基于卷积神经网络和监督核哈希的图像检索方法

柯圣财[1],赵永威[2],李弼程[1],彭天强[3]
(1. 解放军信息工程大学信息系统工程学院,河南郑州 450001;2. 武警工程大学电子技术系,陕西西安 710000;
3. 河南工程学院计算机学院,河南郑州 450001)

摘 要: 当前主流的图像检索方法采用的视觉特征,缺乏自主学习能力,导致其图像表达能力不强,此外,传统的特征索引方法检索效率较低,难以适用于大规模图像数据. 针对这些问题,本文提出了一种基于卷积神经网络和监督核哈希的图像检索方法. 首先,利用卷积神经网络的学习能力挖掘训练图像内容的内在隐含关系,提取图像深层特征,增强特征的视觉表达能力和区分性;然后,利用监督核哈希方法对高维图像深层特征进行监督学习,并将高维特征映射到低维汉明空间中,生成紧致的哈希码;最后,在低维汉明空间中完成对大规模图像数据的有效检索. 在 ImageNet-1000 和 Caltech-256 数据集上的实验结果表明,本文方法能够有效地增强图像特征的表达能力,提高图像检索效率,优于当前主流方法.

关键词: 深度学习; 图像检索; 卷积神经网络; 近似近邻检索; 监督核哈希
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2017)01-0157-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2017.01.022

Image Retrieval Based on Convolutional Neural Network and Kernel-Based Supervised Hashing

KE Sheng-cai[1], ZHAO Yong-wei[2], LI Bi-cheng[1], PENG Tian-qiang[3]
(1. *Institute of Information System Engineering, Information Engineering University, Zhengzhou, Henan* 450001, *China*;
2. *Institute of Electronic Technology, Engineering University of CAPF, Xi'an, Shaanxi* 710000, *China*;
3. *Institute of Computing Science, Henan University of Engineering, Zhengzhou, Henan* 450001, *China*)

Abstract: The visual features of the state-of-the-art image retrieval methods lack of learning ability, which lead to low expression ability. And the efficiency of traditional index methods is fairly low for large image database. In view of this, an image retrieval method based on convolutional neural network and kernel-based supervised Hashing is proposed. Firstly, a large convolutional neural network is employed to learn the intrinsic implications of training images so as to improve the distinguish ability and expression ability of visual feature. Secondly, kernel-based supervised Hashing is applied to learn from the high-dimensional visual feature and map into low-dimensional hamming space and achieve compact Hash codes. Finally, image retrieval is accomplished in low-dimensional hamming space. Experimental results of ImageNet-1000 and Caltech-256 datasets indicate that the expression ability of visual feature is effectively improved and the image retrieval performance is substantially boosted compared with the state-of-the-art methods.

Key words: deep learning; image retrieval; convolutional neural network; approximate nearest neighbor; kernel-based supervised Hashing

1 引言

随着大数据时代的到来,互联网图像资源迅猛增长,如何对大规模图像资源进行快速有效地检索以满足用户需求亟待解决. 图像检索技术由早期的基于文本的图像检索(Text-based Image Retrieval, TBIR)逐渐发展为基于内容的图像检索(Content-based Image Retrieval, CBIR),CBIR 通过提取图像视觉底层特征来实现图像内容表达. 虽然视觉底层特征中,如 GIST[1]、SIFT[2]、SURF[3]等在图像处理领域表现出优良的性能,但是生成这些描述子时

收稿日期:2015-06-25;修回日期:2016-02-29;责任编辑:李勇锋
基金项目:国家自然科学基金(No. 60872142, No. 61301232)

固定的编码步骤使得描述子缺少学习能力,限制了其图像内容表达能力,难以适应多样的图像数据.

为得到大量图像数据的内在隐含关系,生成更具有区分性和代表性的特征,Hinton 等学者[4~6]将深度学习应用于图像处理领域中,为提取更加有效的图像特征提供了新思路. Tang 等[7]将 DBN 第一层采用稀疏化连接,同时利用概率降噪算法提高 DBN 输出特征对噪声的鲁棒性. Lee 等[8]构建了卷积深度置信网络(Convolutional Deep Belief Network,CDBN),利用 CDBN 能从未标注的自然图像中学习有效的高阶特征表示. Huang 等[9]在 CDBN 的基础上提出了局部卷积受限玻尔兹曼机(Local Convolutional Restricted Boltzmann Machines,LCRBM)模型,该模型利用对象类的总体结构学习特征,在人脸识别任务中取得非常好的效果. He 等[10]通过在卷积神经网络(Convolutional Neural Network,CNN)的卷积层和全连接层加入 SPP(Spatial Pyramid Pooling)层,直接对不同大小图像进行学习并生成多尺度特征. 但是,深度学习生成的图像特征维数较高,存在维数灾难问题,当图像数据规模较大时,若采用传统的最近邻检索方法(如 R-tree[11]、KD-tree[12]等)进行检索就会使检索速度急剧下降,难以适用于大规模数据.

