计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2020.20190793
57(11): 2348-2374, 2020

自动图像标注技术综述

马艳春 刘永坚 解 庆 熊盛武 唐伶俐
(武汉理工大学计算机科学与技术学院 武汉 430070)
(mayanchun@whut.edu.cn)

Review of Automatic Image Annotation Technology

Ma Yanchun, Liu Yongjian, Xie Qing, Xiong Shengwu, and Tang Lingli
(*School of Computer Science and Technology, Wuhan University of Technology, Wuhan* 430070)

Abstract As one of the most effective ways to reduce the "semantic gap" between image data and its content, automatic image annotation (AIA) technology has shown its great significance to help people understand image contents and retrieve the target information from the massive image data. This paper summarizes the general framework of AIA models by investigating the literatures about image annotation in recent 20 years, and analyzes the general problems to solve in AIA problems by combining the framework with various specific works. In this paper, the main methods used in various AIA models are classified into 9 types: correlation model, hidden Markov model, topic model, matrix factorization model, neighbor-based model, SVM-based model, graph-based model, CCA (KCCA) model and deep learning model. For each type of image annotation model, this paper provides a detailed study and analysis in terms of "basic principle introduction-specific model differences-model summary". In addition, this paper summarizes some commonly used datasets and evaluation indexes, and compares the performance of some important image annotation models with related analysis on the advantages and disadvantages of various types of AIA models. Finally, some open problems and research directions in the field of image annotation are proposed and suggested.

Key words automatic image annotation; image semantic analysis; image tagging; image content label; image content annotation

摘 要 图像自动标注技术是减少图像数据与内容之间"语义鸿沟"的其中一种最有效途径,对于帮助人类理解图像内容,从海量图像数据中检索感兴趣的信息具有重要现实意义.通过研究近20年公开发表的图像标注文献,总结了图像标注模型的一般性框架;并通过该框架结合各种具体工作,分析出在图像标注研究过程中需要解决的一般性问题;将各种图像标注模型所采用的主要方法归为9种类型,分别为相关模型、隐Markov模型、主题模型、矩阵分解模型、近邻模型、基于支持向量机的模型、图模型、典型相关分析模型以及深度学习模型,并对每种类型的图像标注模型,按照"基本原理介绍—具体模型差异—模型总结"3个层面进行了研究与分析.此外,总结了图像标注模型常用的一些数据集、评测指标,对一些比较著名的标注模型的性能进行了比较,并据此对各种类型的标注模型做了优缺点分析.最后,提出了图像标注领域一些开放式问题和研究方向.

收稿日期:2019-11-18;**修回日期**:2020-04-21
基金项目:国家自然科学基金项目(61602353);中央高校基本科研业务费专项资金(WUT:2017YB028)
This work was supported by the National Natural Science Foundation of China (61602353) and the Fundamental Research Funds for the Central Universities (WUT:2017YB028).
通信作者:解庆(felixxq@whut.edu.cn)

关键词 自动图像标注;图像语义分析;图像标识;图像内容标签;图像内容标注

中图法分类号 TP391

随着计算机软硬件、互联网、大数据及分布式存储等技术的不断成熟和快速发展,图像数据在数量和内容上呈现爆炸式增长.2017 年 1 月中国互联网络信息中心(China Internet Network Information Center, CNNIC)发布的《中国互联网发展状况统计报告》显示,网页中的图片所占比率已达总的多媒体形式的 79.63%,以数字图像作为载体也是文化资源数字化的最主要方式.然而,在数字图像数据保持高速增长的同时,人们对图像数据的利用能力却没有随之增强.究其原因,是计算机难以通过图像的低层视觉特征提取出可供人类理解的高层语义信息,低层视觉特征和高层语义特征之间存在"语义鸿沟"的缺陷.这也导致我们在应对大规模图像数据时缺少有效的检索方案,从而难以获取所需信息.

图像自动标注技术是减少"语义鸿沟"的最有效的途径之一,其以图像为研究目标,以知识为研究手段,利用人工智能和模式识别等方法完成对图像内容的语义解释,使计算机系统自动获取图像蕴含的信息内容,从而协助人们完成对图像信息的获取,检索到感兴趣的内容.因此研究图像的自动标注技术和算法,对于帮助人类从海量图像数据中检索兴趣内容,获取所需信息,具有重要现实意义.

在 2003 年以前,国内外学者对图像自动标注技术的研究仍然处于初级探索阶段,随后广大学者不但加强了对该技术的关注度,同时也取得了一定的研究成果.考察已有研究成果,大部分工作仍是将解决或缩小图像的视觉特征表达与高层语义信息之间的鸿沟问题作为研究重点,主要探索方向为:1)选取鲁棒性强、适应广泛的图像特征;2)建立有效的计算模型;3)设计更加适用的标注算法,使图像标注的上下文信息得到更加充分的利用;4)针对图像本身数据量大、标签空间特征维度高、已有图像标签环境复杂的特点,如何在不影响性能的情况下降低标签空间维度,去除已有图像的标签噪声.到目前为止,对各种已经出现的图像标注模型进行统一分类、梳理的综述性工作仍然相对缺乏,少量的综述性研究工作[1-4]往往存在分类单一、归类模糊以及综合性不强等问题.因此,本文旨在通过深入分析和研究公开发表的图像标注文献,系统归类已有图像标注模型,总结各类模型的优缺点、一般性问题及一般性框架,为后续图像自动标注领域的研究工作提供帮助与思路.

本文的贡献:1)通过研究近 20 年公开发表的图像标注文献,总结了图像标注模型的一般性框架;并通过该框架结合各种具体工作,分析出在图像标注问题中需要解决的一般性问题.2)对各种图像标注模型按照其主要使用的方法类型进行了归类;对每一类方法类型的图像标注模型,首先进行了基本的原理介绍,然后对该方法类型下的图像标注模型之间的差异进行了具体的分析,最后对每一类方法类型的标注模型做简单总结.3)总结了一些比较著名的标注模型的性能和实验数据,并据此对各种方法类型的标注模型做了优缺点分析.4)总结了图像标注模型常用的一些数据集和评测指标.5)给出了图像标注领域一些开放式问题和研究方向.

1 图像标注的一般问题

图像的自动标注是利用人工智能或模式识别等计算机方法对数字图像的低层视觉特征进行分析,从而对图像打上特定语义标签的一个过程.本文通过对近 20 年各种图像标注模型(方法)进行深入分析与研究,总结出图像标注模型的通用框架(如图 1 所示),并依据通用框架中对应的各部分内容,归纳出图像标注技术中存在的一些一般性问题.

如图 1 所示,图像的标注框架总体可分为 3 个模块,包括 2 个特征提取模块和 1 个标注模型模块.其中,2 个特征提取模块表示通过图像的特征提取方法以及词汇(标签)的特征提取方法可分别得到对应的图像低层视觉特征与标注词特征(也称为标签特征).图像的标注模型表示通过需要建立最关键的"图像-标签(I-W)"之间的关联关系,并通过该关联关系和低层视觉特征对未标注图像进行标签预测;同时,更进一步地充分利用"图像-图像(I-I)"、"标签-标签(W-W)"关系对标注模型进行优化,使其得到更加稳健和鲁棒的标注结果.

图像的低层视觉特征提取方法包括全局特征提取方法(如颜色、形状、纹理、直方图等)以及局部特征提取方法(如 SIFT(scale invariant feature transform)角点、斑点等);词汇的特征提取方法包括"one-hot"以及"word2vec"等.根据标注模型的不同,研究人员往往需要选择不同的特征提取方法使选取的特征能够适应特定的应用场景,从而促进模型性能的

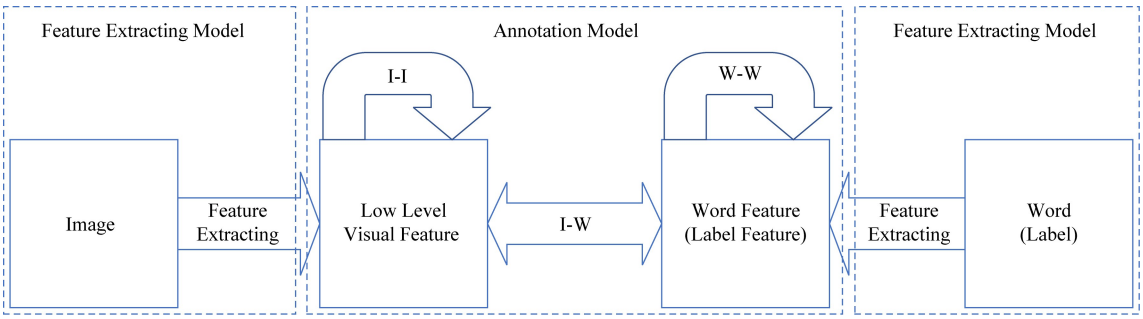

Fig. 1 General framework of the image annotation model

图 1 图像标注模型的通用框架

提升.由于本文关注的重心在于标注模型本身,因此对特征提取方法不做过多讨论.

图像的标注模型最关键的是需要充分利用低层视觉特征和标注词特征,建立起"图像-图像"、"图像-标签"以及"标签-标签"3 种关系.但是由于图像训练集本身存在的固有特点、特征提取算法存在差别以及模型对各种特征的适应性不同,所以标注模型在建立 3 种关系的同时,还需要考虑 7 个一般性问题:

1) 标签的不均衡问题.在图像的训练集中,有少部分标签只存在于较少的图像之中,而另外一些标签则出现频率较高,这种标签分布的不均衡有可能会影响模型的精确性.造成这种问题的原因是在制作训练集时,人们往往倾向用更加广泛和一般性的词汇来进行标注,从而导致标签的频率不尽相同.

2) 弱标签问题.这种问题通常在社交领域图像集上出现,即训练集图像中的标注并不能完整地表现图 1 中反映的所有语义信息,存在标签缺失或错误的情况.产生这种现象的原因是人们在标注时的主观性不同.

3) 特征的高维度问题.从图像中提取的特征往往维度过高,导致模型计算量增加;此外,维度过高也会产生特征的冗余与噪声.

4) 特征内维度不均衡问题.由于标注模型往往需要使用多种低层的图像特征共同作用进行标签预测,而每种特征以及特征内的每一维度对标签预测的贡献程度不一致,影响模型精确性.

5) 特征的选择问题.针对特定标注模型所选取、设计的视觉特征,在其他模型上通常表现较差.因此,在设计新的标注模型时需考虑在多种图像特征之中选取具有更加泛化性能的特征.

6) 3 种关系利用不全的问题.由于在"图像-图像"、"图像-标签"以及"标签-标签"3 种关系中,只需通过最关键的"图像-标签"间相关关系即可对未标注图像进行标签预测.因此,很多模型忽视了"图像-图像"、"标签-标签"2 种关系,影响了标注精度的进一步提升.

7) 标注模型的计算量和运行效率问题.

图像自动标注模型的研究重点基本聚焦于解决上述 7 个问题,从而使标注结果更加精确、模型运行高效并能适用于更多的应用场景.

2 图像标注模型

图像标注领域自本世纪初进入快速发展时期以来,出现了各种不同的方法和模型,也有学者对各种图像标注模型进行综述性研究.然而,需要说明的是,这些研究工作往往存在分类单一和不够清晰的问题.如文献[1-2]主要以基于内容的图像检索角度进行综述,文献[3]主要以统计方法的角度进行综述,文献[4]虽以相对综合的方式进行综述,但是其对方法的类型和标注问题进行了混合,使得归类不够清晰.本文针对这些问题,将众多的图像标注算法和模型依据其主要使用的方法类型分为相关模型、隐 Markov 模型、主题模型、矩阵分解模型、近邻模型、基于支持向量机的模型、图模型、典型相关分析模型以及深度学习模型 9 个大类方法类型,并依次对每种方法类型的图像标注模型进行分析与总结.在每种方法类型的相关小节,首先对共性的原理进行介绍,再分别介绍每一种图像标注模型的差异.

2.1 相关模型

相关模型是图像标注领域受到关注之后最早出现的一类模型,相关模型的代表有 TM(translation model)模型[5]、CMRM(cross-media relevance model)模型[6]、CRM(continuous-space relevance model)模型[7]以及 MBRM(multiple Bernoulli relevance

model)模型[8].其基本思想为:首先将图像分块,假定分块的图像特征和标签之间存在某种特定的概率;然后建立分块图像的特征和标签之间的联合概率密度;最后根据待标注图像的分块信息,求得其针对每个标签的后验概率.

TM模型[5]借鉴了一种词汇翻译模型的思想.词汇翻译模型的原理是:给定一段分别用2种语言翻译过的文档,其在2种语言上的对应关系仅仅在粗粒度(如段落或句子)上已知,而细粒度(如单词与单词)之间的对应关系未知,词汇翻译模型即需要求解这种精确的细粒度的单词与单词之间的对应关系.同理,图像标注问题可类比为:给定了训练集每幅图像以及和其对应的标注词汇,但是细化的图像区域与每个单词之间的对应关系未知,模型即需要求解细粒度的图像区域与单词之间的对应关系.TM模型首先将图像按照分割算法分块,然后用K-means算法对这些块的图像特征进行聚类,每一个聚类称之为1个blob.假定每一个blob对应1个单词,则所有的blob和单词之间存在详细的对应关系,即"probability table".由于训练集中的图像和标注词汇的对应关系已知,因此,可采用EM(expectation-maximization)算法求解"probability table".对于待标注图像,先求出其对应的blob,然后根据"probability table"选择具有最高概率值的单词作为该图像的标注.

CMRM模型[6]同样需要先将图像进行分块,然后根据图像块的特征聚类成blob,并假定每副图像I符合一个潜在的概率分布$P(\cdot|I)$.对于图像标注问题来说,假定图像I可以由blob近似表示为$\{b_1,b_2,\cdots,b_m\}$,单词表示为w,则待求的$P(w|I)$可以近似地由贝叶斯公式得到:

$$P(w|I)\approx P(w|b_1,b_2,\cdots,b_m)=\frac{P(w,b_1,b_2,\cdots,b_m)}{P(b_1,b_2,\cdots,b_m)}. \tag{1}$$

针对训练集τ,如果假定w和$b_1,b_2,\cdots,b_m$相互独立,则$P(w,b_1,b_2,\cdots,b_m)$可根据全概率公式得到:

$$P(w,b_1,b_2,\cdots,b_m)=\sum_{J\in\tau}P(J)P(w,b_1,b_2,\cdots,b_m|J)=\sum_{J\in\tau}P(J)P(w|J)\prod_{i=1}^{m}P(b_i|J). \tag{2}$$

由于训练集为固定大小,所以可认为先验概率$P(J)$为定值.$P(w|J)$和$P(b_i|J)$可根据训练集中的统计信息获得:

$$P(w|J)=(1-\alpha_J)\frac{\#(w,J)}{|J|}+\alpha_J\frac{\#(w,\tau)}{|\tau|}, \tag{3}$$

$$P(b|J)=(1-\beta_J)\frac{\#(b,J)}{|J|}+\beta_J\frac{\#(b,\tau)}{|\tau|}, \tag{4}$$

其中,$\#(w,J)$表示某一个单词在图像J中出现的次数(一般来说为0或者1),$\#(w,\tau)$表示该单词w在整个训练集τ中出现的次数.$\#(b,J)$表示图像J中被标记为b的blob个数,$\#(b,\tau)$表示在整个训练集中被标记为b的blob个数.$|\tau|$为整个训练集的大小,α_J和β_J为平滑因子,控制单词和blob在图像和训练集中的占比.

