经卓勋,刘建明. 面向细粒度分类的预测属性引导的注意力研究[J]. 江西师范大学学报(自然科学版),2022,46(4):379-385.
JING Zhuoxun,LIU Jianming. The method on fine-grained image categorization using predicted-attribute guided channel attention module[J]. Journal of Jiangxi Normal University(Natural Science),2022,46(4):379-385.

文章编号:1000-5862(2022)04-0379-07

面向细粒度分类的预测属性引导的注意力研究

经卓勋,刘建明*
(江西师范大学计算机信息工程学院,江西 南昌 330022)

摘要: 细粒度图像分类任务比一般图像分类任务更具有挑战性,其通常需要对类间差异小、类内差异大的样本进行分类. 现有细粒度分类方法主要依赖视觉特征进行分类,而人类可以根据文本描述等属性描述来辅助识别图像类别. 该文提出了一种通过预测属性引导的通道注意力模块,该模块可以插入到任意的卷积神经网络中,从而让模型学习到更高级的特征表示. 最后,该算法在CUB-200-2011数据集上测试,在使用Resnet-50、VGG-19、Bilinear-CNN作为主干网络训练时的精度分别达到87.1%、82.1%、85.5%,精度得到显著提升.

关键词: 细粒度图像分类; 注意力机制; 属性预测
中图分类号: TP 311 **文献标志码:** A **DOI:** 10.16357/j.cnki.issn1000-5862.2022.04.08

0 引言

细粒度识别分类是计算机视觉领域热门的研究问题,其主要是对数据中在粗粒度的大类下的细粒度子类进行分类. 在细粒度图像分类任务中,不同子类别间存在较高相似度,这种高相似度往往只存在局部的且细微的差异,如不同鸟类间鸟喙的色块差异或羽毛花纹的不同等,这也使得相对于一般的图像分类算法,细粒度图像分类更困难且更具挑战性. 因此,在细粒度识别任务中,通常采用引入注意力机制的方法探索目标图像中细微的、最具有判别力的部位,但此类方法通常只考虑图像样本所提供的特征,从而发掘更多的目标部件,并在单模态场景下做识别,最终忽略了其他模态的信息来源.

最近,基于多模态信息融合的细粒度识别方法引起了越来越多学者的关注,因为人类在学习认识事物的过程中,不仅从图像样本得到知识,还从文本、口头描述中认识事物,所以该类方法就基于这种思路,将高层次的语义信息(如人工标签、属性注释等)作为辅助,结合图像数据建立联合模型. 如双模态渐进式掩模注意力模型(Bi-modal PMA)[1]通过CNN(convolutional neural network)[2]和LSTM(long short-term memory)[3]分别将图像和文本数据编码,通过自注意力模块分别从上述2个单一模态收集信息,并通过一个查询-关系模块建立双模态信息之间的关系; CVL[4]方法则是采取视觉流和语言流同时编码的双分支模型,但这类方法需要构建双分支模型结构,且不易改动,在训练和测试阶段需要双模态数据同时参与,复杂度较高.

为解决上述问题,本文提出了一种可随意插入任何卷积神经网络的、融合预测属性的注意力算法,该算法主要由2个部分组成:(a)属性预测模块APM(attribute-predicted module),该模块通过已知属性语义向量信息构建类别属性向量矩阵,用于预测所有输入图像样本在属性语义空间上的属性语义向量;(b)基于预测属性引导的通道注意力模块PACAM(predicted-attribute-guided channel attention module),该模块以上一层特征图和属性预测模块APM的预测属性语义向量作为辅助输入,挖掘出特

收稿日期:2022-02-18
基金项目:国家自然科学基金(61662034)和江西省省自然科学基金(20202BAB202020)资助项目.
通信作者:刘建明(1981—),男,江西鹰潭人,副教授,博士,主要从事细粒度图像识别、零样本和小样本学习、弱监督学习研究. E-mail:liujianming@jxnu.edu.cn

征图潜在的特征语义表示,且测试阶段不需要属性语义数据.

1 相关工作

1.1 细粒度图像分类

细粒度图像分类是在计算机视觉领域中十分具有挑战性的问题,在现实场景中有着广泛的应用前景. 目前主流的细粒度图像分类方法可以分为 3 类:(a) 端到端的特征学习方法;(b) 先检测对象的多个部件,再从已定位的部件区域中提取图像特征的方法;(c) 基于多模态信息融合的方法.

