DOI: 10.7641/CTA.2015.50063

基于深度网络的可学习感受野算法在图像分类中的应用

王　博, 郭继昌†, 张　艳

(天津大学 电子信息工程学院, 天津 300072)

摘要: 作为图像检索, 图像组织和机器人视觉的基本任务, 图像分类在计算机视觉和机器学习中受到了广泛的关注. 用于目标识别及图像分类的多种基于深度学习的模型同样引发了该领域内的极大兴趣. 本文提出了一种取代尺度不变特征变换(SIFT)和方向梯度直方图(HOG)描述子的算法, 即利用深度分层结构, 按层级学习有效的图像表示, 直接从原始像素点学习特征.该方法分别利用K–奇异值分解(K–SVD)和正交匹配追踪(OMP)进行字典训练和编码.此外, 本文采用了同时学习分类器和用于池化的感受野方案. 实验结果证明, 上述算法在目标(Oxford flowers)和事件(UIUC–sports)图像分类测试集中取得了更好的分类性能.

关键词: 图像分类; 分层结构; 深度网络; 感受野

中图分类号: TP391.4　　**文献标识码**: A

Learnable receptive fields scheme in deep networks for image categorization

WANG Bo, GUO Ji-chang†, ZHANG Yan

(School of Electronic Information Engineering, Tianjin University, Tianjin 300072, China)

Abstract: An increasing interest in computer vision and machine learning has focused on visual categorization as it is a fundamental task for image retrieval, organization and robotic vision. Over the past decade, various deep learning-based models have been proposed and broadly applied to visual recognition and categorization. In this paper, the proposed approach learns features from scratch rather than employ hand-crafted (SIFT) and (HOG) descriptors. Deep hierarchical architecture for learning effective image representations can be built up layer by layer. Specifically, K–SVD and OMP are used for training and encoding phase respectively due to their simplicity and efficiency. In addition, sum, average and max operators are three commonly strategies for pooling in modern categorization models. We aim to apply an improved scheme which learns the receptive fields for pooling together with classifier instead of traditional pooling pattern. We provide a detailed analysis in deep networks for event and object tasks respectively and compare our novel method with several state-of-the-art algorithms comprising kernel-based feature learning and saliency-weighted hierarchical sparse coding. Finally, experimental results show that our algorithm performs better on UIUC–sports and Oxford flowers datasets.

Key words: image categorization; hierarchical architecture; deep networks; receptive fields

1　引言(Introduction)

在识别和分类系统中, 利用现代计算机视觉和机器学习方法, 从观测样本里获得具有判别力的图像表示是非常重要的步骤. 此前的许多方法都利用基于局部图像块的HOG[1]和SIFT[2]描述子进行编码以得到良好的图像表示. 常见的K-means[3]和矢量量化(VQ)可以生成更高级的特征. 近些年, 很多研究都聚焦于训练分层的深度网络, 其中包括用于字典学习的K–SVD[4]和替代矢量量化(VQ)的稀疏编码[5].

上述方法在常用数据测试集中取得了较好的效果, 但在实践中, 依然希望避免对描述子产生严重的依赖, 即可以通过完全自动的像素级方法进行特征学习. 文献[6]证明了基于像素级的分层稀疏编码可以获得与基于SIFT描述子的稀疏编码类似的性能. 文献[7]提出了由稀疏编码, 显著性池化和局部分组组成的分层模型, 结果表明该模型的分类性能优良. 文献[8]论证了相对于卷积深信度网络, 基于SIFT描述子的单层稀疏编码及基于核函数的特征学习算法, 分层匹配追踪算法在3种不同类型的图像分类测试集中均具有最优表现.

与关注编码阶段相比, 极少有关于池化策略的研究, 但池化方法在现代视觉识别和分类任务中是不可或缺的步骤. 特征池化的观点源自于胡贝尔对视觉皮层中复杂细胞的研究, 证实了基于局部邻域的中级特

收稿日期: 2015−01−22; 录用日期: 2015−07−15.
†通信作者. E-mail: jcguo@tju.edu.cn.
高等学校博士学科点专项科研基金项目(20120032110034)资助.
Supported by Specialized Research Fund for the Doctoral Program of Higher Education (20120032110034).