CRM模型[7]没有像TM模型和CMRM模型采取将图像分块的方法,而是将图像看作很多包含了显著物体的区域$\{r_1,r_2,\cdots,r_n\}$叠加的结果.假定1幅图像可以被表示为$R=C^{M\times H}$,每一个区域为一个r,r中的部分像素代表了图像中的某个显著物体,其余部分像素被设置为"透明",多个区域叠加即构成图像R,词汇$\{w_1,w_2,\cdots,w_m\}$用来描述区域$\{r_1,r_2,\cdots,r_n\}$.G为映射函数,其功能为将区域$r\in R$映射成为一个实值特征向量$\boldsymbol{g}\in\mathbb{R}^k$,用来表示区域$r$的图像特征.则训练集$\tau$中的任一图像$J$可表示为1组图像区域和标注词汇的组合$\{R_A,W_B\}$,其中$R_A=\{r_1,r_2,\cdots,r_{n_A}\}$,$W_B=\{w_1,w_2,\cdots,w_{n_B}\}$.$\{R_A,W_B\}$的生成过程可分为3个独立的步骤:

1) 从训练集τ中抽样图像J,图像J在训练集中的先验概率记为$P_\tau(J)$;

2) 在$P_V(\cdot|J)$的分布下,依次对W_B中的每个单词w进行抽样,其中$P_V(\cdot|J)$表示单词在图像J条件下的概率,其中V代表潜在的概率映射函数;

3) 在$P_R(\cdot|J)$的分布下,依次对R_A中的每个区域r进行抽样,其中$P_R(\cdot|J)$表示区域在图像J条件下的概率;

则$\{R_A,W_B\}$的联合概率密度可表示为

$$P(R_A,W_B)=\sum_{J\in\tau}P_\tau(J)\prod_{b=1}^{n_B}P_V(w_b|J)P_R(R_A|J)=\sum_{J\in\tau}P_\tau(J)\prod_{b=1}^{n_B}P_V(w_b|J)\times\prod_{a=1}^{n_A}\int P_R(r_a|\boldsymbol{g}_a)P_G(\boldsymbol{g}_a|J)\mathrm{d}\boldsymbol{g}_a, \tag{5}$$

其中,$P_\tau(J)$为先验概率,取统一值$P_\tau(J)=\frac{1}{N_\tau}$,$N_\tau$为训练集图像数量.$P_R(r|\boldsymbol{g})$用来将特征向量$\boldsymbol{g}$映射为区域$r$,$P_G(\cdot|J)$用一个非参的基于高斯核

的多元密度函数代表特征向量的概率，$P_V(w_b|J)$为加入了狄里克雷先验的概率，用来表示标注词汇的概率.表达式分别为：

$$P_R(r|\boldsymbol{g})=\begin{cases}1/N_{\boldsymbol{g}}, \text{ if } G(r)=\boldsymbol{g};\\ 0, \text{ otherwise.}\end{cases} \tag{6}$$

$$P_G(\boldsymbol{g}|J)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sqrt{2^k\pi^k|\boldsymbol{\Sigma}|}}\exp\{(\boldsymbol{g}-G(r_i))^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{g}-G(r_i))\}, \tag{7}$$

$$P_V(w|J)=\frac{\mu p_w+N_{w,J}}{\mu+\sum_{w'}N_{w',J}}, \tag{8}$$

$N_{\boldsymbol{g}}$是所有可以令$G(r)=\boldsymbol{g}$的区域的总数，由于无法精确估计因此取某一不依赖于$\boldsymbol{g}$的常数.$\boldsymbol{\Sigma}$表示向量$\boldsymbol{g}$每一维特征的协方差矩阵，$N_{w,J}$代表单词w在W_J中出现的次数，p_w是单词w在训练集中出现的频率，w'指图像J中出现的所有单词.

CRM模型直接求取区域与词汇的联合概率密度，而非像TM模型中将单词和图像中分块做一一对应.此外，由于其不需要生成blob，因此不会像TM模型和CMRM模型一样受到聚类算法的影响.最后，由于其直接建模的是连续特征，因此也不会受到特征离散化的干扰.

MBRM模型[8]为基于CRM模型的改进，CRM，TM，CMRM模型都是基于词的多项式分布，而MBRM模型是将词的分布看作为多重伯努利分布.基于词的多重伯努利分布在于解决多项式分布中存在的一个缺点，即在训练集中图像标注词汇长短不一或具有层次结构时会存在索引精度降低的问题.此外，MBRM模型为解决图像分块受聚类算法影响的问题，直接将图像分成固定大小的方块，降低了计算时间，使模型参数更容易估计，更容易关联上下文和模型.MBRM模型对$\{R_A,W_B\}$的联合概率密度可表示为

$$P(R_A,W_B)=\sum_{J\in\tau}\{P_\tau(J)\prod_{a=1}^{n_A}P_G(\boldsymbol{g}_a|J)\prod_{w\in W_B}P_V(w|J)\prod_{w\notin W_B}(1-P_V(w|J))\}. \tag{9}$$

在实际标注过程中，W_B由于不可能遍历所有词汇组合，因此只选取1个词，然后根据式(10)选取所需词汇数量即可：

$$w^*=\arg\max_{w\in\{0,1\}}\frac{P(R_A,w)}{P(R_A)}. \tag{10}$$

$P_V(w|J)$即为采用的多重伯努利分布：

$$P_V(w|J)=\frac{\mu\delta_{w,J}+N_w}{\mu+N}, \tag{11}$$

其中，N_w代表训练集中标注有单词w的图像个数，N代表训练集图像总数.如果w是J的标注，则$\delta_{w,J}=1$否则$\delta_{w,J}=0$，μ是平滑参数.

总体来说，相关模型建立了“图像-标签(I-W)”之间的关联关系，但是此类模型标注精度并不高，其主要原因是相关模型需要假定图像中的目标与标注词汇之间存在某种概率分布关系，而由于训练集中的图片数量有限，所建立的概率分布模型往往只能反映特定的训练集，泛化性能不高.此外，模型对图像中的目标需要依赖于分块、分割以及聚类算法的先序处理，其精度、图像中目标的特征构建也会受分块以及聚类算法的影响.从速度方面来看，相关模型的计算也比较复杂，如EM算法需要大量迭代，因此耗时较高.

2.2 隐Markov模型

隐Markov模型(hidden Markov model, HMM)类似于相关模型，同样需要根据图像块和标注词的联合概率密度来求得最终的标注，但不同之处在于隐Markov模型是通过隐Markov链来建立这种相关关系.HMM的代表模型有文献[9-13]等.

HMM模型[9]利用$I=\{i_1,i_2,\cdots,i_T\}$代表图像中被分割的有序区域，图像区域按照固定个数4×6来进行划分.每一区域的图像特征为d维向量$\boldsymbol{x}_t\in\mathbb{R}^d$，$\{c_1,c_2,\cdots,c_N\}$代表图像的标注词汇.模型将标注词汇看作是隐Markov链，也即链上的隐含状态$s_t\in\{c_1,c_2,\cdots,c_N\}$，$\boldsymbol{x}_t$由潜在的分布函数$f(\cdot|s_t)$产生.根据Markov链公式，由有序区域$r_1=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_T\}$和某一有序标注词汇$S_1=\{s_1,s_2,\cdots,s_T\}$组成的联合概率密度可表示为

$$f(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_T,s_1,s_2,\cdots,s_T|s_0)=f(r_1,S_1|s_0)=\prod_{t=1}^{T}f(\boldsymbol{x}_t|s_t)p(s_t|s_{t-1}), \tag{12}$$

其中s_0认为已知，由于图像的标注词汇是无序的，用最大似然的方法去掉式(12)中标注词汇的序列性即有：

$$f(r_1,C|s_0)=\sum_{S\in C}\prod_{t=1}^{T}f(\boldsymbol{x}_t|s_t)p(s_t|s_{t-1}), \tag{13}$$

其中C为所有的词汇序列组合.对未知图像是否标注某一词汇c，可根据贝叶斯公式得到：

$$P(s_t=c|r_1,s_0)=\frac{f(r_1,s_t=c|s_0)}{f(r_1|s_0)}=\frac{\sum_{S_1:s_t=c}\prod_{t=1}^{T}f(\boldsymbol{x}_t|s_t)p(s_t|s_{t-1})}{f(r_1|s_0)}, \tag{14}$$

$$f(r_1|s_0)=f(r_1,V|s_0)=\sum_{S_1\in V}\prod_{t=1}^{T}f(\boldsymbol{x}_t|s_t)p(s_t|s_{t-1}),\quad(15)$$

式(15)为定值,因此

$$P(s_t=c|r_1,s_0)\propto\sum_{S_1:s_t=c}\prod_{t=1}^{T}f(\boldsymbol{x}_t|s_t)p(s_t|s_{t-1}),\quad(16)$$

$$f(\boldsymbol{x}|c)=\sum_{m=1}^{M}\frac{w_{m,c}}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}_{m,c}|}}\times\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_{m,c})^{\mathrm{T}}\boldsymbol{\Sigma}_{m,c}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{m,c})\right),\quad(17)$$

$f(\boldsymbol{x}|c)$表示隐 Markov 模型的混淆矩阵概率,其中 $w_{m,c}$表示权值,$\boldsymbol{\mu}_{m,c}$代表在所有词汇 V 中 c 所对应的所有区域特征 $\boldsymbol{x}$ 的均值,$\boldsymbol{\Sigma}_{m,c}$代表向量 $\boldsymbol{x}$ 元素间协方差矩阵.

TSVM-HMM (transductive support vector machine based HMM)模型[10]的提出,是为了改善 HMM 模型中存在的问题,即若训练集中已标注的图像区域过少,会导致对 HMM 中混合概率密度 $f(feature|word)$估计不准确.TSVM-HMM 即基于支持向量机(support vector machine, SVM)的半监督学习方法,首先利用训练集中已标注的图像区域训练一个二分类的 SVM 分类器,接着根据该分类器对训练集中未标注区域进行分类,将其中最可能的相关区域和非相关区域分别加入到对应的训练集中;然后根据新扩展的训练集重新训练 SVM,并重复对未标注区域的分类过程,直到重复次数达到预设的最大迭代次数.最后,由于新扩展的训练集之中,具有更多的已标注图像区域,因此对 HMM 中混合概率密度的建模也会更加准确.

SHMM(spatial HMM)模型[11]以及 SMK(spatial Markov kernels)模型[12]为 HMM 模型在垂直方向上的扩展,也即对于图像的分块来说,SHMM 模型分别考虑了水平和垂直方向前一相邻图像分块对当前图像分块的影响:

$$P(q_{l,m}|Q_{l,m\ominus l})=P(q_{l,m}|q_{l\ominus1,m}q_{l,m\ominus1})=\underbrace{P(q_{l,m}|q_{l,m\ominus1})}_{\text{水平方向}}\underbrace{P(q_{l,m}|q_{l\ominus1,m})}_{\text{垂直方向}},\quad(18)$$

l,m 表示图像分块的位置下标,q 代表图像分块的状态,$\ominus$可理解为减号,代表在横向或纵向的位置下标做向前移动的操作.

HMM-SVM 模型[13]首先分别对颜色和纹理特征用 Markov 模型进行建模,据此可分别得到基于颜色和纹理特征的对图像分块区域的标注概率,在此过程之后,每个图像分块都可以得到一个二元的预测组$\{P_{\text{color}},P_{\text{texture}}\}$,其中 P_{color}和 P_{texture}分别代表某图像块基于颜色和基于纹理特征的预测结果.然后,将此二元预测组$\{P_{\text{color}},P_{\text{texture}}\}$作为中级的图像输入特征,标注单词作为分类结果,训练出多个 one-against-all 的 SVM 分类器.最后,根据训练出的 SVM 分类器对图像进行标注.

基于 HMM 的模型用一种很自然的方式建模了每个单词和图像特征之间的关系 $f(feature|word)$,也即"图像-标签(I-W)"之间的关联关系,从可解释性上提供了有效的推导过程;而且相较于相关模型,基于 HMM 的模型更关注抽象信息,如对整个语料库来说,仅保留了如均值和方差等图像的高层次特征,因此模型计算效率较高.然而基于 HMM 的模型也继承了 Markov 模型的固有缺陷,即在给定的标注词条件下,图像的特征是条件独立的,没有利用图像内容上的相关性,并且对于"标签-标签(W-W)"与"图像-图像(I-I)"特征之间所存在的复杂语义关系,仅通过混合矩阵进行建模也不够精确.

2.3 主题模型

LSA(latent semantic analysis)模型是主题建模的基础,其最早的用途是对文档进行检索[14],它的核心思想是把对应的"文档-项"矩阵分解成相互独立的"文档-主题"矩阵和"主题-项"矩阵,从而在隐藏的主题空间建立文档和词汇之间的语义关系.在文档检索领域,"项"即对应检索词汇.在图像标注领域,LSA 模型将图像作为一个独立的文档,标注词汇或者视觉特征等被定义为"项".假定"文档-项"矩阵表示为 $\boldsymbol{A}\in\mathbb{R}^{N\times M}$,$N$ 为文档的数量,M 为项的数量,则矩阵 $\boldsymbol{A}$ 可通过奇异值分解(singular value decomposition, SVD)为

$$\boldsymbol{A}\approx\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}},\quad(19)$$

其中 $\boldsymbol{U}\in\mathbb{R}^{N\times K}$,$\boldsymbol{S}\in\mathbb{R}^{K\times K}$,$\boldsymbol{V}\in\mathbb{R}^{M\times K}$,$K$ 表示降维后的主题空间维度,$\boldsymbol{U}$ 的每一行代表训练集图像在主题空间中的特征表示,该主题空间(也称为隐语义空间)可表示"项"之间的语义关联关系.当利用训练集图像求解 $\boldsymbol{U}$ 与 $\boldsymbol{V}$ 之后,对未标注图像 $\boldsymbol{q}\in\mathbb{R}^{1\times M}$来说,可先将其映射至主题空间

$$\hat{\boldsymbol{q}}=\boldsymbol{q}\boldsymbol{V},\quad(20)$$

然后将 $\hat{\boldsymbol{q}}$ 与 $\boldsymbol{U}$ 的每一行进行相似度计算,利用最相似的前 n 项训练图像标注作为待标注图像的标注词汇.

