·信息工程·

DOI:10.15961/j.jsuese.2017.01.028

基于深度学习编码模型的图像分类方法

赵永威[1],李 婷[2],蔺博宇[3]
(1.武警工程大学 电子技术系,陕西 西安 710000;
2.河南财政金融学院 信息工程系,河南 郑州 451464;3.63618部队,新疆 库尔勒 841000)

摘 要:针对矢量量化编码的量化误差严重,而稀疏编码只是一种浅层学习模型,容易导致视觉词典对图像特征缺乏选择性的问题,提出了一种基于深度学习特征编码模型的图像分类方法。首先,采用深度学习网络无监督的受限玻尔兹曼机(RBM)代替传统的*K*-Means聚类及稀疏编码等方法对SIFT特征库进行编码学习,生成视觉词典;其次,对RBM编码添加正则化项分解组合每个特征的稀疏表示,使得生成的视觉单词兼具稀疏性和选择性;然后,利用训练数据的类别标签信息有监督地自上而下对得到的初始视觉词典进行微调,得到图像深度学习表示向量,以此训练SVM分类器并完成图像分类。实验结果表明,本文方法能有效克服传统矢量量化编码及稀疏编码等方法的缺点,有效地提升图像分类性能。

关键词:图像分类;视觉词典模型;深度学习;稀疏编码;受限玻尔兹曼机
中图分类号:TP39 **文献标志码**:A **文章编号**:2096-3246(2017)01-0213-08

Image Classification Method Based on Deep Learning Coding Model

ZHAO Yongwei[1],*LI Ting*[2],*LIN Boyu*[3]
(1. Dept. of Electrical Eng., Eng. Univ. of PAP, Xi'an 710000, China;
2. Dept. of Info. Eng., Henan Inst. of Finance and Banking, Zhengzhou 451464, China; 3. 63618 Troop, Kuerle 841000, China)

Abstract: For the serious quantization error in vector quantitation coding, the sparse coding is only a shallow learning model which caused the codeword lack selectivity for image features. In this paper, an image classification method based on deep learning coding model was proposed. Firstly, the deep learning network unsupervised RBM was used to encode SIFT features and generate visual dictionary instead of the traditional *K*-means clustering. Then, the unsupervised RBM learning was steered by using a regularization scheme, which decomposes into a combined prior for the sparsity of each feature's representation as well as the selectivity for each codeword. Finally, the initial dictionary was fine-tuned to be discriminative through the supervised learning from top-down labels. To train SVM classifier and complete image classification, the representation features based on image deep learning were obtained. The experimental results demonstrated that the proposed method can overcome the disadvantage of vector quantization coding and sparse coding. Moreover, the classification performance can be boosted effectively.

Key words: image classification; bag of visual words model; deep learning coding model; sparse coding; restricted Boltzmann machine

近十年来,基于视觉词典模型[1-2](bag of visual words model,BoVWM)的图像分类方法在诸多数据集上取得了较好的分类效果。这类方法通常包含以下几步:1)图像局部特征提取;2)对图像库的局部特征进行编码;3)对编码后的特征进行空间融合得到图像表达向量;4)利用SVM等分类器对聚集之后的特征进行训练并将图像划分至相应语义类别。

具体如图1所示。

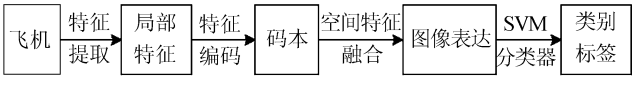

图1 图像分类流程图

Fig. 1 Flow chart of image classification

收稿日期:2016-03-10
基金项目:国家自然科学基金资助项目(61379152;61301232);全军军事学研究生课题资助项目(YJS1062)
作者简介:赵永威(1988—),男,讲师,博士.研究方向:图像分析与处理.E-mail:zhaoyongwei369@163.com

传统视觉词典模型多是采用 K-Means 聚类等矢量量化的编码方法对局部特征进行编码生成视觉码本,也即是视觉词典。然后,将图像中的局部特征特征分配至与之最近的视觉单词,形成表达图像内容的视觉词汇直方图。然而,利用矢量量化编码对图像进行表达和分类的方法存在以下几个问题:1)由矢量量化编码生成的视觉码本会导致空间信息缺失问题;2)视觉词汇直方图构建过程中导致量化误差严重的问题;3)这种图像表达方式只有在利用非线性核 SVM 分类器的情况下才会表现出较好的分类效果,而非线性 SVM 分类器的训练和测试时间复杂度分别为 $O(n^2 \sim n^3)$ 和 $O(n)$,其中,n 为训练图像数目,这无疑会降低其实用性,难以应用到大规模数据集的分类。