征对于适度空间形变具有不变性. 应用在不同系统中的典型池化方案包括: 求平均(average), 求和(sum)以及取最大值(max)运算. 文献[9]提出了一种称为局部约束线性编码的方法, 该方法通过利用取最大值池化获得了较好的分类性能. 文献[10]系统地分析对比了求平均和取最大值池化策略在目标分类中的作用. 然而, 如何学习或设计更好的空间池化方法在近些年很少受到关注, 尤其是在深度网络中. 文献[11]提出了在深度学习背景下对局部感受野的选择, 同时证明了该算法适用于生成大型的特征表示. 文献[12]则创新地采用了与分类器同时训练的更加灵活的全局参数化池化方案. 但因其需要与文献[13]进行对照, 所以该算法仅针对单层网络进行了实验.

本文将提出一种基于两层深度网络的可学习感受野方法. 首先, 本文利用RGB图像类型去提取用于生成有效图像表示的块级特征, 在第1层中统一采用取最大值池化, 而在第2层的预池化阶段中分别采用求平均, 求和及取最大值池化运算方案. 其次, 本文将比较不同类型的预池化方案和其它参数在图像测试数据集中发挥的作用.

2　深度网络中的分层特征提取(Hierarchical feature extraction in deep networks)

2.1　分层的特征提取(Hierarchical feature extraction)

在计算机视觉领域中, 通过未标记的输入数据学习良好的表示以用于高级图像分类任务, 尤其是取代传统描述子算法的方案引起了越来越多的关注. 其中, 利用正交匹配追踪(OMP)的分层结构, 借助贪婪的训练方式可学习多层特征, 且每次只进行一层学习. 然而, 实践中, 批量树正交匹配追踪(BTOMP)总是因为更高效而取代OMP. 正如此前在文献[11]和文献[8]中描述的那样, 深度网络的应用主要包含以下4步. 假定图像是由像素点组成, 流程图如图1所示.

第1层特征提取时, 采用$m \times m$尺寸的感受野, 其间隔设定为1. 通过训练可以得到含有D_1个滤波器的字典. 接着, 利用BTOMP算法获得了形如$(n-m+1)\times(n-m+1)\times D_1$的图像表示.

在邻近的$s \times s$空间块中, 采用了取最大值的池化策略, 然后生成了形如$[(n-m+1)/s]\times[(n-m+11)/s]\times D_1$ 的池化表示.

在所有D_1个层上, 利用$j \times j$尺寸的感受野, 其间隔设定为1, 生成的第2层特征维度是$j \times j \times D_1$, 相应特征量为$\{[(n-m+1)/s]-j+1\}\times\{[(n-m+1)/s]\times D_1-j+1\}$. 通过BTOMP算法得到了形如$\{[(n-m+1)/s]-j+1\}\times\{[(n-m+1)/s]\times D_1-j+1\}\times D_2$的图像表示.

在最后的预训练阶段, 分别采用求平均, 求和与取最大值的池化方法得到最终的图像表示.

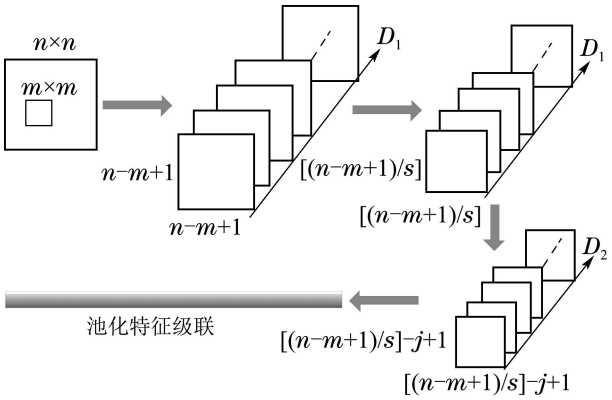


图 1　分层特征提取结构图

Fig. 1　Architecture of hierarchical feature extraction

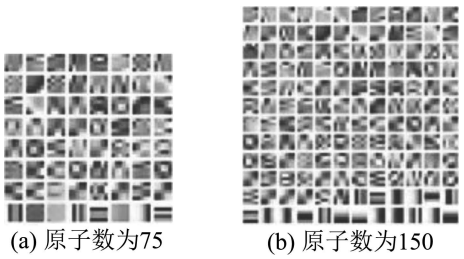
2.2　用于重构和稀疏编码的字典学习(Dictionary learning for reconstruction and sparse coding)