对 LSA 模型的改进一般集中在对"文档-项"矩阵中"项"的表征方面.如文献[15]首先将图像划分

为上、中、下 3 个区域,然后通过图像的标注词以及每个区域的 RGB 颜色直方图、区域号和区域坐标串联起来,形成的向量表征“文档-项”矩阵中“项”.

文献[16]将图像视觉特征和标注词汇特征融合的结果作为“文档-项(document-term)”矩阵中的“项”.对于每幅图像中的视觉特征,其采用了一种视觉词袋(bag of words, BOW)的模型进行构建.首先对每幅图像按矩形栅格以及分割算法分块,同时求出每幅图像的角点;然后,串联每个图像块的质心位置、RGB 直方图、Gabor 系数以及每个角点的 SIFT 特征[17]作为图像的视觉特征;之后再用 K-means 聚类算法将图像集中所有的视觉特征聚类为 k 个簇,每幅训练图像即可根据其视觉特征在聚类簇中出现的次数离散化为一种“次数向量”作为该图像的“视觉项”;最后,训练集中所有图像即可形成新的“文档-视觉项”矩阵 $\boldsymbol{M}_{d,v}$.对于每幅图像中标注词汇的特征,其采用向量空间模型(vector space model, VSM)进行构建,也即获得其词频以及逆文档词频作为“词汇项”,从而使训练集中所有图像形成新的“文档-词汇项”矩阵 $\boldsymbol{M}_{d,t}$.在 $\boldsymbol{M}_{d,v}$ 与 $\boldsymbol{M}_{d,t}$ 获取之后,根据原始标注图像视觉特征与标注词汇之间的对应关系获取“视觉项”与“词汇项”的互相关矩阵 $\boldsymbol{M}_{t,v}$,之后即可将 $\boldsymbol{M}_{t,v}$ 降维到主题空间,尽管取得了一定的成功,基于 SVD 的 LSA 仍然存在计算量较大的缺点,而且从概率的意义上也缺乏可解释性.

PLSA(probabilistic latent semantic analysis)模型[18]是一种在概率意义上具有可解释性的主题模型.模型假定主题 z_k 为基于“文档”和“项”的生成模型中存在的一个隐藏主题空间中的元素,“文档”与“项”之间相互独立.则包含有“文档-项-主题”三者的联合概率密度可表示为

$$P(x_j,z_k,d_i)=P(d_i)P(z_k|d_i)P(x_j|z_k),\quad(21)$$

其中,x,z,d 分别代表“项”、“主题”与“文档”,i,j,k 分别为其对应的索引.通过求解 z 的边缘概率密度即可求得“文档-项”的联合概率密度:

$$P(x_j,d_i)=P(d_i)\sum_{k=1}^{K}P(z_k|d_i)P(x_j|z_k),\quad(22)$$

类似于相关模型,这里 $P(d_i)$ 代表某文档在整个训练集中的占比,$P(z|d)$ 和 $P(x|z)$ 分别代表选定文档包含某一主题的概率以及选定主题包含某一“项”的概率,在训练阶段,可根据训练集由 EM 算法求解.在图像标注领域,$P(d_i)$ 表示图像在训练集中的占比,“项”x 可表示普通变量,也可表示特征向量,其物理含义根据各种模型变体略有不同.如文献[15]中的 PLSA-MIXED 模型,串联单词特征与视觉特征形成 $\boldsymbol{x}=(word,visual)$.在测试阶段,对于未标注图像,则将其标注部分置 0 为 $\boldsymbol{x}_{\text{new}}=(0,visual_{\text{new}})$,然后根据 folding-in 算法[18]求得其 $P(z|d_{\text{new}})$,从而进一步得到 $P(\boldsymbol{x}|d_{\text{new}})$,图像的标注 $P(word|d_{\text{new}})$即可从 $P(\boldsymbol{x}|d_{\text{new}})$抽取得到.对 PLSA 模型的改进大多集中在对多种模态的利用方式上以及多种图像视觉特征的使用方式层面.如文献[19-20]中的 PLSA-WORD 模型,将图像与标注词汇视为 2 种不对等模态,通过不对称的学习算法从标注词汇的数据中学习一个潜在的空间,并将其关联到视觉模态,从而改进了 PLSA-MIXED 模型中将标注词汇与图像视觉特征 2 种模态视为同等重要的缺陷,也即改善了标注词汇数量和图像视觉特征数量之间的不平衡问题.文献[21]提出的 PLSA-FUSION 模型和文献[22]提出的 MM-PLSA(multilayer multimodal PLSA)模型在 PLSA-WORD 模型基础上进一步使用 2 组潜在主题分别从标注词汇和图像视觉特征 2 种模态中学习,然后再融合为一个共同的潜在空间.其中,PLSA-FUSION 模型采用了一种自适应的动态方法进行学习,根据每幅图像各自的视觉词分布确定各不相同的权值对 2 组主题进行融合;MM-PLSA 模型将标注词汇和图像视觉特征的 2 种模态学习到的主题作为 2 片叶子,再通过 1 个顶层的 PLSA 根节点构建 1 个多层次的主题空间树,进而对 2 组主题进行融合.文献[23]提出的 MF-PLSA(multi-feature PLSA)模型在原始模型基础上采用了多特征组合而非融合的方式进行改进,其将图像的 SIFT[17]特征 $\boldsymbol{v}$ 和局部变换颜色直方图特征(local transformed color histogram, LTCH)$\boldsymbol{w}$ 分别作为条件独立于主题 z 的变量,从而求得图像视觉特征(SIFT 和 LTCH)与文档 d 的联合概率密度变为

$$P(\boldsymbol{w},\boldsymbol{v},d)=\prod_{d}P(d)\prod_{(\boldsymbol{w},\boldsymbol{v})}P(\boldsymbol{w},\boldsymbol{v}|d)^{N(\boldsymbol{w},\boldsymbol{v},d)},\quad(23)$$

其中 $N(\boldsymbol{w},\boldsymbol{v},d)$代表$(\boldsymbol{w},\boldsymbol{v},d)$的共现次数.

主题模型最早用于对文档的检索,解决了检索问题中的“一词多义”以及“一义多词”问题.在图像标注领域,主题模型同样通过构建隐藏的主题空间,使得具有语义相似度的模态能够映射到同一主题,或者同一主题可被多种模态所表示.因此,隐藏的主题空间能够较好地建立起图像底层视觉特征同自然语义之间的联系,也即“图像-标签(I-W)”之间的关联关系.但由于主题模型依然是通过选取训练集图像中相应的底层视觉特征和标注词汇来进行概率

运算,因此其概率分布难以有效描述样本外的情况,泛化性能不高.对于选取何种底层视觉特征、标注词汇特征,以及对特征的融合利用也是主题模型需要解决的难题.此外,主题模型中需用到 SVD 分解以及 EM 算法等,都需要耗费大量时间以及运算资源.

2.4 矩阵分解模型

基于矩阵分解的图像标注通过矩阵分解的方式来建立图像、标签等之间的相关关系.本文归类于主题模型中的 LSA 模型,通过 SVD 分解的方式建立隐语义空间,因此也可归类于矩阵分解模型,但由于其分解之后的物理意义更偏重隐藏的语义主题,因此本文将其归类于主题模型.

另一类矩阵分解如非负矩阵分解(nonnegative matrix factorization, NMF)模型最初是用来做人脸识别任务[24],其核心思想是将图像的特征矩阵 $\boldsymbol{V}$ 分解为 2 个非负的矩阵因子 $\boldsymbol{W}$ 和 $\boldsymbol{H}$:

$$\boldsymbol{V}\approx\boldsymbol{W}\boldsymbol{H},$$
$$\text{s.t. } \forall W_{i,j},H_{i,j}\geqslant 0, \tag{24}$$

其中,$\boldsymbol{W}$ 代表特征在新的空间形成的基,$\boldsymbol{H}$ 可代表某样本在基 $\boldsymbol{W}$ 下的坐标,因此如果将 $\boldsymbol{W}$ 形成的空间看做新的语义空间,则 $\boldsymbol{H}_{\cdot,j}$ 可代表样本在语义空间下的特征表示,因此基于 NMF 的图像标注模型关键在于寻找可靠的矩阵分解因子 $\boldsymbol{W}$ 和 $\boldsymbol{H}$.基于矩阵分解的图像标注模型代表有文献[25-30]等.

文献[25]类似于文献[16],首先利用 SIFT 特征建立的视觉词项作为图像特征.接着,对于训练集所有图像来说,根据图像特征即可构建“视觉词汇-文档”矩阵 $\boldsymbol{V}$,对 $\boldsymbol{V}$ 做非负矩阵分解 $\boldsymbol{V}\approx\boldsymbol{W}\boldsymbol{H}$.由于 $\boldsymbol{H}$ 可视为样本在语义空间下的特征,则对于未标注图像的语义特征,可通过 $\boldsymbol{h}_{\text{query}}=\boldsymbol{W}^{-1}\boldsymbol{q}$ 求得,其中 $\boldsymbol{q}$ 为图像的原始特征.然后,通过 cos 距离计算,得到其在语义空间中的近邻图像.最后,将最近邻的图像标签扩散至未标注图像.

文献[26]提出了一种基于多模式非负矩阵分解的标注模型.假定图像训练集表示为 $\boldsymbol{X}=[\boldsymbol{X}_{\text{vision}}^{\text{T}}\ \boldsymbol{X}_{\text{text}}^{\text{T}}]^{\text{T}}$,其中 $\boldsymbol{X}_{\text{vision}}$ 和 $\boldsymbol{X}_{\text{text}}$ 分别表示利用词袋模型(BOW)提取出的视觉特征和词汇特征,对 $\boldsymbol{X}$ 做非负矩阵分解 $\boldsymbol{X}\approx\boldsymbol{W}\boldsymbol{H}$,可将 2 种模式映射到统一空间.对于测试集 $\boldsymbol{Y}_{\text{vision}}$,固定通过训练集得到的 $\boldsymbol{W}$,在保证非负的约束下求解 $\boldsymbol{H}_{\text{vision}}$,即可得到测试集在统一空间内的视觉特征表达.最后,通过 $\boldsymbol{H}_{\text{vision}}$ 和 $\boldsymbol{H}$ 的相似性计算求得近邻,并扩散标签.

文献[27]提出了一种松散的联合框架,同时对图像的视觉特征和词汇特征进行非负矩阵分解.如式(25)所示,统一空间中视觉模式和词汇模式之间的相似性越高,则其距离越近:

$$f=\min_{\boldsymbol{W}_1,\boldsymbol{W}_2,\boldsymbol{H}_1,\boldsymbol{H}_2,\lambda}[\alpha_1\|\boldsymbol{V}-\boldsymbol{W}_1\boldsymbol{H}_1\|_{\text{F}}^2+\alpha_2\|\boldsymbol{T}-\boldsymbol{W}_2\boldsymbol{H}_2\|_{\text{F}}^2+\alpha_3\exp(\gamma\|\Delta\boldsymbol{H}\|^2)],$$
$$\text{s.t. } \boldsymbol{W}_1,\boldsymbol{W}_2,\boldsymbol{H}_1,\boldsymbol{H}_2\geqslant\boldsymbol{0}, \tag{25}$$

其中,$\boldsymbol{V}$ 和 $\boldsymbol{T}$ 分别代表图像的视觉特征和词汇特征.$\alpha_1,\alpha_2,\alpha_3$ 分别代表正则化系数,$\Delta\boldsymbol{H}=\boldsymbol{H}_1-\lambda\boldsymbol{H}_2$ 则用来控制统一空间视觉模式和词汇模式之间的相似性.

文献[28]在文献[27]的基础上,进一步将所有的视觉特征种类包括词汇特征都视为一种视图(view).在同一框架中:1)对每种视图同时做非负矩阵分解;2)同时降低所有视图在统一空间的相互距离;3)对所有视图做拉普拉斯正则化操作.融合 3 种操作即可得到优化目标:

$$f_i=\min_{\boldsymbol{U}^{(i)},\boldsymbol{V}^{(i)},\boldsymbol{V}^*,\alpha}\sum_{i=1}^{nN+1}\{\alpha_i\|\boldsymbol{X}^{(i)}-\boldsymbol{U}^{(i)}(\boldsymbol{V}^{(i)})^{\text{T}}\|_{\text{F}}^2+\beta_i\|\boldsymbol{V}^{(i)}\boldsymbol{Q}^{(i)}-\boldsymbol{V}^*\|_{\text{F}}^2+\gamma_i\text{tr}((\boldsymbol{V}^{(i)})^{\text{T}}\boldsymbol{L}^{(i)}\boldsymbol{V}^{(i)})\}+\lambda\|\alpha\|_{\text{F}}^2,$$
$$\text{s.t. } \boldsymbol{U}^{(i)},\boldsymbol{V}^{(i)},\boldsymbol{V}^*\geqslant\boldsymbol{0},\alpha\geqslant 0,$$
$$\forall i\in\{1,2,\cdots,nN+1\}, \tag{26}$$

其中,n 为训练集样本个数,N 为视觉特征个数,+1 代表将词汇特征也作为一种视图,$\boldsymbol{Q}$ 用来控制在统一空间中模式之间的相似性,$\boldsymbol{L}$ 表示拉普拉斯正则化约束.

传统的 NMF 模型通过非负矩阵分解求得新的特征基之后,往往还需要额外的分类步骤对未标注图像进行标注,导致效率低下.文献[29]针对此问题,提出了一个可以同时进行矩阵分解和分类步骤的框架.具体来说,对图像的特征矩阵 $\boldsymbol{F}$ 进行非负矩阵分解 $\boldsymbol{F}\approx\boldsymbol{W}\boldsymbol{H}$,但有别于传统 NMF 模型,该模型将 $\boldsymbol{H}$ 视为图像的近似标签矩阵,用来决策某一图像是否属于标签 i.因此,$\boldsymbol{H}$ 应当与图像原始标签矩阵 $\boldsymbol{L}$ 具有一致性.据此可得优化目标式为

$$\min\left[\frac{1}{2}\|\boldsymbol{F}-\boldsymbol{W}\boldsymbol{H}\|_{\text{F}}^2+D(\boldsymbol{L},\boldsymbol{H})\right],$$
$$\text{s.t. } \boldsymbol{W},\boldsymbol{H}\geqslant\boldsymbol{0}, \tag{27}$$

其中 D 用来度量 $\boldsymbol{H}$ 和原始标签矩阵 $\boldsymbol{L}$ 的相似性.

