基于深度卷积特征的细粒度图像分类研究综述

罗建豪 [1] 吴建鑫 [1]

摘 要 细粒度图像分类问题是计算机视觉领域一项极具挑战的研究课题, 其目标是对子类进行识别, 如区分不同种类的鸟. 由于子类别间细微的类间差异和较大的类内差异, 传统的分类算法不得不依赖于大量的人工标注信息. 近年来, 随着深度学习的发展, 深度卷积神经网络为细粒度图像分类带来了新的机遇. 大量基于深度卷积特征算法的提出, 促进了该领域的快速发展. 本文首先从该问题的定义以及研究意义出发, 介绍了细粒度图像分类算法的发展现状. 之后, 从强监督与弱监督两个角度对比分析了不同算法之间的差异, 并比较了这些算法在常用数据集上的性能表现. 最后, 我们对这些算法进行了总结, 并讨论了该领域未来可能的研究方向及其面临的挑战.

关键词 细粒度图像分类, 深度学习, 卷积神经网络, 计算机视觉

引用格式 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. 自动化学报, 2017, **43**(8): 1306−1318

DOI 10.16383/j.aas.2017.c160425

A Survey on Fine-grained Image Categorization Using Deep Convolutional Features

LUO Jian-Hao[1] WU Jian-Xin[1]

Abstract Fine-grained image categorization is a challenging task in the field of computer vision, which aims to classify sub-categories, such as different species of birds. Due to the low inter-class but high intra-class variations, traditional categorization algorithms have to depend on a large amount of annotation information. Recently, with the advances of deep learning, deep convolutional neural networks have provided a new opportunity for fine-grained image recognition. Numerous deep convolutional feature-based algorithms have been proposed, which have advanced the development of fine-grained image research. In this paper, starting from its definition, we give a brief introduction to some recent developments in fine-grained image categorization. After that, we analyze different algorithms from the strongly supervised to and weakly supervised ones, and compare their performances on some popular datasets. Finally, we provide a brief summary of these methods as well as the potential future research direction and major challenges.

Key words Fine-grained image categorization, deep learning, convolutional neural networks, computer vision

Citation Luo Jian-Hao, Wu Jian-Xin. A survey on fine-grained image categorization using deep convolutional features. *Acta Automatica Sinica*, 2017, **43**(8): 1306−1318

细粒度图像分类 (Fine-grained image categorization), 又被称作子类别图像分类 (Sub-category recognition), 是近年来计算机视觉、模式识别等领域一个非常热门的研究课题. 其目的是对粗粒度的大类别进行更加细致的子类划分, 但由于子类别间细微的类间差异和较大的类内差异, 较之普通的图像分类任务, 细粒度图像分类难度更大.

细粒度图像分类研究, 从提出到现在, 已经经历了一段较长时间的发展. 早期的基于人工特征的算法, 由于特征的表述能力有限, 分类效果也往往面临很大的局限性. 近年来, 随着深度学习的兴起, 深度卷积特征促进了该领域的快速进步. 另一方面, 由于该课题本身的困难性, 传统的方法不得不依赖于大量的人工标注信息, 严重制约了算法的实用性. 因此, 越来越多的算法倾向于不再依赖人工标注信息, 仅仅使用类别标签来完成分类任务, 这也是该领域逐渐发展成熟的标志.

本文以卷积特征为线索, 从细粒度图像分类的概念出发, 以鸟类数据库[1] 上的发展历程为轴线, 介绍了该领域一些优秀的算法, 并探讨了未来可能的研究方向.

文章剩余部分的内容组织如下: 在第 1 节, 我们将对细粒度图像分类进行简要、系统的介绍. 一些比较常用的数据库将在第 2 节给出, 以便对细粒度分类问题有个更直观的理解. 在第 3 节, 我们将从其发展历程出发, 简要回顾一些基于人工特征的早期算法. 由于本文介绍的大多数算法均基于卷积神经网络, 因此在第 4 节, 我们会对卷积神经网络进行必要的介绍说明. 之后, 在第 5 节和第 6 节, 我们将从

收稿日期 2016-05-25 录用日期 2017-02-03
Manuscript received May 25, 2016; accepted February 3, 2017
国家自然科学基金 (61422203) 资助
Supported by National Natural Science Foundation of China (61422203)
本文责任编委 王亮
Recommended by Associate Editor WANG Liang
1. 南京大学计算机科学与技术系南京大学软件新技术国家重点实验室 南京 210023
1. National Key Laboratory for Novel Software Technology, Department of Computer Science and Technology, Nanjing University, Nanjing 210023

强监督、弱监督两个角度, 对现有的一些基于深度卷积特征的优秀算法展开介绍. 所谓强监督就是指模型中使用了除标签以外的人工标注信息, 而弱监督则仅仅只使用了类别标签. 关于该领域未来可能的发展方向及其挑战, 我们将在第 7 节进行介绍. 最后, 相关的总结会在第 8 节给出.

1　细粒度图像分类概述

图像分类是计算机视觉领域的一个经典研究课题. 传统的图像分类主要处理语义级图像和实例级图像两大类: 前者包括诸如场景识别[2−3]、对象识别[4−5] 等语义级图像分类任务, 其目标是识别不同类别的对象, 如猫和狗等; 后者则是对不同的个体进行分类, 如人脸识别[6−7].

细粒度图像分类则位于这两者之间. 不同于对象识别等粗粒度的图像分类任务, 细粒度图像的类别精度更加细致, 类间差异更加细微, 往往只能借助于微小的局部差异才能区分出不同的类别. 由于其分类边界位于同一类别的不同子类之上, 如不同种类的鸟[1] 等, 故而又被称作子类别分类. 而与人脸识别等对象级分类任务相比, 细粒度图像的类内差异更加巨大, 存在着姿态、光照、遮挡、背景干扰等诸多不确定因素. 因此, 细粒度图像分类是一项极具挑战的研究任务.

细粒度图像分类无论在工业界还是学术界都有着广泛的研究需求与应用场景. 与之相关的研究课题主要包括识别不同种类的鸟[1]、狗[8]、花[9]、车[10]、飞机[11] 等. 以鸟类数据库为例, 单就燕鸥而言, 就存在着北极燕鸥、里海燕鸥等数种不同类别之分. 而这些不同种类的燕鸥之间的差异十分细微, 挖掘出有用信息也更加困难. 如图 1 (a) 所示, 这两种燕鸥的差别仅仅只存在于一些局部区域中, 如翅膀和脚的颜色等. 即便是对于鸟类专家而言, 想要完全识别出这些不同种类的鸟也不是一件容易的事情. 但在实际生活中, 识别不同的子类别又存在着巨大的应用需求. 例如, 在生态保护中, 有效识别不同种类的生物, 是进行生态研究的重要前提. 以往, 这一工作只能依赖于领域专家知识, 造成了研究成本的大幅增加. 如果能够借助于计算机视觉的技术, 实现低成本的细粒度图像识别, 那么无论对于学术界, 还是工业界而言, 都有着非常重要的意义.

不同于普通的图像分类任务, 细粒度图像的信噪比很小, 包含足够区分度的信息往往只存在于很细小的局部区域中. 因此, 如何找到并有效利用这些有用的局部区域信息, 成为了决定细粒度图像分类算法成功与否的关键所在. 目前, 绝大多数的分类算法都遵循这样的流程框架: 首先找到前景对象 (鸟) 及其局部区域 (头、脚、翅膀等), 之后分别对这些区域提取特征. 对所得到的特征进行适当的处理之后, 用来完成分类器的训练和预测.

如前所述, 细粒度图像分类是一项极具挑战的研究课题, 为了达到满意的分类结果, 许多现有的分类算法都严重依赖于人工标注信息 (Annotations). 一些常用的标注信息如图 1 (b) 所示, 其中主要包括标注框 (Bounding box) 和局部区域位置 (Part locations) 等. 借助于标注框能够完成对前景对象的检测, 从而排除掉背景噪声的干扰; 而局部区域位置则可以用来对一些有用的局部区域进行定位, 或者进行姿态对齐等, 以实现局部特征的提取. 然而, 人工标注信息的获取代价十分昂贵, 这在很大程度上制约了这些分类算法的实用性. 近年来, 越来越多的研究倾向于不使用这些标注信息, 仅仅依靠类别标签 (Label) 来完成图像分类任务[12−13], 并取得了相当不错的效果.

(a) 类间差异包含于局部区域中
(a) The intra-class difference exists in subtle local areas

(b) 标注框与局部位置点
(b) Bounding box and part locations

图 1　细粒度图像分类示意图 (取自 CUB200-2011 数据集[1])

Fig. 1　Illustration of fine-grained categorization (sampled from the CUB200-2011 dataset[1])

另一方面, 特征的提取也是决定图像分类准确性的关键因素, 寻找一个更具区分度的特征一直是研究人员所追求的目标[14−16]. 传统的基于人工特征的分类算法往往面临很大的局限性. 这类算法一般是先从图像中提取 SIFT (Scale invariant feature transform)[17] 或者 HOG (Histogram of oriented gradient)[18] 这些局部特征, 之后利用 VLAD (Vector of locally aggrgeated descriptors)[19] 或者 Fisher vector[20−21] 等编码模型进行特征编码, 得到最终所需要的特征表示. 然而, 由于人工特征的描述能力有限, 导致分类效果不佳. 在细粒度图像分类研究的早期, 特征的表示能力成为了制约其性能提升的主要瓶颈.

