ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 30,No. 1,Jan. 2015,pp. 88－98
DOI:10.16337/j.1004-9037.2015.01.008

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于深度学习的图像自动标注算法

杨　阳　张文生

(中国科学院自动化研究所,北京,100190)

摘　要:图像的自动标注是图像检索领域一项基础而又富有挑战性的任务。深度学习算法自提出以来在图像和文本识别领域取得了巨大的成功,是一种解决"语义鸿沟"问题的有效方法。图像标注问题可以分解为基于图像与标签相关关系的基本图像标注和基于标注词汇共生关系的标注改善两个过程。文中将基本图像标注问题视为一个多标记学习问题,图像的标签先验知识作为深度神经网络的监督信息。在得到基本标注词汇的基础上,利用原始图像标签词汇的依赖关系与先验分布改善了图像的标注结果。最后将所提出的改进的深度学习模型应用于 Corel 和 ESP 图像数据集,验证了该模型框架及所提出的解决方案的有效性。

关键词:机器学习;深度学习;神经网络;图像自动标注

中图分类号:TP39　　**文献标志码**:A

Image Auto-Annotation Based on Deep Learning

Yang Yang, Zhang Wensheng

(Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China)

Abstract: Image auto-annotation is a basic and challenge task in the image retrieval work. The traditional machine learning methods have obtained a lot achievements in this field. The deep learning algorithm has achieved great success in image and text learning work since it is presented, so it can be an efficient method to solve the semantic gap problems. Image auto-annotation can be decomposed into two steps, that is, the basic image auto-annotation based on the relationship between image and tag, and the annotation enhanced based on the mutual information of the tags. In this article, the basic image auto-annotation is viewed as a multi-labelled problem. Therefore the prior knowledge of the tags can be used as the supervise information of the deep neural network. After obtained the image tags, the dependent relationship of the tags is used to improve the annotation result. Finally, the model is tested in Corel and ESP datasets, and results prove that the method can efficiently solve the image auto-annotation problems.

Key words: machine learning; deep learning; neural network; image auto-annotation

引　　言

大数据时代,人们可以访问获取的信息资源呈现出爆炸式的增长,互联网上每天都诞生海量的图像

基金项目:国家自然科学基金重点(U1135005)资助项目。
收稿日期:2014-03-25;**修订日期**:2014-05-12

和视频信息。为了有效地组织、查询与浏览如此大规模的图像资源,图像检索技术应运而生。现有的图像检索方式主要分为两种:基于内容的图像检索(Content-based image retrieval, CBIR)和基于文本的图像检索(Text-based image retrieval, TBIR)。对 TBIR 来说,它要求用户提交文本作为查询,因此,图像检索需要事先获取图像的文本语义信息[1]。图像的 Tag 标签是一种有效获取图像语义信息的方法,它应用广泛,例如在社交媒体中大量的图像被用户标记 Tag 信息。然而互联网上同时存在大量的图像未被标记 Tag 信息,因此人们期望利用含有标签图像通过某种算法自动生成缺失 Tag 信息图像的标签。

虽然图像标签的自动生成是一个十分困难的任务,但是通过学者们的不断努力,取得了丰硕的成果。图像标签生成算法按照标注模型的不同主要分成两大类[2]:基于统计分类的自动图像标注模型和基于概率的自动图像标注模型。基于统计分类的方法是将每一个图像的语义概念都被当作一个类别进行分类,自动图像标注就可以转换成图像的多分类问题。代表方法有:支撑向量机(Support vector machine,SVM)方法[3-5]、二维多分辨率马尔可夫模型(2D Multi-resolution hidden Markov model, 2D MHMMs)[6],贝叶斯点学习机[7]和混合分级模型[8]等。基于概率建模的方法尝试推断图像和语义概念(或关键字)之间的相关性或联合概率分布。Mori[9]等提出了一种利用关键字与"视觉词汇"之间的共现关系来标注图像标签的网格区域算法。近些年来流行的主题模型同样应用在图像自动标注领域,例如,狄迪克雷分配模型[10](Latent Dirichlet allocation model, LDA)和一致 LDA 模型[10](Correspondence LDA)。上述模型中,参数的概率分布相对于真实分布仍然过于简单,但参数的估计过程却相对复杂。受到关联语言模型的启发,一些关联模型相继被应用到图像自动标注领域内,如跨媒体相关模型[11](Cross-media relevance model, CMRM)、连续相关模型[12](Continuous relevance model, CRM)和多重伯努利相关模型[13](Multiple Bernoulli relevance model, MBRM)等。稀疏表示在图像与视频处理领域取得了巨大的成绩,Liu[14]等人应用稀疏编码从多视角的角度出发,分析了不同特征标注的平均正确率,从而选取最适宜标注的特征。Feng[15]等人利用核尺度学习(Kernel metric learning, KML)的方法实现图像的自动标注,此方法因为具有很高的效率,特别适用于海量图像。Hu[16]等人提出了一种两阶段的图像标注方法,第一步移除无关标签,第二步常规标注,能大幅提高图像标注正确率与标注效率。