文献[30]提出了一种具有结构化信息的 NMF 模型,其核心思想是具有相同标签的图像特征经分解过后应该处在相同的子空间中,也即分解过后的图像特征矩阵应当具有对角的块状结构.假定 $\boldsymbol{V}$ 为分解过后的特征矩阵,为了保持 $\boldsymbol{V}$ 的块状结构,模型定义了指示矩阵 $\boldsymbol{I}$:

$$\boldsymbol{I}=\begin{pmatrix}\bar{\boldsymbol{0}}_1 & 1 & \cdots & 1\\ 1 & \bar{\boldsymbol{0}}_2 & \cdots & 1\\ \vdots & \vdots & & \vdots\\ 1 & 1 & \cdots & \bar{\boldsymbol{0}}_C\end{pmatrix}, \tag{28}$$

其中 C 表示标签的个数,并以指示矩阵 $\boldsymbol{I}$ 定义最小化正则化项(式):

$$\Omega(\bar{\boldsymbol{V}})=\frac{1}{2}\|\boldsymbol{I}\odot\boldsymbol{V}\|_{\mathrm{F}}^2, \tag{29}$$

其中⊙表示矩阵元素的对应相乘运算.因此,模型总的优化目标可被定义为

$$\min_{\boldsymbol{U},\boldsymbol{V}}\|\boldsymbol{X}-\boldsymbol{U}\boldsymbol{V}\|_{2,\mathrm{p}}^{\mathrm{p}}+\frac{\mu}{2}\|\boldsymbol{I}\odot\boldsymbol{V}\|_{\mathrm{F}}^2,\quad \text{s.t. } \boldsymbol{U},\boldsymbol{V}\geqslant\boldsymbol{0}. \tag{30}$$

由于分解过后的特征矩阵 $\boldsymbol{V}$ 中每个块结构对应一个标签类别,因此特征具有较强的判别能力,更容易区分不同标签类别的图像.

文献[31-32]采用了多层次矩阵分解的方法建立图像视觉特征、标签特征以及其相关关系.文献[31]假定 $\boldsymbol{X}$ 表示图像的视觉特征矩阵,$\boldsymbol{F}$ 表示标签矩阵,则:

$$\begin{aligned}\boldsymbol{F}&\leftarrow\boldsymbol{V}\boldsymbol{U}_M,\\ \boldsymbol{U}_M&\leftarrow\boldsymbol{W}_M\boldsymbol{U}_{M-1},\\ &\vdots\\ \boldsymbol{U}_1&\leftarrow\boldsymbol{W}_1\boldsymbol{X}.\end{aligned} \tag{31}$$

通过多层次的矩阵分解,将 $\boldsymbol{F}$ 与 $\boldsymbol{X}$ 统一到一个潜在的语义空间,也即第 M 层.$\boldsymbol{W}_m(m=1,2,\cdots,M)$ 表示第 m 层的变换矩阵,$\boldsymbol{V}$ 表示潜在的标签特征矩阵,$\boldsymbol{U}_m$ 表示第 m 层的图像特征矩阵.其目标函数(式(32))类似于传统的 NMF 模型:

$$\min_{\boldsymbol{V},\boldsymbol{W}_M,\boldsymbol{W}_{M-1},\cdots,\boldsymbol{W}_1}\frac{1}{2}\|\boldsymbol{F}-\boldsymbol{V}\boldsymbol{U}\|_{\mathrm{F}}^2+\frac{\lambda_1}{2}\sum_{m=1}^{M}\|\boldsymbol{W}_m\|_{\mathrm{F}}^2+\frac{\lambda_2}{2}\|\boldsymbol{W}_1\|_{2,1}, \tag{32}$$

其中后 2 项分别代表正则化操作.文献[32]在文献[31]的基础上,进一步对 $\boldsymbol{U}_{M-1}$,$\boldsymbol{U}$,$\boldsymbol{V}$ 增加了拉普拉斯约束,如式(33)所示,使潜在语义空间中的特征矩阵以及标签矩阵具有更好的聚类特性.

$$\begin{aligned}\min_{\boldsymbol{V},\boldsymbol{W}_M,\boldsymbol{W}_{M-1},\cdots,\boldsymbol{W}_1}&\left\{\frac{1}{2}\|\boldsymbol{F}-\boldsymbol{V}\boldsymbol{U}\|_{\mathrm{F}}^2+\frac{\alpha}{2}\mathrm{tr}[\boldsymbol{U}_{M-1}^{\mathrm{T}}\boldsymbol{T}\boldsymbol{U}_{M-1}]+\right.\\ &\frac{\beta}{2}\mathrm{tr}[\boldsymbol{V}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{V}]+\frac{\mu}{2}\mathrm{tr}[\boldsymbol{U}\boldsymbol{M}\boldsymbol{U}^{\mathrm{T}}]+\\ &\left.\frac{\lambda_1}{2}\left(\|\boldsymbol{V}\|_{\mathrm{F}}^2+\sum_{m=1}^{M}\|\boldsymbol{W}_m\|_{\mathrm{F}}^2\right)+\frac{\lambda_2}{2}\|\boldsymbol{W}_1\|_{2,1}\right\},\\ &\text{s.t. } \boldsymbol{W}_m^{\mathrm{T}}\boldsymbol{W}_m=\boldsymbol{I}_{r_m},\end{aligned} \tag{33}$$

其中,r_m 表示第 m 层矩阵分解的节点个数.

此外,还有其他的矩阵分解模型如文献[33]提出了一种多模态的概率矩阵分解(probabilistic matrix factorization, PMF)模型用来建立“特征-标签”、“特征-特征”以及“标签-标签”之间的相关关系,然后通过分解得到的代表潜在语义空间的矩阵因子来连接 3 种相关关系.文献[34]采用弱监督的方法使隐语义空间中“标签相似矩阵”和“特征相似矩阵”与原始空间中的“标签相似矩阵”和“特征相似矩阵”相似度最小,文献[35]通过低秩矩阵分解来寻找更加本质和符合感官“特征-标签”矩阵.

矩阵分解模型如 NMF 可以很好地表征图像的局部特征,从直观上来说,图像总体是由各局部特征叠加而成,因此具有很好的解释性.此外,矩阵分解模型对于标签具有噪声和缺失的情况具有较高的鲁棒性.然而,如何选取合适的矩阵分解因子中的隐语义空间维度,以及如何有效、快速求解由矩阵分解转化而来的优化目标,仍然是矩阵分解模型面临的问题.

2.5 近邻模型

近邻模型在图像标注模型中思想比较简单,其基本原理是具有相似低层视觉特征的图像应该具有相似的语义.因此,利用近邻模型进行图像标注的一般步骤为:1)构建图像低层视觉特征;2)通过对低层视觉特征采用某种距离度量策略,选择与待标注图像距离较近的已标注图像;3)通过合适的标签扩散方法将已标注图像中的标签应用到待标注图像.相应地,对近邻模型的改进也基本集中在:1)选取或构造更适合用来标签扩散的视觉特征;2)采取更合适的距离度量策略,使图像集中视觉特征的距离远近符合语义空间上的距离远近;3)采用更加优化的标签扩散算法使标签能够良好扩散.近邻模型的代表模型有 JEC(joint equal contribution)模型[36]、TagProp(tag propagation)模型[37]、2PKNN(2-pass k-nearest neighbor)模型[38]、VS-KNN(visual semantic KNN)模型[39]、SNLWL(semantic neighborhood learning for weakly label)模型[40]、SEM(semantic extension model)模型[41]、weight-KNN(weight KNN)模型[42]、AWD-IKNN(adaptive weighted distance method based on improved KNN)模型[43]、NMF-KNN 模型[44]、AL(active learning)模型[45]等.

JEC 模型[36]是一种近邻模型的基线版本,其采用了颜色特征(RGB,HSV,LAB)和纹理特征(Gabor,Haar)作为图像的低层视觉特征,假定图像 i 被提取的 N 种特征表示为 $F_i=f_i^1,f_i^2,\cdots,f_i^N$,则 $d_{(i,j)}^k$

表示 2 幅图像 i,j 关于第 k 种特征的归一化距离,i 与 j 关于所有特征的距离为 $d_{(i,j)}=\sum_{k=1}^{N}\frac{d_{(i,j)}^{k}}{N}$. 求出距离待标注图像的经排序后最近的 K 个近邻$\{I_i|i=1,2,\cdots,K\}$之后,JEC 采取了一种贪婪模式将近邻图像的标签扩散到待标注图像 $\tilde{I}$.首先将 I_1 的标注词汇根据在训练集中出现的频率进行从大到小排序,然后选取 I_1 前 n 个标注词作为图像 $\tilde{I}$ 的标注(其中 n 为标签的需求个数),若图像 I_1 的标注词汇个数小于 n,则继续对 I_2 到 I_K 的标注词汇做频率排序,直到选取的标注词个数达到 n 个.JEC 模型作为近邻模型的基线,将所有图像底层视觉特征视为同等权值,未考虑不同种类图像特征对近邻计算的影响可能不一致的问题.此外,JEC 模型标签扩散采用的贪婪算法存在对 K 个邻居利用不均衡问题,即最近的邻居贡献过高.

TagProp 模型[37]为了解决近邻图像之间权值不均衡以及各种视觉特征之间权值不均衡的问题,提出了一种加权的近邻模型.即假定单词 w 出现在图像 i 的预测值 $p(y_{i,w}=+1)$(+1/−1 表示标签是否存在)可由所有近邻图像 j 对 i 的条件预测值加权得到:

$$p(y_{i,w}=+1)=\sum_{j}\pi_{i,j}p(y_{i,w}=+1|j),\quad(34)$$

从而进一步求得对数似然函数 L:

$$L=\sum_{i,w}^{trainset}c_{i,w}\ln p(y_{i,w}),\quad(35)$$

其中,$c_{i,w}$即为近邻图像间权值,用来解决某一标签对应的负样本远大于正样本的问题,也即对某一单词,不包含该单词的图像个数要远大于包含该单词的图像个数所带来的不均衡问题.其中,若 $y_{i,w}=+1$,则令 $c_{i,w}=1/n^{+}$;若 $y_{i,w}=-1$,则令 $c_{i,w}=1/n^{-}$,n^{+}和 n^{-}分别代表正负样本个数.对于各种视觉特征之间权值不均衡的问题,TagProp 模型定义 $\pi_{i,j}$ 为

$$\pi_{i,j}=\frac{\exp(-\boldsymbol{W}^{\mathrm{T}}\boldsymbol{d}_{i,j})}{\sum_{j'}\exp(-\boldsymbol{W}^{\mathrm{T}}\boldsymbol{d}_{i,j'})},\quad(36)$$

从而使得 $\pi_{i,j}$可以根据图像 j 依据到图像 i 的距离而衰减,$\boldsymbol{W}$ 表示每种视觉特征的权值向量,$\boldsymbol{d}_{i,j}$ 表示图像 i 和图像 j 各种视觉特征的基础距离所构成的向量.此外,为解决标签不平衡问题,也即使用频率较少的标签召回率低的问题,该模型利用 sigmoid 函数改进对图像 i 的预测:

$$p(y_{i,c}=+1)=\sigma(\alpha_w x_{i,w}+\beta_w),\quad(37)$$

使得相对稀少的标签权重增强,而减弱出现频率较高的标签权重.其中 sigmoid 函数 $\sigma(z)=(1+\exp(-z))^{-1}$.

2PKNN 模型[38,46]为解决标签的不平衡问题,采取了一种语义近邻组的方式来构建图像的近邻集合.假定将训练集中每个标签 y_i 看作为一个分类,则训练集 τ 可被划分为 l 个“标签-图像”对,即 $\tau=\{(\tau_1,y_1),(\tau_2,y_2),\cdots,(\tau_l,y_l)\}$,其中 l 代表标签个数,τ_i 表示标签 y_i 对应的图像集合,每一个集合 τ_i 称之为一个语义组.给定未标注图像 J,从每个语义组中选择 K_1 个视觉相似的近邻构成新的集合 τ_J.由此方式构建的近邻集合 τ_J,每个标签在其中至少会出现 K_1 次,因此解决了标签不平衡的问题.根据近邻集合 τ_J 可以定义未标注图像 J 对于标签 $y_k\in Y$ 的后验概率:

$$P(J|y_k)=\sum_{(I_i,Y_i)\in\tau_J}\exp(-D(J,I_i))\delta(y_k\in Y_i),\quad(38)$$

其中,如果标签 y_k 属于某一近邻图像 I_i 的标签集合 Y_i,则 $\delta(y_k\in Y_i)=1$,否则 $\delta(y_k\in Y_i)=0$.$\exp(-D(J,I_i))$表示基于近邻图像间视觉特征的距离对于后验概率的贡献.为了解决各种视觉特征之间以及特征向量内部的不均衡问题,2PKNN 为视觉特征距离 $D(A,B)$分别定义了距离空间权值 $\boldsymbol{w}$ 以及特征空间权值 $\boldsymbol{v}$:

$$D(A,B)=\sum_{i=1}^{n}w_i\sum_{j=1}^{N_i}v_{i,j}d_{i,j}(A,B),\quad(39)$$

n 为视觉特征的个数,N_i 为对应视觉特征的维度,$d_{i,j}(A,B)$表示对应视觉特征的基础距离.然后,根据组间最大、组内最小的原则定义 LOSS 函数,并采用随机梯度下降求解参数.在 $P(J|y_k)$求解之后,最后可根据贝叶斯公式求解后验概率:

$$P(y_k|J)=\frac{P(J|y_k)P(y_k)}{P(J)},\quad(40)$$

其中先验概率 $P(y_k)$和 $P(J)$可认为是定值.

VS-KNN 模型[39]为了更充分地利用标签间的语义相关信息以及解决弱标签问题,在 2PKNN 的基础上进一步扩展了语义近邻组的集合.类似 2PKNN 模型的第 1 步,其将训练集 τ 被划分为 l 个“标签-近邻图像-次近邻图像”组 $\tau=\{(\tau_1,\tau_1',y_1),(\tau_2,\tau_2',y_2),\cdots,(\tau_l,\tau_l',y_l)\}$,其中每个标签对应的次近邻图像集合 τ'可根据式(41)获得:

$$S(x_j,x_k)=\alpha D(x_j,x_k)+(1-\alpha)dis(x_j,x_k),\quad(41)$$

其中 x_j 表示某一标签的语义近邻组内的图片,x_k 表示该标签对应的语义近邻组外的图片,α 表示平滑

参数,D 和 dis 分别表示图像的视觉特征距离和标签距离.根据式(41)求得每一个近邻组内图片 K_2 个近邻组外的邻居,再求并集即可得到对应次近邻图像集合 τ'.

