中图法分类号：TP391　　**文献标识码**：A　　**文章编号**：1006-8961(2018)　-　-
论文引用格式：

聚焦-识别网络架构的细粒度图像分类

王永雄，张晓兵

上海理工大学光电信息与计算机工程学院, 上海　200093

摘　要: 目的 细粒度图像分类是指对一个大类别进行更细致的子类划分，如区分鸟的种类、车的品牌款式、狗的品种等。针对细粒度图像分类中的无关信息太多和背景干扰问题，本文利用深度卷积网络构建了细粒度图像聚焦-识别的联合学习框架，通过去除背景，突出待识别目标，自动定位有区分度的区域，从而提高细粒度图像分类识别率。**方法** 首先基于Yolov2网络快速检测出目标物体，消除背景干扰和无关信息对分类结果的影响，实现聚焦判别性区域，之后将检测到的物体即Yolov2的输出输入双线性卷积神经网络进行训练和分类。此网络框架可以实现端到端的训练，且只依赖于类别标注信息，而无需借助其他的人工标注信息。**结果** 在细粒度图像库CUB-200-2011、Cars196和Aircrafts100上进行实验验证，我们的模型分别达到了84.5%、92%、88.4%的分类精度。我们的方法和同样分类算法得到的最高分类精度相比，准确度分别提升了0.4%，0.7%，3.9%，比使用两个相同D-Net网络的方法分别高出0.5%、1.4%、4.5%。**结论** 使用聚焦-识别深度学习框架提取有区分度的区域对细粒度图像分类有积极作用，能够滤除大部分对细粒度图像分类没有贡献的区域，使得网络能够学习到更多有利于细粒度图像分类的特征，从而降低背景干扰对分类结果的影响，提高模型的识别率。

关键字：细粒度图像分类，Yolov2(YOLO9000:Better, Faster, Stronger)，双线性卷积神经网络，聚焦-识别框架，区分度

Fine-grained image classification with network architecture of focus and recognition

Wang Yong xiong, Zhang Xiao bing

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai, 200093

Abstract: Objective Fine-grained image classification is a hot research topic in the field of computer vision research in recent years. Its purpose is to make a more detailed subdivision of a large category, such as the distinction of bird species, car brand style, dog breed and so on. Fine-grained classification has often smaller difference between classes and larger difference within classes. Thus, compared with the ordinary image classification, fine-grained image classification is more challenging. And there are too many irrelevant information and background interference in fine-grained image classification, which can make the network model difficult to learn the real difference characteristics and finally influence the classification performance in fine-grained image classification. Therefore, finding discriminative regions in the image is very important for fine-grained image classification. In order to solve this problem, a joint deep learning framework of focus and recognition is constructed for fine-grained image classification. This framework can remove the background in the image and highlight the target to be identified, and then automatically locate the discriminative area. Thus, convolutional neural networks can extract more useful and discriminative features and the classification rate of fine-grained images can be improved naturally. **Method** First, the algorithm Yolov2 can detect object in the image quickly and eliminate the

收稿日期:　　；　**修回日期**:
基金项目：国家自然科学基金面上项目（61673276，61603255）
第一作者简介：王永雄（1970-），男，副教授，2013 年于上海交通大学大学获控制理论与控制工程专业博士学位，主要研究方向为智能机器人及视觉，E-mail: wyxiong@usst.edu.cn
Supported by: National Natural Science Foundation of China(61673276, 61603255)

influence of background interference and unrelated information, and then the datasets which include the detected objects is used to train the bilinear convolutional neural network. Finally, the final model can be used for fine-grained image classification. The algorithm Yolov2 is a further improvement of the Yolov1 target detection algorithm, and it is more precise for small objects localization. It can automatically find the target in the picture, so as to filter out most of the regions in the picture that do not contribute to the image classification. Bilinear convolutional neural network is a special network for fine-grained image classification. Its characteristic is that it uses the two convolutional neural networks to extract the features of the same picture at the same time, and the bilinear feature vector is obtained by the ways of bilinear pooling. Finally, the classification is completed by the softmax network layer. And bilinear convolutional neural network is not dependent on additional manual annotation information and can finish end to end training. And it only relies on the class label information. So it greatly reduces the difficulty and complexity of fine-grained image classification. **Result** We do the verification experiments on open standard fine-grained image library CUB-200-2011, Cars196 and Aircrafts100. We use the trained target detection model of Yolov2 algorithm to detect three datasets respectively. And then the bilinear convolutional neural network is trained by the processed datasets. Finally, our bilinear convolutional neural network model achieves classification accuracy of 84.5%, 92% and 88.4% respectively. Compared with the highest classification accuracy obtained by the same classification algorithm that there are not the step of discriminant information extraction, the classification accuracy of the 3 databases is improved by 0.4%, 0.7%, and 3.9% respectively. And the recognition rate is also increased by 0.5%, 1.4%, and 4.5% respectively, compared with the same classification algorithm which extracts features from the two identical D-Net networks. **Conclusion** The experiments show that our method has a positive effect for the fine-grained image classification which uses the network architecture of focus and recognition to detect discriminative region. It can filter out most of the area in the image that does not contribute to the classification of fine-grained images, and thus reduce the influence of background interference to the classification results. So the bilinear convolutional neural network can learn more features which are beneficial to the classification of fine-grained images, Finally, the recognition rate of the model can be improved.

