软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(7):1862−1880 [doi: 10.13328/j.cnki.jos.005112]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于数据均衡的增进式深度自动图像标注*

周铭柯[1,2], 柯 逍[1,2], 杜明智[1,2]

[1](福州大学 数学与计算机科学学院,福建 福州 350116)

[2](福建省网络计算与智能信息处理重点实验室(福州大学),福建 福州 350116)

通讯作者: 柯逍, E-mail: kex@fzu.edu.cn

摘 要: 自动图像标注是一个包含众多标签、多样特征的富有挑战性的研究问题,是新一代图像检索与图像理解的关键步骤.针对传统的基于浅层机器学习标注算法标注效率低下、难以处理复杂分类任务的问题,提出了基于栈式自动编码器(stacked auto-encoder,简称 SAE)的自动图像标注算法,提升了标注效率和标注效果.主要针对图像标注数据不平衡问题,提出两种解决思路:对于标注模型,提出一种增强训练中低频标签的平衡栈式自动编码器(B-SAE),较好地改善了中低频标签的标注效果.并在该模型的基础上提出一种分组强化训练 B-SAE 子模型的鲁棒平衡栈式自动编码器算法(RB-SAE),提升了标注的稳定性,从而保证模型本身具有较强的处理不平衡数据的能力;对于标注过程,以未知图像作为出发点,首先构造未知图像的局部均衡数据集,并判定该图像的高低频属性以决定不同的标注过程,局部语义传播算法(SP)标注中低频图像,RB-SAE 算法标注高频图像,形成属性判别的标注框架(ADA),保证了标注过程具有较强的应对不平衡数据的能力,从而提升整体图像标注效果.通过在 3 个公共数据集上进行实验验证,结果表明,该方法在许多指标上相比以往方法均有较大提高.

关键词: SAE(stacked auto-encoder);深度学习;数据均衡;图像标注;语义传播
中图法分类号: TP391

中文引用格式: 周铭柯,柯逍,杜明智.基于数据均衡的增进式深度自动图像标注.软件学报,2017,28(7):1862−1880. http://www.jos.org.cn/1000-9825/5112.htm

英文引用格式: Zhou MK, Ke X, Du MZ. Enhanced deep automatic image annotation based on data equalization. Ruan Jian Xue Bao/Journal of Software, 2017,28(7):1862−1880 (in Chinese). http://www.jos.org.cn/1000-9825/5112.htm

Enhanced Deep Automatic Image Annotation Based on Data Equalization

ZHOU Ming-Ke[1,2], KE Xiao[1,2], DU Ming-Zhi[1,2]

[1](College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350116, China)

[2](Fujian Provincial Key Laboratory of Networking Computing and Intelligent Information Processing (Fuzhou University), Fuzhou 350116, China)

Abstract: Automatic image annotation is a challenging research problem involving lots of tags and various features. Aiming at the problem that the image annotation based on the traditional shallow machine learning algorithm has low efficiency and is difficult to apply to complex classification task, this paper proposes an automatic image annotation algorithm based on stacked auto-encoder (SAE) to improve both efficiency and effectiveness of annotation. In this paper, two types of strategies are proposed to solve the main problem of unbalanced data in image annotation. For the annotation model itself, to improve the annotation effect of low frequency tags, a balanced and stacked auto-encoder (B-SAE) that can enhance training for low frequency tags is proposed. Based on this model, a robust balanced

* 基金项目: 国家自然科学基金(61502105); 福建省科技引导性项目(2017H0015); 福建省中青年教师教育科研项目(JA15075)

Foundation item: National Natural Science Foundation of China (61502105); Technology Guidance Project of Fujian Province, China (2017H0015); Natural Science Foundation of Fujian Provincial Education Department, China (JA15075)

收稿时间: 2016-01-04; 修改时间: 2016-05-18; 采用时间: 2016-06-12; jos 在线出版时间: 2016-10-11

CNKI 网络优先出版: 2016-10-12 16:26:59, http://www.cnki.net/kcms/detail/11.2560.TP.20161012.1626.026.html

© 中国科学院软件研究所 http://www.jos.org.cn

and stacked auto-encoder algorithm (RB-SAE) is proposed to increase the annotation stability through enhanced training by group in sub B-SAE model. This strategy ensures that the model itself has a strong ability to deal with the unbalanced data. For the annotation process, taking the unknown image as the starting point, the local equilibrium dataset of the unknown image is constructed, and the high and low frequency attribute of the image is discriminated to determine the different annotation process. The local semantic propagation algorithm (SP) annotates the low frequency images and the RB-SAE algorithm annotates the high frequency images. The framework of attribute discrimination annotation (ADA) is formed to improve the overall image annotation effect. This strategy ensures that the labeling process has a strong ability to deal with unbalanced data. Experimental results generated from three public data sets show that many indicators in the presented model are all improved comparing with the previous models.

Key words: SAE (stacked auto-encoder); deep learning; balance data; image annotation; semantic propagation

图像标注是图像检索与图像理解的关键步骤.它利用已标注图像集学习语义概念与视觉特征的关系模型,并用此模型给无标注的图像加上能够反映图像内容的语义关键词.早期自动图像标注的研究主要采用概率统计的方法.比如,“词同现”模型[1]利用网格分割方法将图像分割成子图像单元,并建立每一类图像单元和它所对应的关键词的概率分布,通过图像单元预测关键词.翻译模型[2]对分割后的图像区域进行聚类,构造视觉关键词词汇表,从而将图像的标注问题看作是从图像视觉关键词到语义关键词的翻译过程.相关模型[3−5]认为同一关键词的视觉特征具有一致性,图像可以被分割成一些带有一定语义含义的局部区域,并通过建立视觉关键词与语义关键词的联合概率分布进行标注.这类方法虽然可以很方便地扩展到大数据集,但总体标注效果不够理想.

近几年,图像标注的研究主要集中在两类方法:基于机器学习的标注方法和基于图的标注方法.文献[6,7]采用 *K*NN 的思想,通过计算图像间的视觉相似度来获得相似度最高的 k 幅图,并用标签传播算法获取关键词.Lu 等人[8]利用支持向量机(SVM)训练多类分类器,将图像的底层特征映射为高层语义模型特征,并用核函数方法完成标注过程,相比普通底层特征,该方法提升了自动标注的综合效果.Qiu 等人[9]使用支持向量机(SVM)对部分可确定区域赋予语义标签,并利用区域位置关系帮助标注未知区域.Wang 等人[10]构建结合特征图和标签图的双向图,并用随机游走算法产生图像到顶点的相关性,该模型不仅可用于图像标注,也可用于语义图像检索.Gao 等人[11]利用训练集中的隐含信息(比如,部分标签和多种特征信息)学习最优图结构,从而精确地建立数据点间的关系,大幅提升了标注的平均准确率.Tian 等人[12]针对数据集的弱标签性提出一种直推式标签填充算法来构造局部语义邻域,并用多标签语义嵌入的领域最大边际学习算法提升标注效果.Amiri 等人[13]根据每种特征的视觉形态构建特殊的子图,再将子图联结起来形成超级图,最后用超级图进行图像标注.此外,矩阵和张量的分解理论也被引入到图像标注的问题中来:Kalayeh 等人[14]提出了加权多视图非负矩阵分解的图像标注方法.利用不同特征间潜在的信息构建系数矩阵来解决特征融合不灵活问题,使标注模型扩展训练数据成为可能.Tariq 等人[15]将张量引入到图像标注模型中,利用张量分解来解决图像上下文的非监督的特征独立量化问题,并结合上下文先验知识来预测图像标签,实验结果验证了该方法的有效性.

上述方法虽然在理论和效果等不同层面上取得了进展,但总体而言,这些方法还是围绕于传统的分类、回归等浅层结构学习算法展开研究,利用这些方法处理图像标注任务依然存在如下主要问题.

1) 复杂度较高,标注效率低下.如,基于 *K*NN 模型[7]的标注方法不仅需要两两计算训练集中图像的相似度,还要计算每一张测试图像与所有训练图像的相似度,标注效率低下;基于图模型[10,11,13]的标注方法需要构造复杂的图结构,结点间相关性的度量方式多种多样,且增加结点需要大量的遍历操作,较难应用在真实图像环境中.基于矩阵分解[14,15]的标注方法需要根据训练样本数构建矩阵分解模型,计算复杂度随样本数的增加而呈非线性增长,难以处理大规模数据.

2) 模型鲁棒性较弱,难以有效训练不平衡数据.如,基于 SVM 模型[8,9]的标注方法需要根据类标签规模训练成百甚至上千个分类器,整体标注效果依赖于每个分类器的效果,因图像标注任务的标签分布极不均衡,导致每个分类器分类效果差异较大,极大地影响了整体标注效果;文献[12]虽然提出了一种改善弱标签的算法,但该方法存在两点不足:(1) 直推式算法填充的标签有可能是不正确的标签,在此基础上进行标注,有时反而会影响标注效果;(2) 算法执行过程需要构造数据集的特征矩阵来计算,复杂度较高,难以应用在真实图像环境中.

3) 对于数据不平衡问题,缺乏有效的应对策略,导致整体标注效果受限.传统的图像标注方法采用单一的

© 中国科学院软件研究所 http://www.jos.org.cn

图像标注模型,因图像标注数据集中标签所拥有的训练样本数差异很大,导致训练好的模型预测未知图像时,对于出现频次多的标签,容易获得很高的预测值;而对于出现频次少的标签,则容易获得很低的预测值.因此,用单一的标注过程预测所有图像会导致低频图像的预测结果会被高频图像的预测结果覆盖.

对于第 1 个问题,我们提出用栈式自动编码器(stacked auto-encoder,简称 SAE)训练一个多标签分类器,训练好后即可对新图像进行快速标注,以解决传统方法需要大量计算视觉相似度或大规模矩阵运算的问题.该模型在权重更新的训练过程中可以对训练样本进行分块训练,从而提高训练效率.关于栈式自动编码器的工作过程,本文第 2 节给予详细介绍.

对于第 2 个和第 3 个问题,即训练数据极不平衡导致标注效果不理想的问题,我们从模型内和模型外两个角度进行改善:对于模型本身,我们提出一种增强训练低频样本的模型(B-SAE),并在此模型的基础上提出一种分组训练 B-SAE 模型的鲁棒平衡栈式自动编码器算法(RB-SAE),较好地改善了低频标签的标注效果并得到稳定的标注结果.对于模型本身的改进,即 B-SAE 模型和 RB-SAE 算法,本文第 3 节进行详细介绍;对于模型外,即标注策略上,我们改变传统的一个过程标注所有未知图像的思路,提出一种较好的预测中低频标签的语义传播算法(SP),标注过程中,先利用训练集标签分布信息判别待测图像的高低词频属性,再根据待测图像的不同属性选择不同的标注过程,从而改善不平衡数据集对标注过程的影响,提升整体标注效果,本文第 4 节对该标注框架,即基于属性判别的标注策略(ADA)进行详细说明.

