第18卷 第10期 2018年4月
1671—1815(2018)010-0240-05

科 学 技 术 与 工 程
Science Technology and Engineering

Vol. 18 No. 10 Apr. 2018

引用格式:李新叶 王光陛. 基于卷积神经网络语义检测的细粒度鸟类识别[J]. 科学技术与工程,2018,18(10):240—244
Li Xinye,Wang Guangbi. Fine-grained bird recognition based on convolution neural network semantic detection[J]. Science Technology and Engineering,2018,18(10):240—244

基于卷积神经网络语义检测的细粒度鸟类识别

李新叶 王光陛
(华北电力大学电子与通信工程系,保定 071003)

摘 要 细粒度识别的主要目的是在相同基本类别下对其繁多的子类别进行区分。不只局限于头和躯干的定位现状,提出了一种基于 Faster RCNN 联合语义提取和检测的分类方法。通过引入自上而下的方法来生成七个小语义部位,既大大减少了候选区域的个数,又提高了分类的效率。检测子网可以和区域候选生成网络(RPN)共享卷积特征,结果使得区域建议几乎不花时间,从而可以生成高质量并且具有局部特征的区域建议框,便于 Fast RCNN 的检测。相对于其他鸟类识别研究,实验中鸟类识别准确率达到了 88.37%,提高了识别效率。说明联合语义的 Faster RCNN 网络适用于鸟类的细粒度识别。

关键词 细粒度识别 Faster RCNN 语义特征 鸟类识别
中图法分类号 TP391.41; **文献标志码** A

细粒度分类主要目的是在相同基本类别下对其繁多的子类别进行区分。例如不同种类的鸟类[1]、花卉[2]等;本文针对鸟类进行细粒度识别研究。就细粒度图像分类任务而言,如何找到具有区分能力的关键特征显得尤其重要;而这种细粒度特征需要从复杂的图像背景中搜索到,提取的同时应尽可能减少包括光照、形变、遮挡等环境噪声的干扰。相对于粗粒度特征来说,细粒度特征的获取往往更加复杂、更加依赖图像的标注来确定模型中的复杂参数,从而尽可能地避免少量数据引起的过拟合现象。所以如何更好地提取和检测特征以及选取卷积神经网络的合适结构和网络的连接方式是其中的关键性问题。

传统的 CNN 网络往往缺乏生成对象部分语义的卷积层。在一些目标精细识别分类的情况下,卷积神经网络的低层通常是获取一些初级特征;而高层是获取此类的高级特征。之前有的研究方向是利用低等级的图像特征来达到局部定位和局部特征提取的目的。DPM[3]和 Poselet[4]一直广泛用在不同的位姿中的局部定位。

文献[5]在 CNN 网络基础上在局部定位、局部提取和细分类方面做了改进,明显优于之前依赖于手工特征提取的一些研究。Zhang 等[6]采用自顶而下的选择性搜索(selective search)[7]生成包含部分或对象的区域;采用 RCNN[8]实现定位。但是选择性搜索并不存在几何约束,生成的区域可能过多且很难生成较小的语义部分区域;因此只能定位到鸟类的头和躯体,并且 RCNN 不是 End-to-End 结构,而是采用的 Region + CNN + SVM 框架,无法统一得到训练。Lin 等[9]采用 CNN 特征直接回归得到局部区域;但是,这个方法仍然只能定位到头和躯体。

不只局限于头和躯干的定位结果,提出了一种基于 Faster RCNN 联合语义特征提取和检测的研究方案。引入了类似于 K-NN 几何限制的部位候选方法来产生七个区域候选部位;根据约束条件对细微区别进行特征提取,这种方法在数量级上大大减少了候选区域的个数。然后与区域建议网络(region proposal networks,RPN)共享卷积层,可以得到更加精确的特征和定位结果;二者结合用于回归分类进行训练,共享卷积计算结果,再输入到分类子网中,从而达到对于鸟类细粒度更加准确的分类结果。

1 Faster RCNN 联合语义的网络构建

基于 Faster RCNN 联合语义特征提取和检测的网络的具体结构如图 1 所示。

首先对图像进行相关的图像预处理,对网络(ZF 网络)中的卷积层进行改进,引入语义检测用于提取可用于更好识别的局部特定特征,比如鸟的鸟喙、翅膀、鸟腿等特征。利用检测子网中的卷积神经网络获得更准确的局部边界框,接着通过语义部分 ROI 池化层将检测子网络中检测到的语义部分中的特征提取出来,并且按照预定义的顺序排列好,