Key words: Fine grained image classification; Bilinear convolutional neural network;Yolov1(You Only Look Once:Unified, Real-Time Object Detection);Yolov2 (YOLO9000: better, faster, stronger); Framework of focus and recognition; discrimination;

0 引 言

图像分类是计算机视觉研究领域的一个经典课题。图像分类主要包括粗粒度图像分类和细粒度图像分类。在很多情况下细粒度图像分类更有利用价值，它是对一个大类别进行更精细的子类划分，如区分鸟的种类、车的品牌款式、狗的品种等。因为图像采集中存在姿态、视角、光照、遮挡、背景干扰等差异，所以细粒度分类往往具有细微的类间差异和较大的类内差异。和普通的图像分类相比，细粒度图像分类具有更大的挑战性。

早期基于人工特征的细粒度图像分类算法，一般先从图像中提取 SIFT[1]、 HOG[2]等局部特征，后利用 VLAD[3]或者 Fisher vector[4-5]等编码模型进行特征编码。由于人工特征选择过程繁琐，表述能力有限，因此分类效果不佳。然而，随着深度学习的兴起，从卷积神经网络中自动获得的特征，比人工特征有更强大的描述能力，因此大量基于卷积特征算法的提出，促进了细粒度图像分类算法的快速发展。

按照模型训练时是否需要人工标注信息，基于深度学习的细粒度图像分类算法可分为强监督和弱监督两类，强监督的细粒度图像分类在模型训练时不仅需要图像的类别标签，还需要图像标注框，局部区域位置等人工标注信息，而弱监督的细粒度图像分类在模型训练时仅依赖于类别标签。然而不论是强监督或弱监督的细粒度图像分类算法，大多数细粒度图像分类算法的思路都是先找到前景对象和图像中的局部区域，之后利用卷积神经网络对这些区域分别提取特征，并将提取的特征连接，以此完成分类器的训练和预测。Zhang 等人[6]提出了 Part-based R-CNN 算法，该算法先采用 R-CNN 算法[7]对图像进行检测，得到局部区域，再分别对每一

块区域提取卷积特征，并将这些区域的特征连接，构成一维特征表示，最后用 SVM 训练分类。然而，其利用的选择性搜索算法[8]会产生大量无关的候选区域，造成运算上的浪费。Branson 等人[9]提出了姿态归一化 CNN 算法，它通过原型对图像进行姿态对齐操作，对不同的局部区域提取不同网络层的特征，但该算法利用 DPM[10]算法对关键点进行检测与实际标注的关键点信息差距较大。Xiao 等人[11]提出两级注意力算法，其不依赖额外的标注信息，仅使用类别标签，该模型分为三个处理阶段，分别是预处理、对象级和局部级三个不同的子模型，但是，两级注意力模型利用聚类算法得到局部区域，准确度十分有限。Zhang 等人[12]提出了从候选的卷积特征中选出具有区分度局部区域特征的算法，基于选择性搜索算法产生区域候选框的方法，虽然有效，却面临巨大的计算代价和资源浪费。Simon 等人[13]利用卷积神经网络产生关键点，基于这些关键点得到局部区域，最后通过卷积神经网络对局部区域提取特征。对于前景对象，依然采用传统的选择性搜索算法。然而，以上算法都只是利用卷积神经网络提取特征，各处理步骤之间是一个分散的过程，且未从整体上进行端到端的训练优化。

细粒度图像分类的难点在于各子类之间差异较小，这些差异容易被复杂的背景信息覆盖，使得网络模型难以学习到真正的差异性特征。因此，找到具有区分度的区域，即判别性区域，对细粒度图像分类至关重要。Lin 等人[14]提出了新颖的双线性卷积神经网络 (Bilinear Convolutional Neural Networks，B-CNN)的弱监督细粒度图像分类算法，在三个经典数据集上达到很高的分类精度，能够实现端到端的训练，且仅依赖类别标签，而无需借助其他的图像标注信息，这提高了算法的实用性。双线性网络模型可认为一个网络对物体局部区域进行检测，另一个网络进行特征提取，两个网络相互协调完成细粒度图像分类过程中的区域检测与特征提取，最终完成细粒度图像的分类任务。B-CNN 模型有强大的泛化能力，和以往的细粒度图像分类算法相比，其计算复杂度较低，而且识别效果很好。

同一张图片中有区分度的信息越多或者占比越大，则卷积神经网络就能提取到更多有区分度的特征，分类精度也会更高。这和人类识别细粒度物体聚焦的过程类似。基于此思路，本文提出了基于深度学习的聚焦-识别网络框架实现细粒度图像分类。首先使用 Redmon 等人[15]提出的 Yolov2 算法快速找到物体，聚焦判别性区域，滤除图像中对细粒度图像分类无关的区域，再使用 B-CNN 模型对判别性区域进行特征提取与分类，从而降低背景干扰对分类结果的影响，提高细粒度图像识别率。此方法只需图像的类别标签，减少了大量繁琐的人工标注。