近年来, 深度学习, 尤其是深度卷积神经网络在计算机视觉领域的巨大成功, 引发了人们浓厚的研究兴趣[22−23]. 相对于人工特征而言, 深度学习可以看作是一个表示学习的过程[24], 即针对具体的分类任务学习一个具体的特征表示. 有研究表明, 从深度卷积神经网络中所提取的特征, 比人工特征拥有更强大的描述能力, 将深度卷积特征运用到细粒度图像分类任务中, 能够取得更好的结果[25]. 深度卷积特征的加入, 为细粒度图像分类的发展带来了新的机遇, 使得其研究进入了一个新的阶段.

2 细粒度图像数据库介绍

相对于普通分类任务的数据库而言, 细粒度图像数据库的获取难度更大, 需要更强的专业领域知识才能完成数据的采集与标注. 但近年来, 涌现出了越来越多的细粒度图像数据库, 这也从另一个角度反映了该领域蓬勃的发展趋势与强烈的现实需求.

目前比较常用的细粒度图像数据库主要包括: 1) CUB200-2011[1]: CUB200-2011 是细粒度图像分类领域最经典, 也是最常用的一个数据库, 共包含 200 种不同类别, 共 11 788 张鸟类图像数据. 同时, 该数据库提供了丰富的人工标注数据[1], 每张图像包含 15 个局部区域位置, 312 个二值属性, 1 个标注框, 以及语义分割图像. 2) Stanford Dogs[8]: 该数据库提供了 120 种不同种类的狗的图像数据, 共有 20 580 张图, 只提供标注框这一个人工标注数据. 3) Oxford Flowers[9]: 分为两种不同规模的数据库, 分别包含 17 种类别和 102 种类别的花. 其中, 102 种类别的数据库比较常用, 每个类别包含了 40 到 258 张图像数据, 总共有 8 189 张图像. 该数据库只提供语义分割图像, 不包含其他额外标注信息. 4) Cars[10]: 提供 196 类不同品牌不同年份不同车型的车辆图像数据, 一共包含有 16 185 张图像, 只提供标注框信息. 5) FGVC-Aircraft[11]: 提供 102 类不同的飞机照片, 每一类别含有 100 张不同的照片, 整个数据库共有 10 200 张图片, 只提供标注框信息.

图 2 展示了以上所介绍的几个数据库的部分示意图. 对于每个数据库, 我们随机采集了 4 张来自不同类别的图像. 从这些图像中可以看出, 不同类别之间的差异十分细微, 即便是对于人类自身而言, 也很难完全区分开这些类别. 细粒度图像分类任务的困难性, 由此可见一斑.

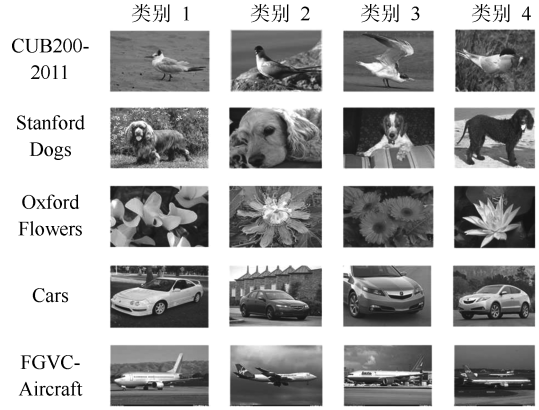


图 2 细粒度图像数据库示意图 (所有图像均取自不同类别)

Fig. 2 Illustration of fine-grained datasets (the images are sampled from different categories)

除了以上介绍的 5 个数据库之外, 相关的数据库还有很多, 这里限于篇幅, 不再一一细述. 需要说明的是, 尽管不同数据库的规模和难易程度不尽相同, 但其背后所蕴含的算法思想却是相类似的. 在一个数据库上能够取得良好性能的分类算法, 在其余数据库上往往也能生效. 而在这众多的细粒度图像数据库中, CUB200-2011 鸟类数据库是最常用, 也是最经典的一个. 因此, 本文将以该数据库为主线, 介绍细粒度图像分类的发展历程.

3 基于人工特征的早期算法简述

如前所述, 相对于普通的图像分类任务, 细粒度图像分类更具挑战性. 其发展的过程也见证了计算机视觉研究领域的一些重要进展. 在本节, 我们将简要地回顾该领域中的一些早期研究成果, 以加深对该领域的认识.

在发布 CUB200-2011 数据库[1] 的技术报告中, Wah 等给出的基准测试的结果仅为 10.3 %. 他们的方法是: 给定一张原始的、未经过裁剪的测试图像, 利用训练得到的模型完成局部区域的定位; 之后, 提取 RGB 颜色直方图和向量化的 SIFT 特征, 经过词包 (Bag of words, BoW) 模型进行特征编码后, 输入到线性 SVM (Support vector machine) 分类器完成分类. 如果在测试时给定了标注框和局部区域位置这些标注信息的话, 利用同样的方法, 得到的基准测试结果为 17.3 %.

从分类准确度上来看, 这个结果并不让人满意. 一方面, 是由于定位不够准确, 局部区域无法归一化对齐; 另一方面, 则是因为特征的描述能力太弱, 不具备足够的区分度. 之后, 研究人员发现, 使用一些更强大的特征, 如 POOF[26]、Fisher-encoded[27] SIFT、KDES (Kernel descriptors)[28] 等, 再利用一定的算法提高定位的精确度, 能够将分类准确度提升至 50 %～62 % 左右[26, 29−31].

其中, Berg 等[26] 提出了一种基于局部区域的特征编码方式, 他们称之为 POOF 特征. 该算法能够自动发现最具区分度的信息, 取得了不错的分类效果. 但该算法对关键点的定位精度要求比较高, 如果用精确的标注信息实现定位的话, 能够达到 73.3 % 的准确率, 但如果利用定位算法去确定关键点的话, 则只有 56.8 % 的准确度. 除了特征之外, 也有针对局部区域的算法研究. 如 Yao[32] 等, Yang[33] 等均尝试使用模板匹配的方法来减少滑动窗口的计算代价.

除此之外, 也有研究工作[34−35] 尝试将人加入到分类任务中来. 用户通过交互式的询问对答, 完成指定的操作, 如给出关键点, 回答一些简单问题等. 其目的在于使用最少的询问次数, 达到最好的分类精度. 这类算法在小样本规模问题上不失为一种折中方案, 对于精度要求比较高的任务而言, 可作为一

[1] 本文将监督信息分成类别标签与人工标注信息两大类. 对于分类任务而言, 类别标签是必不可少的监督信息; 而人工标注信息则主要是指标注框、语义分割图像等额外监督信息.

种合理的补充.

从这一阶段的研究成果上可以看出, 更强大的特征描述和特征编码方式对分类准确度有着显著的影响, 随后关于卷积特征的研究也再次证实了这一点. 其次, 细粒度图像分类有别于其他分类任务的一点就是局部区域的信息是至关重要的. 因此, 设计一个更加精确的定位/对齐模型, 也能带来显著的性能提升. 但同时, 我们也该意识到, 为了实现更精细的局部定位, 很多算法都严重依赖于人工标注信息, 这样的方式在实际应用中存在很大的局限性, 这也是前期研究的一个共性.

4　深度卷积神经网络概述

由于下文所介绍的算法均基于深度卷积特征, 因此有必要对其进行一定的说明. 在本节, 我们将会从网络结构、卷积特征以及模型的训练方法几个方面对卷积神经网络展开必要的介绍.

4.1　卷积神经网络结构

卷积神经网络 (Convolutional neural networks, CNNs) 是神经网络中一个非常经典的模型[36−37], 于上世纪 80 年代受视觉神经运作机制的启发而设计. 其典型的网络结构如图 3 所示:

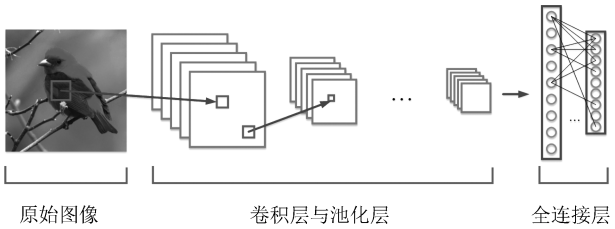


图 3　卷积神经网络框架图

Fig. 3　The framework of convolutional neural networks

在卷积神经网络中, 前若干层由卷积层和池化层组成, 前层的输出作为后层的输入. 其中, 卷积层由一个大小固定的卷积核与输入进行卷积操作, 用来模拟生物视觉系统中的简单细胞, 而池化层则是一种下采样操作. 用来扩大感受野 (Receptive fields), 获得一定的不变性. 随后的若干层由全连接层构成, 其作用相当于一个分类器. 由于其网络层数量较多, 故而称作深度卷积神经网络, 或者深度学习.

4.2　卷积特征

不同于传统的机器学习算法, 深度卷积神经网络将特征提取、模型训练等原本分散的操作结合在一起, 构成了一个端到端 (End-to-end) 的系统进行整体训练, 其巨大的参数数量保证了模型的有效性与强大的表示能力. 卷积层和池化层相当于一个特征提取的操作. 整个系统是一个端到端的训练过程, 即针对特定的分类任务, 利用大量的参数学习得到一个具体的特征表示. 因此, 与人工特征相比, 卷积神经网络获得的特征更加强大, 拥有更强的区分性.