研究人员为解决传统视觉词典模型中的视觉单词空间信息缺失的问题同样做了大量工作,空间金字塔匹配(spatial pyramid matching,SPM)[3]通过在多个尺度下将图像划分为大小相同的多个图像块,并分别生成每个图像块的视觉词汇直方图描述来弥补单词空间信息的不足。Sharma 等[4]对初始的空间金字塔匹配方法进行扩展,将视觉单词的空间位置信息融入到视觉词典模型。赵春晖[5]、赵仲秋[6]等针对量化误差严重的问题,提出了一种稀疏编码多尺度空间的图像分类方法,在弥补空间信息的同时,提高了图像内容表达的鲁棒性。针对量化误差严重的问题,Philbin 等[7]和 van Gemert 等[8]则提出了一种软分配方法,有效降低了量化误差,增强了图像表达的准确性。Weinshall 等[9]则将软分配策略与潜在狄里克雷分布模型相结合(latent dirichlet allocation,LDA),提出了一种软分配的 LDA 模型。

而为了得到非线性的图像表达向量,以便在线性核 SVM 分类器的情况下仍能得到较好的图像分类性能。Yang 等[10]采用快速稀疏编码算法生成基于 SIFT 描述子的视觉词典和稀疏向量,并结合空间金字塔匹配(spatial pyramid matching,SPM)算法,提出基于稀疏编码的空间金字塔匹配方法用于图像分类,该方法利用线性核 SVM 分类器(可以将分类器训练时间复杂度降为 $O(n)$)依旧能取得较好的分类性能,从而有效地降低了分类器训练时间,增强了分类方法的实用性。通过对稀疏编码结果更深入的研究,Yu 等[11]在稀疏编码的基础上,提出一种局部约束线性编码(locality-constrained linear coding,LLC)方法,进一步提高了编码性能。该方法取得较高分类准确率的前提是局部特征与其 k 个近邻点处于同一子空间。然而,k - 近邻算法对噪声非常敏感,难以保证得到的 k 个近邻点都处于同一子空间;为此,庄连生等[12]在 LLC 算法的基础上提出了一种非负稀疏局部线性编码算法(nonnegative sparse locally linear coding,NSLLC)对局部特征进行编码,进而保证了所选近邻点都处于同一子空间,同时,相比于稀疏表示,非负稀疏表示更适合图像分类任务。但是,上述几种稀疏编码模型都属于信号重构误差最小化稀疏表示模型,均以最小化信号的重构误差为目标,忽略了判别性对图像分类任务的重要性。基于此,张瑞杰等[13]在非负稀疏局部线性编码的基础上,加入 Fisher 判别约束准则,提出基于 Fisher 判别约束的非负稀疏局部线性编码模型,使得同一类别图像间的稀疏表示距离更近,不同类别图像间的距离更远,从而有效地增强了图像稀疏表示的判别性。

上述稀疏编码方法在一定程度上克服了传统矢量量化编码方法的空间信息缺失及量化误差严重的问题,并能有效地降低分类器训练时间复杂度,增强图像分类方法的实用性。然而,需要指出的是稀疏编码方法只是一种浅度学习,其只有单层的编码层,而且其需要对每一个局部特征进行编码操作,当特征数量和词典规模较大时,会消耗大量的时间。其次,由稀疏编码学习得到的稀疏向量只具有稀疏性而视觉词典缺乏选择性,因此会降低图像内容的分辨力。针对上述问题,本文提出一种基于深度学习编码模型的图像分类方法。首先,利用非监督 RBM 网络采用自底向上的方式对初始 SIFT 特征进行编码得到视觉词典;然后,采用自顶向下的方式为整个网络参数进行有监督微调;最后,利用误差反向传播对初始视觉词典进行有监督微调,并利用新的图像表达方式,即图像深度学习表示向量训练 SVM 分类器对图像进行分类。

1 深度学习相关理论

深度学习的概念起源于人工神经网络,其基本思想是利用多层非线性运算单元构建深度学习网络,并将较低层的输出作为更高层的输入,以此从大量输入数据中学习得到有效的高阶特征表示,最后将这些高阶特征表示用于解决分类、回归和信息检索等特定问题。得益于深度学习的强大表达能力,它已经被成功应用于文本数据学习和视觉识别任务当中[14-16]。

相较于浅学习而言,深度学习具有更强的特征

表达能力,然而,非凸目标函数产生的局部最优解是造成深度学习困难的主要因素,且情况随着网络深度的增加而越发复杂。针对该问题,2006 年,Hinton 等[17]提出了一种用于深度置信网络的无监督学习算法,有效地解决了深度学习模型训练困难的问题。Ranzato 等[18]提出用无监督学习初始化每一层神经网络的想法。具体的在图像分类领域,2012 年,Srivastava 等[19]提出了一种多模式深度置信网络模型(multimodal deep belief network,MDBN),该模型对图像和图像标注数据分别建立 DBN,在最顶层通过学习联合受限玻尔兹曼机将这两个 DBN 结合起来,取得了较好的图像分类性能。同年,Krizhevsky 等[16]构建了具有 6 千万个参数、65 万神经元的大规模深度卷积神经网络,利用 GPU 加速学习过程,在 ILSVRC－2012 比赛中成功地将图像分类误判率从 26.2% 降到 15.3%,取得了远超其他方法的结果。2013 年,Hayat 等[20]在堆栈自编码网络的基础上提出了基于模板的深度重构模型(template deep reconstruction model,TDRM),该模型利用无监督的贪婪逐层训练算法训练高斯受限波尔兹曼机(gaussian restricted boltzmann machines,GRBM),并将训练好的参数作为 TDRM 的初始值,减少了 TDRM 参数训练时间,在 Pascal VOC 2013 年图像分类竞赛中取得了最好的成绩。