1 基于聚焦-识别的深度学习框架

1.1 聚焦-识别总体框架

首先，利用 Yolov2 检测网络找到图片中的判别性区域，剔除与分类无关的背景信息，之后将得到的结果（Yolov2 的输出）输入双线性卷积神经网络得到最后的分类结果。系统框图如图 1 所示。

1.2 基于 Yolov2 的检测方法

1.2.1 基于 Yolov2 网络聚焦判别性区域

Yolov2 算法是对 Yolov1 目标检测算法[19]的进一步改进。Yolov2 算法首先把输入图像划分成 S×S 的栅格，这样更加细致的栅格划分使得模型对小物体的定位更加精准。经过 Yolov2 检测网络，对每个格子都预测 B 个边界框。由于模型在训练过程中会不断地学习调整预测的边界框的宽高维度，但是，

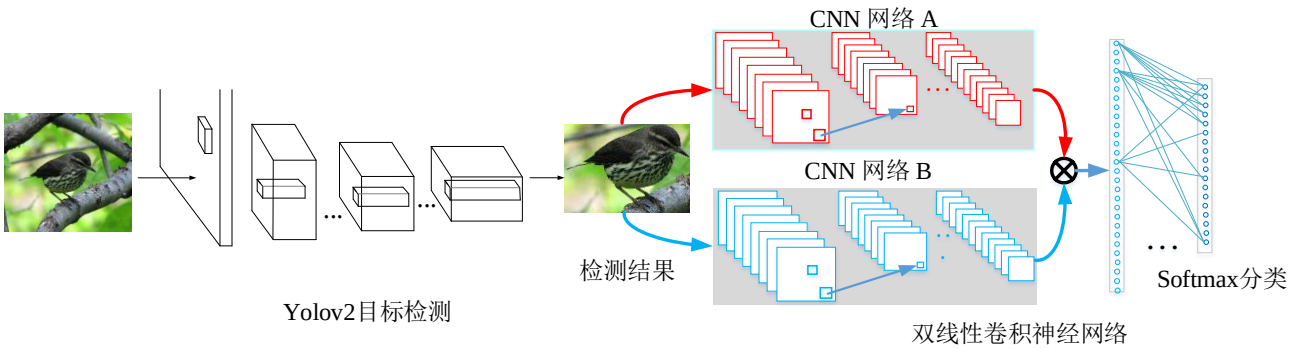

图 1　Yolov2+B-CNN 系统结构

Fig. 1　The system architecture of Yolov2 and B-CNN

如果一开始就选择有代表性的先验框维度，则模型对边界框的预测就更加准确。Yolov2采用k-means的方式对训练集的标注框做聚类，可以找到合适的先验框。在实现k-means聚类时，若选择欧氏距离为测度函数，尺寸较大的边界框会产生比较小的边界框更多的错误，通过引入交并比，用I表示，使得误差和边界框的大小无关，最终距离测度函数公式为：

$$d(\boldsymbol{g},\boldsymbol{h}) = 1 - I(\boldsymbol{g},\boldsymbol{h}), \tag{1}$$

式中 $\boldsymbol{h}$ 表示聚类中心框，$\boldsymbol{g}$ 表示人工标注框。$I(\boldsymbol{g},\boldsymbol{h})$ 表示聚类中心框和标注框的交并比，即二者交集面积与并集面积的比值，表示预测框的准确度，具体公式表示为：

$$I(\boldsymbol{g},\boldsymbol{h}) = \frac{\boldsymbol{g} \cap \boldsymbol{h}}{\boldsymbol{g} \cup \boldsymbol{h}} \tag{2}$$

最后得到的先验框的形状大多为细高型，扁平的居少。为平衡模型复杂度和召回率，选择先验框的个数为5。Yolov2算法使用先验框对检测网络的最后一层的特征图上直接预测，每个格子预测5个边界框，每个边界框都包含5个预测值：t_x，t_y，p_w，p_h 和置信值 t_o。先验框的引入会导致模型在训练过程中不稳定。若 p_w 和 p_h 表示先验框的宽和高，t_x，t_y 经过logistic函数处理后范围在0到1之间，c_x 和 c_y 表示网格偏离图像左上角的偏移量。相应的预测为：

$$\begin{aligned} b_x &= \sigma(t_x) + c_x \\ b_y &= \sigma(t_y) + c_y \\ b_w &= p_w e^{t_w} \\ b_h &= p_h e^{t_h} \\ c &= \sigma(t_o) \end{aligned} \tag{3}$$

