软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2018,29(4):1029−1038 [doi: 10.13328/j.cnki.jos.005404]

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

面向大规模图像分类的深度卷积神经网络优化*

白 琮, 黄 玲, 陈佳楠, 潘 翔, 陈胜勇

(浙江工业大学 计算机科学与技术学院,浙江 杭州 310023)

通讯作者: 白琮, E-mail: congbai@zjut.edu.cn

摘 要: 在图像分类任务中,为了获得更高的分类精度,需要对图像提取不同层次的特征信息.深度学习被越来越多地应用于大规模图像分类任务中.提出了一种基于深度卷积神经网络的、可应用于大规模图像分类的深度学习框架.该框架在经典的深度卷积神经网络 AlexNet 基础上,分别从网络框架和网络内部结构两个方面对网络进行了优化和改进,进一步提升了网络的特征表达能力.同时,通过在全连接层引入隐层,使得网络能够同时具备学习图像特征和二值哈希的功能,从而使该框架具有处理大规模图像数据的能力.通过在 3 个标准数据库中的一系列比对实验,分析了不同优化方法在不同情况下的作用,并证明了所提优化方法的有效性.

关键词: 图像分类;哈希编码;深度卷积神经网络;激活函数;池化

中图法分类号: TP391

中文引用格式: 白琮,黄玲,陈佳楠,潘翔,陈胜勇.面向大规模图像分类的深度卷积神经网络优化.软件学报,2018,29(4):1029−1038. http://www.jos.org.cn/1000-9825/5404.htm

英文引用格式: Bai C, Huang L, Chen JN, Pan X, Chen SY. Optimization of deep convolutional neural network for large scale image classification. Ruan Jian Xue Bao/Journal of Software, 2018,29(4):1029−1038 (in Chinese). http://www.jos.org.cn/1000-9825/5404.htm

Optimization of Deep Convolutional Neural Network for Large Scale Image Classification

BAI Cong, HUANG Ling, CHEN Jia-Nan, PAN Xiang, CHEN Sheng-Yong

(College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract: Features from different levels should be extracted from images for more accurate image classification. Deep learning is used more and more in large scale image classification. This paper proposes a deep learning framework based on deep convolutional neural network that can be applied for the large scale image classification. The proposed framework has modified the framework and the internal structure of the classical deep convolutional neural network AlexNet to improve the feature representation ability of the network. Furthermore, this framework has the ability of learning image features and binary hash simultaneously by introducing the hidden layer in the full-connection layer. The proposal has been validated in showing significance improvement through the serial experiments in three commonly used databases. Lastly, different effects of different optimization methods are analyzed.

Key words: image classification; hash coding; deep conventional neural network; activation function; pooling

图像分类是指利用计算机的特征表达来模拟人类对图像的理解,自动地将图像按照人类能够理解的方式划分到不同的语义空间的技术,其在科学研究、医学应用和工业应用等方面都有广泛的用途.目前,对图像分类

* 基金项目: 国家自然科学基金(61502424, U1509207, 61325019); 浙江省自然科学基金(LY15F020028, LY15F020024, LY18F020032)

Foundation item: National Natural Science Foundation of China (61502424, U1509207, 61325019); Natural Science Foundation of Zhejiang Province, China (LY15F020028, LY15F020024, LY18F020032)

本文由"多媒体大数据处理与分析"专题特约编辑赵耀教授、李波教授、华先胜研究员、文继荣教授、蒋刚毅教授、常冬霞副教授推荐.

收稿时间: 2017-04-28; 修改时间: 2017-06-26; 采用时间: 2017-10-13; jos 在线出版时间: 2017-12-01

CNKI 网络优先出版: 2017-12-04 06:46:49, http://kns.cnki.net/kcms/detail/11.2560.TP.20171204.0646.005.html

的研究主要分为图像特征提取和分类算法研究两部分.尽管传统的图像分类方法,如基于支持向量机(SVM)分类器[1]和视觉词典模型(bag of visual word,简称 BoVW)[2]已在很多数据集上取得了不错的效果,但是仍然存在一个巨大的挑战,即由机器表达出来的底层图像特征和人类所感知的高层语义信息之间存在一个"语义鸿沟".在高层次的图像表达中,这个挑战可以被看成是目前研究的主要挑战,即构建一个能够模拟人类语义理解的计算机工具,而卷积神经网络的出现因其对高层语义特征的强大表达能力,正在试图解决人类与机器之间的"语义鸿沟".

