DOI:10.13878/j.cnki.jnuist.2019.06.001

邓旭冉[1] 闵少波[1] 徐静远[1] 李攀登[1] 谢洪涛[1] 张勇东[1]

深度细粒度图像识别研究综述

摘要

细粒度图像分类是计算机视觉中一项基础且重要的工作,其目的在于区分难以辨别的对象类别(例如不同子类的鸟类、花或动物).不同于传统的图像分类任务可以雇佣大量普通人标注,细粒度数据集通常需要专家级知识进行标注.除了视觉分类中常见的姿态、光照和视角变化因素之外,细粒度数据集具有更大的类间相似性和类内差异性,因此要求模型能够捕捉到细微的类间差异信息和类内公有信息.除此之外,不同类别的样本存在不同程度的获取难度,因此细粒度数据集通常在数据分布中表现出长尾的特性.综上所述,细粒度数据分布具有小型、非均匀和不易察觉的类间差异等特点,对强大的深度学习算法也提出了巨大的挑战.本文首先介绍了细粒度图像分类任务的特点与挑战,随后以局部特征与全局特征两个主要视角整理了目前的主流工作,并讨论了它们的优缺点.最后在常用数据集上比较了相关工作的性能表现,并进行了总结与展望.

关键词

细粒度图像识别;深度学习;局部区域检测;双线性池化

中图分类号 TP183;TP391.41

文献标志码 A

收稿日期 2019-10-09

资助项目 国家重点研发计划(2017YFC0820600)

作者简介

邓旭冉,女,博士,主要研究方向为多媒体技术和网络空间安全.kjj3chu@sina.com

闵少波(通信作者),男,博士,主要研究方向为多媒体技术和计算机视觉.mbobo@mail.ustc.edu.cn

1 中国科学技术大学 信息科学技术学院,合肥,230026

0 引言

使机器能够以视觉方式自动识别物体是计算机视觉的核心挑战之一.其目的是根据视觉线索识别并详细说明图像或视频中的情况.在大多数情况下,此任务会缩小到将输入分为一组给定类别或标签中的一个,因此也被称为分类或标签问题.

细粒度视觉分类(Fine-Grained Visual Categorization,FGVC)是视觉识别系列中一个相对较新的任务.与传统的分类问题[1-3](如区分猫狗、汽车和自行车)不同,FGVC 专注于识别狗的细致种类[4]或具体的汽车模型[5].例如,任何人都可以从车海中自动识别出自行车,但是,如果询问此人图像中所有车的款式、型号和种类的话[6],则大多数人都需要专家或互联网的帮助.由于细粒度或所谓的下属类别通常看起来彼此之间更相似,并且在分类过程中会引起更多混淆,FGVC 是对传统工作的重大改进并大大地提升了实用性.在 FGVC 中,总体数据集被称为域,要区分的下级类别被称为类,而我们要在域中执行分类.有时会用域的特定的名称(例如花[7]或鸟的物种[8]或汽车模型的类型)代替类别.

FGVC 具有广泛的应用,一些比较直观的应用包括:1)自动实地指南:许多手机应用程序可以用于花、鸟、狗识别.无需寻找专家,用户所需要做的就是为花朵拍照,就可以立即获取有关花朵的更多信息,包括名称、购买地点和其他属性.2) 图像检索系统的自动标记:当前,网络上大多数带标签的图像都通过其 HTML 页面上的图片周围的文本进行注释.因此,没有任何文字描述的图像很少出现在搜索结果中.FGVC 可以自动为这些图像标注精细标签,并帮助它们出现在搜索结果中.这不仅适用于网络图像,还适用于个人相册.例如,Google+可让用户搜索相册中的关键字,而无需事先手动标记照片(plus.google.com).3) 一键式购物应用程序:与实地指南类似,购物应用程序旨在使客户更快地找到他们想要的商品,从而直接提高转化率和收入.客户只需为商品拍照就能够快速购买该商品.亚马逊最近推出了一种利用该技术的手机应用(developer.amazon.com/public/solutions/devices/fire-phone/docs/understanding-firefly).4) 机器人:FGVC 可以帮助机器人更准确地了解环境,例如,一个家用机器人需要详细了解冰箱中的东西,才能做出最好的早餐.

尽管 FGVC 有着重要的应用前景,细粒度数据集存在以下几个难

点:1)较小的类间差异和较大的类内差异.细粒度数据集中的许多类别只能通过细微的细节分开,例如鸟头顶上的毛色.图 1 中第 1 行在鸟的域中显示了 4 种这样的情况:尽管外观相似,但每种类别却不同.此外,属于同一类别的不同图像可能具有不同的姿态、视角、形状和颜色等,从而使它们看起来非常不同.例如:图 1 第 2 行包含 4 张属于同一类别的鸟类图像,细粒度分类器需要为所有图像预测相同的标签.因此较大的类内差异和较小的类间差异给 FGVC 带来了非常严峻的挑战.2)冗余背景信息.在一般物体分类问题中,图像的背景通常有助于分类,例如:对飞机和汽车之间进行分类时,作为背景的天空或街道具有十分丰富的信息.但是,在 FGVC 大多数情况下,背景不仅没有价值反而是噪声的来源.例如,在花域中,背景通常由灰尘和树叶组成.3)图像条件.FGVC 同样也具有一般视觉识别中共有的困难,例如亮度、杂波、遮挡等.亮度会随着一天中的时间的变化产生很大差异,而杂波会带来令人混淆的背景.遮挡是指只有部分感兴趣的对象在图像中.更进一步来说,由于 FGVC 的许多最终应用程序都是基于移动端的,因此图像质量也受到分辨率低、聚焦不清晰、模糊等问题的困扰.

综上所述,细粒度视觉分类任务具有广阔的应用前景,但同时也具有难以克服的挑战.

1 数据集

本文介绍 5 种常见的基准 FGVC 数据集:加州理工学院的鸟类(CUB)数据集[8]、斯坦福汽车(Cars)数据集[5]、FGVC 飞机(Aircraft)数据集[9]、斯坦福狗种类(Dogs)数据集[4]和花类数据集(Oxford Flowers)[7].CUB 包含 200 种鸟类的 11 788 张图像,它们之间存在细微的类间差异.所有图像被分为两部分:5 994 张图像用于训练以及 5 794 张图像用于测试.同样,Cars 数据集包含来自 196 类汽车的 8 144 张训练图像和8 041 张测试图像.Aircraft 数据集包含来自 100 类飞机的 6 667 张用于训练的图像和 3 333 张用于测试的图像.Dogs 数据集包含来自 120 个狗品种的数据,其中 12 000 张图像用于训练,8 580 张图像用于测试.最后一个 Flowers 数据集包含 102 种不同的花卉种类,共 8 189 张图像,其中 6 149 训练图片、1 020 张验证图像和 1 020 张测试图像.

2 基于局部检测的细粒度分类方法

根据专家的意见可知,细小的类间差异往往存在于物体特定的局部区域处,比如鸟嘴部形状和尾巴毛色.因此,语义的部位检测可以明确地定位与对象相关的细微外观差异来促进细粒度分类.换句话说,定位局部位置对于抵消对象姿态变化和摄像机视图位置变化之间所带来的类别混淆性是至关重要的.许多方法遵循图 2 中所示的流程:首先定位物体的重要局部位置,例如鸟类的头部和躯干;然后进行局部位置对齐;最后在对齐部位上提取特征并进行分类.在本小节中,我们将介绍近期相关的基于局部检测的 FGVC 方法.