为实现对大规模高维图像数据进行有效检索,研究者提出了近似最近邻搜索策略(Approximate Nearest Neighbor,ANN). 其中,哈希技术是解决近似最近邻检索问题的主流方法,其思想是利用哈希函数族将高维图像特征映射到低维空间中,同时使得原空间中距离较近的点映射到低维空间后仍保持较近的距离. 早期的哈希方法,例如位置敏感哈希[13](Locality Sensitive Hashing,LSH)及其改进算法[14,15],利用随机映射构造哈希函数,为了保证较高的准确率需要生成更长的哈希码,但是随着哈希码的增长,相似图像的哈希码映射到同一哈希桶的概率会逐步减少,导致较低的召回率. LSH 及其改进算法[14,15]构造的哈希函数都是与数据无关的,近年来,研究者们针对如何结合数据特点构造有效、紧致的哈希函数提出了许多算法. Yair 等[16]提出了谱哈希方法(Spectral Hashing,SH),首先对相似图的拉普拉斯矩阵特征值和特征向量进行分析,再通过放宽限制条件,将图像特征向量编码问题转换为拉普拉斯特征图的降维问题进行求解,该方法依赖数据本身生成索引比随机产生哈希函数方法达到更高的准确率. 无监督的方法并没有考虑图像的语义信息,而用户往往更倾向于检索结果的语义信息,为此,Wang 等[17]将图像的语义相似性作为监督信息,提出了半监督哈希方法(Semi-Supervised Hashing,SSH). 在半监督学习方法的基础上研究者们还提出了一些全监督哈希方法,全监督哈希方法相比于非监督方法能达到更高的准确率,但是存在优化过程较为复杂、训练效率低等问题,这严重限制了其在大规模数据集上的应用.

综上所述,为实现更加准确高效的图像检索,本文提出一种基于卷积神经网络和监督核哈希的图像检索方法. 首先,引入卷积神经网络对训练图像进行学习,利用其特殊网络结构隐式地学习得到图像数据的高阶表示,生成具有更强区分性和表达能力的深层特征;然后,引入监督核哈希方法(Kernel-Based Supervised Hashing,KSH)增强对线性不可分数据的分辨力,同时利用哈希码内积与汉明距离的等价关系设计更加简单、有效的目标函数,并结合训练图像的相似性信息对高维图像特征进行监督学习,并生成紧致的哈希码;最后,利用已训练好的哈希函数构造图像索引,实现对大规模图像数据的高效检索.

2 基于卷积神经网络和监督核哈希的图像检索

2.1 基于卷积神经网络的图像深层特征提取

卷积神经网络[18](Convolutional Neural Network,CNN)是第一个真正成功训练多层网络结构的学习算法,并被广泛应用于解决如何提取学习图像数据的深层特征问题[8~10]. CNN 的基本思想:将图像的局部感知区域作为网络的输入,信息再依次传输到不同的层,每层通过一个数字滤波器去获取对平移、旋转和缩放具有不变性的显著特征. 用于提取图像深层特征的卷积神经网络结构[19]如图 1 所示.

该卷积神经网络的输入图像大小为 227×227,输出为 4096×1 的图像深层特征,一共包含 5 个卷积层、3 个子采样层. 在卷积层,前一层的特征图 $x_i^{(l-1)}$ 与可学习的卷积核 K_{ij} 进行卷积,卷积的结果经非线性函数 $g(\cdot)$ 生

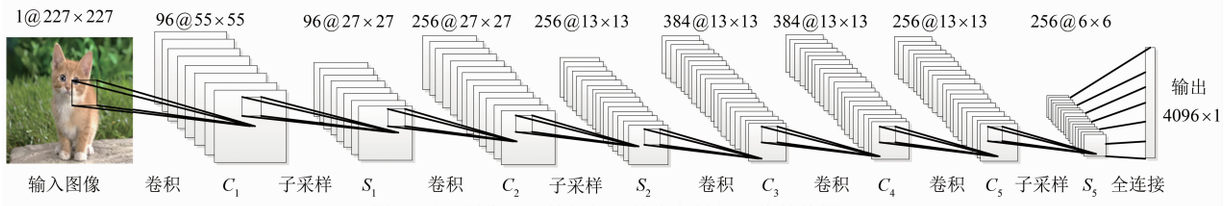


图1 用于图像深层特征提取的卷积神经网络结构图

成这一层的特征图 $y_j^{(l)}$,具体形式如下:

$$y_j^{(l)} = g\left(\sum_{i\in M_j} K_{ij} \otimes x_i^{(l-1)} + b_j\right) \tag{1}$$

其中,$y_j^{(l)}$ 为第 l 个卷积层 C_l 的输出,$\otimes$代表卷积运算,b_j 为偏置,卷积核 K_{ij}可与前一层的一个或多个特征图确定卷积关系,M_j 代表输入特征图集合,常用的非线性函数有 $g(x)=\tanh(x)$ 和 $g(x)=(1+\mathrm{e}^{-x})^{-1}$,与上述非线性函数相比,$g(x)=\max(0,x)$能有效提高训练效率[20]. 卷积层生成的特征图大小 h_l 为:

$$h_l = \frac{h_{l-1}+2\times\rho_l - z_l}{\lambda_l}+1 \tag{2}$$

其中,h_{l-1}为第 $l-1$ 层特征图的大小,z_l 表示第 l 层卷积核的大小,λ_l 是卷积核移动步长,ρ_l 表示卷积运算时对前一层特征图边缘补零的列数. 这里,各层卷积核大小 $Z=\{z_1=11,z_2=5,z_3=z_4=z_5=3\}$,移动步长 $\Lambda=\{\lambda_1=4,\lambda_2=\lambda_3=\lambda_4=\lambda_5=1\}$,特征图边缘补零列数 $P=\{\rho_1=0,\rho_2=2,\rho_3=\rho_4=\rho_5=1\}$. 在子采样层,文献[19]研究表明相对于传统的无重叠采样,使用重叠采样不仅能提高特征的准确性,还可以防止训练阶段出现过拟合. 因此,这里采用重叠采样方法对特征图进行最大值采样,采样区域为 3×3,采样步长为 2 个像素.