研究表明, 前几层网络学习到的特征主要是一些边缘/纹理特征, 而随着神经网络层数的加深, 逐渐从这些低层语义特征过渡到了高层语义特征[38]. 在后几层, 空间信息保留的程度逐渐降低, 而到了全连接层, 则完全丢弃了空间语义信息. 因此, 不同网络层的特征具有不同的描述能力, 卷积特征的抽取需要综合考虑各方面因素[25].

从神经网络特定层提取的输出, 可以作为图像的特征来训练分类模型. Gong 等[39] 抽取全连接层的特征, 与 VLAD[19] 编码相结合, 取得了不错的效果. 考虑到全连接层丢失了空间信息, Cimpoi 等[40] 则尝试使用卷积层的输出作为特征, 并在纹理识别上取得了进步. 在实际应用中, 应该根据特定的需求来选取适当的网络层输出作为卷积特征.

4.3　模型训练方法

在实际应用中, 卷积神经网络的训练方法主要包含以下三种情况: 1) 预训练模型 (Pre-rrained model): 这种方法是直接使用一些在 ImageNet 数据集上已经训练好的模型, 比较常用的模型包括 Alex-Net[22]、VGG-Net[41] 等. 在这种情况下, 这些预训练的模型相当于一个特征提取器; 2) 模型微调 (Fine-tuned model): 由于深度卷积神经网络的特征数量非常庞大, 而特定任务 (如细粒度图像分类) 的数据集规模往往比较小, 若直接进行训练很容易造成过拟合. 一种折中的方法是使用在 ImageNet 上预训练的模型参数, 替换掉最后的 Softmax 层, 在新数据集上进行重新训练, 称之为微调. 在细粒度图像分类研究中, 模型微调是最常用的训练方法; 3) 从头训练 (Training from scratch): 以上两种方法可以被视作为一种迁移学习, 即将模型在 ImageNet 数据集上学习到的知识迁移到特定的数据集 (如 CUB200-2011) 上, 而从头训练则是自行设计网络结构并进行模型训练. 如前所述, 这种情况下很容易造成数据的过拟合, 需要采取一定的方法来避免.

5　强监督的细粒度图像分类研究

所谓强监督的细粒度图像分类算法, 是指在模型训练的时候, 除了图像的类别标签外, 还使用了标注框、局部区域位置等额外的人工标注信息. 如前所述, 由于标注信息的获取代价十分昂贵, 在很大程度上限制了这类算法的实用性. 因此, 也有些算法考虑仅在模型训练的时候使用标注信息, 而在进行图像分类时不使用这些信息. 这在一定程度上提高了算法的实用性, 但与只依赖类别标签的弱监督分类算法相比仍有一定的差距.

5.1　DeCAF

随着深度卷积神经网络在 ImageNet 上的成功,

越来越多的人将目光转向了深度学习. 一个很自然的想法就是, 在 ImageNet 上学习得到的知识能否迁移到其他的具体领域中来? 也就是说, 利用 ImageNet 上预训练的模型, 在其他数据集上提取图像特征, 是否仍然具有强大的区分性? 答案是肯定的.

Donahue 等[25] 通过对在 ImageNet 数据集上所训练得到的卷积网络模型进行分析, 发现从卷积网络中提取的特征具有更强的语义特性, 比人工特征具有更好的区分度. 他们将卷积特征迁移到其他具体领域的任务中, 如场景识别、细粒度分类等, 均获得了更好的分类性能, 从实验上证明了卷积特征强大的泛化性. 他们称之为 DeCAF 特征 (Deep convolutional activation feature).

具体而言, 首先使用标注框对图像进行裁剪, 得到前景对象, 再利用预训练的卷积网络对图像提取 DeCAF 特征. 在文献 [25] 中, 他们提取的是第 6 层网络特征, 即第一个全连接层的输出, 之后训练一个多类别的逻辑回归 (Logistic regression) 模型来进行图像分类. 这样一个简单的框架在 Caltech-UCSD 数据集[42] (CUB200-2011[1] 数据集的早期版本[2]) 上取得了 58.75 % 的分类精度, 超过了很多当时非常优秀的算法. 这也证明了从卷积网络中所提取的特征, 尽管不是为细粒度图像分类专门进行优化设计的, 却捕捉到了更丰富的图像信息.

总的来说, DeCAF 是比较前期的工作, 并不是专门针对细粒度图像分类所优化设计的算法, 其目的在于解释卷积特征的强大泛化性与领域自适应性. DeCAF 的出现, 在卷积特征与细粒度图像分类之间搭起了一座桥梁, 具有十分重要的意义. 如今, 越来越多的算法倾向于使用卷积特征来进行具体领域的图像处理工作, 并取得了很大的进步.

5.2 Part R-CNN

正如我们在前文所描述的那样, 对于细粒度图像分类而言, 图像的局部信息是决定算法性能的关键所在. 对图像进行检测, 并提取出重要的局部信息是大多数细粒度图像分类算法所采用的基本流程. 基于这种观点, Zhang 等提出了 Part R-CNN[43] 算法, 该算法采用了 R-CNN[44] 对图像进行检测. 因此, 在介绍该算法之前, 有必要对 R-CNN 做一个简要的说明.

5.2.1 R-CNN 算法

对象检测 (Objection detection)[45−46] 问题是计算机视觉领域一个非常重要的研究课题, 其目标是判定图像中是否存在特定的对象, 如车、人等, 并给出对象在图像中的位置信息. 基于卷积特征, Girshick 等提出了 R-CNN (Regions with CNN features) 算法[44].

该算法流程十分简单, 首先, 对于输入的图像, 采用自底向上的区域算法 (如 Selective search[47]) 产生 2 000 个区域候选 (Part proposals). 这些候选区域可能包含了想要检测的目标对象, 但绝大多数区域仅仅包含背景信息. 之后, 对每一个候选区域提取卷积特征, 用事先训练好的 SVM 模型来对每一个特征进行分类, 判断该候选区域中是否包含想要检测的对象. 这样, 每一个候选区域都能够计算得到一个相应的评分分值: $score = \omega^{\mathrm{T}}\phi(x)$. 其中 ω 是 SVM 的权重, $\phi(x)$ 是利用卷积网络从候选区域图像 x 中提取的特征. 利用此分值作为评估该候选区域属于某一类别的可能性. 如果某一候选区域与另一分值较高区域之间的 IoU (Intersection-over-union) 重叠值大于某一阈值的话, 则丢弃该低分值的区域, 即采用所谓的非极大抑制 (Non-maximum suppression) 策略. 同时, 分值低于某一阈值的区域也应当被丢弃. 最终所保留下来的区域即为该类的定位检测结果.

在实际应用中, 仍有一些具体的操作细节需要注意, 如卷积网络的微调、训练数据的划分等. 本文由于篇幅限制, 不再一一叙述, 详细可参照文献 [44].

5.2.2 Part R-CNN 算法

顾名思义, Part R-CNN 就是利用 R-CNN 算法进行对象 (鸟) 与局部区域 (头、身体等) 的检测, 图 4 给出了其总体的流程图.

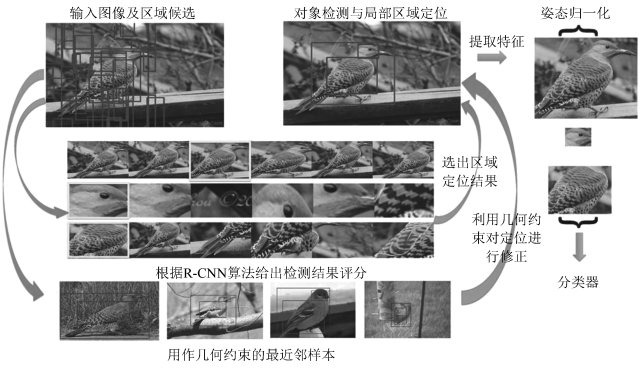


图 4 Part R-CNN 流程图[43]

Fig. 4 Part R-CNN system overview[43]

同 R-CNN 一样, Part R-CNN 也使用自底向上的区域算法 (如 Selective search[47]) 来产生区域候选, 如图 4 左上角所示. 之后, 利用 R-CNN 算法来对这些区域候选进行检测, 给出评分分值. 在这里, Part R-CNN 只检测前景对象 (鸟) 和两个局部区域 (头、身体). 之后, 根据评分分值 (图 4 中间) 挑选出区域检测结果 (见图 4 上方中间). 但 Zhang 等认为, R-CNN 给出的评分分值并不能准确地反映出每个区域的好坏. 例如, 对于头部检测给出的标注框可能会在对象检测的标注框外面, 身体检测的结果与头部检测的结果可能会有重叠等. 这些现象都会影响最终的分类性能. 因此, 需要对检测区域进行修

[2] 如不加说明, 后文均是在 CUB200-2011 数据集上的实验结果.

正.

具体而言, 用 $X = \{x_0, x_1, \cdots, x_n\}$ 表示标注框的位置信息, 其中 x_0 表示对象 (鸟) 的位置, x_1 到 x_n 分别表示 n 个局部区域位置 (头和身体). 通过求解式 (1) 所示的最优化问题来获得最佳的标注框位置:

$$X^* = \arg\max_X \Delta(X) \prod_{i=0}^{n} d_i(x_i) \tag{1}$$

其中, $\Delta(X)$ 表示评分函数, 我们稍后会对其进行介绍, $d_i(x_i) = \sigma(\omega_i^{\mathrm{T}} \phi(x_i))$ 表示对第 i 个区域所对应的 R-CNN 评分值求 Sigmoid 函数值.