令X为一组d维的信号集合, 即$X=[x_1\ \cdots\ x_N]\in\mathbb{R}^{d\times N}$. 为得到关于$X$的稀疏表示, 可通过以下优化问题, 学习带有$P$个原子的可重构字典:

$$\min_{D,S}\|X-DS\|_{\mathrm{F}}^2\ \ \text{s.t.}\ \forall i,\|s_i\|_0\leqslant T_0,\qquad(1)$$

其中: $D=[d_1\ \cdots\ d_P]\in\mathbb{R}^{d\times P}$为学习的字典, $S=[s_1\ \cdots\ s_N]\in\mathbb{R}^{P\times N}$为$X$的稀疏编码, T_0 表示稀疏度, 即非零元素的数目. 字典学习始于给定的输入信号, 目的是同时求得字典D和稀疏表示S. K-SVD算法正适用于解决以上问题, 其可通过迭代方法优化式(1)的能量, 并学习到一个可重构的用于稀疏表示的字典. 给定字典D后, 正交匹配追踪(OMP)算法可利用贪婪的方式计算近似解, 其具体问题如下式:

$$\min_{s_i}\|x_i-Ds_i\|^2\ \ \text{s.t.}\ \forall i,\|s_i\|_0\leqslant T_0.\qquad(2)$$



(a) 原子数为75　　(b) 原子数为150

图 2　第1层训练中原子数分别为75和150的学习字典

Fig. 2　The dictionaries learned with 75 and 150 atoms in the first layer

在某些情况下, 当需要学习一个大规模的字典时, OMP的运算效率将无法满足要求. 然而, 改进后的BTOMP[8]可利用高效的树形结构学习字典. 因此, 从实时运算的角度出发, 在预定义了稀疏度T_0后, BTOMP可以相对更好地执行编码任务. 通过K-SVD算法学习得到的字典如图2所示, 其中图(a)表示

原子数为75的字典, 图(b)表示原子数为150的字典. 以上字典学习均来自于UIUC-sports测试集.

3 可学习的感受野(Learnable receptive fields)

由于在图像分类领域中取得了优异的性能, 空间金字塔匹配算法(spatial pyramid matching, SPM)至今依然受到广泛的关注. 根据SPM算法, 一幅图像首先被划分为不同的子区域, 然后分别计算每一个子区域的码词统计直方图. 最后, 通过将之前得到的统计直方图进行级联, 得到池化的特征. 事实上, 判断某个定义感受野的方式是否合理, 该问题在现存方案中似乎一直没有得到重视. 通常, 求平均和取最大值的池化操作方式可分别由以下式(3)和(4)定义:

$$f(v)=\frac{1}{T}\sum_{m=1}^{\mathrm{T}} v_m, \tag{3}$$

$$f(v)=\max_{1\leqslant m\leqslant T} v_m, \tag{4}$$

其中: v_m表示提取自图像的T个编码块中的一块, m代表了已提取块的空间方位. 池化步骤利用以上定义的空间池化算子f将v_m映射为相应的统计值, 这一过程理论上容易解释, 但却是获得良好分类性能必不可少的阶段. 直观上分析, 一种基于参数化的池化算子能够解决独立于数据本身的池化区域划分问题. 可以首先考虑将α_m作为池化权重, 接着将图像分隔为两个子区域. 例如, 取最大值池化问题可如下表示:

$$\max_{1\leqslant m\leqslant \frac{T}{2}} 1\circ v_m+\max_{\frac{T}{2}+1\leqslant m\leqslant T} 0\circ v_m, \tag{5}$$

$$\max_{1\leqslant m\leqslant \frac{T}{2}} 0\circ v_m+\max_{\frac{T}{2}+1\leqslant m\leqslant T} 1\circ v_m. \tag{6}$$