2017年8月21日收到 河北省教育厅指导性计划(Z2012038)资助
第一作者简介:李新叶(1969—),女,博士,副教授。研究方向:计算机视觉、信息检索。E-mail:lxyyjh@126.com。

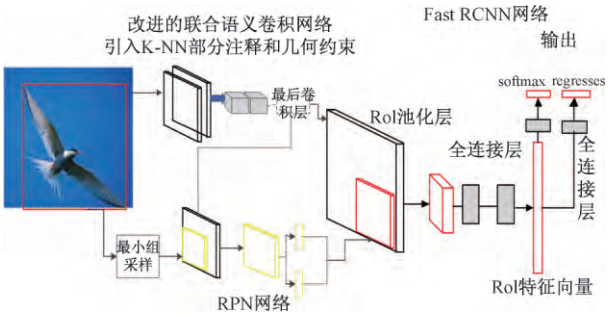

图 1　联合语义检测的 Faster RCNN 网络结构

Fig. 1　Joint semantic parts faster RCNN network structure

送到分类子网中进行分类。

2　网络训练

联合语义检测和提取的 Faster-rcnn 训练的具体步骤如下。

2.1　引入语义的卷积网络

提出了一种自上而下的区域候选方法，并对卷积层(conv_3) 进行几何限制。通过对目标边界框的计算确定梯度直方图(HOG) 特征; 以此作为目标的一个粗略形状; 然后依据 HOG 特征从指定的训练集中选取个相邻图片; 再将得到的个相邻图片中个部件范围调整到合适的大小。$\boldsymbol{H}=[h_{11},h_{12},\cdots,h_{1m},h_{21},h_{22},\cdots,h_{2m},h_{k1},h_{k2},\cdots,h_{km}]$ 表示所有的 k 个近邻 m 个局部的区域。调整后的部件区域依然保留了之前的先验知识，例如部件区域标签以及几何形状。提出以下先验策略并利用得到的部件产生区域候选: 部件的类标签和几何限制。第 i 个部件的候选区域将通过 H 矩阵中的 $[h_{1i},h_{2i},\cdots,h_{ki}]$ 给出，因此每个部件的候选区域为 k 个，最后得到的部件总数为 $N=km$。

由于和的值比较小，相比于自上而下的方法(如选择性搜索) 所获得候选区域个数也将减少一个数量级。

2.2　区域建议网络

选取任意大小的图像作为输入，将目标建议框的集合作为输出。在最后一个共享卷积层输出的卷积特征映射上滑动小网络来生成区域建议框，再将这个网络与输入卷积特征映射为 $n\times n$ 的空间窗口进行全连接。每个滑动窗口对应映射到一个低维向量，这个低维向量再输出给两个同级的全连接层: 边框回归层(reg-layer) 和分类层(reg-layer) 。边框回归层是用来预测候选区域中心锚点相对应的候选框坐标以(x,y) 及宽(w) 、高(h) ; 而分类层则是用来判定该候选区域是前景还是背景的概率。

在每一个滑动窗口的位置，预测 k 个候选区域建议，因此边框回归层有 $4k$ 个输出，即对 k 个边框的坐标进行编码; 分类层则输出 $2k$ 个得分，即对每个候选建议框进行是否为目标的概率估计。

对一个图像的损失函数定义为

$$L(p_i,\boldsymbol{t}_i)=\frac{1}{N_{\text{cls}}}\sum_i L_{\text{cls}}(P_i,P_i^*)+\lambda\frac{1}{N_{\text{reg}}}\sum_i P_i^* L_{\text{reg}}(\boldsymbol{t}_i,\boldsymbol{t}_i^*) \tag{1}$$

分类损失:

$$L_{\text{cls}}(P_i,P_i^*)=-\lg[P_i^* P_i+(1-P_i^*)(1-P_i)] \tag{2}$$

回归损失:

$$L_{\text{reg}}(\boldsymbol{t}_i,\boldsymbol{t}_i^*)=R(\boldsymbol{t}_i-\boldsymbol{t}_i^*) \tag{3}$$