近些年来,深度学习在图像、文本和语音领域取得了巨大的成功。文献[17]对如何进行基于受限制的玻尔兹曼机(Restrict Boltzmann machine, RBM)的深度神经网络(Deep belief network, DBN)的训练提供详细的指导,并应用于 Minst 手写数字识别。Lecun[18]等人提出的卷积神经网络(Convolution neural network, CNN)是第一个真正意义上的多层结构学习算法。Krizhevsky[19]等人利用多层卷积神经网络进行海量图像的分类工作,取得了较好的成绩。Vincent[20]等人提出利用含有噪声的自编码神经网络(Denoise auto-encoder, DAE)来取代 RBM 模型对深度神经网络进行预训练,在 Minst 手写数字识别等常用数据集上取得了超越 RBM 模型的分类结果。Nitish[21]等在含有 Tag 的图像数据集 MIR Flickr 上,应用 DBN 学习得到图像和文本的语义表示并用于分类,同时该网络可以通过图像的单模信息补充遗失的文本数据。同时,深度学习算法开始逐步应用到图像的标注中去。Socher[22]等人提出了一种基于周期神经元网络的方法,用于场景标注。该方法不再将图像做分割处理,也不依赖于图像的视觉特征,而是直接对图像中的每个像素点赋予类别标签。基于自编码神经网络,Wang[23]等改进了双模模型,实现了图像的标注与搜索的并行处理。

虽然上述深度学习模型在图像标签领域取得了一定的成绩,但是也存在一些不足。首先,多模模型是处理图像与本文融合的通用模型,但是对处理图像标注问题往往精度不够。其次,传统方法通常将图像的不同标签等价处理,而没有考虑到图像标签分布的不均匀性。最后,这些方法没有利用图像标签之间的相关性对标注结果作进一步的改进。本文采取基于判别模型的方法,将图像的标签信息视为图像

的类别信息，利用深度神经元网络构建一个图像自动标注的专用模型。

1 模型方法

神经元网络是处理多分类问题的一个有效方法，然而对于深度神经元网络，若给定随机的初始权重，很难将它优化到一个很好的分类结果，因为在优化过程中，它很容易陷入局部最优中。而深度学习的方法通过权重的逐层预训练，将网络权重首先优化到最优解的附近，然后通过反向传播过程对网络权重进行微调，得到整个神经元网络的最优解。常用的预训练方法有 RBM 模型和 DAE 模型。

1.1 受限玻尔兹曼机

受限玻尔兹曼机(Retricted Boltzmann machines，RBM)是由可见层节点($\boldsymbol{v}\in\{0,1\}^D$，$D$ 为输入层节点数目)与隐藏层节点($\boldsymbol{h}\in\{0,1\}^K$，$K$ 为隐藏层节点数目)构成的双向概率图模型。可见层节点与隐藏层节点间有对称的权连接($\boldsymbol{W}_{ij}$)，而可见层节点之间与隐藏层节点之间没有权连接。该模型定义了隐藏层节点 $\boldsymbol{h}$ 与可见层节点 $\boldsymbol{v}$ 的概率分布，相比于全连接的玻尔兹曼机，当给定可见层节点 $\boldsymbol{v}$ 或者隐藏层节点 $\boldsymbol{h}$ 时，这种特殊的模型可以很方便地计算出节点的条件概率分布。定义该模型的能量函数如下

$$E(\boldsymbol{v},\boldsymbol{h};\theta)=-\boldsymbol{v}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{h}-\boldsymbol{v}^{\mathrm{T}}\boldsymbol{B}-\boldsymbol{h}^{\mathrm{T}}\boldsymbol{A}=$$
$$-\sum_{i=1}^{D}\sum_{j=1}^{K}v_i\boldsymbol{W}_{ij}h_j-\sum_{i=1}^{D}b_iv_i-\sum_{j=1}^{K}c_jh_j \tag{1}$$

式中：$\theta=\{b,c,\boldsymbol{W}\}$ 为模型参数。模型关于可见层节点 $\boldsymbol{v}$ 和隐藏层节点 $\boldsymbol{h}$ 的联合概率分布为

$$P(\boldsymbol{v};\theta)=\frac{1}{Z(\theta)}\exp(-E(\boldsymbol{v},\boldsymbol{h};\theta)) \tag{2}$$

式中：$Z(\theta)$ 为标准化项。当给定可见层节点 $\boldsymbol{v}$ 和隐藏层节点 $\boldsymbol{h}$ 后，二者的条件概率为

$$p(v_i=1\mid\boldsymbol{h},\theta)=\sigma\Big(\sum_{j=1}^{K}h_j\boldsymbol{W}_{ij}+b_i\Big) \tag{3}$$

$$p(h_j=1\mid\boldsymbol{v},\theta)=\sigma\Big(\sum_{i=1}^{D}v_i\boldsymbol{W}_{ij}+c_j\Big) \tag{4}$$

