软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2019,30(11):3567−3577 [doi: 10.13328/j.cnki.jos.005570]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

包含跨域建模和深度融合网络的手绘草图检索*

于 邓[1], 刘玉杰[1], 邢敏敏[1], 李宗民[1], 李 华[2]

[1](中国石油大学(华东) 计算机通信工程学院,山东 青岛 266580)

[2](中国科学院 计算计算技术研究所,北京 100190)

通讯作者: 刘玉杰, E-mail: liuyujie@upc.edu.cn

摘 要: 在手绘草图检索(sketch-based image retrieval,简称 SBIR)领域,引入一种手绘草图的新型检索模型.手绘草图与自然图片之间存在巨大的差异性,这是因为,与自然图片相比,手绘草图展现出高度抽象的视觉表达,用现有的方法对手绘草图进行特征提取,其产生的特征描述子对于手绘草图的内容无法进行有效地拟合;对于相同的物体,不同的人群用手绘草图描述方式和表达也存在巨大的差距,这就使得手绘草图-自然图片的匹配更加困难;同时,将手绘草图与自然图片映射到相同视觉域的工作,也是一项具有困难的任务.所以,手绘草图检索技术是公认的比较有挑战性的任务.提出一种将手绘草图与自然图片在多个层次上映射到同一视觉域的策略来解决跨域的问题.同时,引入多层深度融合卷积神经网络(multi-layer deep fusion convolutional neural network)的框架来训练并获得手绘草图和自然彩色图片的多层特征表达.在 Flickr15k 图像数据库进行检索实验,实验结果显示,多层深度融合卷积网络学习到的特征的检索精度超过了现有的手工特征以及由自然图片或者手绘草图训练出来的卷积神经网络(convolutional neural network,简称 CNN)的特征.

关键词: 手绘草图检索;跨域建模;多层深度融合卷积神经网络;特征融合;深度学习

中图法分类号: TP391

中文引用格式: 于邓,刘玉杰,邢敏敏,李宗民,李华.包含跨域建模和深度融合网络的手绘草图检索.软件学报,2019,30(11):3567−3577. http://www.jos.org.cn/1000-9825/5570.htm

英文引用格式: Yu D, Liu YJ, Xing MM, Li ZM, Li H. Sketch-based image retrieval using cross-domain modeling and deep fusion network. Ruan Jian Xue Bao/Journal of Software, 2019,30(11):3567−3577 (in Chinese). http://www.jos.org.cn/1000-9825/5570.htm

Sketch-based Image Retrieval Using Cross-domain Modeling and Deep Fusion Network

YU Deng[1], LIU Yu-Jie[1], XING Min-Min[1], LI Zong-Min[1], LI Hua[2]

[1](College of Computer and Communication Engineering, China University of Petroleum (East China), Qingdao 266580, China)

[2](Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: The purpose of this paper is to introduce a new approach for the free-hand sketch representation in the sketch-based image retrieval (SBIR), where the sketches are treated as the queries to search for the natural photos in the natural image dataset. This task is known as an extremely challenging work for 3 main reasons: (1) Sketches show a highly abstract visual appearance versus natural photos, fewer context can be extracted as descriptors using the existing methods. (2) For the same object, different people provide widely different sketches, making sketch-photo matching harder. (3) Mapping the sketches and photos into a common domain is also a challenging task. In this study, the cross-domain question is addressed using a strategy of mapping sketches and natural photos in multiple layers. For the first time, a multi-layer deep CNN framework is introduced to train the multi-layer representation of free hand sketches and natural photos.

* 基金项目: 国家自然科学基金(61379106, 61379082, 61227802); 山东省自然科学基金(ZR2013FM036, ZR2015FM011)

Foundation item: National Natural Science Foundation of China (61379106, 61379082, 61227802); Natural Science Foundation of Shandong Province (ZR2013FM036, ZR2015FM011)

收稿时间: 2017-06-01; 修改时间: 2017-09-18, 2017-10-25, 2017-11-27, 2017-11-30, 2018-01-08; 采用时间: 2018-01-15

Flickr15k dataset is used as the benchmark for the retrieval and it is shown that the learned representation significantly outperforms both hand-crafted features as well as deep features trained by sketches or photos.

Key words: sketch-based image retrieval (SBIR); crossing-domain modeling; multi-layer deep fusion convolutional neural network; feature fusion; deep learning

近年来,随着互联网技术和移动终端设备的产品更新,触屏技术快速发展,多媒体信息(尤其是图像信息)急剧增加.伴随着触屏技术的不断进步,平板电脑、掌上设备和超宽屏幕手机制造产业和使用正蓬勃发展,可触屏设备逐渐成为人们生活中不可或缺的一部分,用户可以用手绘草图的方式在移动终端快捷方便地描绘出所需物体的外观、特点及其轮廓.而手绘图像检索作为针对可触屏设备的新兴的一个科研领域,具有非常大的潜力和市场价值.如何有效地跨越手绘草图与自然图片之间的语义鸿沟进行检索,是众多研究人员面临的难点.与此同时,当面临高度抽象的手绘图像的问题时,特征的选取、高效的检索算法,成为当前的一个研究热点.手绘草图检索(sketch-based image retrieval,简称 SBIR)是指通过手绘草图,在海量自然图片数据库中找到用户想要的自然图片的过程,如图 1 所示.

Fig.1 Sketch and sketch-based image retrieval

图 1 手绘草图及手绘草图检索

图 1 展示的是使用手绘草图搜索自然图像数据库中的彩色自然图片的结果.自然图片具有丰富的纹理信息和彩色信息,同时,自然图片对于同一物体的刻画基本上是一致的,没有剧烈的形变或者是抽象概括;而与自然彩色图像相比,手绘草图拥有高度的抽象性和概括性,由简单的线条和点组成.所以手绘草图与自然彩色图片之间巨大的视觉差异和语义差距,导致了特征选择的困难,进而导致较低的检索准确率.