然而,上述基于深度学习的图像分类方法都是以训练图像集的像素级数据作为输入,然后学习得到若干维的图像表达向量,这种方法的时间复杂度和计算复杂度都极高,需要耗费大量的人力物力。此外,对学习得到的特征解释性差,也即是这种图像内容的表达方式也仍然停留在底层视觉特征层面。

2　基于深度学习编码模型的图像分类

对训练图像集 而言,基于深度学习编码模型的图像分类流程可由图 2 描述。

具体步骤可如下:

Step 1:提取训练图像库的 SIFT 特征,得训练图像特征库,记为 $R = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_N\}$,其中,$\boldsymbol{r}$ 代表一个 128 的 SIFT 特征,N 为特征点总数;

Step 2:利用无监督受限玻尔兹曼机(RBM)对提取的特征库进行编码,并利用 CD(contrastive divergence)算法[17]不断迭代得到收敛参数码本,生成过完备词典;

Step 3:利用训练数据的类别标签信息对学习得到的词典进行误差反向传播,并对 RBM 网络学习进行有监督的微调,然后,重新对特征库编码得到优化后的视觉词典 W 和稀疏表示系数;

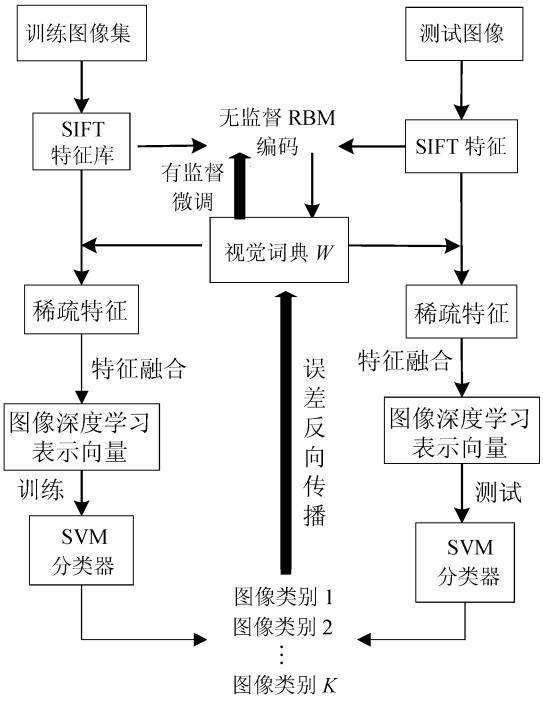

图 2　基于深度学习编码的图像表达与分类方法流程

Fig. 2　Flow chart of image classification method based on deep learning coding model

Step 4:根据词典 W,利用标准梯度优化算法求解每幅图像的稀疏特征,并对稀疏特征进行最大值融合,得到图像的深度学习编码向量表示;

Step 5:利用训练图像的深度学习编码向量表示训练 SVM 分类器,并以此对测试图像进行分类。下面详细介绍如何利用无监督的 RBM 对 SIFT 特征库进行编码以及如何 RBM 网络进行有监督的调整。

2.1　无监督 RBM 编码

这里采用多层的受限波尔兹曼机(restricted boltzmann machines,RBM)对 SIFT 特征库进行编码学习,生成更具有代表性、区分性的视觉词典,具体流程如图 3 所示。

首先,提取 SIFT 特征;其次,结合 SIFT 特征的空间信息,将邻近的 SIFT 特征作为 RBM 的输入,通过 CD 快速算法训练 RBM,得到隐藏层特征;然后,将邻近的隐藏层特征作为下一层 RBM 的输入,得到输出词典。其中,$\boldsymbol{\omega}^1$ 和 $\boldsymbol{\omega}^2$ 是 RBM 的连接权重,RBM 具有一个显层、一个隐层,但是在 RBM 中,同层的神经元之间是无连接的,这样学习使得过程更简单。