第 1 种方法是构建端到端的深度学习框架进行特征编码,在训练中只使用图像级标签构建模型学习更有判别力的特征. 如 Gao Yu 等[5] 提出的通道交互网络,可通过图像的特征图得到通道间的差距信息,再与原始特征图的特征结合,从而增强通道所学习到的判别特征. Zhuang Peiqin 等[6] 提出了一种注意力成对交互网络,通过对输入的 2 张图像得到一个共同特征向量,查找特征通道中包含的对比线索.

第 2 种方法是先检测对象的某些部位,从部位检测特征来识别对象. 早期工作需要数据库的外部标注信息,最近的方法只需提供分类标签,不需要额外标注. 如 Hu Tao 等[7] 提出的 WS-DAN(weakly supervised data augmentation network) 网络,采用双线性注意力池化机制生成注意力图,并使用生成的注意力图来生成 Mask 掩码遮罩,再应用到原图像上实现数据增强. Zheng Heliang 等[8] 提出了三线性注意力采样网络 TASN(tri-linear attention sampling network),在其之前的细粒度识别的方法中,提取的对象部位的数量是预定义的,缺乏灵活性,且计算成本较高,而 TASN 通过一个三线注意力机制模块做细节定位,通过知识蒸馏的方式来优化主网络. Du Ruoyi 等[9] 则致力于将多粒度的图像特征融合,用一个拼图生成器生成不同粒度级别的图像,提出了一个新的渐进式训练策略,在每个训练步骤中融合来自先前粒度级别的数据. Ding Yifeng 等[10] 同时运用高层和低层的特征信息,建立了注意力金字塔卷积神经网络,该模型具有自上而下的特征金字塔结构和自下而上的注意力金字塔结构. A. Behera 等[11] 提出了一种上下文感知注意力池,使模型更好地学习物体各部位特征. Ji Ruyi 等[12] 将决策树和神经网络相结合,提出了一种注意力卷积二叉神经树结构做分类,在树结构的边缘加入卷积计算,使用 Attention Transformer 模块让模型捕捉具有判别力的特征,从而让分支路由模块来决定样本送往左子树还是右子树. 而 He Ju 等[13] 基于 Vision-Transformer 结构设计了 TransFG 模型,考虑到图像空间上的相邻结构,在将图像序列化时采用重叠的 patch 方法,整合注意力权重至注意力映射,保留全局信息,计算区域的对比损失从而定位图像的差异性局部区域,以此进行细粒度分类.

第 3 种方法是基于多模态信息融合建立模型. 人类不仅从物体对象的外观上认识对象,也可从多种渠道获取对象的相关知识. 因此,在拥有图像数据的基础上,将多模态信息数据作为辅助,提高网络的特征表达性能. 在细粒度分类领域中,目前主要使用的辅助分类信息为文本和数据属性. Chen Tianshui 等[14] 提出了知识嵌入式表示学习(knowledge-embedded representation learning,KERL) 的框架,通过引入属性和知识图来处理细粒度图像识别的任务. Sun Liang 等[15] 通过一个文本嵌入网络将图像的文本描述编码为向量表示,并与图像编码网络的向量连接输入双线性网络,得到了较高的准确率. 还有使用更多模态的方法,如 Zhu Linchao 等[16] 利用视频和文字,设计了全局动作、局部区域以及语言描述 3 种模态的输入进行多模态特征学习,提高了人体细粒度动作图像识别的准确率.

但是以上结合多模态信息的方法,通常有着比较复杂的结构,或者在训练与测试阶段都需要额外的属性或者文本信息作为辅助,模型的灵活度较差,适用范围小. 因此,本文提出了基于预测属性引导的注意力机制的细粒度分类方法.

1.2 注意力机制

人类在对接收的视觉信息进行处理时,会选择性地关注某些关键性的细节,而忽略其他信息. 注意力模块将人类认识图像信息的注意力机制应用到计算机视觉领域,让计算机学习人类视觉认知的思维方式处理视觉信息,使模型在处理信息的过程中更多地关注关键部分的特征.

V. Mnih 等[17] 在传统的循环神经网络中应用了注意力机制,通过注意力机制来学习一幅图像中需要关注的部分,而不是直接处理整幅图像的像素,同时也降低了任务的复杂度.

Woo Sanghyun 等[18] 提出了一种卷积注意力模块(convolutional block attention module,CBAM),该模块分为通道注意力模块与空间注意力模块,先后

在通道与空间维度上寻找图像与最终输出结果相关联的区域.

