DOI:10.11830/ISSN.1000-5013.201706028

卷积特征图融合与显著性检测的图像检索

聂一亮,杜吉祥,杨麟

(华侨大学 计算机科学与技术学院,福建 厦门 361021)

摘要: 针对基于深度学习的图像检索提取特征往往包含了复杂的背景噪声,导致图像检索的精确率并不高的问题,提出一种特征图融合与显著性检测的方法.首先,训练用于分类的深度卷积神经网络模型.然后,并将图像卷积之后的特征图谱进行融合,得到图像的显著性区域.最后,通过计算图像显著性特征的余弦距离来进行检索.实验结果证明:相比目前主流的方法,文中方法能够有效提高检测精度,且鲁棒性较高.

关键词: 图像检索;特征图融合;显著性检测;卷积神经网络

中图分类号: TP 391 **文献标志码:** A **文章编号:** 1000-5013(2018)06-0937-05

Image Retrieval Based on Convolution Feature Map Fusion and Saliency Detection

NIE Yiliang, DU Jixiang, YANG Lin

(College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China)

Abstract: Based on an in-depth learning of image retrieval, the features extracted usually contained the complicated background noises, which resulted in a low level of accuracy in image retrieval. The methods of feature map fusion and saliency detection are proposed in this paper. The method firstly trained deep convolutional neural network model used in image classification, and then fused the features of maps after image convolution in order to obtain the salient region of retrieved images. Finally, the retrieved images are calculated using the cosine distance of the salient features. The experiment shows that the proposed methods are able to effectively improve the accuracy of retrieval and that the robustness is relatively high, compared to the current mainstream methods.

Keywords: image retrieval; feature map fusion; saliency detection; convolutional neural network

近些年,深度学习在各种计算机视觉任务中都取得了重大的突破,其中包括基于内容的图像检索(CBIR)[1]任务,目的是通过分析图像内容检索图像.在检索任务中,图像的特征表达与相似性度量成为了图像检索中的关键任务,尽管有很多手工描述子用于提取图像特征,如具有尺度和旋转不变性的尺度不变特征变换(SIFT)[2]算法、在人脸识别广泛应用的局部二值模式(LBP)高效算子[3]、行人检测中的方向梯度直方图(HOG)特征描述子[4],以及基于全局特征的 GIST(gist of the scene)[5]描述子等,但它们是浅层特征,视觉特征的描述能力十分有限.针对大规模图像检索,许多研究旨在解决如何快速和有

收稿日期: 2017-06-28

通信作者: 杜吉祥(1977-),男,教授,博士,主要从事模式识别及图像处理研究. Email:jxdu@hqu.edu.cn.

基金项目: 国家自然科学基金资助项目(61673186, 61370006, 61502183);福建省自然科学基金资助项目(2013J06014, 2014J01237);华侨大学中青年教师科研提升资助计划项目(ZQN-YX108);华侨大学研究生科研创新培育计划资助项目(1511314007)

效地在大规模数据中检索相似图像[6-7],而在大规模图像集中仅仅检索出语义相似的图像并不能满足用户日常的检索需求.在大多数情况下,用户更希望检索出实例级别的相似图像,而不仅仅是语义相似的同类图像.传统的方法是基于视觉词袋模型(BOVW)的图像检索,但由于BOVW最初提取的特征是传统手工描述子,抽取的特征比较低级,无法很好地描述图像的高层语义信息,因此,本文利用融合网络深层特征图谱来获取图像的显著性区域,从而筛去背景干扰信息,并结合阈值分割与局部特征向量编码方法进行实验.

1 图像检索方法

根据卷积神经网络的特性,网络浅层提取的是一些细节特征,深层才开始学习到一些总体特征,并提炼出某些语义信息.文中算法采用GoogLeNet网络模型进行微调,GoogLeNet网络最核心的地方在于其Inception结构,这类结构具有强大的局部细节表征能力.由于图像通常具有总体表征和局部细节这两种特征,只采用一种尺度的卷积核就不能充分捕捉到这两种特征的信息.

