计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2017.20170197
54(8):1824-1832,2017

基于多尺度深度学习的商品图像检索

周 晔 张军平
(复旦大学计算机科学技术学院 上海 200433)
(上海市智能信息处理重点实验室 上海 200433)
(yezhou14@fudan.edu.cn)

Multi-Scale Deep Learning for Product Image Search

Zhou Ye and Zhang Junping
(*School of Computer Science, Fudan University, Shanghai* 200433)
(*Shanghai Key Laboratory of Intelligent Information Processing, Shanghai* 200433)

Abstract Product image search is an important application of mobile visual search in e-commerce. The target of product image search is to retrieve the exact product in a query image. The development of product image search not only facilitates people's shopping, but also results in that e-commerce moves forward to mobile users. As one of the most important performance factors in product image search, image representation suffers from complicated image background, small variance within each product category, and variant scale of the target object. To deal with complicated background and variant object scale, we present a multi-scale deep model for extracting image representation. Meanwhile, we learn image similarity from product category annotations. We also optimize the computation cost by reducing the width and depth of our model to meet the speed requirements of online search services. Experimental results on a million-scale product image dataset shows that our method improves retrieval accuracy while keeps good computation efficiency, comparing with existing methods.

Key words product image search; deep learning; multi scale; metric learning; model compression

摘 要 商品图像检索的目标是检索与图像内容相符的商品,它是移动视觉搜索在电子商务中的重要应用.商品图像检索的发展,既为用户购物提供便利,又促进了电子商务向移动端发展.图像特征是影响商品图片检索性能的重要因素.复杂的图片背景、同类商品之间的相似性和被拍摄商品尺度的变化,都使得商品图像检索对图像特征提出了更高的要求.提出了一种多尺度深度神经网络,以便于抽取对复杂图片背景和目标物体尺度变化更加鲁棒的图像特征.同时根据商品类别标注信息学习图片之间的相似度.针对在线服务对响应速度的要求,通过压缩模型的深度和宽度控制了计算开销.在一个百万级的商品图片数据集上的对比实验证明:该方法在保持速度的同时提升了查询的准确率.

关键词 商品图像检索;深度学习;多尺度;度量学习;模型压缩

中图法分类号 TP391.4

移动视觉搜索是指将移动终端获取的真实世界中的图像或视频作为查询对象,通过移动互联网去搜索视觉对象的关联信息的检索方式[1].电子商务是近年来发展最为迅速的行业之一.商品图像检索

收稿日期:2017-03-20;**修回日期**:2017-05-17
基金项目:国家自然科学基金项目(61673118);上海市浦江人才计划项目(16PJD009)
This work was supported by the National Natural Science Foundation of China (61673118) and Shanghai Pujiang Program (16PJD009).
通信作者:张军平(jpzhang@fudan.edu.cn)

是移动视觉搜索在电子商务中的重要应用.通过智能手机终端与移动视觉搜索技术的结合,用户可以随时在街上、商店中、家中拍摄自己看到的商品照片,并在电子商务网站中检索对应的商品.随着越来越多的电子商务请求从桌面端转移到移动端,商品图像检索的广泛应用可以为用户提供精准的个性化服务,从而为电子商务网站产生巨大的经济价值,这使得商品图像检索成为了一个全新的热门研究领域.如何通过移动设备拍摄的图片精确查找对应的商品,是一个非常困难的问题.首先,移动设备的感光元件与拍摄时的光照条件各不相同,同时,目标商品的视点和尺度的变化、遮挡和模糊等,都使得精确匹配的难度显著加大.不仅如此,同类商品之间的外观可能非常接近,例如服装类的商品,不同的款式之间可能只有颜色、图案等的微小差别.如何区分这些细粒度的物体类别是一个具有挑战性的问题.

商品图像检索可以看作一种限定的基于内容的图像检索(content based image retrieval, CBIR)[2].在基于内容的图像检索系统中,图像特征是影响性能最重要的因素之一[3].由于商品图像检索问题的一些特殊难点,使得商品图像检索对于图像特征的敏感度和判别力提出了更高的要求.如何提取更加有效的图像特征,成为商品图像检索问题研究的主要方向之一.在商品图像检索的研究工作中,尺度不变特征变换(scale invariant feature transform, SIFT)[4]等图像局部特征和 Fisher Vector[5-6]、局部聚合描述符(vector of locally aggregated descriptors, VLAD)[7-8]等传统图像全局特征等均被广泛使用.近年来,使用深度学习方法[8-9]抽取的图像特征在商品图像检索问题上取得了巨大的性能提升.在深度卷积神经网络中,层数越深、每层过滤器(filter)数量越多的网络,通常具有更强的特征表示能力,同时需要更多的运算量.由于在线商品图像检索通常由服务器端进行全部的计算操作,而图像特征的抽取、相似度的计算等,通常耗时巨大.控制模型的复杂度、做到查询准确率与查询速度之间的平衡是在线商品图像检索需要克服的另一个重要难点.