卷积神经网络的训练主要分前向传播和后向传播两个阶段:

(1)前向传播阶段. 从训练样本中选取一个样本 (X,Y_p),X 从输入层经逐级变换传送到输出层,计算相应的实际输出:

$$O_p = F_n(\cdots(F_2(F_1(XW^{(1)})W^{(2)})\cdots)W^{(n)}) \tag{3}$$

(2)后向传播阶段,也称误差传播阶段. 计算实际输出 O_p 与对应理想输出 Y_p 的误差:

$$E_p = \frac{1}{2}\sum_{j=1}^{m}(y_{pj}-o_{pj})^2 \tag{4}$$

将误差 E_p 反向逐层后推得到各层的误差,并按最小化误差方法调整神经元权值,当总误差 $E\leqslant\varepsilon$ 时,完成该批次训练样本的训练. 当所有批次训练完成后,将图像输入卷积神经网络中,图像数据逐级通过各个网络层后,在输出端即可得到图像的深层特征.

2.2　基于监督核哈希的图像检索

为增强哈希函数对线性不可分的高维数据 $X=\{x_1,\cdots,x_n\}\subset\mathbb{R}^d$ 的分辨能力,利用核函数 $\kappa:\mathbb{R}^d\cdot\mathbb{R}^d\to\mathbb{R}$ 构建哈希函数 $h:\mathbb{R}^d\to\{1,-1\}$①,对高维数据进行映射生成哈希码,哈希函数具体形式为:

$$f(x) = \sum_{j=1}^{m}\kappa(x_{(j)},x)a_j - b \tag{5}$$

$$h(f(x))=\mathrm{sgn}(f(x))=\begin{cases}1, & f(x)>0\\ -1, & f(x)\leqslant 0\end{cases} \tag{6}$$

其中,$a_j\in\mathbb{R}$,$b\in\mathbb{R}$,$x_{(1)},\cdots,x_{(n)}$是从 χ 中随机选取的 m 个样本,为了实现快速哈希映射,m 是远小于 n 的常数. 哈希函数 $h(x)$除了满足低维汉明空间与原始高维空间的相似一致性外,还应保证生成的哈希码是均衡的,即哈希函数 $h(x)$ 应满足 $\sum_{i=1}^{n}h(x_i)=0$,则偏置 $b=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\kappa(x_{(j)},x)a_j$,将 b 的值代入式(5)可得:

$$\begin{aligned} f(x) &= \sum_{j=1}^{m}\left(\kappa(x_{(j)},x)-\frac{1}{n}\sum_{i=1}^{n}\kappa(x_{(j)},x)\right)a_j\\ &= \boldsymbol{a}^{\mathrm{T}}\bar{\boldsymbol{k}}(x)\end{aligned} \tag{7}$$

其中,$\boldsymbol{a}=[a_1,\cdots a_m]^{\mathrm{T}}$,$\bar{\boldsymbol{k}}:\mathbb{R}^d\to\mathbb{R}^m$ 是映射向量:

$$\bar{\boldsymbol{k}}=[\kappa(x_{(1)},x)-\mu_1,\cdots,\kappa(x_{(m)},x)-\mu_m]^{\mathrm{T}} \tag{8}$$

这里,$\mu_j=\frac{1}{n}\sum_{i=1}^{n}\kappa(x_{(j)},x_i)$ 可通过预先计算得到,传统哈希方法中系数向量 $\boldsymbol{a}$ 是通过随机抽样得到的 m 维向量,为增强生成哈希码的区分性,提高检索查准率,本文利用训练数据的相关性信息进行监督学习得到系数向量 $\boldsymbol{a}$,构造与数据相关的哈希函数.

给定哈希码的维数 r,则需要 r 个向量 $\boldsymbol{a}_1,\cdots,\boldsymbol{a}_r$ 构造哈希函数 $\mathcal{H}=\{h_k(x)=\mathrm{sgn}(\boldsymbol{a}^{\mathrm{T}}\bar{\boldsymbol{k}}(x))\mid k\in[1,r]\}$. 训练图像的标签信息可通过图像的语义相关性和空间距离获得,$lable(x_i,x_j)=1$ 表示图像 x_i,x_j 是相似的;反之,$lable(x_i,x_j)=-1$ 代表图像 x_i,x_j 差异很大. 为描述标签图像集 $\chi_l=\{x_1,\cdots,x_l\}$ 中元素之间的相互关系,定义监督矩阵 $\boldsymbol{S}\in\mathbb{R}^{l\times l}$:

$$s_{ij}=\begin{cases}1, & lable(x_i,x_j)=1\\ -1, & lable(x_i,x_j)=-1\\ 0, & \text{otherwise}\end{cases} \tag{9}$$