针对手绘草图与自然彩色图片之间的视觉差异,本文提出了“分层”跨域匹配的框架;同时,针对手绘草图与自然图片之间的语义差距,本文设计了多层深度融合卷积神经网络(multi-layer deep fusion convolutional neural network)来学习手绘草图和自然图片的多层跨域特征.该方法还探索了对于多层深度特征的融合技术的研究.为了实现更高精度的和更高效的检索,本文方法主要研究的是手绘草图与自然彩色图片之间的跨域建模(cross-domain modeling)和与模型相适应的深度特征学习.对本文的算法框架的简介如下.

在手绘草图检索领域中提出了独特的基于“层次”属性的检索模型.本文检索模型旨在跨越手绘草图和自然图像的视觉域的差距;对于手绘草图和自然图片的不同层次的特征性质进行了建模;与此同时,提出了与多层模型相适应的多层深度融合卷积神经网络,并且展示了将多层提取的特征融合成最终特征表示的过程.

本文第 1 节介绍相关工作.第 2 节阐述本文的核心理论.第 3 节对实验细节和实验结果进行展示.第 4 节总结本文方法,并对未来的工作进行初步的探讨.

1 相关工作

目前对于手绘草图语义的定义,有些工作给出其独到的定义方法.Eitz 等人[1]在手绘草图的线条长度方面,研究了人类使用手绘草图来描述物体的过程,并且公布了一个拥有“时序”属性的 TU-Berlin 手绘草图数据库.Fu

等人[2]提出了一项用以挖掘手绘草图中的手绘时序属性的工作,对手绘草图中的线条设计了先后顺序的规则,以获得一幅手绘图像中的每一条笔画(stroke)的时序,再将一幅手绘图像中的线条以动画的形式一条一条地绘制出来,这项工作主要用来模拟动画绘制的过程.Yu 等人[3]提及到从手绘草图中获取手绘草图的时序信息,并对这些时序信息进行分离,将它们输入到多通道多尺度的卷积神经网络(convolutional neural network,简称 CNN)框架中,即 Sketch-A-Net.Sangkloy 等人[4]提出了一个数据量更大的手绘草图数据库,同时,在数据库中为每一幅手绘草图添加了 Sketchablity 属性,用来表明该手绘草图在绘制时的难易程度.

在手绘草图的图像检索(SBIR)中,对于将彩色自然图片与手绘草图映射到同一视觉域的跨域方法,主要是将自然图片转化成类似于手绘图片的边缘图,以保证两者在高层的视觉上的可比性.主流的边缘检测方法分为 2 类:第 1 类是基于显著性检测的方法,包括 Canny Detector 边缘检测方法和 Robert Detector 边缘检测方法等;第 2 类是基于边缘感应的方法,例如 Sketch-token[5]、Berkeley detection[6]等方法.这两类边缘检测方法广泛地被研究人员所认可和使用,但是它们处理后的边缘图像中仍然不可避免地存在着大量的背景噪声,所以,由这些方法得到的自然图片的边缘图(edge map)的背景并不像手绘草图图像一样背景纯净.

本文的深度多层卷积神经网络主要是受到了以下工作的启发:Yu 等人[7]在细粒度手绘草图检索中,将三元损失(triple loss)应用到 3 层的卷积神经网络中;Su 等人[8]提出了一种新型的组合和融合多层深度网络学习特征的方法.

现有的手绘草图和自然图片的特征描述子,主要分为手工特征和深度学习特征两类.

- 第 1 类:手工特征.如 Sift[9]、Shape Context[10]等局部特征和近期专门为手绘草图图片或者边缘图设计的特征描述子——Cao[11]提出的 binary HOG 和 Gradient Field HOG(GF-HOG)[12]以及 shape words[13]等手工特征.
- 第 2 类:深度学习特征.由于卷积神经网络(CNN)对于数据的强大拟合能力以及对于特征提取的深度,随着对于网络的深度的增加,深度学习框架 AlexNet[14]、LeNet[15]、VGG[16]以及它们的变体 VGG-16、VGG-19 等网络的出现,使得计算机视觉领域图像特征的提取方式发生了巨大的变化.

本文将在第 4 节实验部分对于这些网络的特征刻画能力进行实验比较和实验验证.

近年来,针对手绘草图和自然图像跨域建模的方法,大多采用了比较新型的深度学习框架及模型.如 Bui 等人[17]提出的 sketch-photo model、Qi 等人[18]提出的 Siamese CNN 模型以及 Seddati 等人[19]提出的 Quardruplet network 等.现有的跨域深度模型致力于挖掘更深的网络特征,提取更强的特征描述子.通过对现有工作进行实验对比,我们发现,手绘草图检索领域的研究,从最初的浅层手工特征的研究到目前的深度学习框架的加入,尽管在检索的精度上有了较大的提升,但对于手绘草图本身的探究并未达到令人满意的程度.所以本文从手绘草图本身的绘制机制出发,进行了更深一步的探索并发现手绘草图的层次规律,最终实现了对检索精度的进一步提升.

2 研究框架

为了计算同一视觉域中手绘草图与自然图片之间的相似度,采用多层深度融合卷积神经网络来捕捉和提取手绘草图与自然图片的特征.图 2 展示了多层深度融合卷积神经网络的整个流程.

如图 2 所示.