除上述章节外,本文第 1 节介绍深度学习的背景及深度学习在图像领域的其他问题上的应用.第 5 节为实验,包括实验设置和实验结果分析.第 6 节对本文进行总结.

1 相关工作

深度神经网络包含多组非线性隐层,可以在输入和输出之间学习非常复杂的关系.然而,在输入和输出之间的非线性映射容易陷入局部最优,并且很难通过后向传播算法达到收敛状态[16].为了克服深度神经网络的学习问题,Hinton 等人[17]最早提出了基于受限玻尔兹曼机(restrict Boltzmann machine,简称 RBM)的非监督的逐层贪心训练算法,通过定义输入层和隐层的概率分布,并用多次迭代的求解方式来减小模型能量,使网络达到稳定状态,在字符识别问题上取得很好的效果.随后,深度学习模型出现了多种变形结构,比如,深度信念网络(deep belief net,简称 DBN)[18]、卷积神经网络(convolution neural network,简称 CNN)[19]和自动编码器 SAE[20]等.DBN 利用“先验互补”信息消除对信念网络推理造成困难的解释距离的影响,并使用变形的唤醒休眠算法进行慢速调优,得到一个深层的产生式模型,该模型的分类效果比同时期最好的判别模型更好.CNN 利用卷积和采样的方法训练深层模型,并用权值共享的方式大大减少了网络参数,在图像识别问题上,取得比 RBM 更好的效果.

深度学习除了在传统图像分类和识别任务上取得惊人的效果外,近几年,许多研究人员也将深度学习算法应用在图像领域的其他问题上,比如,场景识别,Zhou[21]等人用深度学习模型进行图像对象识别,再根据对象上下文来识别场景.Pinheiro[22]等人提出一种不依赖图像分割和指定特征的递归卷积神经网络的场景分割方法,对每一个像素点使用合理大小的上下文块进行输入,进而得到场景类别的归属.图像分割,Dan 等人[23]利用深度人工神经网络作为像素级分类器,对电子显微镜图像中的细胞薄膜进行分割.以每一个像素点为中心的方形领域内的像素值来预测各个像素的标签(细胞薄膜或非细胞薄膜).Luo 等人[24]使用深度信念网络训练人脸各个成分(眼睛、鼻子、嘴等)的识别器,再利用深度自动编码器将识别出的人脸成分映射成标签图,最后再用归一化分割方法完成人脸分割.研究人员还试图将深度学习理论应用于图像标注中,Ryan 等人[25]利用深度学习模型将图像的像素级原始特征进行多层深度表达,免去了工程上的特征提取及特征选择,并用 TagPro[7]算法验证了该方法的有效性.Wu 等人[26]分别将图像和网页中文本标签作为实例集,构建用于弱监督多实例学习的深度模型,在图像分类和单一对象的图像标注中取得比普通神经网络更好的效果.

根据上述研究,虽然深度学习在图像分类、图像识别、场景识别、图像分割等模式识别与计算机视觉领域得到了成功的应用,但鲜有直接将深度学习应用在图像标注问题上.文献[25]仅将深度模型用于图像标注的预处理,文献[26]仅实现单一对象的图像标注,因此,本文提出直接用深度学习解决多对象多标签的复杂图像标注

© 中国科学院软件研究所 http://www.jos.org.cn

问题.

DBN 和 CNN 这两个模型在标签较少、特征简单、特征完整的识别任务中可以取得较好的效果,而图像标注问题标签众多、图像特征复杂、每张图像包含多个类别标签、每张图像的类别标签数不等极大地影响了 DBN 和 CNN 的应用效果.而 SAE 网络,更加注重特征间的近似表达,容易调整模型将复杂的输入表达为理想的输出并应用于特定情形,比如,降噪自动编码器(DAE)[20],对 SAE 的输入添加适当的噪声,将带有噪声的特征表达为原始的完整特征,提升了 SAE 模型的泛化能力.又因为 SAE 的分类效果与 DBN、CNN 相当,因此,针对传统浅层模型存在泛化能力弱和难以收敛到最佳值等问题,本文选用 SAE 模型解决图像标注任务.

2 传统自动编码器

我们将图像标注任务当作多标签分类问题,图像特征作为模型的输入,图像标签作为监督信息.首先用自动编码器 AE 逐层预训练权值,然后将得到的权值赋给深度神经网络进行初始化,最后再整体调优完成模型训练.

2.1 问题定义

用 $X=\{x_1,x_2,...,x_N\}$, $x_i\in\mathbb{R}^d$ 表示 N 幅图像, $Y=\{y_1,y_2,...,y_M\}$ 表示 M 个关键词.图像标注任务可以表示为图像和关键词对 $P=\{(x_1,Y_1),(x_2,Y_2),...,(x_N,Y_N)\}$, $Y_i\subseteq Y$.为了讨论方便,我们将 Y_i 表示成 M 维向量 $Y_i\in\{0,1\}^M$, 在图像关键词 Y^j 中, $Y_i^j=1$ 表示第 i 幅图像 x_i 标注了关键词 y_j, 而 $Y_i^j=0$ 表示未标注关键词 y_j.

2.2 栈式自动编码器SAE

传统的神经网络用随机初始参数来调整整个网络,而深度学习分为两个阶段:逐层非监督预训练和整体调优,如图 1 所示.第 1 个阶段,图像特征 x 用非监督学习模型 AE 学习第 1 层参数 θ_1,当第 1 层 AE 训练好后, θ_1 用于产生隐层 h^1 的输出,并作为第 2 个 AE 模型的输入,像学习 θ_1 一样,逐层学到 $\theta_2,\theta_3,...,\theta_L$. 在第 2 个阶段,深度神经网络用学习到的参数 $\theta_1,\theta_2,...,\theta_L$ 进行初始化,并用反向传播算法优化整个网络.最终优化的参数可以写成 θ_1^*, $\theta_2^*,...,\theta_L^*$, 表示在预训练参数 $\theta_1,\theta_2,...,\theta_L$ 的基础上调优后的结果.一般来说,通过 AE 预训练的深度神经网络被称为栈式自动编码器(SAE).

AE 模型由两部分组成,编码器 f_θ 和解码器 $g_{\theta'}$, 如图 2 所示.编码器 f_θ 将输入图像 x 转换为隐层表达 h,解码器 $g_{\theta'}$ 将 x 重构为和 x 维度一致的向量 x', 并用损失函数 $L(x,x')$ 来优化重构误差.

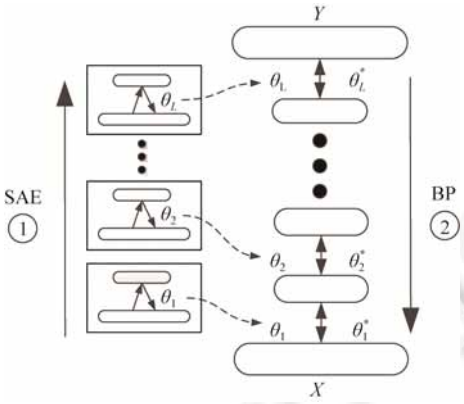


Fig.1 Work process of the SAE

图 1 SAE 工作过程

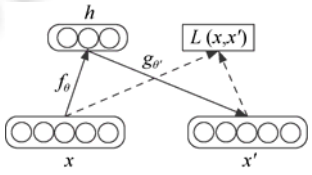


Fig.2 Work principle of an AE

图 2 AE 工作原理

f_θ 表达形式如下:

$$f_\theta(x)=\sigma(W\bullet x+b) \tag{1}$$

© 中国科学院软件研究所 http://www.jos.org.cn

其中, $\theta=\{W,b\}$, W 为网络权重,满足 $W'=W^T$, b 为偏置向量, $\sigma(x)=1/(1+e^{(-x)})$ 为激活函数.

$g_{\theta'}$ 表达形式如下:

$$g_{\theta'}(h)=\begin{cases}\sigma(W'\bullet h+b'), & x\in[0,1]\\ W'\bullet h+b', & x\in\mathbb{R}\end{cases} \tag{2}$$

其中, $\theta'=\{W',b'\}$.

AE 模型学习一个函数使输出 $x'=g_{\theta'}(f_\theta(x))$ 和 x 近似.我们定义损失函数为 $L(x,x')=(x-x')^2$, 则该模型可通过最小化损失函数进行学习.

$$\theta^*,\theta'^*=\arg\min_{\theta,\theta'}\frac{1}{N}\sum_{i=1}^{N}L(x_i,g_{\theta'}(f_\theta(x_i))) \tag{3}$$

假设用于图像标注的 SAE 模型有 L 层,用序号 $l\in\{1,...,L\}$ 表示.用 h^l 表示第 l 层的输出向量($h^0=x$ 表示输入, h^L 表示输出). W^l 和 b^l 表示第 l 层的网络权重和偏置.根据前面所述, $\{W^l,b^l\}$, $l\in\{1,...,L\}$ 使用 AE 逐层预训练.SAE 的前馈过程可表述如下:

$$h^{l+1}=\sigma(W^{l+1}h^l+b^{l+1}),\ l\in\{0,...,L-1\} \tag{4}$$

整个模型用后向传播算法调优:

$$\theta^*=\arg\min_{\theta}\sum_{i=1}^{N}L(F_\theta(x_i),Y_i) \tag{5}$$

其中, $F_\theta(x)=\sigma_{\theta_L}(...(\sigma_{\theta_1}(x)))$ 是多个 AE 模型的合成函数,而 θ_l 为参数 $\{W^l,b^l\}$, $l\in\{1,...,L\}$, 损失函数定义为 $L(x,y)=(x-y)^2$.

当模型训练好后,SAE 的最后一层 h^L 的输出即为预测图像的关键词的可能性分布 D.通过对分布 D 排名得到图像的预测关键词 Y^*.

3 基于数据均衡的自动编码器

用传统栈式自动编码器(SAE)模型处理图像标注问题是有效的,但总体效果不够理想,主要有两个原因:(1) 数据不平衡,一部分标签出现次数多,训练充分;而另一部分标签出现次数少,训练不充分,导致低频标签准确率比高频标签准确率低很多;(2) 单个 SAE 模型参数多,标注效果易随参数变化而变化,鲁棒性差,很难在实际中得到应用.为了让模型更好地训练不平衡数据,我们提出平衡栈式自动编码器(B-SAE),并在 B-SAE 基础上提出鲁棒平衡栈式自动编码器算法(RB-SAE)以提升整个模型的鲁棒性.

3.1 平衡栈式自动编码器(balanced and stacked auto-encoder,简称B-SAE)

由于低频标签 F_1 值低的一个重要原因是训练不充分,因此,我们希望 SAE 可以加重对样本的训练,以提升整个模型的泛化能力,降噪自动编码器(DAE)[20]可以在一定程度上提升模型的泛化能力,但该方法有一定局限性:(1) 以一定概率将特征向量上的某些维度的值强制置为 0,并没有真正改变特征向量,难以达到加重训练的效果;(2) 不能针对某些特定样本,比如,包含较多低频标签的样本,进行针对性的训练.为此,我们提出平衡式自动编码器(B-AE)模型,如图 3 所示, $\Phi(x)$ 表示让模型在训练过程中对训练样本进行判断,若样本 x 包含低频标签的个数多于 k 个,则对该样本添加适当的噪声. $\Gamma(x)$ 表示对样本 x 的训练强度,若该样本所包含标签的出现次数低于一定阈值,则增加它的训练次数.