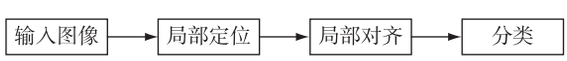

图 2 基于局部检测的一般流程

Fig. 2 General flow chart of part-based FGVC methods

图 1 视觉上相似的不同鸟类图片和类内差异明显的同一种鸟类图片

Fig. 1 Visually similar bird images of different categories (a-d), and visually different bird images of the same category(e)

2.1 基于局部检测的 R-CNN

在许多方法中被广泛使用的深度卷积局部检测器能够有效地改善 FGVC 性能.一种常见的方法 R-CNN 在文献[10]中被提出:通过自下而上区域合并的方法来学习部位检测器.R-CNN 扩展了 RCNN [11]来检测对象,并利用几何先验来提升定位精度.整个过程如图 3 所示:首先使用选择性搜索来定位候选区域;随后提取深度卷积特征训练部位检测器.在测试期间,所有候选区域都由检测器评分,并且使用非参数几何约束来选择最佳对象和部件检测.最后对区域进行姿态标准化来进行细粒度识别.

不同于基于局部区域的 R-CNN,多候选框集合[12]使用改进的 AlexNet[2]架构来同时预测任何给定图像的所有关键点位置及其可见性(头部、躯干、身体).具体来说,AlexNet 的 FC8 层被两个独立的输出层替换,分别用于关键点定位和可见性分析.网络在候选框区域[13]上进行训练,并使用 ImageNet[1]数据集上预训练的模型进行初始化.在获得关键点预测及其可见性之后,具有低可见性的信息将被删除,从而只保留具有区分性的视觉信息

姿态归一化网络[14]首先计算对象姿态的估计值用于辅助局部图像特征提取,而这些局部特征最终会反过来用于分类.在训练过程中,姿态标准化网络使用 DPM[15]来预测 2D 位置和 13 个语义部分关键点的可见性,或者直接使用预先提供的对象边界框和部分注释来学习姿态原型.实验证明,将底层区域特征与高层标准化全局特征相结合的模型会得到更好的分类效果.

2.2 基于视觉注意力的方法

注意力的存在是人类视觉系统最令人好奇的方面之一.注意力系统不是将整个图像压缩为静态表示,而是使重要特征根据需求动态地排列在前面.当图像中有许多干扰时,这一点尤其重要.视觉注意力机制还用于许多细粒度图像分类系统中.

2.2.1 基于强化学习的注意力定位

FCN 注意力[16]是基于强化学习的全卷积注意力定位网络,用于自适应选择多个任务驱动的视觉注意力区域.与以前的基于强化学习的模型[17-18]相比,该方法运用全卷积的架构来提升训练和测试过程中的计算效率,并且能够同时将注意力集中在多个视觉注意力区域上.具体框架如图 4 所示,它可以使用注意力机制来定位多个物体部位.不同的部位可以具有不同的预定义尺寸.该网络包含部位定位部分和分类部分.

部位定位部分使用全卷积神经网络来定位重要部位的位置.给定输入图像,该方法首先使用 ImageNet 预训练的 VGG 16 模型提取卷积特征图,并针对目标细粒度数据集进行微调;随后使用注意力定位网络为每个部位生成一张分数图来定位多个部位;最后选择具有最高概率的注意力区域作为部位位置.分类部分为每个部位以及全局图像都训练一个 CNN 分类器.不同的部位可能具有不同的大小,并且根据部位的大小在每个部位位置周围裁剪局部图像区域.每个局部图像区域的分类器以及整个图像的分类器分开进行训练.最终分类结果为各个分类器结果的平均值.

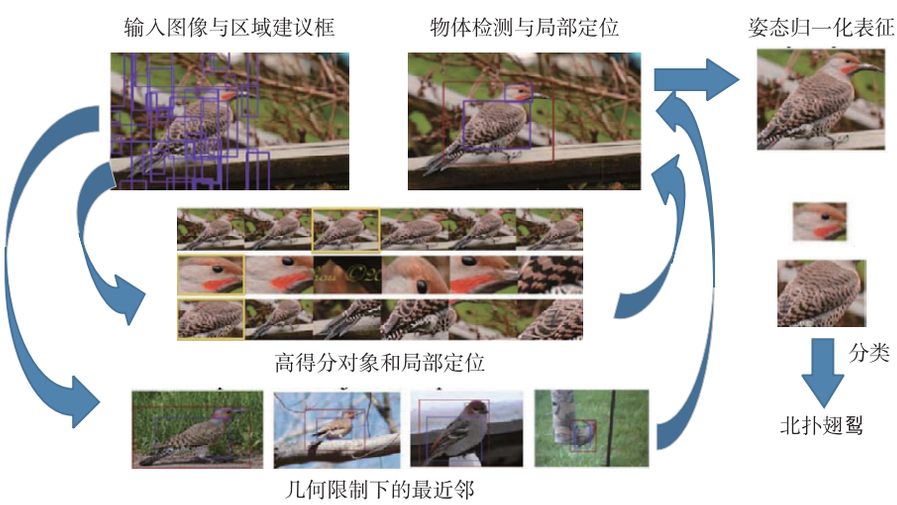

图 3 用于细粒度图像分类的基于局部的 R-CNN[10]

Fig. 3 Part-based R-CNN for FGVC[10]

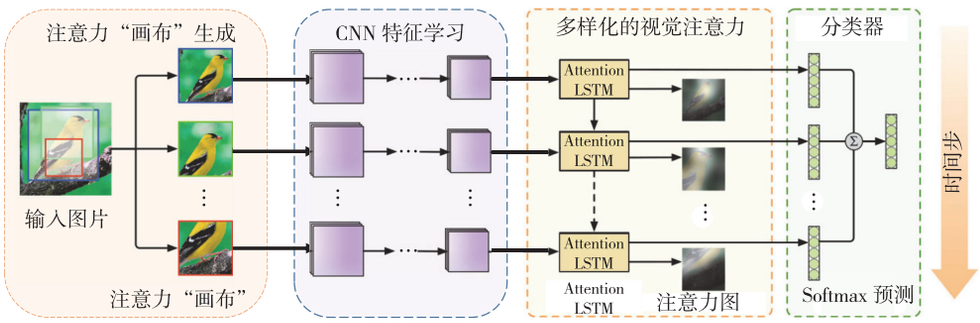

图 4　多样化的注意力视觉网络[19]

Fig. 4　Framework of diversified visual attention networks[19]

2.2.2　多样化的注意力

除此之外,多样化的视觉注意力网络在文献[19-20]中被提出,并用于追求多样性的注意力机制来最大程度地收集可区分性信息.在图 4 中描述的多样化注意力网络模型包括 4 个组成部分:注意力区域生成、CNN 特征学习、多样化的视觉注意力和分类模块.多样化注意力网络首先以不同的比例定位输入图像的多个区域,并将作为之后视觉注意力的区域.然后采用卷积神经网络(即 VGG-16[3])从每张注意力区域中学习卷积特征.为了在每个局部区域中定位物体的重要部位或组成部分,引入了多样化的视觉注意组件以预测注意力图,以便突出每个“画布”中的重要位置,并使多个注意力“画布”的信息得到最大化.与专注于单个区分性位置的传统注意力模型不同,多样化注意力网络能够通过损失函数的精心设计来确定不同的位置.根据生成的注意力图,卷积特征将被动态池化并累积到多样化的注意力模型中.