以卷积神经网络为代表的深度学习技术近些年来已在很多方面取得了重大突破,特别是在计算机视觉领域,如图像分类[3]、目标识别[4]、图像检索[5]等,都取得了很好的效果.LeCun[6]首先成功地实现了采用有监督反向传播网络进行数字识别.8 层的深度卷积网络 AlexNet[3]在 ImageNet 大规模视觉识别挑战 2012(ILSVRC-2012)的分类任务中获得冠军.VGG[4]将卷积网络的深度提高到 19 层,并分别获得 ILSVRC-2014 的定位和分类的第 1 名和第 2 名.GoogleNe[5]提出了 Inception 深层架构,构建了 22 层深层网络,获得了 ILSVRC-2014 的分类冠军.MSRA[7]通过研究线性整流函数,在性能上比 GoogleNet 有了 26%的提升.网络的深度对计算机视觉任务的性能有着很大的影响,但仅线性地增加网络深度会造成梯度消失,这不仅不会提升网络精度,还会降低网络的性能.ResNet[8]引入了残差网络结构,在加深网络的同时解决了梯度消失的问题.在此基础上,Densnet[9]设计了一种新的深度网络架构改善梯度消失的问题:在保证网络中层与层之间最大程度的信息传输的前提下,直接将所有网络层连接起来.但从另一个角度看,这些网络框架都是趋向于往更深层次的方向发展.网络越深,意味着需要训练的参数越多,需要的存储空间也会越大,计算花费的时间也会更多.这对于实际应用来讲,会存在一些问题.目前已有一些研究致力于降低网络运行的计算开销[10,11],常用的方法就是用一个预训练网络模型,在此基础上,用很少的参数在特定数据集上训练目标神经网络.同时,还有一些研究通过改善网络的结构,采用优化类别间相似性度量的算法来进行图像分类[12],也有研究针对多标签的图像分类,提出了输入输出更灵活的 HCP 网络[13].

随着近些年网络上可获得的信息量的增大,在大数据集上进行图像信息计算不仅在时间开销上,在计算开销上也都是不乐观的.哈希算法因其在速度和存储方面存在的优势,近些年来被广泛地应用于大规模数据集的视觉任务中[14,15].目前,基于哈希的方法主要分为两大类:有监督哈希方法[16,17]和无监督哈希方法[18,19].其中,最具代表性的是局部感知哈希(local-sensitive hashing,简称 LSH)[16],使用随机映射使相似的数据匹配到相近的二进制编码的概率最大化.另一个具有代表性的方法是谱哈希(spectral hashing,简称 SH)[18],通过非线性函数沿着数据的主成分分析(principal component analysis,简称 PCA)方向设定阈值产生二进制编码.在卷积神经网络的基础上,文献[20]首先提出了一种监督哈希方法 CNNH 和 CNNH+,该方法把训练数据成对的语义相似度矩阵因式分解成近似哈希编码,然后利用这些近似哈希编码和图像标签训练出网络模型,取得了不错的性能.文献[14]提出了一种简单、高效的深度学习框架,在 AlexNet 框架的基础上,提出隐层概念,能够同时学习图像特征表示和哈希函数,在图像检索性能上取得了卓越的表现.

本文提出一种基于深度卷积神经网络 AlexNet 的二值哈希图像分类方法.采用有监督的学习方式同时学习不同层次的图像特征和哈希编码.在网络中采用扩大局部感受野和减小卷积滤波器尺寸的方法,获得了更具区分力和表达力的深层特征;然后,在全连接层中引进隐层并对隐层神经元用二值激活函数获得二值哈希编码,通过计算不同类别间的二值哈希编码的汉明距离对图像进行分类.相比于其他图像分类方法,本文提出的方法有以下特点.

(1) 提出了一种简单、高效的有监督学习的图像分类框架,能在提高分类精度的同时降低计算开销;

(2) 该框架在原有的 AlexNet 框架上进行了改进,在池化阶段采用最大-均值池化(max-ave pooling)方式,在扩大局部感受野的同时保留更精确的图像特征信息;

(3) 在全连接层采用最大值(maxout)激活输出,使网络表达更精确的高维特征信息;

(4) 通过在全连接层引入隐层来学习哈希编码,提高分类效率,使得网络能够同时学习图像特征表达和二值哈希编码,并可应用于大规模图像数据.

实验结果表明,本文提出的优化方法可以明显地提升深度卷积神经网络在大规模图像分类任务上的性能,

性能优于现有的方法.

1 卷积神经网络与哈希算法概述

1.1 卷积神经网络

卷积神经网络(convolutional neural network)[6]是第一个成功训练多层网络结构的学习算法,通过提取图像特征[21],最终能够获得一幅图像的高级语义特征.网络越靠近输出层,图像的特征表示就越抽象,高级语义特征越丰富,就越能够表现图像主题,在图像分类任务中的识别能力也就越强.

AlexNet 是一种被经常用到的深度卷积神经网络,该网络包括 5 个卷积层、3 个池化层和 3 个全连接层.卷积层和池化层实现图像特征的提取,全连接层放在卷积层后面,将二维的特征图压缩为一维的特征向量.网络的预训练过程分为前向传播和反向传播两个阶段.

(1) 前向传播阶段.传播过程中对每层输入特征的运算如下:

$$y^{(l)}=f\left(\sum_{i\in m}W_i^l\otimes x_i^{(l-1)}+b^l\right) \tag{1}$$

其中,$y^{(l)}$ 为第 l 个卷积层的输出,$x^{(l)}$ 为输入向量,$\otimes$ 为卷积运算,b^l 为偏置,W_i 为该层对应的卷积核权值,m 代表输入特征图集合,$f(x)$代表非线性激活函数,常用的有 Sigmoid、Tanh 和 Relu 等,最近也有一些研究致力于激活函数的使用,如 PRelu、Maxout[22]等.