其中,$lable(x_i,x_i)=1$,$S_{ii}=1$,$S_{ij}=0$ 表示图像 x_i,x_j 之间的相似性不确定. 为增强哈希码的区分能力,使得在汉明空间中能高效地判断图像之间的相似性,应尽量使图像 x_i,x_j 的汉明距离 $D_h(x_i,x_j)$满足:

$$D_h(x_i,x_j)=\begin{cases}0, & S_{ij}=1\\ r, & S_{ij}=-1\end{cases} \tag{10}$$

由于汉明距离计算公式形式复杂,很难直接对其进行优化,因此本文利用向量内积运算计算哈希码之间的距离. 记图像 x 的哈希码为 $code_r(x)=[h_1(x),\cdots,h_r(x)]\in\{1,-1\}^{1\times r}$,则图像 x_i,x_j 的距离 $D(x_i,x_j)$为:

$$\begin{aligned} D(x_i,x_j) &= code_r(x_i)\cdot code_r(x_j)\\ &= |\{k\mid h_k(x_i)=h_k(x_j),1\leqslant k\leqslant r\}|\\ &\quad -|\{k\mid h_k(x_i)\neq h_k(x_j),1\leqslant k\leqslant r\}|\\ &= r-2|\{k\mid h_k(x_i)\neq h_k(x_j),1\leqslant k\leqslant r\}|\\ &= r-2D_h(x_i,x_j)\end{aligned} \tag{11}$$

① 在训练阶段使用“-1”代替“0”比特,在对数据进行哈希编码时仍使用“0”。

式(11)表明了通过哈希码内积运算与汉明距离运算是一致的,且 $D(x_i,x_j)\in[-r,r]$,对 $D(x_i,x_j)$ 归一化后得到 $S'_{ij}=\dfrac{D(x_i,x_j)}{r}\in[-1,1]$. 为使得相似矩阵 $\boldsymbol{S}'=\dfrac{1}{r}\boldsymbol{H}_l\boldsymbol{H}_l^{\mathrm{T}}$ 与监督矩阵 $\boldsymbol{S}$ 距离最小,定义目标函数:

$$\min \quad \Gamma=\left\|\frac{1}{r}\boldsymbol{H}_l\boldsymbol{H}_l^{\mathrm{T}}-\boldsymbol{S}\right\|_{\mathrm{F}}^2 \tag{12}$$

其中,$\|\cdot\|_{\mathrm{F}}^2$ 表示求矩阵 Frobenius 范数,$\boldsymbol{H}_l=\begin{bmatrix}code_r(x_1)\\ \cdots\\ code_r(x_l)\end{bmatrix}\in\{1,-1\}^{l\times r}$ 为标签图像集 χ_l 的哈希码矩阵. 将 sgn(·)推广到矩阵形式,根据公式(7),$\boldsymbol{H}_l$ 可表示成:

$$\boldsymbol{H}_l=\begin{bmatrix}h_1(x_1) & \cdots & h_r(x_1)\\ \cdots & \cdots & \cdots\\ h_1(x_l) & \cdots & h_r(x_l)\end{bmatrix}=\mathrm{sgn}(\bar{K}_lA) \tag{13}$$

其中,$\bar{K}_l=[\bar{k}(x_1),\cdots,\bar{k}(x_l)]^{\mathrm{T}}\in\mathbb{R}^{l\times m}$,$A=[\boldsymbol{a}_1,\cdots,\boldsymbol{a}_r]\in\mathbb{R}^{m\times r}$,将 $\boldsymbol{H}_l$ 代入式(12)得

$$\min_{A\in\mathbf{R}^{m\times r}} \quad \Gamma(A)=\left\|\frac{1}{r}\mathrm{sgn}(\bar{K}_lA)(\mathrm{sgn}(\bar{K}_lA))^{\mathrm{T}}-S\right\|_{\mathrm{F}}^2 \tag{14}$$

$$\Rightarrow \quad \min_{A\in\mathbf{R}^{m\times r}} \quad \left\|\sum_{k=1}^{r}\mathrm{sgn}(\bar{K}_l\boldsymbol{a}_k)(\mathrm{sgn}(\bar{K}_l\boldsymbol{a}_k))^{\mathrm{T}}-rS\right\|_{\mathrm{F}}^2 \tag{15}$$

与 BRE 相比,目标函数 $\Gamma(A)$ 通过内积计算相似性,对参数 A 建模更加直观. 假定在 $t=k$ 时刻,已知向量 $\boldsymbol{a}_1^*$,$\cdots$,$\boldsymbol{a}_{k-1}^*$,需要估算 $\boldsymbol{a}_k$,定义矩阵 $R_{k-1}=rS-\sum_{t=1}^{k-1}\mathrm{sgn}(\bar{K}_l\boldsymbol{a}_t^*)(\mathrm{sgn}(\bar{K}_l\boldsymbol{a}_t^*))^{\mathrm{T}}$,其中 $R_0=rS$,则可通过贪婪算法最小化式(15)逐步估算 $\boldsymbol{a}_k$:

$$\begin{aligned}&\left\|\mathrm{sgn}(\bar{K}_l\boldsymbol{a}_k)(\mathrm{sgn}(\bar{K}_l\boldsymbol{a}_k))^{\mathrm{T}}-R_{k-1}\right\|_{\mathrm{F}}^2\\ &=((\mathrm{sgn}(\bar{K}_l\boldsymbol{a}_k))^{\mathrm{T}}\mathrm{sgn}(\bar{K}_l\boldsymbol{a}_k))^2-\\ &\quad 2(\mathrm{sgn}(\bar{K}_l\boldsymbol{a}_k))^{\mathrm{T}}R_{k-1}\mathrm{sgn}(\bar{K}_l\boldsymbol{a}_k)+\mathrm{tr}(R_{k-1}^2)\\ &=-2(\mathrm{sgn}(\bar{K}_l\boldsymbol{a}_k))^{\mathrm{T}}R_{k-1}\mathrm{sgn}(\bar{K}_l\boldsymbol{a}_k)+l^2+\mathrm{tr}(R_{k-1}^2)\\ &=-2(\mathrm{sgn}(\bar{K}_l\boldsymbol{a}_k))^{\mathrm{T}}R_{k-1}\mathrm{sgn}(\bar{K}_l\boldsymbol{a}_k)+const\end{aligned} \tag{16}$$

去掉常数项,可以得到更简洁的目标函数:

$$\vartheta(\boldsymbol{a}_k)=-(\mathrm{sgn}(\bar{K}_l\boldsymbol{a}_k))^{\mathrm{T}}R_{k-1}\mathrm{sgn}(\bar{K}_l\boldsymbol{a}_k) \tag{17}$$

由于目标函数中的 sgn(x)函数使得 $\vartheta(\boldsymbol{a}_k)$ 不连续,而且 $\vartheta(\boldsymbol{a}_k)$ 也不是凸函数,很难直接对 $\vartheta(\boldsymbol{a}_k)$ 最小化,文献[21]研究表明,当 $|x|>6$ 时,连续函数 $\varphi(x)=2/(1+\exp(-x))-1$ 能很好地近似 sgn(x). 因此,本文利用 $\varphi(x)$ 替换 sgn(x),得到近似目标函数 $\tilde{\vartheta}(\boldsymbol{a}_k)$:

$$\tilde{\vartheta}(\boldsymbol{a}_k)=-(\varphi(\bar{K}_l\boldsymbol{a}_k))^{\mathrm{T}}R_{k-1}\varphi(\bar{K}_l\boldsymbol{a}_k) \tag{18}$$

可通过梯度下降法对 $\tilde{\vartheta}(\boldsymbol{a}_k)$ 最小化,求 $\tilde{\vartheta}(\boldsymbol{a}_k)$ 关于 $\boldsymbol{a}_k$ 求梯度得:

$$\nabla\tilde{\vartheta}(\boldsymbol{a}_k)=-\bar{K}_l^{\mathrm{T}}((R_{k-1}\boldsymbol{b})\odot(1-\boldsymbol{b}\odot\boldsymbol{b})) \tag{19}$$

其中,$\boldsymbol{b}=\varphi(\bar{K}_l\boldsymbol{a}_k)\in\mathbb{R}^l$,$\mathbf{1}=[1,\cdots,1]\in\mathbb{R}^l$,$\odot$ 表示 Hadamard 内积运算. 经平滑处理后的 $\tilde{\vartheta}$ 不是凸函数,无法求得全局最优解,为了加速 $\tilde{\vartheta}$ 收敛,本文利用谱哈希[16]中的谱分析方法生成初始值 $\boldsymbol{a}_k^0$,再用文献[22]中的方法加速梯度寻优过程.

通过监督学习得到向量系数 $\boldsymbol{a}$ 后,即可生成哈希函数 $\mathcal{H}$ 和哈希表 H,对查询图像的深层特征进行哈希映射得到 $code_r(x_q)$,计算 $code_r(x_q)$ 与哈希表 H 中哈希码的距离,返回距离较近的图像作为检索结果.

3 实验结果与性能分析

3.1 实验设置与性能评价

本文在 ImageNet-1000[23]、和 Caltech-256[24]图像集上对本文方法进行了评估,ImageNet-1000 图像集是 ImageNet 图像集的一个子集,是大尺度视觉识别竞赛(Large Scale Visual Recognition Challenge,LSVRC)的评测数据集,包含 1000 个类别共计 120 万张图像;Caltech-256 图像集是目标分类任务中常用数据集,包含 256 个类别共计 30608 张图像,其中每个类别中至少包含 80 张图像. 实验硬件配置为内存为 6G 的 GPU 设备 GTX Titan 和 Intel Xeon CPU、内存为 16G 的服务器. 图像检索性能指标采用平均查准率均值(Mean Average Precision,MAP)和查全率(Recall),其定义如下:

$$\mathrm{MAP}=\frac{\text{多次图像检索的平均查准率}}{\text{检索次数}}\times100\% \tag{20}$$

$$\text{查全率}=\frac{\text{检索出的相关图像}}{\text{全部相关图像}}\times100\% \tag{21}$$

3.2 实验结果与分析

为验证监督核哈希方法(简称 KSH)检索性能的优越性,并与当前的一些主流哈希方法在 ImageNet-1000 图像集上进行了实验比较,包括 LSH[13](Locality Sensitive Hashing)、SKLSH[25](LSH with Shift-Invariant Kernels)、SH[16](Spetral Hashing)、DSH[26](Density Sensitive Hashing)、PCA-ITQ[27](Iterative Quantization of PCA)、BRE[28](Binary Reconstructive Embedding)等方法. 实验首先从 ImageNet-1000 图像集中随机选取 50 类,并对这 50 类图像提取 GIST 特征;然后,从每个类中随机选取 1000 张图像的特征(共计 50000 张图像的特征)作为监督训练哈希函数的训练集,其余图像作为查询用例;最后,引入哈希方法进行检索,得到实验结果如图 2 所示.