SNLWL 模型[40]构建语义近邻的方式类似于 2PKNN 模型[38,46],但是为了解决弱标签和标签不平衡问题,通过在构建语义近邻组之前求解其定义的学习误差函数(如式(42))来平衡训练集样本标签,进而使低频标签可以更有效地参与标注.

$$E=\min_f\Big\{\frac{1}{2}\sum_{j=1}^{q}\sum_{i=1}^{l}\mu_{i,j}(y_{i,j}-f_{i,j})^2+\frac{1}{2}\lambda\sum_{i=1}^{l}\sum_{k=1}^{l}w_{i,k}\left\|\frac{f_i}{\sqrt{d_i}}-\frac{f_k}{\sqrt{d_k}}\right\|^2\Big\},\tag{42}$$

其中,q,l 分别代表标签个数和训练集样本数量,y,f 分别代表真实以及待平衡的训练集样本标签.$\mu_{i,j}$ 表示样本 x_i 相对于标签 j,在 x_i 不包含标签 j 的时候取值为 $\tau(0\leqslant\tau\leqslant1)$,否则取值为 1.$w_{i,k}=\exp\{-\|x_i-x_k\|^2/(2\sigma^2)\}$,其中 σ 为样本平均距离,令 $d_i=\sum_{k=1}^{l}w_{i,k}$,$d_k=\sum_{i=1}^{l}w_{i,k}$,$\lambda$ 为平衡系数,通过求解 f 即可得到平衡后的训练集样本标签.

SEM 模型[41]和 weight-KNN 模型[42]针对传统模型需要手动设计视觉特征并且视觉特征性能泛化程度不高的问题,采用了从预训练的经典深度学习模型中抽取特征的方法来构建具有更泛化的视觉特征,此外,weight-KNN 模型为了解决视觉特征元素之间的不平衡问题以及充分利用标签之间的相关信息,采用了一种多标签线性判别(multiple linear discriminant analysis, MLDA)方法[47],在 KNN 的距离计算中给元素赋值适当的权重 $\boldsymbol{W}$.假定图像的每个标签为 1 个类别,则权重 $\boldsymbol{W}$ 需要满足条件使得类内距离最近、类间距离最远.因此,根据 MLDA 方法可定义目标函数为

$$\arg\max_{\boldsymbol{W}}\operatorname{tr}\left(\frac{\boldsymbol{W}^{\mathrm{H}}\boldsymbol{S}_{\mathrm{b}}\boldsymbol{W}}{\boldsymbol{W}^{\mathrm{H}}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{W}}\right).\tag{43}$$

其中,类内散布矩阵 $\boldsymbol{S}_{\mathrm{b}}$、类间散布矩阵 $\boldsymbol{S}_{\mathrm{w}}$ 以及总散布矩阵 $\boldsymbol{S}_{\mathrm{t}}$ 分别定义为

$$\begin{aligned}\boldsymbol{S}_{\mathrm{b}}&=\sum_{k=1}^{K}N^k(\boldsymbol{m}^k-\boldsymbol{m})(\boldsymbol{m}^k-\boldsymbol{m})^{\mathrm{H}},\\ \boldsymbol{S}_{\mathrm{w}}&=\sum_{k=1}^{K}\sum_{i=1}^{N^k}(\boldsymbol{f}_i-\boldsymbol{m}^k)(\boldsymbol{f}_i-\boldsymbol{m}^k)^{\mathrm{H}},\\ \boldsymbol{S}_{\mathrm{t}}&=\sum_{k=1}^{K}\sum_{i=1}^{N^k}(\boldsymbol{f}_i-\boldsymbol{m})(\boldsymbol{f}_i-\boldsymbol{m})^{\mathrm{H}},\end{aligned}\tag{44}$$

K 和 k 分别为标签个数和索引,N 和 N^k 分别为训练集图像数目和第 k 个标签对应的图像数目,$\boldsymbol{m}^k$ 和 $\boldsymbol{m}$ 分别代表第 k 个标签对应图像的特征向量均值和总的图像特征向量均值,$\boldsymbol{f}_i$ 表示第 i 幅图像的特征向量.对于标签之间相关信息的利用,weight-KNN 模型重定义了标签矩阵 $\boldsymbol{L}$:

$$\boldsymbol{L}=\boldsymbol{Y}\boldsymbol{C},\tag{45}$$

其中,$\boldsymbol{Y}$ 为原始标签矩阵,$\boldsymbol{C}\in\mathbb{R}^{K\times K}$ 代表标签之间的相关关系矩阵,元素 $C_{k,l}$ 表示任意第 k 个和第 l 个标签之间的相关性:

$$C_{k,l}=\cos(\boldsymbol{y}^k,\boldsymbol{y}^l)=\frac{\langle\boldsymbol{y}^k,\boldsymbol{y}^l\rangle}{\|\boldsymbol{y}^k\|_2\|\boldsymbol{y}^l\|_2}.\tag{46}$$

将包含有标签关系的新标签矩阵引入之后可得到新的散布矩阵为

$$\begin{aligned}\boldsymbol{S}_{\mathrm{b}}&=\widetilde{\boldsymbol{F}}\boldsymbol{T}\boldsymbol{D}^{-1}\boldsymbol{T}^{\mathrm{H}}\widetilde{\boldsymbol{F}}^{\mathrm{H}},\\ \boldsymbol{S}_{\mathrm{t}}&=\sum_{i=1}^{N}\Big((\boldsymbol{f}_i-\boldsymbol{m})(\boldsymbol{f}_i-\boldsymbol{m})^{\mathrm{H}}\sum_{k=1}^{K}\boldsymbol{T}_{i,k}\Big),\\ \boldsymbol{S}_{\mathrm{w}}&=\boldsymbol{S}_{\mathrm{t}}-\boldsymbol{S}_{\mathrm{b}},\end{aligned}\tag{47}$$

$\boldsymbol{T}$ 为 $\boldsymbol{L}$ 的行归一化变体,$\widetilde{\boldsymbol{F}}$ 为训练集特征矩阵 $\boldsymbol{F}$ 的中心化变体,$\boldsymbol{D}=\operatorname{diag}(\cdots,\sum_{i=1}^{N}\boldsymbol{T}_{i,k},\cdots)$,$\boldsymbol{D}\in\mathbb{R}^{K\times K}$.由于目标函数式(43)融合了标签相关关系矩阵 $\boldsymbol{C}$,因此求解该目标函数得到的权值 $\boldsymbol{W}$ 能有效解决特征向量元素之间的不平衡问题以及利用“标签-标签”之间的相关关系.

AWD-IKNN 模型[43]集成了图像的 CNN(convolutional neural networks)特征与传统视觉特征,并且针对标签的不平衡问题设计了一种加权模型(如式(48)):

$$D(i,j)=\sum_{k=1}^{N}\frac{\sum_{l\in\Omega_1}W(l,k)+\sum_{l\in\Omega_2}W(l,k)+1}{|\Omega_1|+|\Omega_2|+1}d_k(x_{i,k},x_{j,k}),\tag{48}$$

$D(i,j)$ 表示图像 i 和 j 之间的视觉特征距离,Ω_1 和 Ω_2 分别表示图像 i 和图像 j 对应的标签集合,$W(l,k)$ 表示在对应图像中,针对第 k 种特征,标签 l 对应的权值,N 为图像特征数目,$d_k(x_{i,k},x_{j,k})$ 为第 k 种特征的基础距离.由式(48)可知,如果 2 个图像之间存在共同的标签,则该标签的权重将被重复计算,也即表明这些标签的相应特征在图像之间的距离运算中权重更大.所有标签都会被同等对待即使某些标签出现频率较小,如果 2 个图像有相同标签,则 2 个图像之间的距离会更短,从而解决标签不

平衡的问题.根据正样本(具有相同标签的图像)间距最小、负样本间距最大的原则来定义 LOSS 函数:

$$L=\sum_{i=1}^{m}[D(i,j)^2-D(i,k)^2+\alpha],\quad x_j\in x_i^{\mathrm{p}},x_k\in x_i^{\mathrm{n}}, \tag{49}$$

权值 W 即可通过梯度下降法求解 LOSS 函数式(49)得到,其中 x_i^{p} 与 x_i^{n} 分别表示正样本与负样本集合.

NMF-KNN 模型[44]借鉴了 NMF 多视图聚类[48]的思想,将每一种视觉特征以及图像的标签特征视作为一种视图(view).训练集所有图像针对任意一种特征都可构成特征矩阵 $\boldsymbol{X}^{(f)}$,$\boldsymbol{X}^{(f)}$ 可通过 NMF 分解成为 1 组由基 $\boldsymbol{U}^{(f)}$ 和系数相关表达式 $\boldsymbol{V}^{(f)}$ 组成的矩阵.在 NMF 分解之后,系数表达式 $\boldsymbol{V}^{(f)}$ 应当与潜在的表达空间 $\boldsymbol{V}^*$ 有近似的语义.由此,可定义 LOSS 函数为

$$L=\sum_{f=1}^{F+1}\|\boldsymbol{T}(\boldsymbol{X}^{(f)}-\boldsymbol{U}^{(f)}\boldsymbol{V}'^{(f)})\boldsymbol{W}\|_{\mathrm{F}}^2+\sum_{f=1}^{F+1}\lambda_f\|\boldsymbol{W}'(\boldsymbol{V}^{(f)}\boldsymbol{Q}^{(f)}-\boldsymbol{V}^*)\|_{\mathrm{F}}^2,$$
$$\text{s.t. }\forall 1\leqslant f\leqslant F+1,\boldsymbol{U}^{(f)},\boldsymbol{V}^{(f)},\boldsymbol{V}^*\geqslant\boldsymbol{0}, \tag{50}$$

其中,$\boldsymbol{Q}^{(f)}=\mathrm{diag}\left(\sum_{m=1}^{M}\boldsymbol{U}_{m,1}^{(f)},\sum_{m=1}^{M}\boldsymbol{U}_{m,2}^{(f)},\cdots,\sum_{m=1}^{M}\boldsymbol{U}_{m,K}^{(f)}\right)$,用来限制 NMF 分解不唯一的情况发生.$F$ 为视觉特征类型个数,$F+1$ 表示将图像标签特征也视作为一种视图.此外,通过式(50),NMF-KNN 模型分别为标签不平衡问题以及图像的不平衡问题引入了权值 $\boldsymbol{T}$ 和 $\boldsymbol{W}$.$T_{i,i}$ 为第 i 个标签对应权值,被设置为 $1/frequency$,用来强制 $\boldsymbol{U}$ 可以准确地建模稀有标签;$W_{j,j}$ 为第 j 幅图像的近邻图像集合中所有标签的 $1/frequency$ 之和,其中 $frequency$ 表示标签出现次数.对于 LOSS 函数的求解,可采用文献[48]中模型进行求解.最后,对于待标注图像,可以先估计出其系数相关矩阵 $\widetilde{\boldsymbol{V}}$,然后根据其对应的 $\boldsymbol{U}\widetilde{\boldsymbol{V}}$ 值来将分值最高的标签作为其标注.

AL 模型[45]类似于 2PKNN 模型,仍然采取了语义近邻组的方法来避免标签不平衡问题;此外,与传统方法[36-38]选择固定数量的近邻数量有所不同,AL 模型采取了一种动态阈值的方法选取近邻图像.其中阈值定义为

$$\tau=S_m-\varepsilon, \tag{51}$$

其中,S_m 为图像视觉特征相似度均值,ε 为控制召回率大小的容忍值参数.

此外,还有其他的基于 KNN 的图像标注模型,如文献[49]定义了一个最优化的框架,分别集成了标签集合与图像之间的相关关系以及图像集合之间的相关关系 2 种因素,用以提高标注精度.文献[50]采用了 SIFT 以及 SURF(speeded up robust features)的图像局部特征进行单标签图像分类.文献[51]在 JEC 模型基础上增加了一种标签过滤机制,用以去除大部分图像中不相关标签,从而提高标注的精度.文献[52]进一步将近邻图像划分为强相关图像近邻与弱相关图像近邻,同时提出了基于范围约束视觉邻域的标签相关随机搜索方法,可以找出每个候选标签的可信部分,从而增强注释性能的鲁棒性.

近邻模型的基础思想相对比较简单,通常聚焦于解决图像标注的一般性问题(如弱标签、标签不平衡、特征的选取与权值赋值等),并且与其他方法的结合也比较灵活.由于其标注效果相对较好,近年来得到了广泛的研究和应用.近邻模型由于需要与训练集中的图片进行轮询对比,因此只适用于较小或中型的数据集,在面对超大规模数据集时往往耗时较长.

2.6 基于 SVM 的模型

SVM 是专门用于解决二分类问题的分类器,而图像的标注问题可被视为图像的多分类问题.因此,最基本的基于 SVM 的图像标注模型[53]是通过训练多个 SVM 分类器并采取一定的策略结合多个分类器的结果来完成分类.常用的 2 种策略包括“one-against-all”和“one-against-one”2 种,假定将每个标签作为 1 个类别,“one-against-all”策略是对每个标签训练 1 个“本标签相对于其他标签”的 SVM,每个 SVM 分类器都会定义 1 个判别函数 f_i 用来区分某一图像是否属于标签 i 或属于其他标签,多分类器的输出即为具有最大输出的判别函数 f_i 所对应的标签类别.最常用的判别函数可表示为

$$f_i(\boldsymbol{x})=\frac{\boldsymbol{w}_i\boldsymbol{x}+\boldsymbol{b}_i}{\|\boldsymbol{w}_i\|}, \tag{52}$$

其中,$\boldsymbol{x}$ 表示图像的视觉特征,$\boldsymbol{w}_i$ 和 $\boldsymbol{b}_i$ 分别表示为第 i 个标签所对应 SVM 分类器的权值和偏置.对待标注图像的标注为:

$$c(\boldsymbol{x})=\arg\max_{i\in\{1,2,\cdots,K\}}f_i(\boldsymbol{x}), \tag{53}$$

其中 K 表示标签个数.

“one-against-one”策略是对任意 2 个标签都训练 1 个 SVM.然后对未标注图像的分类结果进行投票,投票最多的标签即认为是图像标签.