式(3) 中为鲁棒损失函数:

$$smooth_{L_1}(x)=\begin{cases}0.5x^2, & |x|>1\\ |x|-0.5, & \text{其他}\end{cases} \tag{4}$$

i 是一个最小采样中锚点的指数，P_i 是对锚点是否为目标的预测概率。如果锚点生成框为正，则对应标签(ground-truth) 下的 P_i 就是 1。相反如果为负，则对应标签(ground-truth) 下的 P_i 就是 0。是一个向量，表示预测的边界框的 4 个参数化坐标，$\boldsymbol{t}_i$ 是与锚点生成框为正时所对应的边界框的坐标向量。

对于回归，主要采用 4 个坐标:

$$\begin{cases}t_x=(x-x_a)/w_a\\ t_y=(y-y_a)/h_a\\ t_w=\lg(w/w_a)\\ t_h=\lg(h/h_a)\end{cases} \tag{5}$$

$$\begin{cases}t_x^*=(x^*-x_a)/w_a\\ t_y^*=(y^*-y_a)/h_a\\ t_w^*=\lg(w^*/w_a)\\ t_h^*=\lg(h^*/h_a)\end{cases} \tag{6}$$

式(6) 中 x、y、w、h 指的是候选框中心的坐标、宽和高，其中 x、x_a、x^* 分别指的是预测建议框、锚点边界框和正确标注边界框(ground-truth box) 的 x 坐标。同理对于 y、w、h 也是一样的，由此可以认为是从锚点边界框到附近正确标注边界框的回归。

以前的特征映射方法中，通过在任意大小区域中已经池化到的特征来进行边界框回归，这些尺寸大小不同的区域共享回归权重。而上述特征映射方法中那些用来进行回归的特征都具有相同的空间大小($n\times n$) 。鉴于空间大小差异的原因，要学习一系列 k 个边框回归量，各个回归量都有其相对应的尺度和长宽比，它们之间不共享权重。尽管特征具有固定的尺寸或者尺度，用来预测各种尺寸或尺度的边界框仍然是可能存在的。

2.3　区域建议与目标检测共享卷积特征

以上已经描述了如何通过改进卷积层生成小语

义候选部分和生成区域建议训练网络，通过交替优化来学习共享的特征。

步骤 1 需要用 ImageNet 预训练的模型对 RPN 网络其进行初始化，然后需要对 end-to-end 网络进行微调，以完成生成区域候选的任务。

步骤 2 用 Fast R-CNN 来训练一个单独的检测网络，这个检测网络的第三层卷积网引入了小语义部分，这个检测网络也需要由 ImageNet 预训练的模型初始化的，到目前这两个网络其实还没有共享卷积层。

步骤 3 用检测网络来对 RPN 网络进行初始化训练，然后去固定它们之间共有的卷积层，对 RPN 网络中独有的层进行微调，此时这两个网络得以共享卷积层了。

通过 Fast-RCNN 对上面共享的卷积检测结果进行区域回归，回归出每一个区域的边界包围框和一个置信度。因为每个目标具有 m 个部件，所以回归网络则具有 $m+1$ 个输出，其中包含背景区域在内。

通过 Fast-RCNN 对上面共享的卷积检测结果进行区域回归，回归出每一个区域的边界包围框和一个置信度。因为每个目标具有个部件，所以回归网络则具有个输出，其中包含背景区域在内。

通过 RCNN 回归整个网络使用多任务 Loss 函数:

$$L(s,b,c,b^{gt}) = L_{cls}(s,c) + \lambda [c>0] L_{loc}(b^c,b^{gt}) \tag{7}$$

式(7)中 c 是部件边界框的正确类别分类，取值为 $0\sim m$，$L_{cls}(s,c)$ 代表正确类别分类的损失函数，L_{loc} 表示部件边框损失，b^c 为正确类别的边框回归，b^{gt} 是正确标注框的回归。

3 统一网络

将检测子网和分类子网放在一起，利用轮流训练的方式进行训练，具体步骤如下。

首先，两个检测子网和分类子网使用需要使用 ImageNet 的预训练模型进行初始化训练以及微调。三个子网络有着不同的卷积层，其中中层卷积层还需要改进。由于这个原因，将分类子网络中前 n(n 取值为 5)个卷积层用来代替检测子网络中相对应的卷积层，再对检测子网络中中层卷积网络层进行调整引入小语义部分，对其中细微部分进行检测。最后，利用检测子网络的检测结果对分类子网络中共享卷积层之外的网络层进行微调，即上一步中所述的前 n 个卷积层不再需要微调。至此两个子网络将拥有相同的卷积层，并最终构成一个统一的网络。