其中式(5)和(6)分别代表对第1和第2子区域进行的相应运算. 由于需要同时学习α和分类器的参数, 因此可以采用反向传播算法训练稠密联通的多层感知器. 所以, 每一个编码的坐标都可看作是多层感知器的一个输入值. 假设有K层用于编码训练, 那么第l个池化单元u_l^k可与第k层的第m个输入单元相关联, 利用矢量描述如下所示:

$$u_l:=f_{m=1}^{T}(\alpha_m^l\circ v_m)=\Theta_{\alpha^l}(V). \tag{7}$$

假定$V^{(i)}$属于第i幅图像且$u_l^{(i)}:=\Theta_{\alpha^l}(V^{(i)})$, 那么$u^{(i)}$就表示栈式池化单元对相应图像的响应. 下一步, 可以直接对相关联的池化单元使用softmax回归, 那么代价函数可记为:

$$J(\Theta):=-\frac{1}{D}\sum_{i=1}^{D}\sum_{j=1}^{C}1\{y^{(i)}=j\}\log\frac{\mathrm{e}^{\Theta_j^T u^{(i)}}}{\sum\limits_{r=1}^{C}\mathrm{e}^{\Theta_r^T u^{(i)}}}, \tag{8}$$

其中: $y^{(i)}$代表第i幅输入图像的标签, C代表图像的类别, D代表所有图像的数量.

最后, 利用梯度下降算法迭代计算出包含分类器参数Θ和池化权重矩阵A的参数H. 在实践中, 一般引入一个正则项以防止出现过度拟合. 另外, 由于池化区域的非平滑性, 可以通过惩罚权重对平滑性进行约束. 为此, l_2正则项$\sum\limits_k\left\|A^k\right\|^2$和$\|\Theta\|^2$, 以及空间变化测度项$\|\nabla_x A\|_{\mathrm{F}}^2+\|\nabla_y A\|_{\mathrm{F}}^2$均加入到代价函数中. 然而, 根据文献[12]的结论, 仅使用平滑性正则项可以取得更高的分类精度, 所以l_2正则项无需引入到最终的目标函数中. 综上所述, 目标函数可记为:

$$\begin{aligned}&\min_{A,\Theta}J(A,\Theta):=\\&-\frac{1}{D}\sum_{i=1}^{D}\sum_{j=1}^{C}1\{y^{(i)}=j\}\log\frac{\mathrm{e}^{\Theta_j^T u^{(i)}}}{\sum\limits_{r=1}^{C}\mathrm{e}^{\Theta_r^T u^{(i)}}}+\frac{\lambda_1}{2}\|\Theta\|^2+\\&\frac{\lambda_2}{2}\left(\|\nabla_x A\|_{\mathrm{F}}^2+\|\nabla_y A\|_{\mathrm{F}}^2\right)\ \text{s.t.}\ A\in[0,1]^{K\times M\times L}.\end{aligned} \tag{9}$$

其中$\|\cdot\|_{\mathrm{F}}$代表弗罗贝尼乌斯范数, 这组约束条件实质是通过在单位立方体上的投影到来保持池化区域的可判读性.

因此, 本文通过基于深度网络的可学习感受野构建图像级表示的算法框架如图3所示.

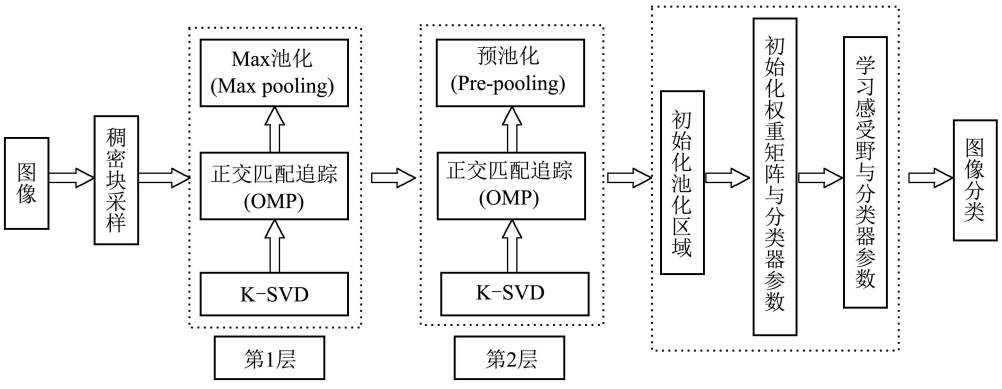


图 3 基于深度网络的可学习感受野算法框架

Fig. 3 An algorithm framework of learnable receptive fields scheme using deep networks

4　时间复杂度分析(Time complexity analysis)