文献[54]结合了多实例学习的方法,针对图像的“包特征(bag-features)”与全局特征分别构建

2 类 SVM,并针对 2 类 SVM 的分类结果进行融合得到最终的标注结果.假定依据“包特征”所构建 SVM 的输出作为概率向量表示为 $\boldsymbol{p}_m$,依据全局特征所构建 SVM 的输出作为概率向量表示为 $\boldsymbol{p}_g$,则融合2 类 SVM 的最终概率向量 $\boldsymbol{p}$ 可表示为

$$\boldsymbol{p}=\boldsymbol{w}\odot\boldsymbol{p}_m+(1-\boldsymbol{w})\odot\boldsymbol{p}_g, \tag{54}$$

其中⊙表示对应元素相乘,概率向量的长度与标签个数相同,概率向量 $\boldsymbol{p}_m$ 与 $\boldsymbol{p}_g$ 的每个元素表示图像属于某一标签的概率.$\boldsymbol{w}$ 表示对应的权值,可以通过“包特征”和“全局特征”分别对应的权值向量 $\boldsymbol{w}_m$ 与 $\boldsymbol{w}_g$ 进行估计:

$$\boldsymbol{w}=\frac{\boldsymbol{w}_m}{\boldsymbol{w}_m+\boldsymbol{w}_g}, \tag{55}$$

其中除法表示为对应元素相除.$\boldsymbol{w}_m$,$\boldsymbol{w}_g$ 中的元素可通过式(56)进行估计:

$$\begin{aligned}w_{m,k}&=\frac{\dfrac{1}{NK}\displaystyle\sum_{n=1}^{N}\sum_{c=1}^{K}L_m(n,c)}{\dfrac{1}{N}\displaystyle\sum_{n=1}^{N}L_m(n,k)},\\ w_{g,k}&=\frac{\dfrac{1}{NK}\displaystyle\sum_{n=1}^{N}\sum_{c=1}^{K}L_g(n,c)}{\dfrac{1}{N}\displaystyle\sum_{n=1}^{N}L_g(n,k)},\end{aligned} \tag{56}$$

其中 $L_m(n,c)$与 $L_g(n,c)$分别表示图像 n 依据“包特征”属于标签 c 的概率和依据“全局特征”属于标签 c 的概率.

文献[55]为改善传统 SVM 在训练过程中因约束过强导致过拟合的问题,提出了一种松弛的 SVM 分类方法,允许有少量样本不受约束;同时为了更好地利用标签与图像特征的一致性,在框架中增加了对标签特征和图像特征的拉普拉斯约束,如式(57)的后 2 项.其标注框架的总体目标函数可表示为:

$$\begin{aligned}\min\sum_{i=1}^{m}\frac{1}{2}\boldsymbol{err}_i^{\mathrm{T}}\boldsymbol{err}_i+\frac{1}{2}\mathrm{tr}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})+\\ \beta(\boldsymbol{ZLZ}^{\mathrm{T}})+\gamma\mathrm{tr}(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{HZ}),\end{aligned} \tag{57}$$

其中,$\boldsymbol{err}_i$ 向量由元素 $f_{i,q}-y_{i,q}$ 组成,q 表示在约束区域之外的少量样本点的索引,也即 $q=\{k\mid 1\leqslant k\leqslant n, f_{i,k}y_{i,k}<(1-\xi_{i,k})\}$,$f_{i,q}$表示映射函数 $f_{\boldsymbol{w}_i,\boldsymbol{b}_i}(\boldsymbol{x}_q)=\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{x}_q+\boldsymbol{b}_i$.

文献[56]为了增强用于标签传播的已知图像的相关性,采取了生成模型与 SVM 相结合的方式进行标注.具体来说,先通过生成模型找到具有最大生成概率的图像集合作为图像的近邻;接着,为了从近邻图像集合中寻找到具有更大相关性的图像子集,采用 SVD 分解得到 D 个隐藏主题,获知每幅图像从属于哪一个主题;然后,针对 D 个隐藏主题,训练 $\dfrac{D(D-1)}{2}$个 SVM 分类器,则针对未标注图像 I,则可根据投票策略找到其近邻图像集中具有最大相关性隐藏主题,属于该隐藏主题的近邻图像被认为是具有最大相关性的图像子集 S;最后,根据相关模型的方法求得未标注图像 I 的最大生成概率:

$$P(I,w)=\sum_{i=1}^{|S|}P(I\mid J_{S_i})P(w\mid J_{S_i})P(J_{S_i}). \tag{58}$$

此外,还有文献[57]对 SVM 的分类输出通过标签关系矩阵重新加权作为标注结果;文献[58]通过为生成模型和 SVM 分类输出设置不同的权重,用来作为标注结果.

由于 SVM 专门用于解决二分类问题,因此大多基于 SVM 的语义标注模型也继承了 SVM 的缺点,在标注词(也意味着分类个数)较多时,需要训练大量的分类器,因此模型在训练阶段的速度往往比较低下.此外,由于 SVM 模型本身的构建方式比较固定,对性能的提升主要集中在选取更加泛化和鲁棒的特征之上,而对图像标注的一般性问题如标签不平衡、弱标签、特征的权值赋值以及对“图像-图像(I-I)”、“标签-标签(W-W)”之间的关系利用等,难以融合其他方法进行有效解决.

2.7 图模型

图模型的基本思想是通过图来集成样本间的相似关系,包括样本间视觉特征之间的相似性、标签特征之间的相似性以及视觉特征和标签特征的对应关系,然后再利用相关的图论技术建立图结构中样本以及各种特征的关联模型,从而对标签进行预测.因此,大多基于图的标注模型的区别在于图的构建方式以及选择的图论技术存在差别.描述基于图的图像标注模型的代表文献有[59-72]等.

文献[59]首先利用分割算法将图像分块,提取每个图像块的视觉特征.构建图的时候,将图像的视觉特征和标签特征统一视为图的节点,并将这些节点分为 3 层.其中,作为视觉特征的节点为 1 层,图像节点作为 1 层,标签特征作为 1 层,视觉特征节点和标签特征节点作为图像节点的属性节点.将图像节点与其对应的视觉特征节点和标签特征节点通过边连接起来即可构成“图像-属性”的连接;再通过 KNN 方法对每个视觉特征的近邻通过边连接起来,构成“近邻连接”关系.据此,即可得到对应的关系图.对于待标注图像 I 节点来说,通过对建立的关系图采用重启随机游走算法(random walking with

restart, RWR),即可获得图中所有节点的稳定概率,具有最高稳定概率的标签特征节点,即可作为图像 I 的标注结果.文献[67]只是将节点分为 2 层,并按照图像和标签的对应关系进行双向连接,更加细致地定义了从标签到标签、标签到视觉特征、视觉特征到标签的 3 种转移概率.

文献[60,62,64]首先用基于图像特征的图学习方法得到每幅图像的备选标注,然后再采用基于标签特征的图学习方式来优化图像和标注之间的关联关系,进而得到最终的标注结果.其中,文献[60,64]的新颖之处在于,为了利用图像间的结构化信息,提出了一种最近生成链(nearest spanning chain, NSC)的图模型用来建立节点间的相似性关系.NSC 的构建原则为:1)N 个节点通过 $N-1$ 条边构成 1 条链,除了链的 2 个端点只有 1 条边相连,其余节点均有 2 条边相连;2)每个节点在构建链的时候,选择剩余节点中最近的节点作为相连节点.因此,若将图像视作节点构建 NSC 图,则通过统计建立的多条 NSC 的统计信息,即可得到图像 i 和图像 j 间的统计性描述关系 $C_{i,j}$:

$$\begin{aligned} C_{i,j} &= \sum_{n=1}^{N} seq\_w_{i,j}^{n} \times \delta_{i,j}^{n}, \\ seq\_w_{i,j}^{n} &= \exp\left(\frac{\lambda_1}{id(x_i)+id(x_j)}\right), \end{aligned} \tag{59}$$

其中,$seq\_w_{i,j}^{n}$ 表示图像 i 和 j 在第 n 条链中的排序权重,$id(x_i)$ 和 $id(x_j)$ 分别表示图像 i 和 j 在第 n 条链中的排序值,$\delta_{i,j}^{n}$ 表示图像 i 和 j 在链中如果直连,则值为 1,否则为 0.新的图像间相似性关系可表示为

$$\begin{aligned} W_{i,j}^{*} &= C_{i,j} W_{i,j}, \\ W_{i,j} &= \exp\left(-\frac{dis(x_i,x_j)}{\sigma^2}\right), \end{aligned} \tag{60}$$

其中 $W_{i,j}$ 表示原始的图像间相似关系,x_i 和 x_j 分别表示图像 i 和 j 的视觉特征.将 $\boldsymbol{W}$ 归一化为 $\boldsymbol{S}$ 之后,即可通过迭代式(61)直到收敛从而得到每幅图像的备选标注:

$$\begin{aligned} R(t+1) &= \alpha \boldsymbol{S} \cdot R(t) + (1-\alpha)\boldsymbol{Y}, \\ R(0) &= \boldsymbol{Y}, \end{aligned} \tag{61}$$

其中,$\boldsymbol{Y}$ 为原始标签矩阵,t 表示迭代次数.从标签关系的角度,为了使低频且具有更高判别性能的标签有更大的权值,可设计标签之间的统计性描述关系 K^{*} 为

$$K^{*}(x,y) = K(x,y) \lg \frac{N_T}{n_x}, \tag{62}$$

$K(x,y)$ 为原始的标签相似性矩阵 $\boldsymbol{K}$ 中的元素,表示标签 x 与标签 y 的相似性;n_x 表示标签 x 在训练集图像中出现的次数;N_T 表示训练集总数.类似地,再继续采用式(61)中的迭代方法,预测最终的标注结果,其中把式(61)中 $\boldsymbol{S}$ 用 $\boldsymbol{K}^{*}$ 代入,$\boldsymbol{Y}$ 用收敛后的图像备选标注代入即可.

文献[61]采用类似谱聚类的方式分别为图像和标签建立相似性关系图,然后再将 2 种相似性关系整合到一个统一的标注框架.以图像本身作为图的节点,图像 i 和 j 的相似度 $W_{i,j}$ 作为边($W_{i,j}$ 的定义见式(60)),根据相似的图像标签也应该类似的原则,即可定义二次能量函数 $E(f)$ 为

$$E(f) = \frac{1}{2}\sum_{i,j=1}^{n} W_{i,j} \| f_i - f_j \|^2, \tag{63}$$

其中 f 代表对图像的实值预测函数.同理,可定义标签的二次能量函数 $E'(\boldsymbol{g})$ 为

$$E'(\boldsymbol{g}) = \frac{1}{2}\sum_{i,j=1}^{k} W'_{i,j} \| \boldsymbol{g}_i - \boldsymbol{g}_j \|^2, \tag{64}$$

其中 $F=(f_1,f_2,\cdots,f_n)^{\mathrm{T}}=(\boldsymbol{g}_1,\boldsymbol{g}_2,\cdots,\boldsymbol{g}_k)$,$W'_{i,j}$ 表示标签 i 和 j 的 cosin 相似度.根据 2 种能量函数即可定义统一框架的目标函数为

$$\min \sum_{i=1}^{n} \| f_i - \boldsymbol{y}_j \|^2 + \mu E(f) + v E'(\boldsymbol{g}), \tag{65}$$

通过对式(65)的解 f 进行排序,即可得到未标注图像的标注.

文献[63]的模型原理与文献[61]类似,但是整合了图像的单实例和多实例 2 种特征描述.单实例特征也即图像的总体特征,多实例特征也即以标签作为包构建的多实例特征.由于多实例特征不易整合,因此文献提出了一种将多实例特征转换为单实例特征的方法.假定包含某一标签 i 的图像子集作为一个正包 B_i^{+},剩余图像作为负包 B_i^{-},则每一个正包中应当存在一幅图像 x^c 使得其与正包中其他图像的距离最小,同时与负包中图像的距离最大.寻找该图像的方法可归纳为

$$x^c = \arg\max_{x_p \in B_i^{+}} \prod_{i=1}^{l_+} P(x_p \mid B_i^{+}) \prod_{i=1}^{l_-} P(x_p \mid B_i^{-}), \tag{66}$$

其中:

$$\begin{aligned} P(x_p \mid B_i^{+}) &\propto \\ 1 - \prod_j \left(1 - \exp\left(-\frac{dis(x_{i,j}^{+},x_p)^2}{\sigma^2}\right)\right), \\ P(x_p \mid B_i^{-}) &\propto \\ \prod_j \left(1 - \exp\left(-\frac{dis(x_{i,j}^{-},x_p)^2}{\sigma^2}\right)\right). \end{aligned} \tag{67}$$

据此,任意多实例包 B_i 都通过映射得到一个单实例特征描述:

$$[s(x^1,B_i),s(x^2,B_i),\cdots,s(x^m,B_i)], \quad (68)$$

其中 m 代表标签个数,$s(x^c,B_i)=\min\limits_{x_{i,j}\in B_i}\{dis(x^c,x_{i,j})\}$.再分别对总体特征和通过多实例包映射得到的单实例特征依照文献[61]的方式构建相似关系图 G^{g} 和 G^{b},并将其统一到最终的目标函数式:

$$\min\frac{1}{2}[\alpha\sum_{i,j}W^{\mathrm{g}}_{i,j}(f_i-f_j)^2+(1-\alpha)W^{\mathrm{b}}_{i,j}(f_i-f_j)^2+\mu\sum_{i\in L}(f_i-\boldsymbol{y}_i)], \quad (69)$$

其中,$W^{\mathrm{g}}_{i,j}$ 和 $W^{\mathrm{b}}_{i,j}$ 分别表示归一化的单实例和多实例特征,f 表示标签的预测值函数,$\boldsymbol{y}$ 为标签的真实值.

文献[65]针对每个标签类别分别构建图,包含该标签的图像作为正节点,反之为负节点,其中正节点之间和负节点之间通过实线边进行连接,表示其具有标签一致性,正负节点用虚线连接,表示具有标签的非一致性.对于某一类标签,其目标就是构建具有最大数量实线边的图,使其具有最大一致性.考虑到节点之间的相似性 $W_{i,j}$,其目标函数可被定义为

$$\max\left(\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}W_{i,j}y_iy_j\right)=\max_{\boldsymbol{y}}(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{y}),$$
$$\text{s.t. } y_i=1,\ \forall i\in S_+,y_i=-1,\ \forall i\in S_-, \quad (70)$$

其中,S_+ 和 S_- 分别表示正节点和负节点.求解式(70)即可得到针对每一个标签的二分类分类器,之后再利用每个标签的分类器对未标注图像进行标注.

文献[66]针对每幅图像设计了一种基于 KNN 的稀疏图用来建立图像与近邻的潜在映射关系 $\boldsymbol{W}$,并假定该图像的标签与其近邻的标签具有相同的映射关系,然后通过这种映射关系来建立预测标签与真实标签的损失函数,从而对待标注图像进行标注.其中,$\boldsymbol{W}$ 可通过式(71)进行求解:

$$\min_{\boldsymbol{w}_i}\|\boldsymbol{w}_i\|_1,\quad \text{s.t. } \boldsymbol{x}_i=\boldsymbol{B}_i\boldsymbol{w}_i, \quad (71)$$

其中,$\boldsymbol{x}_i$ 表示图像 i 的视觉特征向量,$\boldsymbol{B}_i$ 表示由图像 i 的近邻所构成的特征矩阵,$\boldsymbol{w}_i$ 表示映射关系.由于标签具有相同的映射关系,因此有:

$$\min_{\boldsymbol{f}}\{\|\boldsymbol{f}-\boldsymbol{W}\boldsymbol{f}\|^2+\lambda_1\|\boldsymbol{f}_{\mathrm{L}}-\hat{\boldsymbol{f}_{\mathrm{L}}}\|^2+\lambda_2\|\hat{\boldsymbol{f}_{\mathrm{L}}}-\boldsymbol{y}\|_1\}, \quad (72)$$

其中,$\boldsymbol{f}_{\mathrm{L}}$ 代表图像经模型预测之后的标签,而 $\hat{\boldsymbol{f}_{\mathrm{L}}}$ 代表理想的标签,$\boldsymbol{y}$ 代表实际有噪声的标签.后 2 项用于对标签噪声的正则化.