同样，还有受基于 CBAM 的工作启发后的延伸方法：BAM（bottleneck attention module）[19]，该模块的构思是将 CBAM 中通道与空间注意力模块的连接方式从串联改为并联，但同时 CBAM 需要通过通道间的池化处理，而 BAM 则全部通过卷积或空洞卷积完成，融合了多感受野的信息，比 CBAM 计算量更大，但是包含更多的信息.

本文提出注意力引导的通道注意力模块（PACAM），通过属性预测模块（APM）将图像特征图数据映射到一个属性向量语义空间，并对输入的每个样本做属性预测，得到一个属性语义向量（以下简称属性向量）作为注意力机制的辅助特征，在特征图的通道维度上应用注意力机制. 在应用到细粒度分类任务常用的骨干网络框架下，其准确率得到了明显提升.

2　预测属性引导的注意力机制算法

本文提出的基于预测属性引导的通道注意力算法是通过属性信息来引导通道注意力，模块结构由属性预测模块（APM）与预测属性引导的通道注意力模块（PACAM）组成.

本文算法使用 VGG-19、Resnet-50、Bilinear-CNN 3 种在图像细粒度识别领域中常用的卷积神经网络结构作为骨干（Backbone）网络，可以在任意的 CNN 网络结构后添加本文提出的 PACAM 与 APM 进行端到端的训练. 在通道维度上使用属性语义向量作为辅助引导注意力模块学习，从而增强网络的表达能力.

设在模型训练时从训练集 S 中提取一个 Batch 的样本数据 $S_1 = \{x_1, x_2, \cdots, x_N\}$，总数为 N，则以在 S_1 中提取出的图像数据来随机裁剪、翻转，得到 $S_2 = \{x'_2, x'_2, \cdots, x'_N\}$，其中 x'_i 代表 S_1 中第 i 幅图像经过随机裁剪、翻转后的图像，则每个 Batch 是 S_1 与 S_2 的合并，总共有 $2N$ 幅图像.

算法整体结构图如图 1 所示，上一层特征图作为模块输入，C、H、W 分别为特征图通道数、特征图高度、宽度，D_{attr} 为预测属性向量维度，K 为总类别数. 其中 APM 用于学习一个属性语义空间，将类别和样本映射到语义空间中，让视觉分类器学习一个特征图到属性向量的映射. 以图像的特征图作为输入，得到预测的样本的属性向量.

PACAM 以输入特征图以及从 APM 的输出属性向量作为输入，用于将预测的属性向量与图像特征图的向量表示连接，寻找图像特征中更具有判别力的部分，得到加权后的特征图作为下一层输入.

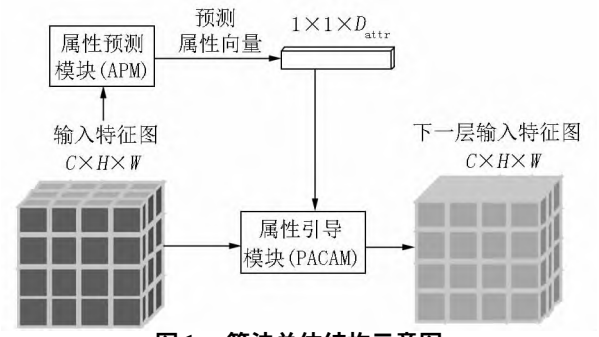

图 1　算法总体结构示意图

2.1　属性预测模块

属性预测模块将学习一个属性语义空间，将样本特征图映射到这个空间中，得到特征图在属性语义空间中的属性语义表示. 属性预测方法使用卷积神经网络作为骨干网络，网络输出的特征图作为输入. 本文的属性预测模块的主要结构图如图 2 所示.