根据图3的算法框架提示, 完整的算法实现过程主要是由分层特征提取和模型参数学习组成的, 这两部分基本控制了算法的运算时间. 首先, 利用BTOMP进行稀疏编码是特征提取中最关键的部分, 其时间复杂度可以分为预计算和单一信号计算两部分. 参照第2.2节中对信号, 字典以及稀疏度的定义, 用$T_{\rm p}$表示预计算时间, $T_{\rm s}$表示单一信号计算时间, 则$T_{\rm p}=dP^2$, $T_{\rm s}=2dP+T_0^2P+3T_0P+T_0^3$. 若不考虑高效稀疏编码, 则直接利用OMP算法, 其时间复杂度为$T_{\rm omp}=2dPT_0+2dT_0^2+2T_0(P+d)+T_0^3$. 通常, 针对超完备的K-SVD字典, 假设$T_0=\dfrac{\sqrt{d}}{2}$且$P=2d$, num为用于计算的信号的数量, $d\ll$ num, 稀疏编码时间复杂度对比如表1所示.

表 1　稀疏编码时间复杂度对比

Table 1　Comparison of time complexity for sparse coding

$T_{\rm BTOMP}\approx$	$T_{\rm OMP}\approx$
$4d^3+\text{num}\times(4.5d^2)$	$\text{num}\times 2d^{2.5}$
K-SVD字典: $d\times P$, 稀疏度: T_0	

因此, 结合本文采用稠密块采样时, 对于200 × 200的图像, 感受野设定为6 × 6, 那么$d=108$, num = 34225. 此时, $T_{\rm OMP}\approx 7d^{4.5}$, 而$T_{\rm BTOMP}\approx 15.8d^4$, 所以利用基于BTOMP的编码方式体现了更高效的编码效率. 其次, 在分类器参数优化时, 分别设定图像类别为class, 隐层数为hiddennum, 池化区域数为poolingregion, 采样数量为samplenum, 那么参数Θ可记为

$$\Theta=(\text{classnum}\times\text{hiddennum}\times\text{poolingregion})+(\text{samplenum}\times\text{hiddennum}\times\text{poolingregion}),$$

所以, 用于训练的隐层数量将会直接决定参数Θ的规模. 例如, UIUC-sports数据集共有8种类别, 采样数量和池化区域数均设定为4, 则$\Theta=48$ hiddennum. 此时, 用于训练的隐层数由第2层字典的原子数目决定.

5　实验结果与分析(Experiment results and analysis)

为了验证以上提出的算法, 本文利用两个目前广泛使用的图像分类基准测试库作为标准, 即UIUC-sports和oxford flowers. 针对两个数据集, 首先, 本文算法均只使用基于RGB 图像的块级特征提取方案,即稠密地进行块采样. 其次, 实验中发现, 仅针对一层网络进行学习, 分类性能相对较低. 如果进行三层深度网络训练, 时间开销较大, 且分类性能并没有显著提升. 因此, 最终采用两层的深度网络学习方案.

文献[3]和文献[8]中分别对影响单层和多层网络性能的因素进行了详尽的分析, 所以本次实验中的网络参数设置可参照其中对于原子数量, 采样间隔, 感受野尺寸的讨论结果. 其中, 图像块的采样间隔设定为1时, 分类准确率最优, 随着间隔的增加, 性能递减, 因此以下两组实验的图像块采样间隔均统一设定为1. 基于超完备字典的稀疏表示除了对噪声有更好的鲁棒性外, 还可显著提高图像分类性能, 所以本文设定的原子数量需大于采样特征维度. 理论上, 大尺寸的感受野可有助于识别更复杂的特征, 但是同时也会增加数据特征的维度. 通常这个尺寸选定为5 × 5或6 × 6以得到更好的分类性能. 关于池化类型对分类性能的影响, 文献[2]描述max池化算子使得目标类别数据库分类性能更优, 而使用规模较大的字典时, average和max池化算子对于场景类别数据库的分类性能基本相同. 所以本文选择使用3种不同类型算子作为预池化的方案, 分别用于验证针对不同类型数据库分类性能的影响.

在实验中, 所有的图像尺寸都被调整为200 × 200, 每幅图像第1层特征平均提取时间为0.6 s, 第2层特征提取平均时间约为0.5 s. 利用K-SVD训练字典时, 第1层与第2层迭代次数分别设置为50和10, 全部训练时间约为20 min. 模型参数的学习使用L-BFGS算法, 迭代次数设定为500. 实验平台选择Windows 7, 64位操作系统, 主频为3.0 GHz的Intel i5处理器.