2.3　半监督局部区域检测算法

尽管上述方法都在 FGVC 中取得了不错的效果,但它们依赖于额外的密集局部标记(也称为关键点定位)来定位对象的语义的关键部位(例如头部和躯干).然而,获得如此密集的局部标记对于数据标注者来说意味着巨大的工作量,这限制了现实生活中细粒度应用的可扩展性和实用性.因此,最近的工作更倾向于只需要图像级标签就能够捕捉细粒度类别之间共享的语义部分(例如,头部和躯干).例如,Jaderberg 等[21]介绍了一种新的空间迁移模块,能够主动将特征映射变换后来进行区域定位.Simon 等[22]通过使用神经网络找到一系列相似的神经激活模态来学习部位检测模型.为了定位可区分的对象所在区域,He 等[23]提出通过联合显著性提取和共同分割来学习针对整个对象的检测器.在最近的研究中人们发现,局部定位和特征生成可以相互增强彼此得到更好的综合表现,因此 Fu 等[24]使用注意力模块以相互增强的方式递归地定位感兴趣区域并提取到重要的视觉图像特征.Zheng 等[25]提出了一种新的基于半监督多注意力模块的局部学习方法,能够同时进行局部生成和特征学习.Yao 等[26]设计两个互补的局部级和对象级的视觉描述模块,用来提取鲁棒的和具有区分性的视觉描述.Peng 等[27]采用多视点和多尺度特征融合的方式,通过两级(对象级和局部级)注意力机制来增强特征表达.虽然这些方法能够融合全局和局部的特征生成更强大的图像表达,但它们需要采用额外的部位检测器的多阶段训练,具有更复杂的训练过程.接下来我们具体介绍几种最新、最常用的半监督局部区域检测算法.

2.3.1　两级注意力

如图 5 所示,两级注意力模型[17]包含 3 种类型的注意力:基于候选区域的自下而上注意力、基于目标相关区域的自上而下的注意力以及可区分性区域的自上而下注意力.这些注意力类型被组合在一起用来训练特定域的深度网络,并用于查找前景物体或对象区域来提取可判别的特征.该模型易于泛化,因为它不需要对象边界框或区域标注.

类似的方法在文献[18](深度循环神经网络)中也被提到.深度循环神经网络会对输入图像执行多分辨率裁剪.网络使用“瞥(glimpse)”的信息来更新输入的内部表示,并输出序列中的下一个“瞥(glimpse)”位置以及可能的下一个物体.有了序列化的注意力机制,深度循环神经网络 RNN 和强大的视觉网络(GoogLeNet)进行细粒度分类.整个系统将任何大小的图像作为输入,并使用 softmax 分类器输出 N 路分类得分.

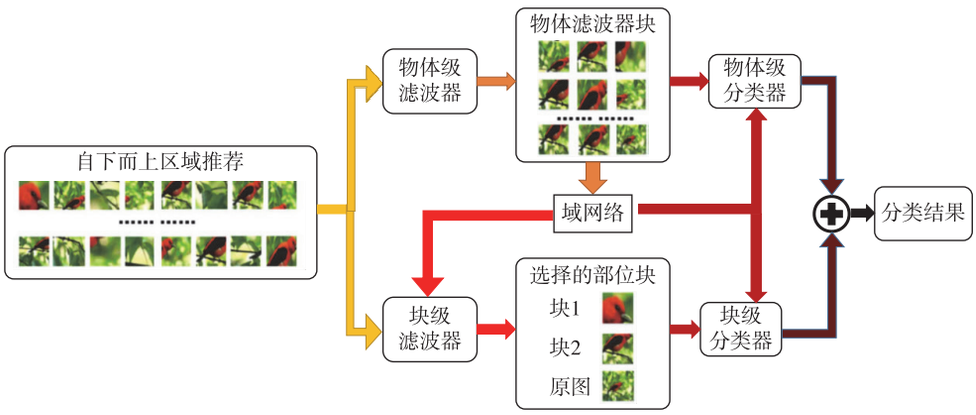

图 5　两级注意力的框架[17]

Fig. 5　Bi-level attention framework[17]

2.3.2　半监督多注意力卷积神经网络

Zheng 等[25]提出了一种基于多注意力卷积神经网络 MA-CNN(Multi-Attention Convolutional Neural Network)用于无边界框/局部标注的细粒度识别. MA-CNN 学习并结合每个局部候选区域和局部特征表示来得到描述性更强的图像特征.与人类定义的语义局部区域不同,这里的语义区域被定义为图像中具有较强辨别能力的多个注意力区域. MA-CNN 由卷积、通道分组和局部分类 3 个子网络组成,它们输入全图像并生成多个候选区域,其具体架构如图 6 所示.

1)网络特征的各个通道通常对应于某种类型的视觉响应模式.因此,通道分组子网络根据通道的峰值响应出现的位置来聚集并加权空间相关模式成为局部注意力图.多样化的高响应位置进一步构成了多部分注意力图,并将其裁剪为固定大小来提取多个局部建议.2)一旦获得了局部建议,分类网络将进一步根据局部特征对图像进行分类.这些局部特征是从全卷积特征映射中经过空间池化得到的.特别地,这样的设计可以通过去除对其他区域的依赖性来优化与某一部分区域相关的一组特征通道,从而可以更好地学习该区域上的细粒度特征.3)两种优化损失函数共同指导通道分组和局部分类的学习,这促使 MA-CNN 根据特征通道生成更多的具有区分性的局部区域,并以相互增强的方式从局部中学习更多细粒度特征.具体来说,他们提出了一种通道分组损失函数来优化通道分组子网络,将空间区域上的高类内相似性和类间可分性的通道簇作为局部注意力考虑,从而可以产生紧凑和多样的局部建议.

2.3.3　半监督循环注意力卷积神经网络

Fu 等[24]提出了一种新的循环注意力卷积神经网络 RA-CNN(Recurrent Attention Convolutional

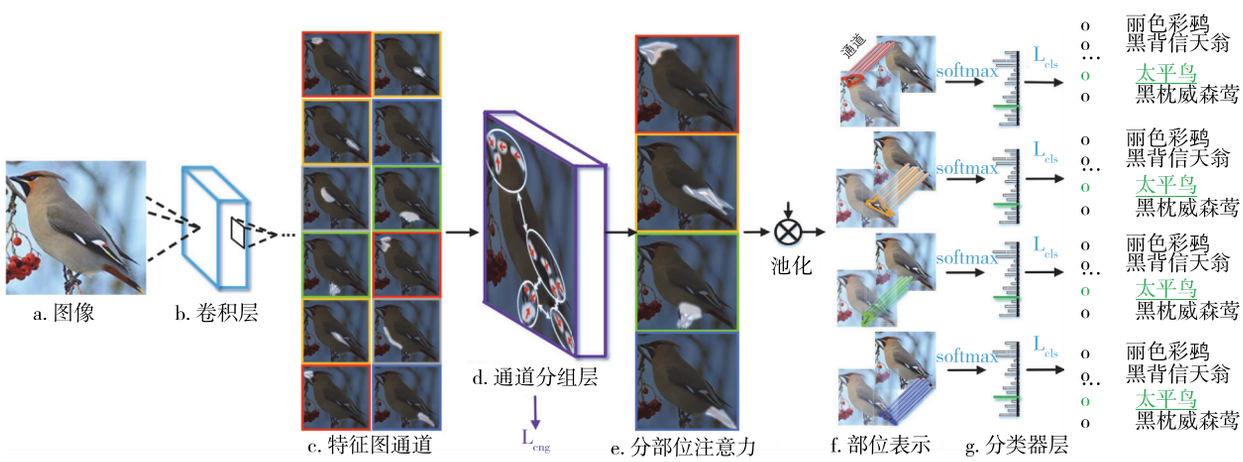

图 6　多注意力神经网络(MA-CNN)[25]

Fig. 6　Framework of multi-attention convolutional neural network (MA-CNN)[25]