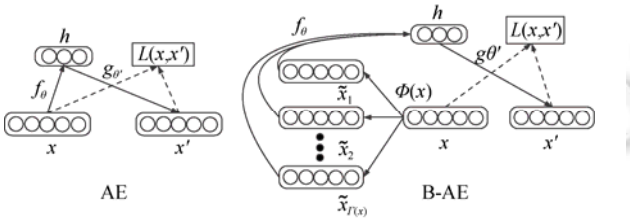


Fig.3 Comparison between the traditional AE and the B-AE

图 3 传统 AE 和 B-AE 对比

这样,等式(3)可以调整为如下形式:

© 中国科学院软件研究所 http://www.jos.org.cn

$$\theta^*,\theta'^* = \arg\min_{\theta,\theta'} \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{1}{\Gamma(x_i)}\sum_{j=1}^{\Gamma(x_i)} L(\Phi(x_i), g_{\theta'}(f_\theta(\Phi(x_i))))\right\} \tag{6}$$

接下来,我们说明 $\Gamma(x_i)$ 和 $\Phi(x_i)$ 是如何定义的.

令向量 $C=(c_1,c_2,...,c_M)$, $c_i\in\mathbb{Z}^+$ 表示关键词 y_i 在训练集 P 中出现的次数, $\Pi=\frac{1}{M}\sum_{j=1}^{M}c_j$ 表示关键词的平均出现次数.这样,我们可以得到一个向量,表示第 i 幅图像 x_i 的每个关键词 $Y_i^j, j\in\{1,2,...,M\}$ 在训练集中出现的次数 $Y_{C,i}=C*Y_i$ (*表示两个向量对应点相乘得到一个新向量).从而得到在图像 x_i 中出现次数最低的关键词为 $\Lambda_{x_i}=\min_j(Y_{C,i}^j)$, 于是得到式子 $\Gamma(x_i)$ 和 $\Phi(x_i)$ 分别为

$$\Gamma(x_i)=\begin{cases}\alpha\bullet\dfrac{\Pi}{\Lambda_{x_i}}=\alpha\bullet\dfrac{\frac{1}{M}\sum_{j=1}^{M}c_j}{\min_j(Y_{C,i}^j)}, & \Lambda_{x_i}<=\beta\bullet\Pi\\ 1, & \text{Others}\end{cases} \tag{7}$$

其中,α和β为常系数,β用于确定哪些样本需要加重训练,α用于控制需要加重训练的样本的训练强度.

$$\Phi(x_i)=\begin{cases}\chi\bullet\left(\dfrac{1}{d}\sum_{j=1}^{d}x_i^j\right)\bullet Ran(\cdot), & \Lambda_{x_i}<=\beta\bullet\Pi\\ x_i, & \text{Others}\end{cases} \tag{8}$$

其中,β和式(7)中的β是一致的,χ为常系数,用于控制噪声添加的强度, d 为图像 x_i 特征的维度, x_i^j 表示图像 x_i 第 j 个维度的值, $Ran(\cdot)$ 为随机向量函数,该向量的维度和 x_i 一致,每个维度的值是随机的,并且随机值的取值分布服从 $N(0,1)$ 分布或服从 $U[0,1]$ 分布.

与传统 SAE 模型类似,得到预训练权值后用后向传播算法进行调优,但调优过程会对特定样本(含低频标签的个数多于 k 个)加重训练,整个模型的优化等式(5)变为

$$\theta^*=\arg\min_\theta\sum_{i=1}^{N}\sum_{j=1}^{\Gamma(x_i)}L(F_\theta(\Phi(x_i)),Y_i) \tag{9}$$

根据上述描述的方法可知,对于任意一幅图像 x_i, 若 x_i 包含多于阈值个数的低频标签,则平衡栈式自动编码器(B-SAE)在迭代训练时,会给 x_i 添加噪声加重训练,隐式地扩充了低频标签数据集,从而提升低频词的 F_1 值.

3.2 鲁棒平衡栈式自动编码器算法(robust balanced and stacked auto-encoder,简称RB-SAE)

通过实验,我们验证了改进后的平衡栈式自动编码器(B-SAE)较好地提高了低频词的 F_1 值,但整个 B-SAE 模型较为复杂,需要调整的参数很多(比如,隐层神经元个数的确定、随机函数 $Ran(\cdot)$ 的选择、迭代次数的控制等),每个参数的设置都会对结果产生很大影响,坏的情况甚至让整个模型失效.为了提高模型的鲁棒性,我们提出了一种适合图像标注的鲁棒平衡栈式自动编码器算法(RB-SAE),算法的主要过程如图 4 所示,第 1 步,输入训练图像特征,并按序分组训练子 B-SAE 模型;第 2 步,对每一组内的各个 B-SAE 计算分类误差率,得到各组分类误差率最小的 B-SAE^t 模型;第 3 步,根据分类误差率对各组间的 B-SAE^t 计算权重,并组合成一个带权 B-SAE 模型;第 4 步,将待测图像输入组合后的带权 B-SAE 模型;第 5 步,输出预测分布 D,并对分布 D 排名得到预测结果.

算法按不同的加噪方式划分不同的组,每组内根据不同的隐层神经元个数划分子模型 B-SAE$_k^t$, t 表示模型 B-SAE 采用第 t 种加噪方式,k 表示第 k 个子 B-SAE 模型设置的隐层神经元个数.为了计算组内 B-SAE$_k^t$ 的分类误差率以及组间模型 B-SAE^t 的权重,我们需对训练数据设置权值,初始权值设置如下:

$$W=(w_{11},...,w_{1i},...,w_{1N}), w_{1i}=\frac{1}{N}, i=1,2,...,N \tag{10}$$

这样, B-SAE$_k^t$ 的分类误差率可按下式计算:

© 中国科学院软件研究所 http://www.jos.org.cn

$$e_k^t = \sum_{i=1}^{N} w_{ti} \bullet Sgn(\text{B-SAE}_k^t(x_i) \neq Y_i) \tag{11}$$

其中, $Sgn(x) = \begin{cases} 1, & x = \text{true} \\ 0, & x = \text{false} \end{cases}$, $\text{B-SAE}_k^t(x_i) \neq Y_i$ 表示的意思为,假设图像 x_i 的真实标签集 Y_i 包含 c 个关键词,并通过模型 B-SAE_k^t 预测得到标签集 Y_i^* 的个数也为 c 个,如果 $Y_i = Y_i^*$, 则 $\text{B-SAE}_k^t(x_i) \neq Y_i$ 为 false,否则为 true.

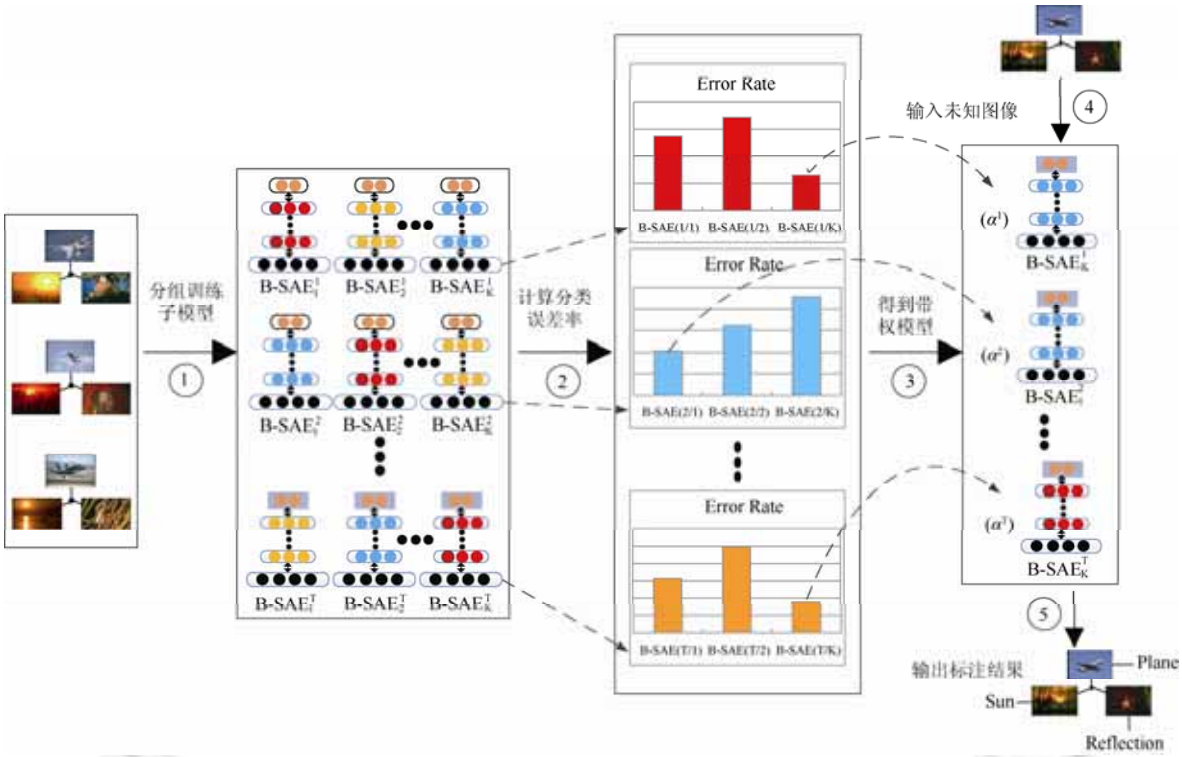


Fig.4 Workflow of the RB-SAE algorithm

图 4 RB-SAE 算法工作流程

根据组内所有子 B-SAE_k^t 模型的分类误差率,我们可以得到该组分类误差率最低的模型 B-SAE^t 和对应的分类误差率 e^t,B-SAE^t 的权重可按如下方式计算:

$$\alpha^t = \frac{1}{2}\log\frac{1-e^t}{e^t} \tag{12}$$

当第 t 组的模型训练完后,我们需要更新训练数据的权值,以便更好地获得下一组模型的权重,更新训练数据权值的方式如下:

$$W_{t+1} = \{w_{t+1,1},...,w_{t+1,i},...,w_{t+1,N}\},\ w_{t+1,i} = \frac{w_{ti} \bullet e^{(-\alpha^t \bullet Y_i \bullet \text{B-SAE}^t(x_i))}}{\sum_{i=1}^{N} w_{ti} \bullet e^{(-\alpha^t \bullet Y_i \bullet \text{B-SAE}^t(x_i))}}, i = 1,2,...,N \tag{13}$$

当所有组都训练完毕后,我们就可以得到如下关键词预测分布:

$$D = \sum_{t=1}^{T} \alpha^t \bullet \text{B-SAE}^t(x) \tag{14}$$

根据算法的执行过程,我们可以知道,分组训练模型后,算法会选取组内效果最好的子模型,并对选取出来的组间子模型计算权重,保证好的模型可以获得较大的权重,较差的模型获得较小的权重,从而提升整个模型的鲁棒性.