- 首先,按照分层规则(本文第 2.1 节、第 2.2 节详述),同时对手绘草图和自然图片进行分层操作,即 3 层视觉表达,使得它们之间的视觉表达在同一层次上达到统一.
- 然后,基于多层视觉表达,将手绘草图与彩色图片的边缘图作为训练图片,分别输入多层深度融合卷积神经网络中,进行分层训练.多层深度融合卷积神经网络的框架中,手绘草图的 3 层视觉表达的特征由 3 个 CNN 网络进行训练,直至收敛.这里,3 个 CNN 网络是相同的 Image-very-deep-19 网络,简称 VGG-19[16],网络训练是基于 MatConvNet 的环境.通过将 3 个网络训练收敛后,提取网络的全连接层 FC-7 的 512 维特征向量作为学习特征.然后,把由 3 层视觉表达学习到的特征向量融合成 1 个特征,作

为手绘草图的特征.通过这种方式,学习到的特征就具有了比原来单层训练学习的特征更加丰富的语义、层次、空间信息.对于自然图片采用相同的操作,实现了更好的跨域检索效果.

- 最后,本文利用手绘草图的特征描述子来检索自然图片库中的彩色自然图片.

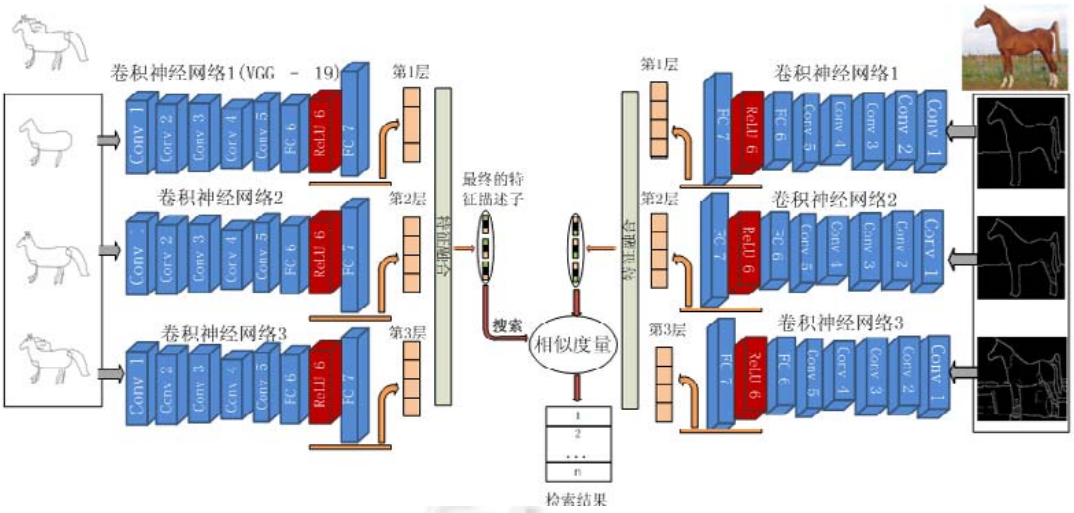

Fig.2 Cross-domain retrieval and deep fusion model

图 2 跨域检索与深度融合模型

将特征表达扩展到多层表达,是为了从手绘草图和自然图片的边缘图中获取更多抽象性、细节性的语义信息.在特征相似度计算方面,能够更多地考虑到语义含义.同时,对于不同层次特征的融合操作,是为了得到手绘草图和彩色自然图片唯一的特征表达.

2.1 手绘草图语义的分层定义

本文将手绘草图的“时序”特质作为分析的语义特征.时序性是手绘草图所固有的属性.通过调研文献[20]并受实验测试结果的启发,本文通过实验发现手绘草图的分层属性,并提出对手绘草图的分层方法.我们发现,人们在创作手绘草图时,通常通过第 1 层视觉表达,人们就可以大致地绘制出该物体的类别;接下来,人们往往在第 2 层视觉表达上添加一些细节信息,来具体标识出描绘的物体;最后在第 3 层的视觉表达上,人们用更加具体和细节的纹理来绘制出在大脑中所认知的物体,如图 3 所示.

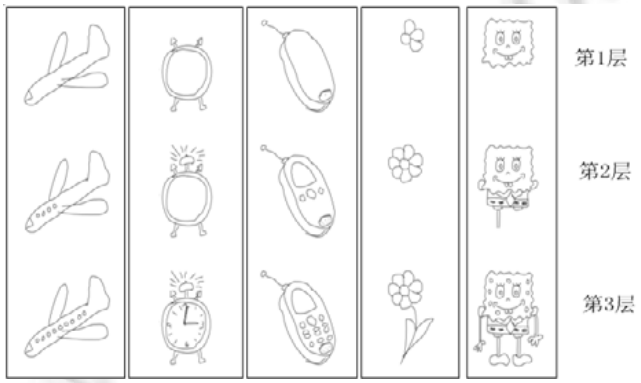

Fig.3 Layer-based semantic definition of the sketch

图 3 手绘草图的多层语义定义

与此同时,人们在识别手绘草图时,也是一个由粗到细、由外至内、由轮廓至细节的过程.所以,将手绘草图进行分层操作符合神经网络学习的特质,即模拟人类的思维和感官感受方式来进行学习.对于这种层次性绘制和理解的手绘草图,本文相应地使用层次视觉表达来描述这种层次性的语义.然后,使用卷积神经网络提取出不同层次的语义特征,并加以融合.