Neural Network)用于无边界框/局部标注的细粒度识别.RA-CNN 以一种相互增强的方式循环地学习有区分性的区域注意力和基于区域的特征表示.RA-CNN 是一个多层网络,它输入完整图像和多个尺度的细粒度局部区域.1)多尺度网络共享相同的网络架构,但在每个尺度上具有不同的参数,以适应具有不同分辨率的输入(例如,图 7 中的粗粒度尺度和细粒度尺度).在每个尺度上的模型由一个分类子网和一个注意力建议子网(APN)组成,这可以保证每个尺度上有足够的辨别能力并为下一个更精细的尺度生成一个精确的注意区域.2)专用于高分辨率区域的一个更精细的网络将放大的注意区域作为输入,以提取更细粒度的特征.3)递归网络交替优化尺度内的分类的 softmax 损失和尺度间的注意力建议网络的成对排序损失.排序损失优化更精细的网络,在正确的类别上产生比先前预测更高的置信分数.由于更精细的网络可以以循环的方式堆叠,RA-CNN 可以从粗粒度到细粒度(例如,从身体到头部,然后到鸟喙)逐渐关注最具辨别力的区域.注意,精确的区域定位有助于基于区域的特征识别,反之亦然.因此,该网络可以受益于区域定位和特征学习之间的相互加强.为了进一步发挥整体学习的优势,网络学习一个全连接融合层,对多尺度特征进行深度融合,最终对图像进行分类.

3 基于端到端的视觉编码方法

与基于局部的方法不同,端到端的视觉特征编码倾向于通过增强图像的全局视觉信息,来直接学习到更具有辨别力的全局表征.本章首先介绍几种传统的端到端特征编码方法,然后介绍目前主流的双线性模型.

3.1 多特征融合网络

子集特征学习网络[28]由一个域通用卷积神经网络和几个特定的卷积神经网络所组成.域通用卷积神经网络首先在与目标数据集相同域的大规模数据集上进行预训练,然后再在目标数据集上进行微调.该网络使用带有线性判别分析(LDA)的 fc6 特征来降低其维数,把在视觉上相似的种类聚集成 K 个子集,以训练第 2 部分中的多个特定 CNN.需要注意的是,K 个预聚类子集中的每一个都用一个单独的 CNN 来进行学习,其目的是为学习每个子集的特征用于区分视觉上相似的物体.每个单独 CNN 的 fc6 特征用于学习不同的子集特征.

与子集特征学习网络类似,MixDCNN[29]系统也将学习 K 个特定的 CNN.但是,它不需要将图像预先划分为 K 个相似图像子集.图像将被输入到所有的 K 个 CNN 中,每个 CNN 的输出组合成一个单独的分类决策.与子集特征学习网络不同,MixDCNN 采用响应概率方程对 K 个 CNNs 同时进行端到端联合训练.

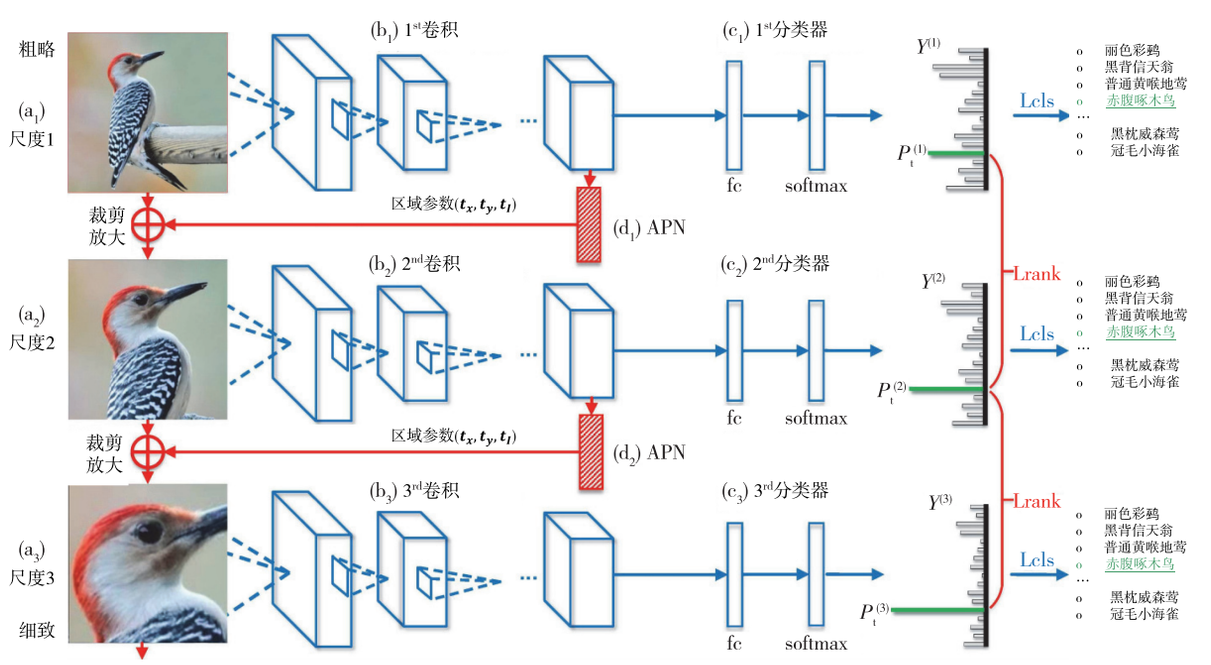

图 7 循环注意力卷积神经网络[24]

Fig. 7 Framework of recurrent attention convolutional neural network (RA-CNN)[24]

响应概率定义为

$$\alpha_k = \frac{e^{C_k}}{\sum_{c=1}^{K} e^{C_c}}, \tag{1}$$

其中 C_k 是第 k 个 CNN 的最佳分类结果.占用概率给予对置信度大的模块更高的权重.每个子集的占用概率是基于每个组件的分类置信度,这使得联合训练 K 个 DCNN 组件成为可能,而不必像子集特征学习网络那样估计一个单独的标签向量 $\boldsymbol{y}$ 或训练一个单独的门控网络.分类是通过将每个组件的最终层输出乘以响应概率,然后将 K 个分量求和来完成的.最终网络输出混合在一起,然后通过应用 softmax 函数生成每个类的概率.

3.2 多粒度 CNN

可以观察到,从属级标签带有一个隐含的标签层次结构,每个标签对应于数据域中的一个级别.例如,melanerpes formicivorus 也被称为橡树啄木鸟,在属级也可以被称为 melanerpes,在科级也可以被称为啄木鸟科.这些标签可以方便地提取它们相应有判别性的图像块和特征.这些标签可用于训练一系列基于 CNN 的分类器,每个分类器专门用于一个粒度级别.这些网络的内部表示具有不同的兴趣区域,构造多粒度描述符,这些描述符编码涵盖所有粒度级别的信息和鉴别特性.

基于此思想,多粒度 CNN[30] 包含一组并行的深度卷积神经网络,每个并行的神经网络都被优化以在给定的粒度下进行分类.换句话说,多粒度 CNN 由一组单粒度描述符组成.从自底向上图像块的公共池中,隐藏层的显著性引导了感兴趣的区域(ROI)的选择.因此,按照定义来看 ROI 选择与粒度有关,因为所选的图像块是给定粒度的相关分类器的结果.同时,ROI 的选择也依赖于跨粒度:更细粒度的 ROI 通常从较粗粒度的 ROI 中采样.最后,将每个粒度的 ROI 输入框架的第 2 阶段,以提取每个粒度的描述符,然后合并这些描述符以给出分类结果.

