



JOURNAL OF IMAGE AND GRAPHICS

主办: 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2021

04

VOL.26

ISSN1006-8961
CN11-3758/TB



细粒度图像分类 P847



第26卷第4期 (总第300期)

2021年4月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

版权声明

凡向《中国图象图形学报》投稿, 均视为同意在本刊网站及CNKI等全文数据库出版, 所刊载论文已获得著作权人的授权。本刊所有图片均为非商业目的使用, 所有内容, 未经许可, 不得转载或以其他方式使用。

Copyright

All rights reserved by Journal of Image and Graphics, Institute of Remote Sensing and Digital Earth, CAS. The content (including but not limited text, photo, etc) published in this journal is for non-commercial use.

主管单位 中国科学院

主办单位 中国科学院空天信息创新研究院

中国图象图形学学会

北京应用物理与计算数学研究所

主 编 吴一戎

编辑出版 《中国图象图形学报》编辑部

通信地址 北京市海淀区北四环西路19号

邮 编 100190

电子信箱 jig@radi.ac.cn

电 话 010-58887035

网 址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号

总 发 行 北京报刊发行局

订 购 全国各地邮局

海外发行 中国国际图书贸易集团有限公司

(邮政信箱: 北京399信箱 邮编: 100048)

印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian

Monthly, Started in 1996

Superintended by Chinese Academy of Sciences

Sponsored by Aerospace Information Research Institute, CAS

China Society of Image and Graphics

Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong

Editor, Publisher Editorial Office of Journal of
Image and Graphics

Address No. 19, North 4th Ring Road West, Haidian District,
Beijing, P. R. China

Zip code 100190

E-mail jig@radi.ac.cn

Telephone 010-58887035

Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals

Domestic All Local Post Offices in China

Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048, P.R.China)

Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB

ISSN 1006-8961

CODEN ZTTXFZ

国外发行代号 M1406

国内邮发代号 82-831

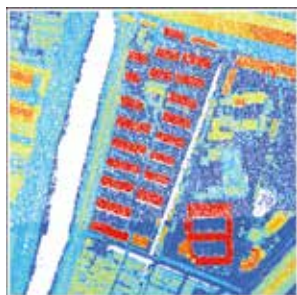
国内定价 60.00元



YOLOv3剪枝模型的多人姿态估计(第0837页)



融入时序和速度信息的自适应更新目标跟踪(第0883页)



可变半径Alpha Shapes提取机载LiDAR点云建筑物轮廓(第0910页)

综述

从图像到语言: 图像标题生成与描述

谭云兰, 汤鹏杰, 张丽, 罗玉盘 0727

CT影像肺结节分割研究进展

董婷, 魏珑, 聂生东 0751

图像处理和编码

全局注意力门控残差记忆网络的图像超分重建

王静, 宋慧慧, 张开华, 刘青山 0766

多层次感知残差卷积网络的单幅图像超分重建

何蕾, 程佳豪, 占志钰, 杨雯博, 刘沛然 0776

采用双尺度图像分解的水下彩色图像增强

李健, 张显斗, 李熹飞, 吴子朝 0787

图像分析和识别

姿态特征结合2维傅里叶变换的步态识别

王新年, 胡丹丹, 张涛, 白桂欣 0796

复杂背景下的手势识别

王银, 陈云龙, 孙前来 0815

结合形变模型与图像修复的人脸姿态矫正

吴从中, 郑荣生, 臧怀娟, 刘明威, 徐甲甲, 詹曙 0828

YOLOv3剪枝模型的多人姿态估计

蔡哲栋, 应娜, 郭春生, 郭锐, 杨鹏 0837

YOLOv3和双线性特征融合的细粒度图像分类

闫子旭, 侯志强, 熊磊, 刘晓义, 余旺盛, 马素刚 0847

深度多模态融合服装风格检索

苏卓, 柯司博, 王若梅, 周凡 0857

超像素条件随机场下的RGB-D视频显著性检测

李贝, 杨铀, 刘琼 0872

融入时序和速度信息的自适应更新目标跟踪

尹宽, 李均利, 胡凯, 李丽 0883

融合逆密度函数与关系形状卷积神经网络的点云分析

邱云飞, 朱梦影 0898

图像理解和计算机视觉

可变半径Alpha Shapes提取机载LiDAR点云建筑物轮廓

伍阳, 王丽妍, 胡春霞, 程亮 0910

3D遮挡模型引导的光场图像深度获取

吴迪, 张旭东, 张骏, 范之国, 孙锐 0924

计算机图形学

基于全正基的非均匀三次加权B样条

姬佩佩, 张贵仓, 汪凯, 孟建军 0939

遥感图像处理

LBP特征分类的极化SAR图像机场跑道检测

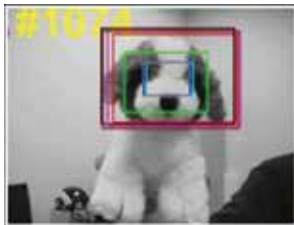
韩萍, 万义爽, 刘亚芳, 韩宾宾 0952

CONTENTS

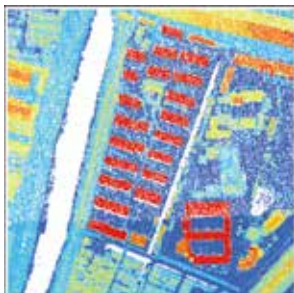
JOURNAL OF IMAGE AND GRAPHICS



Research on multiperson pose estimation combined with YOLOv3 pruning model(P0837)



Adaptive update object tracking algorithm incorporating timing and speed information(P0883)



Extraction of building contours from airborne LiDAR point cloud using variable radius Alpha Shapes method(P0910)

Review

From image to language: image captioning and description

Tan Yunlan, Tang Pengjie, Zhang Li, Luo Yupan 0727

Research progress of lung nodule segmentation based on CT images

Dong Ting, Wei Long, Nie Shengdong 0751

Image Processing and Coding

Learning global attention-gated multi-scale memory residual networks for single-image super-resolution

Wang Jing, Song Huihui, Zhang Kaihua, Liu Qingshan 0766

Single image super-resolution reconstruction based on multi-level perceptual residual convolutional network

He Lei, Cheng Jiahao, Zhan Zhiyu, Yang Wenbo, Liu Peiran 0776

Underwater color image enhancement based on two-scale image decomposition

Li Jian, Zhang Xiandou, Li Shangfei, Wu Zizhao 0787

Image Analysis and Recognition

Gait recognition using pose features and 2D Fourier transform

Wang Xinnian, Hu Dandan, Zhang Tao, Bai Guixin 0796

Hand gesture recognition in complex background

Wang Yin, Chen Yunlong, Sun Qianlai 0815

Face pose correction based on morphable model and image inpainting

Wu Congzhong, Zheng Rongsheng, Zang Huaijuan, Liu Mingwei, Xu Jiajia, Zhan Shu 0828