基于深度学习的图像检索一般都是提取最后一个卷积层或全连接层的特征直接进行相似度计算[8],导致最后检索出的结果虽然是语义同类的图像,但是图像间的局部细节并不相似.例如,在电商平台进行纹饰服装或箱包图像搜索时,不同的服装或箱包可能会因为局部的某一个细节(如衣服的纹理、领口和袖口)而区分[9],使用户检索不到相同款型的服装和箱包.因此,在高层语义保持一致的情况下,尽可能多地让特征包含局部细节信息显得尤为重要.基于卷积特征图融合概要图,如图1所示.

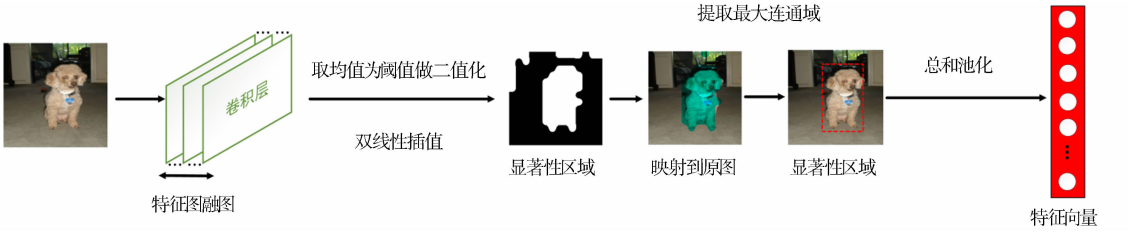


图1 特征图融合概要图

Fig. 1 Overview image of feature map fusion

首先,将GoogLeNet网络中经过前向运算得到的Inception结构的输出特征图谱进行叠加融合,其计算过程表示为

$$\boldsymbol{f}_{\text{add}} = \sum_{i=1}^{K} \text{feature}_i. \tag{1}$$

式(1)中:K为Inception结构输出的特征图谱的个数;feature_i为一个平面二维向量,大小为$H\times W$,它表示输出的第i个特征图谱;$\boldsymbol{f}_{\text{add}}$为这$K$个特征图谱叠加的结果,由于每个特征图谱的大小是相同的,所以最终得到一个大小为$H\times W$的二维平面向量.$\boldsymbol{f}_{\text{add}}$特征图谱中值越大的区域,网络的激励越高,说明此区域有网络学习到的对象.

为了得到特征图谱中高响应区域的具体位置,筛去干扰的背景信息,利用图谱的均值进行阈值分割,计算过程可表示为

$$\boldsymbol{f}_{\text{mask}}(m,n) = \begin{cases} 1, & \text{if } \boldsymbol{f}_{\text{add}}(m,n) \geqslant \text{mean}(\boldsymbol{f}_{\text{add}}), \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

式(2)中:$\boldsymbol{f}_{\text{add}}(m,n)$为特征图谱中的对应的每一个像素点;$\text{mean}(\boldsymbol{f}_{\text{add}})$为整个特征图谱的均值,将小于均值的位置取0,反之,取1,从而得到整个特征图中激励比较大的区域;$\boldsymbol{f}_{\text{mask}}$为阈值分割后的标记图谱,它的大小与$\boldsymbol{f}_{\text{add}}$相同.

对标记图谱使用双线性插值将其变换到原图大小,双线性插值可以使区域边缘更加地圆滑.然后,将标记图谱映射到原图上显示,实验发现需要检索的实例对象几乎都在区域里面.为了进一步得到原图的显著性区域,提取出标记图谱的最大连通域,这样做可以除去其他小面积非显著性区域的影响.为了方便观察,在实验中将得到的显著性区域用蓝绿色进行渲染,并用红色虚线矩形框标记出来.