从图 2 中可以看出,随着哈希码位数 r 的增加,各方法的 MAP 值有所提高,然而,当 r 增加到一定值后,MAP 值增加幅值逐渐变小趋于饱和. 对比各哈希方法

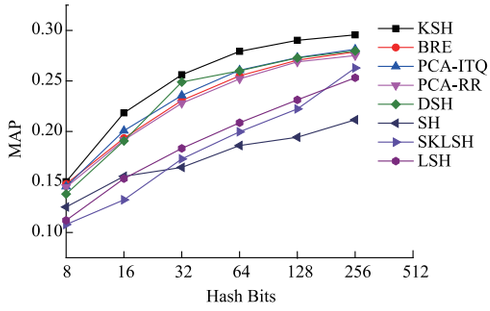


图 2　在 ImageNet-100 数据库上图像检索MAP值

的图像检索 MAP 值可知,采用本文方法(KSH)进行检索较之其它主流方法有更好的性能,这是因为无监督哈希方法(例如 LSH,SH,DSH,PCA-ITQ 等)和监督哈希方法 BRE 没有很好地利用图像的语义信息构造哈希函数,导致检索性能较低;而 KSH 引入核函数构造哈希函数加强了对线性不可分数据的分辨能力,同时结合了图像的相似性信息对哈希函数进行训练,生成更加紧致的哈希码,从而提高了图像检索性能.

实验又将 KSH 与当前主流哈希方法的哈希函数训练时间消耗进行了对比,具体如图 3 所示. 从图 3 中不难看出,采用无监督哈希方法的时间消耗均比监督哈希方法少,无监督哈希方法大都是以保留原始特征的位置敏感性为优化目标,而监督哈希方法是以图像语义近邻信息作为监督信息,其寻优过程更为复杂,相比无监督哈希方法时间开销更大. 对比监督哈希方法 KSH 和 BRE 的训练时间开销可以看出,当哈希码位数较少时,BRE 相比 KSH 训练时间较短,而随着哈希码长度增加,BRE 训练时间快速增长,使得 BRE 训练时间远远超过 KSH. 由于 KSH 以最小化相似图像的哈希码汉明距离同时使得不相似图像的哈希码汉明距离达到最大作为目标函数,并利用哈希码内积与汉明距离的等价关系对目标函数进行简化,引入贪婪算法和梯度下降法加速寻优过程,相比 BRE 最小化编码的重构误差作为优化目标,KSH 寻优过程更加简单.

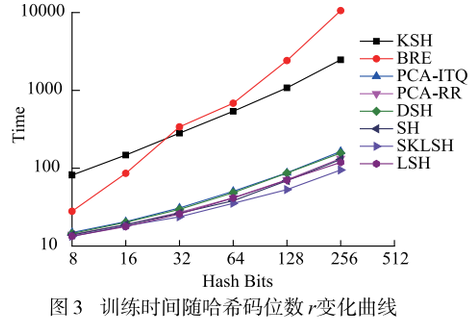


图 3　训练时间随哈希码位数 r 变化曲线

为验证基于卷积神经网络和监督核哈希的图像检索方法(简称 CNN + KSH)的有效性,在 ImageNet-1000 图像集每个类别中随机选取 500 张图像的特征作为监督训练哈希函数的训练集,其余图像作为查询用例进行实验得到表 1. 从表 1 中不难看出,基于 CNN 提取深层特征进行检索的 MAP 值比基于全局 GIST 特征进行检索的 MAP 值高出 10% 以上,说明利用 CNN 提取的图像深层特征具有更强的区分性和表达能力. 这是因为 GIST 特征提取步骤固定,不具有自主学习能力,从而使得其图像表达能力受限,而 CNN 能模仿大脑处理数据的模式对图像进行特征提取,而且其深层的网络结构能有效地挖掘图像内在隐含关系,增强了特征的图像表达能力. 其中,CNN + KSH 的在线检索时间为 1.843 $\times 10^{-4}$ S,与其它主流方法相当,而 MAP 值达到了 38.57%,检索性能明显高于其它方法,因此本文方法在大数据环境下具有较强的适用性.

表 1　不同方法的图像检索 MAP 值和检索时间对比(64bits)

	MAP (%)	Time $\times 10^{-4}$s		MAP (%)	Time $\times 10^{-4}$s
GIST + LSH	17.73	1.351	CNN + LSH	32.27	1.417
GIST + SKLSH	17.12	1.384	CNN + SKLSH	34.05	1.406
GIST + SH	15.64	134.2	CNN + SH	32.16	131.4
GIST + DSH	23.21	1.411	CNN + DSH	34.51	1.452
GIST + PCA-RR	22.44	2.154	CNN + PCA-RR	35.34	2.088
GIST + PCA-ITQ	23.35	1.943	CNN + PCA-ITQ	35.91	1.987
GIST + BRE	22.51	1.897	CNN + BRE	34.85	1.912
GIST + KSH	25.07	1.905	CNN + KSH	38.57	1.843

为进一步验证 CNN + KSH 的有效性,又在 Caltech-256 图像集上进行了实验,并与文献[13]中的 LSH 方法,文献[25]中的 SKLSH 方法,文献[16]中的 SH 方法,文献[26]中的 DSH 方法,文献[27]中的 PCA-ITQ 和 PCA-RR 方法以及文献[28]中的 BRE 方法进行实验对比. 首先,利用已训练好的 CNN 对 Caltech-256 图像集所有图像提取 4096 维图像深层特征;然后,从每个类中随机选取 80 张图像(共计 20480 张图像)的深层特征作为监督训练哈希函数的训练集,其余图像作为测试集,得到各方法图像检索 MAP 曲线与 Precision-Recall 曲线如图 4、图 5 所示.