Research on multiperson pose estimation combined with YOLOv3 pruning model

Cai Zhedong, Ying Na, Guo Chunsheng, Guo Rui, Yang Peng 0837

Fine-grained classification based on bilinear feature fusion and YOLOv3

Yan Zixu, Hou Zhiqiang, Xiong Lei, Liu Xiaoyi, Yu Wangsheng, Ma Sugang 0847

Fashion style retrieval based on deep multimodal fusion

Su Zhuo, Ke Sibao, Wang Ruomei, Zhou Fan 0857

RGB-D video saliency detection via superpixel-level conditional random field

Li Bei, Yang You, Liu Qiong 0872

Adaptive update object tracking algorithm incorporating timing and speed information

Yin Kuan, Li Junli, Hu Kai, Li Li 0883

Point cloud analysis model combining inverse density function and relation-shape convolution neural network

Qiu Yunfei, Zhu Mengying 0898

Image Understanding and Computer Vision

Extraction of building contours from airborne LiDAR point cloud using variable radius Alpha Shapes method

Wu Yang, Wang Liyan, Hu Chunxia, Cheng Liang 0910

Light field depth estimation guided by 3D occlusion model

Wu Di, Zhang Xudong, Zhang Jun, Fan Zhiguo, Sun Rui 0924

Computer Graphics

Non uniform weighted cubic B-spline based on all positive basis

Ji Peipei, Zhang Guicang, Wang Kai, Meng Jianjun 0939

Remote Sensing Image Processing

Airport runway detection based on LBP feature classification in PolSAR images

Han Ping, Wan Yishuang, Liu Yafang, Han Binbin 0952

中图法分类号: TP391 文献标识码: A 文章编号: 1006-8961(2021)04-0847-10

论文引用格式: Yan Z X, Hou Z Q, Xiong L, Liu X Y, Yu W S and Ma S G. 2021. Fine-grained classification based on bilinear feature fusion and YOLOv3. Journal of Image and Graphics, 26(04): 0847-0856 [闫子旭, 侯志强, 熊磊, 刘晓义, 余旺盛, 马素刚. 2021. YOLOv3 和双线性特征融合的细粒度图像分类. 中国图象图形学报, 26(04): 0847-0856] [DOI: 10.11834/jig.200031]

YOLOv3 和双线性特征融合的细粒度图像分类

闫子旭¹, 侯志强^{1*}, 熊磊², 刘晓义¹, 余旺盛³, 马素刚¹

1. 西安邮电大学计算机学院, 西安 710121; 2. 西安交通大学电信学院, 西安 710049;
3. 空军工程大学信息与导航学院, 西安 710077

摘要: 目的 细粒度图像分类是计算机视觉领域具有挑战性的课题, 目的是将一个大的类别分为更详细的子类别, 在工业和学术方面都有着十分广泛的研究需求。为了改善细粒度图像分类过程中不相关背景干扰和类别差异特征难以提取的问题, 提出了一种将目标检测方法 YOLOv3 (you only look once) 和双线性融合网络相结合的细粒度分类优化算法, 以此提高细粒度图像分类的性能。**方法** 利用重新训练过的目标检测算法 YOLOv3 粗略确定目标在图像中的位置; 使用背景抑制方法消除目标以外的信息干扰; 利用融合不同通道、不同层级卷积层特征的方法对经典的细粒度分类算法双线性卷积神经网络 (bilinear convolutional neural network, B-CNN) 进行改进, 优化分类性能, 通过融合双线性网络中不同卷积层的特征向量, 得到更加丰富的互补信息, 从而提高细粒度分类精度。**结果** 实验结果表明, 在 CUB-200-2011 (Caltech-UCSD Birds-200-2011)、Cars196 和 Aircrafts100 数据集中, 本文算法的分类准确率分别为 86.3%、92.8% 和 89.0%, 比经典的 B-CNN 细粒度分类算法分别提高了 2.2%、1.5% 和 4.9%, 验证了本文算法的有效性。同时, 与已有细粒度图像分类算法相比也表现出一定的优势。**结论** 改进算法使用 YOLOv3 有效滤除了大量无关背景, 通过特征融合方法来改进双线性卷积神经网络, 丰富特征信息, 使分类的结果更加精准。**关键词:** 细粒度图像分类; 目标检测; 背景抑制; 特征融合; 双线性卷积神经网络 (B-CNN)

Fine-grained classification based on bilinear feature fusion and YOLOv3

Yan Zixu¹, Hou Zhiqiang^{1*}, Xiong Lei², Liu Xiaoyi¹, Yu Wangsheng³, Ma Sugang¹

1. College of Computer, Xi'an University of Posts and Telecommunications, Xi'an 710121, China;
2. College of Telecommunications, Xi'an Jiaotong University, Xi'an 710049, China;
3. College of Information and Navigation, Air Force Engineering University, Xi'an 710077, China

Abstract: **Objective** Image classification is a classic topic in the field of computer vision. It can be divided into coarse-grained classification and fine-grained classification. The purpose of coarse-grained classification is to identify objects of different categories, whereas that of fine-grained image classification is to subdivide larger categories into more fine-grained categories, which in many cases have greater use value. Fine-grained image classification is a challenging research topic in computer vision. There are extensive research needs and application scenarios of fine-grained image classification in the industry and academia. Due to background interference and difficulty in extracting effective classification features, problems still exist in fine-grained classification. Compared with general image classification, fine-grained classification experiences

收稿日期: 2020-02-11; 修回日期: 2020-10-08; 预印本日期: 2020-10-15

* 通信作者: 侯志强 hzq@xupt.edu.cn

基金项目: 国家自然科学基金项目 (61703423, 61473309, 61379104)

Supported by: National Natural Science Foundation of China (61703423, 61473309, 61379104)

background interference. This problem can be addressed by object detection methods. The task of object detection is to find the objects of interest in the image and determine their position and size. At present, more and more target detection methods are based on deep learning. These methods can be divided into two categories: one-stage detection method and two-stage detection method. One-stage detection method has fast detection speed, but its accuracy is slightly lower. Examples of one-stage detection method mainly include you only look once (YOLO) and single shot multibox detector (SSD). Two-stage detection method first uses region recommendation to generate candidate targets, and then it uses a convolutional neural network (CNN) to process this condition. Some of the examples of this method include R-CNN (region CNN), SPP-NET (spatial pyramid pooling convolutional network), and Faster R-CNN. Among them, YOLOv3 of the YOLO series has achieved a better balance in detection accuracy and speed compared with other commonly used target detection frameworks.