最终只需提取出图像显著性区域对应的特征,再将其编码生成最终的特征向量进行检索.首先,将

标记图谱 $\boldsymbol{f}_{\text{mask}}$ 依次与 K 个 feature$_i$ 取交集，得到 K 个显著性特征图 $\boldsymbol{S}(\boldsymbol{S}_1,\boldsymbol{S}_2,\cdots,\boldsymbol{S}_k)$，其计算过程为

$$\boldsymbol{S}_i = \boldsymbol{f}_{\text{mask}} \cap \text{feature}_i. \tag{3}$$

由于每张图像的显著性区域大小不一，故采用文献[10]聚合处理可以消除大小影响，使得每张图像最后的向量维度都一致. 这种局部向量聚合方法不仅计算简便，而且使融合后的特征表达效果更好，其计算可表示为

$$v_i = \sum_{y=1}^{a}\sum_{x=1}^{b}\boldsymbol{S}_i(x,y). \tag{4}$$

式(4)中：$\boldsymbol{S}_i$ 为上一步得到的显著性特征图；a,b 分别为显著性特征图的高度与宽度；x,y 为 $\boldsymbol{S}_i$ 上的空间坐标. 文献[10]实际上是对得到的每一个显著性特征图 $\boldsymbol{S}_i$ 上的所有元素进行累加求和，这里有 K 个特征图，最终会生成一个 $K\times1\times1$ 的向量 $\boldsymbol{V}(v_1,v_2,\cdots,v_k)$，即得到图像检索需要的特征向量.

图像显著性检测结果的好坏直接影响图像分类与检索任务的性能[11]，在实验中选择 GoogLeNet 网络中哪一层 Inception 结构输出的特征图融合成为了需要研究的问题. 为了便于观察，实验将 GoogLeNet 网络中的 11 个卷积层生成的特征图谱分别进行融合，可视化 GoogLeNet 网络中融合各个卷积层之后所得的标记图谱示意图，如图 2 所示. 由图 2 可知：浅层的特征更加侧重于一些角点和边缘信息，并且随着层数的增加，网络高响应的区域往往是实例对象存在的区域[12]，使得提取的深层特征更加具有针对性，便于更好地检索. 在经过大量的显著性检测后发现，Inception4d 与 Inception4e 两个层生成的特征图谱进行融合之后的效果都比较好，并且有互补的效果. 所以，实验将这两个层阈值分割之后的特征图谱 $\boldsymbol{f}_{\text{mask-4d}}$ 与 $\boldsymbol{f}_{\text{mask-4e}}$ 取交集，得到 $\boldsymbol{f}_{\text{mask}}$，然后，再提取交集的最大连通域，得到更加准确的标记图谱，其计算可表示为

$$\boldsymbol{f}_{\text{mask}} = \boldsymbol{f}_{\text{mask-4d}} \cap \boldsymbol{f}_{\text{mask-4e}}. \tag{5}$$

图 2　标记图谱示意图

Fig. 2　Image of marker map

2　实验结果与分析

2.1　评估标准

为评估图像的检索性能，采用平均精度均值(MAP)度量方法，计算过程分为两步. 第一步，计算平均准确度(AP)，对不同召回率上的准确度进行平均. 假设检索有 N 个结果，K 个相关图像，返回的排序位置分别为 $x_1,x_2,\cdots,x_K$，则单个类别的平均准确度 AP$_i$ 表示为 AP$_i=(1/x_1+2/x_2+\cdots+K/x_K)/K$. 第二步，对 AP 进行算术平均，假设检索总次数为 M，则平均精度均值 MAP 表示为 $\text{MAP}=\frac{1}{M}\sum_{i=1}^{M}\text{AP}_i$.

2.2　数据集

为验证文中方法的有效性，实验分别在 CUB200-2011 和 Stanford Dogs 两个细粒度数据集上进行. CUB200-2011 鸟类数据集含 200 类，共 11 788 张图像，其中，训练集有 5 994 张，测试集有 5 794 张. Stanford Dogs 狗类数据集含 120 类，共 20 580 张图像，其中，训练集有 12 000 张，测试集有 8 580 张.

