片2，但为了增强其与注意力学习模块的区分度，对注意力图A_{2k}采用K-Max pooling处理，即保留前K个响应最大的注意力图。由经过K-Max pooling层的注意力图去生成剪切图片。

2.3 目标定位模块

在测试阶段，为了降低模型对分类图片的误判，模型通过构建一个目标定位模块，定位原图中的分类目标，并将其放大至原图得到目标定位图片。具体步骤：首先可以从注意力学习模块和信息增益模块分别得到经过CBAM模块的聚焦特征F_1和F_2。对于特征$F_1\in\mathbb{R}^{C\times H\times W}$，沿着通道方向对特征$F_1$进行深度求和，得到一个二维的深度描述子$S(i,j)\in\mathbb{R}^{H\times W}$，由$S(i,j)$可以得到其均值$\bar{a}$，对于$S(i,j)$中大于均值$\bar{a}$的值设定为1，其他则设定为0。由此，最终可以从注意力学习模块得到掩模图$M_1(i,j)$。同理，可以从信息增益模块中的特征F_2中得到掩模图$M_2(i,j)$。将这2个掩模图分别对应原图，取其重叠的部分，最终将重叠部分放大至原图大小得到目标定位图片，如式(8)所示：

$$L^{\text{Location\_map}}=S(I, M_1(i,j))\cap S(I, M_2(i,j)) \tag{8}$$

其中：S为采样函数；I为原图。

2.4 损失函数

由以上提出模型，可以分别得到分类特征P_1和P_2，对其采用交叉熵损失函数指导模型训练，与此同时，模型另外沿通道联接特征P_1和P_2，得到特征P，同样采用交叉熵损失函数。对于注意力学习模块和信息增益模块，模型采用双线性注意力汇合的思想，引入中心损失函数Center Loss，迫使最终特征P_1和P_2的每一维能对应分类目标的一关键部位。在测试阶段，模型实验只取联接特征P用于得到预测值。本模型实验损失函数最终如式(9)所示：

$$L=(L_{\text{cross-entropy}}(P)+L_{\text{cross-entropy}}(P_1)+L_{\text{cross-entropy}}(P_2))/3+\lambda(L_{\text{center}}(P_1)+L_{\text{center}}(P_2)) \tag{9}$$

3 实验分析与结果

本节通过实验证明各模块分别及其组合对模型分类准确率的贡献，同时在3个通用实验数据集上对比其他主流算法，最后通过可视化实验给出注意力学习模块和信息增益模块得到的不同语义层次的剪切图片，及其测试时经过目标定位后得到的剪切图片。本文模型由pytorch深度学习框架所搭建，训练环境为英伟达P40 GPU。

3.1 实验数据集

本次实验采用细粒度图像识别领域3个通用实验数据集：CUB-200-2011鸟类数据集[19]，FGVC Aircraft飞机数据集[20]，Stanford Cars车类数据集[21]。这3个数据集的详细信息如表1所示。

表1 细粒度图像分类数据集

Table 1 Fine-grained image classification datasets

数据集	分类隶属	类别数	训练样本数	测试样本数
CUB-200-2011	鸟类	200	5 974	5 794
FGVC Aircraft	飞机	100	6 667	3 333
Stanford Cars	车	196	8 144	8 041

3.2 实验细节

实验参数设置：本次实验模型采用通用网络模型Inception V3作为特征提取器，取Mix6d层特征映射作为注意力学习模块的特征图，取Mix6e层特征映射作为信息增益模块的特征图。注意力图由特征图经若干个1×1卷积核得到，其中注意力学习模块和信息增益模块本实验均设置为64个1×1卷积核，即生成64张注意力图。剪切图片模块θ阈值设为：random(0.4,0.6)。中心损失函数参数λ设为1。

训练参数设置：批量样本数设为16，学习率设为0.01。实验采用随机梯度下降法(SGD)来训练模型，动量设为0.9，权重衰减设为0.000 01。最大迭代次数设为180。在训练阶段，剪切模块剪切图片大小

均为256像素×256像素。

测试参数设置:批量样本数设为12,目标定位模块图片大小设为448像素×448像素。

实验数据集预处理:实验训练过程将所有图片调整尺寸为448像素×448像素,统一将图片进行随机翻转、随机调整亮度、标准化。测试过程将调整尺寸为448像素×448像素,统一将图片标准化。