(2) 反向传播阶段(也称为误差传播阶段).对于有 m 个样本的数据集,网络的前向传播阶段会输出每个类别线性预测的结果,根据这个结果和网络期望的输出,定义网络的整体目标函数为

$$E(W)=\min\sum_{i=1}^{M}L(z_i)+\lambda\|W\|^2 \tag{2}$$

其中,$L(z_i)$ 是网络对应的损失函数.通过迭代训练最小化损失函数来降低网络分类误差,z_i 为网络反向传播的输入,即公式(1)中最后一层网络的输出.W 代表网络在本次迭代中所占的权值,λ代表相应的归一项所占比重.损失函数 $L(z_i)$ 的选择需要根据具体的分类目标来确定.对于本文中的多类别图像分类任务,我们直接采用 Softmax 分类器的输出并最小化交叉熵损失函数.Softmax 归一化概率函数定义如下:

$$z_i=z_i-\max(z_1,z_2,\ldots,z_m) \tag{3}$$

$$\sigma_i(z)=\frac{\exp(z_i)}{\sum_{j=1}^{m}\exp(z_j)},i=1,\ldots,m \tag{4}$$

其中,z_i 是每一个类别线性预测的结果,减去最大值的目的是为了保持计算时的数值稳定性,因为网络最后会做归一化处理,故此处减去一个最大值从形式上是不会改变最终结果的.同时,根据 $\sigma_i(z)$ 来预测输入 z_i 属于每一个类别的概率.在此基础上,我们定义损失函数为

$$L(z_i)=-\log\sigma_i(z) \tag{5}$$

通过梯度下降算法对公式(1)中每一层的参数 W 和 b^l 求导,得到网络参数的更新值,最小化损失函数.

1.2 哈希算法

哈希(hash),也被称作散列,是指把任意长度的输入通过哈希算法变换成固定长度的散列输出.这种转换实质上是进行数据降维.目前,主流的哈希算法均基于二进制编码(binary code).这可以压缩高维特征向量,从而具有计算效率高和存储空间小等优势.本文采用的哈希编码方法受到基于监督核的哈希算法[23]的启发,其核心是利用核函数 $\kappa:R^d\bullet R^d\to R$ 构建哈希函数 $h:R^d\to\{0,1\}$,将高维特征映射到低维空间,并保持原特征点相对空间位置不变.同时,因二进制哈希码采用汉明距离计算,使得最后的特征处理哈希函数的具体表示如下:

$$f(x)=\sum_{j=1}^{m}\kappa(x_{(i)},x)a_j-b \tag{6}$$

$$h(f(x))=\mathrm{sgn}(f(x))=\begin{cases}1, & f(x)>0\\ 0, & f(x)\leqslant 0\end{cases} \tag{7}$$

其中,$a_j\in R, b_j\in R$, $x_{(i)}$ 是从数据中随机选取的 m 个样本.

哈希函数 $h(x)$除了满足低维空间与高维空间的相似一致性以外,还应保证生成均衡哈希码,使学习到的哈希码中保存的信息量最大,即满足 $\sum_{i=1}^{n}h(x_i)=0$, 则偏置 $b=\sum_{i=1}^{n}\sum_{j=1}^{m}\kappa(x_{(j)},x_i)a_j/n$. 将 b 代入式(6)中,可得:

$$f(x)=\sum_{j=1}^{m}(\kappa(x_{(j)},x)-\frac{1}{n}\sum_{i=1}^{n}\kappa(x_{(j)},x_i))a_j=a^T\bar{k}(x) \tag{8}$$

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

$$\bar{k}=[\kappa(x_{(1)},x)-\mu_1,...,\kappa(x_{(m)},x)-\mu_m]^T \tag{9}$$

2 基于深度卷积神经网络的图像分类

2.1 AlexNet网络框架的优化

AlexNet 框架如图 1 所示.该模型的输入为图像及相应的标签,输出为图像属于某一类的概率.我们从两个方面对框架进行了改进与优化,分为网络框架和网络内部优化两个方面.网络框架的优化分为两种:(1) 在卷积层后接最大值-均值池化方式,替代原先的最大值池化;(2) 在全连接层 FC-7 与 FC-8 之间增加一个新的采用全连接的隐层 H.如图 1 中虚线框所示.对于网络内部的优化也有两种方案:(1) 在每一层卷积的输出之后对数据做批规范化处理(batch normalization)[24],再送入下一个网络层;(2) 全连接层的激活函数采用 Maxout 激活函数,替换原来的 Relu 激活函数.

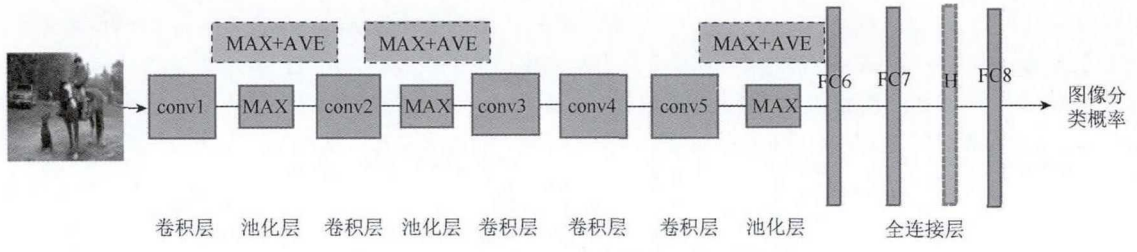

Fig.1 AlexNet architecture and the optimizations

图 1 AlexNet 网络结构图及优化

2.2 Maxout+Dropout在卷积神经网络中的特征表达

传统的激活函数,如 Sigmod、ReLu 只能拟合二维函数,而研究[22]表明,Maxout 能够拟合任意维度的函数.