4 实验

采用 ZF 网络对其进行训练，共有 5 个卷积层、2 个池化层和 2 个全连接层。为了验证提出的方法在鸟类数据集分类识别中的效果，进行了三个部分实验。

(1) 将引入的候选区域方法用于小部分语义检测，并与传统方法进行平均正确率比较。

(2) 将提出的联合语义检测和分类应用在数据集识别中，比较引入小语义部件后识别效果。

(3) 使用传统的网络进行实验，并与本文方法的识别效果进行比较。

4.1 数据集

实验所使用的鸟类数据集 CUB-200-2001 包含 200 种鸟类，比如黑背信天翁、美洲乌鸦等，共 11 788张图片。训练集包含 52 种鸟类，共 3 068 张图片，测试集包含 100 种鸟类，共 5 898 张图片。验证集包含 48 种鸟类，共 2 822 张图片。

4.2 参数设置

首先要将所用数据集图片改成相对应的 XML 格式的文件。

卷积层 C1、C2、C3、C4、C5 中卷积核大小分别为 7×7、5×5、3×3、3×3、3×3。池化区域为 3×3。

卷积层采用的激活函数是修正线性单元(rectified linear units，ReLU)[10]。采用带有冲量项的随机梯度下降法来优化网络。冲量项设置为 0.95。学习率的初始值设置为 0.001，训练后期逐步接近最优值，根据迭代次数调整为 0.000 1。权值设置为 0.000 5。通过实验验证，将最近邻算法的取值定为 20。

对于训练集中的每张图像进行如下处理。

(1) 把每个标出的真值候选区域，与其重叠比例最大的锚点矩形框进行结合，这就记为前景样本。

(2) 对步骤(1)剩余的锚点矩形框进行比对，假如其与某个标定区域重叠比例大于 0.7，就可以记为前景样本；如果其与任意一个标定的区域重叠比例都小于 0.3，则记为背景样本。

(3) 对前两步训练剩余的锚点矩形框，弃去不用。

(4) 超过图像边界的锚点矩形框也弃去不用。

4.3 实验结果与分析

实验结果如表 1 所示。相对于其他部位，头部的正确率最高，其次是腿部。提出的引入 K-NN 候选区域方法相对于其他方法在提取特征方面有了显著效果，平均正确率达到了 73.05%，比边缘信息框(edge box)[11] 和选择性搜索[8] 方法都有所提高，但是通过其他细小部件进行提取特征的正确率[12] 还有待进一

步提高，总体来说联合语义的区域候选方法（K-NN）在鸟类特征提取方面展现出了很好的效果。这是由于之前的方法采用自下向上法，忽略了几何限制导致小语义部件区域候选产生不是很理想。小语义部位检测如图 2 所示，红色框为正确标签框，提出的语义检测（蓝色框）达到了明显效果。

表 1　与先进方法结果比较

Table 1　Comparison with the results of advanced methods

部件	正确度/%		
	边缘信息框[11]	选择性搜索[12]	引入 K-NN
头部	90.54	90.80	90.89
胸部	50.08	51.79	67.78
腹部	48.61	50.98	63.82
背部	35.66	56.07	76.13
翅膀	53.03	62.09	65.03
尾部	43.28	63.87	67.68
腿部	66.28	66.26	80.03
平均	55.35	63.12	73.05

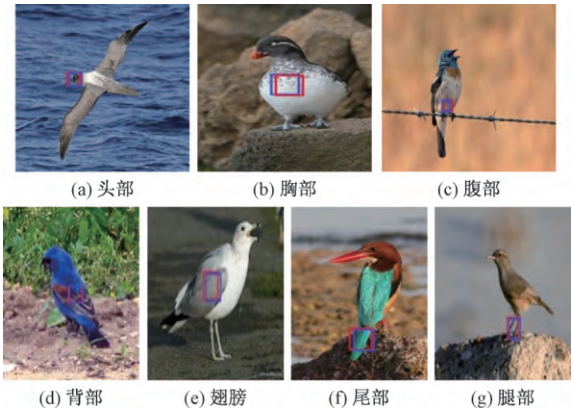

图 2　小语义部位检测

Fig. 2　Joint small semantic parts detection

引入小语义的 faster RCNN 方法与其他特征提取方法的识别结果比较如表 2 所示。由表 2 可以看出，联合语义来进行特征检测和提取的细粒度识别取得了很好的效果，关注更多具有区分特征的部件使得识别正确率达到了 88.37%，实验说明联合小语义区域候选的方法可以更好地融合到网络中来提高识别精度，更加说明了局部特征对图像识别任务的重要性。