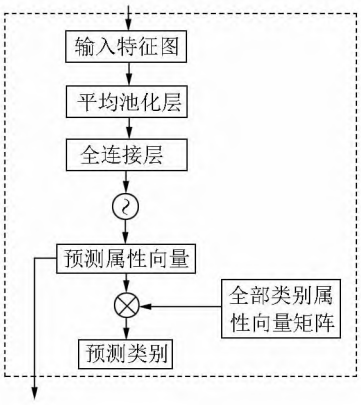

图 2　属性预测模块结构示意图

假设输入的某一层特征图为 $\boldsymbol{F}^{\text{In}}$，其维度为 $C \times H \times W$. 属性预测模块的类别属性向量矩阵由所有类别的属性向量按序排列得到，类别的属性向量维度为 $1 \times 1 \times D_{\text{attr}}$，而类别属性向量矩阵 $\boldsymbol{M}_{\text{attr}}$ 的维度为 $1 \times K \times D_{\text{attr}}$，即每个类别存在一个 $1 \times D_{\text{attr}}$ 维度的向量表示，D_{attr} 为属性向量维度，大小和所使用数据集中样本提供属性数目相等.

在 APM 模块中，在构建类的全部类别属性向量矩阵（以下简称为类别属性矩阵）时，若数据库没有给出每个类的属性向量，则取该类所有样本实数值向量的平均值. 考虑到每个属性对于类别目标的重要程度不同，将每个类的属性向量归一化到 0 ~ 1 之间，每个属性向量为连续性实值向量，则重要的属性能够获得更大的权值，不重要的属性的权值较小.

相比于不重要的属性,重要的属性在训练中能发挥更大的影响.

在本模块中,获得预测的属性向量的公式为

$$\boldsymbol{a}_{\mathrm{ttr}}(\boldsymbol{F}^{\mathrm{In}};\theta) = \sigma(f_c(\mathrm{avg}(\boldsymbol{F}^{\mathrm{In}};\theta))),$$

其中 $\mathrm{avg}(\cdot)$ 为特征图 $\boldsymbol{F}^{\mathrm{In}}$ 经过的全局平均池化层,f_c 为全连接层函数,θ 为全连接层可学习参数,最终结果将向量映射到 $1\times1\times D_{\mathrm{attr}}$ 维度,σ 为 Sigmoid 激活函数,$\boldsymbol{a}_{\mathrm{ttr}}(\boldsymbol{F}^{\mathrm{In}};\theta)$ 为预测出的属性向量,其将作为 PACAM 的输入,记为 $\boldsymbol{V}_p$.

在训练阶段,最终预测结果的向量表示为

$$P(\boldsymbol{F}^{\mathrm{In}};\theta) = \boldsymbol{M}_{\mathrm{attr}}^{\mathrm{T}}\boldsymbol{a}_{\mathrm{ttr}}(\boldsymbol{F}^{\mathrm{In}};\theta), \quad (1)$$

其中 $P(\boldsymbol{F}^{\mathrm{In}};\theta)$ 为最后输出的每个类别的概率分布. APM 在模型训练阶段最小化 $P(\boldsymbol{F}^{\mathrm{In}};\theta)$ 与真实类别的 One-Hot 标签向量的交叉熵损失,而在训练阶段优化 APM,该交叉熵损失函数可表示为

$$L_{\mathrm{ce}} = -\frac{1}{n}\sum_{j=1}^{n}\sum_{c=1}^{K} y_{jc}\log(p_{jc}),$$

其中 n 为样本数,j 表示第 j 个样本,y_{jc} 为符号函数,若样本 j 真实类别为 c 则 y_{jc} 取 1,否则 y_{jc} 取 0,p_{jc} 样本 j 属于 c 类别的概率,即式(1)中得到的 $P(\boldsymbol{F}^{\mathrm{In}};\theta)$. 而在测试阶段,只输出预测出的属性向量作为通道注意力模块的输入.

属性预测模块使用对比损失函数来缩短当前样本属性向量与同一类样本属性向量之间的距离,远离其他类别的样本属性向量. 该对比损失函数可以表示为

$$L_{\mathrm{cons}} = \mathrm{e}^{1+\sum_{i=0,i\neq z,a\neq b}^{2N}|\cos\langle \boldsymbol{v}_{a,z},\boldsymbol{v}_{b,i}\rangle|} \Big/ \mathrm{e}^{\sum_{i=0,i\neq z}^{2N}|\cos\langle \boldsymbol{v}_{a,z},\boldsymbol{v}_{a,i}\rangle|},$$

其中 $2N$ 表示在一个 Batch 中所有的样本数量,z 为当前样本编号,a 表示当前预测属性样本的类别,$\boldsymbol{v}_{b,i}$ 表示在一个 epoch 中编号为 i 的属于 b 类别的样本,$b\in K$,$\cos\langle\cdot,\cdot\rangle$ 表示计算 2 个预测出的属性向量之间的余弦相似度.