关于评分函数 $\Delta(X)$ 有两种选择, 分别表示边框约束与几何约束, 其定义如下所示:

1) 边框约束: 该约束的出发点在于, 所有的局部区域的范围不能超出对象区域的某个阈值:

$$\Delta_{box}(X) = \prod_{i=1}^{n} c_{x_0}(x_i) \tag{2}$$

当局部区域 x_i 超出对象区域 x_0 的像素点个数不超过 ϵ 时, $c_{x_0}(x_i) = 1$; 否则, 取 0.

2) 几何约束: 由于单个检测器的结果不一定可靠, 几何约束在边框约束的基础上增加了额外的约束信息:

$$\Delta_{\text{geometric}}(X) = \Delta_{box}(X) \left(\prod_{i=1}^{n} \delta_i(x_i) \right)^{\alpha} \tag{3}$$

其中, α 是超参, δ_i 是对区域 i 位置的评分, 考虑两种不同的形式:

a) $\delta_i^{MG}(x_i)$ 对区域 x_i 求在训练数据上的混合高斯模型的值;

b) $\delta_i^{NP}(x_i)$ 首先找到与 x_0 最接近的 K 个近邻, 然后使用这 K 个近邻来训练混合高斯模型, 并求 x_i 的值.

利用如上所述的约束条件对 R-CNN 检测的位置信息进行修正之后, 再分别对每一块区域提取卷积特征, 将不同区域的特征相互连接起来, 构成最后的特征表示, 用来训练 SVM 分类器. 这里, 在进行网络训练时, 利用检测到的局部图像对网络进行了微调. 实验结果显示, 如果只在训练时提供标注框与局部区域信息, 测试时不提供任何信息的情况下, Part R-CNN 在 CUB200-2011 数据集上能够达到 73.89 % 的分类精度. 进行几何约束后可以带来 1 % 左右的效果提升, 而且 δ_i^{NP} 的效果最好.

相对于只是简单地引入卷积特征的 DeCAF 算法[25] 而言, Part R-CNN 的进步是明显的. 从局部区域的检测定位, 到特征的提取, 该算法均基于卷积神经网络, 并针对细粒度图像的特点进行改进优化, 以改进通用物体定位检测算法在该任务上的不足, 达到了一个相对比较高的准确度. 同时, 该算法进一步放松了对标记信息的依赖程度, 在测试时无需提供任何标记信息, 大大增强了算法的实用性. 其不足之处在于, 利用自底向上的区域产生方法, 会产生大量无关区域, 这会在很大程度上影响算法的速度. 另一方面, 该算法本身的创新性十分有限, 既然局部区域对于细粒度图像而言是关键所在, 那么对其进行定位检测则是必要的途径. 只是引入现有的通用定位算法, 似乎并不能很好地解决该问题.

5.3 姿态归一化 CNN (Pose normalized CNN)

在细粒度图像分类任务中, 除了至关重要的局部区域信息之外, 还有一个十分显著的特点: 其巨大的类内方差会对最终的分类性能造成很大的影响. 而在这些不同的干扰信息中, 姿态问题则是一个普遍存在的影响因素. 有鉴于此, Branson 等提出了姿态归一化 CNN (Pose normalized CNN) 算法[48]. 他们所采取的方案是: 对于每一张输入图像, 利用算法完成对局部区域的定位检测, 根据检测的标注框对图像进行裁剪, 提取出不同层次的局部信息 (鸟、头部), 并进行姿态对齐操作. 之后, 针对不同部位的局部信息, 提取出不同层的卷积特征. 最后, 将这些卷积特征连接成一个特征向量, 进行 SVM 的模型训练, 达到了 75.7 % 的分类精度. 其具体流程如图 5 所示.

整个算法流程中, 首先要解决的就是如何检测局部区域的问题. 对于输入图像, Branson 等利用预先训练好的 DPM (Deformable part model) 算法[49] 完成关键点的检测. DPM 算法能够给出预先定义好的关键位置点的坐标, 以及该点是否可见等信息. 之后, 利用这些关键点进行姿态对齐操作.

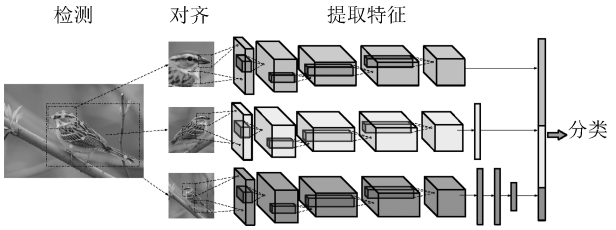


图 5　姿态归一化 CNN 流程图[48]

Fig. 5　Pose normalized CNN system overview[48]

具体而言, 给定 n 张训练图像, 每张图像包含 K 个关键点. 首先使用这些训练数据训练 P 个原型 (Prototype) $R_p = \{i_p, b_p, S_p\}$, 原型的个数代表不同局部区域的数量. 其中 i_p 表示一张参考图像, b_p 是相应的标注框, S_p 则是一系列关键点的位置信息. 给定一张测试图像 X_t, 利用 DPM 算法检测出关键点位置 Y_t 之后, 将其与原型中的位置点对齐. 这可以通过一个变换函数 $W(y_{tj}, \omega)$ 完成:

$$\omega_{tp}^* = \arg\min_{\omega \in W} \sum_{j \in S_p} E(y_{tj}, R_p, \omega) \tag{4}$$

其中, $E(y_{tj}) = \|\hat{y}_{i_pj} - W(y_{tj}, \omega)\|^2$ 表示像素对齐误差, 即变换后的坐标与原型里的坐标之间的误差, $\hat{y}_{i_pj}$ 表示原型进行归一化 (减去标注框左上角坐标, 再除以长/宽) 之后的新坐标, ω 是该变换函数的参数. 这样的变化函数有很多种选择, 例如简单变换、相似变换、仿射变换等. 这些变换都存在着闭式解, 因此式 (4) 能够十分高效地求解.

下面的问题变成了如何对 P 个原型 R_p 进行训练. 文献 [48] 给出的思路是使用受约束的最小化对齐误差, 其约束条件是训练集中的每一个关键点 y_{tj} 至少与一个原型对齐. 这一约束可以形式化地描述为

$$R^* = \arg\min_R \lambda P + \frac{1}{nK}\sum_{t=1}^{n}\sum_{j=1}^{K}\min_p E(y_{tj}, R_p, \omega_{tp}^*) \tag{5}$$

其中, 第一项表示对原型个数的惩罚项, 后一项是使得每一张图里的每一个关键点与原型的像素对齐误差最小化. 通过对该函数进行优化求解即可完成原型的训练过程.

由于不同网络层提取的特征包含不同的语义信息, Branson 等认为应该针对不同的局部区域提取不同网络层的卷积特征. 为了证明这一点, 他们比较了不同的局部区域在各个网络层提取的特征所能达到的分类准确度. 实验结果表明, 对于低层对齐图像 (原始图像与前景对象) 而言, 后层的卷积特征更具区分度, 能够实现更高的准确度, 相对浅层特征具有绝对的优势. 但对于高层对齐图像 (头部图像) 来说, 情况却恰恰相反. 因此, 对于不同的局部区域应当提取不同网络层的特征.

姿态归一化 CNN 的创新之处在于使用原型对图像进行了姿态对齐操作, 并针对不同的局部区域提取不同网络层的特征, 以试图构造一个更具区分度的特征表示, 这一方案在先前的研究工作中并不常见. 它在原有的局部区域模型的基础上, 进一步考虑了鸟类的不同姿态的干扰, 减轻了类内方差造成的影响, 从而取得了较好的性能表现. 但是, 该算法对于关键点的检测精度较为敏感, 利用 DPM 算法对关键点进行检测, 其精度为 75.7 %. 而如果在测试时使用真实的关键点标注信息, 则可以达到 85.4 %, 达到了一个相当高的分类水平.

5.4 其他

除了以上所介绍的算法之外, 还有很多优秀的算法, 如 Krause 等[50] 将协同分割[51−52] 引入到细粒度图像分类中来, 提出了一种新颖的局部区域检测算法. 该算法无需借助局部区域标注信息, 只依靠标注框, 便可完成分割与对齐操作, 实现了 82 % 的分类精度. 相类似的, Lin 等[53] 设计了一个新颖的系统, 在单个网络结构中同时实现了局部区域的定位、对齐与分类任务, 通过梯度回传的机制达到共同优化训练的目的, 实现了 80.26 % 的精度.

另一方面, 由于细粒度图像数据库的规模较小, 即便是对预训练的网络进行微调, 也难以避免过拟合带来的问题. 因此, 也有研究人员考虑使用数据增强的方式来扩大细粒度图像数据库的规模. 如 Xu 等[54] 提出利用网络图片来进行数据增强. 由于数据库的规模得到了扩充, 得到的网络也更加强大, 从而能够带来性能上的提升. 但网络图片包含了大量的干扰信息, 因此, Xu 等利用细粒度图像数据库上的标注信息来学习相应的检测器, 并利用检测器来对噪声图片进行过滤, 实现了 84.6 % 的分类精度.

借助于丰富的人工标注信息, 辅以精确的检测技术, 实现更高的分类精度已不再是难事. 但考虑到现实应用的实际需求, 随着研究的深入, 越来越多的算法不再依赖于这些强监督信息, 仅仅使用类别标签来完成分类任务, 这就是我们以下要介绍的弱监督的细粒度图像分类.