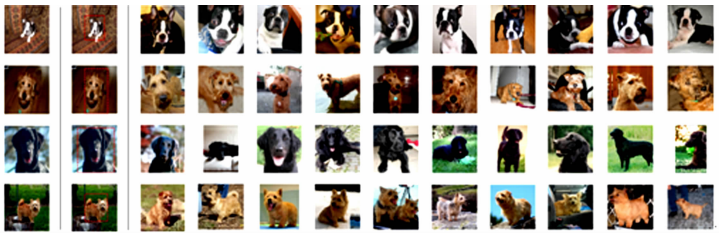
2.3 实验配置

在GPU为Tesla K40c且内存大小为16 GB的机器上,使用开源的Caffe[13]深度学习框架搭建网络,并采用随机梯度下降进行微调训练,设置初始学习率为0.01,批大小为50,共迭代50 000次.数据预处理阶段,在输入层使用Crop对图像进行裁剪,并将每个图像的大小统一重置为224 px×224 px.微调之后,提取GoogLeNet中Inception4d与Inception4e层的特征并进行融合,将融合后特征图谱的均值mean(f_{add})作为分割阈值,在提取显著性区域特征之后,使用拓展查询进行检索,拓展查询是对返回的前M个结果,包括查询样本本身,对它们的特征求和取平均,再做一次查询,实验中取M值为10.

2.4 结果分析

在CUB200-2011和Stanford Dogs数据集上Top 10的显著性检测,以及图像检索结果,如图3所示.图3中:第1栏为检索输入的原图像;第2栏为原图像的显著性检测结果;最后1栏为检索返回的Top 10图像.由图3可知:在对两个数据集图像进行检索时,算法不仅可以检测出同属类别的相关图像,包含颜色相近、外形相似的鸟与狗,并且检索返回鸟与狗的图像在姿态上更加相似.

(a) CUB200-2011



(b) Stanford Dogs

图3 显著性检测以及图像检索结果

Fig. 3 Significance detection and image retrieval results

为验证算法的高效性,将实验结果与其他主流图像检索方法在CUB200-2011和Stanford Dogs数据集上进行比较,包括SPoC(sum pooling of convolution)[10],CroW(cross-dimensional weighting)[14]及SCDA(selective convolutional descriptor aggregation)[12]方法等.CUB200-2011和Stanford Dogs数据集上MAP值对比,如表1所示.

表1 MAP值对比

Tab. 1 MAP values comparisons

方法	维度	MAP(CUB)	MAP(Dogs)
fc8_im	4 096	0.481	0.727
fc8_gtBBox	4 096	0.553	0.766
fc8_predBBox	4 096	0.531	0.741
SPoC[10](w/o cen.)	256	0.425	0.559
SPoC[10](with cen.)	256	0.473	0.557
CroW[14]	256	0.597	0.684
SCDA[12]	1 024	0.658	0.752
文中算法	528	0.683	0.772

由表1可知:与其他方法相比,文中算法的优点在于检索的同时能检测出图像的显著性区域,筛去了背景干扰信息,使得图像特征的表达能力更突出,而fc8_im,SPoC等方法都是直接提取整张图像特征进行检索;实验得到的MAP值在CUB200-2011和Stanford Dogs数据集上分别为0.683与0.772,相较于无监督方法SPoC,CroW与有监督方法SCDA上表现最优.算法的不足之处在于提特征之前需要进行有监督微调,让网络学习出图像的语义特征,从而找出图像的显著性区域.实验最后提取的特征维度为528,该参数由特征提取层的通道数决定,虽然在检索精度上比其他方法高,但在特征维度上相较于SPoC与CroW方法的256维存在一定劣势,未来的工作将从网络结构与维度缩减上进行优化.

3　结束语

提出一种基于深层卷积特征图融合与显著性检测的图像检索方法，相较于主流算法，其优点在于检索的同时能检测出图像的显著性区域. 筛去了背景干扰信息，可有效提高检索的精确度. 在算法复杂度方面，较之前方法，仅增加特征图谱的矩阵线性相加计算，使得算法具有良好的计算性能. 通过有监督的分类标签将图像准确分类，网络模型中响应值较大的区域即为类别实例区域. 与直接提取整图特征相比，计算实例区域的特征相似度有助于提高检索精度与速度. 算法不足之处是没有实现端到端的训练出特征，而是对卷积特征重新进行编码，得到比较好的效果. 实验结果表明：文中算法在 CUB200-2011 和 Stanford Dogs 细粒度数据集上具有良好的性能与检索精度，优于现有大多数图像检索方法.