图 3 展示的是不同的手绘草图的多层视觉表达.手绘草图的分层方法是将手绘草图中的时序特质转化为层次语义属性,即把一幅手绘草图转换成一组多层的视觉表达.在本文中,手绘草图具体的分层操作是通过获取手绘草图中的线条的时序信息,再按照先后顺序对手绘草图的笔画线条进行分类,即分成前期笔画、中期笔画、最终笔画这 3 个视觉层次的集合.最终,笔画的组合就是图 3 中 3 个层次的手绘草图.

图 3 中的分层算法步骤如下.

(1) 对于一幅手绘草图,本文提取草图中笔画的总数目 N 和每一笔笔画的时序$\{T_1,T_2,\ldots,T_n\}$,即一幅手绘草图中众多笔画的先后顺序,其中,每一笔笔画的长度是已知的.

(2) 设置手绘草图的层次,本文设置为 3 层.因为在实验测试时,当手绘草图的多层视觉表达少于 3 层时,无法显示出分层网络的优势;然而,当多层视觉表达多于 3 层时,对于不同人的绘制习惯就失去了普适性和统计性.在 3 层的范围内,人们的绘制习惯大致达到统一,虽然在细节上微微有些出入,但却极大地提高了对图像的描述力.所以,本文选择 3 层视觉表达,记为 v=3,如图 3 所示.

(3) 每一层的笔画数目由笔画总数目除以手绘草图的层次数,并向上取整所得,即 $m=\lceil N/v \rceil$.此时,

- 前期笔画集合(第 1 层视觉表达)为$\{T_1,T_2,\ldots,T_m\}$,因此,第 1 层视觉表达绘制出来为$\{T_1,T_2,\ldots,T_m\}$;
- 中期笔画集合(第 2 层视觉表达)为$\{T_{m+1},T_{m+2},\ldots,T_{2m}\}$,因此,第 2 层视觉表达绘制出来为$\{T_1,T_2,\ldots,T_{2m}\}$;
- 最终的笔画集合(第 3 层视觉表达)为原始的手绘草图$\{T_{n-2m+1},T_{n-2m+2},\ldots,T_n\}$,因此,第 3 层视觉表达绘制出来为$\{T_1,T_2,\ldots,T_n\}$,即包含所有笔画的手绘草图.

2.2 跨域建模(cross-domain modeling)

基于手绘草图的多层视觉表达,对自然彩色图片采用相同的分层策略,以达到更好的跨域建模(cross domain modeling)效果.首先,对传统的边缘检测方法进行分析,使用如图 4 所示的 Canny Detection 边缘检测和 Robert Detection 边缘检测方法以及 Sketch-Token 边缘检测方法和 Berkeley 边缘检测方法.

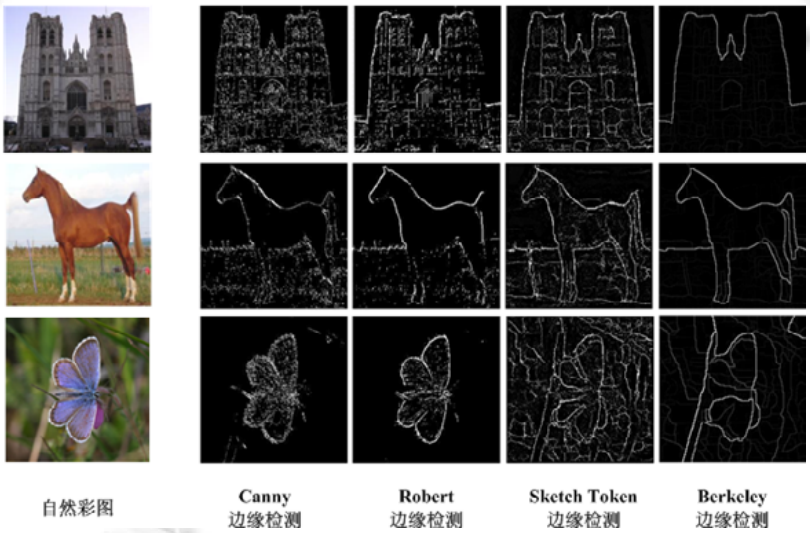

Fig.4 Different methods of edge detection on natural photos

图 4 不同方法对于自然图片的边缘提取

从图 4 中这些边缘提取方法的表现中可以看出,基于显著性检测的 Canny 和 Robert 检测方法对于自然图片中的背景噪声具有较强的抗噪能力.然而这两种方法对于目标的边缘感应较差,产生了较差的类手绘边缘图.而与前两种方法相比,Sketch-Token 和 Berkeley 边缘检测方法对于自然图片中的物体具有比较令人满意的边缘感应能力.然而,在最后提取的边缘图片中却保留了大量的背景噪音.

本文在得到手绘草图的 3 个层次的视觉表达后,也对自然图像进行了相似的实验探索.由于手绘草图 3 个层次的表达顺序为抽象-轮廓-纹理,所以针对自然图像也按照相同的规律和策略进行处理:首先,对于自然图像进行边缘提取,生成边缘图.实验中发现,通过 Berkeley 边缘检测算法提取的边缘图对于自然图像的轮廓具有不同的感应能力:对于外围的轮廓,该方法有着较强的感应能力;而对于内部的纹理的感应能力较弱.所以,本文通过调节 Berkeley 边缘感应的阈值,并通过聚类方法进行约束,最后也生成了自然图像的 3 层视觉表达,即最外围边缘(边缘感应最强)-大体结构(边缘感应中等)-细节纹理(边缘感应最弱).这样,本文的手绘草图 3 个层次的视觉表达与自然图像 3 个视觉表达的对应与统一,如图 5 所示.