式中 $\sigma(t_o)$ 为置信值，σ 为logistic激活函数，b_x，b_y，b_w，b_h 表示预测框的中心坐标和宽高。通过对位置预测进行限制后，模型参数更容易学习，使得模型更加稳定。相应的损失函数表示如下：

$$\begin{aligned} &\lambda_{\text{coord}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} \Delta_{ij}^{\text{obj}} \left[(b_{xi} - \hat{b}_{xi})^2 + (b_{yi} - \hat{b}_{yi})^2 \right] \\ &+ \lambda_{\text{coord}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} \Delta_{ij}^{\text{obj}} \left[(\sqrt{b_{wi}} - \sqrt{\hat{b}_{wi}})^2 + (\sqrt{b_{hi}} - \sqrt{\hat{b}_{hi}})^2 \right] \\ &+ \sum_{i=0}^{s^2} \sum_{j=0}^{B} \Delta_{ij}^{\text{obj}} (C_{ij} - \hat{C}_{ij})^2 \\ &+ \lambda_{\text{noobj}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} \Delta_{ij}^{\text{noobj}} (C_{ij} - \hat{C}_{ij})^2 \\ &+ \sum_{i=0}^{s^2} \Delta_i^{\text{obj}} \sum_{c \in \text{classes}} (p_i(c) - \hat{p}_i(c))^2 \end{aligned} \tag{4}$$

式中的 s^2 表示将图像划分的栅格数，B 表示每个栅格的预测框个数，$\hat{C}_{ij}$ 表示预测框中物体的置信值，C_{ij} 表示人工标注框内物体的置信值，$\hat{p}_i(c)$ 表示预测的栅格中包含物体且物体是某一类别的概率，$p_i(c)$ 表示栅格真实条件类别概率。Δ_{ij}^{obj} 表示第 i 个单元格存在目标，且该单元格预测的第 j 个边界框负责预测该目标，Δ_i^{obj} 表示物体是否出现在第 i 个栅格里，λ_{coord}，λ_{noobj} 分别表示位置预测和物体预测正则化惩罚系数。

1.2.2 判别性区域检测的网络结构

表1　Darknet19 网络结构

Table 1　network structure of Darknet19

类型	卷积核/个	尺寸/步长	输出
卷积	32	3×3	224×224
最大池化		2×2/2	112×112
卷积	64	3×3	112×112
最大池化		2×2/2	56×56
卷积	128	3×3	56×56
卷积	64	1×1	56×56
卷积	128	3×3	56×56
最大池化		2×2/2	28×28
卷积	256	3×3	28×28
卷积	128	1×1	28×28
卷积	256	3×3	28×28
最大池化		2×2/2	14×14
卷积	512	3×3	14×14
卷积	256	1×1	14×14
卷积	512	3×3	14×14
卷积	256	1×1	14×14
卷积	512	3×3	14×14
最大池化		2×2/2	7×7
卷积	1024	3×3	7×7
卷积	512	1×1	7×7
卷积	1024	3×3	7×7
卷积	512	1×1	7×7
卷积	1024	3×3	7×7
卷积	1000	1×1	7×7
平均池化		全局	1000
Softmax分类			1000

Yolov2算法提出了Darknet19的分类网络，该网络借鉴了VGG[17]分类网络结构，网络结构包括19个卷积层和5个全连接层。Darknet19大多采用3×3

卷积核，且在每一次池化操作后把通道数翻倍，其网络结构如表1所示。

Yolov2的检测网络是在改进分类网络结构的基础上得到的，首先使ImageNet数据集训练darknet19，输入图像大小为224×224，迭代160次，然后使用分辨率为448×448的图像微调网络，迭代10次，这可以使网络的卷积核更好地适应高分辨率图像的输入。去掉原网络最后一个卷积层，增加了3个3×3的卷积层，每层卷积核个数为1024，并且在每一个卷积层后面跟一个1×1的卷积层，在最后一个1×1的卷积层，卷积核个数为125，即每个格子检测所需要的数量。为了得到模型的细粒度特征，添加pass through层，将最后一个3×3×512的卷积层和倒数第二个卷积层特征堆叠，形成不同的通道，这样有利于小目标（细粒度特征）的检测。

1.3 基于双线性卷积神经网络的特征提取和识别

1.3.1 双线性卷积神经网络模型的结构

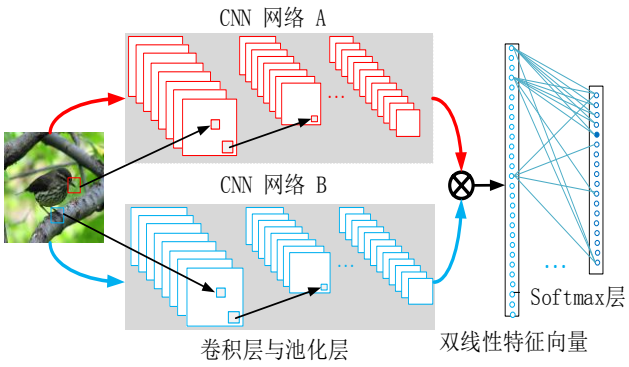

图 2 双线性卷积神经网络结构

Fig.2 Bilinear convolutional neural network architecture

双线性卷积神经网络模型由一个四元组 $\beta=(f_A, f_B, P, C)$ 组成。其中，f_A 和 f_B 是2个基于卷积神经网络的特征提取函数，分别对应于图2中的网络A和网络B，P 是一个池化函数，C 则是分类函数。特征提取函数 f 可以看成它接收一个 $\boldsymbol{i}\in\boldsymbol{I}$ 的图像块，相应区域位置满足 $\boldsymbol{l}\in\boldsymbol{L}$，其输出K×D大小的特征图，通过矩阵外积将每一个位置点的特征输出汇聚，也就是在 $\boldsymbol{l}$ 区域 f_A 和 f_B 的双线性特征的融合，公式如下：