3.3 双线性网络模型

双线性模型[31](Bilinear CNN)是目前最主流的端到端视觉编码框架.Bilinear CNN 由两个特征提取器组成,其输出在图像的每个位置使用外积相乘,并汇集在一起以获得图像描述符.如图 8 所示,该体系结构可以以平移不变的方式对局部成对特征相互作用建模,这对于细粒度分类特别有用.用于图像分类的双线性模型由四元组 $\beta = (f_A, f_B, P, C)$ 组成.这里 f_A 和 f_B 是特征函数,P 是池化函数,C 是分类函数.特征函数表示映射 $f: L \times I \rightarrow \mathbf{R}^{c \times D}$,其将图像 I 和位置信息 L 作为输入并输出大小为 $c \times D$ 的特征,$\mathbf{R}$ 为实数集.位置信息通常包括方位和尺度.在每个位置使用矩阵外积组合得到特征输出,即位置 l 处 f_A 与 f_B 的双线性特征组合由

$$\text{bilinear}(l, I, f_A, f_B) = f_A(l, I)^{\mathrm{T}} f_B(l, I) \tag{2}$$

给出.f_A 和 f_B 输出的特征维度为 c.为了获得图像描述符,池化函数 P 聚合了图像中所有位置的双线性特征.池化的一种选择是简单地对所有双线性特征求和,即

$$\phi(I) = \sum_{l \in L} \text{bilinear}(l, I, f_A, f_B). \tag{3}$$

另一种选择是 max-pooling.两者都忽略了特征的位置,因此是无序的.如果 f_A 和 f_B 分别提取大小为

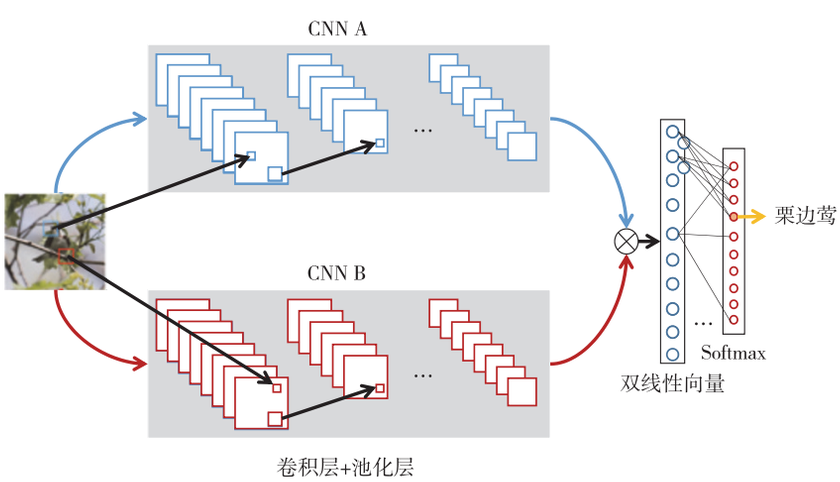

图 8 双线性 CNN 模型框[31]

Fig. 8 Framework of bilinear CNN[31]

$C \times M$ 和 $C \times N$ 的特征,则 $\phi(I)$ 的大小为 $M \times N$.变换 $\phi(I)$ 使其大小为 $MN \times 1$,而获得的双线性向量是一个通用的图像描述符,它可以与分类函数 C 一起使用.直观地说,通过考虑与二次核扩展类似的所有成对相互作用,双线性形式允许特征提取器 f_A 和 f_B 的输出彼此调节.

特征函数 f 的天然候选者是由卷积层和池化层的层次结构组成的 CNN.在很多论文中,作者使用了两个不同的仅包含非线性项的 CNN 作为特征抽取器,并在 ImageNet 数据集上预先训练.通过预训练,双线性深度网络模型将在特定域数据稀缺的情况下受益于额外的训练数据.从物体检测、纹理识别到细粒度分类的许多识别任务都证明预训练对于双线性 CNN 是至关重要的[32-33].仅使用卷积层的另一个优点是所得到的 CNN 可以在单个前向传播过程中处理任意大小的图像,并产生由图像和特征信道中的位置索引的输出.

Bilinear CNN 是第 1 个可以端到端训练的协方差池化网络模型[34].它对协方差矩阵进行 $L2$ 归一化之后采用了元素平方根归一化,在细粒度识别任务上达到很好的结果.紧凑的双线性池化(CBP)[6]阐明了双线性池化与二阶多项式核密切相关,并通过低维特征映射给出了两种紧凑的核逼近表示.内核池[35]近似于高斯 RBF 核,通过紧凑的显式特征映射到一个给定的阶,旨在刻画高阶特征间的相互作用.Cai 等[36]提出了一种基于多项式核的预测器,用于多层卷积特征的高阶统计建模.改良的 B-CNN[37]指出在原有归一化操作之前进行额外的矩阵平方根归一化能够达到更好的效果.在训练阶段,他们在前向传播中使用牛顿-舒尔兹迭代或者使用 SVD 分解,在反向传播通过求解李雅普诺夫方程或者计算 SVD 相关的梯度,执行反向传播.其前向传播公式为

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^{\mathrm{T}}, \tag{4}$$

其中 $\boldsymbol{U}$ 是双线性特征的特征向量,$\boldsymbol{\Lambda}$ 为双线性特征的特征值.在这种情况下,通过矩阵平方根优化后的双线性特征具有更好的特征稳定性.然而改良后的 B-CNN 由于 SVD、SCHUR 和 EIG 方法的使用,并不能很好地利用 GPU.因此迭代式的矩阵平方根协方差池化(Iterative Matrix Square Root Normalization of Covariance Pooling,iSQRT-COV)[38]提出了一种更高效的矩阵正则化算法.相较于文献[31],iSQRT-COV 有 3 点不同:1)他们的前向传播和反向传播都是基于牛顿-舒尔兹迭代的,由于只涉及 GPU 矩阵乘法,所以网络训练高效,其具体前向和反向传播公式为

$$\boldsymbol{Y}_k = \frac{1}{2}\boldsymbol{Y}_{k-1}(3\boldsymbol{I} - \boldsymbol{Z}_{k-1}\boldsymbol{Y}_{k-1}),$$

$$\boldsymbol{Z}_k = \frac{1}{2}(3\boldsymbol{I} - \boldsymbol{Z}_{k-1}\boldsymbol{Y}_{k-1})\boldsymbol{Z}_{k-1}, \tag{5}$$

其中,$\boldsymbol{Y}$ 是优化后的双线性特征,$\boldsymbol{Z}$ 为辅助变量.2)他们在牛顿-舒尔兹迭代前后使用了预归一化和后补偿,这对深度 ConvNets 的训练至关重要.3)iSQRT 在大规模的 Imagenet 数据集和 3 个通用的细粒度数据集上进行了评估.

针对大规模视觉识别问题,文献[39]提出了一种矩阵幂归一化协方差池化方法(MPN-COV).在 AlexNet、VGG-Net 和 ResNet[40]架构上,它超越了一阶池化方法实现了重大的提升.MPN-COV 指出,在给定少量高维特征的情况下,矩阵幂与鲁棒协方差估计的收缩原理一致,通过冯诺依曼正则的最大似然估计[41],可以得到作为鲁棒协方差估计的矩阵平方根.结果表明,矩阵幂归一化近似而有效地利用了协方差矩阵流形的几何性质,在高维特征上会优于矩阵对数归一化.除了需要在 CPU 上运行 EIG,MPN-COV 层的所有计算都可以利用英伟达 cuBLAS 库在 GPU 上运行.