式中：$\sigma(x)=1/(1+\exp(-x))$ 为逻辑斯蒂克函数。

1.2 高斯-伯努利受限玻尔兹曼机

当模型中的可见层节点为实数($\boldsymbol{v}\in\mathbf{R}^D$)，隐藏层节点($\boldsymbol{h}\in\{0,1\}^K$)为二元随机数时，原有模型失效。于是定义高斯-伯努利受限玻尔兹曼机(Gaussian-Bernoulli retricted Boltzmann machines，GRBM)模型，其能量函数为

$$E(\boldsymbol{v},\boldsymbol{h};\theta)=\sum_{i=1}^{D}\frac{(v_i-b_i)^2}{2\sigma^2}-\sum_{i=1}^{D}\sum_{j=1}^{K}v_i\boldsymbol{W}_{ij}h_j-\sum_{j=1}^{K}c_jh_j \tag{5}$$

式中：$\theta=\{b,c,W,s\}$ 为模型参数，该模型下可见层与隐藏层节点的条件概率为

$$p(v_i=x\mid\boldsymbol{h})=\frac{1}{\sqrt{2\pi}s_i}\exp\Big(-\frac{(x-s_i\sum_{j=1}^{K}h_j\boldsymbol{W}_{ij})^2}{2s_i^2}\Big) \tag{6}$$

$$p(h_j=1\mid\boldsymbol{v})=\sigma\Big(\sum_{i=1}^{D}\boldsymbol{W}_{ij}\frac{v_i}{s_i}\Big) \tag{7}$$

受限波尔兹曼机通常采用 Hinton[17] 等人提出的对比离差(Contrastive divergence，CD)方法进行近似求解，得到模型参数。

1.3　带噪声的自编码神经网络

自编码神经网络(Auto-encoder,AE)是一种无监督的学习算法。自编码神经网络尝试学习一个恒等函数 $y=g_{\theta}(f(x))\approx x$,使得输出 y 接近于输入 x。如式(8)所示,模型通过优化最小损失函数 $L(x^i,y^i)$,学习得到模型的参数 θ^*,其中 $f(x)=\sigma(\boldsymbol{W}x+b)$,$y=g_{\theta'}(h)=\sigma(\boldsymbol{W}'\boldsymbol{h}+b')$,$\sigma$ 是逻辑斯蒂克函数。当权重紧致时,变换的参数 θ,θ' 对称。Vincent[20] 认为如果网络的输入数据完全随机,比如每一个样本 x_i 都是一个跟其他样本完全无关的独立同分布高斯随机变量,那么这一压缩表示将会非常难学习。但是如果输入数据中隐含着一些特定的结构,比如某些输入特征彼此相关,那么这一算法就可以发现输入数据中的这些相关性。事实上,这一简单的自编码神经网络通常可以学习出一个跟主元分析(Principal component analysis,PCA)结果非常相似的输入数据的低维表示。

$$\theta^*,\theta'^*=\arg\min_{\theta,\theta'}\frac{1}{n}\sum_{n=1}^{N}L(x^i,y^i)=$$

$$\arg\min_{\theta}\frac{1}{n}\sum_{n=1}^{N}L(x^i,g_{\theta'}(f_{\theta}(x^i))) \tag{8}$$

为了提高模型参数对输入数据的鲁棒性,Vincent[20] 提出了带噪声的自编码神经网络(Denoise auto-encoder,DAE)如图 1 所示。在原有自编码神经网络的基础上,对输入数据 x 加入部分噪声 q_D,得到含有噪声的数据 $\bar{x}$,然后从含有噪声的数据 $\bar{x}$ 重建得到一个干净的输入数据 x,因此恒等函数由 $g_{\theta'}(f_{\theta}(x))\approx x$ 变为 $g_{\theta'}(f_{\theta}(\bar{x}))\approx x$,优化目标变为式(9)。常用的污染函数 q_D 有高斯随机噪声、白噪声、椒盐噪声等。模型参数的学习可用随机梯度下降法得到,为了让模型更好的学习得到输入数据的特征,常常在模型中加入 $L2$ 正则化与稀疏惩罚。

$$\theta^*,\theta'^*=\arg\min_{\theta}\frac{1}{n}\sum_{n=1}^{N}L(x^i,g_{\theta'}(f_{\theta}(\bar{x}^i))) \tag{9}$$