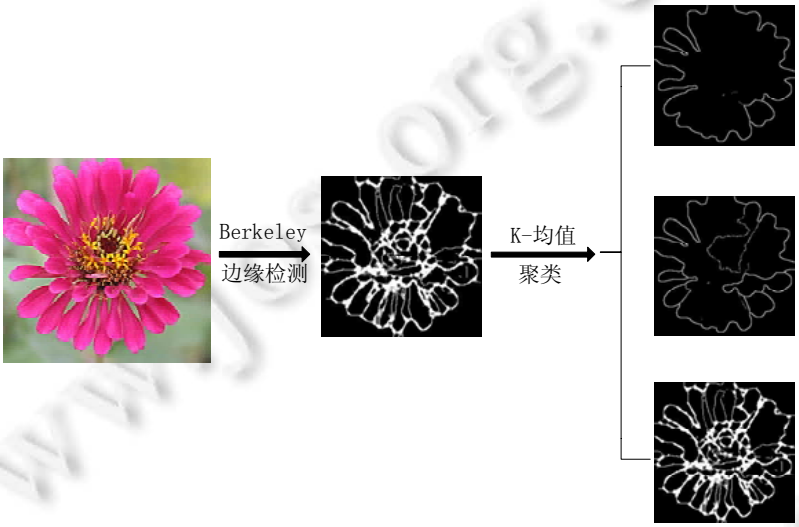

Fig.5 Layer-based representations of edge map from the natural photo

图 5 自然图片边缘图的多层表达

最终,根据文献[21,22]中 Canny 检测、Robert 检测、Sketch-Token 与 Berkeley 边缘检测方法的对比效果以及本文的实现效果,如图 4 所示,本文提出了一种新方法来解决跨域建模阶段背景噪音和物体边缘看似互相矛盾的问题.基于 Berkeley 边缘检测方法的理论:边缘图中的每一个像素的灰度值对应的是该点被分类成边缘的概率,即边缘图中的一条线灰度值越高,这条线就越容易被分类成边缘,进而这条线的亮度也就越高.进一步来说,由 Berkeley 边缘检测方法产生的边缘图阐述了一个边缘图的上每个点的概率分布.因此,针对 Berkeley 边缘检测的产生的边缘图,根据边缘图上每个点的灰度值,对于自然图像产生多层视觉表达.和手绘草图的层次策略相同,本文的方法对于一张彩色图片产生 3 层视觉表达,对于自然图片分层跨域方法的阐述如下.

(1) 采用 *K*-均值(*K*-means)方法对一张自然图片的边缘图上的每个点的灰度值进行 3 聚类中心的聚类.

(2) 第 1 层视觉表达被定义为聚类中心中最小的聚类中心所在的簇;接下来,第 2 层视觉表达表示的是中层灰度值的聚类簇;最后,拥有最高的灰度值的聚类簇被定义为第 3 层.方法如图 5 所示.

通过采用与处理手绘草图相同的分层策略,获得了自然图片的 3 层视觉表达,如图 5 所示.到目前为止,可以认为手绘草图与彩色自然图片是出于同一视觉域.之后,多层视觉表达就可以输入到多层深度融合卷积神经网络来提取特征表示.

多层深度融合卷积神经网络中的多层框架为手绘草图和彩色图片的跨域建模提供了新的跨域思路,并且

提供了更多的细节信息和空间信息以及多层的抽象的语义信息.在特征学习阶段,这些细节信息不仅丰富了本文特征的描述力,而且保证了跨域检索的稳定性.

2.3 深度特征融合

在多层深度融合卷积神经网络的框架中,单一的卷积神经网络(CNN)采用的是在 MatConvNet 环境所训练的“Image-very-deep-19”网络结构.在训练完 Flickr15k 数据库中手绘草图和彩色自然图片的多层视觉表达之后,提取每一层网络的全连接第 7 层 Convfc-7(fully connected layer 7)的特征作为每一层视觉表达的特征表示.

如上所述,本文的关注点是获取手绘草图和彩色自然图片基于层次的特征表示.这个层次特征是可训练的,能够产生语义和区分力的特征表示;同时,这种框架是保证可被高效实现的.基于层次的特征表示从跨域建模阶段的多层视觉表达开始,由深度多层融合卷积神经网络产生.但是如何融合多层特征,是手绘草图检索(SBIR)阶段的关键一步,因为在最后的相似度计算过程中,需要手绘草图和自然图片的唯一的特征表示.

特征融合的方式有均值融合、权重融合、串联融合.

(1) 一种简单的方式就是对不同层次的特征进行均值化,把每一层的特征的权重视为同等重要,这种方法称为均值融合.

$$feature_{fusion}=\sum_{i=1}^{N}\frac{1}{N}feature_i \tag{1}$$

其中,$feature_{fusion}$ 表示融合 3 层视觉表达的最终融合特征,$feature_i$ 表示第 i-th 层的特征,$\frac{1}{N}$ 表示对不同层次的特征进行均值化的操作.

(2) 第 2 种方式是使用类似于贝叶斯特征融合的方式[20],将不同层次的特征按照其不同的概率数据分布进行融合,称为权重融合,如公式(2)和公式(3)所示.

$$feature_{fusion}=\sum_{i=1}^{3}P_i\times feature_i \tag{2}$$

$$feature_{fusion}=\sum_{i=1}^{3}P_i\times PCA\_feature_i \tag{3}$$

在公式(2)中,$feature_{fusion}$ 表示融合 3 层视觉表达的最终融合特征;$feature_i$ 表示第 i-th 层的特征;P_i 是一个依赖于不同层次的检索表现的值,即在第 i 层的检索精度越高,P_i 的值越高.在本文的实验中,P_i 的值定义为 0.4101、0.4582、0.5298.为了使最后的特征表达更加紧凑,对融合特征进行了 L2 规范化,并使用 PCA 算法对融合特征的维度进行降维,如公式(2)所示.将特征的维度减小到既能保证特征计算的高效性,又能减少特征向量中的零元素.因此 PCA 的系数需要谨慎选择,以保证融合特征的区分力和紧凑型之间的平衡.本文将融合特征的维度减小到 128 维.