在网络的训练过程中,RBM 的隐层与显层间之间是通过条件概率分布相关联的,显层和隐层的条

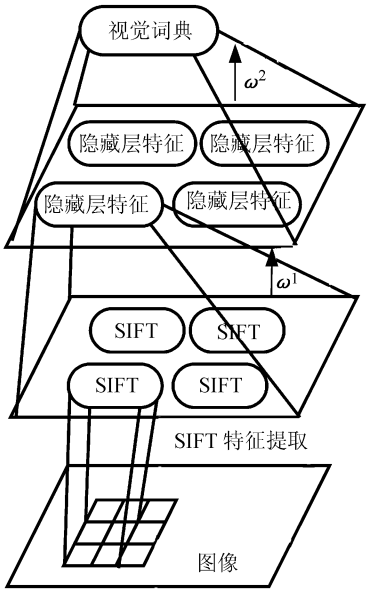

图 3 基于无监督 RBM 特征编码示意图

Fig. 3 Illustration of feature coding based on unsupervised RBM

件概率为：

$$p(z_j \mid \boldsymbol{x}) = \text{sigmoid}(b_j + \sum_{i=1}^{I} \omega_{ij} x_i) \tag{1}$$

$$p(x_i \mid \boldsymbol{z}) = \text{sigmoid}(c_i + \sum_{i=1}^{I} \omega_{ij} z_j) \tag{2}$$

式中：$\text{sigmoid}(\boldsymbol{x}) = \dfrac{1}{1+\mathrm{e}^{-x}}$；$x_i$、$z_j$ 分别代表特征层和编码层，即 RBM 中的显层与隐层；ω_{ij} 为特征层 x_i 与编码层之间的连接权重系数。给定权重系数矩阵 $\boldsymbol{\omega}$ 和隐层偏置向量，输入层特征 $\boldsymbol{x}$ 就可被编码为视觉词典 $\boldsymbol{z}$；相应地，给出 $\boldsymbol{\omega}$ 和显层偏置矩阵 $\boldsymbol{c}$ 就可以由视觉词典 $\boldsymbol{z}$ 重构出特征 $\boldsymbol{x}$。对于 RBM 中一组给定的输入层和编码层 $(\boldsymbol{x},\boldsymbol{z})$，其能量函数可计算如下：

$$E(\boldsymbol{x},\boldsymbol{z}) = -\lg p(\boldsymbol{x},\boldsymbol{z}) = -\sum_{i=1}^{I}\sum_{j=1}^{J} x_i \omega_{ij} z_j - \sum_{i=1}^{I} c_i x_i - \sum_{j=1}^{J} b_j z_j \tag{3}$$

基于能量函数，可得到 $(\boldsymbol{x},\boldsymbol{z})$ 的联合概率分布函数：

$$p(\boldsymbol{x},\boldsymbol{z}) = \frac{\mathrm{e}^{-E(\boldsymbol{x},\boldsymbol{z})}}{\sum_{\boldsymbol{x},\boldsymbol{z}} \mathrm{e}^{-E(\boldsymbol{x},\boldsymbol{z})}} \tag{4}$$

进而得到联合分布的边缘分布——特征输入节点的概率分布，即：

$$p(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{z}} \mathrm{e}^{-E(\boldsymbol{x},\boldsymbol{z})}}{\sum_{\boldsymbol{x},\boldsymbol{z}} \mathrm{e}^{-E(\boldsymbol{x},\boldsymbol{z})}} \tag{5}$$

而 RBM 网络训练的目的就是使 $p(\boldsymbol{x})$ 的值最大化，为此，对式(5)求其梯度得：

$$\frac{\partial \lg p(\boldsymbol{x})}{\partial \omega_{ij}} = \langle x_i z_j \rangle_{\text{data}} - \langle x_i z_j \rangle_{\text{model}} \tag{6}$$

式中，$\langle x_i z_j \rangle_{\text{dist}}$ 表示在分布 dist 下的期望，$\langle x_i z_j \rangle_{\text{data}}$ 为训练数据集经验概率分布下的期望值，$\langle x_i z_j \rangle_{\text{model}}$ 为该模型下概率分布的期望值。通常可由蒙特卡罗马尔可夫链(Monte-Carlo Markov chain，MCMC)方法来得到模型样例：

$$x_i = f_{\text{dec}}(\boldsymbol{z},\boldsymbol{\omega}_i) = \sigma \sum_{j=0}^{J} \omega_{ij} z_j \tag{7}$$

通过 CD 算法对 RBM 进行快速学习，加快参数的收敛，可得到权值 ω_{ij} 的更新量为：

$$\Delta\omega_{ij} = \varepsilon(\langle x_i z_j \rangle_{\text{data}} - \langle x_i z_j \rangle_{\text{model}}) \tag{8}$$

式中，ε 为学习速率。通过 CD 算法，可以得到不断更新的参数，一直到参数收敛，得到初始的视觉词典。

2.2 稀疏性与选择性调整

特征编码后的词典能否准确表达图像内涵，对图像数据进行中层语义表达至关重要。在 RBM 目标优化函数中加入一个正则项 $h(\boldsymbol{z})$，将目标函数 $\boldsymbol{z}^* = \arg\min_{z} \| \boldsymbol{x} - \boldsymbol{\omega}\boldsymbol{z} \|_2^2 + \lambda \| \boldsymbol{z} \|_1$ 调整如下：

$$\arg\min_{\boldsymbol{\omega},\boldsymbol{c},\boldsymbol{b}} -\sum_{k=1}^{K} \lg(\sum_{z} p(x_k,\boldsymbol{z}_j)) + \lambda h(\boldsymbol{z}) \tag{9}$$