3.3　模块及其组合准确率贡献

表2给出在数据集CUB-200-2011上模块及其组合对模型分类准确率的贡献,可以看出,各模块能有效地提高模型分类准确率。本文构建的注意力模块和信息增益模块所提取的分类特征较好地表征了细粒度图像。表3给出2种语义数据增强下上模块及其组合对模型分类准确率的贡献,可以看出,结合语义数据训练能大幅提高模型分类准确率,且组合2种语义数据辅助模型训练达到了模型最高的准确率。数据增强可以提高细粒度图像分类模型的准确率,并且双语义数据增强的设置下可以使模型性能达到最优。

表2　模块及其组合贡献程度

Table 2　Contribution of module and their combinations　%

注意力学习模块	信息增益模块	图片剪切模块1	图片剪切模块2	目标定位模块	准确率
√					86.7
√	√				87.2
√	√	√			88.5
√	√	√	√		89.1
√	√	√	√	√	89.5

表3　2种语义数据及其组合贡献程度

Table 3　Contribution of two kinds of semantic data and their combinations　%

第1类语义数据增强		第2类语义数据增强		准确率
注意力学习模块	图片剪切模块1	信息增益模块	图片剪切模块2	
√				86.7
		√		86.8
√	√			88.1
		√	√	88.3
√	√	√	√	89.1

3.4　与其他先进算法的对比

本文设置实验在数据集CUB-200-2011、FGVC Aircraft和Stanford Cars上对比其他同期先进算法,实验结果分别如表4~表6所示。可以看出,对比本实验的基准网络Inception-V3,本文算法在CUB-200-2011鸟类数据集上准确率提高了5.8%,对比本文采用的双线性注意力池化特征聚合方式,本文算法在CUB-200-2011鸟类数据集上准确率提高了3.1%。对比其他同期先进细粒度图像分类算法,本文模型在数据集CUB-200-2011、FGVC Aircraft和Stanford Cars上均表现出了更优越的性能。此外,本文模型在实验细节参数设置下,模型复杂度为185.71 MB。

表4　在CUB-200-2011数据集上的分类性能对比

Table 4　Comparison of classification performance on CUB-200-2011 dataset　%

算法	准确率
VGG-19[3]	77.8
Inception-V3[5]	83.7
BAP[2]	86.4
B-CNN[14]	84.1
RA-CNN[12]	85.4
MA-CNN[13]	86.5
DFL-CNN[22]	87.4
NTS-Net[23]	87.5
WS-DAN[17]	89.4
本文算法	89.5

表5　在FGVC Aircraft数据集上的分类性能对比

Table 5　Comparison of classification performance on FGVC Aircraft dataset　%

算法	准确率
VGG-19[3]	80.5
Inception-V3[5]	87.4
BAP[2]	84.1
B-CNN[14]	88.4
RA-CNN[12]	89.9
MA-CNN[13]	91.7
DFL-CNN[22]	91.4
NTS-Net[23]	93.0
WS-DAN[17]	93.6
本文算法	80.5

表6　在Stanford Cars数据集上的分类性能对比

Table 6　Comparison of classification performance on Stanford Cars dataset　%

算法	准确率
VGG-19[3]	85.7
Inception-V3[5]	90.8
BAP[2]	92.5
B-CNN[14]	92.8
RA-CNN[12]	93.1
MA-CNN[13]	93.9
DFL-CNN[22]	94.5
NTS-Net[23]	94.7
WS-DAN[17]	85.7
本文算法	90.8

3.5　双语义增强数据及目标定位剪切图片可视化

由图片剪切模块1和图片剪切模块2得到的剪切图片如图4所示。可以看出,在双语义数据增强

模型的设置下,模型可以由此得到2种不同语义层次的剪切图片。其中剪切图片1更关注于分类目标局部细节信息例如鸟的眼睛等,剪切图片2更关注目标重要的有区分度的轮廓,结合这2种语义层次的剪切图片可以有效提高模型分类准确率。

图4 原图及其对应双语义下的剪切图片

Fig.4 Original image and its corresponding cut image under dual semantics

本文给出经过目标定位模块的图片,如图5所示。可以看出,经过目标定位模块模型可以准确地定位于分类目标整体,从而忽视图片背景等无关信息的干扰。

图5 原图及其对应目标定位图片

Fig.5 Original image and its corresponding target location image

4 结束语

针对细粒度图像分类类内差异大、类间差异小的特点,本文基于双线性注意力融合提出注意力学习模块和信息增益模块,分别关注目标局部细节信息和目标整体重要轮廓,由此得到2种语义层次的增强数据辅助模型训练,并在测试阶段提出目标定位模块用于定位目标整体,进一步提高分类准确率。实验结果表明,本文算法在CUB-200-2011、FGVC Aircraft和Stanford Cars数据集上分别达到89.5%、93.6%和94.7%的分类准确率,性能优于对比算法。本文设计的2种语义特征学习模块可以得到2种语义层次的增强数据,但得到的2种语义层次的剪切图片区分度不够明显,有可能成为冗余数据,无法为模型带来增益。下一步将增加模块间的区分度,减少冗余信息。此外,本文算法包含了特征间的外积运算,对比基准网络Inception-V3复杂度较高,这局限了模型在移动端的应用范围,后续将考虑降低模型复杂度,构建轻量型细粒度图像分类网络。