5.1　UIUC-sports分类(UIUC-sports categorization)

UIUC-sports数据库由文献[14]提出作为基准测试集, 其可被视为典型的事件类数据集. 该图像库由8个不同运动类别组成, 例如: 室外地滚球、马球、攀岩、单板滑雪等项目. 每类中包括137∼250幅图像, 全部图像数量合计为1579. 因为该图像库带有复杂的背景, 且每类图像的大小以及内容变化较大, 所以利用该库测试分类准确率具有一定的挑战性. 为了保持公平的测试条件, 根据常规的实验设定, 本文随机从每类中抽取70幅用作训练图像, 60幅用作测试图像. 同时, 为比较不同的池化算子, 本文在预池化阶段分别使用了求平均, 求和以及取最大值的操作.

首先, 实验确定了第1和第2层中字典的尺寸大小, 即利用5 × 5的块提取特征以训练具有150个原子的字典, 并且将OMP稀疏度设定为4. 在第2层中, 选择3×3的块提取特征以训练具有1600原子的字典, 但将OMP稀疏度设定为10. 其中有3类不同的池化算子用于测试分类准确率. 最终的实验结果表明, 在第1组实验中, 预池化阶段利用求平均算子可以取得最优的分类准确率(见表2). 所以, 第2组实验将预池化类型固定为求平均, 而第1层池化类型始终保持取最大值方法不变. 关于池化类型的不同组合对分类性能的影响, 如图4所示.

表 2 基于3种预池化算子的UIUC-sports图像集分类准确率

Table 2 Classification accuracy on UIUC-sports dataset (3 pre-pooling opertations)

预池化类型	第1层字典大小	第2层字典大小	分类准确率/%
average	150	1600	76.9
sum	150	1600	76.1
max	150	1600	74.6

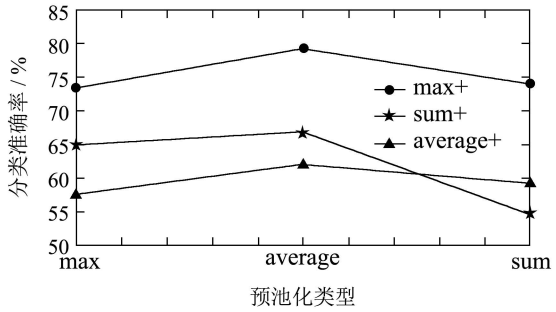


图 4 不同池化方式组合下分类准确率的比较(UIUC-sports)

Fig. 4 Classification accuracy comparison based on types of different combinations of pooling for UIUC-sports in the first and pre-pooling stage

特别注意的是, K-SVD利用一个超完备的离散余弦变换对字典进行初始化操作, 因此, 在第1层训练中, 本文设置的原子数量大于或等于两倍的字典维度. 然后, 在减小并固定第2层字典大小设置的基础上, 逐渐增加第1层字典的大小以比较分类准确率(见表3). 第2组实验结果表明, 当采用本文算法, 将第1层字典原子数量设定为其大小2倍, 选择池化类型分别是取最大值(max)和求平均(average)时, 获得了关于UIUC- sports 测试集的最优分类准确率79.2%. 相反, 对比表2和表3中第2层字典, 证明其大小设定为1000时, 已可获得最优分类准确率, 若继续增大, 分类准确率会出现下降. 原因主要是训练样本的数量比较有限, 因此在第2层网络中若采用超完备字典, 在具有冗余度的情况下会出现过度拟合, 从而直接导致了分类性能的下降.

表 3 UIUC-sports图像集分类准确率(70张训练图像)

Table 3 Classification accuracy on UIUC-sports dataset (70 images for training)

第1层池化+预池化类型	第1层字典大小	第2层字典大小	分类准确率/%
max + average	75	1000	71.8
max + average	150	1000	79.2

经过10次实验后取得平均分类准确率如表4所示. 通过对比发现, 本文的算法可以在分类准确率方面大幅超越基于SIFT描述子的生成图形化模型(GGM)算法及基于目标(object bank) 的方法.