表 2　引入小语义的 faster RCNN 方法与其他特征提取方法的识别结果比较表

Table 2　Introducing small semantic part method compared with the traditional methods of identification

	分类	正确率/%
1	目标整体	67.02
2	一个部位	71.68
3	两个部位	75.45
4	七个部位（fast rcnn）	85.14
5	七个部位（faster rcnn + ZF）	88.37

5　结论

为了更好地解决细粒度图像识别分类问题，以鸟类识别为例，提出了一种联合语义部分进行特征检测和提取的 Faster rcnn 方法，通过在卷积层引入自上而下的候选区域建议方法，用于鸟类的细粒度识别分类，并取得了良好的识别效果。得到的识别准确率比原有特征提取的卷积神经网络方法得到了很大的提高。实验中使用的数据集、参数的调节以及获取关键特征能力的限制，对实验结果有一定影响。因此下一步的努力方向是如何更好地发掘有效特征来便于识别。

参　考　文　献

1　Wah C，Branson S，Welinder P，*et al*. The Caltech-UCSD birds-200-2011 dataset. San Diego：California Institute of Technology，2011

2　Angelova A，Zhu S，Lin Y. Image segmentation for large-scale subcategory flower recognition. 2013 IEEE Workshop on Applications of Computer Vision（WACV）. New York：IEEE，2013：39—45

3　Felzenszwalb P F，Girshick R B，Mcallester D，*et al*. Object detection with discriminatively trained part-based models. IEEE Transactions on Software Engineering，2014；32（9）：1627—1645

4　Farrell R，Oza O，Zhang N，*et al*. Birdlets：Subordinate categorization using volumetric primitives and pose-normalized appearance. IEEE International Conference on Computer Vision. New York：IEEE Computer Society，2011：161—168

5　Krause J，Jin H，Yang J，*et al*. Fine-grained recognition without part annotations. IEEE Conference on Computer Vision and Pattern Recognition（CVPR）. New York：IEEE，2015：5546—5555

6　Zhang N，Donahue J，Girshick R，*et al*. Part-based R-CNNs for fine-grained category detection. European Conference on Computer Vision，2014；8689：834—849

7　Girshick R，Donahue J，Darrell T，*et al*. Rich feature hierarchies for accurate object detection and semantic segmentation. IEEE Conference on Computer Vision and Pattern Recognition（CVPR），New York：2014：580—587

8　Uijlings J R，Sande K E，Gevers T，*et al*. Selective search for object recognition. International Journal of Computer Vision，2013；104（2）：154—171

9　Lin D，Shen X，Lu C，*et al*. Deep LAC：deep localization，alignment and classification for fine-grained recognition. IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE Computer Society，2015：1666—1674

10　Dahl G E，Sainath T N，Hinton G E. Improving deep neural networks for LVCSR using rectified linear units and dropout. IEEE International Conference on Acoustics，Speech and Signal Processing. New York：IEEE，2013：8609—8613

11　Zitnick C L，Dollár P. Edge boxes：locating object proposals from edges. European Conference on Computer Vision. Berlin：Springer International Publishing，2014：391—405

12　吴正文. 卷积神经网络在图像分类中的应用研究. 成都：电子科技大学，2015

Wu Zhengwen. Application of convolution neural network in image classification. Chengdu: University of Electronic Science and technology of China, 2015

Fine-grained Bird Recognition Based on Convolution Neural Network Semantic Detection

LI Xin-ye, WANG Guang-bi

(Department of Electronic and Communication Engineering, North China Electric Power University, Baoding 071003, China)

[Abstract] The main purpose of fine-grained identification is to distinguish between its many subcategories under the same basic categories. The localization of head and torso was not only confined, and proposed a classification method based on Faster RCNN joint semantic extraction and detection. By introducing the top-down method to generate seven small semantic parts, it greatly reduces the candidate region number, but also improve the efficiency of classification. Since the detection subnet can share the convolution characteristics with the regional proposals network (RPN), the result is that the area is proposed to take almost no time, so that it can generate high quality and local characteristics of the regional proposal boxes to facilitate the detection of Fast RCNN. Compared with other bird recognition studies, the accuracy of birds recognition in the experiment reached 88.37%, which improved the recognition efficiency. The joint semantic Faster RCNN network is suitable for fine-grained recognition of birds.

[Key words] fine-grained identification Faster RCNN semantic parts features birds recognition