对比损失函数 L_{cons} 采用余弦相似度来计算 2 个向量之间的相似度,从而缩小同一类预测属性向量的距离,远离其他类别的预测向量.

APM 的总体损失函数为对比损失函数加上交叉熵损失函数 L_{ce},即 $L_{\mathrm{apm}} = L_{\mathrm{ce}} + \alpha L_{\mathrm{cons}}$,其中 α 为根据经验设定的超参数.

2.2 属性引导的通道注意力模块

本文的算法是在 CBAM 的基础上提出改进的 PACAM,改进自 CBAM 在通道维度上的注意力模块,并将由 APM 预测的属性向量 $\boldsymbol{V}_p$ 作为引导输入,通过属性向量让神经网络模型关注图像信息中更具有判别力的部分. PACAM 的算法流程如图 3 所示.

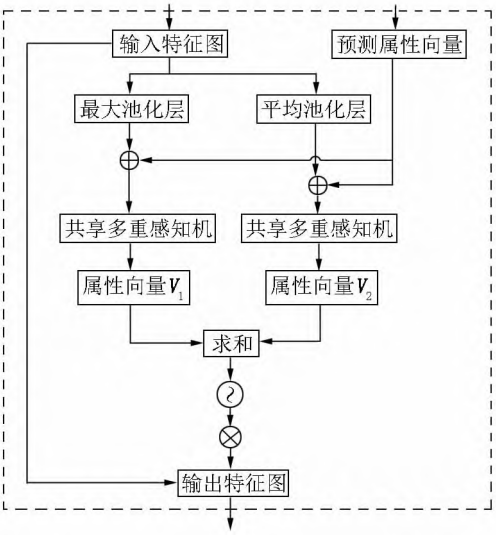

图 3 属性引导的通道注意力模块结构示意图

此时输入的特征图 $\boldsymbol{F}^{\mathrm{In}}\in\mathbf{R}^{C\times H\times W}$,记输出的特征图 $\boldsymbol{F}^{\mathrm{Out}}\in\mathbf{R}^{C\times H\times W}$,PACAM 对输入的特征图 $\boldsymbol{F}^{\mathrm{In}}$ 分别做全局最大池化与全局平均池化,得到 2 个维度相同的向量 $\boldsymbol{V}_{\max}$ 和 $\boldsymbol{V}_{\mathrm{avg}}$,向量维数均为 $1\times1\times C$.

将这 2 个向量和得到的预测属性向量拼接,得到总维数为 $1\times1\times(C+D_{\mathrm{attr}})$ 的 2 个维度相同的向量,再将这 2 个向量通过一个共享权值的多重感知机得到 2 个属性向量 $\boldsymbol{V}_1$ 和 $\boldsymbol{V}_2$,它们的维数均为 $1\times1\times C$. 按元素求和并通过 sigmoid 函数后,输出压缩维数为 $1\times1\times C$ 的通道注意力向量 $\boldsymbol{V}_a$,将其与 $\boldsymbol{F}^{\mathrm{In}}$ 做乘法,得到重组后的下一层特征图 $\boldsymbol{F}^{\mathrm{Out}}$.

通道注意力向量 $\boldsymbol{V}_a$ 的计算公式为

$$\boldsymbol{V}_a(\boldsymbol{F}^{\mathrm{In}}) = \sigma(M_{LP}([\mathrm{avg}(\boldsymbol{F}^{\mathrm{In}});\boldsymbol{V}_p]) + M_{LP}([\max(\boldsymbol{F}^{\mathrm{In}});\boldsymbol{V}_p])),$$

其中 M_{LP} 为共享权重的多重感知机,有一个隐藏层. 进一步可得

$$\boldsymbol{V}_a(\boldsymbol{F}^{\mathrm{In}}) = \sigma(W_1(W_0([\mathrm{avg}(\boldsymbol{F}^{\mathrm{In}});\boldsymbol{V}_p])) + W_1(W_0\cdot([\max(\boldsymbol{F}^{\mathrm{In}});\boldsymbol{V}_p]))),$$

其中 W_0、W_1 分别为多重感知机中的可学习的权重参数. $W_0\in\mathbf{R}^{(C/r)\times C)}$,$W_1\in\mathbf{R}^{C\times C/r}$,$r$ 是可设定的系数且为削减参数开销的比率. 最后得到的通道注意力向量 $\boldsymbol{V}_a(\boldsymbol{F}^{\mathrm{In}})$.