在公式(3)中,$PCA\_feature_i$ 表示的是多层融合卷积神经网络中的第 i 层规范后的特征.

(3) 第 3 种特征融合方式就是把所有层次的特征按序排列,按照一定的顺序串联起来,这种方法称为串联融合.多层深度融合卷积神经网络框架中采用的就是串联融合方式,将第 1 层直至第 3 层所学习到的深度特征进行有序的串联,得到最终的特征向量.

以上 3 种特征融合方式的有效性验证结果在第 4 节展示.

3 实 验

3.1 数据库

本文使用 Flickr15k 数据集.该数据集包含 60 类,总共 14 460 张彩色自然图片;并且,该数据集还包含由非专业手绘草图人员绘制的 33 类,总共 329 张手绘草图.

3.2 数据扩增

对于训练深度卷积神经网络(deep CNN),关键的一步是提供充足的训练数据.对于训练数据较少的数据库,我们需要进行数据扩增(data augment)操作.手工特征或者 ImageNet[23]所训练的数据库基本不需要进行数据扩

增.对于在手绘草图数据集上训练的对比深度网络,本文使用的手绘草图训练库是 TU-Berlin[1].这个手绘草图数据集是目前最大的和最通用的数据集,包含了 20 000 张手绘草图.但是,数据量不足以驱动深度卷积神经网络,容易对数据产生过拟合现象.所以,采用与文献[20]相同的数据扩增方式,即通过镜像、(水平、垂直)平移、或者旋转,将数据扩增到 32×32×11×2 倍.

在本文的实验中,训练的实验数据集有 ImageNet、TU-Berlin 手绘数据集和 Flickr15k 数据集.实验中,本文将实验数据集大体按照 7:3 的比例进行分割.所有模型在训练集完成训练并收敛之后,再在测试集上进行测试,从而获得每个模型的检索精度.

3.3 评价标准

本文使用的评价标准是 mAP、平均准确率(mean average precision).本文中对于 SBIR 中的 mAP 定义如下.

$$mAP=\frac{1}{N}\sum_{q\in N_q}\sum_{r\in R}\frac{RT_i/IT_i}{N_R} \tag{4}$$

其中,q 表示在检索序列里面的检索手绘草图,N_q 是检索序列中的手绘草图的总数,RT_i 是指检索出来的正样本的索引号,IT_i 表示的是在检索序列中正样本的排序索引,N_R 表示正样本的总数.

3.4 实验结果

本文方法与现有方法的实验对比结果见表 1.为了充分验证本文网络框架的有效性,对于现有的深度学习网络,设置了 3 种类型的训练方式来使现有的 CNN 网络的学习能力达到最大化.

(1) 第 1 类是单独在自然彩色图片 ImageNet 上训练的深度卷积神经网络特征,其中,使用 LeNet 的数据集 MINIST 是因为手写数字数据集和我们的手绘草图在视觉上有一定的相似性.

(2) 第 2 类是单独在手绘草图数据库 TU-Berlin 上训练的深度卷积神经网络,用所学习的特征进行检索.

(3) 第 3 类是使用训练好的深度卷积网络,在 Flickr15k 数据集上微调,然后使用再次收敛的网络所生成的特征进行检索.

Table 1 Retrieval performance of different methods

表 1 不同方法的检索表现

名称	描述	训练数据集	检索数据集	mAP
VGG-16	深度网络(自然图片数据集上训练)	ImageNet	Flickr15k	0.179 3
VGG-19		ImageNet		0.226 1
VGG		ImageNet		0.288 2
Alexnet		ImageNet		0.293 5
Sketch-A-Net(single)	深度网络(手绘草图数据集上训练)	TU-Berlin	Flickr15k	0.153 8
AlexNet-Sketch		TU-berlin		0261 6
VGG-Sketch		TU-berlin		0.265 8
Siamese CNN[18]	深度网络(数据集 Flickr15k 上训练)	Flickr15k	Flickr15k	0.195 4
Sketch-A-Net(single)		Flickr15k		0.237 4
Sketch-Photo Mode[1,17]		Flickr15k		0.361 7
AlexNet		Flickr15k		0.381 1
VGG		Flickr15k		0.429 3
Quardruplet Network[19]		Flickr15k		0.433 0
VGG-16		Flickr15k		0.458 2
VGG-19		Flickr15k		0.501 7
多层卷积神经网络	本文框网络框架	Flickr15k	Flickr15k	0.557 4

表 1 中,Sketch-A-Net(single)网络结构是参考引用文献[1]中的网络参数实现出来的单一尺度的深度网络,其中,本文使用单一图片尺度为 256×256.

从表 1 可以得到的结论如下:多层融合卷积神经网络的跨域建模的视觉表达和多层深度融合卷积神经网络的特征表达模型对于手绘草图检索(SBIR)领域中手绘草图和自然图片具有较强的拟合能力;本文模型所生成的特征向量的检索准确率远超基于自然图片训练的模型,如主流的 Alexnet 网络和 VGG 框架等,同时,对于基于手绘草图数据集训练的深度卷积神经网络的性能也有较大的提升.另外,在第 3 类实验中,本文通过与同领域

基于手绘草图跨域检索方法进行比较,如 Sketch-Photo Model[17]、Siamese CNN[18]与 Quardruplet Network[19]等,实验结果表明,本文的多层卷积神经网络比同类方法的检索准确率大致提升了 12%~20%.从表 1 的实验结果可以看出本文方法相对于现有方法在网络性能上的提升.