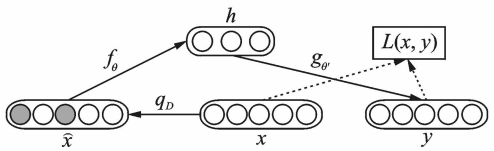


图 1　含有噪声的自编码神经网络

Fig. 1　Denoising auto-encoder model

1.4　模型结构

传统的 BP 神经元网络采用单一隐藏层对数据进行建模,深度学习方法为了解决更复杂的问题,通常采用多层隐藏层神经元网络。为了避免神经元网络求解陷入局部最优的问题,首先采用 RBM 或者 DAE 模型对网络权重进行预训练,然后将预训练得到的每层权重值作为反向传播算法的权重初始值赋予整个神经元网络,最后采用反向传播算法更新整个网络的权重。

如图 2 所示,本文采用 3 层深度神经元网络架构,输入层单元输入图像的特征 v,将图像的标签信息作为网络的输出节点 t。当采用 RBM 模型时,由于神经网络输入单元 $v\in\mathbf{R}^N$(N 为输入图像特征维数),故应当采用 GBRBm 作为深度神经网络的第一层网络结构 h_1,随后两层采用常规 RBM 作为深度网络的第二层 h_2 和第三层 h_3 网络;当采用 DAE 模型时,第一层自编码神经网络的反向激励函数运用线性函数,第二层 h_2 和第三层 h_3 网络反向激励函数应用逻辑斯蒂克函数。

神经网络模型在处理分类问题时,设定监督向量的维数目为类别数目 M,对应于样本所属类别 k,

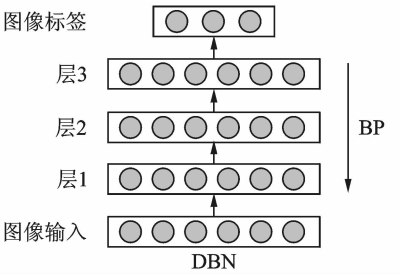


图 2 深度神经元网络模型

Fig. 2 Model of deep belief network

那么输出层的第 k 维为 1,其余维数为零。在反向传播算法中,计算神经网络实际输出与监督向量的之间的差值来衡量网络的收敛程度,当满足训练次数要求时停止训练。训练完成后,将测试数据组输入神经网络,取输出层最大节点的位置为样本的预测类别。在处理多分类问题时,可以将样本多个类别信息所对应的监督向量的维数均设为 1,输出层的激励函数由 softmax 函数替换为逻辑斯蒂克函数,对输出层的结果做排序,排序靠前的类别为该神经网络对样本类别的预测结果。

虽然图像标注问题类似于多分类问题,一个图像可能所属多个标签,但是与常见的多分类问题有很大的不同。多分类问题对应的类别信息通常是均匀分布的,也就是说每个类别所属的图像数量通常是均匀分布。然而图像标注问题的标注信息通常不是均匀分布的,某个标签可能所属的图像较多,也可能较少。例如"天空"、"大海"所属的图像一般较多,而"猫"、"城墙"所属的图像数目会很有限。倘若同等考虑不同的标签信息,那么标注频率低的标签会淹没于标注频率高的标签之中,而无法对给定的图像给予准确的标注。

针对图像标签分布不均匀的问题,本文将图像的标签频率引入到监督信息中,通过改善神经网络的监督向量改进模型的准确度。如式(10)所示,新的监督信息 $\tilde{y}$ 在原有监督信息 y 的基础上除以标签所属图像的和 n。同时为了避免监督信息 $\tilde{y}$ 过小,再除以 $\tilde{y}$ 中的最大值做标准化处理。通过上述处理保证了图像的低频标签有较大的返回值,能够有效改善图像的标注精度。

$$\tilde{y}=\frac{(y/n)}{\max(y/n)} \tag{10}$$

为了增加标注的准确性,更加有效地返回低频标签。本文改进了神经网络的损失函数,每一维的监督信息不再平等对待。一方面,含有低频标签的样本具有更大的权重,另一方面,低频标签所对应的监督信息具有更大的权重。如式(11)所示

$$\tilde{L}=f_1 f_2 L \tag{11}$$

$$f_1=f_2^T=1/n \tag{12}$$

式中:L 为原损失函数,f_1 为考虑词频的样本的权重,f_2 为考虑词频的损失函数的权重。通常权重取标签所含图像的总数的倒数。对于 f_1 通常取样本对应具有最低词频的标签的权重。

2 图像标注改善

利用深度学习模型得到图像标注信息,主要是利用了图像间的视觉相似性,但由于语义鸿沟的存在,每幅图像很难保证得到的标签与原图像的语义一致性。文献[24]指出依据朴素贝叶斯的思想,利用文本的先验与后验关系,可以提高文本的分类结果。因此本文将标签的分布作为先验,并引入图像标签