在图像检索中,通常将整个查询图像视为一个整体处理.而商品图像检索问题中,查询图像中只包含一个特定的商品区域,其余部分均可视为背景.被拍摄商品的尺度和图像的背景噪声是影响商品图像检索性能的另外 2 个重要的因素.背景杂乱或被拍摄的商品在图像中的比例过小,都会严重影响查询性能.在商品图像检索中,一些研究工作使用人工标记目标区域[10],另一些使用了图像分割[11]等自动方法,从查询图像中截取包含商品主体的区域后进行处理.与这些方法不同,在我们的方法中,查询图像被视为一个整体进行处理,通过多尺度方法解决商品区域的尺度问题.具体来说,我们提出了一种多尺度的神经网络模型.它可以使用同样的模型参数来接受不同尺寸的输入尺寸.通过对不同尺寸的输入图像进行整合得到的多尺度特征,相对于单尺度特征更有利于提升特征的鲁棒性,减少复杂的图像背景对特征的影响.

除此以外,在互联网图像搜索引擎中,获得有效的标签通常需要消耗巨大的人力,因而通常采用无监督的方法.而在商品图像检索问题中,2 幅图像是否包含同一个商品比较容易确定.因而可以通过人工标记部分数据的标签,使用监督方法学习图像之间的相似度.图像相似度学习在人脸验证等领域中有着广泛的应用.主成分(principal component analysis, PCA)、线性判别分析(linear discriminant analysis, LDA)等均为广泛使用的传统方法[12].近年来,使用深度神经网络的图像相似度学习方法[13-14]被广泛应用.深度图像相似度学习同样应用于商品图像检索中,Wang 等人[15]使用了孪生网络(siamese neural networks)学习商品图片间的相似度.我们使用了 LDA 学习商品图片间的相似度,进一步增强特征的判别能力.

我们的贡献主要有 3 方面.1)提出了一种多尺度深度神经网络模型,在不需要更改模型参数的情况下,我们的多尺度模型可以接受不同尺寸的输入.通过整合图像的全局和局部信息,可以提升对物体尺度的鲁棒性.2)商品图像检索对模型运算速度非常敏感.我们通过对卷积神经网络模型进行压缩,提出了一种更小尺寸的网络模型,可以在压缩模型运算量接近一半的同时基本保持特征的判别力.3)我们通过图像相似度学习的方法进一步提升了特征的判别性能.在一个百万级别的大规模商品图像检索数据集 ALISC 上,我们同时验证了我们提出的方法的准确率和响应速度.在单张图片特征抽取不超过 1 s 的限制内,与现有的其他方法相比,我们的方法取得了最好的检索性能.

1　多尺度商品图像检索方法

在图像检索系统中,最关键的部分是图像之间相似度的计算.在我们的方法中,图像相似度的计算流程如图 1 所示.首先,我们在图片中心截取一些可

能包含目标商品的区域,然后使用神经网络抽取特征.之后,我们使用 LDA 对提取的图像特征进行变换.最后,我们使用余弦相似度对 2 张图片的相似度进行度量.

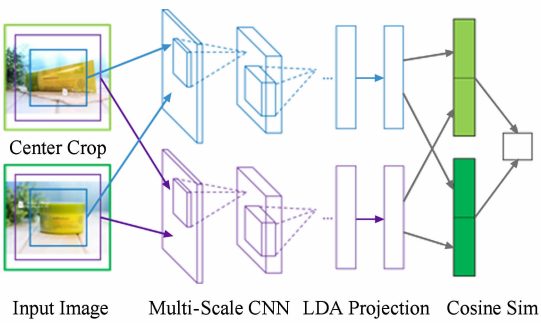


Fig. 1 The pipeline of our proposed method

图 1 方法流程示意

我们首先介绍多尺度卷积神经网络模型.在我们提出的多尺度模型中,同一个网络模型可以接受不同尺寸的输入.之后,我们将介绍使用的模型压缩方法.最后,我们将介绍图片相似度学习与图片相似度的度量方法.