为了揭示手绘草图和自然图片不同层次之间的内在联系,对不同视觉层次的检索结果进行分析,实验的对比结果见表 2.

Table 2 Retrieval results on different layers

表 2 不同层次的检索效果

层次	层次 1	层次 2	层次 3
mAP	0.410 1	0.458 2	0.529 8

在表 2 中,实验中使用的检索特征是 VGG-19 网络 FC7 的特征向量,在拥有最少的空间、语义信息的第 1 层,获得了相对较低的检索精度;而在具有最细致语义信息和空间信息的第 3 层,获得了 52.9%的检索精度,仅仅比本文的多层融合卷积神经网络少 3%的精度.同时,这也是本文的特征融合策略有效性的证明.

为了探索不同的特征融合策略对于手绘草图检索(SBIR)的影响,我们对不同的融合策略进行了实验对比,实验结果见表 3.

Table 3 Impacts of different feature fusion strategies on retrieval results

表 3 特征融合策略对于检索结果的影响

特征融合策略	均值融合	串联融合	权重融合
mAP	0.360 4	0.557 4	0.438 1

从表 3 可知,不同的特征融合策略对于提升或者拉低本文最终的融合特征向量的辨别力产生了非常大的影响,尤其是一个较差的融合策略能够将本文的多层特征的检索精度从 55.74%降低到 36.04%.所以,一个合适的特征融合策略对于多层融合卷积神经网络能力的提升具有非常重要的意义.

最后,多层融合卷积神经网络对 Flickr15k 数据集的检索结果如图 6 所示.

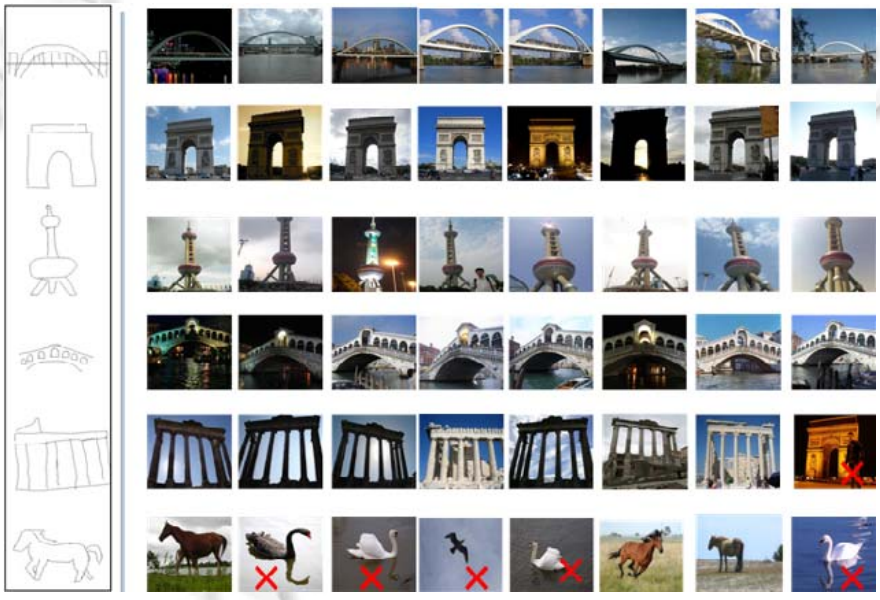

Fig.6 Retrieval performance of multi-layer fusion CNN on Flickr15k

图 6 多层融合卷积神经网络在 Flickr15k 上的检索表现

如图6所示,“×”表示错误的检索结果.其中最后一行展示的是多层融合卷积神经网络最差的检索表现,主要是因为 Flickr15k 数据集中这类图片的训练数据远远不足的缘故.

4 总结与未来工作

在本文中,“多层”的概念一直是模型中至关重要的主题,而在实验中发现,本文的多层视觉表达方法对于手绘草图或者自然图片具有数据扩增的潜力.与此同时,本文的理论中仍存在着一个猜想:对于手绘草图和自然图片的边缘图,是否它们的抽象层次越高,它们之间的相似度就越大?在实验中,本文的3层视觉表达的确显示了这样的趋势.

关于设置多少层才能最佳地利用好手绘草图和彩色自然图片中的层次信息的问题,本文的实验仅仅证明了多层表达能够进一步提高检索表现,所以仍然需要进行更多的实验来解答这个问题.正如本文的实验部分所示,抽象的程度越高,在自然图片的边缘图中的背景噪音就越少.所以我们猜想:在某一个合适的抽象层次,能够获得一个最好的视觉表达和特征表示.

本文的特征采用的特征融合策略都比较简单,与此同时,特征融合阶段也是提高多层卷积神经网络的最终特征的辨别力的关键环节.所以,一种更好的特征融合方法将会为手绘草图检索(SBIR)的检索精度带来巨大的提升.