© 中国科学院软件研究所 http://www.jos.org.cn

4 基于数据均衡的标注策略

虽然我们提出了在模型内加重训练低频标签的平衡栈式自动编码器模型(B-SAE),并在B-SAE的基础上提出了鲁棒平衡栈式自动编码器算法(RB-SAE),在一定程度上提升了低频标签的F_1值,但仍然难以改变中低频标签训练不足的本质,中低频标签的 F_1 值依旧较低,从而影响了整个模型的标注效果.为此,我们提出一种基于数据均衡的标注策略,首先为测试图像构造一个局部均衡数据集,然后再用此数据集判别该图像的属性(高频图像或低频图像).若为高频图像,则执行 RB-SAE 算法进行预测;若为低频图像,则利用局部均衡数据集以语义传播的方式进行预测,从而提升整个模型的标注效果.

4.1 局部数据均衡与语义传播(semantic propagation,简称SP)

传统的标签传播算法[6,7]将整个训练集作为训练的数据集,对于每一幅测试图像 I,从训练集中获取相似度最高的 k 幅图,k 幅图中出现次数多的标签最可能传递给测试图像 I.由于图像标注数据集的标签分布极不均衡,这类方法存在如下问题:(1) 若从训练集中获取数量较少的 k 幅图,则较难覆盖到所有标签(图像标注问题的标签规模较大),导致部分正确的标签不在 k 幅图内,因此,这部分标签肯定无法被准确预测到,从而影响到整体标注效果;(2) 若从训练集中获取数量较大的k幅图,因为训练集标签不均,k幅图中的标签肯定也不均,导致数据集中的高频标签在k幅图中也为高频标签,使得这部分标签更容易传递给测试图像,而另一部分标签(中低频)的标注结果容易被高频标签覆盖.

为了解决上述两个问题,我们构建一种理想的局部均衡数据集,要求该数据集覆盖所有标签且每个标签的出现频次一致.为此,我们划分每一个标签所包含的所有图像为一个语义组(不同语义组间允许有相同的图像),对于每一幅测试图像 I,从每一个语义组中选取 n 幅和图像 I 视觉相似度最高的图像构造子训练集,并且把每一幅选出来的图像看成一个图像对象,仅表示一个语义概念,即其所在语义组的语义概念.因此,图像 I 的子训练集包含所有关键词,并且每个关键词的出现频次一致,得到了局部均衡数据集.基于此数据集,我们提出一种语义传播算法,具体实现描述如下:

令 $G=\{(y_1,X_1),(y_2,X_2),...,(y_M,X_M)\}, X_i \subseteq X$ 表示每一个关键词和该关键词所包含的所有图像, X_i 和 X_j 中的图像可重复, $i,j \in \{1,...,M\}, i \neq j$.我们用条件概率 $P(x \mid y_i)$ 建立给定关键词 $y_i \in Y$ 的图像 x 的特征分布.这样,我们将图像标注转化为求解后验概率的问题:

$$P(y_i \mid x)=\frac{P(x \mid y_i)P(y_i)}{P(x)} \tag{15}$$

其中, $P(y_i)$ 和 $P(x)$ 为先验概率,取固定值.

因此,对于测试图像 I,它的最佳标签可以这样得到:

$$y^*=\arg\max_i P(y_i \mid I) \tag{16}$$

令 $G_i=(y_i,X_i)$, 对于测试图像 I,我们从 G_i 中的 X_i 中挑选 n 幅和 I 视觉距离最近的图像构成子集 $G_{I,i} \subseteq G_i$. 每一个集合 $G_{I,i}$ 是图像 I 对应于标签 y_i 的语义组.一旦 $G_{I,i}$ 确定后,我们将它们合并为一个集合 $G_I=\{G_{I,1} \cup ... \cup G_{I,M}\}$, 各 $G_{I,i}$ 间的关键词不同,但可以有相同的图像.通过这一方法,我们得到一个针对图像 I 的局部均衡数据集 $G_I \subseteq G$.这样就容易知道,在集合 G_I 中的每一个关键词都将出现 n 次($G_{I,i}$ 中的图像数少于 n 的情况除外).我们定义给定标签 $y_k \in Y$ 的图像 I 的后验概率为

$$P(I \mid y_k)=\sum_{(y_i,X_i) \in G_I} \theta_{I,x_i} \cdot P(y_k \mid x_i) \tag{17}$$

其中, $P(y_k \mid x_i) \in \{0,1\}$ 表示当图像 x_i 所在语义组 $G_{I,i}$ 的语义概念等于标签 y_k 时为 1,否则为 0; θ_{I,x_i} 表示图像 x_i 的权重,按如下方式计算:

$$\theta_{I,x_i}=\frac{e^{(-Dis(I,x_i))}-e^{(Dis(I,x_i))}}{e^{(-Dis(I,x_i))}+e^{(Dis(I,x_i))}}+1.0 \tag{18}$$

其中, $Dis(I,x_i)$ 表示图像 I 和 x_i 之间的距离,距离的计算采用文献[7]中的方法.

© 中国科学院软件研究所 http://www.jos.org.cn

由公式(17)可知,若局部均衡数据集中的图像 x_i 和图像 I 的相似度越高,则图像 x_i 的语义传递给图像 I 的可能性就越大,将公式(17)代入公式(15)和公式(16)即可得到未知图像 I 的预测标签.

4.2 属性判别的标注策略(attribute discrimination annotation,简称ADA)

上一节提出的语义传播算法通过构造局部均衡数据集改善了传统标签传播算法难以覆盖所有标签以及中低频标签难以被预测的问题,提升了中低频标签的预测可能性.然而,由于训练集中中低频标签的训练样本数极其有限,导致我们提出的鲁棒平衡栈式自动编码器算法(RB-SAE)对中低频标签的预测能力仍不够理想,因此,利用语义传播算法能够较好预测中低频标签的准确率这一优势,在局部均衡数据集的基础上,我们提出一种判别测试图像属性的图像标注方法,以提升整个模型的标注效果.

基于属性判别的标注过程如图 5 所示.第 1 步,根据每一个关键词对训练集划分语义组,每一个关键词包含的所有图像构成一个语义组;第 2 步和第 3 步,构造测试图像的局部均衡数据集,即,根据输入的测试图像从每一个语义组中获取指定数量且视觉上距离最近的图像构成局部均衡数据集,每一幅图像仅表示一个语义概念;第 4 步,训练鲁棒平衡栈式自动编码器模型(RB-SAE),首先训练平衡栈式自动编码器模型(B-SAE),然后再在 B-SAE 基础上加入 RB-SAE 算法来完成模型的训练;第 5 步和第 6 步,分别从语义组和局部均衡数据集获得全局词频信息和局部词频信息(根据测试图像从局部均衡数据集中选取指定数量且视觉相似度最高的图像,再从中统计词频信息),用于判别测试图像的属性;第 7 步,对全局高频词信息和局部高频词信息取交集,交集个数大于阈值,则判定测试图像的属性为高频,反之为中低频;若测试图像的判别结果为高频,则进入第 8 步,将测试图像输入 RB-SAE 模型,得到输出结果;若测试图像的判别结果为中低频,则进入第 10 步,通过语义传播算法从局部均衡数据集中获取标签.

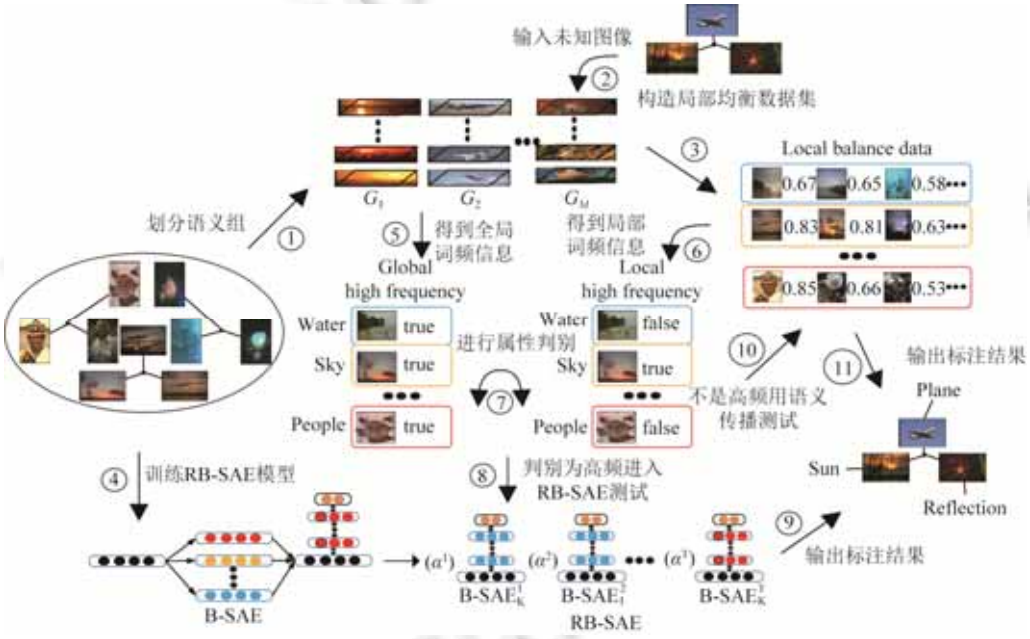


Fig.5 Workflow of the image annotation based on attribute discrimination (ADA)

图 5 基于属性判别的图像标注(ADA)流程

基于属性判别标注策略的理论,详细描述如下.