Method To improve the accuracy of these detection methods, a fine-grained classification algorithm based on the fusion of YOLOv3 and bilinear features is proposed in this study. The algorithm first uses the retrained target detection algorithm YOLOv3 to coarsely locate the target. Then, a background suppression method is used to remove irrelevant background interference. Finally, the feature fusion method is used to bilinear convolutional neural networks in the classic fine-grained classification algorithm. It can find that the convolutional neural network (referred to as B-CNN (bilinear CNN)) is greatly improved. By merging the features of different convolutional layers, more abundant complementary information is obtained. We use this method to improve the accuracy. The specific operation steps are as follows: 1) enter the image; 2) use YOLOv3 pre-trained model to generate discriminative regions; 3) the background suppression method removes irrelevant background interference outside the discrimination box; 4) construct a bilinear fine classification network of feature fusion, and use deep convolutional neural networks to extract features at the multi-layer convolution stage on the image; 5) the outer product operation is used to fuse the features of the convolution layers at different stages, and the obtained fusion features of the three different levels of features are connected by the concat method to obtain the final bilinear vector. Finally, the Softmax layer is used to achieve fine-grained classification. **Result** After adding the YOLOv3 algorithm with background suppression, the classification accuracy rates on the three datasets are 0.7%, 0.5%, and 3.1% higher than those of B-CNN, respectively, indicating that removing background interference using the YOLOv3 algorithm can effectively improve classification. After using feature fusion to optimize the B-CNN network structure, we use three datasets (namely, CUB-200-2011 (Caltech-UCSD Birds-200-2011), Stanford Cars, and fine-grained visual categorization (FGVC) Aircraft) to test the performance. The results are 1.4% and 1.2% higher than B-CNN, which indicates that the fusion of the features of different convolutional layers and the strengthening of the spatial relationship of the features can effectively improve the classification accuracy rate. After using YOLOv3 for background suppression and fusion of B-CNN, the accuracy rates reach 86.3%, 92.8%, and 89.0% in the three datasets, respectively. Compared with B-CNN algorithm, the proposed algorithm improves the accuracy by 2.2%, 1.5%, and 4.9% in the three datasets, respectively, indicating its effectiveness. For the purpose of analyzing the classification performance of the algorithm, the improved algorithm classification results also have certain advantages compared with the mainstream algorithms. **Conclusion** The fine-grained classification algorithm based on YOLOv3 and bilinear feature fusion proposed in this study not only uses YOLOv3 to effectively filter out several irrelevant backgrounds to obtain discriminative regions on the image, but also improves the bilinear fine-grainedness by means of feature fusion. Classification network, so as to extract richer fine-grained features, and make the results of fine-grained image classification more accurate. This study proposes a fine-grained classification algorithm based on YOLOv3 and bilinear feature fusion, which can remove interference from irrelevant backgrounds. At the same time, the improved feature fusion B-CNN can learn richer features, which improves to a certain extent the accuracy of fine-grained classification. Compared with the classic B-CNN algorithm, the three fine-grained datasets are better than some mainstream algorithms. On the other hand, some new fine-grained classification algorithms are constantly changing. They use a host of different deep learning models to perform fine classification in fine-grained classification, but do not use background suppression and feature fusion to extract richer fine-grained features. In the future, we will apply fusion to the new network and use different types of fusion to further improve the accuracy of fine-grained classification in this study.

Key words: fine-grained image classification; target detection; background suppression; feature fusion; bilinear convolutional neural network (B-CNN)

0 引言

图像分类是计算机视觉领域一个经典课题(黄凯奇等,2014),可以分为粗粒度图像分类和细粒度图像分类。粗粒度图像分类的主要任务是对不同类别的物体进行识别判断,例如区分车和人等;细粒度图像分类是对目标进行更加详细的种类划分,如区别汽车的车型、鸟的品种等。在很多情况下,细粒度图像分类具有更大的使用价值。

深度学习的优良性能使得基于深度学习的细粒度图像分类算法成为趋势。根据网络模型训练时是否需要人工标注信息,细粒度图像分类算法大致分为强监督细粒度分类算法和弱监督细粒度分类算法两类(罗建豪和吴建鑫,2017)。强监督细粒度分类算法不但需要图像的类别标注,而且需要人工标注框和局部区域位置等信息。在强监督细粒度分类算法中,Zhang 等人(2014)提出 Part-Based R-CNN 算法对图像进行细粒度分类,但会产生大量不相关的候选区域,造成计算资源浪费;Branson 等人(2014)提出基于姿态归一化的卷积神经网络算法(pose normalized convolutional neural networks (CNN)),使用 DPM(deformable parts model)产生候选框与图像的中心点区域进行匹配并进行信息提取。然而,DPM 获取的信息与标注信息之间存在一定差距,导致分类结果并不理想。弱监督细粒度图像分类在进行模型训练时只依赖类别标签,大幅减少了人工标注的昂贵代价,变得越来越受推崇。在弱监督学习方法中,Xiao 等人(2015)提出了一种两级注意力算法,使用聚类计算近似得到目标的局部区域,但分类性能一般,得到的识别率有限;Zhang 等人(2016)提出利用选择性搜索方法得到目标的区域候选框,虽然分类准确度有一定提升,但计算成本巨大。Lin 等人(2018)提出双线性卷积神经网络(bilinear convolutional neural network, B-CNN),在 CUB-200-2011(Caltech-UCSD Birds-200-2011)、Cars196 和 Aircrafts100 等标准的细粒度分类数据集上都有很好的表现。双线性分类模型可以视为用一路网络检测和定位目标,而用另一路网络对目标进行更精细的特征提取,这两路卷积神经网络相互协作配合,提取更多图像的特征信息,完成细粒度图像的识别。与一般的细粒

度分类算法相比,B-CNN 特有的双线性模型能更充分地提取特征,识别效果很好,但在提取更多细致特征信息上还存在一些局限。为了提取更丰富有效的图像特征,进一步提高双线性网络的表征力,本文利用特征融合的思想对细粒度算法 B-CNN 进行改进。

与一般的图像分类相比,细粒度分类还存在背景干扰等问题,对此可以运用目标检测的相关算法解决。目标检测的任务是寻找出判别区域,定位目标在图像中的位置并进行识别(罗海波等,2017)。当前,基于深度学习的检测方法越来越多,分为单阶段检测方法和两阶段检测方法两大类型。单阶段检测方法直接对图像进行特征提取,检测速度快,但检测精度略低。代表算法有 YOLO(you only look once)(Redmon 等,2016)和 SSD(single shot multibox detector)(Liu 等,2016)等。两阶段检测方法首先生成目标的预选框,然后利用卷积神经网络对预选框的信息进行提取,用于检测和回归,检测精度高但速度慢。代表算法包括 SPP-NET(spatial pyramid pooling convolutional network)和 Faster R-CNN(region CNN)(Ren 等,2015)等。其中,YOLO 系列的 YOLOv3(Redmon 和 Farhadi,2018)与其他常用目标检测算法框架相比,可以兼顾精度和速度,检测目标效果很有优势。