此外,还有文献[68,70-71]同时利用了图像的视觉信息和标签信息,以及文献[72]同时利用了图像搜索引擎的返回结果和候选标注词在 Web 页面中的重要程度进行特征相似图的构建.文献[68]更进一步地在目标框架中嵌入了潜语义空间中的信息.文献[69]利用图和 KNN 相结合的方法对图像特征进行降维,使相关联的图像特征距离更近,而不相关的图像特征距离更远.

由于多数基于图的标注模型是通过与其他方法相结合的方式来融合各自的优点,因此对于解决一般性的图像标注问题(如特征的不平衡、特征选取、"图像-图像(I-I)"、"标签-标签(W-W)"之间的关系利用等)更加取决于与其结合的方法的特性,并且构建关系图的方式非常多样,使得基于图的标注模型具有相当的灵活性.

2.8 典型相关分析模型

典型相关分析(canonical correlation analysis, CCA)模型与 KCCA(kernel CCA)模型的本质是用来寻找 2 组特征变量的最大相关关系,最早被用于基于语义的图像检索领域[73-74],其基本思想为:假定图像的视觉特征与对应的标签特征分别为异构的 2 种特征,则 CCA 模型通过 2 组对应的基可将 2 种异构特征分别映射到一个具有最大相关性的可对比的隐藏语义空间,进而再通过适当的距离运算或比较模型,获得与图像最相关的标签.由于 KCCA 模型是在 CCA 模型基础上,通过核函数的方式增强了模型的非线性特征,其本质与 CCA 并无区别,因此,本文统一将 CCA 模型和 KCCA 模型都称之为 CCA 模型.基于 CCA 的图像标注模型代表有文献[75-81]等.

由于基于 CCA 的图像标注模型的基本思想类似,区别大多集中在特征的选择、距离运算以及比较模型方面.因此,本文先以文献[79]为例,介绍 CCA 模型方法,在此基础上进一步描述各模型的区别.

假定训练集中的 N 幅图像与其对应的标注词汇特征可表示为特征对$\{(\boldsymbol{v}_1,\boldsymbol{t}_1),(\boldsymbol{v}_2,\boldsymbol{t}_2),\cdots,(\boldsymbol{v}_N,\boldsymbol{t}_N)\}$,$\boldsymbol{v}_i\in\mathbb{R}^n$,$\boldsymbol{t}_i\in\mathbb{R}^m$,则 CCA 的目标是寻找到映射方向 $\boldsymbol{w}_v^*$ 和 $\boldsymbol{w}_t^*$,从而使得将特征 $\boldsymbol{v}$ 和特征 $\boldsymbol{t}$ 分别映射至 $\boldsymbol{w}_v^*$ 和 $\boldsymbol{w}_t^*$ 之后,相关系数能达到最大.$\boldsymbol{w}_v^*$ 和 $\boldsymbol{w}_t^*$ 分别计算为:

$$\boldsymbol{w}_v^*=\arg\max_{\boldsymbol{w}_v}\frac{\hat{E}[\langle\boldsymbol{v},\boldsymbol{w}_v\rangle\langle\boldsymbol{t},\boldsymbol{w}_t\rangle]}{\sqrt{\hat{E}[\langle\boldsymbol{v},\boldsymbol{w}_v\rangle^2]\hat{E}[\langle\boldsymbol{t},\boldsymbol{w}_t\rangle^2]}}=\arg\max_{\boldsymbol{w}_v}\frac{\boldsymbol{w}_v^{\mathrm{T}}\boldsymbol{C}_{v,t}\boldsymbol{w}_t}{\sqrt{\boldsymbol{w}_v^{\mathrm{T}}\boldsymbol{C}_{v,v}\boldsymbol{w}_v\boldsymbol{w}_t^{\mathrm{T}}\boldsymbol{C}_{t,t}\boldsymbol{w}_t}},$$

$$\boldsymbol{w}_t^* = \arg\max_{\boldsymbol{w}_t} \frac{\hat{E}[\langle \boldsymbol{v},\boldsymbol{w}_v\rangle\langle \boldsymbol{t},\boldsymbol{w}_t\rangle]}{\sqrt{\hat{E}[\langle \boldsymbol{v},\boldsymbol{w}_v\rangle^2]\hat{E}[\langle \boldsymbol{t},\boldsymbol{w}_t\rangle^2]}} = \arg\max_{\boldsymbol{w}_t} \frac{\boldsymbol{w}_v^{\mathrm{T}}\boldsymbol{C}_{v,t}\boldsymbol{w}_t}{\sqrt{\boldsymbol{w}_v^{\mathrm{T}}\boldsymbol{C}_{v,v}\boldsymbol{w}_v\boldsymbol{w}_t^{\mathrm{T}}\boldsymbol{C}_{t,t}\boldsymbol{w}_t}}, \tag{73}$$

其中,$\hat{E}$ 表示期望,$\boldsymbol{C}_{v,v}$ 和 $\boldsymbol{C}_{t,t}$ 分别表示特征 $\boldsymbol{v}$ 与特征 $\boldsymbol{t}$ 的协方差矩阵,$\boldsymbol{C}_{v,t}$ 表示 $\boldsymbol{v}$ 与 $\boldsymbol{t}$ 这 2 个特征之间的协方差矩阵.通过文献[82]中的方法,可将式(73)转化为广义特征值问题进行求解.$\boldsymbol{w}_v^*$ 和 $\boldsymbol{w}_t^*$ 可分别将图像的视觉特征以及标签特征映射至一个统一的隐藏空间,在该空间中 2 种异构特征具有最大的相关关系,因而也获取了可比较性.KCCA 为 CCA 模型的非线性版本,通过分别为图像的视觉特征和标签特征定义核函数来增加非线性关系.由此,定义最优映射方向向量 $\boldsymbol{\alpha}^*$ 和 $\boldsymbol{\beta}^*$ 分别为:

$$\boldsymbol{\alpha}^* = \arg\max_{\boldsymbol{\alpha}} \frac{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}_v\boldsymbol{K}_t\boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}_v^2\boldsymbol{\alpha}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{K}_t^2\boldsymbol{\beta}}},$$

$$\boldsymbol{\beta}^* = \arg\max_{\boldsymbol{\beta}} \frac{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}_v\boldsymbol{K}_t\boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}_v^2\boldsymbol{\alpha}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{K}_t^2\boldsymbol{\beta}}}, \tag{74}$$

其中 $\boldsymbol{\alpha}$ 和 $\boldsymbol{\beta}$ 分别为新的映射方向,$\boldsymbol{K}_v$ 和 $\boldsymbol{K}_t$ 分别表示对训练集 N 幅图片经由视觉特征和标签特征各自的核函数映射过后的内积矩阵.经正则化[82]之后,可转化为标准的特征向量问题:

$$(\boldsymbol{K}_v+k\boldsymbol{I})^{-1}\boldsymbol{K}_t(\boldsymbol{K}_t+k\boldsymbol{I})^{-1}\boldsymbol{K}_v\boldsymbol{\alpha}=\lambda^2\boldsymbol{\alpha}, \tag{75}$$

最大的 D 个特征值分别对应的特征向量组成的矩阵 $\boldsymbol{A}=(\boldsymbol{\alpha}^{(1)},\boldsymbol{\alpha}^{(2)},\cdots,\boldsymbol{\alpha}^{(D)})$ 和 $\boldsymbol{B}=(\boldsymbol{\beta}^{(1)},\boldsymbol{\beta}^{(2)},\cdots,\boldsymbol{\beta}^{(D)})$ 即可作为图像特征与标签特征的基矩阵,映射 2 种特征到具有最大相关性的隐藏空间.

文献[75]采用 SIFT 特征[17]构成的视觉词袋(BOW)作为图像的视觉特征,对图像的标签采用“词频-逆文档频率(term frequency inverse document frequency, TFIDF)”$P_{\{\mathrm{TFIDF}\}}$ 作为特征:

$$P_{\{\mathrm{TFIDF}\}}(d_i,w_j)=|\{w_j\in d_i\}|\lg(l\,|\,d_i\in D:w_j\in d_i\,|^{-1})|, \tag{76}$$

利用 TFIDF 作为标签特征是一种扩大在文档中经常出现但在整个集合中较少出现的单词影响的方法.其中 d_i 代表视作为文档的图像,w_j 代表标签,l 为训练集中总的图像数量.

文献[76]中提出的 PLSA+CCA 模型,首先利用 CCA 模型将图像视觉特征与标签特征映射至隐藏的语义空间,然后再通过 PLSA 模型[18]中的方法,建立起联合 2 种特征之间的联合概率密度.

文献[79]将 Gist、颜色直方图以及由 SIFT 特征构成的视觉词袋作为图像的视觉特征分别来表征图像的全局和局部特征.针对每种视觉特征的距离采用 χ^2 核函数增强其非线性关系:

$$K_{\chi^2}(h_i^f,h_j^f)=\exp\left(-\frac{1}{2A}\sum_{k=1}^{d}\frac{(h_i^f(k)-h_j^f(k))^2}{(h_i^f(k)+h_j^f(k))}\right), \tag{77}$$

其中,A 为训练集中所有特征的 χ^2 距离均值,d 为某种视觉特征 f 的维度,h 表示其视觉特征的表示函数.从而,针对 2 幅图像总的视觉特征距离,其核函数可定义为

$$K_v(I_i,I_j)=\frac{1}{|\mathcal{F}|}\sum_{f\in\mathcal{F}}K_{\chi^2}(h_i^f,h_j^f), \tag{78}$$

$\mathcal{F}$ 表示所有视觉特征.图像的标签特征表函数为 l,V 为标签总数.则可类似的为 2 幅图像 i,j 的标签特征距离定义核函数为

$$K_t(l_i,l_j)=\langle l_i,l_j\rangle=\sum_{k}^{V}l_i(k)l_j(k). \tag{79}$$

最后通过 KCCA 方法建立隐藏空间.对于待标注图像,则可通过图像在隐藏空间的视觉特征与标签特征的内积来寻找近邻标签并进行标签扩散.

文献[77]在文献[79]的视觉特征基础上,将训练图像中每幅图像的标签顺序进行建模,从而使被标注的图像能够更符合人类的感官.其基本依据是,人们在对图像进行标注时往往会先对图像中最显著的目标进行标注,从而使得每幅图像的标注词列表中带有目标的重要性信息,也即越显著的物体其标注词在列表中越靠前.标签的相对顺序与绝对顺序分别定义为

$$r_i=1-\frac{\sum_{k=1}^{J}w_{i,k}}{\sum_{k=1}^{N}w_{i,k}}, \tag{80}$$

$$\boldsymbol{A}=\left[\frac{1}{\mathrm{lb}(1+a_1)},\frac{1}{\mathrm{lb}(1+a_2)},\cdots,\frac{1}{\mathrm{lb}(1+a_V)}\right], \tag{81}$$

其中,$w_{i,k}$ 表示第 i 个标签在第 k 幅图像中出现的次序,N 为训练集图像总数,$J=\min(a_i,H)$,H 为选定的阈值,为了防止 $w_{i,k}$ 出现在超长列表中排序过于靠后的情况.a_i 代表第 i 个标签在所有图像中的平均排序.其对于核函数的建立与文献[79]保持一致.

文献[78]通过 K-means 算法对所有图像的标签特征先进行聚类,并假定聚类可得到 c 个具有语义相似性的簇.针对所有的 n 幅训练图像,可得到 $\boldsymbol{C}\in\mathbb{R}^{n\times c}$,其中 $C_{i,j}\in\{1,0\}$ 表示第 i 幅图像是否属于第 j 个聚类簇,每幅图像也就对应有了一个 c 维向量作为其语义特征.然后,类似于 CCA,建立“视觉特征-标签特征”、“视觉特征-语义特征”、“标签特征-语义特征”三者之间的最大相关关系:

$$\arg\min_{\boldsymbol{W}_1,\boldsymbol{W}_2,\boldsymbol{W}_3}\|\varphi_1(\boldsymbol{V})\boldsymbol{W}_1-\varphi_2(\boldsymbol{T})\boldsymbol{W}_2\|_{\mathrm{F}}^2+\|\varphi_1(\boldsymbol{V})\boldsymbol{W}_1-\varphi_3(\boldsymbol{C})\boldsymbol{W}_3\|_{\mathrm{F}}^2+\|\varphi_2(\boldsymbol{T})\boldsymbol{W}_2-\varphi_3(\boldsymbol{C})\boldsymbol{W}_3\|_{\mathrm{F}}^2, \tag{82}$$

其中 $\varphi_1(\boldsymbol{V})$,$\varphi_2(\boldsymbol{T})$,$\varphi_3(\boldsymbol{C})$分别代表图像的视觉特征、标签特征以及语义特征.

文献[80-81]都采用从预训练的 CNN 模型中提取的特征作为视觉特征,其中文献[80]选用了词向量(word2vec[83])作为标签特征.

CCA,KCCA 模型的构建模式比较固定,最关键的步骤在于寻找到映射 2 种异构特征到具有最大相关性的潜语义空间中的基.对于图像标注的一般性问题来说,由于与其他类型方法结合的灵活性也较差,使得可供解决这些问题的方法不够充分,从而也导致相关的研究工作不够充足.此外,由于解决 CCA,KCCA 模型的核心问题最终会演变为寻找矩阵的特征值和特征向量,因此在面对大规模数据集时,同样也会遭遇到运算效率低下的问题.

2.9 深度学习模型

近年来,由于硬件运算设备如 GPU,NPU(neural-network process unit)等运算性能的大幅提升,以深度学习为基础的模型克服了早期存在的运算瓶颈,并且在计算机视觉、文本处理、电子商务[84]等各应用领域,以高泛化性和优异的性能得到了广泛的应用和发展.深度学习模型通过若干层的卷积神经网络(CNN)、非线性激活函数和池化层相连接,直接建立从图像原始像素到图像标签的端到端关系映射.深度学习模型具有 2 个重要优势:1)与传统方法中手工设计的图像特征相比,通过预训练的深度学习模型提取出的图像特征具有更高的泛化性以及抽象性[85-88];2)通过基于深度学习的文本处理模型提取出的标签特征[83]具有高层的语义相关关系.