对于一幅测试图像 I,构造标签出现次数均匀的局部均衡数据集 G_I,再从 G_I 取 m 幅和 I 距离最近的图像 $\tilde{G}_I=\{(y_1,\tilde{X}_1),(y_2,\tilde{X}_2),...,(y_M,\tilde{X}_M)\}$,$\tilde{X}_i\subseteq X_i$,需要注意的是,$\tilde{X}_i$ 可能为 $\varnothing$(因为从 G_I 中取出的 m 幅图像可能都不包含在 $\tilde{X}_i$ 中),且 $\tilde{X}=\{\tilde{X}_1\cup...\cup\tilde{X}_M\}$ 包含的图像幅数为 m.又因为从集合 C 中,我们可以知道哪些关键词属于高

© 中国科学院软件研究所 http://www.jos.org.cn

频关键词,哪些属于低频关键词.因此,结合集合 C,我们可以从 $\tilde{G}_I$ 中知道,在与图像 I 距离最近的 m 幅图像中,属于高词频图像和中词频图像的比例分别是多少.为了讨论方便,我们将 $\tilde{X}_i$ 表示成 N 维向量 $\tilde{X}_i \in \{0,1\}^N$, 当 $\tilde{X}_i^j=1$ 时,表示 $\tilde{X}_i$ 包含图像 x_j, 当 $\tilde{X}_i^j=0$ 时,表示 $\tilde{X}_i$ 不包含图像 x_j. 这样,我们可以通过下式求得图像的预测分布 D.

$$D=\begin{cases}\sum_{t=1}^{T}\alpha^t\text{•B-SAE}^t(x), & \left(\sum_{(y_i,\tilde{X}_i)\in\tilde{G}_I}\phi(y_i)\text{•}\varphi(\tilde{X}_i)\right)\geq\varepsilon\\ P(y_k\mid x), & \text{Others}\end{cases} \tag{19}$$

其中, $\sum_{t=1}^{T}\alpha^t\text{•B-SAE}^t(x)$ 对应公式(14), $P(y_k\mid x)\propto P(x\mid y_k)=\sum_{(y_i,X_i)\in G_I}\theta_{I,x_i}\bullet P(y_k\mid x_i)$ 对应公式(15),ε为常系数,用于控制图像 I 的属性的判别, $\phi(y_i)$ 用于判定在集合 P 中哪些关键词属于高频关键词,哪些不属于,定义如下:

$$\phi(y_i)=\begin{cases}1, & c_i\geq\eta\text{•}\dfrac{1}{M}\sum_{j=1}^{M}c_j\\ 0, & \text{Others}\end{cases} \tag{20}$$

η为常系数,当 $\phi(y_i)=1$时,表示 y_i 在 P 中为高频关键词,当 $\phi(y_i)=0$ 时,则相反.

$\varphi(\tilde{X}_i)$ 用于判定在集合 $\tilde{G}_I$ 中哪些关键词属于高频关键词,哪些不属于,定义如下:

$$\varphi(\tilde{X}_i)=\begin{cases}1, & \sum_{j=1}^{N}\tilde{X}_i^j\geq\tau\text{•}\left(\dfrac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{N}\tilde{X}_i^j\right)\\ 0, & \text{Others}\end{cases} \tag{21}$$

τ为常系数,当 $\varphi(\tilde{X}_i)=1$时,表示 y_i 在 $\tilde{G}_I$ 中为高频关键词,当 $\varphi(\tilde{X}_i)=0$ 时,则相反.

由此可知,若在与图像 I 距离最近的 m 个邻居中,含高频词的邻居超过了一定的比例,则图像 I 的预测结果由 $\sum_{t=1}^{T}\alpha^t\text{•B-SAE}^t(x)$ 决定,反之,图像 I 的预测结果由 $\sum_{(y_i,X_i)\in G_I}\theta_{I,x_i}\bullet P(y_k\mid x_i)$ 决定.

5 实 验

这一部分,我们验证基于数据均衡的自动编码器(包含 B-SAE 和 RB-SAE 算法)和基于数据均衡的标注策略中的属性判别标注模型(ADA)在 3 个数据集上的效果.首先,详细介绍我们所使用的数据集、图像特征和评价标准.其次,展示每一种方法的实验结果并作简要分析.最后,将我们的模型和目前效果较好的几个图像标注模型进行比较,并展示了一些图像的实际标注例子.

5.1 实验设置

5.1.1 数据集

我们使用图像标注领域广泛使用的 3 个公共数据集:Corel5k、Espgame 和 Iaprtc12.Corel5k 是图像标注最常用的数据集,已成为图像标注的基准,包含 5 000 幅自然风光和日常生活照片.Espgame 数据集由 20 770 幅类别广泛的图像构成,比如,商标、绘图和个人照片,这些图像从在线交互式游戏的任务中获取.Iaprtc12 数据集由 19 627 幅图像构成,涵盖了运动、行为、人、动物、城市、风景和当代人生活的其他方面,很多关键词是从图像中的标语或字幕中获取的.详细的数据信息见表 1.

Table 1 Information on the three public datasets

表 1 3 个公共数据集的信息

数据集	图像数	标签数	训练图像数	测试图像数	每幅图像平均标签数	每个标签平均出现次数
Corel5k	5 000	260	4 500	500	3.4	58.6
Espgame	20 770	268	18 689	2 081	4.7	326.7
Iaprtc12	19 627	291	17 665	1 962	5.7	347.7

© 中国科学院软件研究所 http://www.jos.org.cn

5.1.2 图像特征

我们使用与文献[7]中相似的图像特征,该特征是由局部特征和全局特征组合而成的.局部特征包含 Sift 和 Hue 描述符,并通过多尺度网格和 Harris 拉普拉斯兴趣点两种方式获取.全局特征包含 RGB、HSV 和 LAB 这 3 个颜色空间上的直方图,以及 Gist 特征.为了对图像的空间信息进行编码,还对每一幅图像水平 3 等分后提取除了 Gist 以外的所有特征.

对于栈式自动编码器(SAE)模型(包含普通 SAE、平衡栈式自动编码器模型(B-SAE)和鲁棒平衡栈式编码器算法(RB-SAE)),我们使用 DenseHue、DenseHueV3H1、DenseSift、Gist、HarrisHue、HarrisHueV3H1 和 HarrisSift 这 7 种特征,共计 3 312 维.对于构建局部均衡数据集和实现语义传播算法,我们使用全部 15 种特征,图像间视觉相似度的计算也采用文献[7]中的方法,即, L_1 距离用于计算颜色直方图, L_2 距离用于计算 Gist 特征, χ^2 距离用于计算 Sift 和 Hue 特征.

5.1.3 评价指标

我们采用与文献[7]相同的评价标准.首先,对所有测试图像标注 5 个相关关键词.其次,为每一个关键词 y_j 计算准确率 P^j 和召回率 R^j, $j\in\{1,...,M\}$, 具体计算公式为

$$P^j=\frac{Precision(y_j)}{Prediction(y_j)},\ R^j=\frac{Precision(y_j)}{Ground(y_j)} \tag{22}$$

其中, $Precision(y_j)$ 为关键词 y_j 的准确预测次数, $Prediction(y_j)$ 为关键词 y_j 的预测次数, $Ground(y_j)$ 为关键词 y_j 的真实标注次数.最后,对所有关键词的准确率和召回率求平均值,分别记为 P 和 R.

$$P=\frac{1}{M}\sum_{j=1}^{M}P^j,\ R=\frac{1}{M}\sum_{j=1}^{M}R^j \tag{23}$$

我们还对 P 和 R 计算 F_1 值,用以反映 P 和 R 的平衡性(P 过小或 R 过小都会使 F_1 变小),并记录至少准确预测过 1 次的关键词个数 N^+,用以反映整个数据集中有多少关键词被准确预测过,有多少未被准确预测过,F_1 和 N^+分别按如下方式计算:

$$F_1=\frac{2\bullet P\bullet R}{P+R},\ N^+=\sum_{j=1}^{M}Sgn(R^j) \tag{24}$$

其中, $Sgn(x)=\begin{cases}1, & x>0\\ 0, & x=0\end{cases}$.

5.2 实验结果

对于栈式自动编码器(SAE)模型(包含普通 SAE、平衡栈式自动编码器(B-SAE)和鲁棒平衡栈式自动编码器算法(RB-SAE)),我们都设计 4 层的网络结构(这里的 4 层是指包含特征层和标签层的层数,原因在第 5.2.1 节中说明),层与层之间的网络权值,用各自对应的编码器进行预训练,最后再整体调优.各模型中的参数较多,我们设置不同的参数进行实验,取实验效果最好时所设定的参数作为以下实验结果的参数.

5.2.1 SAE 层数的选择

图像标注的 3 个公共数据集相比大数据集而言属于小规模数据集,Corel5k 数据集的训练图像数仅为 4 500 张,而 Espgame 和 Iaprtc12 两个数据集的训练图像数也才 18 000 张左右,因此,为了防止过拟合,我们对 SAE 模型设计相对少的层数进行实验以指导本文提出的其他类 SAE 模型(包含 B-SAE 和 RB-SAE)的结构设计.从表 2 可以看出,设计 4 层的 SAE 网络结构效果最好,传统的 3 层神经网络因模型表达能力较弱导致标注效果受限.随着层数的增加,模型复杂度也随之增加,出现了一定程度的过拟合现象,标注效果反而减弱了,小数据集 Corel5k 上的准确率下降得较为明显.因 4 层 SAE 网络结构较适合处理图像标注问题,所以本文提出的 B-SAE 和 RB-SAE 模型都采用 4 层网络结构.

© 中国科学院软件研究所 http://www.jos.org.cn

Table 2 Experimental results obtained from different number of the SAE layers

表 2 不同 SAE 层数得到的实验结果

层数	Corel5k				Espgame				Iaprtc12			
	P	*R*	F_1	*N*+	*P*	*R*	F_1	*N*+	*P*	*R*	F_1	*N*+
3	0.17	0.22	0.19	104	0.22	0.18	0.20	193	0.30	0.19	0.23	203
4	**0.20**	**0.23**	**0.21**	**110**	**0.23**	**0.18**	**0.20**	**201**	**0.30**	**0.20**	**0.24**	**208**
5	0.14	0.19	0.16	102	0.18	0.16	0.17	188	0.27	0.19	0.22	198
6	0.14	0.18	0.16	93	0.18	0.16	0.17	180	0.25	0.18	0.21	196

5.2.2 B-SAE 的有效性

B-SAE 是针对数据不平衡问题以模型本身作为出发点而提出的.该模型使传统 SAE 模型拥有训练不平衡数据的能力,从而提升了低频标签的 F_1 值.需要设定 4 个参数α、β、χ和 $Ran(\cdot)$, α和β取固定值,分别为 1 和 0.5,我们在 3 个数据集上设置不同的χ(噪声强度)和 $Ran(\cdot)$ (随机向量函数,其中,Gaussian 表示随机向量每一维的取值服从 $N(0,1)$ 分布,Uniform 表示取值服从 $U[0,1]$ 分布)进行测试,结果见表 3.

Table 3 Results of the B-SAE model with different parameters

表 3 不同参数的 B-SAE 模型的测试结果

数据集	常系数	随机函数	*P*	*R*	F_1	*N*+
Corel5k	**χ=1.2**	**Gaussian**	**0.25**	**0.31**	**0.28**	**128**
	χ=1.5	Gaussian	0.24	0.31	0.27	126
	χ=1.2	Uniform	0.21	0.28	0.24	121
	χ=1.5	Uniform	0.23	0.30	0.26	125
Espgame	**χ=1.2**	**Gaussian**	**0.25**	**0.21**	**0.23**	**221**
	χ=1.5	Gaussian	0.25	0.20	0.22	222
	χ=1.2	Uniform	0.25	0.21	0.23	220
	χ=1.5	Uniform	0.25	0.21	0.23	215
Iaprtc12	χ=1.2	Gaussian	0.31	0.21	0.25	215
	χ=1.5	Gaussian	0.30	0.21	0.25	218
	χ=1.2	**Uniform**	**0.31**	**0.22**	**0.26**	**223**
	χ=1.5	Uniform	0.30	0.21	0.25	222

根据表 3,我们取各数据集得到的最好结果(加粗部分)作为以下分析的数据.