Maxout 模型是一种前向传播结构,采用最大输出的激活形式.给定输入 $x\in\Re^d$ (其中,x 可以是给定的输入向量,或者是隐层状态),Maxout 输出本层一个节点的表达式为

$$h_i(x)=\max_{j\in[1,k]}z_{ij} \tag{10}$$

其中,$z_{ij}=x^TW_{...ij}+b_{ij}, W\in\Re^{d\times m\times k}, b\in\Re^{m\times k}$.

在卷积神经网络中,Maxout 函数从 k 个隐层节点的延伸节点中取最大值输出作为该节点的输出.单个的 Maxout 激活函数可以被看作是一个分段线性函数,可以拟合任意的凸函数.因为 k 个隐层节点是线性的,故它在每一处都是局部线性的.只要 Maxout 单元含有多个延伸的隐层节点,那么理论上只需要两个 Maxout 输出就可以实现任意连续函数的拟合.在拟合函数的同时,网络还可以通过训练学习得到隐层之间的关系.

为了解决 Maxout 输出的非稀疏性问题,Dropout 函数被加在 Maxout 之后,以达到输出稀疏的效果.本文在全连接层利用 Maxout+Dropout 代替传统的 Relu 激活函数,提取到了更抽象、精确的图像特征.考虑到加入

Maxout 激活函数之后,全连接层的参数变多了,为了在不加大计算负担的前提下还能保证提取到更为精确的图像特征,FC6 和 FC7 层的输出维度由 4 096 维降到 2 048 维,并减小 Dropout,以达到 50%的稀疏性.

2.3 最大-均值池化扩大图像局部感受野

很多流行的计算机视觉识别算法[4,5,25]中都包含了空间池化这一步骤,即把特定区域内的几个特征检测器的输出整合成一个局部或者全局的特征包,以这种方式保留与目标任务相关的特征信息,同时丢弃相关性小的细节信息.池化的主要优点在于可以实现图像变换不变性、特征表达的紧凑性、对噪声和扰乱的鲁棒性以及扩大局部感受野.文献[26]提出了一种局部约束线性编码方法,通过最大池化方法取得了很好的分类性能.文献[27]对所提系统分析、比较了最大值池化和平均值池化在目标分类中的作用效果,表明池化的细节处理会在很大程度上影响任务的性能.但是,在不同的分类任务中,如何学习或者设计更好的池化方式,尤其是在深度网络中,却仍然是一个待解决的问题.本文提出的最大-均值池化结合了最大池化和平均池化各自的有点,在扩大局部感受野的同时保留了更精确的图像特征信息.

最大池化使得提取的特征具有平移不变性,而平均池化使得提取的特征对微小变形鲁棒,这与人类视觉感知中的复杂细胞功能类似.公式(11)、公式(12)分别定义了最大池化和平均池化:

$$f(v)=\frac{1}{T}\sum_{m=1}^{T}v_m \tag{11}$$

$$f(v)=\max_{1\leqslant m\leqslant T}v_m \tag{12}$$

其中,v_m 表示提取自图像的滑动窗口中 T 个像素点的第 m 个像素点,m 表示该元素在滑动窗口中的空间方位,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.

本文结合两者的优势,把最大值池化和平均值池化分别以权重 1 相加作为新的最大-平均池化方法,如公式(13)所示:

$$f(v)=\left(\frac{1}{T}\sum_{m=1}^{T}v_m\right)+\max_{1\leqslant m\leqslant T}v_m \tag{13}$$

该最大-平均池化方法用于替代原有的最大池化方法,如图 1 中虚线框所示.

2.4 隐层的二值哈希编码学习

卷积神经网络提取的特征是基于有标签数据的,这与传统的手工特征提取不同.最近的研究[3,28,29]表明,全连接层 FC6-8 能够很好地表达图像信息.这些中层的图像特征在图像分类、检索和其他任务中都有很好的表现.目前的研究都是沿用 Hinton 的方法[3]用 FC7 层表达图像特征.但 FC7 层输出的特征是高维的,势必会增加计算负担.有研究[14]提出,在 FC7 层后增加一个新的隐层 H,隐层 H 是一个全连接层,其神经元的活动由后续的 FC8 层的语义编码和分类调节.隐层的引出不仅能够从 FC7 层提取丰富的抽象特征,同时也能将特征表达更贴近高层次语义表达.

在本文中,为解决大规模图像分类的计算负担问题,受文献[14]启发,隐层模式被采用,即在 FC7 层之后添加一个隐层 H,其激活函数为

$$a_n^H=\sigma(a_n^7W^H+b^H) \tag{14}$$

其中,$\sigma(\cdot)$ 是 Sigmoid 逻辑回归函数,把输出控制在(0,1)之间.a_n^7 为给定图像在第 7 层的输出特征向量,W^H 是该层网络的权值,b^H 是隐层的偏置参数.在此基础上,定义二值编码函数为

$$b_n=\begin{cases}1, & a_n^H>0.5\\ 0, & a_n^H\leqslant 0.5\end{cases} \tag{15}$$

通过以上操作把输入图片对应特征向量编码为用 0,1 表示的二值码,通过比较二值码的相似程度,对图像进行分类.这样,一方面有利于让网络在迭代过程中学习图像到二值码的特征映射图;另一方面,使网络的计算过程更简单,即直接采用汉明距离计算两张图片的距离,减小计算量的同时降低电脑的内存占用.