文献[42]在 ConvNets 中添加了全局高斯分布,并可以端到端训练.在全局高斯分布嵌入网络(Global Gaussian Distribution Embedding Network,G^2 DeNet)中,根据高斯流形[43]的 Lie 群结构,将每个高斯分布定义为对称正定矩阵的平方根.矩阵平方根能够获得有竞争力的结果,其具体网络框图如图 9 所示.

4 用外部信息辅助 FGVC

为了识别各种细粒度类别之间的细微差别,模型需要足够多地标记良好的训练图像.然而,由于标注的困难(总是需要领域专家)和细粒度类别的繁多(即,在一个元类别中有数千个从属类别),获取精确的细粒度类别的人类标注并非易事.

因此,部分细粒度识别方法试图利用免费但有噪声的 web 数据来提高识别性能.这一流派现有的工作大致可以分为两个方向.一种方法是将测试类别中带有噪声的标记 web 数据作为训练数据,这种方法称之为 webly 监督学习[44-45].这些方法的主要工作集中在:1)克服了易获取的 web 图像与有良好标注的标准数据集之间的差距;2)减少噪声数据带来的负面影响.为了解决上述问题,常用的方法有使用

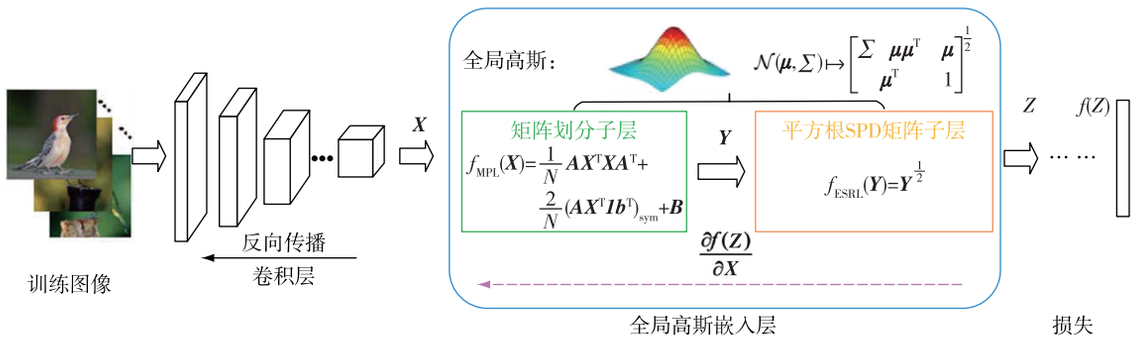

图 9　全局高斯嵌入网络的结构示意图[42]

Fig. 9　Overview of the Global Gaussian Distribution embedding Network (G^{2}DeNet)[42]

对抗学习的深度学习技术[45]和注意机制[44].使用 web 数据的另一种方法是将具有良好标记的训练数据的辅助类别的知识转移到测试类别,而测试类别通常采用 zero-shot learning[46]或 meta learning[47]来实现这一目标.

随着多媒体数据(如图像、文本、知识库等)的快速增长,多模态分析受到了广泛的关注.在细粒度识别中,需要多模态数据建立联合表示/嵌入合并多模态信息.它能够提升细粒度识别的准确性.特别是,包括文本描述(如自然语言的句子和短语)和图形结构知识库在内频繁使用的多模态数据.与细粒度图像的强监督(如注释部分)相比,文本描述则是弱监督.此外,文本描述可以由普通人准确地标注,而无需某个特定领域的专家.此外,高层的知识图谱是一种现有的资源,包含丰富的专业知识,例如 DBpedia[48].实际上,文本描述和知识库都可以作为额外指导,有效地进行细粒度图像表示学习.

具体而言,文献[49]收集了文本描述,并引入了一种结构化的联合嵌入方法,将文本和图像信息结合起来进行零样本细粒度图像识别.后来,文献[50]以端到端的方式联合训练,将视觉和语言结合起来,以用于生成互补的细粒度表达.对于基于知识库的细粒度识别,有一些工作[51-52]引入知识库信息(始终与属性标签关联,参见图 10)来隐式地学习嵌入空间,并同时对细粒度对象的判别属性进行推理.

5　实验分析与性能比较

本文使用上述深度学习方法在最常见的 CUB200-2011 数据集上进行评估实验,并在表 1 中列出了他们的 Top-1 分类准确度.分类准确度定义为类别分类准确度的平均值.需要注意的是,我们只比较了基于类别标签监督的方法.同时,一些方法由于未在该数据集上报告结果或者效果不佳,因此被省略了.

从表 1 中可以看出,这些方法分为两组.第一组中的方法基于部位检测和定位,第二组中的方法基于端

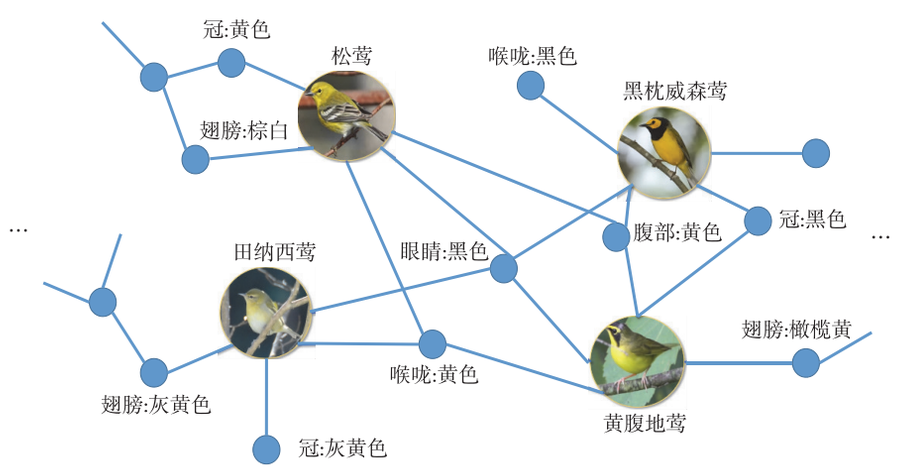

图 10　在 CUB200-2011 数据集上对类别属性关联性建模的知识图谱例子[51]

Fig. 10　An example of knowledge graph about category relationship in CUB200-2011 dataset[51]

到端视觉编码以提高分类性能.这些方法基于不同的基础神经网络,例如 VGG[3]、DenseNet[54] 或 ResNet[40].从结果中我们可以发现,基于局部检测的细粒度识别算法取得了不错的结果.其原因在于:局部区域往往包含更丰富和可区分的视觉信息,因此能更准确地区分细小的类间差异.然而大部分基于局部检测的方法都无法做到端到端的训练,他们需要先进行区域检测,然后再将全局和局部信息进行有效的融合,来得到更强的视觉表征.基于端到端视觉编码的方法往往具有更好的泛化能力.由于省掉了区域检测的步骤,这些方法具有更简单的训练策略,更快的运行速度,同时也拥有不俗的识别效果.如表 1 中所示,最新的 iSQRT 方法达到了目前第一梯队的识别精度,并且该方法也被移植到了很多其他方法中,证明了端到端视觉编码方法优异的可迁移性.

6 总结与展望

近年来,基于深度学习的细粒度图像识别(FGVC)取得了很大进展.本文通过深度学习对 FGVC 的最新进展进行了广泛的调查.主要介绍了 FGVC 的相关问题及其面临的挑战,讨论了该领域的重大改进,并介绍了与 FGVC 相关的一些领域应用.尽管该领域已经取得了巨大成功,但仍有许多未解决的问题.因此,本章节将明确指出这些问题,并介绍一些未来发展的研究趋势.希望这篇综述论文不仅可以帮助读者了解 FGVC,还可以促进该领域未来的研究活动和应用开发.从该领域的发展趋势来看,有以下几个研究热点:

1)自动细粒度模型.如今,自动机器学习(AutoML)[55] 和神经架构搜索(neural architecture searching-NAS)[56] 正在引起人工智能社区的热切关注,特别是在计算机视觉领域.AutoML 旨在自动化地将机器学习应用于实际任务的端到端流程.最近的 AutoML 和 NAS 方法可以在各种计算机视觉应用中产生具有可比性或甚至优于手工设计的架构.因此,通过 AutoML 或 NAS 技术开发的自动细粒度模型可以找到更好、更多量身定制的深度模型,同时它可以依次推进 AutoML 和 NAS 的研究.