$$b(\boldsymbol{l},\boldsymbol{i},f_A,f_B)=f_A(\boldsymbol{l},\boldsymbol{i})^{\mathrm{T}} f_B(\boldsymbol{l},\boldsymbol{i}) \quad \boldsymbol{l}\in\boldsymbol{L},\boldsymbol{i}\in\boldsymbol{I} \tag{5}$$

式中 f_A 和 f_B 必须具有相同的特征维度 K，K 的值取决于具体的模型。池化函数 P 的作用则是将所有位置的双线性特征汇聚以获得图像的全局特征 $\boldsymbol{\phi}(\boldsymbol{I})$，表示如下：

$$\boldsymbol{\phi}(\boldsymbol{I})=\sum_{\boldsymbol{l}\in\boldsymbol{L}} b(\boldsymbol{l},\boldsymbol{i},f_A,f_B)=\sum_{\boldsymbol{l}\in\boldsymbol{L}} f_A(\boldsymbol{l},\boldsymbol{i})^{\mathrm{T}} f_B(\boldsymbol{l},\boldsymbol{i}) \quad \boldsymbol{i}\in\boldsymbol{I} \tag{6}$$

在池化过程中，由于特征的位置信息被忽略，因此双线性特征 $\boldsymbol{\phi}(\boldsymbol{I})$ 是一个无序的特征表示。如果 f_A 和 f_B 提取的特征维度分别为 $K\times M$ 和 $K\times N$，则 $\boldsymbol{\phi}(\boldsymbol{I})$ 的大小为 $M\times N$ 的矩阵，将其转化为一个 MN×1 的列向量，作为最终的双线性向量。最后，通过 softmax 网络层进行分类。

1.3.2 特征提取

B-CNN可采用由卷积和池化层组成的M-Net[16]和D-Net[17]网络结构。由于目标数据集较小，我们在ImageNet[18]数据集上预训练B-CNN模型。分别从B-CNN网络的M-Net的relu5层和D-Net的relu5_3层分别提取特征。在网络的前向运算时，模型 f_A 和 f_B 可以是独立的、完全共享或部分共享，如图3所示。实验中采用完全共享的模型，即采用两个相同的截断在relu5_3层的D-Net网络。我们对输入图像，归一化为448×448大小，D-Net网络的输出大小为512个28×28的特征矩阵。经过外积运算和reshape操作得到一个512×512的双线性特征向量。

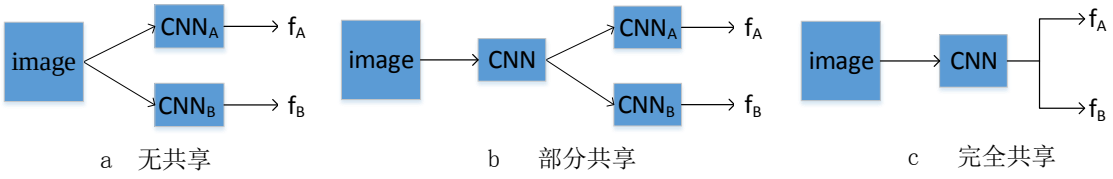

图3 双线性卷积神经网络前向运算方式

Fig.3 Forward operation ways of bilinear convolution neural network

1.3.3 归一化和分类

为提高分类精度，在得到双线性特征 $\boldsymbol{x}=\boldsymbol{\phi}(\boldsymbol{I})$ 后，对双线性特征进行带符号的开平方根及 l_2 归一化处理，公式如下：

$$\boldsymbol{y}=\operatorname{sgn}(\boldsymbol{x})\sqrt{|\boldsymbol{x}|} \tag{7}$$

$$\boldsymbol{z}=\boldsymbol{y}/\|\boldsymbol{y}\|_2 \tag{8}$$

经过较精确的检测和特征提取之后，可以使用常规的分类方法进行识别，如logistic回归或线性SVM。这里使用softmax分类层进行分类。

1.3.4 端到端的训练

B-CNN网络结构是一个有向无环图。通过分类损失函数的梯度反向传播完成网络参数的训练，如交叉熵。双线性形式简化了梯度运算。

如果两个网络的输出为矩阵$\boldsymbol{A}$和$\boldsymbol{B}$，其大小分别为 $\mathrm{L}\times\mathrm{M}$ 和 $\mathrm{L}\times\mathrm{N}$，则双线性特征为 $\boldsymbol{x}=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{B}$，大小为 $\mathrm{M}\times\mathrm{N}$。令 $\mathrm{d}l/\mathrm{d}\boldsymbol{x}$ 表示损失函数 l 对 $\boldsymbol{x}$ 的梯度，由梯度的链式法则，有：

$$\frac{\mathrm{d}l}{\mathrm{d}\boldsymbol{A}}=\boldsymbol{B}\left(\frac{\mathrm{d}l}{\mathrm{d}\boldsymbol{x}}\right)^{\mathrm{T}},\frac{\mathrm{d}l}{\mathrm{d}\boldsymbol{B}}=\boldsymbol{A}\left(\frac{\mathrm{d}l}{\mathrm{d}\boldsymbol{x}}\right) \tag{9}$$