1.1 多尺度卷积神经网络

卷积神经网络(convolutional neural networks, CNN)近年来在图像分类和识别中取得了巨大的成功.LeCun 等人[16]将卷积神经网络成功应用于手写数字识别上.Krizhevsky 等人提出了一个在 ImageNet 数据集上图像分类性能超越传统方法的卷积神经网络模型 AlexNet[17],该模型共有 8 层.Simonyan 等人提出了一个 16 层的卷积神经网络模型[18].Szegedy 等人提出了一个 22 层的卷积神经网络 GoogleNet[19],其中借鉴了多尺度的思想.通常而言,随着 CNN 模型深度和宽度的增加,模型分类性能和特征表示能力均有明显的提升[3].针对 GoogleNet 多个损失函数较难学习的问题,Ioffe 等人提出了与 GoogleNet 结构非常相近,但使用单一损失函数 Inception-6 网络[20].在 GoogleNet 和 Inception-6 网络中,除了通常的卷积、池化等操作,还引入 Inception 模块.在 Inception-6 网络的一个 Inception 模块中,上层特征经过 1×1、3×3、双 3×3、池化等一系列变换后,将特征进行连接作为下一层的输入.在 Inception-6 网络中,使用双 3×3 卷积代替了 GoogleNet 的 Inception 模块中使用的 5×5 卷积,进一步加大了模型的深度.我们使用 Inception-6 网络作为基准模型.Inception-6 模型的输入尺寸为 224×224,完整结构如表 1 中 Output Size(Large)所示:

Table 1 Multi-Scale Inception-6 Model

表 1 多尺度 Inception-6 模型

Type	Size/Stride	Output Size (Large)	Output Size (Small)	1×1	3×3 Reduce	3×3	Double 3×3 Reduce	Double 3×3	Pooling & Projection
Input		224×224×3	160×160×3						
Convolution	7×7/2	112×112×96	80×80×96						
Max Pool	3×3/2	56×56×96	40×40×96						
Convolution	7×7/2	56×56×288	40×40×288		128	288			
Max Pool	3×3/2	28×28×288	20×20×288						
Inception(3a)		28×28×384	20×20×384	96	96	96	96	144	Avg/48
Inception(3b)		28×28×480	20×20×480	96	96	144	96	144	Avg/96
Inception(3c)	Stride 2	14×14×864	10×10×864		192	240	96	144	Max/Pass Through
Inception(4a)		14×14×576	10×10×576	224	64	96	96	128	Avg/128
Inception(4b)		14×14×576	10×10×576	192	96	128	96	128	Avg/128
Inception(4c)		14×14×608	10×10×608	160	128	160	128	160	Avg/128
Inception(4d)		14×14×608	10×10×608	96	128	192	160	192	Avg/128
Inception(4e)	Stride 2	7×7×960	5×5×960		128	192	192	256	Max/Pass Through
Inception(5a)		7×7×1 024	5×5×1 024	352	192	320	160	144	Avg/128
Inception(5b)		7×7×1 024	5×5×1 024	352	192	320	192	224	Max/128
Average Pool		1×1×1 024	1×1×1 024						
Softmax		1×1×21 841	1×1×21 841						

在商品图像检索中,被拍摄的商品的尺度可能差别较大,而除了被拍摄的商品区域外,其他区域均为杂乱的背景噪声.尺度的差别为特征提取带来了难度.我们希望可以复用现有模型的权重信息,使得同一个卷积神经网络模型可以接受不同尺寸的输入数据,并通过后续的模型整合,整合不同输入尺寸的特征,提升对于尺度的鲁棒性.

在 Inception-6 网络中,Inception(5b)两层的输出大小为 7×7,而在 Inception 模块中,双 3×3 卷积需要输入尺寸至少为 5×5.我们将 Inception(5b)的输出尺寸缩减为 5×5,计算可得图片的初始输入尺寸应为 160×160.输入尺寸 160×160 的模型参数与输出尺寸如表 1 中 Output Size(Small)所示.缩减了输入大小后的模型,与原始的模型具有完全一致的权重矩阵大小.即我们可以将同样的模型参数应用到 224×224 与 160×160 两种不同的输入尺寸中.

我们提出的多尺度方法本质是只计算原始图像对应区域的特征.由神经网络卷积层的计算公式可以得出,如果不考虑池化的影响,160×160 小尺寸的输入相当于使用原始的 224×224 输入,但在每一个中间层中,都只保留与中心 160×160 区域对应的输出值,其余值置为 0.即在特征计算的过程中不考虑中心 160×160 以外的图片背景部分.因而这样的计算方式不仅减少了运算量,而且保留了大部分的特征表示能力,减少了背景噪声对于图像特征的影响.在实验中,我们将会验证 224×224 与 160×160 两种不同的输入尺寸的性能.