2)细粒度的少样本学习.人类能够在很少的监督下学习一个新的细粒度概念,但是最好的深度学习细粒度系统需要数百或数千个标记的例子.更糟糕的是,对细粒度图像进行标记监督既耗时又昂贵,因为细粒度数据始终由相应领域专家准确标记.因此,开发细粒度的少样本学习算法[57] 具有着广泛的

表 1 在鸟类、汽车和飞机数据集上的评估结构,评估指标是分类准确率(%),VGG-D 包含 16 和 19 层,DenseNet-D 包含 161 和 201 层,ResNet-D 包含 50 和 101 层

Table 1 Evaluation results of different methods on CUB,Cars,and Aircraft datasets,the metric is classification accuracy (%), VGG-D contains 16 and 19 layers,DenseNet-D contains 161 and 201 layer settings,ResNet-D contains 50 and 101 layer settings

Methods		Backbone	CUB	Cars	Aircraft
基于局部检测的方法	TLAN[16]	AlexNet	77.9		
	MG-CNN[30]		81.7	86.6	
	ST-CNN[21]		84.1		89.1
	FCAN[18]		82.0		
	RA-CNN[24]	VGG-D	85.3	88.2	92.5
	MA-CNN[25]		86.5	89.9	92.8
端到端视觉编码方法	CBP[6]		84.3	91.2	84.1
	LR-BCNN[53]		84.2	90.9	87.3
	KP[35]		86.2	92.4	86.9
	DeNeT[42]	VGG-D	87.1	92.5	89.0
	Impro.BCNN[37]		85.8	90.0	88.5
	HIHCA[36]		58.3	91.7	88.3
	iSQRT[35]		87.2	92.5	90.0
	CBP[41]	ResNet	83.1	87.9	83.2
	iSQRT[38]		88.5	92.1	90.9
	CBP[6]		81.6	88.6	81.6
	KP[35]	ResNet-D	84.7	91.1	85.7
	iSQRT[38]		88.7	93.3	91.4

社区需求.

3)细粒度的哈希方法.随着对细粒度图像检索越来越多的关注,更多构造良好的大规模细粒度数据集[58-60]已经被公开发布.在诸如细化图像检索之类的实际应用中,自然会产生这样的问题:在参考数据库非常大的情况下,找到精确最近邻居的成本非常高.哈希[59,61]作为近似最近邻搜索的最流行和最有效的技术之一,具有处理大规模精细化数据的潜力.因此,细粒度哈希是一个值得进一步探索的有前途的方向.

4)向更现实的环境中进行细粒度分析.在过去十年中,很多开发的细粒度图像分析相关技术在传统环境中实现了良好的性能[4-5,62].然而,仅在这些环境下成功无法满足当今各种现实应用的日常需求,例如,通过在受控图像环境中训练的产品识别模型[60]难以识别/检测野外[61]的自然物种.因此,新颖的细粒度图像分析趋势,应该是一些域自适应的模型算法,即用知识转移进行细粒度分析,用长尾分布进行细化分析,以及在资源受限的嵌入式设备上运行细粒度分析.

参考文献
References

[1] Deng J,Dong W,Socher R,et al.ImageNet:a large-scale hierarchical image database[C]// IEEE Conference on Computer Vision and Pattern Recognition,2009:248-255
[2] Krizhevsky A,Sutskever I,Hinton G.ImageNet classification with deep convolutional neural networks[C]// Advances in Neural Information Processing Systems,2012:1097-1105
[3] Simonyan K,Zisserman A.Very deep convolutional networks for large-scale image recognition[J]. arXiv Preprint,2014,arXiv:1409.1556
[4] Khosla A, Jayadevaprakash N, Yao B P, et al. Novel dataset for fine-grained image categorization[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2011
[5] Krause J,Stark M,Jia D,et al.3D object representations for fine-grained categorization[C]// IEEE International Conference on Computer Vision Workshops, 2013:554-561
[6] Gao Y, Beijbom O, Zhang N, et al. Compact bilinear pooling[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2016:317-326
[7] Nilsback M E, Zisserman A. Automated flower classification over a large number of classes[C]// Sixth Indian Conference on Computer Vision, Graphics & Image Processing,2008:722-729
[8] Wah C,Branson S,Welinder P,et al.The Caltech-UCSD birds-200(2011 dataset)[R]. Computation & Neural Systems Technical Report,CNS-TR-2011-001,California Institute of Technology,Pasadena,CA,2011
[9] Maji S, Rahtu E, Kannala J, et al. Fine-grained visual classification of aircraft[J]. arXiv Preprint, 2013, arXiv:1306.5151
[10] Zhang N, Donahue J, Girshick R, et al. Part-based R-CNNs for fine-grained category detection[C]// European Conference on Computer Vision,2014:834-849
[11] Girshick R,Donahue J,Darrell T,et al.Rich feature hierarchies for accurate object detection and semantic segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition,2014:580-587
[12] Shih K J,Mallya A,Singh S,et al.Part localization using multi-proposal consensus for fine-grained categorization[J].arXiv Preprint,2015,arXiv:1507.06332
[13] Zitnick C L, Dollár P. Edge boxes: locating object proposals from edges[C]// European Conference on Computer Vision,2014:391-405
[14] Branson S, Van Horn G, Belongie S, et al. Bird species categorization using pose normalized deep convolutional nets[J].arXiv Preprint,2014,arXiv:1406.2952
[15] Branson S, Beijbom O, Belongie S. Efficient large-scale structured learning[C]// IEEE Conference on Computer Vision and Pattern Recognition,2013:1806-1813
[16] Liu X,Xia T,Wang J,et al.Fully convolutional attention localization networks: efficient attention localization for fine-grained recognition[J]. arXiv Preprint, 2016, arXiv:1603.06765
[17] 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]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2015:842-850
[18] Mnih V, Heess N, Graves A, et al. Recurrent models of visual attention[J]. arXiv Preprint, 2014, arXiv:1406.6247
[19] Zhao B,Wu X,Feng J S,et al.Diversified visual attention networks for fine-grained object classification[J].IEEE Transactions on Multimedia,2017,19(6):1245-1256
[20] Ba J,Mnih V,Kavukcuoglu K,et al.Multiple object recognition with visual attention[J].arXiv Preprint,2014,arXiv:1412.7755
[21] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks[C]// Advances in Neural Information Processing Systems,2015:2017-2025
[22] Simon M,Rodner E.Neural activation constellations:unsupervised part model discovery with convolutional networks[C]// IEEE International Conference on Computer Vision (ICCV),2015:1143-1151
[23] He X T,Peng Y X.Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification[C]// Thirty-First AAAI Conference on Artificial Intelligence,2017:4075-4081
[24] Fu J L,Zheng H L,Mei T.Look closer to see better:recurrent attention convolutional neural network for fine-grained image recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2017:4438-4446