将向量 $\boldsymbol{V}_a(\boldsymbol{F}^{\mathrm{In}})$ 广播到和特征图相同的维度 $C\times H\times W$,得到通道注意力图 $\boldsymbol{M}_a$.

最后得到的下一层特征图 $\boldsymbol{F}^{\mathrm{Out}}$ 的公式为

$$\boldsymbol{F}^{\mathrm{Out}} = \boldsymbol{M}_a \otimes \boldsymbol{F}^{\mathrm{In}}.$$

APM 的总损失函数是由 APM 的损失函数 L_{apm} 和主干网络的交叉熵损失函数 L_{ce} 之和，即

$$L_{\mathrm{all}} = L_{\mathrm{ce}} + L_{\mathrm{apm}}.$$

3　实验

3.1　实验数据集 CUB-200-2011

本文采用鸟类细粒度图像数据集 CUB-200-2011[20] 作为实验数据集，因为含有属性信息的细粒度数据集较少，所以在细粒度图像数据集中只有 CUB 鸟类数据集具有可靠清晰的属性语义数据.

该数据集包含 200 种鸟类，共有 11 788 幅图，5 994幅训练图，5 794 幅测试图，共设有羽毛颜色、背上颜色等 312 种属性. 在数据集中，每种鸟类具有连续的实数的 312 维属性向量，某些种类的鸟类间的差异可能非常微小，如大天鹅和小天鹅，它们只有鸟喙的图案判断差异，观察者很难准确判断图像中鸟的类别，因此十分具有挑战性.

3.2　模型训练

本文的实验用的 GPU 为 NVIDIA GTX 1080Ti，显存为 11 GB，CPU 为 Inter i7-7700，所用的实验系统为 Ubuntu 16.04 LTS，CUDA 10.1，深度学习框架为 Pytorch-1.6.0.

本文提出的预测属性引导的注意力模块能插入任意 CNN 网络实现端到端的训练. 在训练中，Batch Size 设为 16(即 $2N$) ，α 采用人工设置为 0.1，采用 SGD 优化器，初始学习率为 0.001，每训练 40 个 epoch 将学习率乘以 0.1 衰减，因此在进行至 40 个 epoch 以后总损失不再变化，判断学习率过大，故将学习率乘以 0.1 并使得总损失函数值继续收敛，共训练了 130 个 epoch.

3.3　PACAM + APM 性能测试

为了测试本文所提出的 PACAM 的性能，选择了在细粒度识别领域比较常用的 3 种骨干网络(VGG-19、Resnet-50、Bilinear-CNN) 作为实验对象.

表 1 显示了在使用不同的骨干网络时设置的图像大小以及模块插入位置，考虑到本文提出的模块加在不同网络层数上会得到不同结果，因此在使用 Resnet-50 作为骨干网络进行实验时考虑了多种情况.

表 1　模块性能测试实验设置

骨干网络	输入图像大小	插入位置
VGG-19	448 × 448	conv5_4
Resnet-50-b4	448 × 448	conv4_6
Resnet-50-b5	448 × 448	conv5_3
B-CNN	448 × 448	conv_fusion

实验图像的输入分辨率统一变换至 448 × 448，在实验中采用了左右随机翻转、随机裁剪 2 种图像增广操作. 使用了 3 种不同骨干网络，Resnet-50-b4/5 表示将 PACAM 和 APM 添加到第 4 或第 5 组的 Block 的最后一个 Bottleneck 结构之后(如 conv5_3 表示 Block 中第 3 个 Bottleneck 结构位置) ，VGG-19 网络则加在第 16 个卷积层之后，其中 conv5_4 表示在第 16 个卷积层之后的位置.

B-CNN 意为双线性池化卷积神经网络(biline-ar-CNN) ，采用 Resnet-34 网络结构作为其分支网络的骨干网络，在表 1 中 conv_fusion 意味着将 PACAM 加在 B-CNN 双分支的特征融合层之后. 其在 CUB 数据集上的表现如表 2 所示.