1.2　模型压缩

Inception-6 网络结构复杂,计算复杂度非常高,为了加速图像特征的计算,我们希望在 Inception-6 网络的基础上进行压缩,构造一个更小更快速的模型.神经网络模型的压缩通常有 2 种可行的方法:压缩模型的深度和压缩模型的宽度.压缩模型的深度,是指通过去掉一些隐含层,使神经网络的层数减少.压缩模型的宽度,是指减少每一层的过滤器个数,使得每一层抽取的特征数量减少.

我们同时使用压缩模型的深度和压缩模型的宽度这 2 种方法.对比压缩后的模型和原始的 Inception-6 模型,我们分别去掉了 Inception(4)和 Inception(5)中的一个 Inception 模块,同时每一层的过滤器个数也有所减少.经过压缩的模型记作 Inception-6-Small 网络,完整的结构如表 2 所示:

Table 2　Multi-Scale Inception-6-Small Model

表 2　多尺度 Inception-6-Small 模型

Type	Size/Stride	Output Size (Large)	Output Size (Small)	1×1	3×3 Reduce	3×3	Double 3×3 Reduce	Double 3×3	Pooling & Projection
Input		224×224×3	160×160×3						
Convolution	7×7/2	112×112×48	80×80×48						
Max Pool	3×3/2	56×56×48	40×40×48						
Convolution	7×7/2	56×56×128	40×40×128		48	128			
Max Pool	3×3/2	28×28×128	20×20×128						
Inception(3a)		28×28×164	20×20×164	40	40	40	40	64	Avg/20
Inception(3b)		28×28×208	20×20×208	40	40	64	40	64	Avg/40
Inception(3c)	Stride 2	14×14×400	10×10×400		96	120	48	72	Max/Pass Through
Inception(4a)		14×14×576	10×10×576	224	64	96	96	128	Avg/128
Inception(4b)		14×14×608	10×10×608	160	128	160	128	160	Avg/128
Inception(4c)		14×14×608	10×10×608	96	128	192	160	192	Avg/128
Inception(4d)	Stride 2	7×7×960	5×5×960		128	192	192	256	Max/Pass Through
Inception(5)		7×7×1024	5×5×1024	352	192	320	192	224	Max/128
Average Pool		1×1×1024	1×1×1024						
Softmax		1×1×21841	1×1×21841						

我们进一步对深度压缩和宽度压缩对模型参数规模的影响进行了定量分析,对比了压缩前和压缩后卷积层参数的数量.结果表明:深度压缩的过程减少了约 27%的卷积层参数,而宽度压缩的过程减少了约 6%的卷积层参数.

模型参数规模的减少将会一定程度地影响模型

的性能. 在实验章节中,我们将会对比经过模型压缩的 Inception-6-Small 模型和原始的 Inception-6 模型的性能. 我们的实验结果表明:经过模型压缩的 Inception-6-Small 模型只有很小的性能损失,但是大大节省了抽取特征所需的时间.

1.3 图像相似度学习与度量

卷积神经网络模型承担了抽取图像特征的功能. Inception-6 网络模型的平均池化层(average pooling)的输出,可以直接作为图像的一个 1 024 维的特征. 但卷积神经网络模型训练时,损失函数通常为图像分类的误差,2 幅图像特征之间的距离并没有具体的物理意义,因而抽取得到的特征向量之间的相似度难以有效度量. 我们使用了线性判别分析(LDA)对特征向量进行进一步的相似度学习,同时,LDA 还可以增强特征的判别性能. LDA 的目标是学习特征不同维度间的一个线性组合. LDA 的目标函数定义为[12]

$$\hat{\boldsymbol{w}}=\arg\max_{\boldsymbol{w}}\frac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{w}},$$

其中,$\boldsymbol{S}_b$ 与 $\boldsymbol{S}_w$ 分别为类间与类内的散布矩阵,分别定义为

$$\boldsymbol{S}_b=\sum_{k=1}^{m}n_k(\boldsymbol{\mu}_k-\boldsymbol{\mu})(\boldsymbol{\mu}_k-\boldsymbol{\mu})^{\mathrm{T}},$$

$$\boldsymbol{S}_w=\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{\mu}_{y_i})(\boldsymbol{x}_i-\boldsymbol{\mu}_{y_i})^{\mathrm{T}},$$

其中,n 为总样本数,m 为总类别数,n_k 为对应类别中的样本数,$\boldsymbol{\mu}$ 为所有样本的均值,$\boldsymbol{\mu}_k$ 为对应类别样本的均值. 即 LDA 的优化目标为类间与类内散布比值的最大化. 在商品图像检索中,不同商品之间可能极为相似,经过 LDA 后,相似的商品类别被尽量区分开,同类的商品尽量接近,进一步增强了特征的判别性能.