对比图 4、图 5 可知,CNN + KSH 在等长度哈希码编码下 MAP 值均高于其它主流方法,而且在保证相同查准率的条件下,CNN + KSH 能达到比其它方法更高的查全率. 文献[13,16,25~28]中的方法都是对图像提取 GIST 特征后再引入哈希方法构造索引,虽然 GIST 特征通过多尺度多方向 Gabor 滤波器组对图像滤波后得到了全局结构和空间上下文信息,但是 GIST 特征粒度较为粗糙,而且缺乏自主学习能力,限制了其图像表达能力,并且这些

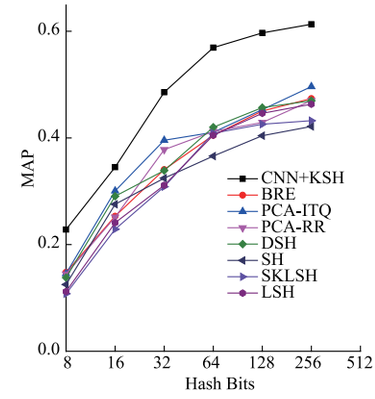


图4 在Caltech-256数据库上图像检索MAP值

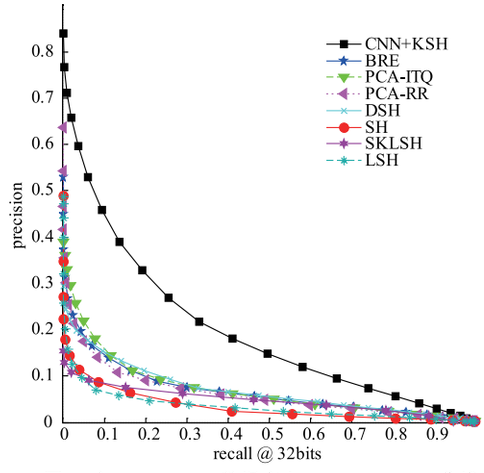


图5 在Caltech-256数据库上Precision-Recall曲线

文献中的哈希方法只是考虑如何结合数据特点构造哈希函数,并没有很好地利用图像的语义信息,导致其图像检索性能较低,难以适应于大规模图像检索;CNN + KSH 方法利用 ImageNet-1000 图像集对 CNN 进行训练,大量的训练图像使得 CNN 模型参数训练比较充分,能够有效挖掘图像内在隐含关系,有效增强了特征的图像表达能力,而且 KSH 利用图像相似性信息对哈希函数进行监督训练,同时采用贪婪算法和梯度下降法加速寻优过程,使得 CNN + KSH 优于当前主流方法. 实验结果也证明了本文方法(CNN + KSH)在 Caltech-256 图像集上的检索性能明显优于其它方法.

4 结论

本文提出了一种基于卷积神经网络和监督核哈希的图像检索方法. 首先,针对传统图像特征表达能力差、适应性不强等问题,引入卷积神经网络逐层地对训练图像进行学习,利用其特殊网络结构有效挖掘训练图像内容的内在隐含关系,得到图像数据的高阶表示,增强了图像特征的区分性和表达能力;然后,引入监督核哈希方法有效克服了高维图像数据检索性能低问题,并结合图像相似性信息提出简单、有效的目标函数,降低寻优过程复杂度,实现了对大规模图像集的高效检索. 实验结果有效地验证了本文方法的图像检索性能优于当前主流方法. 如何加快卷积神经网络的训练速度,有效防止出现过拟合现象是本文的下一步研究方向. 此外,如何更加有效利用标注数据的标注信息,降低优化过程复杂度也是今后亟待解决的问题.

参考文献

[1] Oliva A, Torralba A. Modeling the shape of the scene a holistic representation of the spatial envelope [J]. International Journal in Computer Vision, 2001, 42(3): 145 – 175.

[2] Lowe D. Distinctive image features from scale-invariant keypoints [J]. Int J Comput Vis, 2004, 60(2): 91 – 110.

[3] Bay H, Ess A, Tuytelaars T, et al. Speeded-up robust features [J]. Computer Vision and Image Understanding, 2008, 110(3): 346 – 359.

[4] Hinton G E, Osindero S, The Y. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7): 1527 – 1539.

[5] Bengio Y, Lamblin P, Popovici D, et al. Greedy layer-wise training of deep networks[A]. Proceedings of Advances in Neural Information Processing Systems [C]. Vancouver, Canada: ACM, 2007. 154 – 156.

[6] Hinton G E. To recognize shapes, first learn to generate images[J]. Progress in Brain Research, 2007, 165(6): 539 – 547.