表 4 UIUC-sports图像集分类准确率对比

Table 4 Classification accuracy comparison on UIUC-sports dataset

对比算法	分类准确率/%
SIFT + GGM[14]	73.4
Object bank[15]	76.3
本文方法	79.2

5.2 Oxford flowers分 类(Oxford flowers categorization)

Oxford flowers测试库包含了1360幅图像, 总共17个不同类型, 每类中有80幅. 在这个关于植物的基准测试库中, 类内差异有时甚至大于类间差异, 而且存在两个不同类型花种间极具相似度的情况, 因此该测试集具有一定挑战性. 为得到公平的评价, 本文与此前的实验设置保持一致. 根据文献[7], 分别随机抽取40和60幅用于训练的图像. 本文进行10次实验, 然后通过取平均, 得到最终的分类准确率.

在第1层中, 用于训练字典的特征取自6×6的图像块, 原子数设定为200, OMP稀疏度设置为5. 针对第2层, 依然取6×6的块用作字典训练, 并将原子数定为3200, OMP稀疏度为10, 与前一项实验保持一致. 如表5所示, 在预池化方面, 取最大值的池化方法得到了明显最优的性能, 而该组实验中, 求平均的方法性能此时严重下降. 关于不同池化类型对分类准确率的影响对比如图5所示.

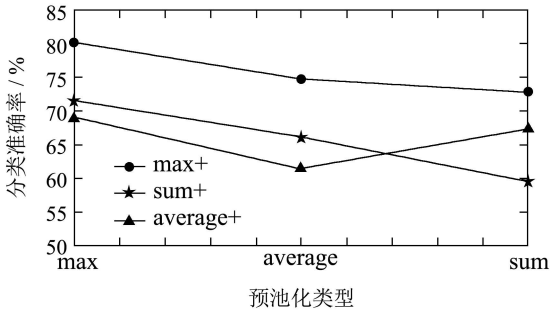


图 5 不同池化方式组合下分类准确率的比较(oxford flowers)

Fig. 5 Classification accuracy comparison based on types of different combinations of pooling for oxford flowers in the first and pre-pooling stage

此外, 从表6中可以发现, 第2层字典大小设定为1600时, 分类准确率已可以达到最优值, 若固定第1层字典原子数为200, 继续增加第2层字典大小, 会引起准确率的严重下降. 其原因同前一组实验类似, 当训练样本的数量十分有限时, 在第2层网络采用带有冗余度的字典会导致过度拟合的发生. 同时, 第1层字典大小设置为其维度的3倍时, 得到最佳分类准确率. 在这两组实验中, 训练的图像数量均为60.

表 5 基于3种预池化算子的oxford flowers图像集分类准确率

Table 5 Classification accuracy on oxford flowers dataset (3 pre-pooling opertations)

预池化类型	第1层字典大小	第2层字典大小	分类准确率/%
average	200	3200	72.4
sum	200	3200	70.6
max	200	3200	76.2

表 6 Oxford flowers图像集分类准确率(60张训练图像)

Table 6 Classification accuracy on oxford flowers dataset (60 images for training)

第1层池化+预池化类型	第1层字典大小	第2层字典大小	分类准确率/%
max + max	100	1600	70.1
max + max	200	1600	76.5
max + max	300	1600	80.2

与其他算法的对比实验结果由表7所示, 当随机抽取40幅图像用于训练集时, 本文算法性能优于利用稀疏性, 显著性和局部性的分层模型(HSSL). 同时, 相对于文献[16]提出的基于形状、颜色和纹理描述子的方法, 本文算法大幅提高了性能. 当用于训练集的图像数目增加到60时, 本文算法性能依然优于HSSL, 且高于文献[17]中提出的基于颜色的可视化字典算法6.5%.

表 7 Oxford flowers图像集分类准确率对比

Table 7 Classification accuracy comparison on oxford flowers dataset

训练图像40	本文方法	HSSL[7]	文献[16]
分类准确率/%	74.1	69.7±2.7	Color shape texture 59.7±2.0, 68.9±2.0, 59.0±2.1
训练图像60	本文方法	HSSL[7]	文献[17]
分类准确率/%	80.2	76.2±3.8	Color shape texture 73.7, 71.8, 55.5

6 结论(Conclusions)

本文提出了联合可学习感受野的深度网络算法, 并将其用于图像分类任务. 通过在目标函数中引入控制平滑性的正则项, 同时学习分类器和用于池化的感受野, 该算法在目标和事件图像测试集中取得了良好的分类准确率. 实验分析证明, 在基于本文算法的深度网络中, 不仅是字典大小、编码稀疏度、感受野大小等大量参数, 预池化类型同样在很大程度上影响了最终的分类准确率. 针对目标和事件图像库, 分别采用求平均(average)和取最大值(max)的池化算子可获得较好的分类性能.