LDA 可以转化为对一个广义特征值问题进行求解. 相对其他图像相似度学习方法,LDA 的优势之一是求解速度快,在现有研究[12]中广泛用于高维数据的监督判别投影. 在本研究中它有助于快速对大批量的监督数据进行学习. 在文献[12]中,作者通过舍弃掉 LDA 中特征值较小的维度,学习一个低维的特征空间,但在我们的方法中,LDA 不用于降维,而是保留了完整的 LDA 系数矩阵 $\hat{\boldsymbol{w}}$,即原特征空间到目标特征空间的一个仿射变换. 经过 LDA 变换后的特征维度保持不变,是我们的工作与其他工作的另一个区别.

我们发现,通过 LDA 相似度学习,我们还扩大了图像特征分布的空间. 如图 2 所示,由于 CNN 的激活函数为 ReLU,在 CNN 提取出的特征中没有负值存在,所有的特征向量都集中在第 1 象限. 经过 LDA 之后,特征空间从第 1 象限扩大到了整个空间,有利于提升特征的判别性能.

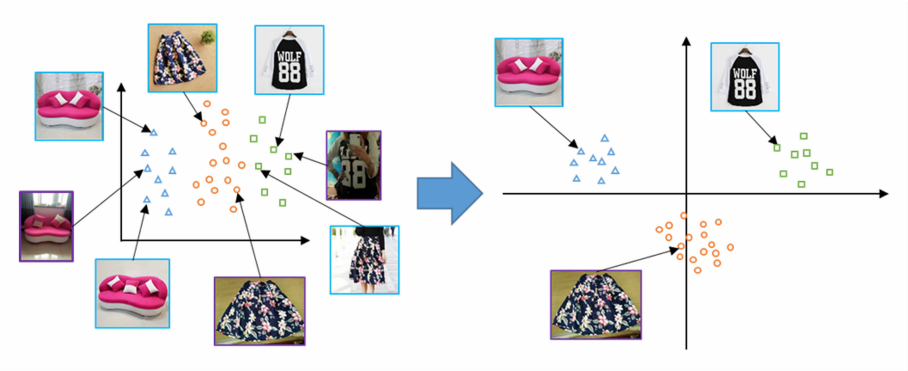


Fig. 2 Illustration of CNN and LDA feature spaces

图 2 CNN 特征空间与 LDA 特征空间示意

图像相似度常用的计算方法为 L2 距离与余弦相似度(cosine similarity)等. 余弦相似度的物理意义是 2 个特征向量间的夹角. 对于 2 张图片的特征向量 $\boldsymbol{a}$ 和 $\boldsymbol{b}$,余弦相似度定义为

$$sim(\boldsymbol{a},\boldsymbol{b})=\frac{\boldsymbol{a}\cdot\boldsymbol{b}}{|\boldsymbol{a}|\times|\boldsymbol{b}|}.$$

2 数据集介绍

我们的实验在 ALISC(Alibaba Large-scale Image Search Challenge)① 数据集上进行. ALISC 数据集分为 3 部分. 训练数据集包含约 195 万张由

① ALISC 数据集来自阿里巴巴集团.

卖家上传的商品描述图片.这些图片可以分为10个商品大类和676个商品子类.验证数据集包含1 417张手机拍摄的查询图片和约320万张备选商品描述图片.测试数据集包含3 567张查询图片和验证数据集共用备选图片.测试数据集的标签信息不公开.

我们使用$MAP@n$作为检索性能的标准.对于单条查询,我们计算检索结果的$AP@n$.$AP@n$的计算为

$$AP@n = \sum_{k=1}^{n} P(k)/\min(m,n),$$

其中,如果第k条是一条正确的结果,$P(k)$表示查询结果排序中到第k条为止的正确结果个数,否则$P(k)=0$.m表示该查询在数据库中的所有正确结果总数.$MAP@n$定义为所有查询$AP@n$的平均值.在商品检索的实际应用中,最受用户关注的首页检索结果通常包含20条左右的商品.根据商品检索的应用特点,我们使用$MAP@20$作为检索性能的标准.

由于测试数据集的标签不对外公开,为了验证模型的性能,我们进一步随机地将验证数据集切分为1 000张训练图片与417张测试图片.在第3节中,我们的部分实验将会报告在验证数据集,即417张测试图片上的$MAP@20$结果.

在我们的实验中,还使用了2个辅助数据集,分别为ImageNet与ImageNet-21K[21].ImageNet数据集是应用最广泛的图像分类数据集之一,包含100多万张图片,分为1 000个类别.ImageNet-21K数据集为ImageNet数据集的扩充,包含1 400多万张图片,涵盖了21 000多个更加细致的类别.