本文提出的网络框架的网络的权值采用 ImageNet 上预训练的网络权值,隐层和 FC8 层的权值采用随机初

始化的方式.通过在全连接层引入隐层,一个可以同时学习特征表达和哈希编码的卷积神经网络模型得以构建.将特征学习和哈希编码相结合,其好处在于,不仅利用了深度学习提取图像特征的能力和哈希编码对高维特征的压缩处理方式,还能对卷积神经网络的高维特征输出进行特征压缩,使得网络输出既符合高层语义特征,且生成的特征向量具有紧凑性,能在很大程度上解决内存占用大以及计算时间长等问题.

3 实验结果与分析

为了验证本文所提方法的有效性,本文利用 Caffe[30]实现了如图 1 所示的经过优化的深度学习框架.采用预训练网络模型的方式,即用在 ImageNet 数据集上预训练好 AlexNet 的权值来初始化网络,并对隐层和输出层的权值采用随机初始化的方式,通过反向传播算法,在目标数据集上微调网络参数,实现了网络模型的迁移学习.实验在配置有 i7-6800K CPU,32G 内存和 GeForce GTX TITAN X 显卡的工作站上进行.

验证实验在 3 个公共数据集:MNIST、CIFAR-10、CIFAR-100 上进行.每个数据集都进行了如下 3 种深度卷积神经网络优化方法的比对实验:(1) 在网络的全连接层用 Maxout 替换 Relu 非线性激活函数且在 FC-7 与 FC-8 之间引入隐层 H;(2) 在图 1 所示网络框架的基础上在卷积层加入批规范化处理;(3) 用最大-均值池化方法来代替传统的最大值池化方法,且在全连接层采用 Maxout 和在 FC-7 与 FC-8 之间引入隐层 H,即如图 1 虚线框所示的网络框架,也是本文最终提出的深度卷积神经网络框架的最终优化方法.

图像分类性能评价指标为误差率,相关定义如下:

$$\text{误差率}=\frac{\text{被错误分类的图像数}}{\text{被分类的图像总数}}\times 100\% \tag{16}$$

3.1 数据集

MNIST[31]数据集是由 0~9 之间的灰度手写数字组成的数据集,共分为 10 类.包含 60 000 张训练图像和 10 000 张测试图像.图像大小为 28×28 像素.

CIFAR-10[32]数据集共分为 10 类,每一类由 6 000 张彩色图像组成.其中,包含 50 000 张训练图像和 10 000 张测试图像.所有图像的大小都是 32×32 像素.

CIFAR-100[32]数据集中的图像大小与 CIFAR-10 一样,同为 32×32 像素分布,但是该数据库有 100 类图像,每一类包括 600 张彩色图片,分别为 500 张训练图片和 100 张测试图片.

3.2 MNIST数据集结果分析

MNIST 数据集是由 10 个类别的图像组成的,所以网络的输出设定为 10 通道,即输出图像属于 10 个类别的概率.同时,设定隐层神经元的个数 n=48,*batchsize*=64,初始学习率 LR=0.001,并采用随机梯度下降法(SGD)训练数据.3 种比对实验的结果和目前性能比较好的方法[12,24,33,34]的比较结果见表 1.

Table 1 Performance comparison of error rates on the MNIST dataset

表 1 错误分类率在 MNIST 数据集的比较结果

分类方法	误差率(%)
2-Layer CNN+2-Layer NN[33]	0.53
Stochastic Pooling[33]	0.47
NIN+Dropout[34]	0.47
Conv.maxout+Dropout[22]	**0.45**
AlexNet-Fine-tuning[14]	0.47
AlexNet+FC.maxout	0.66
AlexNet+FC.maxout+Batch Normalization	0.67
AlexNet+FC.maxout+Max-Ave-pooling	0.50

从上述实验结果可以看出,在本文提出的 3 种方案中,全连接层使用 Maxout 激活函数的错误分类率和卷积层使用 Batch Normalization 的错误分类率都比采用最大-均值池化的方法的错误分类率要大,两种框架的性能分别为 0.66%和 0.67%.同时,采用最大-均值池化方案的实验结果与比较基准 AlexNet[14]和目前性能较好的方

法,如NIN[34]和Conv.maxout+Dropout[22]基本持平.一方面,因MNIST数据集简单,一张图片包含的像素信息比较少,在经过几个卷积层的特征提取之后像素信息均能被很好地表达出来,因此,Maxout 在复杂图像信息中能够很好地进行特征表达的特点在本数据集中并没有发挥很大作用;另一方面,Batch Normalization 是针对复杂的网络框架进行快速训练的,但在本文的网络训练过程中我们发现,大约迭代1 000次之后网络就能达到约2%的错误分类率,所以 Batch Normalization 的引入对分类效果所起的作用也不大,但可以在一定程度上加快网络训练速度,减轻计算负担.

3.3 CIFAR-10数据集实验结果分析

CIFAR-10 数据库中图像的类别也分为 10 类,所以把网络的输出定为 10 通道,以此来预测 10 个类别的CIFAR-10 数据图像的分类.同时,设定隐层神经元的个数 n=48,*batchsize*=64,初始学习率 LR=0.001,并采用随机梯度下降法(SGD)训练数据.在此数据库上的比对实验性能与 Stochastic Pooling[33]、CNN+Spearmint[35]、Conv.maxout+Dropout[22]、MCDNN[36]、NIN[34]、CNN[14]在MNIST数据集上作了比较,结果见表2.