本文提出一种基于 YOLOv3 和特征融合的双线性细粒度分类算法。具体工作如下:1)使用经过重新训练的 YOLOv3 算法粗检测出物体,获取目标区域,然后通过背景抑制滤除图像中与细粒度图像分类无关的区域,进一步提升算法性能;2)通过融合不同通道不同卷积层的特征改进 B-CNN 的双线性网络,加强双路网络间的空间联系,丰富层与层之间的交互,充分表达特征信息,提高识别率;3)在 CUB-200-2011、Cars196 和 Aircrafts100 等标准的细粒度数据集上进行实验,分类准确率分别为 86.3%、92.8% 和 89.0%,与原始的 B-CNN 模型相比,分别提高了 2.2%、1.5% 和 4.9%。

1 相关工作

本文算法属于基于双线性卷积神经网络(B-CNN)和目标检测算法 YOLOv3 的细粒度分类算法。

1.1 双线性卷积神经网络 (B-CNN)

B-CNN 的整体架构如图 1 所示。可以看出, B-CNN 模型 F 是一个四元组, 具体为

$$F = (f_A, f_B, P, C) \quad (1)$$

式中, f_A 和 f_B 表示图 1 中两个双线性卷积神经网络 A 和 B 的特征提取函数, P 是池化函数, C 是一个经由 Softmax 归一化层后对双线性特征向量进行分类和识别的分类函数。

该模型的特征提取 f 是一个函数映射过程, $f: L \times I \rightarrow R^{c \times D}$, 该过程将输入图像 L 和位置区域 I 通过映射关系转换成一个尺寸大小为 $c \times D$ 维的向量, 再将经过特征提取过程得到的特征函数 f_A 和 f_B 的输出进行外积运算, 从而获得对应位置上的双线性特征向量。

池化函数 P 将得到的双线性特征进行整合, 目的是得到用于细粒度分类的特征函数。其中, 池化过程中采用将图像上每个相应位置的双线性特征累加的方式进行, 计算为

$$\varphi(I) = \sum_{l \in L} Bi(I, l, f_A, f_B) \quad (2)$$

式中, l 表示区域中面积池化核大小的子区域, Bi 表示双线性操作, 并且

$$Bi(I, l, f_A, f_B) = f_A(I, l) \cdot f_B(I, l) \quad (3)$$

通过式 (3) 可以得到每个位置上的特征。

尽管 B-CNN 具有一定的特征表征力, 但存在对图像中目标位置获取不敏感、输出的卷积特征信息不充分等问题, 不利于提高细粒度分类精度。

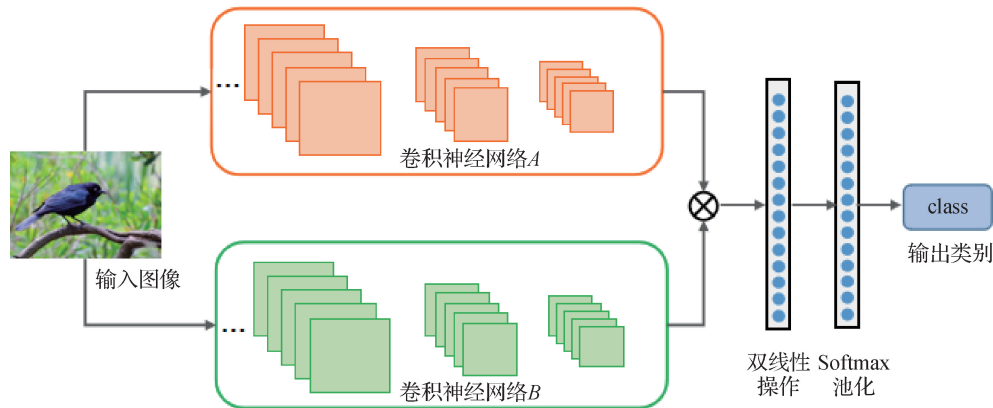


图 1 双线性卷积神经网络 B-CNN 模型结构

Fig. 1 B-CNN model architecture of bilinear convolutional neural network

1.2 YOLOv3 检测方法

YOLOv3 目标检测算法是 YOLO 系列算法的第 3 个版本, 使用的网络模型为 DarkNet-53, 结构如图 2 所示。DarkNet-53 网络共 75 层, 使用一系列 3×3 、 1×1 的卷积, 包括 53 层卷积层, 其余为残差层, 采用跳层连接的方法组成残差模块, 在很好地平衡速度和精度的情况下完成对图像目标的检测。

2 本文算法

利用目标检测方法 YOLOv3 检测定位目标在图像上的大致位置, 通过背景抑制方法剔除目标区域以外的背景, 防止无关信息干扰; 然后将去除了背景干扰的图像输入到添加了特征融合的双线性卷积神经网络 B-CNN 中, 以获得最终的分类结果。

	类型	卷积核数量	卷积核尺寸	输出
1×	卷积层	32	3×3	256×256
	卷积层	64	3×3/2	128×128
	卷积层	32	1×1	
	卷积层	64	3×3	128×128
2×	残差层			128×128
	卷积层	128	3×3/2	64×64
	卷积层	64	1×1	
	卷积层	128	3×3	64×64
4×	残差层			64×64
	卷积层	256	3×3/2	32×32
	卷积层	128	1×1	
	卷积层	256	3×3	32×32
8×	残差层			32×32
	卷积层	512	3×3/2	16×16
	卷积层	256	1×1	
	卷积层	512	3×3	16×16
8×	残差层			16×16
	卷积层	1 024	3×3/2	8×8
	卷积层	512	1×1	
	卷积层	1 024	3×3	8×8
4×	残差层			8×8