6 弱监督的细粒度图像分类研究

仅仅依赖于类别标签完成分类是近年来细粒度图像研究的一大趋势. 得益于深度学习的发展, 以及相关研究工作的深入, 不借助人工标注信息, 也能实现良好的分类性能. 如 Jaderberg 等[55] 和 Lin 等[13] 均实现了 84.1 % 的分类精度, 超过了绝大多数依赖于人工标注的分类算法.

从前文的讨论中可以看出, 对于细粒度图像分类算法而言, 局部区域信息是至关重要的, 这也正是大多数算法依赖于标注信息的一大原因. 因此, 要实现更好的弱监督的细粒度图像分类, 首先要解决的就是如何检测并定位这些局部区域.

6.1 两级注意力 (Two level attention) 算法

两级注意力 (Two level attention) 算法[56] 是第一个尝试不依赖额外的标注信息, 而仅仅使用类别标签来完成细粒度图像分类的工作, 由 Xiao 等提出, 取得了不错的分类效果. 顾名思义, 该模型主要关注两个不同层次的特征, 分别是对象级 (Object-level) 和局部级 (Part-level), 即在以往强监督工作中所使用的标注框和局部区域位置这两层信息.

该模型主要包含三个处理阶段, 对应于如下三个不同的子模型:

1) 预处理模型: 在预处理阶段, 主要是从原始图像中检测并提取前景对象, 以减少背景信息带来的干扰. 与 R-CNN[44] 相类似, Xiao 等使用一个卷积网络来对 Selective search[47] 产生的所有区域候选进行筛选, 检测该区域的图像中是否包含鸟类. 不同之处在于, R-CNN 只是用卷积网络来提取特征, 并针对具体检测目标专门训练一个 SVM, 根据评分结果来给出标注框的位置. 而 Xiao 等采取的方案是: 仅仅使用卷积网络来对背景区域进行过滤. 这样导致的结果是, 对于一张输入图像, 可能对应许多包

含前景对象的候选区域.

2) 对象级模型: 此模型的主要作用是对对象级图像进行分类. 经过预处理后, 得到了许多包含前景对象的图片, 可以用来从头开始训练一个卷积神经网络 (Training from scratch). 由于一张图像包含多个候选区域, 因此, 最终对一张图片的输出结果是一个集成 (Ensemble). 具体而言, 就是一张图的一个区域候选, 经过卷积网络之后, 得到一个 Softmax 层的输出. 对所有区域的输出求平均, 作为该图像最终的 Softmax 层输出. 值得注意的是, 对象级模型本身就是一个完整的分类方案, 但对于细粒度分类任务而言, 局部信息更加重要. 因此, 在对象级模型的基础上, 需要与局部级模型相结合, 才能实现最终的分类目标.

3) 局部级模型: 由于预处理模型选择出来的这些候选区域大小不一, 有些可能包含了头部, 有些可能只有脚. 因此, 局部级模型的作用就是为了选出这些局部区域. 首先利用对象级模型得到的网络来对每一个候选区域提取特征. 对这些特征进行谱聚类, 得到 k 个不同的聚类簇, 每个簇代表一个局部信息, 如头部、脚等. 于是, 每个簇都可以被看作一个区域检测器, 可以对测试样本的局部区域进行检测.

将不同局部区域的特征级联成一个特征向量, 用来训练 SVM, 作为局部级模型给出的分类器. 最后, 将对象级模型的预测结果与局部级模型的结果相结合, 作为模型的最终输出, 达到了 69.7 % 的精度. 需要说明的是, 这是在 Alex-Net[22] 上的实验结果, 如果采用更强大的网络结构如 VGG-Net[39], 则能将分类准确率提升到 77.9 %. 这也从另一个角度说明了特征对于图像分类算法的重要性.

总体上来看, 两级注意力模型较好地解决了在只有类别标签的情况下, 如何对局部区域进行检测的问题. 但是, 利用聚类算法所得到的局部区域, 准确度十分有限. 在同样使用 Alex Net 的情况下, 其分类精度要低于强监督的 Part R-CNN 算法[43].

6.2　基于局部区域的图像表示

以上所介绍的算法都只是简单地将卷积网络的输出作为特征表示来使用. 事实上, 卷积特征的每一个位置点, 都对应于原图中的一个局部的感受野 (Receptive fields), 即卷积特征的一些局部区域对应于原图中的局部区域.

基于这种思想, Zhang 等[12] 提出了一种能够从卷积特征中挑选出具有分辨力的局部区域特征的算法, 与传统算法相比, 减少了产生局部区域所需的计算量. 首先对于输入图像, 利用 Selective search[47] 产生对象区域候选. 对于每一个候选, 利用 MMP (Multi-max pooling) 方法, 直接从候选的卷积特征中产生局部区域的特征. 之后, 对这些特征做聚类, 并计算每一个聚类簇的重要性, 选择重要的聚类簇来构造最终的图像特征表示. 其算法流程图如图 6 所示.

对于每一个候选区域, 提取其卷积特征为一个 $N \times N \times d$ 的张量, Zhang 等采用 MMP 方法从卷积特征中, 直接提取出局部区域的特征, 得到若干 d 维的特征. 该方法利用一个 $M \times M$ 大小的滑动窗口, 从卷积特征的左上角向右下角扫描, 每次扫描都对窗口内的特征做一次 Max pooling 编码, 得到一条 d 维特征. 同时, 通过变化 M 的取值, 可以得到不同大小的区域的特征表示, 这里 $M \in [1, N]$.

这样, 利用 MMP 方法就能够直接得到局部候选的特征表示, 避免了基于 Selective search 方法的巨大计算开销. 但是这些特征中, 包含着大量无关信息, 需要对其进行选择, 去除噪音.

首先, 利用 FV (Fisher vector) 编码[20] 将每一张图像的所有局部区域候选表示成一个向量. 由于 FV 编码使用了高斯混合模型 (Gaussian mixture model, GMM) 进行聚类, 因此, 每一个聚类簇可以认为是一种局部区域 (如头部、翅膀、爪子等). 于是, 接下来的任务就是从众多的聚类簇中, 选择那些重要的聚类簇. 这可以通过计算每一个类的相互信息值 (Mutual information, MI) 作为该簇的重要程度分值[57]. 通过这样的方式能够选择出那些重要的聚类簇.

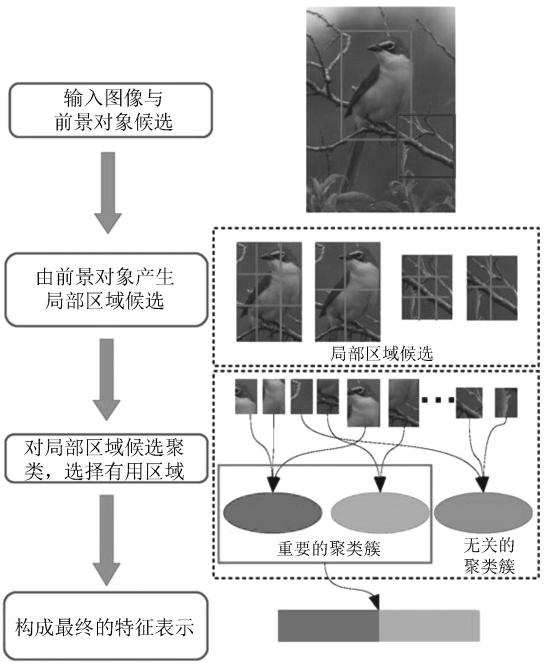


图 6　算法流程图[12]

Fig. 6　System overview[12]

最后, Zhang 等提出使用一种改进的 FV 编码方式 (ScPM 编码), 将不同的规模的局部特征编码为最终的特征表示, 用来训练 SVM 分类器, 达到了 79.34 % 的分类精度.

6.3 星座 (Constellations) 算法

正如我们在上一节曾经提到的, 基于 Selective search[47] 产生区域候选的方法, 尽管有效, 却面临巨大的计算代价和资源浪费. 因此, 有研究人员尝试采用其他方式来产生足够的局部区域.

Simon 等[58] 设计了一种新颖的局部区域检测与提取的方案, 在 CUB200-2011 数据集上达到 81.01 % 的分类精度. 他们利用卷积网络特征产生一些关键点, 并基于这些关键点来提取局部区域信息. 通过对卷积特征进行可视化分析, Simon 等发现响应比较强烈的区域往往对应于原图中一些潜在的局部区域点. 从这一角度来看, 卷积特征还可以被视为一种检测分数, 响应值高的区域代表着原图中检测到的局部区域.

但是, 特征输出的分辨率与原图相差悬殊, 很难对原图中的区域进行精确定位. 受前期研究工作[59−60] 的启发, Simon 等采用的方法是通过计算梯度图来产生区域位置.

具体而言, 卷积特征的输出是一个 $W \times H \times P$ 维的张量, P 表示通道的数量, 每一维通道可以表示成一个 $W \times H$ 维的矩阵. 通过计算每一维通道 p 对每一个输入像素的平均梯度值, 可以得到与原输入图像大小相同的特征梯度图:

$$m_{x,y}^{(p)}(I) = \frac{\partial}{\partial I_{x,y}} \sum_{j,j'} f_{j,j'}^{(p)}(I) \tag{6}$$

式 (6) 可以通过反向传播高效地完成计算[59]. 这样, 每一个通道的输入, 都可以转换成与原图同样大小的特征梯度图. 于是, 在特征梯度图里响应比较强烈的区域, 即代表原图中的一个局部区域. 通过计算每一个梯度图里响应最强烈的位置, 作为原图中的关键点:

$$\mu_{i,p} = \arg\max_{x,y} |m_{x,y}^{(p)}(I_i)| \tag{7}$$

卷积层的输出共有 P 维通道, 通过计算特征梯度图的方式能够产生 P 个关键点位置. 但这些关键点中仍然存在一些无关的背景信息, 因此, 需要对关键点进行选择. 这可以通过随机选择或者星座 (Constellations) 算法来完成.