表 2　在 CUB-200-2011 数据库上各算法性能比较

骨干网络	测试集准确率/%
VGG-19	81.6
VGG-19 + PACAM + APM	82.1
Resnet-50	84.5
Resnet-b4 + PCBAM + APM	86.7
Resnet-b5 + PCBAM + APM	87.1
B-CNN	83.2
B-CNN + PACAM + APM	85.6

只使用 Resnet-50 网络在 CUB 数据集上分类的结果为 84.5%，在第 5 个 block 之后加入了本文所述的 PACAM 和 APM 以后，其鸟类分类的准确率提高了 2.6%. 而 VGG-19 网络使用 PACANM 比不使用 PACAM 仅提高了 0.5%，双线性网络B-CNN在使用 Resnet-34 作为骨干网络后提高了2.4% 的准确率.

在使用 VGG-19 的情况下提升情况较小，推测原因是通道维数的差异产生的影响，如在 Resnet-50 网络的 b4 和 b5 这 2 种设置下 b5 的通道维数更大，也得到了更高的准确率. 且在实验中的 Resnet-50 骨干网络输入 PACAM 的特征图通道维数比 VGG-19 网络的特征图通道维数更大，能够在更多的通道中更好地寻找到其中应当得到关注的部分.

3.4 消融实验

为了验证本文提出的通过预测属性引导注意力的方法有效性,在性能测试的基础上设计了消融实验(见表 3).

表 3 采用 Resnet-50 作为骨干网络的消融实验结果

CBAM	PACAM + MAP	PACAM + APM	测试集准确率/%
√			85.5
	√		84.9
		√	87.1

在表 3 中的消融实验是将 3 种模块设计加入 Resnet-50 的 conv5_3 位置得到的实验结果,其中 CBAM 是使用没有加入属性语义信息的 CBAM 插入 Resnet-50 进行的网络实验,准确率仅有85.5%.

MAP 是不使用预测属性模块,仅仅通过一个多重感知机模块将图像直接映射到属性相同维度的向量,且未使用类属性矩阵引导直接产生预测结果的方法. 在 PACAM + MAP 的组合中,使用 MAP 替换本文的 APM,将 MAP 方法得到的向量输入 PACAM 得到的结果显示: 其在测试集上的准确率比只使用 CBAM 的方法的准确率更低,仅有84.9%,而在使用了本文提出的 PACAM 和 APM 的组合后,其测试集上的准确率有了明显提高. 这说明加入属性向量引导注意力的方法是有效的.

在实验中直接将图像映射到向量维度造成识别精度的下降,以上结果说明: 类别的属性编码确实能够成为辅助信息做注意力引导,关注重要的特征通道. 而在进行细粒度识别分割任务中,为了准确分类差异小的子类对象,通常需要大量而准确的各类标签,但让模型学习这类信息较为困难. 由于属性信息带有精确的类别差异知识,所以本文实验将其编码的实值属性语义向量和图像特征结合,提高了模型分类的性能和准确率.

4 结论

本文提出了一种基于预测属性引导的通道注意力算法,该算法包含 PACAM 和 APM,可以插入任意一种 CNN 网络. 一方面,利用已有的类属性向量构成类属性向量矩阵,通过 APM 建立属性语义向量的预测模型,输出预测属性向量引导主干网络的训练和测试; 另一方面,通过 PACAM 模块连接属性向量和通道注意力向量,引导通道注意力算法寻找更具有判别力区域的特征通道. 本文最后通过实验验证了该模块的性能.

最后,本文算法未来考虑从如下一些方面改进: 考虑属性语义信息与图像特征在空间方面的联系,将空间信息结合语义向量进一步做出研究和改进; 考虑更多形式的语义信息,而属性信息通常需要人工标注,代价较高,可以考虑结合其他更易取得的数据,如数据集的文本描述,将其与属性结合、与图像数据建立联合判断模型.

5 参考文献

[1] SONG Kaitao, WEI Xiushen, SHU Xiangbo, et al. Bi-modal progressive mask attention for fine-grained recognition [J]. IEEE Transactions on Image Processing, 2020, 29: 7006-7018.

[2] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.

[3] SHI Xingjian, CHEN Zhourong, WANG Hao, et al. Convolutional LSTM network: a machine learning approach for precipitation nowcasting [EB/OL]. [2021-12-17]. https://arxiv.org/pdf/1506.042.pdf.

[4] HE Xiangteng, PENG Yuxin. Fine-grained image classification via combining vision and language [EB/OL]. [2021-11-13]. https://arxiv.org/abs/1704.02792v1.