其中，λ 为正则项的加权系数。深度学习编码能够使得学习得到的视觉词典具有较强的选择性，并使得图像表达向量具有较好的稀疏性。

这里，说明一下深度学习编码中的稀疏性和选择性概念。稀疏性的核心思想[21]是使用少量的基向量来有效而简洁地表示图像内容。具体为稀疏向量中大部分分量为零，只有少数分量为非零，而少数非零系数则揭示了图像数据的内在结构和本质属性。它是编码向量对输入特征响应的一种度量准则。选择性是度量一个单一视觉单词对输入特征向量的响应。稀疏性与选择性具体如图 4 所示。

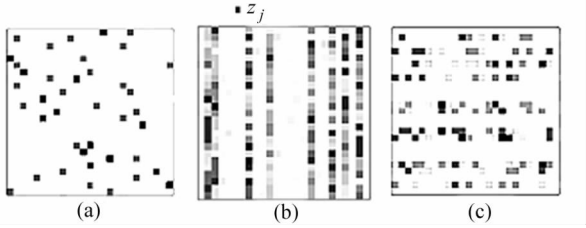

图 4 视觉词典稀疏性与选择性示意图

Fig. 4 Illustration of visual dictionary sparsity and selectivity

图 4(a)中视觉词典兼具稀疏性和选择性，可以

认为词典中的单词具有多样性,且单词之间不具有相关性。图 4(b)中视觉词典只具有选择性,如此会导致某些输入特征向量被忽视或者过表达的现象。图 4(c)中视觉词典只具有稀疏性,它会导致视觉词典中的单词相关性较强,加大冗余。

利用视觉词典对每一维特征响应的均值就可以定量分析稀疏性与选择性,即:

$$h(\boldsymbol{z}) = \sum_{j=1}^{J} \| \hat{p} - \frac{1}{K} \sum_{k=1}^{K} p_{jk} \|^2 \tag{10}$$

式中,$\hat{p}$ 是每个单词针对 K 个特征平均激活概率的期望值,单词 z_j 对特征 $\boldsymbol{x}_k$ 响应概率的期望值可标记为 $p_{jk} \in (0,1)$,那么,整个词典对 K 个输入特征的响应期望值可记为矩阵 $\boldsymbol{P} \in \mathbb{R}^{J \times K}$,矩阵中的每一行元素 $\boldsymbol{p}_j$ 代表了单词 z_j,$(1 \leq j \leq J)$ 对 K 个输入特征向量响应的期望值,向量 $\boldsymbol{p}_k$ 则代表了输入特征 $\boldsymbol{x}_k$ 在整个视觉词典上的分布。因此,为了对整个 RBM 网络进行有监督地微调,定义交叉熵损失函数 $h(\boldsymbol{z})$ 如下:

$$h(\boldsymbol{z}) = -\sum_{j=1}^{J} \sum_{k=1}^{K} p_{jk} \lg p_{jk} + (1 - p_{jk}) \lg (1 - p_{jk}) \tag{11}$$

学习得到视觉词典的稀疏性和选择性与目标矩阵 $\boldsymbol{P}$ 密切相关,对视觉词典 $\boldsymbol{z} \in \mathbb{R}^N$ 而言,矩阵 $\boldsymbol{P}$ 中元素为:

$$p_n = (\mathrm{rank}(z_n, \boldsymbol{z}))^{\frac{1}{\mu}-1} \tag{12}$$

式,参数 $\mu \in (0,1)$。这样就可以获得兼具稀疏性和选择性的视觉词典,进而既能保证各视觉单词的多样性又能兼顾图像局部特征表达之间的差异性,更加准确地表达图像内容。

2.3　有监督微调

由于深度学习编码需要对多层网络进行训练学习,而无监督 RBM 网络在训练时存在一个问题就是,若对所有层同时训练,时间复杂度会太高;如果每次训练一层,偏差就会逐层传递,从而导致严重的欠拟合问题。为此,在利用深度学习对 SIFT 特征编码时,首先采用自底向上的非监督 RBM 分层训练各层参数每层网络生成视觉词典,训练时逐层学习每一层参数,降低时间复杂度。此外,由于非监督 RBM 学习模型的限制以及稀疏性约束使得模型能够学习到训练数据本身的结构,从而得到比输入更有表示能力的特征;然后根据训练数据的标签类别,误差自顶向下传播,对网络各层参数进行微调如下:

$$\tilde{z}_{j,\mathrm{target}}^{(l)} = f_{\mathrm{dec}}(\varphi^{(l+1)} \tilde{z}_{j,\mathrm{target}}^{(l)} + (1 - \varphi^{(l+1)}) z_{\mathrm{data}}^{(l+1)}, \boldsymbol{\omega}_j^{(l+1)}) \tag{13}$$

$$\Delta w_{ij}^{(l)} = \gamma \langle z_i^{(l-1)} z_j^{(l)} \rangle_{\mathrm{data}} + \eta \langle z_{i,\mathrm{data}}^{(l-1)} \tilde{z}_{j,\mathrm{target}}^{(l)} \rangle - \varepsilon \langle z_i^{(l-1)} z_j^{(l)} \rangle_{\mathrm{recon}} \tag{14}$$