计算得到特征$\boldsymbol{A}$和$\boldsymbol{B}$的梯度，则整个模型可以进行端到端的训练，梯度更新如图4。其他部分的训练和常规的CNNs网络相同。

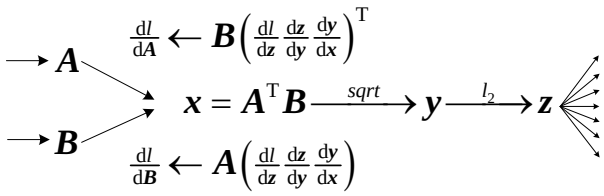

图4 双线性卷积神经网络梯度流

Fig.4 Gradient flow of B-CNN

2 实验结果与分析

2.1 数据集

为了验证改进后算法的效果，本文采用目前较常用的三个标准细粒度图像分类数据集CUB-200-2011[20]、Cars196[21]和Aircrafts100[22]进行测试。CUB-200-2011是细粒度图像分类领域最经典，也是最常用的一个数据集，共包含200种不同类别，共11788张鸟类图像。该数据集被分为训练集和测试集，且两者包含的图像数量大致相同。同时，该数据集提供了丰富的人工标注数据，每张图像包含15个局部区域位置，1个标注框和类别标签。由于该数据集的图像背景大多比较杂乱，且目标在图像中的占比不定，及诸多遮挡及姿态的因素，使得其分类

图 5 数据库图像示例

Fig.5 Database image example. Each row represents Arctic terns,Caspian swallows, AMM General Hummer SUV 2000 ,Acura RL Seddan 2012, Boeing737-300 and Boeing737-400 respectively.

任务具有一定的挑战性。Aircrafts100数据集提供100类不同的飞机照片，每一类包含有100张不同的照片，整个数据集共10000张图片，该数据集分为训练集，验证集和测试集，大小分别为：3334，3333，3333，只提供标注框信息和类别标签，由于数据集类内差距大，类别间的差距较小，如类别Boeing737-300与Boeing737-400在形状、颜色等方面很相似。Stanford Cars数据集共包含有16185张图像，提供196类不同品牌不同年份不同车型的车辆图像数据, 该数据集也分两个部分：训练集和测试集，大小分别为8144和8041,只提供标注框信息和类别标签。图5列出了这三个数据集的类别的若干样本。

2.2 实验结果和分析

首先，使用Yolov2预训练模型分别从CUB-200-2011、Cars196和Aircrafts100三个标准数据集中提取具有判别性的区域，并将其分辨率归一化为 448×448，然后微调双线性卷积神经网络模型。我们分两个步骤进行微调，首先将分类层的类别数替换为细粒度数据集相应的类别数，对最后一层的参数进行随机初始化，且只训练最后一层，接下来设置相对较小的学习率0.001使用随机梯度下降法通过反向传播微调整个模型，迭代次数大约在45-100。

最终在公开的细粒度图像集CUB-200-2011、Cars196和Aircrafts100上分别达到了84.5%、92%、88.4%的分类精度，和没有进行有区分度信息提取的数据集相比，分类准确度比原方法的最高分类精度高出0.4%，0.7%，3.9%。在表2中给出了我们的方法和多个国内外先进方法对比结果。

表2中，FV-SIFT[14]的字典大小为256，空间金字塔有1层。而FV+SIFT[23]字典大小为1024，使用的是多尺度SIFT特征，空间金字塔有2 层，为1×1，3×1区域。从表2可以看出，使用传统FV特征的分类正确率明显低于本文模型。FV-CNN[14]方法对CNN提取的特征进行FV编码，此方法使用两种不同的卷积神经网络模型M-Net和D-Net分别提取relu5和relu5_3的特征，由于FV采用混合高斯模型构建码本，且FV编码后的向量通常不是稀疏的，分类正确率也不高。FC-CNN[14]使用两个不同的卷积神经网络模型M-Net和D-Net对目标数据集分类，分类准确率有限。