Table 2 Performance comparison of error rates on the CIFAR-10 dataset

表2 错误分类率在CIFAR-10数据集上的比较结果

分类方法	误差率(%)
Stochastic Pooling[33]	15.13
CNN+Spearmint[35]	14.98
Conv.maxout+Dropout[22]	11.68
MCDNN[36]	11.21
NIN+Dropout[34]	10.41
AlexNet+Fine-tuning[14]	10.60
AlexNet+FC.maxout	11.10
AlexNet+FC.maxout+Batch Normalization	11.37
AlexNet+FC.maxout+Max-Ave-pooling	**9.80**

从实验结果可以看出,单独在全连接层使用Maxout激活函数并不能取得性能的最大提升,在所有比较方法中处于中游水平.相对于在卷积层及全连接层都使用 Maxout[22]能达到 11.68%的错误分类率,本文提出的方法只在全连接层使用了 Maxout 激活函数也可以取得 11.10%的表现,虽然性能提升并不明显,但需训练的网络参数及计算开销大为减少.而在网络的卷积层使用 Batch Normalization 处理,也不能很好地降低图像分类误差;而加入最大-均值池化之后,错误分类率有了明显的降低,达到了约 9.80%.由此可见,最大-均值池化方式相比于单独使用最大值池化,结合了平移不变性和微小形变不变性的优点,对噪声及其他干扰具有更高的鲁棒性,保留了更重要的图像特征,这使得网络在训练过程中有更好的学习样本,最终学习到的网络权值也会更加合适.Batch Normalization 批规范化处理的引入,则是为了克服深度神经网络训练困难的弊端,加快网络训练速度,在网络结构不是特别复杂、网络深度不是很深的情况下,Batch Normalization 的作用是可以被忽略的.

3.4 CIFAR-100数据集实验结果分析

为了验证本文提出的网络模型在多类别的大数据集上的图像分类的能力,CIFAR-100 数据集被用来进行对比实验.为匹配该数据集的输出类别 100,网络模型的输出被调整为 100 个通道,并设定隐层神经元个数为128.同时,设定 *batchsize*=64,初始学习率 LR=0.001,采用随机梯度下降法(SGD)训练数据.本文提出的网络框架与Learning Pooling[37]、Stochastic Pooling[33]、Conv.maxout+Dropout[22]、Tree based priors[38]和NIN[34]在CIFAR-100数据集上的比较结果见表3.

从实验结果可以看出,本文提出的网络优化方法具有很好的泛化能力,在多类别大规模数据集 CIFAR-100上的测试误差均低于比较基准 AlexNet+Fine-tuning[14],也优于当前性能最好的图像分类方法.作为自身实验对比,本文提出的在卷积层采用最大-均值池化和全连接层采用Maxout激活函数及隐层H的优化网络模型依然表现出最好的分类性能,将测试误差降低到了29.15%.

Table 3 Performance comparison of error rates on CIFAR-100 dataset
表 3 错误分类率在 CIFAR-100 数据集上的比较结果

分类方法	误差率(%)
Learning Pooling[37]	43.71
Stochastic Pooling[33]	42.51
Conv.maxout+Dropout[22]	38.57
Tree based priors[38]	36.85
NIN+Dropout[34]	35.68
AlexNet+Fine-tuning[14]	32.62
AlexNet+FC.maxout	31.75
AlexNet+FC.maxout+Batch Normalization	32.34
AlexNet+FC.maxout+Max-Ave-pooling	**29.15**

3.5 隐层神经元个数对图像分类结果的影响

为了验证隐层神经元个数对图像分类结果的影响,我们分别在上述 3 个数据集上测试不同神经元个数的图像分类精度.针对 MNITST、CIFAR-10 数据集的 10 分类任务,设定隐层神经元个数 n={12,32,48,64,128}.因为 CIFAR-100 数据集共有 100 类,为了避免网络维度不匹配的情况,设定隐层神经元个数 n={100,128,256}.在本文提出的最终优化模型 AlexNet+FC.maxout+Max-Ave-pooling 上进行实验,*batchsize*=64,初始学习率 LR=0.001,采用随机梯度下降法(SGD)训练数据.实验结果见表 4 和表 5.

Table 4 Comparison of error rates on different hidden unit (I)
表 4 错误分类率在不同隐层神经元个数上的比较结果(I)

数据集	12	32	48	64	128
MNIST (%)	0.45	0.48	0.50	0.55	0.58
CIFAR-10 (%)	12.19	11.00	9.80	10.05	10.55

Table 5 Comparison of error rates on different hidden unit (II)
表 5 错误分类率在不同隐层神经元个数上的比较结果(II)

数据集	12	32	48	64	100	128	256
CIFAR-100 (%)	77.62	51.10	39.46	34.6	32.16	29.15	29.54

从上述两个表格可以看出,隐层神经元数量对图像分类精度是有影响的.对于 MNIST 数据集,随着隐层神经元个数的增加,分类精度逐渐降低,但误差都保持在 0.1%左右.主要是因为 MNIST 数据集包含的图像信息较为简单,增加神经元个数会使网络过拟合,降低分类精度.对于 CIFAR-10 数据集,当隐层神经元个数小于 48 时,随着个数增加,图像分类精度越来越高,但在 48~64 之间时,分类精度变化不明显,到 128 时,网络有轻微的过拟合现象,分类精度开始降低.对于 CIFAR-100 数据集,因为分类目标有 100 类,可以明显地看出,在隐层神经元个数小于 100 时网络处于欠拟合状态,在 128 和 256 时,网络的分类精度基本持平.综合考虑图像分类精度以及计算资源和计算时间,本文在 MNIST 数据集和 CIFAR-10 数据集中所用神经网络中隐层神经元个数为 48,在 CIFAR-100 数据集中隐层神经元个数为 128 个.