图 2 DarkNet-53 网络结构

Fig. 2 DarkNet-53 network structure

2.1 基于 YOLOv3 网络的聚焦判别性区域

尽管 YOLOv3 具有优秀的目标检测性能,但仅能进行物体的粗检测,要用于细粒度分类,还需要进一步训练。

将图像送入 YOLOv3 网络进行目标的判别性区域提取。首先利用 ImageNet 数据集的预训练模型对 YOLOv3 的 DarkNet-53 网络进行初始化参数调试,然后利用本文的细粒度数据集进行微调(fine-tuning),训练整个 YOLOv3 模型,将输入图像缩放至统一像素大小,训练过程中学习率设置为 0.000 1,迭代次数设置为 10 000 次。YOLOv3 将输入图像划为 $S \times S$ 个大小均等的网格,当目标落入某一网格时,该网格就担任相应的检测任务。同时,计算 3 个不相同尺度的预测框(anchorbox),用来检测不同大小的目标。每个预测框包含的对应信息值为 $5 + C$, C 代表所用数据集中的目标类别总数,5 代表 5 个属性信息值,即图像的目标中心点位置坐标(x, y)、预测框的宽高度大小(w, h)以及置信度(confidence)。网络预测的分类置信度为

$$Pr(class_i | object) \times Pr(object) \times IoU_{pred}^{truth} \quad (4)$$

式中, $Pr(class_i | object)$ 表示预测第 i 类目标落到该网格的可能性, $Pr(object) = 1$ 表示有目标的中心点位置在网格内,若没在网格内则 $Pr(object) = 0$ 。 IoU_{pred}^{truth} 表示预测目标的边界框与实际位置框的交并比(intersection over union, IoU),用来判定两个框的相近程度。利用非极大值抑制(nonmaximum suppression, NMS)方法选择最有可能作为最终检测框的预测框。

背景抑制是指对检测框之外的背景进行裁剪剔除,具体步骤为:1)读取有检测框的图像;2)搜索图像的轮廓;3)仅保留目标检测出的最大判别检测框;4)向上下左右扩展 90 个像素,滤除大量背景,只保留目标。利用背景抑制方法可以去除背景干扰,更利于细粒度特征的提取,提取流程如图 3 所示。



图 3 判别区域的提取流程图

Fig. 3 Extraction flow of discriminative region

2.2 特征融合的双线性卷积神经网络 B-CNN

在深度卷积神经网络中,不同卷积层提取的特

征各不相同,表达的语义信息也不相同。例如浅层特征注重描述图像的边缘信息,而深层特征更注重描述图像的细节信息。在细粒度图像分类时,除了处理最后一层卷积层特征信息,还要考虑到更加丰富的细粒度图像特征,利用不同特征之间的空间联系性来进行特征增强。受这种特征融合思想(赵浩如等,2019;秦兴和宋各方,2019)的启发,本文通过融合特征来优化网络结构,使网络间不同卷积层特征的优势互补,从而提高分类的精度。

2.2.1 改进思路

首先,将双线性网络 A 和 B 的卷积层 conv4 和 conv5 中的每一层都经过 add 操作连接到一个 conv 层,即分别提取网络 A 和网络 B 中 conv4_1、conv4_2、conv4_3 和 conv5_1、conv5_2、conv5_3 的特征向量进行 add 操作(图 4 中 $\oplus$)。add 操作是将同维度的特征像素点进行相加,以此获得更多信息的特征,对图像分类任务大有裨益。

其次,增加了两个新的双线性层,用于对 add 提取的特征向量进行外积运算,得到双线性向量。如图 4 所示,通过对网络 A 中 conv4 和 conv5 的融合特征进行外积操作(图 4 中 $\otimes$)得到双线性特征 B_2 ,通过对网络 B 中 conv4 和 conv5 的融合特征进行外积操作得到双线性特征 B_3 。外积操作就是将卷积层输出的特征向量进行一个转置运算,达到与不同空间维度特征

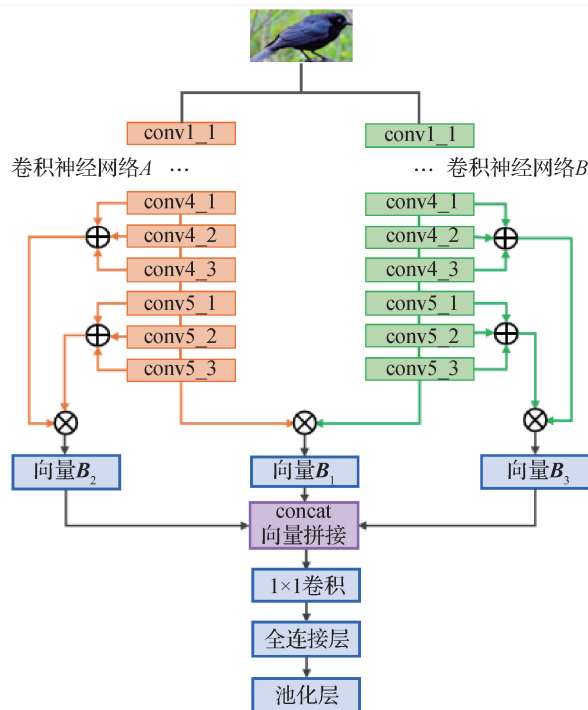


图 4 改进的 B-CNN 算法流程图

Fig. 4 The flowchart of improved B-CNN algorithm

交互的目的。因为在不同通道网络中,不同卷积层表达的信息均不相同,所以通过增加不同层的外积操作,可以得到图像的不同特征,使双线性网络模型的表征能力得到提高。

然后,添加一个 concat 拼接层。concat 为张量拼接,是直接连接不同通道的特征,将其融合转换成同一特征信息,可以加强不同通道之间的特征联系,实现特征信息互补。本文利用 concat 将得到的 3 组不同双线性向量进行连接融合,即将上一步融合卷积层得到的特征 B_2 、 B_3 与原始双线性网络中的特征 B_1 拼接(concat)在一起得到特征 B 。

最后,将拼接得到的特征向量 B 通过 1×1 卷积核进行降维表示,并在池化层得到最终的细粒度分类结果。

2.2.2 网络训练

本文对利用特征融合方法改进后的 B-CNN 进行训练,具体步骤如下:

1) 使用 YOLOv3 预训练模型从常用的 3 个细粒度分类标准数据集 CUB-200-2011、Cars196 和 Aircrafts100 中分别提取目标的大体位置区域,经过背景抑制操作后,输出只包含目标的图像,并将大小归

一化为 448×448 像素。

2) 对改进的 B-CNN 模型的相关参数进行微调,包括将分类类别数目改为 3 个细粒度数据集对应的类别数目和利用随机赋值的方法对最后一层的参数进行初始化并训练。

3) 设置较小的学习率为 0.001,因为在实验中发现设置过大的学习率会使模型不收敛,导致振荡。

4) 采用随机梯度下降(stochastic gradient descent,SGD)和反向传播(back propagation,BP)算法实现对整个网络模型的调整训练,最大迭代次数设置为 100,batch size 设置为 16。

2.3 本文算法框架

本文算法流程如图 5 所示,具体步骤如下:

1) 输入图像;

2) 利用 YOLOv3 预训练模型,生成具有判别性的目标区域;

3) 通过背景抑制,剔除判别框之外的无关背景干扰;

4) 添加 add 操作和 3 个双线性层,构建基于特征融合的双线性细粒度分类网络,从图像中提取更多丰富的细粒度特征;