[25] Zheng H L,Fu J L,Mei T,et al.Learning multi-attention convolutional neural network for fine-grained image recognition [C] // IEEE International Conference on Computer Vision(ICCV),2017:5209-5217
[26] Yao H T,Zhang S L,Yan C G,et al.AutoBD:automated Bi-level description for scalable fine-grained visual categorization[J].IEEE Transactions on Image Processing,2018,27(1):10-23
[27] Peng Y X,He X T,Zhao J J.Object-part attention model for fine-grained image classification [J]. IEEE Transactions on Image Processing, 2018, 27 (3): 1487-1500
[28] Ge Z Y,McCool C,Sanderson C,et al.Subset feature learning for fine-grained category classification[C] // IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),2015:46-52
[29] Ge Z Y,Bewley A,McCoolC,et al.Fine-grained classification via mixture of deep convolutional neural networks [C]//IEEE Winter Conference on Applications of Computer Vision (WACV),2016:1-6
[30] Wang D Q,Shen Z Q,Shao J,et al.Multiple granularity descriptors for fine-grained categorization[C]//Proceedings of IEEE International Conference on Computer Vision,2015:2399-2406
[31] Lin T Y,Roychowdhury A,Maji S.Bilinear CNN models for fine-grained visual recognition [C] // IEEE International Conference on Computer Vision (ICCV), 2015:1449-1457
[32] Cimpoi M,Maji S,Kokkinos I,et al.Describing textures in the wild[C]//IEEE Conference on Computer Vision and Pattern Recognition,2014:3606-3613
[33] Donahue J,Jia Y Q,Vinyals O,et al.DeCAF:a deep convolutional activation feature for generic visual recognition [J].arXiv Preprint,2013,arXiv:1310.1531
[34] Ionescu C,Vantzos O,Sminchisescu C.Matrix back-propagation for deep networks with structured layers[C] // IEEE International Conference on Computer Vision(ICCV),2015:1-2,6
[35] Cui Y,Zhou F,Wang J,et al.Kernel pooling for convolutional neural network [C] // IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017:2921-2930
[36] Cai S J,Zuo W M,Zhang L.Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization[C]//IEEE International Conference on Computer Vision(ICCV),2017:511-520
[37] Lin T Y,Maji S.Improved bilinear pooling with CNNs [J].arXiv Preprint,2017,arXiv:1707.06772
[38] Li P H,Xie J T,Wang Q L,et al.Towards faster training of global covariance pooling networks by iterative matrix square root normalization[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2018:947-955
[39] Li P H,Xie J T,Wang Q L,et al.Is second-order information helpful for large-scale visual recognition? [C]// Proceedings of the IEEE International Conference on Computer Vision,2017:2070-2078
[40] He K M,Zhang X Y,Ren S Q,et al.Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016:770-778
[41] Wang Q L,Li P H,Zuo W M,et al.RAID-G:robust estimation of approximate infinite dimensional Gaussian with application to material recognition [C] // IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2016:4433-4441
[42] Wang Q L,Li P H,Zhang L.G2DeNet:Global Gaussian distribution embedding network and its application to visual recognition [C] // IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 2730-2739
[43] Li P H,Wang Q L,Zeng H,et al.Local log-Euclidean multivariate Gaussian descriptor and its application to image classification[C] // IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39 (4): 803-817
[44] Zhuang B H,Liu L Q,Li Y,et al.Attend in groups:a weakly-supervised deep learning framework for learning from web data [C] // IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 1878-1887
[45] Sun X X,Chen L Y,Yang J F.Learning from web data using adversarial discriminative neural networks for fine-grained classification[J].Proceedings of the AAAI Conference on Artificial Intelligence,2019,33:273-280
[46] Niu L,Veeraraghavan A,Sabharwal A.Webly supervised learning meets zero-shot learning:a hybrid approach for fine-grained classification[C] // IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 7171-7180
[47] Zhang Y B,Tang H,Jia K.Fine-grained visual categorization using meta-learning optimization with sample selection of auxiliary data[C]//European Conference on Computer Vision,2018:233-248
[48] Lehmann J,Isele R,Jakob M,et al.DBpedia:a large-scale,multilingual knowledge base extracted from Wikipedia[J].Semantic Web Journal,2015,6(2):167-195
[49] Reed S,Akata Z,Lee H,et al.Learning deep representations of fine-grained visual descriptions[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2016:49-58
[50] He X T,Peng Y X.Fine-grained image classification via combining vision and language[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017:5994-6002
[51] Chen T S,Lin L,Chen R Q,et al.Knowledge-embedded representation learning for fine-grained image recognition [C]//Twenty-Seventh International Joint Conference on Artificial Intelligence,2018:627-634
[52] Xu H P,Qi G L,Li J J,et al.Fine-grained image classification by visual-semantic embedding [C] // Twenty-Seventh International Joint Conference on Artificial Intelligence,2018:1043-1049
[53] Kong S,Fowlkes C.Low-rank bilinear pooling for fine-grained classification [C] // IEEE Conference on Computer Vision and Pattern Recognition (CVPR),

2017:365-374
[54] Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017:4700-4708
[55] Feurer M, Klein A, Eggensperger K, et al. Efficient and robust automated machine learning[C]// Advances in Neural Information Processing Systems, 2015:2962-2970
[56] Elsken T, Metzen J H, Hutter F. Neural architecture search: a survey [J]. arXiv Preprint, 2018, arXiv:1808.05377
[57] Wei X S, Wang P, Liu LQ, et al. Piecewise classifier mappings: learning fine-grained learners for novel categories with few examples[J]. IEEE Transactions on Image Processing, 2019, 28(12):6116-6125
[58] Berg T, Liu J X, Lee S W, et al. Birdsnap: large-scale fine-grained visual categorization of birds[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014:2019-2026
[59] van Horn G, Aodha O M, Song Y, et al. The iNaturalist species classification and detection dataset[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017:8769-8778
[60] Wei X S, Cui Q, Yang L, et al. RPC: a large-scale retail product checkout dataset [J]. arXiv Preprint, 2019, arXiv:1901.07249
[61] Wang J D, Zhang T, Song J K, et al. A survey on learning to Hash[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4):769-790
[62] Li W J, Wang S, Kang W C. Feature learning based deep supervised hashing with pairwise labels[C]// Twenty-Fifth International Joint Conference on Artificial Intelligence, 2016:1711-1717

A survey of deep fine-grained visual categorization

DENG Xuran[1] MIN Shaobo[1] XU Jingyuan[1] LI Pandeng[1] XIE Hongtao[1] ZHANG Yongdong[1]
1 University of Science and Technology of China, School of Information Science and Technology, Hefei 230026

Abstract Fine-grained image classification is a fundamental and important task in field of computer vision. The purpose of the task is to distinguish between object categories that have subtle inter-class differences (e.g., birds, flowers, or animals of different sub-categories). Different from traditional image classification tasks that can employ a large number of common people for image annotations, fine-grained image classification usually requires expert-level knowledge. In addition to the common classification challenges of pose, lighting, and viewing changes, fine-grained datasets have larger inter-class similarity and intra-class variability. Therefore, it puts a high demand on the models to capture the subtle visual differences between classes and common intra-class characteristics. Furthermore, owing to the difficulty in obtaining samples of different categories, fine-grained datasets suffer from long-tail distribution problem. In summary, fine-grained data distribution has the characteristics of small, non-uniform, and indistinguishable inter-class differences, which also poses a huge challenge to the powerful deep learning algorithms. In this paper, we first introduce the formulation and challenges of fine-grained visual categorization tasks, and then illustrate two mainstream methods about local features and global features, as well as their advantages and disadvantages. Finally, we compare the performance of related works on common used datasets, and we make the required summarization and forecast.

Key words fine-grained visual categorization; deep learning; part region detection; bilinear pooling