参考文献(Reference):

[1] PANG Y W, YUAN Y, LI X L, et al. Efficient HOG human detection [J]. *Signal Processing*, 2011, 91(4): 773 – 781.

[2] BOUREAU Y L, BACH F, LECUN Y, et al. Learning mid-level features for recognition [C] //*Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. San Francisco: IEEE, 2010: 2559 – 2566.

[3] COATES A, LEE H, NG A Y. An analysis of single-layer networks in unsupervised feature learning [C] //*Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*. Fort Lauderdale: JMLR, 2011: 1 – 9.

[4] LI Q, ZHANG H G, GUO J, et al. Reference-based scheme combined with K-SVD for scene image categorization [J]. *IEEE Signal Processing Letters*, 2013, 20(1): 67 – 70.

[5] YANG J C, YU K, GONG Y H, et al. Linear spatial pyramid matching using sparse coding for image classification [C] //*Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. Miami: IEEE, 2009: 1794 – 1801.

[6] YU K, LIN Y Q, LAFFERTY J. Learning image representations from the pixel level via hierarchical Sparse coding [C] //*Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. Providence: IEEE, 2011: 1713 – 1720.

[7] YANG J M, YANG M H. Learning hierarchical image representation with sparsity, saliency and locality [C] //*Proceedings of the British Machine Vision Conference*. Dundee: BMVA, 2011: 19.1 – 19.11.

[8] BO L F, REN X F, FOX D. Hierarchical matching pursuit for image classification: architecture and fast algorithms [C] //*Advances in Neural Information Processing Systems 24*. Granada: NIPS Foundation, 2011: 1 – 9.

[9] WANG J J, YANG J C, YU K, et al. Locality-constrained linear coding for image classification [C] //*Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. San Francisco: IEEE, 2010: 3360 – 3367.

[10] BOUREAU Y L, PONCE J, LECUN Y. A theoretical analysis of feature pooling in visual recognition [C] //*Proceedings of the 27th International Conference on Machine Learning*. Haifa: IMLS, 2010: 111 – 118.

[11] COATES A, NG A Y. Selecting receptive fields in deep networks [C] //*Advances in Neural Information Processing Systems*. Granada: NIPS Foundation, 2011: 1 – 9.

[12] MALINOWSKI M, FRITZ M. Learning smooth pooling regions for visual recognition [C] //*Proceedings of the British Machine Vision Conference*. Bristol: BMVA, 2013: 118.1 – 118.11.

[13] COATES A, NG A Y. The importance of encoding versus training with sparse coding and vector quantization [C] //*Proceedings of the 28th International Conference on Machine Learning*. Bellevue: IMLS, 2011: 921 – 928.

[14] LI L J, LI F F. What, where and who? Classifying events by scene and object recognition [C] //*Proceedings of 11th IEEE International Conference on Computer Vision*. Rio de Janeiro: IEEE, 2007: 1 – 8.

[15] LI L J, SU H, XING E, et al. Object bank: a high-level image representation for scene classification and semantic feature sparsification [C] //*Advances in Neural Information Processing Systems 23*. Vancouver: NIPS Foundation, 2010: 1 – 9.

[16] VARMA M, RAY D. Learning the discriminative power-invariance trade-off [C] //*Proceedings of 11th IEEE International Conference on Computer Vision*. Rio de Janeiro: IEEE, 2007: 1 – 8.

[17] NILSBACK M E, ZISSERMAN A. A visual vocabulary for flower classification [C] //*Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. New York: IEEE, 2006: 1447 – 1454.

作者简介:

王　博　(1984–), 男, 博士, 主要研究方向为模式识别、计算机视觉与机器学习, E-mail: neuwb@tju.edu.cn;

郭继昌　(1966–), 男, 博士, 教授, 博士生导师, 主要研究方向为数字图像处理、模式识别, E-mail: jcguo@tju.edu.cn;

张　艳　(1982–), 女, 博士, 主要研究方向为数字图像处理、计算机视觉及压缩感知, E-mail: yanzhang0910@163.com.