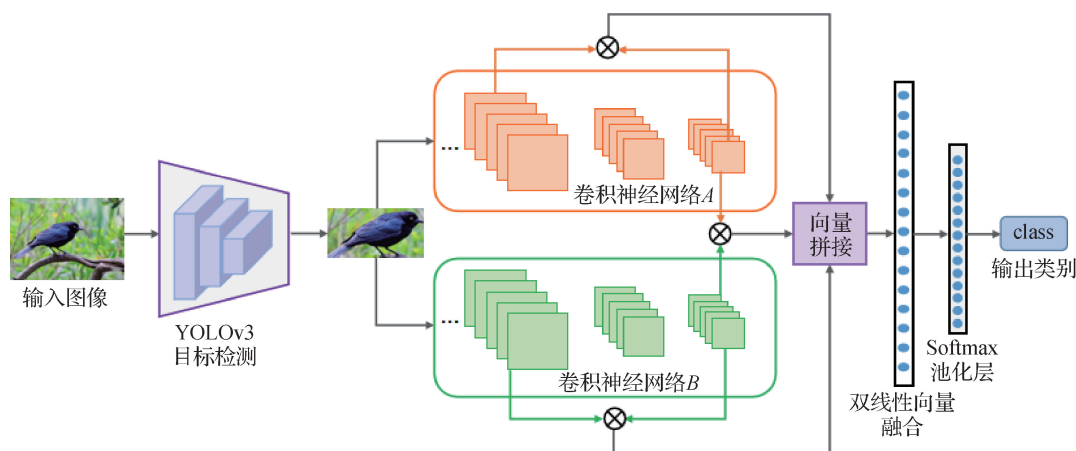


图 5 本文算法框架

Fig. 5 The algorithm framework of this paper

5) 利用外积操作融合 3 个不同层次的双线性层特征,通过拼接(concat)的方法联接在一起,得到最终的融合特征向量并进行细粒度图像分类结果的输出。

3 实验比较与分析

本文算法在处理器为 Intel(R) Corei5

8400CPU、GPU 为 GTX1080Ti 的计算机上运行,使用的开源深度学习框架为 PyTorch。

3.1 数据集介绍

为测试改进算法的细粒度分类性能,采用常用的 3 个细粒度图像分类数据集 CUB-200-2011(Wah 等,2011)、Cars196(Krause 等,2013)和 Aircrafts100(Maji 等,2013)进行实验。

CUB-200-2011(简称 Birds200)数据集包括 200 种不同种类、姿态和环境下的鸟类图像 11 788 幅,其中训练集 5 994 幅,测试集 5 794 幅,呈基本持平状态,每幅图像包含 1 个标注框、1 个类别标签和 15 个局部区域位置。Cars196 数据集包括 196 类不同型号和品牌的车辆图像 16 185 幅,其中训练集

8 144 幅,测试集 8 041 幅,比例接近 1 : 1,每幅图像提供标注信息和类别标签。Aircrafts100 数据集包括 10 000 幅飞机图像,共有 100 种飞机模型,其中训练图像 6 667 幅,测试图像 3 333 幅,每幅图像提供标注信息和类别标签。3 个数据集的部分图像如图 6 所示。



图6 本文数据集部分示例

Fig.6 Examples of the datasets

((a) CUB-200-2011 dataset; (b) Cars196 dataset; (c) Aircrafts100 dataset)

3.2 算法性能分析

采用分类精度(accuracy)作为评估结果的指标进行算法性能分析。分类精度是最常用的图像分类评估指标之一,其定义为分类正确的图像数量占总数据集图像数量的比重,即

$$Acc = \frac{I_c}{I} \quad (5)$$

式中, I_c 代表细粒度分类正确的图像数量, I 代表图像总数。

为了验证提出融合方案的性能,在常用的细粒度标准数据集 CUB-200-2011 上,与秦兴和宋各方(2019)和赵浩如等人(2019)提出的两种融合方案进行对比。

秦兴和宋各方(2019)运用的融合方法是将网络 A 的 conv4_1 和 conv5_1 的特征向量提取出来分别与网络 B 的 conv5_3 的特征向量进行外积操作,再将得到的特征向量与原 B-CNN 的双线性特征向量进行 concat 融合,得到最终特征向量进行细粒度分类。

赵浩如等人(2019)运用的融合方法是将网络

A 的 conv4_3 和 conv5_3 的特征向量进行外积操作,网络 B 的 conv5_1 和 conv5_3 的特征向量进行外积操作,再将得到的两组双线性特征向量与原 B-CNN 的双线性向量进行 concat 操作,得到细粒度分类结果。

实验对比结果见表 1。从表 1 可以看出,本文融合方案得到的分类准确度比其他两个融合方案和原 B-CNN 分别提高了 0.7%、0.5% 和 1.4%,表明融合更多卷积层特征有助于提高细粒度图像分类结果。

表1 不同融合方案结果对比

Table 1 Comparison of results of different fusion schemes

融合方案	精度 /%
B-CNN	84.1
赵浩如等人(2019)	84.8
秦兴和宋各方(2019)	85.0
本文	85.5

注:加粗字体表示最优结果。

将YOLOv3背景抑制与B-CNN的多层特征融合改进结合起来,在3个数据集上进行识别率对比实验,结果如图7所示。其中,B-CNN表示不使用YOLOv3去除背景干扰,不改变B-CNN网络结构;YOLOv3+B-CNN表示使用YOLOv3去除背景干扰,但不改变B-CNN网络结构;Fusion B-CNN表示不使用YOLOv3去除背景干扰,但利用特征融合优化B-CNN网络结构;YOLOv3+Fusion B-CNN表

示使用YOLOv3去除背景干扰,且利用特征融合优化B-CNN网络结构。由图7可以看出,随着训练步数的增加,分类准确率持续增加后趋于平稳,并且YOLOv3+Fusion B-CNN曲线一直处于最高状态,B-CNN曲线处于最低状态。表明无论是去除背景抑制还是只融合B-CNN对分类效果都有一定程度的提升,两者结合改进呈现出最好的分类效果。