之间的关系作为后验来改善算法得到的标注结果。

2.1　基于共生关系的图像标注改善

图像标注的目的是为了得到反映图像语义信息的一组相关词汇,而词汇间存在着各式各样的语义关系。一般来讲,在训练集中,同一个样本内同时出现的词汇具有较强的语义相关性。这是由于共生频率高的词汇往往代表了两个关系密切的概念或者事物,从而存在很大的可能性被标注在同一幅图像中。生活中有很多这样的例子,"森林"与"树木","城市"与"建筑"等。因此利用共生词汇在同一幅图像出现的相关性可以有效地提供词汇之间的语义相关信息,从而提高图像标注的准确率。然而简单的依据共生关系进行词频数的统计,不能有效地考虑到不同词汇的不同特性。因此参照文献中给出的共生关系的度量,通过式(13)来衡量词汇的共生关系

$$K(v_1,v_2)=K_c(v_1,v_2)\times\frac{1}{n_1} \tag{13}$$

式中:v_1,v_2,为词汇,$K_c(*)$为二者共生出现的次数,n_1 为包含 v_1 作为标注的图像数目。根据以上定义可以发现,$K(v_1,v_2)$和 $K(v_2,v_1)$并不相等,说明它们具有不对称性。考虑到 v_1,v_2 是两种出现频率差异较大的词汇,若 v_1 与 v_2 之间存在着一定的相互依赖关系,比如 v_1 依赖于 v_2 存在。那么比较容易从 v_1 得到 v_2,但很难从 v_2 的存在来推断 v_1 是否存在。例如词汇"水"和"鱼"之间的关系,很容易能从"鱼"中推断"水"的存在,但给定"水"很难断定"鱼"是否存在,因为"水"与更多是事物相关联。

2.2　基于词频的图像标注改善

在图像标注的改善中,本文同样考虑到了词频对标注结果的影响。因此,定义词频系数 $K_F=1/n$ 来进一步增强低频词汇的返回率。那么对于深度学习模型得到的神经网络的实际输出 R,通过式(14)得到模型的最终标注结果为 R^*。其中,α 用来平衡基本图像标注与标注改善后的结果。

$$\begin{aligned} R^* &= \alpha R+(1-\alpha)R' \qquad \alpha\in(0,1)\\ R' &= KK_FR \end{aligned} \tag{14}$$

3　实验分析

为了验证本文方法的有效性,并同其他算法进行合适的比较,采用了图像自动标注工作中普遍使用的 Corel 和 ESP 图像集作为实验数据集。深度神经网络的算法用 Matlab 实现。

3.1　数据集

Corel-5K 图像集共包含科雷尔(Corel)公司收集整理的 5 000 幅图像。该数据集可用于科学图像实验的分类、检索等,Corel-5k 数据集是图像实验的事实标准数据集。Corel 数据集的标签信息的字典长度为 260,每幅图像包含 1～5 个标签,图像的平均标签数目为 3.5 个。在实验中选取4 000个数据作为训练集,500 个数据作为模型参数的评价集,500 个数据作为测试集。

ESP game 图像集包含了 20 770 幅图像。它覆盖面很广,包括绘画与个人肖像等。所有的图像被标记为 268 个标签,其中标签最多的图像有 15 个标签,平均每幅图像有 4.6 个标签。

3.2　特征提取

本文选用高层视觉特征作为深度神经网络的输入信息。为了与相关实验进行对比,选用图像的全局特征包括 1 个 Gist 特征;局部特征包括 SIFT 描述子和 HUE 描述子。所有特征均以词包的形式存储,组合特征一共 4 912 维。计算前,对所有输入数据进行标准化。

3.3 评价指标

本文选用基于分类学习的方法来实现图像的自动标注，因此首先选用分类正确率来衡量模型的训练程度。定义图像原有的标签数目为 N，正确匹配的标签数目为 n，那么模型对图像标注准确率为 n/N。然后对所有图像求均值，那么得到数据集的标注准确率。为了衡量模型的训练情况，给出了训练集的图像准确率与测试集的图像准确率。

本文采用了最常见的几个指标来衡量图像标注方法的性能。正确率与召回率是以某单一关键词作为查询，在标注好的测试图像集上进行检索，假设标注正确的图像数为 N_c，可检索到的所有图像数为 N_s，测试集中与该词相关的所有图像数为 N_r，于是可得

$$\text{Precision}=\frac{N_c}{N_s} \qquad \text{Recall}=\frac{N_c}{N_r} \tag{15}$$

本文对每幅图像标注 5 个最相关的关键词，然后针对每个词的正确率 P 与召回率 R 求均值。为了更加容易对比实验结果，选取了上述两个指标的联合函数($F1=2P*R/(P+R)$)作为另一个评价指标。除此之外，还统计了被正确标注的词汇数量，即至少被正确标注一次的关键词数量，这一数值反映了标注算法对词汇的覆盖程度，记为 $N+$。每幅图像都被标记为 5 个关键词，无论图像本身的标签数目或多或少。因此，即使一个模型可以对图像本身的关键词给予精确的预测，仍无法得到一个完美的正确率和召回率。