进行特征选择之后, 关键位置点的个数就从 P 个减少到了 M 个. 得到这些关键点之后, 将其作为标注框的中心, 取大小为 $\sqrt{\lambda \cdot W' \cdot H'}$, 其中 $\lambda \in 1/5, 1/16\}$ 是一个超参数, W' 和 H' 是原图的大小. 这样就能够利用标注框来从原图中提取出局部区域, 再利用卷积网络来提取特征.

至于前景对象, Simon 等并未提出更好的解决方案, 他们采用的仍是传统的局部区域候选的方法, 即利用 Selective search[47] 产生候选区域, 再利用卷积神经网络对其进行分类, 取置信度最高的区域作为前景对象. 最后的特征向量由三部分信息构成: 原图的特征、前景对象的特征以及局部区域的特征. 在训练时, 对 VGG-Net[39] 进行了微调, 并将所有训练数据进行水平翻转, 用来进行数据增强, 最终结果为 81 %.

6.4 双线性 CNN (Bilinear CNN)

同样是回答如何在不依赖于标记信息的情况下, 完成对局部区域的检测问题, 以上介绍的两种算法均给出了让人满意的解决方案. Zhang 等通过对卷积特征进行多尺度的划分来产生局部区域, 而星座算法则是直接从卷积特征中反推原图中的关键点, 进而确定局部区域. 但这两种算法都只是把卷积网络当做一个特征提取器, 各个步骤之间的处理仍然是一个分散的过程, 并未从整体上进行端到端 (End-to-end) 的训练优化. 与此不同的是, Lin[13] 等设计了一种新颖的网络模型双线性 CNN (Bilinear CNN), 在 CUB200-2011 数据集上实现了 84.1 % 的分类精度. 其网络结构如图所示:

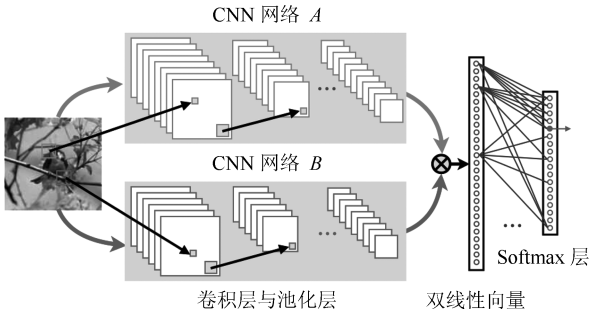


图 7 双线性 CNN 网络结构图[13]

Fig. 7 Illustration of Bilinear CNN[13]

顾名思义, 双线性 CNN 中最重要的就是双线性 (Bilinear) 模型. 一个双线性模型 $\mathcal{B}$ 由一个四元组组成: $\mathcal{B} = (f_A, f_B, \mathcal{P}, \mathcal{C})$. 其中, f_A, f_B 代表特征提取函数, 即图 7 中的网络 A、网络 B, $\mathcal{P}$ 是一个池化函数 (Pooling function), $\mathcal{C}$ 则是分类函数.

特征提取函数 $f(\cdot)$ 的作用可以看作一个函数映射, $f : \mathcal{L} \times \mathcal{I} \rightarrow R^{c \times D}$, 将输入图像 $\mathcal{I}$ 与位置区域 $\mathcal{L}$ 映射为一个 $c \times D$ 维的特征. 而两个特征提取函数的输出, 可以通过一个双线性操作进行汇聚, 得到一个双线性特征: $\text{bilinear}(l, \mathcal{I}, f_A, f_B) = f_A(l, \mathcal{I})^{\text{T}} f_B(l, \mathcal{I})$. 而池化函数 $\mathcal{P}$ 的作用则是将所有位置的双线性特征汇聚成一个特征. 文章所采用的池化函数是将所有位置的双线性特征累加起来: $\phi(\mathcal{I}) = \sum_{l \in \mathcal{L}} \text{bilinear}(l, \mathcal{I}, f_A, f_B)$. 如果两个特征函数 f_A, f_B 提取的特征维度分别是 $C \times M$ 与 $C \times N$ 的话, 则池化函数 $\mathcal{P}$ 的输出将是一个 $M \times N$ 的矩阵, 将其转化为一个 $MN \times 1$ 的列向量, 作为所提取的特征. 最后, 分类函数的作用是对提取的特征进行分类, 可以采用逻辑回归或者 SVM 分类器.

当双线性模型应用到实际的网络中时, 特征提取函数 f_A, f_B 的输出是一个 $M \times N \times P$ 维的张

量, 这时位置 $\mathcal{L}$ 定义为 $M \times N$ 维矩阵上的每一个位置点, 共有 MN 个位置. 每个位置经过双线性操作后转化为一个 $P \times P$ 维的矩阵, 经过池化函数之后, 最终得到一个 $PP \times 1$ 的特征向量.

最后, 是关于模型端到端的训练过程. 从图 7 中可以看出, 模型的前半部分是普通的卷积层与池化层, 因此, 只要求得后半部分的梯度值, 即可完成对整个模型的训练. 假设对于每个位置 l, 特征提取函数 f_A, f_B 的输出分别是 $A \in \mathbf{R}^{L\times M}$ 与 $B \in \mathbf{R}^{L\times N}$, 则池化的双线性特征是 $x = A^{\mathrm{T}}B$. 令 $\mathrm{d}l/\mathrm{d}x$ 表示损失函数对特征 x 的梯度值, 则根据链式法则, 可以得到损失函数对两个网络输出的梯度值, 从而完成模型的端到端的训练:

$$\frac{\mathrm{d}l}{\mathrm{d}A} = B\left(\frac{\mathrm{d}l}{\mathrm{d}x}\right)^{\mathrm{T}}, \quad \frac{\mathrm{d}l}{\mathrm{d}B} = A\left(\frac{\mathrm{d}l}{\mathrm{d}x}\right)^{\mathrm{T}} \tag{8}$$

一种对双线性 CNN 模型的解释是, 网络 A 的作用是对物体进行定位, 即完成传统算法的对象与局部区域检测工作, 而网络 B 则是用来对网络 A 检测到的物体位置进行特征提取. 两个网络相互协调作用, 完成细粒度图像分类过程中两个最重要的任务: 区域检测与特征提取.

6.5 其他

弱监督的分类算法, 是当前细粒度图像研究的发展趋势. 除了以上所介绍的若干算法之外, 相关的研究领域中还存在着如下重要工作:

在文献 [55] 中, Jaderberg 等提出了一种端到端的模型, 他们称之为空间转换网络 (Spatial transformer networks). 该模型只使用类别标签就能完成对象的定位与对齐, 同样实现了 84.1 % 的分类精度. 整个系统由两部分组成: 对象检测器与空间转换器. 前者用来完成前景对象的检测工作, 后者则是对检测结果进行对齐操作.

Wang 等[61] 则提出应当进行多层次的图像分类. 他们根据生物学上的分类方法, 将数据库重新划分为科、属、种等多个不同的层次. 对于每个不同层次的网络, 使用不同尺度的图像和不同的监督信息进行训练, 以达到粗细互补的目的. 最后的特征由多个不同层次网络的输出拼接而成, 实现了 81.7 % 的分类精度.

相类似的研究成果还有很多, 也都取得了不错的效果, 本文限于篇幅, 不再一一介绍.

7 未来研究方向

本文介绍了近年来基于卷积特征的细粒度图像分类算法的发展状况. 我们在表 1 总结了其中若干优秀算法在 CUB200-2011[1] 数据集上的性能表现, 给出了训练和测试阶段所使用的标注信息, 并简要地描述了算法的大致流程: 如 SIFT + BoW + SVM 指的是, 先对图像提取 SIFT 特征, 并用 BoW 对局部特征进行编码, 最后使用 SVM 进行分类.

该表主要分为 4 个部分, 第一部分是数据库发布之时的分类精度, 受限于当时的技术水准, 传统分类算法的表现不尽人意. 第二部分是基于人造特征的早期算法, 借助于特殊的特征描述与编码方式, 以及人工标注信息, 这类算法能够实现一定的突破. 第三部分是基于卷积特征的强监督的分类算法, 相比于人造特征, 卷积特征提供了更好的图像描述. 其中, Alex-Net + Fine-Tune 表示使用 Alex-Net[22] 预训练网络模型, 并在数据集上进行了微调. 最后一部分是基于卷积特征的弱监督的分类算法, 这类算法不借助任何标注信息, 仅仅依靠类别标签, 实现了更高的分类精度. 其中, *Flip* 表示在训练时对图像进行了水平翻转操作, 这是一种常用的用于数据增强的方式, 能够改善因训练数据不足而带来的过拟合问题.