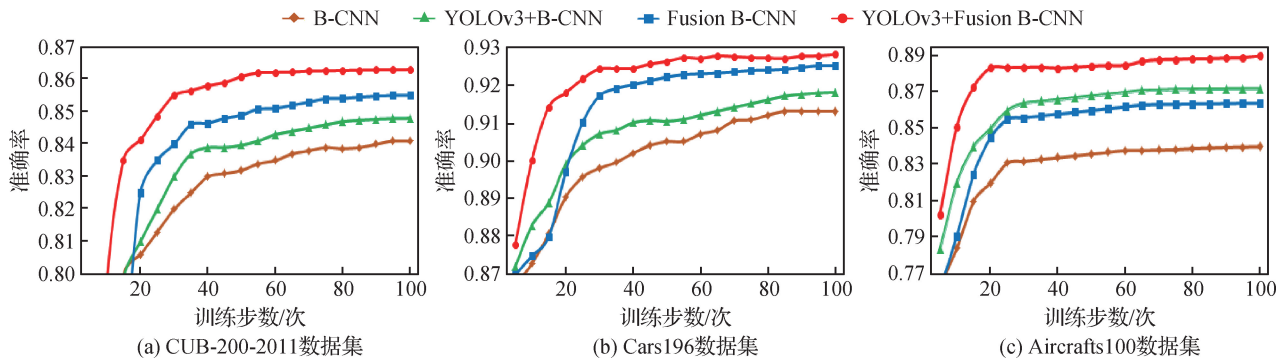


图7 不同模型在3个数据集上的识别准确率比较

Fig. 7 Comparison of recognition accuracies of different models on three datasets
((a) CUB-200-2011 dataset; (b) Cars196 dataset; (c) Aircrafts100 dataset)

表2展示了不同算法的消融实验结果对比。可以看出,在3个数据集上,加入背景抑制的YOLOv3+B-CNN算法的分类准确率比B-CNN分别提高了0.7%、0.5%和3.1%,表明利用YOLOv3算法去除背景干扰后可以有效提高细粒度图像分类的精度。使用特征融合改进的Fusion B-CNN网络结构的细粒度分类准确率有明显提升,比B-CNN分别提高了1.4%、1.2%和2.3%,表明融合不同通道和不同卷积层的特征加强了特征空间之间的联系,得到了更多有助于分类的信息,有效提高了分类准确率。使用YOLOv3进行背景抑制且融合B-CNN

的YOLOv3+Fusion B-CNN的分类准确率分别为86.3%、92.8%和89.0%,与B-CNN算法相比,分别提高了2.2%、1.5%和4.9%,提升效果显著,证明了该算法的有效性。

为了更进一步分析本文算法的分类性能,将本文算法与近年来常用的two-level attention(Xiao等,2015)、DVAN(diversified visual attention networks)(Zhao等,2017)、RA-CNN(recurrent attention-CNN)(Fu等,2017)、CSGLML(cascaded softmax and generalized large-margin losses)(Shi等,2019)、YOLOv2+B-CNN(王永雄和张晓兵,2019)、RPN(region proposal network)+B-CNN(赵浩如等,2019)和空间关系利用算法(exploiting spatial relation)(Qi等,2019)等弱监督算法模型在3个细粒度数据集上分别进行对比,分类结果如表3所示。

从表3可以看出,本文算法在3个细粒度数据集上的精度均较高。在Birds200数据集上,本文算法的细粒度分类准确率为86.3%,比two-level attention、DVAN、B-CNN、RA-CNN、CSGLML、YOLOv2+B-CNN、RPN+B-CNN和exploiting spatial relation算法分别提高了8.4%、7.3%、2.2%、1.0%、0.9%、0.8%、0.8%和0.8%。在Cars196数据集

表2 不同算法在3个数据集上的分类准确率对比

Table 2 Comparison of accuracies among different algorithms on three datasets

算法	/%		
	CUB-200-2011	Cars196	Aircrafts100
B-CNN	84.1	91.3	84.1
YOLOv3+B-CNN	84.8	91.8	87.2
Fusion B-CNN	85.5	92.5	86.4
YOLOv3+Fusion B-CNN	86.3	92.8	89.0

注:加粗字体表示各列最优结果。

表3 本文算法在3个数据集上与
各种弱监督算法的分类准确率对比Table 3 Comparison of classification accuracies among
various weakly supervised algorithms
and ours on three datasets

算法	CUB-200-2011	Cars196	Aircrafts100
two-level attention (Xiao 等,2015)	77.9	-	-
DVAN(Zhao 等,2017)	79.0	87.1	-
B-CNN(Lin 等,2018)	84.1	91.3	84.1
RA-CNN(Fu 等,2017)	85.3	92.5	-
CSGLML(Shi 等,2019)	85.4	91.8	85.1
YOLOv2 + B-CNN (王永雄和张晓兵, 2019)	84.5	92.0	88.4
RPN + B-CNN(赵浩如 等,2019)	85.5	-	-
exploiting spatial relation (Qi 等,2019)	85.5	-	86.9
本文	86.3	92.8	89.0

注:加粗字体表示各列最优结果,“-”表示对应文献未进行实验。

上,本文算法的分类准确率为 92.8%,比 DVAN、B-CNN、RA-CNN、CSGLML 和 YOLOv2 + B-CNN 算法分别提高了 5.7%、1.5%、0.3%、1.0% 和 0.8%。在 Aircrafts100 数据集上,本文算法的分类准确率为 89.0%,比 B-CNN、CSGLML、YOLOv2 + B-CNN 和 exploiting spatial relation 算法分别提高了 4.9%、3.9%、0.6% 和 2.1%。从表 3 可以看出,本文算法相较近几年的一些细粒度分类算法,如 exploiting spatial relation(Qi 等,2019)也具有一定的优势。

4 结 论

本文提出了一种将目标检测方法 YOLOv3 与双线性特征融合相结合的细粒度图像分类算法,能够很好地去除无关背景的干扰,同时改进后的特征融合 B-CNN 网络能够学习到更丰富的特征。实验结果表明,在 3 个标准的细粒度数据集 CUB-200-

2011、Cars196 和 Aircrafts100 上,本文算法的分类精度得到了显著提高,比原 B-CNN 算法分别提高了 2.2%、1.5% 和 4.9%。

另外,最近出现的一些细粒度分类算法(Zheng 等,2019;Ge 等,2019)使用了更深的网络模型进行细粒度分类,得到了更好的结果,但是并未使用背景抑制和特征融合来提取更丰富的细粒度特征。在下一步工作中,将考虑将融合的思想用于新的网络,尝试更多不同的融合方式,进一步优化细粒度分类算法的性能。

参考文献 (References)