[7] Tang Yi-chuan, Eliasmith C. Deep networks for robust visual recognition [A]. Proceedings of the 27th International Conference on Machine Learning[C]. Haifa, Israel: ACM, 2010. 1055 – 1062.

[8] Lee H, Grosse R, Ranganath R, et al. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations[A]. Proceedings of the 26th International Conference on Machine Learning [C]. New York: ACM, 2009. 609 – 616.

[9] Huang G B, Lee H, Learned-Miller E. Learning hierarchical representations for face verification with convolutional deep belief networks[A]. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition[C]. Providence, USA: IEEE, 2012. 2518 – 2525.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904 – 1916.

[11] Gong Jun. A new implementation mechanics of database for integrating R-tree and levels of detail[A]. Proceedings

of International Conference on Audio Language and Image Processing[C]. Shanghai, China: IEEE, 2010. 484 – 488.

[12] Liu Qiang, Huang Hao, Wang Yongmin, et al. The KD-tree-based nearest-neighbor search algorithm in GRID interpolation[A]. Proceedings of International Conference on Image Analysis and Signal Processing[C]. Hangzhou, China: IEEE, 2012. 1 – 6.

[13] Indyk P, Motwani R. Approximate nearest neighbors: towards removing the curse of dimensionality[A]. Proceedings of the Symposium on Theory of Computing[C]. Dallas, Texas, USA: ACM, 1998. 604 – 613.

[14] Brian Kulis, Kristen Grauman. Kernelized locality-sensitive Hashing[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(6): 1092 – 1104.

[15] Datar M, Immorlica N, Indyk P, et al. Locality sensitive hashing scheme based on p-stable distributions [A]. Proceedings of the ACM Symposium on Computational Geometry[C]. New York, USA: ACM, 2004. 253 – 262.

[16] Yair Weiss, Antonio Torralba, Rob Fergus. Spectral Hashing[A]. Proceedings of Neural Information Processing Systems [C]. Vancouver, Canada: ACM, 2008. 1753 – 1760.

[17] Wang J, Kumar S, Chang SF. Semi-supervised Hashing for large scale search[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(12): 2393 – 2406.

[18] Fukushima K. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position[J]. Biological Cybernetics, 1980, 36(4): 193 – 202.

[19] Alex Krizhevsky, Ilya Sutskever, Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks [A] Proceedings of Advances in Neural Information Processing Systems[C]. Lake Tahoe, Nevada, USA: ACM, 2012. 1106 – 1114.

[20] Nair V, Hinton G E. Rectified linear units improve restricted boltzmann machines[A]. Proceedings of International Conference on Machine Learning[C]. Haifa, Israel: ACM, 2010. 807 – 814.

[21] Wei Liu, Jun Wang, Rongrong Ji, et al. Supervised Hashing with kernels[A]. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition[C]. Providence, USA: IEEE, 2012. 2074 – 2081.

[22] Nesterov Y. Introductory Lectures on Convex Optimization: A Basic Course [M]. Kluwer Academic Publishers, 2003.

[23] Jia Deng, Wei Dong, Richard Socher, et al. ImageNet: a large-scale hierarchical image database[A]. Proceedings of Computer Vision and Pattern Recognition[C]. Miami, Florida, USA: IEEE, 2009. 248 – 255.

[24] Li F F, Fergus R, Perona P. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories[J]. Computer Vision and Image Understanding, 2007, 106(1): 59 – 70.

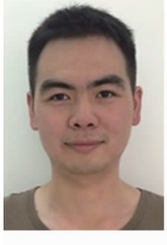
[25] Raginsky M, Lazebnik S. Locality sensitive binary codes from shift-invariant kernels[A]. Proceedings of Neural Information Processing System[C]. New York, USA: ACM, 2009. 1509 – 1517.

[26] Zhongming Jin, Cheng Li, Yue Lin, et al. Density sensitive Hashing[J]. IEEE Transactions on Cybernetics, 2014, 44(8): 1362 – 1371.

[27] Yunchao Gong, Svetlana Lazebnik, Albert Gordo, et al. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2916 – 2929.

[28] Kulis B, Darrell T. Learning to hash with binary reconstructive embeddings[A]. Proceedings of Neural Information Processing System[C]. Vancouver, Canada: ACM, 2009. 1042 – 1050.

作者简介



柯圣财　男,1991 年生于湖北黄石市. 2013 年毕业于解放军信息工程大学. 现为解放军信息工程大学硕士研究生,主要从事图像分析与处理方面的研究.
E-mail: keshengcai0705@163.com

赵永威　男,1988 年 1 月生于河南省周口市. 2016 年毕业于解放军信息工程大学获博士学位,现为武警工程大学讲师,主要从事视频图像分析及处理方面的研究.
E-mail: zhaoyongwei369@163.com

李弼程(通信作者)　男,1970 年生于湖南省衡阳市. 教授、博士生导师,1998 年毕业于国防科技大学获博士学位. 现为解放军信息工程大学教授. 主要从事智能信息处理方面研究.
E-mail: lbclm@163.com

彭天强　男,1978 年生于湖北省随州市. 2008 年毕业于解放军信息工程大学获博士学位,现为河南工程学院副教授. 主要从事智能信息处理、模式识别方面研究.