细粒度图像分类的研究方兴未艾, 亟待后续研究的深入进行. 关于未来可能的研究方向, 我们认为可从以下几个方面进行考虑:

1) 构建更高质量的标准数据库: 当前主流研究所采用的细粒度图像数据库, 尽管可供选择的余地很大, 但都存在一个共同的不足之处: 数据规模与精细程度都不太高, 标注质量与类别数量也十分有限. 众所周知, 深度学习的性能与数据库的规模呈正相关性, 训练图像越丰富, 所能带来的性能提升越明显, 实用性也越强. 因此, 如何构建更高质量的标准数据库成为了未来研究急需解决的一个问题.

2) 有效地利用局部区域信息: 细粒度图像识别有别于普通图像分类任务的一大特点, 便是具有区分度的信息隐藏在局部区域中. 如何更有效地利用这些局部信息, 将成为未来研究一大突破点. 其中主要包含两个方面的问题, 一是何谓 "有用的" 局部信息, 二是如何获取这些信息. 前者主要依赖于人工经验, 由人来指定所需要提取的局部区域. 其不足之处在于, 我们很难概括所有的有用区域, 而这些区域在不同的子类上往往是不同的. 后者则寄希望于更高效的区域检测算法, 这可以从通用的物体定位检测任务中获取灵感. 但需要注意的是, 弱监督的细粒度图像分类是未来研究的主要方向, 如何在只有类别标记的前提下, 有效地完成对局部区域的定位检测工作, 这无疑是个不小的挑战.

3) 构造更强大的特征表示: 诚然, 一个更强大的特征表示离不开深度学习相关研究工作的突破. 但对于细粒度图像分类而言, 最终的特征表示往往是由多个不同的局部区域特征组合而成. 简单的特征拼接, 尽管有效, 但似乎并不是最佳选择. 另一方面, 双线性 CNN[13] 的成功也为我们提供了新思路: 进行端到端的训练, 构造一个整体的系统, 将特征提取与定位检测任务相结合, 以达到相互促进的目的.

表 1 CUB200-2011[1] 数据库上的算法性能比较 (其中 BBox 指标注框信息 (Bounding Box), Parts 指局部区域信息)
Table 1 Performance of different algorithms in CUB200-2011[1] (where BBox refers to bounding box, Parts means part annotations)

算法	BBox (训练)	Parts (训练)	BBox (测试)	Parts (测试)	简要描述	准确率 (%)
CUB[1]	✓	✓			SIFT + BoW + SVM	10.3
CUB[1]	✓	✓	✓	✓	SIFT + BoW + SVM	17.3
POOF[26]	✓	✓	✓		POOF + SVM	56.8
POOF[26]	✓	✓	✓	✓	POOF + SVM	73.3
Alignment[31]	✓		✓		Fisher + SVM	62.7
Symbiotic[30]	✓		✓		Fisher + SVM	61.0
DeCAF[25]	✓		✓		Alex-Net + Logistic Regression	61.0
Part R-CNN[43]	✓	✓			Alex-Net + Fine-Tune + SVM	73.9
Pose Normalized CNN[48]	✓	✓			Alex-Net + Fine-Tune + SVM	75.7
Pose Normalized CNN[48]	✓	✓	✓	✓	Alex-Net + Fine-Tune + SVM	85.4
Two-level Attention[56]					Alex-Net	69.7
Two-level Attention[56]					VGG16-Net	77.9
Zhang et al.[12]					VGG16-Net + Fine-Tune + SVM	79.3
Constellations[58]					VGG19-Net + Fine-Tune + Flip + SVM	81.0
Bilinear CNN[13]					VGG19-Net/VGG-M + Flip	84.1
Spatial Transformer Net[55]					Inception[62] + Flip	84.1

4) 自然场景下的图像识别: 细粒度图像分类是一门与实际应用密切相关的研究课题, 其最终目的应当是服务于实际生活. 但目前学术研究中所用的数据库, 普遍具有前景对象突出, 背景单一的特点, 这样的图片在实际生活中其实并不常见. 若想使细粒度图像识别系统在自然场景下得到广泛应用, 就不得不考虑诸如光照、模糊、遮挡、低分辨率, 物体干扰等复杂场景下的图像识别问题, 而这些因素在当前的系统中往往是欠缺的. 另外, 除了静态图片之外, 视频中的细粒度识别[63] 也是一项极具挑战的研究任务. 目前, 这方面的研究工作并不丰富, 但其在智能监控, 生态研究等领域具有更强烈的实际需求, 值得未来工作的展开.

5) 向其他领域的拓展: 事实上, 细粒度图像是一个综合性的研究课题, 不应局限于图像分类一个领域, 需要向计算机视觉的其他研究方向进行拓展, 如图像检索[64−65]、对象检测[66] 等. 在这方面, 我们看到了一些初步尝试, 如有研究人员提出细粒度图像检索的任务[67−68], 并取得了一定的效果, 但更多的研究内容仍然有待进一步挖掘.

8 总结

细粒度图像分类算法是计算机视觉领域的一个热门研究课题, 深度卷积特征的出现为其带来了新的发展机遇. 本文从强监督、弱监督两个角度, 对近年来基于卷积特征的细粒度图像分类算法的发展状况给予了介绍. 针对细粒度分类中的两个核心任务: 局部信息的检测与特征提取, 进行了详细讨论, 并总结了该领域未来可能的发展机遇.

References

1 Wah C, Branson S, Welinder P, Perona P, Belongie S. The Caltech-UCSD Birds-200-2011 Dataset, Technical Report CNS-TR-2011-001, California Institute of Technology, Pasadena, CA, USA, 2011

2 Bosch A, Zisserman A, Muñoz X. Scene classification using a hybrid generative/discriminative approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008, **30**(4): 712−727

3 Wu J X, Rehg J M. CENTRIST: a visual descriptor for scene categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, **33**(8): 1489−1501

4 Gehler P, Nowozin S. On feature combination for multiclass object classification. In: Proceedings of the 12th IEEE International Conference on Computer Vision. Kyoto, Japan: IEEE, 2009. 221−228

5 Jarrett K, Kavukcuoglu K, Ranzato M, LeCun Y. What is the best multi-stage architecture for object recognition? In: Proceedings of the 12th IEEE International Conference on Computer Vision. Kyoto, Japan: IEEE, 2009. 2146−2153

6 Wright J, Yang A Y, Ganesh A, Sastry S S, Ma Y. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, **31**(2): 210−227

7 Li Xiao-Li, Da Fei-Peng. A rapid method for 3D face recognition based on rejection algorithm. *Acta Automatica Sinica*, 2010, **36**(1): 153−158
(李晓莉, 达飞鹏. 基于排除算法的快速三维人脸识别方法. 自动化学报, 2010, **36**(1): 153−158)

8 Khosla A, Jayadevaprakash N, Yao B P, Li F F. Novel dataset for fine-grained image categorization. In: Proceedings of the 1st Workshop on Fine-Grained Visual Categorization (FGVC), IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Springs, USA: IEEE, 2011.

9 Nilsback M E, Zisserman A. Automated flower classification over a large number of classes. In: Proceedings of the 6th Indian Conference on Computer Vision, Graphics & Image Processing. Bhubaneswar, India: IEEE, 2008. 722−729