表 2　birds，cars 和 aircrafts 数据集分类结果

Table 2　Classification results on birds，cars and aircrafts dataset

方 法	birds 分类结果(%)	cars 分类结果(%)	aircrafts 分类结果(%)
FV-SIFT[14]	18.8	59.2	61.0
FV+SIFT[23]	-	82.7	80.7
FV-CNN(M)[14]	64.1	77.2	71.2
FV-CNN(D)[14]	74.7	85.7	78.7
FC-CNN(M)[14]	58.8	58.6	63.4
FC-CNN(D)[14]	70.4	79.8	76.6
B-CNN(M)[14]	78.1	86.5	79.5
B-CNN(D)[14]	84.0	90.6	83.9
B-CNN(M,D)[14]	84.1	91.3	84.5
Part-based RCNN[6]	73.9	-	-
STNs[24]	84.1	-	-
BaseNet + SegNet[25]	-	86.74	83.43
Yolov2+B-CNN(D)	84.5	92.0	88.4

B-CNN模型需要同时学习两个卷积神经网络，这两个模型可以是对称的，例如学习两个相同的M-Net或两个相同的D-Net网络，也可以同时学习M-Net和D-Net这两个不对称的网络，B-CNN算法在CUB-200-2011、Cars196和Aircrafts100这三个数据集上得到的最高分类准确率分别为84.1%、91.3%、

84.5%，分类准确率比本文方法分别低0.4、0.7、3.9个百分点。和使用两个相同D-Net网络的B-CNN算法相比，在此三个数据集上，本文方法分别高出0.5、1.4、4.5个百分点。Part-based RCNN算法利用选择性搜索算法会产生大量无关的候选区域，造成计算上的浪费，在鸟类数据集上，其分类准确率比本文低10.6个百分点。STNs[24]算法的创新之处在于提出了带参数且可学习的网络层，根据任务自己学习图片或特征的空间变换参数，加入到已有的CNN或者FCN网络，能够提升网络的学习能力，但分类效果依然低于本文。BaseNet+SegNet[25]算法利用卷积神经网络对细粒度数据集进行初分类，得到基本网络模型。利用学习好的BaseNet生成自上而下注意图，再用注意图初始化GraphCut[26]算法，分割出关键的目标区域，从而提高图像的判别性。最后，对分割图像提取CNN特征实现细粒度分类，在Cars196和Aircrafts100数据集上分类准确率为86.74%和83.43%，但算法处理过程较为繁杂，且准确率低于本文。本文方法使用Yolov2预训练的模型快速找出判别区域，然后仅使用图像的类别标签完成B-CNN算法的学习和分类，其分类性能明显优于其他优秀的算法。

2.3 时间复杂度分析

为验证本方法的复杂度，我们与传统模型、改进前的B-CNN 模型进行测试时间复杂度分析和运行时间比较。

传统特征样本的特征维数为m，n为样本个数，则线性SVM 的测试时间复杂度为O(m*n) 。底层特征为128维的SIFT，则文献[14]中每张图像的特征维数分别为65536，而本文模型，每张图像的特征长度为512*512*k 维，k取决于数据集的种类，其特征维度远高于传统特征的维数。CNN分类模型中，VGG网络模型共包含138M参数，而本文采用的BCNN模型则包含58.9M 参数，其参数数量明显少于VGG网络模型。在Tesla K40 GPU中，B-CNN(M)模型的提取图像特征的运行时间为87帧/秒，B-CNN(D)模型特征提取的时间为10帧/秒，B-CNN(M,D)模型提取特征的时间为8帧/秒，由于本文方法是在B-CNN上的基础上增加了判别性区域检测部分，在Tesla K40 GPU中Yolov2检测图像中的判别性区域的速度为4帧/s。本文模型的运行时间相比于未改进的B-CNN算法而言，增加的运行时间部分为Yolov2检测判别性区域的时间。特征提取时间如表3所示，从中可以看出本文方法只是小幅度地提高了时间复杂度。

表3 特征提取时间比较

Table 3 The speeds of feature extraction

方法	时间（帧/秒）
FV-SIFT[14]	10
Part-based RCNN[6]	1/47
B-CNN(M)[14]	87
B-CNN(D)[14]	10
B-CNN(M,D)[14]	8
本文方法	4

3 结 论

本文针对细粒度图像分类中的背景干扰问题，提出基于深度学习的聚焦-识别框架，先检测图像中判别性区域再进行识别。具体是利用Yolov2预训练的模型聚焦图像中有区分度的区域，将聚焦结果输入B-CNN模型来对目标区域进行学习与分类，构造更具判别性的特征表示，从而提高分类性能。实验结果表明，在识别模型基础上引入聚焦方法能够进一步提高细粒度图像分类的正确率。相对于原B-CNN算法，Yolov2算法的引进能够过滤掉很多对细粒度图像分类没有贡献的图片内容，使得B-CNN能够学习到更多有利于细粒度图像分类的判别性特征，从而增强模型的鲁棒性，且整个模型仅使用图像的类别标签信息，对于其他的图像库也能适用，增加了模型的通用性。

参考文献(References)

[1] Lowe D G. Object recognition from local scale-invariant features [C]//Computer vision,1999. The proceedings of the seventh IEEE international conference on. Ieee, 1999, 2: 1150-1157. [DOI: 10.1109/ICCV.1999.790410]

[2] Dalal N, Triggs B. Histograms of oriented gradients for human detection [C]//Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, 2005, 1: 886-893. [DOI: 10.1109/CVPR.2005.177]