- Branson S, Van Horn G, Belongie S and Perona P. 2014. Bird species categorization using pose normalized deep convolutional nets [EB/OL]. [2020-02-11]. <http://arxiv.org/pdf/1406.2952.pdf>
- Fu J L, Zheng H L and Mei T. 2017. Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 4476-4484 [DOI: 10.1109/CVPR.2017.476]
- Ge W F, Lin X R and Yu Y Z. 2019. Weakly supervised complementary parts models for fine-grained image classification from the bottom up//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 3029-3038 [DOI: 10.1109/CVPR.2019.00315]
- Huang K Q, Ren W Q and Tan T N. 2014. A review on image object classification and detection. Chinese Journal of Computers, 37(6): 1225-1240 (黄凯奇,任伟强,谭铁牛. 2014. 图像物体分类与检测算法综述. 计算机学报, 37(6): 1225-1240) [DOI: 10.3724/SP.J.1016.2014.01225]
- Krause J, Stark M, Deng J and Li F F. 2013. 3d object representations for fine-grained categorization//Proceedings of 2013 IEEE International Conference on Computer Vision Workshops. Sydney, Australia: IEEE: 554-561 [DOI: 10.1109/ICCVW.2013.77]
- Lin T Y, RoyChowdhury A and Maji S. 2018. Bilinear convolutional neural networks for fine-grained visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(6): 1309-1322 [DOI: 10.1109/TPAMI.2017.2723400]
- Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y and Berg A C. 2016. SSD: single shot MultiBox detector//Proceedings of European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 21-37 [DOI: 10.1007/978-3-319-46448-0_2]
- Luo H B, Xu L Y, Hui B and Chang Z. 2017. Status and prospect of target tracking based on deep learning. Infrared and Laser Engineering, 46(5): #0502002 (罗海波,许凌云,惠斌,常铮. 2017. 基于深度学习的目标跟踪方法研究现状与展望. 红外与激光工

- 程, 46(5): #0502002 [DOI: 10.3788/IRLA201746.0502002]
- Luo J H and Wu J X. 2017. A survey on fine-grained image categorization using deep convolutional features. *Acta Automatica Sinica*, 43(8): 1306-1318 (罗建豪, 吴建鑫. 2017. 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*, 43(8): 1306-1318) [DOI: 10.16383/j. aas. 2017. c160425]
- Maji S, Rahtu E, Kannala J, Blaschko M and Vedaldi A. 2013. Fine-grained visual classification of aircraft [EB/OL]. [2020-02-11]. <https://arxiv.org/pdf/1306.5151.pdf>
- Qi L, Lu X Q and Li X L. 2019. Exploiting spatial relation for fine-grained image classification. *Pattern Recognition*, 91: 47-55 [DOI: 10.1016/j. patcog. 2019. 02. 007]
- Qin X and Song G F. 2019. Pig face recognition algorithm based on bilinear convolution neural network. *Journal of Hangzhou Dianzi University (Natural Sciences)*, 39(2): 12-17 (秦兴, 宋各方. 2019. 基于双线性卷积神经网络的猪脸识别算法. *杭州电子科技大学学报(自然科学版)*, 39(2): 12-17) [DOI: 10.13954/j. cnki. hdu. 2019. 02. 003]
- Redmon J, Divvala S, Girshick R and Farhadi A. 2016. You only look once: unified, real-time object detection//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, USA; IEEE: 779-788 [DOI: 10.1109/CVPR.2016.91]
- Redmon J and Farhadi A. 2018. YOLOv3: an incremental improvement [EB/OL]. [2020-02-11]. <https://arxiv.org/pdf/1804.02767.pdf>
- Ren S Q, He K M, Girshick R and Sun J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks//*Proceedings of the 28th International Conference on Neural Information Processing Systems*. Cambridge, USA; MIT Press: 91-99
- Shi W W, Gong Y H, Tao X Y, Cheng D and Zheng N N. 2019. Fine-grained image classification using modified DCNNs trained by cascaded softmax and generalized large-margin losses. *IEEE Transactions on Neural Networks and Learning Systems*, 30(3): 683-694 [DOI: 10.1109/TNNLS.2018.2852721]
- Wah C, Branson S, Welinder P, Perona P and Belongie S. 2011. The Caltech-UCSD birds-200-2011 dataset [EB/OL]. [2020-02-11]. http://www.vision.caltech.edu/visipedia/papers/CUB_200_2011.pdf
- Wang Y X and Zhang X B. 2019. Fine-grained image classification with network architecture of focus and recognition. *Journal of Image and Graphics*, 24(4): 493-502 (王永雄, 张晓兵. 2019. 聚焦—识别网络架构的细粒度图像分类. *中国图象图形学报*, 24(4): 493-502) [DOI: 10.11834/jig.180423]
- Xiao T J, Xu Y C, Yang K Y, Zhang J X, Peng Y X and Zhang Z. 2015. The application of two-level attention models in deep convolutional neural network for fine-grained image classification//*Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition*. Boston, USA; IEEE: 842-850 [DOI: 10.1109/CVPR.2015.7298685]
- Zhang N, Donahue J, Girshick R and Darrell T. 2014. Part-based R-CNNs for fine-grained category detection//*Proceedings of European Conference on Computer Vision*. Zurich, Switzerland; Springer: 834-849 [DOI: 10.1007/978-3-319-10590-1_54]
- Zhang Y, Wei X S, Wu J X, Cai J F, Lu J B, Nguyen V A and Do M N. 2016. Weakly supervised fine-grained categorization with part-based image representation. *IEEE Transactions on Image Processing*, 25(4): 1713-1725 [DOI: 10.1109/TIP.2016.2531289]
- Zhao B, Wu X, Feng J S, Peng Q and Yan S C. 2017. Diversified visual attention networks for fine-grained object classification. *IEEE Transactions on Multimedia*, 19(6): 1245-1256 [DOI: 10.1109/TMM.2017.2648498]
- Zhao H R, Zhang Y and Liu G Z. 2019. Fine-grained image classification algorithm based on RPN and B-CNN. *Computer Applications and Software*, 36(3): 210-213, 264 (赵浩如, 张永, 刘国柱. 2019. 基于RPN与B-CNN的细粒度图像分类算法研究. *计算机应用与软件*, 36(3): 210-213, 264) [DOI: 10.3969/j. issn. 1000-386x. 2019. 03. 038]
- Zheng H L, Fu J L, Zha Z J and Luo J B. 2019. Looking for the devil in the details: learning Trilinear attention sampling network for fine-grained image recognition//*Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, USA; IEEE: 5007-5016 [DOI: 10.1109/CVPR.2019.00515]

作者简介



闫子旭, 1995年生, 女, 硕士研究生, 主要研究方向为图像处理和细粒度图像分类。

E-mail: 15667020597@163.com



侯志强, 通信作者, 男, 教授, 博士生导师, 主要研究方向为图像处理、计算机视觉和信息融合。

E-mail: hzq@xupt.edu.cn

熊磊, 男, 副教授, 主要研究方向为智能系统和模式识别。

E-mail: xionglei12@sina.com

刘晓义, 女, 硕士研究生, 主要研究方向为目标检测。

E-mail: 844807356@qq.com

余旺盛, 男, 硕士, 讲师, 主要研究方向为图像处理、计算机视觉与模式识别。E-mail: xing_fu_yu@sina.com

马素刚, 男, 博士研究生, 主要研究方向为计算机视觉、机器学习。E-mail: msg@xupt.edu.cn