我们将 3 个数据集划分高频标签和低频标签进行统计(高频标签是指标签的出现频次高于所有标签的平均出现频次,低频标签则相反),并计算高、低频标签的 F_1 值和 N^+.从表 4 中我们可以知道,3 个数据集的低频标签数占总标签数的比例将近 75%,而高频仅占 25%.相比普通 SAE,B-SAE 在 3 个数据集上低频标签的平均 F_1 值分别提升了 50.4%、21.2%和 15.5%,低频标签至少准确预测 1 次的关键词个数 N^+分别增加了 16、20 和 15 个.对于训练数据较少的 Corel5k 数据集,高频标签的平均 F_1 值和 N^+也得到改善.

Table 4 Comparison of predicting high and low frequency labels between SAE and B-SAE

表 4 SAE 和 B-SAE 预测高低频标签的对比

数据集	高频标签数	低频标签数	模型	高频标签平均 F_1	低频标签平均 F_1	高频标签 *N*+	低频标签 *N*+
Corel5k	65	195	SAE	0.385	0.123	62	48
			B-SAE	0.464	0.185	64	64
Espgame	67	201	SAE	0.287	0.151	67	134
			B-SAE	0.315	0.183	67	154
Iaprtc12	74	217	SAE	0.354	0.181	74	134
			B-SAE	0.341	0.209	74	149

表 4 仅从整体上统计了 SAE 和 B-SAE 的效果,我们又从更细的角度进行统计,即对每个数据集上的低频关键词统计 F_1 值,具体情况如图 6~图 8 所示.

在图 6~图 8 中,我们分别从 3 个数据集上抽出 20 个低频关键词(均低于平均关键词出现频次),粉色条状图对应左纵轴,代表标签的出现频次,蓝色和紫色折线图对应右纵轴,分别代表模型 SAE 的 F_1 值和模型 B-SAE 的 F_1 值.3 个数据集上的数据表明,B-SAE 提升了低频标签的 F_1 值,对于一些在 SAE 模型上未准确预测的词,在 B-SAE 模型上也得到了预测.

© 中国科学院软件研究所 http://www.jos.org.cn

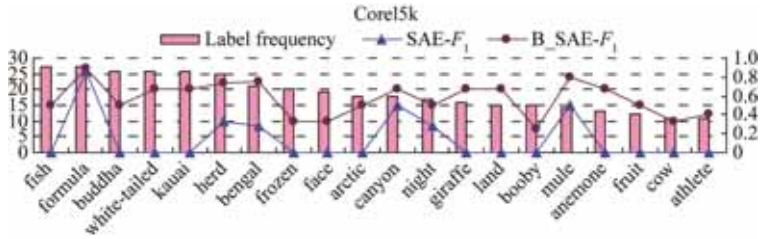


Fig.6 F_1 value of the low frequency labels of SAE and B-SAE on Corel5k

图 6 在 Corel5k 上测试 SAE 和 B-SAE 预测低频标签的 F_1 值

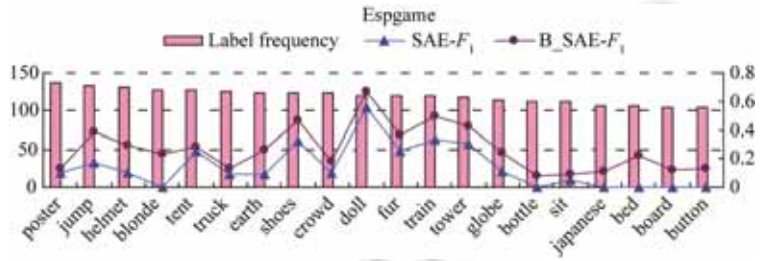


Fig.7 F_1 value of the low frequency labels of SAE and B-SAE on Espgame

图 7 在 Espgame 上测试 SAE 和 B-SAE 预测低频标签的 F_1 值

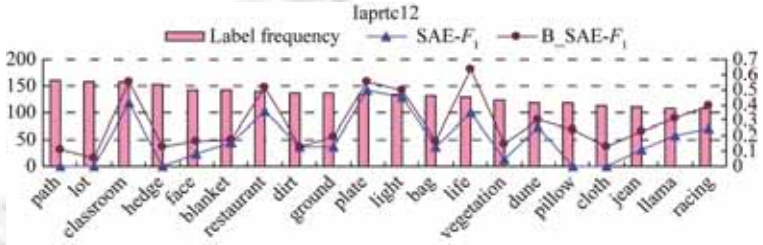


Fig.8 F_1 value of the low frequency labels of SAE and B-SAE on Iaprtc12

图 8 在 Iaprtc12 上测试 SAE 和 B-SAE 预测低频标签的 F_1 值

5.2.3 RB-SAE 算法的有效性

RB-SAE 算法是在 B-SAE 模型基础上,针对单个 B-SAE 标注结果不稳定、容易随参数变动的问题,而提出的改善算法,从而提升 B-SAE 的鲁棒性.为了说明单个 B-SAE 模型存在不稳定现象并验证该模型的效果,我们分 4 组进行测试, $Ran(\cdot)$ 产生的随机向量的分量取值服从 $N(0,1)$ 分布和 $U[0,1]$ 分布.每种加噪方式又设置两个 χ 参数,1.2 和 1.5.组内又分 3 个子 B-SAE 模型,依次设置隐层神经元个数为 450、500 和 550.总共 12 个子模型,参数设定见表 5.

Table 5 Parameter settings for testing the RB-SAE algorithm

表 5 测试 RB-SAE 算法的参数设定

子模型	常系数	随机函数	隐层神经元个数
B-SAE(1/1)	χ=1.2	Gaussian	450
B-SAE(1/2)	χ=1.2	Gaussian	500
B-SAE(1/3)	χ=1.2	Gaussian	550
B-SAE(2/1)	χ=1.5	Gaussian	450
B-SAE(2/2)	χ=1.5	Gaussian	500
B-SAE(2/3)	χ=1.5	Gaussian	550
B-SAE(3/1)	χ=1.2	Uniform	450
B-SAE(3/2)	χ=1.2	Uniform	500
B-SAE(3/3)	χ=1.2	Uniform	550
B-SAE(4/1)	χ=1.5	Uniform	450
B-SAE(4/2)	χ=1.5	Uniform	500
B-SAE(4/3)	χ=1.5	Uniform	550

© 中国科学院软件研究所 http://www.jos.org.cn

使用上述 12 个 B-SAE 子模型的参数设定,分别在 3 个数据集上的执行过程如图 9~图 11 所示.

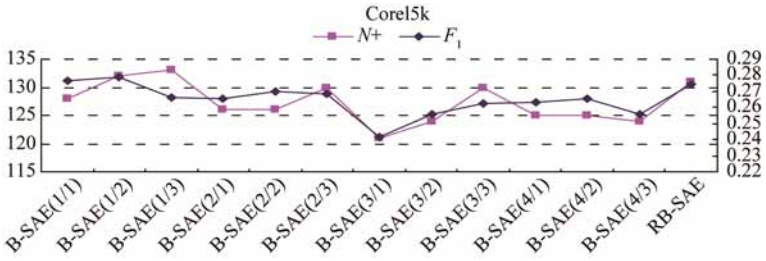


Fig.9 Results of the RB-SAE algorithm on Corel5k

图 9 在 Corel5k 上测试 RB-SAE 算法

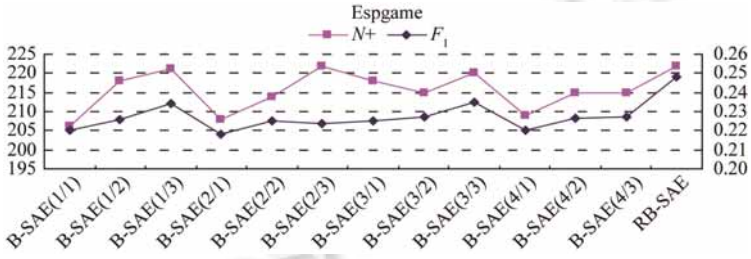


Fig.10 Results of the RB-SAE algorithm on Espgame

图 10 在 Espgame 上测试 RB-SAE 算法

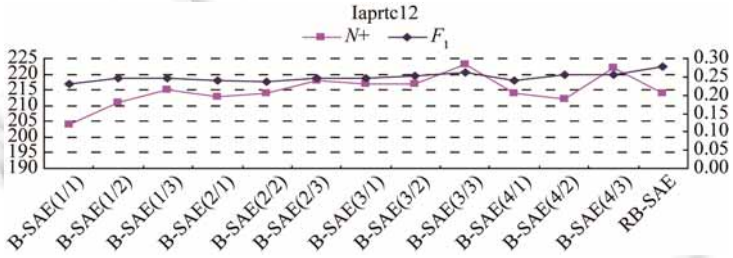


Fig.11 Results of the RB-SAE algorithm on Iaprtc12

图 11 在 Iaprtc12 上测试 RB-SAE 算法

在图 9~图 11 中,粉色折线图对应左纵轴,代表子模型至少准确预测 1 次的关键词个数 N^+,紫色折线图对应右纵轴,代表子模型的 F_1 值,每个图的最后一列为执行 RB-SAE 算法后得到的结果.从图中我们可以看出,参数设定不同,单个 B-SAE 模型得到的数据结果不稳定,要么 N^+ 高一点,F_1 值低一点(如图 10 中的 B-SAE(2/3),N^+ 较高但 F_1 值较低);要么 F_1 高一点,N^+ 低一点(如图 11 中的 B-SAE(1/1),F_1 值较高但 N^+ 较低),通过执行 RB-SAE 算法后,N^+ 和 F_1 值都可以取到相对较为稳定的结果.

为了方便对比 SAE 系列模型的整体效果,我们把 3 个模型的结果记录在表 6 中.SAE 为传统栈式自动编码器.B-SAE 为本文提出的平衡栈式自动编码器,这里数据对应于表 3 中加粗的部分.RB-SAE 为本文提出的鲁棒平衡栈式自动编码器算法,所给出的数据对应于图 9~图 11 的执行结果,即每个图的最后一列.

Table 6 Comparison of three auto-encoder models

表 6 3 个自动编码器模型的测试对比

模型	Corel5k				Espgame				Iaprtc12			
	P	*R*	F_1	*N*+	*P*	*R*	F_1	*N*+	*P*	*R*	F_1	*N*+
SAE	0.20	0.23	0.21	110	0.23	0.18	0.20	201	0.30	0.20	0.24	208
B-SAE	0.25	0.31	0.28	128	0.25	0.21	0.23	221	0.31	0.22	0.26	223
RB-SAE	0.24	0.32	0.27	131	0.27	0.23	0.25	222	0.35	0.23	0.28	214

© 中国科学院软件研究所 http://www.jos.org.cn

5.2.4 ADA 的实验效果

ADA 是针对数据集不平衡问题,以模型外,即标注过程,作为出发点提出来的一种根据测试图像的高低频属性来选择标注过程的标注策略.为了说明该标注框架的合理性,我们对仅用鲁棒平衡栈式自动编码器算法(RB-SAE)和仅用语义传播算法(SP)的结果进行对比,发现大多数高频标签用 RB-SAE 得到的 F_1 值比用 SP 算法得到的 F_1 值要高,而中低频标签用 SP 算法得到的 F_1 值比用 RB-SAE 得到的 F_1 值要高,我们分别在 3 个数据集上取 20 个关键词进行统计,其中 10 个为高频词,另外 10 个为低频词,具体情况如图 12~图 14 所示.