3　实验与分析

3.1　基准模型

我们使用在ImageNet-21K数据集上训练的Inception-6模型[22]作为基准模型,抽取Inception-6模型的平均池化层直接作为图像特征.作为对比,我们还在AlexNet模型[17]上进行实验,在AlexNet模型上,使用最后一层全连接层的输出作为特征.对于输入图片,我们将短边压缩到256像素,之后截取中央的224×224作为CNN的输入.我们在ALISC验证数据集上测试了不同模型的准确率和运行时间.测试模型运行时间的环境为Xeon E5 2650 v2 CPU,主频为2.6 GHz.运行时间为在单核CPU上进行一次模型特征提取需要的时间.我们观测到,使用余弦相似度计算CNN特征的相似度,普遍比使用L2距离的准确度更高,因此我们在之后的实验中均使用余弦相似度.预训练的Inception-6模型的实验结果如表3所示.结果表明Inception-6模型的特征相比AlexNet模型的特征具有更强的表示能力,但是Inception-6模型消耗了更多的运行时间.

Table 3　The Results of Our Baseline Model on Validation Set
表3　基准模型在验证集上的实验结果

Methods	$MAP@20$	Runtime/s
Fine-tuned AlexNet	5.1	0.29
Pre-trained Inception-6	19.4	0.83
Fine-tuned Inception-6	**20.5**	0.83

我们进一步在ALISC数据集上对Inception-6模型进行微调(fine-tune).我们根据ALISC数据集的676个商品子类,训练Inception-6模型在商品子类上的分类性能.我们将21K的softmax层替换为676类的softmax层,使用1e-4的学习率(原模型的初始学习率为1e-3)训练到损失函数收敛为止.之后,我们调整学习率至1e-5,继续学习到模型收敛.预训练的Inception-6模型与经过微调的Inception-6模型在验证数据集上的性能对比如表3所示.结果表明:微调的过程可以提升模型在商品图像检索问题上的判别性能.在之后的实验中,我们采用微调Inception-6模型(之后记作Inception-6)作为基准模型.

3.2　模型压缩

本节中,我们对经过模型压缩的Inception-6-Small模型进行实验,对Inception-6-Small模型在ImageNet-21K数据集上进行训练.与模型微调的过程类似,我们使用1e-3,1e-4,1e-5三种阶梯学习率学习到模型收敛为止.在训练过程中,我们使用了Batch Normalization[20]对模型进行归一化,提升模型收敛速度.训练过程在一台双路GTX Titan X的服务器上进行,耗时约2周.训练后的模型在ImageNet-21K训练集上的Top-1准确率为37.8%.对比预训练的Inception-6模型在ImageNet-21K训练集上的Top-1准确率为37.1%,证明我们提出的Inception-6-Small模型具有与Inception-6模型相近的特征表示能力.

我们在验证数据集上对比了Inception-6与Inception-6-Small模型的检索性能与运行时间.如表4所示,Inception-6-Small模型的性能接近Inception-6模型,但是模型消耗的运算时间减少了

近一半.我们推测性能损失是由于减少了每一层特征抽取的过滤器数量,导致 Inception-6-Small 模型虽然在图像分类问题上的性能与 Inception-6 相似,但是在图像检索问题上的性能有一定的损失.

Table 4 The Results of Compressed Model on Validation Set
表 4 压缩后的模型在验证集上的实验结果

Methods	*MAP*@20	Runtime/s
Inception-6	20.5	0.83
Inception-6-Small	**19.1**	0.43

3.3 多尺度模型测试

我们对提出的多尺度模型测试方法进行实验验证,使用 Inception-6 与 Inception-6-Small 两个模型进行多尺度测试.我们对比了 224×224 的原始输入尺寸与 160×160 的输入尺寸下,模型的准确率与耗时.为了区别 2 种输入尺寸,使用 160×160 的输入尺寸的结果以"-160"结尾,实验结果如表 5 所示.实验结果表明:在 160×160 输入尺寸下模型的计算时间大约减少了一半,但是也带来了一些性能损失.观察一些测试图片之后我们发现,截取图像中心的 160×160 部分之后,虽然截取图像的中心区域可以裁剪掉了一部分背景,从而减少输入图像的噪声,但如果被拍摄的商品在图片中的位置不在正中央,或被拍摄的商品过大超出了图像中心区域,则商品也有一部分会被裁剪掉,我们猜测这是导致160×160 的小尺寸输入产生性能损失的主要原因.