式中:$\varphi^{(l)}$ 是一个超参数函数;γ、η、ε 代表学习速率,且有 $\gamma = \varepsilon - \eta$。那么对于第一层网络而言,$\boldsymbol{z}^{(0)}$ 为图像 SIFT 特征输入向量 $\boldsymbol{x}$,且 $\boldsymbol{z}_{\mathrm{data}}^{(3)} = \boldsymbol{z}_{\mathrm{target}}^{(3)} = \boldsymbol{y}$,即深度学习表示向量。那么,顶层网络的参数就可更新如下:

$$\Delta \boldsymbol{\omega}_{ic}^{(3)} = \varepsilon(\langle z_i^{(2)} y_c \rangle_{\mathrm{data}} - \langle z_i^{(2)} y_c \rangle_{\mathrm{recon}}) \tag{15}$$

式中,y_c 是指顶层输出向量被判别为图像类别 c。在上述微调的过程中,采用最大交叉信息熵损失代表基于特征的分类误差,然后该误差反向传播至每层网络中。

综上所述,整个基于深度学习编码模型的图像语义表达方法可以分为 3 个阶段:第 1 阶段是利用非监督 RBM 网络采用自底向上的方式对初始 SIFT 特征进行编码得到视觉词典;第 2 阶段是利用自顶向下的方式为整个网络参数进行有监督微调;第 3 阶段是利用误差反向传播对初始视觉词典进行有监督微调,获得新的图像表达方式,即图像深度学习表示向量训练 SVM 分类器用以对图像进行分类。

3　实验结果与性能分析

3.1　实验设置

分别在常用的 Caltech-256 图像集[22] 和 15-Scenes 图像集[23] 对本文方法性能进行评估。Caltech-256 图像集作为 Caltech-101 图像集的扩充,包含 256 个物体类别,共 30 607 幅图像,其中,每个类别至少包含 80 幅图像。15-Scenes 图像集是由 Lazebnik、Schmid 和 Ponce 共同提出的,15-Scenes 图像集库是在之前 Li Feifei 等建立的 13 类场景数据集的基础上继续扩展得到的数据集,增加了 241 幅 Store 场景和 311 幅 Industrial 场景。它包含 15 类自然图像场景是目前公开的最大场景分类数据集。在有监督微调阶段从每个图像类别中随机选取 50 幅图像用以有监督地微调整个 RBM 网络,并用同样的数据训练线性 SVM 分类器[10],每个类别中的剩余图像用作测试图像集。图像分类性能评价指标为平均准确率(average precision,AP)以及时间开销。相关定义如下:

$$\text{准确率} = \frac{\text{被正确分类的图像数}}{\text{被分类的图像总数}} \times 100\% \tag{16}$$

$$\text{平均准确率} = \frac{\text{各图像类别分类准确率之和}}{\text{图像类别总数}} \times 100\% \quad (17)$$

3.2 实验结果分析

首先,在 Caltech-256 图像集上进行分类实验,分析不同视觉词典规模对本文方法的影响,结果如图 5 所示。从图 5 中不难看出,在一定范围内,随着视觉词典规模的增加目标分类准确率有着明显的提升,然而,当视觉词典规模达到一定数量时,目标分类准确率增长缓慢甚至有所降低,这是因为当词典规模较小时,视觉词典中的单词不足以表达全部的图像内容,而当视觉词典规模过大时会导致词典中有一定的冗余信息,降低视觉词典的语义分辨能力。因此,针对不同的数据只有选择合适的词典规模才能达到较好的分类效果,后续实验中选取词典规模为 1 024。