[3] Jégou H, Douze M, Schmid C, et al. Aggregating local descriptors into a compact image representation[C]//Computer Vision and Pattern Recognit-

ion (CVPR), 2010 IEEE Conference on. IEEE, 2010: 3304-3311. [DOI: 10.1109/CVPR.2010.5540039]

[4] Perronnin F, Dance C. Fisher kernels on visual vocabularies for image categorization[C] //Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007: 1-8. [DOI: 10.1109/CVPR.2007.383266]

[5] Sánchez J, Perronnin F, Mensink T, et al. Image classification with the fisher vector: theory and practice[J]. International Journal of Computer Vision, 2013, 105(3): 222-245. [DOI: https://doi.org/10.1007/s11263-013-0636-x]

[6] Zhang N, Donahue J, Girshick R, et al. Part-based R-CNNs for fine-grained category detection[C]// European conference on computer vision. Springer, Cham, 2014: 834-849. [DOI: https://doi.org/10.1007/978-3-319-10590-1_54]

[7] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and Semantic segmentation [C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587. [DOI: 10.1109/ CVPR.2014.81]

[8] Uijlings J R R, Van De Sande K E A, Gevers T, et al. Selective search for object recognition [J]. International Journal of Computer Vision, 2013, 104(2):154-171 . [DOI**:** 10.1007/s11263-013-0620-5]

[9] Branson S, Van Horn G, Belongie S, et al. Bird species categorization using pose normalized deep convolutional nets [J]. arXiv Preprint ArXiv: 1406.2952, 2014.

[10] Branson S, Beijbom O, Belongie S. Efficient large-scale structured learning [C] // Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, 2013: 1806-1813. [DOI: 10.1109/CVPR.2013.236]

[11] Xiao T, Xu Y, Yang K, et al. The application of two-level attention models in deep convolutional neural network for fine-grained image classification [C] // Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on. IEEE, 2015: 842-850.

[12] Zhang Y, Wei X S, Wu J, et al. Weakly supervised fine-grained categorization with part-based image representation [J]. IEEE Transactions on Image Processing, 2016, 25(4): 1713-1725. [DOI: 10.1109/TIP.2016.2531289]

[13] Simon M, Rodner E. Neural activation constellations: unsupervised part model discovery with convolutional networks [C] //Proceedings of the IEEE International Conference on Computer Vision. 2015: 1143-1151. [DOI: 10.1109/ICC V. 2015.136]

[14] Lin T Y, RoyChowdhury A, Maji S. Bilinear convolutional neural networks for fine-grained visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(6): 1309-1322. [DOI: 10.1109/TPAM I. 2017.2723400]

[15] Redmon J, Farhadi A. Yolo9000: better, faster, stronger [J]. arXiv Preprint, 2017. [DOI:10.1109/ CVPR.2017.690]

[16] Chatfield K, Simonyan K, Vedaldi A, et al. Return of the devil in the details: delving deep into convolutional nets[J]. arXiv preprint arXiv:1405.3531, 2014.

[17] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014..

[18] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database [C] // Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009: 248-255.［DOI: 10.1109/CVPR.2009.5206848］

[19] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788. .[DOI: 10.1109/CVP R.2016.91]

[20] Wah C, Branson S, Welinder P, et al. The Caltech-UCSD birds 200-2011 dataset [J]. California Institute of Technology, 2011.

[21] Krause J, Stark M, Jia D, et al. 3D Object Representations for Fine-Grained Categorization [C] // IEEE International Conference on Computer Vision Workshops. IEEE, 2014:554-561.［DOI:

10.1109/ICCVW.2013.77］

[22] Maji S, Rahtu E, Kannala J, et al. Fine-grained visual classification of aircraft [J]. arXiv preprint arXiv:1306.5151, 2013.

[23] Gosselin P H, Murray N, J égou H, et al. Revisiting the fisher vector for fine-grained classification [J]. Pattern Recognition Letters, 2014, 49: 92-98.［DOI: 10.1016/j.patrec.2014.06.011］

[24] Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks [C]//Advances in neural information processing systems. 2015: 2017- 2025.

[25] Feng Y S，Wang Z L. Fine-grained image categorization with segmentation based on top-down atention map [J]. Journal of Image and Graphics，2016，21(9) :1147-1154．［冯语姗，王子磊. 自上而下注意图分割的细粒度图像分类[J]. 中国图象图形学报，2016，21(9) : 1147-1154.］［DOI:10. 11834/ jig. 20160904］

[26] Boykov Y Y, Jolly M P. Interactive graph cuts for optimal boundary & region segmentation of objects in ND images [C]//Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on. IEEE, 2001, 1: 105 – 112.［DOI: 10.1109/ICCV.2001.937505］

作者简介：

王永雄，第一作者，通信作者，1970年生，男，副教授，2013年12月上海交通大学大学获控制理论与控制工程专业博士学位，主要研究方向为智能机器人及视觉。E-mail: wyxiong@ usst.edu.cn。

张晓兵，女，1991年生，硕士，研究方向为机器视觉、图像处理。E-mail: bingzxcn@163.com。