3.4 实验结果

(1) 基于分类的图像标注准确率

本文首先从多分类的角度来衡量模型的标注准确率。表 1 给出了基于 RBM 模型与 DAE 模型的深度学习模型对图像的标注结果。为了衡量模型的性能，本文分别给出了训练集的标注准确率与测试集的标注准确率。由表 1 中可以看出，两个模型对训练集具有较好的学习能力，图像的标注准确率为 1，意味着对于训练集，预测的图像标签内容与给定的标签内容完全相符。但由于深度学习算法对于小样本训练集容易造成过拟合，导致模型在测试集上的准确率表现不佳。RBM 模型与 DAE 模型泛化能力有差异，对于测试集，DAE 的结果明显好于 RBM 模型得到的结果。

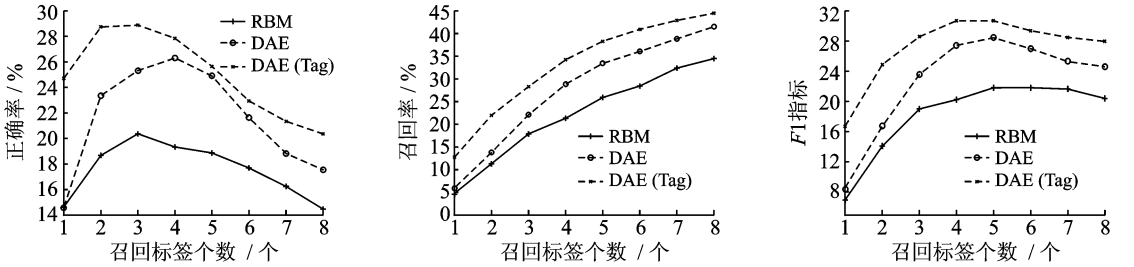
表 1 不同模型的图像标注准确率

Table 1 Annotation accuracy of different models

名称	训练	测试
RBM	1	0.41
DAE	1	0.49

(2) 不同标签数目对图像标注的影响

为了进一步对比不同方法的标注性能与标签数目的关系，本文还考虑到对于不同图像标签数目对图像自动标注的影响，如图 3 所示。本文给出了基于 RBM 和 DAE 模型预训练的传统深度神经网络模型，及改进监督信息后的深度神经网络模型得到的图像标签的准确率、召回率和 $F1$ 数值随返回标签数目的变化曲线。从图 3 中可以得到：(1)在评价图像标注指标的三个数据上，基于 DAE 模型的深度神经网络的图像标注结果明显好于基于 RBM 模型的结果。针对图像标注问题，改进的深度神经网络得到的标注结果最好。(2)图像标注的正确率随着标签数目的增加，先上升后下降，而召回率随着标签数目的增加而不断上升。二者综合指标 $F1$ 在标签数目为 5 时，达到最优。这是因为在返回标签数目较少时，返回得到的正确标签数目占图像原标签数目的比率在不断上升，当超过一定限度时，模型会返回一



(a) 不同标签数目对精度的影响　(b) 不同标签数目对召回率的影响　(c) 不同标签数目对$F1$指标的影响

(a) Precise result of different return tag number　(b) Recall result of differnet return tag number　(c) $F1$ result of different return tag number

图 3　不同返回标签数目对标注正确率与召回率的影响

Fig. 3　Annotation results of different return tag number

些无关标签,图像标注的准确度会下降;而模型返回正确标签的数目仍在不断增加,从而召回率不断上升。当二者的综合指标 $F1$ 达到最大时,得到的图像标注最有意义。

(3) 不同图像自动标注算法的比较与分析

本文对比了深度神经网络方法与其他典型图像自动标注算法的结果,来验证基于深度学习方法的有效性。这里涉及到的方法包括:leastsquares,CRM[12],InfNet[13],NPDE[21],SML[8],MBRM[13],TGLM[25],JEC[26]。表 2 给出了深度学习方法与其他方法的在 Corel 数据集上的实验结果详细对比,表 3 给出了在 ESP 数据集上的试验结果详细对比。由此可以得出:

(1)在描述由图到词的关系时,基本图像标注过程更适合按照多标记问题来解决,而非多类别分类问题来假设它的先验分布。具体而言,当采用多分类问题来假设此的先验分布时,采用传统的深度学习算法得到的图像标签的效果与 JEC 的方法相当(标注准确率略低,召回率提高)。当采用多标记问题来处理图像标签的先验分布时,图像的标注效果有明显的提高。在 Corel 数据集上,相对于 JEC 方法,精度提高了 7%,召回率提高 22%,返回标签词汇数目提高了 14%。在 ESP 数据集上,在精度不变的情况下,召回率提高 30%,返回标签的数据提高了 10%。说明基于图像标签先验分布的深度学习模型可以更好地解决图像的自动标注问题。