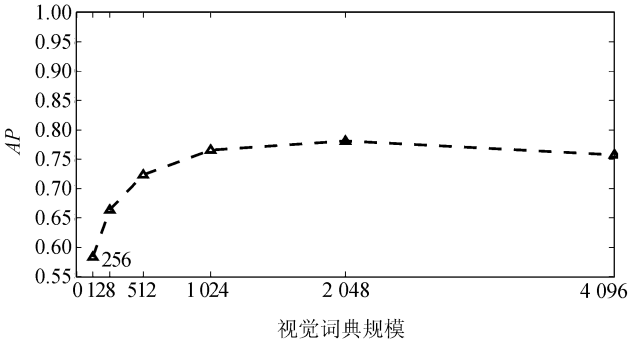

图 5 不同视觉词典规模对目标分类准确率的影响

Fig. 5 Relationship between the average precision and the dictionary scale

其次,为了验证有监督微调对目标分类的效果,在同样的数据和词典规模下分别采用有监督的微调和不进行微调分别进行 10 次分类实验,得分类结果如图 6 所示。

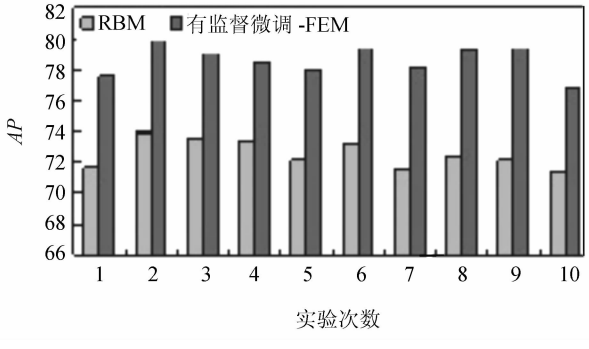

图 6 有监督微调对目标分类结果的影响

Fig. 6 Influence of supervised fine for *AP*

从图 6 可以看出,利用 RBM 对 SIFT 特征进行编码时,在有监督微调之后可以明显地改善目标分类性能。这是因为,有监督微调可以利用误差反向传播的方式更好地调整网络各层参数,这与 3.3 节分析是一致的。

然后,分别在 Caltech-256 数据集和 15-Scenes 数据集上进行分类实验,将本文方法与其他几种经典的目标分类方法包括基于传统视觉词袋模型的方法以及基于稀疏编码模型的方法进行比较,以验证本章方法性能,分类 *AP* 值如表 1 所示。

从表 1 可以看出,ScSPM 方法[10]和 LLC 方法[24]由于得到了图像语义的稀疏表达,其分类性能要优于传统的基于硬分配的视觉词袋模型方法(HA)[25]和基于软分配的视觉词袋模型方法(SA)[26]。本文方法由于利用 RBM 对 SIFT 特征进行深度编码,并利用训练数据的标签类别信息对整个编码网络进行有监督微调,使得视觉词典具有很好的选择性且图像表示向量具有稀疏性,因此,其分类性能要优于 ScSPM 方法以及 LLC 方法。

表 1 不同方法在 Caltech-256 数据集与 15-Scenes 数据集上的分类结果

Tab. 1 Object classification results of different methods for Caltech-256 database and 15-Scenes database

%

目标分类方法	视觉词典规模	Caltech-256	15-Scenes
HA[25]	1 000	66.3	69.85
SA[26]	1 000	70.8	73.89
ScSPM[10]	1 024	71.9	81.5
LLC[24]	1 024	73.5	83.5
本文方法	1 024	79.6	87.6

最后,在数据集 15-Scenes 数据集上进行实验,将本文方法与其他方法之间的目标分类时间效率进行分析对比,得平均训练时间和平均测试时间如表 2 所示。从表 2 可以看出,由于 ScSPM 方法、LLC 方法以及本文方法采用线性 SVM 分类器进行分类,因此,它们的训练时间要远低于 SA 方法。由于 LLC 方法叫较之于 ScSPM 方法作了一些优化工作,所以其训练和测试时间要高于 ScSPM 方法。综合表 1 和 2 可以看出,本文方法可以取得较好分类性能的情况下,消耗最少的分类时间,尤其适用于大规模数据下的目标分类。

表 2 不同方法在数据集 15-Scenes 上的时间效率对比

Tab. 2 Time efficiency comparison of different methods on 15-Scenes database

分类方法	SA[26]	ScSPM[10]	LLC[24]	本文方法
训练时间/s	62.6	0.8	1.4	0.74
测试时间/ms	36.3	5.2	4.5	3.7

4 结　论

为增强视觉词典的语义分辨能力、获取非线性图像表达向量,进而增强图像语义表达能力并降低图像分类时间开销,本文引入深度学习模型,提出了一种基于深度学习编码模型的图像分类方法。该方法利用 RBM 网络逐层对 SIFT 特征进行编码,实现对图像特征的深度编码,并利用训练图像的标签类别对 RBM 编码网络进行有监督微调,使得编码学习得到的视觉词典具有独特性,且图像表示向量具有稀疏性,从而使得在线性 SVM 分类器下依然能够保持较高的目标分类性能。实验结果表明,本文方法的分类效率要远高于其他方法,且能取得较好的分类准确率,在大数据环境下具有很好地适用性。

参考文献:

[1] Sivic J, Zisserman A. Video google: A text retrieval approach to object matching in videos[C]//Proceedings of 9th IEEE International Conference on Computer Vision. Nice: IEEE, 2003: 1470 – 1477.

[2] Otávio A, Penatti B, Fernanda B S, et al. Visual word spatial arrangement for image retrieval and classification[J]. Pattern Recognition, 2014, 47(1): 705 – 720.

[3] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories[C]//Proceedings of the IEEE Conference on Computer Vision and PatternRecognition. New York: IEEE, 2006: 2169 – 2178.