References:

[1] 1Eitz M, Hays J, Alexa M. How do humans sketch objects? ACM Trans. on Graph, 2012,31(4):44:1−44:10. [doi: 10.1145/2185520.2185540]

[2] Fu H, Zhou S, Liu L, *et al*. Animated construction of line drawings. ACM Trans. on Graphics, 2011,30(6):1−10. [doi: 10.1145/2070781.2024167]

[3] Yu Q, Yang Y, Song YZ, *et al*. Sketch-a-Net that beats humans. arXiv preprint arXiv:1501.07873, 2015. [doi: 10.5244/C.29.7]

[4] Sangkloy P, Burnell N, Ham C, *et al*. The sketchy database: Learning to retrieve badly drawn bunnies. ACM Trans. on Graphics (TOG), 2016,35(4):Article No.119. [doi: 10.1145/2897824.2925954]

[5] Lim JJ, Zitnick CL, Dollár P. Sketch tokens: A learned mid-level representation for contour and object detection. In: Proc. of the 2013 IEEE Conf. on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2013. 3158−3165. [doi: 10.1109/CVPR.2013.406]

[6] Arbelaez P, Maire M, Fowlkes C, *et al*. Contour detection and hierarchical image segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2011,33(5):898−916. [doi: 10.1109/TPAMI.2010.161]

[7] Yu Q, Liu F, Song YZ, *et al*. Sketch me that shoe. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2016. 799−807. [doi: 10.1109/CVPR.2016.93]

[8] Su H, Maji S, Kalogerakis E, *et al*. Multi-View convolutional neural networks for 3D shape recognition. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Washington: IEEE Computer Society, 2015. 945−953. [doi: 10.1109/ICCV.2015.114]

[9] Lowe DG. Distinctive image features from scale-invariant keypoints. Int'l Journal of Computer Vision, 2004,60(2):91−110. [doi: 10.1023/B:VISI.0000029664.99615.94]

[10] Mori G, Belongie S, Malik J. Efficient shape matching using shape contexts. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005,27(11):1832−1837. [doi: 10.1109/TPAMI.2005.220]

[11] 11Cao Y, Wang C, Zhang L, *et al*. Edgel index for large-scale sketch-based image search. In: Proc. of the 2011 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Washington: IEEE Computer Society, 2011. 761−768. [doi: 10.1109/CVPR.2011.5995460]

[12] 12Hu R, Collomosse J. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. Computer Vision and Image Understanding, 2013,117(7):790−806. [doi: 10.1016/j.cviu.2013.02.005]

[13] 13−Xiao C, Wang C, Zhang L, *et al*. Sketch-Based image retrieval via shape words. In: Proc. of the 5th ACM Int'l Conf. on Multimedia Retrieval. New York: ACM Press, 2015. 571−574. [doi: 10.1145/2671188.2749360]

[14] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Proc. of the 2012 Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2012. 1097−1105. [doi: 10.1145/3065386]

[15] Le Cun Y, Bottou L, Bengio Y, *et al*. Gradient-Based learning applied to document recognition. Proc. of the IEEE, 1998,86(11): 2278−2324. [doi: 10.1109/5.726791]

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

[17] Bui T, Ribeiro L, Ponti M, *et al*. Generalisation and sharing in triplet convnets for sketch based visual search. arXiv preprint arXiv: 1611.05301, 2016.

[18] Qi Y, Song YZ, Zhang H, *et al*. Sketch-Based image retrieval via Siamese convolutional neural network. In: Proc. of the 2016 IEEE Int'l Conf. on Image Processing (ICIP). Los Alamitos: IEEE Computer Society Press, 2016. 2460−2464. [doi: 10.1109/ICIP.2016.7532801]

[19] Seddati O, Dupont S, Mahmoudi S. Quadruplet networks for sketch-based image retrieval. In: Proc. of the 2017 ACM Int'l Conf. on Multimedia Retrieval. New York: ACM Press, 2017. 184−191. [doi: 10.1145/3078971.3078985]

[20] Yu Q, Yang Y, Song YZ, *et al*. Sketch-a-Net that beats humans. arXiv preprint arXiv:1501.07873, 2015.

[21] Feng GH, Sun ZX, Viard-Gaudin C. Stroke fragmentation using geometry features and hidden Markov model. Ruan Jian Xue Bao/Journal of Software, 2009,20(1):1−10 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3383.htm [doi: 10.3724/SP.J.1001.2009.03383]

[22] Liu YJ, Pang YP, Lu ZQ, *et al*. Sketch based image retrieval based on chamfer distance transform and bag of mid maps descriptor. Journal of Computer-Aided Design & Computer Graphics, 2016,28(12):2168−2174 (in Chinese with English abstract). [doi: 10.3969/j.issn.1003-9775.2016.12.017]

[23] Deng J, Dong W, Socher R, *et al*. Imagenet: A large-scale hierarchical image database. In: Proc. of the 2009 IEEE Conf. on Computer Vision and Pattern Recognition. Washington: IEEE Computer Society, 2009. 248−255. [doi: 10.1109/CVPR.2009.5206848]

附中文参考文献:

[21] 冯桂焕,孙正兴,Viard-Gaudin C.使用几何特征与隐 Markov 模型的手绘笔画图元分解.软件学报,2009,20(1):1−10. http://www.jos.org.cn/1000-9825/3383.htm [doi: 10.3724/SP.J.1001.2009.03383]

[22] 刘玉杰,庞芸萍,路子奇,等.结合距离变换和隐层图词包的手绘图像检索方法.计算机辅助设计与图形学学报,2016,28(12): 2168−2174. [doi: 10.3969/j.issn.1003-9775.2016.12.017]

于邓(1992−),男,山东潍坊人,硕士,主要研究领域为计算机图形学,模式识别,机器学习,手绘检索与识别.

刘玉杰(1971−),男,博士,副教授,CCF 专业会员,主要研究领域为计算机图形图像处理,多媒体数据分析,多媒体数据库,多媒体数据压缩.

邢敏敏(1992−),女,硕士,主要研究领域为图像处理,行人检测.

李宗民(1965−),男,教授,博士生导师,CCF 高级会员,主要研究领域为计算机图形学,图像处理,模式识别.

李华(1956−),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为计算机图形图像处理.