4 结束语

本文提出一种基于深度卷积神经网络 AlexNet 的二值哈希图像分类框架,通过在卷积层使用最大-均值池化方式和在全连接层采用 Maxout 激活函数以及在隐层实现二值哈希编码,能够同时学习特征的精确表达和高效二值哈希编码.同时,本文也探索了把 Batch Normalization 应用在卷积层,以及单独使用 Maxout 激活函数等其他可能的优化方案,并分析了各自的特点.通过在 3 个常用数据库上与最近报告的性能较好的方法及 AlexNet 方法进行比对实验和理论分析,其结果表明,在卷积层使用最大-均值池化方式,在全连接层采用 Maxout 激活函数和隐层二值哈希编码具有较好的性能,并有很好的应用前景.未来工作将在目前图像分类的基础上,进一步调整网络框架,学习更精确的特征表达,并尝试应用于图像检索、目标识别等其他多媒体分析任务中.

References:

[1] Christopher JCB. A tutorial on support vector machines for pattern recognition. ACM Trans. on Data Mining and Knowledge Discovery, 1998,2(2):121-167. [doi: 10.1023/A:1009715923555]

[2] Penatti OAB, Silva FB, Valle E, Gouet-Brunet V, Torres RDS. Visual word spatial arrangement for image retrieval and classification. ACM Trans. on Pattern Recognition, 2014,47(2):705–720. [doi: 10.1016/j.patcog.2013.08.012]

[3] Krizhevsky A, Sutskever I, Hinton GE. ImageNet: Classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. Lake Tahoe: Curran Associates, Inc., 2012. 1097–1105.

[4] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2016. 779–788. [doi: 10.1109/CVPR.2016.91]

[5] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. IEEE, 2015. 1–9. [doi: 10.1109/CVPR.2015.7298594]

[6] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-Based learning applied to document recognition. Proc. of the IEEE, 1999, 86(11):2278–2324. [doi: 10.1109/5.726791]

[7] He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proc. of the Int'l Conf. on Computer Vision. 2015. 1026–1034. [doi: 10.1109/ICCV.2015.123]

[8] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proc. of the Computer Vision and Pattern Recognition. IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

[9] Huang G, Liu Z, van der Maate L, Weinberger KQ. Densely connected convolutional networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017. [doi: 10.1109/CVPR.2017.243]

[10] Wang J, Kumar S, Chang SF. Semi-Supervised hashing for large-scale search. IEEE Trans. on Pattern Analysis & Machine Intelligence, 2012,34(12):2393. [doi: 10.1109/TPAMI.2012.48]

[11] Ba LJ, Caruana R. Do deep nets really need to be deep. In: Advances in Neural Information Processing Systems. Montreal: Curran Associates, Inc., 2013. 2654–2662.

[12] Qu Y, Li L, Shen F, Lu C, Wu Y, Xie Y, Tao DC. Joint hierarchical category structure learning and large-scale image classification. IEEE Trans. on Image Processing, 2017,99:1. [doi: 10.1109/TIP.2016.2615423]

[13] Wei Y, Wei X, Lin M, Huang JS, Ni BB, Dong J, Zhao Y, Yan SC. HCP: A flexible CNN framework for multi-label image classification. IEEE Trans. on Pattern Analysis & Machine Intelligence, 2015,38(9):1901–1907. [doi: 10.1109/TPAMI.2015.2491929]

[14] Yang HF, Lin K, Chen CS. Supervised learning of semantics-preserving hash via deep convolutional neural networks. IEEE Trans. on Pattern Analysis & Machine Intelligence, 2017,99:1. [doi: 10.1109/TPAMI.2017.2666812]

[15] Wang CF, Su L, Zhang WG, Huang QM. No reference video quality assessment based on 3D convolutional neural network. Ruan Jian Xue Bao/Journal of Software, 2016,27(S2):103–112 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/16025.htm

[16] Gionis A, Indyk P, Motwani R. Similarity search in high dimensions via hashing. In: Proc. of the Int'l Conf. on Very Large Data Bases. San Francisco: Morgan Kaufmann Publishers, 2000. 518–529.

[17] Mao XJ, Yang YB. Semantic hashing with image subspace learning. Ruan Jian Xue Bao/Journal of Software, 2014,25(8): 1781–1793 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4488.htm [doi: 10.13328/j.cnki.jos.004488]

[18] Weiss Y, Torralba A, Fergus R. Spectral hashing. In: Proc. of the Conf. on Neural Information Processing Systems. Vancouver: Curran Associates, Inc., 2008. 1753–1760.

[19] Norouzi M, Fleet DJ. Minimal loss hashing for compact binary codes. In: Proc. of the Int'l Conf. on Machine Learning. Washington: Omnipress, 2011. 353–360.