参考文献

[1] 罗建豪,吴建鑫. 基于深度卷积特征的细粒度图像分类研究综述[J]. 自动化学报,2017,43(8):1306-1318.
LUO J H, WU J X. Review of fine-grained image classification based on deep convolution feature[J]. Acta Automatica Sinica, 2017, 43(8): 1306-1318. (in Chinese)

[2] HU T, XU J, HUANG C, et al. Weakly supervised bilinear attention network for fine-grained visual classification [EB/OL]. (2018-08-06)[2020-10-21]. http://arxiv. org/pdf/1808. 02152. pdf.

[3] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2020-10-21]. https://arxiv. org/pdf/1409. 1556. pdf.

[4] HE K M, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016: 770-778.

[5] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision [C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016: 2818-2826.

[6] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, Inception-ResNet and the impact of residual connections on learning[EB/OL]. (2016-08-23)[2020-10-21]. https://arxiv. org/pdf/1602. 07261. pdf.

[7] 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. Berlin, Germany: Springer, 2014: 834-849.

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

[9] BRANSON S, HORN G, BELONGIE S, et al. Bird species categorization using pose normalized deep convolutional nets[EB/OL]. (2014-06-11)[2020-10-21]. http://de. arxiv. org/pdf/1406. 2952.

[10] LIN D, SHEN X, LU C, et al. Deep LAC: deep localization, alignment and classification for fine-grained recognition [C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2015: 1666-1674.

[11] JADERBERG M, SIMONYAN K, ZISSERMAN A. Spatial transformer networks [C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. New York, USA: ACM Press, 2015: 2017-2025.

[12] FU J, ZHENG H, MEI T. Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 4438-4446.

[13] ZHENG H, FU J, MEI T, et al. Learning multi-attention convolutional neural network for fine-grained image recognition[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2017: 5209-5217.

(下转第249页)

（上接第242页）

[14] LIN T Y, ROYCHOWDHURY A, MAJI S. Bilinear CNN models for fine-grained visual recognition[C]//Proceedings of 2015 IEEE International Conference on Computer Vision. Washington D. C. , USA: IEEE Press, 2015: 1449-1457.

[15] LIN T Y, MAJI S. Improved bilinear pooling with CNNs [EB/OL]. (2017-07-21)[2020-10-21]. https://arxiv. org/pdf/1707. 06772. pdf.

[16] LI P, XIE J, WANG Q, et al. Towards faster training of global covariance pooling networks by iterative matrix square root normalization[C]//Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. , USA: IEEE Press, 2018: 947-955.

[17] HU T, QI H, HUANG Q, et al. See better before looking closer: weakly supervised data augmentation network for fine-grained visual classification[EB/OL]. (2019-03-23) [2020-10-21]. https://arxiv. org/pdf/1901. 09891. pdf.

[18] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//Proceedings of 2018 European Conference on Computer Vision. Berlin, Germany: Springer, 2018: 3-19.

[19] WAH C, BRANSON S, WELINDER P, et al. The caltech-ucsd birds-200-2011 dataset[EB/OL]. [2020-10-21]. http://citeseerx. ist. psu. edu/viewdoc/download;jsessionid=374669091E8C13903183C647B249A20B? doi=10. 1. 1. 372. 852&rep=rep1&type=pdf.

[20] MAJI S, RAHTU E, KANNALA J, et al. Fine-grained visual classification of aircraft[EB/OL]. (2013-06-21) [2020-10-21]. https://arxiv. org/pdf/1306. 5151. pdf.

[21] KRAUSE J, STARK M, DENG J, et al. 3d object representations for fine-grained categorization [C]// Proceedings of 2013 IEEE International Conference on Computer Vision. Washington D. C. , USA: IEEE Press, 2013: 554-561.

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

[23] YANG Z, LUO T, WANG D, et al. Learning to navigate for fine-grained classification [C]//Proceedings of 2018 European Conference on Computer Vision. Berlin, Germany: Springer, 2018: 420-435.

编辑　金胡考