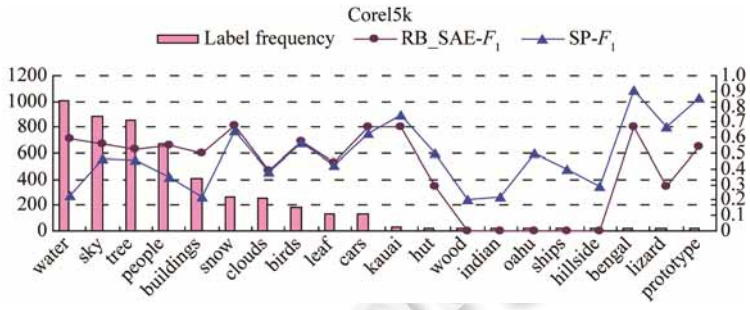


Fig.12 F_1 value of the high and low frequency labels on Corel5k computed by RB-SAE and SP

图 12 在 Corel5k 上测试 RB-SAE 和 SP 预测高低词频标签 F_1 值

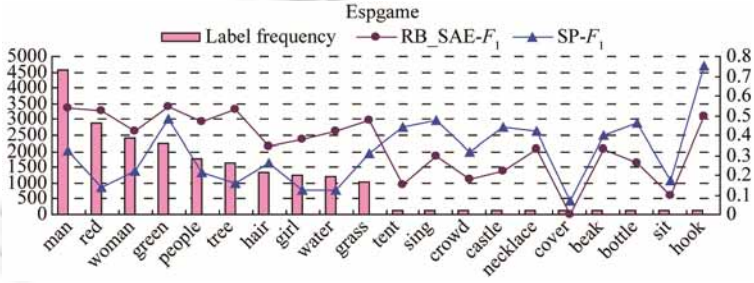


Fig.13 F_1 value of the high and low frequency labels on Espgame computed by RB-SAE and SP

图 13 在 Espgame 上测试 RB-SAE 和 SP 预测高低词频标签 F_1 值

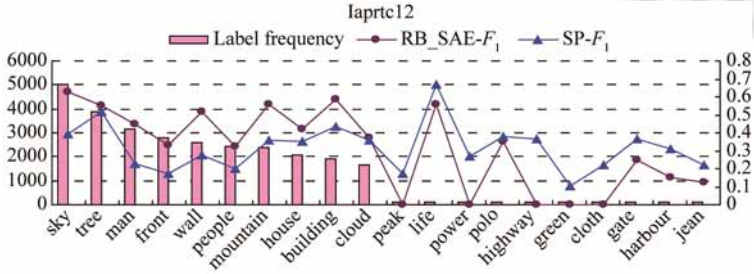


Fig.14 F_1 value of the high and low frequency labels on Iaprtc12 computed by RB-SAE and SP

图 14 在 Iaprtc12 上测试 RB-SAE 和 SP 预测高低词频标签 F_1 值

我们把提出的模型和一些经典模型以及近几年效果较好的模型在 3 个数据集上进行比较,从表 7 可以看出,我们的模型的效果大幅超过了一些经典模型,且比得上目前较好的模型.由于属性判别标注模型(ADA)综合了鲁棒自动编码器算法(RB-SAE)可以较好地预测高频属性图像和语义传播算法(SP)可以较好地预测中低频属性图像的特点(如图 12~图 14 所示),使得整体标注效果得到较大提升,在数据集 Corel5k 和 Iaprtc12 上得到的 F_1

© 中国科学院软件研究所 http://www.jos.org.cn

值是最好的.又由于局部均衡数据集平衡了高频标签和低频标签的出现频次,增大了低频标签被预测的可能性,因此至少准确预测 1 次的关键词个数 N^+得到很大提升,在 3 个数据集上分别达到 172、251 和 280 个,大幅超过其他模型.在 3 个数据集上,我们的模型的准确率 P 和召回率 R 也均达到当前较好水平,说明我们提出的 ADA 模型是有效的.

Table 7 Comparison between the ADA model and other models

表 7 本文 ADA 模型与其他模型的对比

模型	Corel5k				Espgame				Iaprtc12			
	P	R	F_1	$N+$	P	R	F_1	$N+$	P	R	F_1	$N+$
MBRM[5]	0.24	0.25	0.24	122	0.18	0.19	0.18	209	0.24	0.23	0.23	223
GS[27]	0.30	0.33	0.31	146	–	–	–	–	0.32	0.29	0.30	252
JEC[6]	0.27	0.32	0.29	139	0.24	0.19	0.21	222	0.29	0.19	0.23	211
TagProp-ML[7]	0.31	0.37	0.34	146	**0.49**	0.20	0.28	213	**0.48**	0.25	0.33	227
LM3L[28]	**0.33**	0.37	0.35	146	0.40	0.26	**0.32**	239	0.44	0.28	0.34	242
NW-RNN[29]	0.29	0.32	0.30	149	–	–	–	–	0.28	0.30	0.29	259
RNN[29]	0.31	0.34	0.32	149	–	–	–	–	0.33	0.31	0.32	255
χ^2Kernel[30]	0.31	0.39	0.35	153	0.38	0.21	0.27	214	0.42	0.24	0.31	239
ANNOR-G[31]	0.22	0.29	0.25	129	0.36	**0.29**	0.32	231	0.38	0.31	0.34	242
FFSS[32]	0.27	0.33	0.30	141	0.21	0.23	0.22	221	0.29	0.29	0.29	251
MLRank[33]	0.32	0.37	0.34	151	–	–	–	–	0.38	**0.32**	0.35	259
ADA	0.32	**0.40**	**0.36**	**172**	0.35	0.21	0.26	**251**	0.42	0.30	**0.35**	**280**

5.2.5 结果展示

下文所给出的表 8 中的 9 幅图是本文提出的 ADA 模型的标注实例.在图像自动标注结果一列中,我们用黑色加粗的字体表示自动标注结果与人工标注结果相同,而使用斜体字体的关键词表示可以描述图像的内容,但是并没有被人工标注出来.在这里,我们并没有选择完全被标注正确的那些图像,而是选择了部分能够比较好地反映本文模型特点的一些图像.从图中可以看出,本文一些图的标注结果虽然与原始图像上的人工标注结果有区别,但却是对原始图像标注结果的有益补充,能够更加准确地描述图像的语义信息.例如,第 1 幅图像的人工标注并未将 clouds 这一关键词给标注上,而从图像的场景来看,clouds 显然要作为一个重要的关键词来描述该幅图像的场景.在第 2 幅图像中,从人的视觉角度分析,显然用 sea 这个关键词相比原始图像中的 water 更有说服力,并且原始图像中也疏漏了 beach 等从图像中可以直接得到的关键词.此外,在对抽象概念 old,area 等的描述上,原始图像中的信息并不能对其进行准确的描述,或者说,单从人的视觉角度出发,无法从图像上得到这些信息.因此,这也从另一个角度说明了人工标注存在的一些问题,可能存在漏标注,并且不同人对同一幅图像的认识也存在一定的主观差异.同一幅图像,不同的人可能给出不同的标注结果.

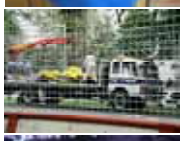
6 总 结

本文将栈式自动编码器(SAE)应用于图像标注任务,改善了传统的基于浅层机器学习模型标注效率低下、模型泛化能力弱等问题,并针对图像标注数据不平衡问题提出了两种解决方案:(1) 对于模型本身,我们提出了平衡栈式自动编码器(B-SAE)模型,该模型可以针对特定样本进行增强训练,解决了传统模型难以有效训练不平衡数据的问题,有效提升了低频标签的 F_1 值,并在 B-SAE 的基础上提出鲁棒平衡栈式自动编码器算法(RB-SAE),有效地解决了 B-SAE 模型本身不稳定的问题,并得到稳定的标注结果;(2) 对于标注过程,我们以待测图像作为出发点,首先为每一幅待测图像构造局部均衡数据集,并在此数据集上实现了一种有效提升中低频标签标注效果的语义传播算法(SP);然后通过判定待测图像的高低频属性来分步标注未知图像(ADA).一方面,用 RB-SAE 算法标注高频图像;另一方面,用 SP 算法标注中低频图像,标注过程取长补短,提升了整个模型的标注效果,3 个数据集上的实验验证了我们方法的有效性.此外,如何让标注模型更好地解决弱对象、背景信息、水印等噪声干扰问题是下一步的主要研究内容.

© 中国科学院软件研究所 http://www.jos.org.cn

Table 8 Annotation instances of the ADA model

表 8 本文 ADA 模型的标注实例

数据集	图像	人工标注结果	ADA 模型标注结果
Corel5k		mountain, sky, sun, water	*clouds*, *peaks*, *sunset*, elephant, silhouette
		sun, water, clouds, birds	**sun**, **water**, *sea*, *beach*, **birds**
		sky, water, people, sand	*beach*, **sand**, *shadows*, coyote, maui
Espgame		blue, desert, game, man, people, yellow	**game**, **man**, **people**, soldier, **yellow**
		old, sky, stone	*brown*, *mouth*, nose, **sky**, smile
		band, group, man, red	*black*, **group**, **man**, *people*, **red**
Iaprtc12		boy, door, front, jumper	**boy**, couch, **door**, **front**, *pullover*
		area, car, fence, racetrack, racing, tree	**car**, *man*, **racing**, spectator, **tree**
		grandstand, lawn, people, round, stadium, team, uniform	**lawn**, **round**, **stadium**, **team**, **uniform**

References:

[1] Mori Y, Takahashi H. Image-to-Word transformation based on dividing and vector quantizing images with words. In: Proc. of the 1st Int'l Workshop on Multimedia Intelligent Storage and Retrieval Management. 1999. 405–409. http://xueshu.baidu.com/s?wd=Image-to-Word+transformation+based+on+dividing+and+vector+quantizing+images+with+words&rsv_bp=0&tn=SE_baiduxueshu_c1gjeupa&rsv_spt=3&ie=utf-8&f=8&rsv_sug2=1&sc_f_para=sc_tasktype%3D%7BfirstSimpleSearch%7D&rsv_n=2

[2] Duygulu P, Barnard K, Freitas JFGD, Forsyth DA. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. European Conf. on Computer Vision, 2002,2353:97–112. [doi: 10.1007/3-540-47979-1_7]

[3] Wang M, Zhou XD, Zhang JQ, Xu HT, Shi BL. Image auto-annotation via an extended generative language model. Ruan Jian Xue Bao/Journal of Software, 2008,19(9):2449–2460 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/2449.htm [doi: 10.3724/SP.J.1001.2008.02449]