参考文献：

[1] LEW M S，SEBE N，DJERABA C，*et al*. Content-based multimedia information retrieval：State of the art and challenges[J]. ACM Trans Multimedia Comput Commun Appl，2006，2(1)：1-19. DOI：10. 1145/1126004. 1126005.

[2] LOWE D G. Object recognition from local scale-invariant features[C]// IEEE International Conference on Computer Vision. Kerkyra：IEEE Press，1999：1150. DOI：10. 1109/ICCV. 1999. 790410.

[3] OJALA T，PIETIKAINEN M，MAENPAA T. Gray scale and rotation invariant texture classification with local binary patterns[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2002，24(7)：971-987. DOI：10. 1109/TPAMI. 2002. 1017623.

[4] DALAL N，TRIGGS B. Histograms of oriented gradients for human detection[C]// IEEE Conference on Computer Vision and Pattern Recognition. San Diego：IEEE Press，2005：886-893. DOI：10. 1109/CVPR. 2005. 177.

[5] OLIVA A，TORRALBA A. Modeling the shape of the scene：A holistic representation of the spatial envelope[J]. Intl J of Computer Vision，2001，42(3)：145-175. DOI：10. 1023/A：1011139631724.

[6] LIN K，YANG H F，HSIAO J H，*et al*. Deep learning of binary hash codes for fast image retrieval[C]// IEEE Conference on Computer Vision and Pattern Recognition Workshops. Boston：IEEE Press，2015：27-35. DOI：10. 1109/CVPRW. 2015. 7301269.

[7] NIE Yiliang，DU Jixiang，FAN Wentao，*et al*. Visual categorization with bags of keypoints[C]// Workshop on Statistical Learning in Computer Vision Eccv. Liverpool：Springer，2004：1-22. DOI：10. 1007/978-3-319-63309-1_19.

[8] NG Y H，YANG F，DAVIS L S. Exploiting local features from deep networks for image retrieval[C]// Computer Vision and Pattern Recognition Workshops. Boston：IEEE Press，2015：53-61. DOI：10. 1109/cvprw. 2015. 7301272.

[9] LIU Ziwei，LUO Ping，QIU Shi，*et al*. Deepfashion：Powering robust clothes recognition and retrieval with rich annotations[C]// The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas：IEEE Press，2016：1096-1104. DOI：10. 1109/cvpr. 2016. 124.

[10] BABENKO A，LEMPITSKY V. Aggregating deep convolutional features for image retrieval[C]// The IEEE International Conference on Computer Vision (ICCV). Santiago：IEEE Press，2015：2380-7504. DOI：10. 1109/iccv. 2015. 150.

[11] 祝军，赵杰煜，董振宇. 融合显著信息的层次特征学习图像分类[J]. 计算机研究与发展，2014，51(9)：1919-1928. DOI：10. 7544/issn1000-1239. 2014. 20140138.

[12] WEI Xiushen，LUO Jianhao，WU Jianxin. Selective convolutional descriptor aggregation for fine-grained image retrieval[J]. IEEE Transactions on Image Processing，2017，26(6)：2868. DOI：10. 1109/tip. 2017. 2688133.

[13] JIA Y，SHELHAMER E，DONAHUE J，*et al*. Caffe：Convolutional architecture for fast feature embedding[C]// ACM International Conference on Multimedia. Florida：ACM，2014：675-678.

[14] KALANTIDIS Y，MELLINA C，OSINDERO S. Cross-dimensional weighting for aggregated deep convolutional features[C]// European Conference on Computer Vision. Amsterdam：Springer，2016：685-701. DOI：10. 1007/978-3-319-46604-0_48.

（**责任编辑**：陈志贤　　**英文审校**：吴逢铁）