(2)DAE 给出在基于图像标签先验知识的深度学习模型得到的图像标签经过“标注性能改善”后得到的图像标注结果。实验结果表明,通过考虑标注词汇的相关性与词频得到的图像标注标签具有最优的标注效果。在 Corel 数据集上,它在略微降低标注精度的情况下,大幅提高标签的召回率(26%)与返回标签的数目(25%)。同样在 ESP 数据集上,精度也略微的下降,但召回率(40%)和返回标签的数目(16%)也得到的很大提高。在考虑正确率和召回率时需要做一个平衡,当过度考虑低频词汇的召回时,会对整体标注的正确率造成不利影响。

(3)图像自动标注在实际中的表现

本文给出了图像自动标注的实际结果,每幅图像根据模型给出最靠前的 5 个标签作为图像的生成标签,并按照标注评价指标分为两个层次,标注准确率高,标注准确率低。从图 4 中可以看出对于标注准确率较高的图像,模型自动标注得出的标签不但与原标签匹配的较好,而且得到的新的标签能对原图像标签进行有益的补充,能够更加准确地描述原图像的语义信息。对于标注表现不好的图像,模型得到的标签与原图像相关程度低,甚至有些与原图不符,但是也存在部分标注补充的原图像的语义信息。如第 2 排图像的第 3 幅图像中的模型生成的标签“日落”和“水”,第 4 幅图像中生成的标签“草”和“树”等与图像本身的语义相符。

表 2 Corel 数据深度学习与其他实验结果的对比

Table 2 Comparison of annotation results using deep learning to other methods in Corel dataset

名称	*P*	*R*	*F*1	*N*+
Least Squares	29	32	30	125
CRM	16	19	17	107
Inf Net	17	24	20	112
NPDE	18	21	19	114
SML	23	29	26	137
MBRM	24	25	24	122
TGLM	25	29	27	131
JEC	27	32	29	139
RBM	19	26	22	120
DAE	25	34	29	141
DAE(Tag)	29	39	34	159
DAE(Enhence)	24	43	30	174

表 3 ESP 数据深度学习与其他实验结果的对比

Table 3 Comparison of annotation results using deep learning to other methods in ESP dataset

名称	*P*	*R*	*F*1	*N*+
Least Squares	35	19	25	215
MBRM	18	19	18	209
JEC	24	19	21	222
DAE	21	20	21	223
DAE(Tag)	23	25	24	244
DAE(Enhence)	20	27	22	257

原始标注	sky, sun, clouds	sky, jet, plane	Coral, ocean, reefs	wall, cars, tracks formula	sky, water, beach, sand	swimmers, pool, people, water
自动标注	sky, sun, clouds, water, sunset	sky, jet, plane smoke, formation	Coral, ocean, reefs, flowers, waves	wall, cars, tracks formula, bear	sky, water, beach, sand, ships	swimmers, pool, people, water

(a) 好的自动标注

(a) Good auto-annotation

原始标注	flowers, petals, grass	sky, road, cars tower	light, shore	water, plane, lion	ice, plants, frost	people
自动标注	tree, ice, plants ocean, frost	leaf, ice, close-up plants, frozen	city, sky, water, bridge, sunset	tree, grass, field hut, landscape	branch, leaf, close-up, display	flowers, close-up costume, tulip

(b) 差的自动标注

(b) Bad auto-annotation

图 4 图像自动标注的实际效果

Fig. 4 Real effect of image automatic annotation

4　结束语

针对图像自动标注问题,本文将图像标记问题分解为基于图像与标签关系的基本图像标注和基于标签间相互关系的标注改善。在基本图像标注过程中,本文将图像标注视为基于图像标签先验知识的多标记的问题,利用图像标签的词频信息改进深度学习模型的监督信息。在标注改善过程中,利用标记词汇的共生关系与词频先验知识来改善已经得到的图像标记结果。最后,选取合适的数据集 Corel 和 ESP,并提取图像的语义特征作为模型的输入,对图像标注问题进行时实验。实验结果表明,(1)相比于 RBM 模型,DAE 模型能够更好地优化深度学习模型;(2)图像标注问题更适合用于基于标签先验的多标记模型,而非多分类模型;(3)在得到图像标签的基础上,利用标签的共生关系与先验知识可以有效改善图像标注的结果;(4)本文提出的模型对于解决图像标注问题有效。

参考文献:

[1] 卢汉卿,刘静. 基于图学习的自动图像标注[J]. 计算机学报,2008, 31(9): 1629-1639.
Lu Hanqing, Liu Jing. Image annotation based on graph learning [J]. Chinese Journal of Computers, 2008, 31(9): 1629-1639.

[2] 许红涛,周向东,向宇,等. 一种自适应的 Web 图像语义自动标注方法[J]. 软件学报,2010, 21(9):2183-2195.
Xu Hongtao, Zhou Xiangdong, Xiang Yu, et al. Adaptive model for web image semantic automatic annotation [J]. Journal of Software, 2010, 21(9): 2183-2195.

[3] Cusano C, Ciocca G, Schettini R. Image annotation using SVM[C]// International Society for Optics and Photonics. [S. l.]: SPIE, 2004: 330-338.