[4] Sharma G, Jurie F. Learning discriminative spatial representation for image classification[C]//Proceedings of the 22nd British Machine Vision Conference. Dundee: WILEY Press, 2011: 1 – 11.

[5] Zhao Chunhui, Wang Ying, Kaneko M. An optimized method for image classification based on bag of words model [J]. Jounal of Electronics & Information Technology, 2012, 34(9): 2064 – 2070. [赵春晖, 王莹, Kaneko M. 一种基于词典模型的图像优化分类方法[J]. 电子与信息学报, 2012, 34(9): 2064 – 2070.]

[6] Zhao Zhongqiu, Ji Haifeng, Gao Jun, et al. Sparse coding based on multi-scale spatial latent semantic analysis for image classification [J]. Chinese Journal of Computers, 2014, 37(6): 1251 – 1260. [赵仲秋, 季海峰, 高隽, 等. 基于稀疏编码多尺度空间潜在语义分析的图像分类[J]. 计算机学报, 2014, 37(6): 1251 – 1260.]

[7] Philbin J, Chum O, Isard M, et al. Lost in quantization: Improving particular object retrieval in large scaleimage databases[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Anchorage: IEEE, 2009: 278 – 286.

[8] van Gemert J C, Veenman C J, Smeulders A W M, et al. Visual word ambiguity[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 7(32): 1271 – 1283.

[9] Weinshall D, Levi G, Hanukaev D. LDA topic model with soft assignment of descriptors to words[C]//Proceedings of the 30th International Conference on Machine Learning. Atlanta: ACM, 2013: 711 – 719.

[10] Yang J, Yu K, Gong Y, 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.

[11] Yu K, Zhang T, Gong Y H. Nonlinear learning using local coordinate coding[C]//Proceedings of Advances in Neural Information Processing Systems. Whistler: MIT Press, 2009: 2223 – 2231.

[12] Zhuang Liansheng, Gao Haoyuan, Liu Chao, et al. Non-negative sparse locally linear coding[J]. Journal of Software, 2011, 22(2): 89 – 95. [庄连生, 高浩渊, 刘超, 等. 非负稀疏局部线性编码[J]. 软件学报, 2011, 22(2): 89 – 95.]

[13] Zhang Ruijie, Wei Fushan. Image scene classification based on fisher discriminative analysis and spars codeing [J]. Journal of Computer-Aided Design & Computer Graphics, 2015, 27(5): 808 – 814. [张瑞杰, 魏福山. 结合 Fisher 判别和稀疏编码的图像场景分类[J]. 计算机辅助设计与图形学学报, 2015, 27(5): 808 – 814.]

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

547.

[15] Srivastava N, Salakhutdinov R. Learning representations for multimodal data with deep belief nets[C]//Proceedings of International Conference on Machine Learning. Edinburgh: ACM, 2012: 113 – 131.

[16] Krizhevsky A, Sutskever I, Hinton G E. Image net classification with deep convolutional neural networks[C]// Proceedings of 2012 Advances in Neural Information Processing Systems (NIPS 2012). Lake Tahoe: ACM, 2012, 1(2): 412 – 425.

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

[18] Ranzatom M A, Poultney C, Chopra S, et al. Efficient learning of sparse representations with an energy-based model[C]//Proceedings of Advances in Neural Information Processing Systems. Vancouver: MIT, 2007: 1137 – 1144.

[19] Srivastava N, Salakhutdinov R. Learning representations for multimodal data with deep belief nets[C]//Proceedings of International Conference on Machine Learning. New York: ACM, 2012: 321 – 332.

[20] Hayat M, Bennamoun M, An S. Deep reconstruction models for image set classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(4): 713 – 727.

[21] Olshausen B A, Field D J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images[J]. Nature, 1996, 381: 607 – 609.

[22] Griffin G, Holub A, Perona P. Caltech-256 object category dataset[M]. Pasadena: California Inst Technol, 2007.

[23] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2006: 2169 – 2178.

[24] Wang J, Yang J, Yu K, et al. Locality-constrained linear coding for image classification[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Sanfrancisco: IEEE, 2010: 3360 – 3367.

[25] Philbin J, Chum O, Isard M, et al. Object retrieval with large vocabularies and fast spatial matching[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis: IEEE, 2007: 1 – 8.

[26] Koniusz P, Mikolajczyk K. Soft assignment of visual words as linear coordinate coding and optimisation of its reconstruction error[C]//Proceedings of 18th IEEE International Conference on Image Processing. Brussels: IEEE, 2011: 2413 – 2416.

(编辑 杨 蓓)

引用格式: Zhao Yongwei, Li Ting, Lin Boyu. Image classification method based on deep learning coding model[J]. Advanced Engineering Sciences, 2017, 49(1): 213 – 220. [赵永威, 李婷, 蔺博宇. 基于深度学习编码模型的图像分类方法[J]. 工程科学与技术, 2017, 49(1): 213 – 220.]