[5] GAO Yu, HAN Xintong, WANG Xun, et al. Channel interaction networks for fine-grained image categorization [J]. AAAI-20 Technical Track 7, 2020, 34(7): 10818-10825.

[6] ZHUANG Peiqin, WANG Yali, QIAO Yu. Learning attentive pairwise interaction for fine-grained classification [J]. AAAI-20 Technical Track 7, 2020, 34(7): 13130-13137.

[7] HU Tao, Qi Honggang. See better before looking closer: weakly supervised data augmentation network for fine-grained visual classification [EB/OL]. [2021-11-12]. https://arxiv.org/abs/1901.09891v1.

[8] ZHENG Heliang, FU Jianlong, ZHA Zhengjun, et al. Looking for the devil in the details: learning trilinear attention sampling network for fine-grained image recognition [EB/OL]. [2021-11-15]. https://arxiv.org/abs/1903.06150v2.

[9] DU Ruoyi, CHANG Dongliang, BHUNIA A K, et al. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches [EB/OL]. [2021-10-19].

https://arxiv.org/abs/2003.03836v2.

[10] DING Yifeng，MA Zhanyu，WEN Shaoguo，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] BEHERA A，WHARTON Z，HEWAGE P，et al. Context-aware attentional pooling (cap) for fine-grained visual classification [EB/OL]. [2021-10-13]. https://doi.org/10.48550/arXiv.2101.06635.

[12] JI Ruyi，WEN Longyin，ZHANG Libo，et al. Attention convolutional binary neural tree for fine-grained visual categorization [EB/OL]. [2021-11-10]. https://ieeexplore.ieee.org/document/9157539.

[13] HE Ju，CHEN Jieneng，LIU Shuai，et al. TransfG：a transformer architecture for fine-grained recognition [EB/OL]. [2021-12-03]. https://doi.org/10.48550/arXiv.2103.07976.

[14] CHEN Tianshui，LIN Liang，CHEN Riquan，et al. Knowledge-embedded representation learning for fine-grained image recognition [EB/OL]. [2021-12-09]. https://doi.org/10.48550/arXiv.1807.00505.

[15] SUN Liang，GUAN Xiang，YANG Yang，et al. Text-embedded bilinear model for fine-grained visual recognition [EB/OL]. [2021-10-13]. https://doi.org/10.1145/3394171.3413638.

[16] ZHU Linchao，YANG Yi. ActBERT：learning global-local video-text representations [EB/OL]. [2021-10-18]. https://doi.org/10.48550/arXiv.2011.07231.

[17] MNIH V，HEESS N，GRAVES A. Recurrent models of visual attention [EB/OL]. [2021-10-18]. http://de.arxiv.org/pdf/1406.6247.

[18] WOO S，PARK J，LEE J Y，et al. CBAM：convolutional block attention module [EB/OL]. [2021-10-16]. https://arxiv.org/pdf/1807.06521.pdf.

[19] PARK J，WOO S，LEE J Y，et al. BAM：bottleneck attention module [EB/OL]. [2021-11-13]. https://arxiv.org/pdf/1807.06514.pdf

[20] WAH C，BRANSON S，WELINDER P，et al. The CALTECH-UCSD birds-200-2011 dataset [EB/OL]. [2021-11-16]. https://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=FF4D49EB69FF280011EEF1AA90D8050A?doi=10.1.1.372.852&rep=rep1&type=pdf.

The Method on Fine-Grained Image Categorization Using Predicted-Attribute Guided Channel Attention Module

JING Zhuoxun，LIU Jianming*

(School of Computer Information Engineering，Jiangxi Normal University，Nanchang Jiangxi 330022，China)

Abstract: The fine-grained image classification task is more challenging than the general image classification task. Its task usually needs to classify the samples with low inter-class but high intra-class variation. The existing fine-grained classification methods mainly rely on visual features for classification，but human beings can recognize image categories according to text attribute description. To this end，the predicted-attribute guided channel attention module is proposed. The module can insert any convolutional neural network，which is intended to make the model to learn more advanced feature representation. Finally，the algorithm proposed in this paper tests on CUB-200-2011 dataset. The algorithm achieves 87.1%，82.1%，85.5% when training using Resnet-50，VGG-19，Bilinear-CNN as backbone network. It can observe a significant improvement in accuracy.

Key words: fine grained image classification; attention n mechanism; attribute prediction

(责任编辑：冉小晓)