© 中国科学院软件研究所 http://www.jos.org.cn

[4] Lavrenko V, Manmatha R, Jeon J. A model for learning the semantics of pictures. In: Advances in Neural Information Processing Systems. 2004. 553–560. http://xueshu.baidu.com/s?wd=A+model+for+learning+the+semantics+of+pictures&tn=SE_baiduxueshu_c1gjeupa&cl=3&ie=utf-8&bs=Image-to-Word+transformation+based+on+dividing+and+vector+quantizing+images+with+words&f=8&rsv_bp=1&rsv_sug2=1&sc_f_para=sc_tasktype%3D%7BfirstSimpleSearch%7D&rsv_spt=3&rsv_n=2

[5] Feng SL, Manmatha R, Lawenko V. Multiple Bernoulli relevance models for image and video annotation. In: Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. 2004,2:1002–1009. [doi: 10.1109/CVPR.2004.1315274]

[6] Makadia A, Pavlovic V, Kumar S. A new baseline for image annotation. In: Proc. of the European Conf. on Computer Vision. 2008. [doi: 10.1007/978-3-540-88690-7_24]

[7] Guillaumin M, Mensink T, Verbeek J, Schmid C. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. IEEE Int'l Conf. on Computer Vision, 2009,30(2):309–316. [doi: 10.1109/ICCV.2009.5459266]

[8] Lu J, Ma SP. Automatic image annotation based on concept indexing. Journal of Computer Research and Development, 2007,44(3):452–459 (in Chinese with English abstract). [doi: 10.1360/crad20070313]

[9] Qiu ZY, Fang Q, Sang JT, Xu CS. Regional context-aware image annotation. Chinese Journal of Computers, 2014,37(6):1390–1397 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2014.01390]

[10] Wang H, Huang H, Ding C. Image annotation using bi-relational graph of images and semantic labels. IEEE Conf. on Computer Vision and Pattern Recognition, 2011,42(7):793–800. [doi: 10.1109/CVPR.2011.5995379]

[11] Gao L, Song J, Nie F, Sebe N, Shen H. Optimal graph learning with partial tags and multiple features for image and video annotation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. [doi: 10.1109/CVPR.2015.7299066]

[12] Tian F, Shen XK. Image semantic annotation method for weakly labeled dataset. Ruan Jian Xue Bao/Journal of Software, 2013, 24(10):2405–2418 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4424.htm [doi: 10.3724/SP.J.1001.2013.04424]

[13] Amiri SH, Jamzad M. Efficient multi-modal fusion on supergraph for scalable image annotation. Pattern Recognition, 2015, 48(7):2241–2253. [doi: 10.1016/j.patcog.2015.01.015]

[14] Kalayeh MM, Idrees H, Shah M. NMF-KNN: Image annotation using weighted multi-view non-negative matrix factorization. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2014. 184–191. [doi: 10.1109/CVPR.2014.31]

[15] Tariq A, Foroosh H. Feature-Independent context estimation for automatic image annotation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. [doi: 10.1109/CVPR.2015.7298806]

[16] Erhan D, Bengio Y, Courville A, Vincent P. Why does unsupervised pretraining help deep learning. The Journal of Machine Learning Research, 2010,11(3):625–660.

[17] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science, 2006, 504–507. [doi: 10.1126/science.1127647]

[18] Hinton GE, Osindero S, Yw T. A fast learning algorithm for deep belief nets. Neural Computation, 2006,18(7):1527–1554. [doi: 10.1162/neco.2006.18.7.1527]

[19] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. 2012. http://xueshu.baidu.com/s?wd=ImageNet+classification+with+deep+convolutional+neural+networks&tn=SE_baiduxueshu_c1gjeupa&cl=3&ie=utf-8&bs=A+model+for+learning+the+semantics+of+pictures&f=8&rsv_bp=1&rsv_sug2=1&sc_f_para=sc_tasktype%3D%7BfirstSimpleSearch%7D&rsv_spt=3&rsv_n=2

[20] Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol PA. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. The Journal of Machine Learning Research, 2010,11(12):3371–3408.

[21] Zhou B, Garcia AL, Xiao J, Torralba A, Oliva A. Learning deep features for scene recognition using places database. In: Advances in Neural Information Processing Systems. 2014. http://xueshu.baidu.com/s?wd=Learning+deep+features+for+scene+recognition+using+places+database&tn=SE_baiduxueshu_c1gjeupa&cl=3&ie=utf-8&bs=ImageNet+classification+with+deep+convolutional+neural+networks&f=8&rsv_bp=1&rsv_sug2=1&sc_f_para=sc_tasktype%3D%7BfirstSimpleSearch%7D&rsv_spt=3&rsv_n=2

[22] Pinheiro P, Collobert R. Recurrent convolutional neural networks for scene labeling. In: Proc. of the 31st Int'l Conf. on Machine Learning. 2014. 82–90. http://xueshu.baidu.com/s?wd=Recurrent+convolutional+neural+networks+for+scene+labeling&tn=SE_baiduxueshu_c1gjeupa&cl=3&ie=utf-8&bs=Learning+deep+features+for+scene+recognition+using+places+database&f=8&rsv_bp=1&rsv_sug2=1&sc_f_para=sc_tasktype%3D%7BfirstSimpleSearch%7D&rsv_spt=3&rsv_n=2

[23] Dan CC, Giusti A, Gambardella LM, Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. In: Advances in Neural Information Processing Systems. 2012. 2852–2860. http://xueshu.baidu.com/s?wd=+Deep+neural+networks+segment+neuronal+membranes+in+electron+microscopy+images&tn=SE_baiduxueshu_c1gjeupa&cl=3&ie=utf-8&bs=Recurrent+convolutional+neural+networks+for+scene+labeling&f=8&rsv_bp=1&rsv_sug2=1&sc_f_para=sc_tasktype%3D%7BfirstSimpleSearch%7D&rsv_spt=3&rsv_n=2

© 中国科学院软件研究所 http://www.jos.org.cn

[24] Luo P. Hierarchical face parsing via deep learning. IEEE Conf. on Computer Vision and Pattern Recognition, 2012,157(10): 2480–2487. [doi: 10.1109/CVPR.2012.6247963]

[25] Ryan K, Csaba S. Deep representations and codes for image auto-annotation. In: Advances in Neural Information Processing Systems. 2012. http://xueshu.baidu.com/s?wd=Deep+representations+and+codes+for+image+auto-annotation&tn=SE_baiduxueshu_c1gjeupa&cl=3&ie=utf-8&bs=Deep+neural+networks+segment+neuronal+membranes+in+electron+microscopy+images&f=8&rsv_bp=1&rsv_sug2=1&sc_f_para=sc_tasktype%3D%7BfirstSimpleSearch%7D&rsv_spt=3&rsv_n=2

[26] Wu JJ, Yu YN, Huang C, Yu K. Deep multiple instance learning for image classification and auto-annotation. In: Proc. of the IEEE Int'l Conf. on Computer Vision and Pattern Recognition. 2015. [doi: 10.1109/CVPR.2015.7298968]

[27] Zhang S, Huang J, Huang Y, Yu Y, Li H. Automatic image annotation using group sparsity. In: Proc.of the IEEE Conf. on Computer Vision and Pattern Recognition. 2010. 3312–3319. [doi: 10.1109/CVPR.2010.5540036]

[28] Hariharan B, Zelnik-Manor L, Varma M, Vishwanathan SVN. Large scale max-margin multi-label classification with priors. In: Proc. of the 27th Int'l Conf. on Machine Learning. 2010. 423–430. http://xueshu.baidu.com/s?wd=Large+scale+max-margin+multi-label+classification+with+priors&tn=SE_baiduxueshu_c1gjeupa&cl=3&ie=utf-8&bs=Deep+representations+and+codes+for+image+auto-annotation&f=8&rsv_bp=1&rsv_sug2=1&sc_f_para=sc_tasktype%3D%7BfirstSimpleSearch%7D&rsv_spt=3&rsv_n=2

[29] Cui C, Ma J, Lian T, Wang X, Ren Z. Ranking-Oriented nearest-neighbor based method for automatic image annotation. In: Proc. of the 36th Int'l ACM SIGIR Conf. on Research and Development In Information Retrieval. 2013. 957–960. http://xueshu.baidu.com/s?wd=Ranking-Oriented+nearest-neighbor+based+method+for+automatic+image+annotation&tn=SE_baiduxueshu_c1gjeupa&cl=3&ie=utf-8&bs=Large+scale+max-margin+multi-label+classification+with+priors&f=8&rsv_bp=1&rsv_sug2=1&sc_f_para=sc_tasktype%3D%7BfirstSimpleSearch%7D&rsv_spt=3&rsv_n=2

[30] Wang Y, Dawood H, Yin Q, Guo P. A comparative study of different feature mapping methods for image annotation. In: Proc. of the Int'l Conf. on Advanced Computational Intelligence. 2015. 340–344. [doi: 10.1109/ICACI.2015.7184726]

[31] Kuric E, Bielikova M. ANNOR: Efficient image annotation based on combining local and global features. Computers and Graphics, 2015,47:1–15. [doi: 10.1016/j.cag.2014.09.035]

[32] Zhang X, Liu C. Image annotation based on feature fusion and semantic similarity. Neurocomputing, 2015,149:1658–1671. [doi: 10.1016/j.neucom.2014.08.027]

[33] Li Z, Liu J, Xu C, Lu H. MLRank: Multi-Correlation learning to rank for image annotation. Pattern Recognition, 2013,46(10): 2700–2710. [doi: 10.1016/j.patcog.2013.03.016]

附中文参考文献:

[3] 王梅,周向东,张军旗,许红涛,施伯乐.基于扩展生成语言模型的图像自动标注方法.软件学报,2008,19(9):2449–2460. http://www.jos.org.cn/1000-9825/19/2449.htm [doi: 10.3724/SP.J.1001.2008.02449]

[8] 路晶,马少平.基于概念索引的图像自动标注.计算机研究与发展,2007,44(3):452–459. [doi: 10.1360/crad20070313]

[9] 邱泽宇,方全,桑基韬,徐常胜.基于区域上下文感知的图像标注.计算机学报,2014,37(6):1390–1397. [doi: 10.3724/SP.J.1016.2014.01390]

[12] 田枫,沈旭昆.一种适合弱标签数据集的图像语义标注方法.软件学报,2013,24(10):2405–2418. http://www.jos.org.cn/1000-9825/4424.htm [doi: 10.3724/SP.J.1001.2013.04424]

周铭柯(1990－),男,福建三明人,硕士,主要研究领域为模式识别,计算机视觉,深度学习,机器学习.

杜明智(1988－),男,硕士,主要研究领域为图像处理,人工智能.

柯逍(1983－),男,博士,讲师,主要研究领域为模式识别,计算机视觉,图像处理,机器学习.

© 中国科学院软件研究所 http://www.jos.org.cn