[20] Xia R, Pan Y, Lai H, Liu C, Yan S. Supervised hashing for image retrieval via image representation learning. In: Proc. of the American Association for Artificial Intelligence. 2014. 2156–2162. https://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8137

[21] Lecun Y, Kavukcuoglu K, Farabet C. Convolutional networks and applications in vision. In: Proc. of the IEEE Int'l Symp. on Circuits and Systems. 2010. 253–256. [doi: 10.1109/ISCAS.2010.5537907]

[22] Goodfellow IJ, Wardefarley D, Mirza M, Courville A, Bengio Y. Maxout networks. In: Proc. of the Int'l Conf. on Machine Learning. Atlanta, 2013. 1319–1327.

[23] Liu W, Wang J, Ji RR, Jiang YG, Chang SF. Supervised hashing with kernels. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2012. 2074–2081. [doi: 10.1109/CVPR.2012.6247912]

[24] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proc. of the Int'l Conf. on Machine Learning. 2015. 448–456.

[25] Boureau YL, Bach F, Lecun Y, Ponce J. Learning mid-level features for recognition. In: Proc. of the IEEE Int'l Conf. on Computer Vision and Pattern Recognition. 2010. 2559–2566. [doi: 10.1109/CVPR.2010.5539963]

[26] Wang JJ, Yang JC, Yu K, Lü FJ, Huang T, Gong YH. Locality-Constrained linear coding for image classification. In: Proc. of the IEEE Int'l Conf. on Computer Vision and Pattern Recognition. 2010. 3360–3367. [doi: 10.1109/CVPR.2010.5540018]

[27] Boureau YL, Ponce J, Lecun Y. A theoretical analysis of feature pooling in visual recognition. In: Proc. of the Int'l Conf. on Machine Learning. Haifa, 2010. 111–118.

[28] Wang H, Cai Y, Zhang Y, Pan HX, Lü WF, Han H. Deep learning for image retrieval: What works and what doesn't. In: Proc. of the Int'l Conf. on Data Mining Workshop. 2015. 1576–1583. [doi: 10.1109/ICDMW.2015.121]

[29] Donahue J, Jia Y, Vinyals O, Hoffma J, Zhang N, Tzeng E, Darrell T. DeCAF: A deep convolutional activation feature for generic visual recognition. In: Proc. of the Int'l Conf. on Machine Learning. Atlanta, 2013. 815–830.

[30] Jia YQ, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T. Caffe: Convolutional architecture for fast feature embedding. In: Proc. of the 22nd ACM Int'l Conf. on Multimedia. 2014. 675–678. [doi: 10.1145/2647868.2654889]

[31] Lecun Y, Cortes C. The MNIST database of handwritten digit. 1998. http://yann.lecun.com/exdb/mnist

[32] Krizhevsky A. Learning multiple layers of features from tiny images. Technical Report, Computer Science Department, University of Toronto, 2009. http://www.cs.toronto.edu/~kriz/cifar-10-binary.tar.gz

[33] Zeiler MD, Fergus R. Stochastic pooling for regularization of deep convolutional neural networks. In: Proc. of the Int'l Conf. on Learning Representation. 2013. http://arxiv.org/abs/1301.3557

[34] Lin M, Chen Q, Yan S. Network in network. In: Proc. of the 2nd Int'l Conf. on Learning Representations. 2014, arXiv:1312.4400. https://arxiv.org/abs/1312.4400

[35] Snoek J, Larochelle H, Adams RP. Practical Bayesian optimization of machine learning algorithms. In: Advances in Neural Information Processing System. Lake Tahoe: Curran Associates, Inc., 2012. 2951–2959.

[36] Schmidhuber J, Meier U, Ciresan D. Multi-Column deep neural networks for image classification. In: Proc. of the Computer Vision and Pattern Recognition Workshops. 2012,157(10):3642–3649. [doi: 10.1109/CVPR.2012.6248110]

[37] Malinowski M, Fritz M. Learnable pooling regions for image classification. In: Proc. of the Int'l Conf. on Learning Representations Workshop. 2013. http://arxiv.org/abs/1301.3516

[38] Srivastava N, Salakhutdinov R. Discriminative transfer learning with tree-based priors. In: Advances in Neural Information Processing Systems. Lake Tahoe: Curran Associates, Inc., 2013. 2094–2102.

附中文参考文献:

[15] 王春峰,苏荔,张维刚,黄庆明.基于3D卷积神经网络的无参考视频质量评价.软件学报,2016,27(增刊(2)):103–112. http://www.jos.org.cn/1000-9825/16025.htm

[17] 毛晓蛟,杨育彬.一种基于子空间学习的图像语义哈希索引方法.软件学报,2014,25(8):1781–1793. http://www.jos.org.cn/1000-9825/4488.htm [doi: 10.13328/j.cnki.jos.004488]

白琮(1981－),男,山东泰安人,博士,讲师,CCF 专业会员,主要研究领域为计算机视觉,多媒体信息处理.

黄玲(1994－),女,学士,CCF 学生会员,主要研究领域为计算机视觉,多媒体信息处理.

陈佳楠(1990－),男,硕士生,主要研究领域为计算机视觉,多媒体信息处理.

潘翔(1977－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为计算机视觉.

陈胜勇(1973－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为计算机视觉.