Table 5 The Results of Multi-scale Model on Validation Set
表 5 多尺度模型在验证集上的实验结果

Methods	*MAP*@20	Runtime/s
Inception-6	20.5	0.83
Inception-6-Small	19.1	0.43
Inception-6-160	**18.6**	0.46
Inception-6-Small-160	**17.3**	0.22

3.4 图像相似度学习与模型整合

我们使用验证数据集的标签训练 LDA 模型.验证数据集的 1000 张训练图片,总共包含约 6 万个正确查询结果,平均每张查询图片有 60 个结果.我们将每一个查询作为子类,使用 CNN 特征训练了一个 1000 类的多类 LDA 模型.在验证数据集上的实验结果如表 6 所示.我们发现,LDA 对于所有的 CNN 模型提取的特征均可带来不同程度的性能提升,同时,LDA 在模型测试时会带来约 0.03 s 的额外时间消耗.

Table 6 The Results of LDA Features on Validation Set
表 6 LDA 特征在验证集上的实验结果

Methods	*MAP*@20	Runtime/s
Inception-6	28.6	0.86
Inception-6-Small	27.8	0.47
Inception-6-160	26.2	0.49
Inception-6-Small-160	25.5	0.26
Inception-6 + Inception-6-Small	29.2	1.32 (Timeout)
Inception-6-Small + Inception-6-160	**32.6**	0.96
All 4 models	36.2	2.06 (Timeout)

我们继续实验了不同模型整合的效果.不同模型可以通过对多个余弦相似度取均值来实现整合.我们首先实验了 Inception-6 与 Inception-6-Small 模型进行整合,2 个模型的输入尺寸均为 224×224.同时,我们对比不同输入尺寸的模型整合的效果.如表 6 所示,我们发现,将 224×224 与 160×160 两个不同的尺度的模型进行整合,不同尺度的特征信息互相补充,可以比 2 个 224×224 大小的模型带来更为明显的性能提升,显示了多尺度的重要性.若将我们提出的 4 种模型进行整合,性能可以进一步提升.但在线检索服务对图片特征提取消耗的时间非常敏感,在测试中,每张图片特征提取的时间被限制在 1 s 以内,使用的模型过多则不能满足时间限制的要求.Inception-6-Small 与 Inception-6-160 两个不同尺度模型的整合,同时满足了性能和速度的要求.

最后,我们在封闭的测试数据集上测试了本文提出的方法.实验结果如表 7 所示.使用我们提出的多尺度测试方法,将 Inception-6-Small 与 Inception-6-160 两个模型进行整合,在与 Inception-6 单模型的计算速度相近的情况下,取得了较大的性能提升,证明了多尺度方法在商品图像检索问题上的有效性.使用 4 种模型进行整合可以取得最好的效果,但超过了图片特征抽取的时间限制.在时间限制内,Inception-6-Small 与 Inception-6-160 模型的组合取得了最好的效果.

我们同时对比了所提方法与 Qi 等人[8]提出的方法.在实验结果中可以看到,在我们的方法与 Qi 等人的方法中,深度学习方法的性能均全面超过了 SIFT 与 VLAD 等传统方法.同时,我们的图像特征在测试集上取得了更好的效果.

Table 7　The Results of Different Methods on Testing Set
表 7　不同方法在测试数据集上的实验结果

Methods	*MAP*@20
SIFT (VLAD) [8]	11.1
Inception-6	25.3
Inception-6 (LDA)	32.0
NIN + CaffeNet (VLAD)[8]	32.0
NIN + CaffeNet + SIFT (VLAD) [8]	35.2
Inception-6-Small + Inception-6-160 (LDA)	**37.4**
All 4 Models (LDA) (Timeout)	40.5

如图 3 所示,我们选取了测试集中的一些代表的商品类型,展示了查询图片与结果图片.结果图片的标记表示对应的图片在数据集标注的正确结果中出现过.结果表明,我们的方法在食品、化妆品等大类中,均可以取得较好的检索效果.但对于服装类等商品,商品种类繁多,不同视角、环境拍摄的商品可能外观差异极大,难以保证结果的绝对准确.我们的方法仍然可以检索到外观非常相似的商品供用户参考,但精确度还有待于进一步的提升.

Fig. 3　Some query images and results
图 3　部分查询图片与检索结果

4　总　　结

本文提出了一种多尺度方法解决在线商品图像检索问题.我们提出了一种多尺度网络,可以在不修改模型参数的条件下接受不同尺寸的图像输入.同时,我们在 Inception-6 模型的基础上,将模型运算时间压缩了近一半,同时取得了与原始模型相似的性能.我们使用了 LDA 进行图像相似度学习,进一步提升了特征的判别力.在 ALISC 数据集上,我们的方法在相近的运行时间下,相对其他方法性能提升明显,同时保证了检索准确率和在线响应速度.