10 Krause J, Stark M, Deng J, Li F F. 3D object representations for fine-grained categorization. In: Proceedings of

the 2013 IEEE International Conference on Computer Vision Workshops (ICCVW). Sydney, Australia: IEEE, 2013. 554−561
11 Maji S, Rahtu E, Kannala J, Blaschko M, Vedaldi A. Fine-grained visual classification of aircraft [Online], available: https://arxiv.org/abs/1306.5151, June 21, 2013
12 Zhang Y, Wei X S, Wu J X, Cai J F, Lu J B, Nguyen V A, Do M N. Weakly supervised fine-grained categorization with part-based image representation. *IEEE Transactions on Image Processing*, 2016, **25**(4): 1713−1725
13 Lin T Y, RoyChowdhury A, Maji S. Bilinear CNN models for fine-grained visual recognition. In: Proceedings of the 15th IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 1449−1457
14 Zhang Lin-Bo, Wang Chun-Heng, Xiao Bai-Hua, Shao Yun-Xue. Image representation using bag-of-phrases. *Acta Automatica Sinica*, 2012, **38**(1): 46−54
(张琳波, 王春恒, 肖柏华, 邵允学. 基于 Bag-of-phrases 的图像表示方法. 自动化学报, 2012, **38**(1): 46−54)
15 Yu Wang-Sheng, Tian Xiao-Hua, Hou Zhi-Qiang. A new image feature descriptor based on region edge statistical. *Chinese Journal of Computers*, 2014, **37**(6): 1398−1410
(余旺盛, 田孝华, 侯志强. 基于区域边缘统计的图像特征描述新方法. 计算机学报, 2014, **37**(6): 1398−1410)
16 Yan Xue-Jun, Zhao Chun-Xia, Yuan Xia. 2DPCA-SIFT: an efficient local feature descriptor. *Acta Automatica Sinica*, 2014, **40**(4): 675−682
(颜雪军, 赵春霞, 袁夏. 2DPCA-SIFT: 一种有效的局部特征描述方法. 自动化学报, 2014, **40**(4): 675−682)
17 Lowe D G. Object recognition from local scale-invariant features. In: Proceedings of the 7th IEEE International Conference on Computer Vision. Kerkyra, Greece: IEEE, 1999. 1150−1157
18 Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA: IEEE, 2005. 886−893
19 Jégou H, Douze M, Schmid C, Pérez P. Aggregating local descriptors into a compact image representation. In: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, USA: IEEE, 2010. 3304−3311
20 Perronnin F, Dance C. Fisher kernels on visual vocabularies for image categorization. In: Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA: IEEE, 2007. 1−8
21 Sánchez J, Perronnin F, Mensink T, Verbeek J. Image classification with the Fisher vector: theory and practice. *International Journal of Computer Vision*, 2013, **105**(3): 222−245
22 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada, USA: MIT Press, 2012. 1097−1105
23 Gao Ying-Ying, Zhu Wei-Bin. Deep neural networks with visible intermediate layers. *Acta Automatica Sinica*, 2015, **41**(9): 1627−1637
(高莹莹, 朱维彬. 深层神经网络中间层可见化建模. 自动化学报, 2015, **41**(9): 1627−1637)
24 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*, 2015, **521**(7553): 436−444
25 Donahue J, Jia Y Q, Vinyals O, Hoffman J, Zhang N, Tzeng E, Darrell T. DeCAF: a deep convolutional activation feature for generic visual recognition. In: Proceedings of the 31st International Conference on Machine Learning. Beijing, China: ACM, 2014. 647−655
26 Berg T, Belhumeur P N. POOF: part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Portland, USA: IEEE, 2013. 955−962
27 Perronnin F, Sánchez J, Mensink T. Improving the fisher kernel for large-scale image classification. In: Proceedings of the 11th European Conference on Computer Vision. Berlin Heidelberg, Germany: Springer, 2010. 143−156
28 Bo L, Ren X, Fox D. Kernel descriptors for visual recognition. In: Proceedings of the 24th Annual Conference on Neural Information Processing Systems. Vancouver, Canada: MIT Press, 2010. 244−252
29 Branson S, Van Horn G, Wah C, Perona P, Belongie S. The ignorant led by the blind: a hybrid human-machine vision system for fine-grained categorization. *International Journal of Computer Vision*, 2014, **108**(1−2): 3−29
30 Chai Y N, Lempitsky V, Zisserman A. Symbiotic segmentation and part localization for fine-grained categorization. In: Proceedings of the 14th IEEE International Conference on Computer Vision (ICCV). Sydney, Australia: IEEE, 2013. 321−328
31 Gavves E, Fernando B, Snoek C G M, Smeulders A W M, Tuytelaars T. Fine-grained categorization by alignments. In: Proceedings of the 14th IEEE International Conference on Computer Vision (ICCV). Sydney, Australia: IEEE, 2013. 1713−1720
32 Yao B P, Bradski G, Li F F. A codebook-free and annotation-free approach for fine-grained image categorization. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, USA: IEEE, 2012. 3466−3473
33 Yang S L, Bo L F, Wang J, Shapiro L. Unsupervised template learning for fine-grained object recognition. In: Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, USA: MIT Press, 2012. 3122−3130
34 Branson S, Wah C, Schroff F, Babenko B, Welinder P, Perona P, Belongie S. Visual recognition with humans in the loop. In: Proceedings of the 11th European Conference on Computer Vision. Berlin Heidelberg, Germany: Springer, 2010. 438−451
35 Wah C, Branson S, Perona P, Belongie S. Multiclass recognition and part localization with humans in the loop. In: Proceedings of the 13th IEEE International Conference on Computer Vision (ICCV). Barcelona, Spain: IEEE, 2011. 2524−2531
36 LeCun Y, Boser B, Denker J S, Henderson D, Howard R E, Hubbard W, Jackel L D. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1989, **1**(4): 541−551
37 LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998, **86**(11): 2278−2324
38 Zeiler M D, Fergus R. Visualizing and understanding convolutional networks. In: Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014. 818−833
39 Gong Y C, Wang L W, Guo R Q, Lazebnik S. Multi-scale orderless pooling of deep convolutional activation features. In: Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014. 392−407
40 Cimpoi M, Maji S, Vedaldi A. Deep filter banks for texture recognition and segmentation. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015. 3828−3836
41 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [Online], available: https://arxiv.org/abs/1409.1556, April 10, 2015
42 Welinder P, Branson S, Mita T, Wah C, Schroff F, Belongie S, Perona P. Caltech-UCSD Birds 200, Technical Report CNS-TR-2010-001, California Institute of Technology, Pasadena, CA, USA, 2010
43 Zhang N, Donahue J, Girshick R, Darrell T. Part-based R-CNNs for fine-grained category detection. In: Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014. 834−849

44 Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, USA: IEEE, 2014. 580−587

45 Viola P, Jones M J. Robust real-time face detection. *International Journal of Computer Vision*, 2004, **57**(2): 137−154

46 Wu J X, Liu N N, Geyer C, Rehg M J. C^4: a real-time object detection framework. *IEEE Transactions on Image Processing*, 2013, **22**(10): 4096−4107

47 Uijlings J R R, van de Sande K E A, Gevers T, Smeulders A W M. Selective search for object recognition. *International Journal of Computer Vision*, 2013, **104**(2): 154−171

48 Branson S, Van Horn G, Belongie S, Perona P. Bird species categorization using pose normalized deep convolutional nets [Online], available: https://arxiv.org/abs/1406.2952, June 11, 2014

49 Branson S, Beijbom O, Belongie S. Efficient large-scale structured learning. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Portland, USA: IEEE, 2013. 1806−1813

50 Krause J, Jin H L, Yang J C, Li F F. Fine-grained recognition without part annotations. In: Proceedings of the 15th IEEE International Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015. 5546−5555

51 Guillaumin M, Küttel D, Ferrari V. Imagenet auto-annotation with segmentation propagation. *International Journal of Computer Vision*, 2014, **110**(3): 328−348

52 Kuettel D, Guillaumin M, Ferrari V. Segmentation propagation in imagenet. In: Proceedings of the 12th European Conference on Computer Vision. Berlin Heidelberg, Germany: Springer, 2012. 459−473

53 Lin D, Shen X Y, Lu C W, Jia J Y. Deep LAC: deep localization, alignment and classification for fine-grained recognition. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015. 1666−1674

54 Xu Z, Huang S L, Zhang Y, Tao D C. Augmenting strong supervision using web data for fine-grained categorization. In: Proceedings of the 15th IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 2524−2532

55 Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K. Spatial transformer networks. In: Proceedings of the 29th Annual Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 2015. 2017−2025

56 Xiao T J, Xu Y C, Yang K Y, Zhang J X, Peng Y X, Zhang Z. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015. 842−850

57 Zhang Y, Wu J X, Cai J F. Compact representation for image classification: to choose or to compress. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, USA: IEEE, 2014. 907−914

58 Simon M, Rodner E. Neural activation constellations: unsupervised part model discovery with convolutional networks. In: Proceedings of the 15th IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 1143−1151

59 Simon M, Rodner E, Denzler J. Part detector discovery in deep convolutional neural networks. In: Proceedings of the 12th Asian Conference on Computer Vision. Singapore: Springer, 2014. 162−177

60 Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps [Online], available: https://arxiv.org/abs/1312.6034, April 19, 2014

61 Wang D Q, Shen Z Q, Shao J, Zhang W, Xue X Y, Zhang Z. Multiple granularity descriptors for fine-grained categorization. In: Proceedings of the 15th IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 2399−2406

62 Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015. 1−9

63 Hall D, Perona P. Fine-grained classification of pedestrians in video: benchmark and state of the art. In: Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015. 5482−5491

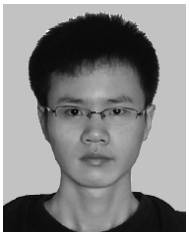
64 Liu Y, Zhang D S, Lu G J, Ma W Y. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 2007, **40**(1): 262−282

65 Datta R, Joshi D, Li J, Wang J Z. Image retrieval: ideas, influences, and trends of the new age. *ACM Computing Surveys*, 2008, **40**(2): Article No. 5

66 Felzenszwalb P F, Girshick R B, McAllester D, Ramanan D. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, **32**(9): 1627−1645

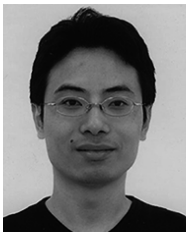
67 Wei X S, Luo J H, Wu J X. Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE Transactions on Image Processing*, 2017, **26**(6): 2868−2881

68 Xie L X, Wang J D, Zhang B, Tian Q. Fine-grained image search. *IEEE Transactions on Multimedia*, 2015, **17**(5): 636−647



罗建豪 南京大学计算机科学与技术系博士研究生. 2015 年获得吉林大学计算机科学与技术学院学士学位. 主要研究方向为计算机视觉与机器学习.
E-mail: luojh@lamda.nju.edu.cn

(**LUO Jian-Hao** Ph. D. candidate in the Department of Computer Science and Technology, Nanjing University. He received his bachelor degree from the College of Computer Science and Technology, Jilin University in 2015. His research interest covers computer vision and machine learning.)



吴建鑫 南京大学计算机科学与技术系教授. 分别于 1999 年, 2002 年获得南京大学计算机科学与技术系学士, 硕士学位. 于 2009 年获得美国佐治亚理工学院博士学位. 曾担任新加坡南洋理工大学计算机工程学院助理教授. 主要研究方向为计算机视觉与机器学习. 本文通信作者. E-mail: wujx2001@nju.edu.cn

(**WU Jian-Xin** Professor in the Department of Computer Science and Technology, Nanjing University. He received his bachelor and master degrees from Nanjing University in 1999 and 2002, respectively. In 2009, he received his Ph. D. degree in computer science from the Georgia Institute of Technology, USA. He was an assistant professor at the Nanyang Technological University, Singapore. His research interest covers computer vision and machine learning. Corresponding author of this paper.)