[4] Gao Y, Fan J, Xue X, et al. Automatic image annotation by incorporating feature hierarchy and boosting to scale up SVM classifiers[C]// Proceedings of the 14th Annual ACM International Conference on Multimedia. [S. l.]: ACM, 2006: 901-910.

[5] Verma Y, Jawahar C V. Exploring SVM for image annotation in presence of confusing labels[C]// Proceedings of the 24th British Machine Vision Conference. London, British: BMVC, 2013: 25. 1-25. 11.

[6] Li J, Wang J Z. Automatic linguistic indexing of pictures by a statistical modeling approach [J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2003, 25(9): 1075-1088.

[7] Chang E, Goh K, Sychay G, et al. CBSA: content-based soft annotation for multimodal image retrieval using Bayes point machines [J]. Circuits and Systems for Video Technology, IEEE Transactions on, 2003, 13(1): 26-38.

[8] Carneiro G, Chan A B, Moreno P J, et al. Supervised learning of semantic classes for image annotation and retrieval [J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2007, 29(3): 394-410.

[9] Mori Y, Takahashi H, Oka R. Image-to-word transformation based on dividing and vector quantizing images with words[C]// First International Workshop on Multimedia Intelligent Storage and Retrieval Management. Florida, USA: ACM, 1999.

[10] Blei D M, Jordan M I. Modeling annotated data[C]// Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. Toronto, Canada: ACM, 2003: 127-134.

[11] Jeon J, Lavrenko V, Manmatha R. Automatic image annotation and retrieval using cross media relevance models [J]. ACM, 2003: 119-126.

[12] Lavrenko V, Manmatha R, Jeon J. A model for learning the semantics of pictures[C]// Advances in Neural Information Processing Systems. British Columbia, Canada: NIPS, 2003.

[13] Feng S L, Manmatha R, Lavrenko V. Multiple Bernoulli relevance models for image and video annotation [J]. 2004, 2: 1002-1009.

[14] Liu W, Tao D, Cheng J, et al. Multi-view Hessian discriminative sparse coding for image annotation[J]. Computer Vision and Image Understanding, 2014, 118: 50-60.

[15] Feng Z, Jin R, Jain A. Large-scale Image Annotation by Efficient and Robust Kernel Metric Learning[C]// International Conference on Computer Vision. Darling Harbour, Sydney: IEEE, 2013: 1609-1616.

[16] Hu J, Lam K M. An efficient two-stage framework for image annotation[J]. Pattern Recognition, 2013, 46(3): 936-947.

[17] Hinton G. A practical guide to training restricted Boltzmann machines [J]. Momentum, 2010, 9(1): 926.

[18] 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.

[19] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]// Advances in Neural Information Processing systems. Nevada, USA: NIPS, 2012: 1097-1105.

[20] Vincent P, Larochelle H, Lajoie I, et al. Stacked denoising auto-encoders: Learning useful representations in a deep network with a local denoising criterion [J]. The Journal of Machine Learning Research, 2010, 9999: 3371-3408.

[21] Srivastava N, Salakhutdinov R. Learningrepresentationsfor multimodal data withdeep beliefnets [C] // International Conference on Machine Learning Workshop. Edinburgh, Scotland: ICML, 2012.

[22] Socher R, Lin C C, Manning C, et al. Parsing natural scenes and natural language with recursive neural networks[C]// Proceedings of the 28th International Conference on Machine Learning. Washington, USA: ICML, 2011: 129-136.

[23] Wang W, Ooi B C, Zhang D. Effective multi-modal retrieval based on stacked auto-encoders[C] // Proceedings of International Conference on Very Large Data Bases. Hangzhou, China: VLDB, 2014: 649-660.

[24] 邸鹏,段利国. 一种新型朴素贝叶斯文本分类算法[J]. 数据采集与处理, 2014, 29(1): 71-75.
Di Peng, Duan Liguo. New naive Bayes test classification algorithm [J]. Journal of Data Acquisition and Processing, 2014, 29(1): 71-75.

[25] Yavlinsky A, Schofield E, Rüger S. Automated image annotation using global features and robust nonparametric density estimation[C]// Conference on Image and Video Retrieval. Berlin Heidelberg: Springer, 2005: 507-517.

[26] Liu J, Li M, Liu Q, et al. Image annotation via graph learning [J]. Pattern Recognition, 2009, 42(2): 218-228.

[27] Makadia A, Pavlovic V, Kumar S. A new baseline for image annotation[C]// The 10th European Conference on Computer Vision. Berlin Heidelberg: Springer, 2008: 316-329.

作者简介:**杨阳**(1986-),男,博士研究生,研究方向:跨域异构数据处理,深度学习,E-mail: yang.yang@ia.ac.cn;**张文生**(1965-),男,研究员,研究方向:模式识别与机器学习,大数据处理,E-mail: wensheng.zhang@ia.ac.cn。