参　考　文　献

[1] Duan Lingyu, Huang Tiejun, Alex C K, et al. Mobile visual search: Technical bottlenecks and challenges [J]. Communications of the CCF, 2012, 8(12): 8-15 (in Chinese)
(段凌宇, 黄铁军, Alex C K, 等. 移动视觉搜索技术瓶颈与挑战[J]. 中国计算机学会通讯, 2012, 8(12): 8-15)

[2] Datta R, Joshi D, Li Jia, et al. Image retrieval: Ideas, influences, and trends of the new age [J]. ACM Computing Surveys, 2008, 40(2): 5

[3] Jiang Shuqiang, Min Weiqing, Wang Shuhui. Survey and prospect of intelligent interaction-oriented image recognition techniques [J]. Journal of Computer Research and Development, 2016, 53(1): 113-122 (in Chinese)
(蒋树强, 闵巍庆, 王树徽. 面向智能交互的图像识别技术综述与展望[J]. 计算机研究与发展, 2016, 53(1): 113-122)

[4] Lowe D G. Object recognition from local scale-invariant features [C] //Proc of the 7th IEEE Int Conf on Computer Vision (ICCV), Volume 2. Piscataway, NJ: IEEE, 1999: 1150-1157

[5] Perronnin F, Liu Y, Sánchez J, et al. Large-scale image retrieval with compressed fisher vectors [C] //Proc of the 23rd IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2010: 3384-3391

[6] Perronnin F, Sánchez J, Mensink T. Improving the fisher kernel for large-scale image classification [C] //Proc of the 11th European Conf on Computer Vision (ECCV). Berlin: Springer, 2010: 143-156

[7] Jégou H, Douze M, Schmid C, et al. Aggregating local descriptors into a compact image representation [C] //Proc of the 23rd IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2010: 3304-3311

[8] Qi Shuhan, Zawlin K, Zhang Hanwang, et al. Saliency meets spatial quantization: A practical framework for large scale product search [C/OL] //Proc of IEEE Int Conf on Multimedia & Expo (ICME) Workshops. Piscataway, NJ: IEEE, 2016 [2017-05-20]. http://ieeexplore.ieee.org/document/7574756

[9] Wan Ji, Wang Dayong, Hoi S C H, et al. Deep learning for content-based image retrieval: A comprehensive study [C] //Proc of the 22nd ACM Int Conf on Multimedia (MM). New York: ACM, 2014: 157-166

[10] Hadi Kiapour M, Han Xufeng, Lazebnik S, et al. Where to buy it: Matching street clothing photos in online shops [C] //Proc of the 15th IEEE Int Conf on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2015: 3343-3351

[11] Shen Xiaohui, Lin Zhe, Brandt J, et al. Mobile product image search by automatic query object extraction [C] //Proc of the 12th European Conf on Computer Vision (ECCV). Berlin: Springer, 2012: 114-127

[12] Belhumeur P N, Hespanha J P, Kriegman D J. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1997, 19(7): 711-720

[13] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering [C] //Proc of the 28th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2015: 815-823

[14] Wang Jiang, Song Yang, Leung T, et al. Learning fine-grained image similarity with deep ranking [C] //Proc of the 27th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2014: 1386-1393

[15] Wang Xi, Sun Zhenfeng, Zhang Wenqiang, et al. Matching user photos to online products with robust deep features [C] //Proc of the 18th ACM on Int Conf on Multimedia Retrieval (ICMR). New York: ACM, 2016: 7-14

[16] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324

[17] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C] // Proc of Advances in Neural Information Processing Systems (NIPS). Montreal: NIPS Foundation, 2012: 1097-1105

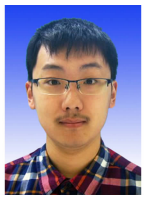
[18] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [J]. arXiv preprint arXiv: 1409.1556, 2014

[19] Szegedy C, Liu Wei, Jia Yangqing, et al. Going Deeper With Convolutions [C] //Proc of the 28th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2015: 1-9

[20] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift [J]. arXiv preprint arXiv: 1502.03167, 2015

[21] Deng Jia, Dong Wei, Socher R, et al. Imagenet: A large-scale hierarchical image database [C] //Proc of the 22nd IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2009: 248-255

[22] Chen Tianqi, Li Mu, Li Yutian, et al. MXNet: A flexible and efficient machine learning library for heterogeneous distributed systems [J]. arXiv preprint arXiv: 1502.01274, 2015



Zhou Ye, born in 1992. Master candidate at the School of Computer Science, Fudan University. Student member of CCF. His main research interests include deep learning and computer vision.

Zhang Junping, PhD, born in 1970. Professor at the School of Computer Science, Fudan University. Member of CCF. His main research interests include machine learning, image processing, biometric authentication, and intelligent transportation systems.