深度学习模型近年来成为计算机视觉领域研究热点,源于以 AlexNet(Alex net)模型[89]、VGG(visual geometry group)模型[90]、GoogLeNet(Google net)模型[91]、ResNet(residual network)模型[92]等著名深度学习模型在大规模图像分类竞赛 ILSVRC(image-net large scale visual recognition challenge)中的大放异彩.ILSVRC 竞赛任务为单一类别图像分类问题,而图像标注从分类角度可归类于图像的多分类问题,因此这些深度学习模型的模型结构大多适用于图像标注任务,区别仅在于对每一类别的预测得分选取方法有所不同.此外,各模型均基于前馈神经网络[93],差异大多集中在模型结构、激活函数、超参数调节等方面.因此,本文仅以 AlexNet 模型为例介绍深度模型结构,并在此基础上进一步描述各模型的区别.

AlexNet 模型[89]的输入为 RGB 图像,包括 5 个卷积层、3 个全连接层,最后通过 Softmax 函数输出预测,并通过预测结果与标签建立的损失函数对模型进行训练.其中,在第 1,2,5 卷积层之后使用了最大池化操作,避免了传统 CNN 模型使用平均池化所带来的模糊化效果.模型采用了 ReLU 作为每层卷积之后的非线性激活函数,从而减缓传统 CNN 模型采取 sigmoid 函数所面对网络较深时的梯度弥散问题,ReLU 函数定义为

$$f(\boldsymbol{x})=\max(0,\boldsymbol{x}). \tag{83}$$

同时,提出了局部响应归一化(local response normalization, LRN)层,用来抑制反馈较小的神经元,从而使相应比较大的神经元值变得相对更大,从而增加模型泛化能力.对在 ReLU 层第 i 个卷积核 (x,y)位置的 LRN 值定义为

$$b_{x,y}^{i}=\frac{a_{x,y}^{i}}{\left(k+\alpha\sum\limits_{j=\max\left(0,i-\frac{n}{2}\right)}^{\min\left(N-1,i+\frac{n}{2}\right)}(a_{x,y}^{j})^2\right)^{\beta}}, \tag{84}$$

其中,$a_{x,y}^{i}$表示第 i 个卷积核在(x,y)位置经 ReLU 函数作用后的输出值,n 表示为在同一空间位置选取的近邻数,N 为该层卷积核总数,k,α,β 为超参数.在防止模型过拟合的问题上,首先采用随机从原始图像裁切部分区域以及对训练集图像进行镜像操作,增强训练集数据;其次,采取在训练时使用 dropout(即随机忽略一部分神经元)的方法;最后,为处理深度学习训练时的大量矩阵运算,AlexNet 模型利用了 GPU 进行并行计算从而加速训练过程.

VGG 模型[90]相对 AlexNet 模型采用了更小的卷积核,也即 3×3 和 1×1 卷积核,而 AlexNet 模型中包括 11×11 以及 5×5 卷积核.采用小卷积核的原因是经试验验证,在卷积步长相同的情况下,不同卷积核大小对网络参数量差别影响不大,但大的卷积核会导致计算量增大,而大的卷积核所覆盖的感受野可通过小卷积核的叠加实现.同时,模型将卷积层增加到了 19 层,另有一个分支模型为 16 层,其中分支的 VGG16 模型中的若干层采用了 1×1 卷积核做线性映射.

GoogLeNet[91]采用了一种 inception 模型作为构建深度模型的基本模块,通过多个 inception 的模块串联形成最终的深度网络结构.inception 模块由 4 个分支组成,每个分支都包含一个 1×1 且步长

为1的卷积核,第2,3个分支分别在1×1卷积核后串联3×3以及5×5卷积核来增加感受野,第4个分支在1×1卷积核之前采用3×3最大池化层获取局部关键信息;最后,将4个分支的输出在维度上进行串联,作为下一个inception模块的输入.其中,采用4分支的并行结构依据为:首先在直观感觉上对多个尺度同时进行卷积,可提取到不同尺度的特征,使得最后分类判断更加准确;其次,根据Hebbian原则,4个分支分别代表4种具有高度聚类相关性的神经元,由于训练收敛的最终目的就是要提取出独立的特征,所以预先把相关性强的特征汇聚,就能起到加速收敛的作用.而在每个分支采用的1×1卷积核,其采用的依据为在相同尺寸的感受野中叠加更多的卷积,可提取到更丰富的特征[94],同时也降低了特征的维度,减少了计算复杂度.此外,由于每个卷积核后串联ReLU函数,因此还可增强模型的非线性关系.

ResNet模型[92]针对深度网络模型的层数加深达到一定数量之后模型准确率不升反降的问题,提出采用恒等映射(identity mapping)的方式使得网络加深的同时不会导致误差增加,也即通过"shortcut connections(捷径链接)"之间将输入 $\boldsymbol{x}$ 传到输出作为结果.假设某一段网络的输入为 $\boldsymbol{x}$,期望输出为 $H(\boldsymbol{x})=F(\boldsymbol{x})+\boldsymbol{x}$,则当网络已经学习到较饱和准确率时,$F(\boldsymbol{x})=0$,从而相当于 $H(\boldsymbol{x})=\boldsymbol{x}$,也即此时学习的目标为恒等映射.$F(\boldsymbol{x}):=H(\boldsymbol{x})-\boldsymbol{x}$ 即代表网络中的残差.ResNet通过引入残差的结构,打破了传统神经网络第 n 层的输出只能作为第 $n+1$ 层输入的惯例,使得某一层的输出可以直接跨过几层作为后面某一层的输入,在某种程度上解决了网络模型叠加多层之后精度不升反降的问题.

在深度学习的基础上,文献[95]针对社交网络图像的标注问题,有效结合了图像领域的元数据信息作为原始图像信息的补充.模型将图像的标签、图像上传时所属的集合以及图像的分组信息作为图像的元数据,然后利用这些元数据的相似性可获取某一图像的近邻集合.对图像及其近邻利用深度模型提取特征之后,采用最大池化的方式对所有特征进行融合,并将融合后的特征作为图像特征,最后训练模型.文献[96]类似于文献[95],也是利用了图像的拍摄时间和位置作为上下文信息.不同的是,文献[96]针对上下文信息建立了一个和原图像并行的网络,合并由原图像作为输入的深度网络输出结果和上下文信息作为输入的深度网络输出结果,作为总的网络结构的预测结果,从而训练深度学习模型,进而完成标注任务.

文献[97]将不同类型的视觉特征看作不同的视图(view),首先为不同的view建立深度模型作为encoder,并以此encoder输出的分布作为图像的标签特征,然后再利用图像本身结合标签特征作为输入,训练另外一个深度模型,从而得到最终的图像预测结果.

文献[98]为了建模提取标签之间的相关关系,将RNN(recurrent neural networks)模型与CNN模型相结合形成了一个统一的框架.其中,CNN用来提取图像的视觉信息并作为RNN模型在每一个节点的输入,从而使得RNN模型可提取到"图像-标签"、"标签-标签"之间的相互关系.文献[99]同样是为了利用标签之间的相关关系,但不同于文献[98],文献[99]将原始图像和标签特征统一到了一个深度学习框架中,让原始图像和标签特征统一作为模型的输入.

文献[100]首先利用分块算法将图像中包含有目标的区域划分为很多个候选块,由于多个候选块会对应至少一个标签,因此将这种对应关系转化为多实例学习问题,最后再利用多实例学习的模型框架来完成图像标注.文献[101]类似于文献[100],也是先将图像分解成包含有目标区域的多个候选块,多个候选块分别进入深度网络会形成多个预测向量,模型再将多个预测向量做最大池化操作,形成最终的多分类预测结果.

文献[102-103]分别在VGG网络[90]和GAN网络[104]的基础上引入了一种行列式点过程(DPP[105]).为了去除候选标签集合中的同义词等冗余标签,将最终的标注词的选定视为一种标签子集选择过程,在所有可能的标记中检索 K 个大小的标记子集作为最终的图像标注.文献[106]针对网络社交图像进行语义标注,利用具有相似视觉特征、相似主题、相似地标信息等元数据的图像构建近邻关系图,并引入到ResNet[92]网络,进一步提升视觉特征不明显的图像标注精度.文献[107]在VGG的训练过程中,加入了视觉特征一致性、标签关系一致性以及用户误标注的稀疏性作为每一批训练样本的限制,从而强化"图像-图像(I-I)"、"标签-标签(W-W)"关系.

由于传统的深度学习模型大多为针对特定领域大型数据集的分类任务而设计,在面对新领域视觉任务、小数据集或新数据集标签类别较少等特定情况下,重新设计或训练网络会造成网络过拟合、增加

运算负担等一系列问题.因此,有学者提出利用深度迁移学习的思想将某个领域或任务上学习到的知识或模式(如样本特征、网络的部分层次结构等)应用到不同领域或问题中,如文献[86,108-111]等.

文献[108]设计了一系列有 2 个不同但相关领域的分类任务 A 和 B,通过对网络结构不同层数迁移、微调、预训练参数等实验,表明了神经网络的前几层通常表示图像的通用特征.此外该工作与文献[86]的实验结果都可反映,经大型数据集(如 ImageNet)训练的网络结构,其网络参数可用来初始化其他未经训练的同构网络.这种优势使模型在面对相似领域但数据集不同的视觉任务(如图像标注)时,仅通过新的数据集对网络某些层进行微调(fine-tune)即可达到较好的效果,从而避免了重新训练的代价,是小数据集、标签数量少等问题的有效解决方案.例如文献[109]即采用了同 AlexNet 模型相同的网络结构,针对图像的标注问题,将最后的 Softmax 层输出进行排序,排序结果最大的 top N 个向量元素对应的标签作为图像的标注结果.

此外,实验还表明由预训练模型的全连接层提取出的向量,可作为其他视觉任务的中级特征,并且具有良好的泛化性能.如本文归类于近邻模型的 SEM 模型[41]和 weight-KNN 模型[42],以及归类于 CCA 模型的文献[80-81]都利用了从深度学习的预训练模型中提取图像特征的方法.

文献[110-111]更进一步地针对数据集的领域不同问题,利用处于 2 个不同领域之间的若干领域,将知识传递式地完成迁移.

基于深度学习的图像标注模型大多聚焦于对网络结构的变换以及利用深度模型提取“标签-标签(W-W)”之间相关关系.由于深度学习中的深层网络结构较之其他模型建立的相关关系具有更加复杂、非线性的特性,因此具有良好的标注性能.此外,从深度学习网络结构中提取出的图像特征也具有较强的泛化性能,因此近年来得到了广泛的应用和发展.但基于深度学习的模型也存在 4 方面不足:1)缺乏可解释性,即对于模型结构的调参以及如何使模型收敛缺乏理论指导依据,也是基于此原因,近年来的深度学习模型大部分集中于对模型结构的改进;2)传统深度模型需要较大的训练集,对于某些样本难以获取的图像领域任务来说代价及开销过大;3)模型过度依赖于硬件设备的性能如 GPU,NPU 等;4)现有的深度学习模型始终无法跳出人类设计的框架,无法自动化地生成具有高性能、高泛化性能的网络结构.

3 模型特点及对比

本文选取了一些有代表性的图像标注模型,针对其在公用数据集上的性能指标进行对比(如表 1 和表 2 所示,其中相关指标 P(precision),R(recall),$F1$(balanced score),N_+ 在本文 4.2 节中有详细说明),并将各种图像标注模型所采用的主要方法类型归为 9 种类型,分别为相关模型、隐 Markov 模型、主题模型、矩阵分解模型、近邻模型、基于支持向量机的模型、图模型、典型相关分析模型以及深度学习模型,通过表 3 概要列出了每种方法类型所对应的优缺点,其中,“相关文献(年份)”列表示本文所引用的各种类型相关文献公开发表的起始年份和该类模型活跃的平均年份.

Table 1 Comparison of Traditional Models(1)

表 1 经典模型指标对比(1)

Dataset&Model	Corel5k				ESP Game				IAPRTC-12			
	P/%	R/%	$F1$/%	N_+	P/%	R/%	$F1$/%	N_+	P/%	R/%	$F1$/%	N_+
TM[5]	6	4	4	49								
CMRM[6]	10	9	9	66								
CRM[7]	16	19	17	107								
MBRM[8]	24	25	24	122	18	19	18	209	24	23	23	223
HMM[9]	19	18	18	107								
TSVM-HMM[10]	27	35	30	156								
SMK[12]	29	33	31									
LSA[16]	7	9	8									
PLSA[18]	17	20	19									
PLSA-WORD[20]	14	20	16.5	105								
PLSA-FUSION[21]	19	22	20.4	112								

Continued (Table 1)

Dataset&Model	Corel5k				ESP Game				IAPRTC-12			
	P/%	*R*/%	*F*1/%	*N*+	*P*/%	*R*/%	*F*1/%	*N*+	*P*/%	*R*/%	*F*1/%	*N*+
LJNMF[27]	35.5	43	39.1		41	27	32.5					
MvNMF[28]	40.3	45.8	42.9	182	41.5	29.2	34.3	248	41.5	29.2	34.3	248
MPMF[33]	27	34	30.1	135								
JEC[36]	27	32	29.3	139	22	25	23.4	224	28	29	28.5	250
TagProp[37]	33	42	37	160	39	27	31.9	239	46	35	39.8	266
2PKNN[38]	39	40	39.5	177	51	23	31.7	245	49	32	38.7	274
2PKNN+ML[38]	44	46	45	191	53	27	35.8	252	54	37	43.9	278
VS-KNN[39]	39.8	41.8	40.7	187	32.7	33.3	33	255	44.6	37	40.5	278
SNLWL[40]	43	45	44	187	51	30	37.8	249	52	37	43.2	276
SEM[41]	37	52	43		38	42	40		41	39	40	
weight-KNN[42]	22	15	18		46	22	30		42	17	24	
NMF-KNN[44]	38	56	45.3	150	33	26	29.1	238				
AL+2PKNN[45]	36	34	35	67								
Label+Filter[51]	31	40	35	151	35	25	29.2	228				
TS[52]	32	35	33.4	149								
KSVMMN[55]					29.3	35	31.9					
GDM[56]	20.1	19.9	20									
SVM-DMBRM[58]	36	48	41	197	55	25	34	259	56	29	34.4	283
TGLM[64]	25	29	26.9	131								
GDR[69]	38.8	41.2	40									
KCCA[80]	39	53	44.9	184	30	36	32.7	252	38	39	38.5	273
KCCA+2PKNN[79]	42	46	43.9	179					59	30	39.8	259
CCA+KNN[80]	42	52	46	201	46	36	41	260	45	38	41	278
MVSAE[97]	37	47	42	175	47	28	34	246	43	38	40	283

Table 2 Comparison of Traditional Models(2)

表 2 经典模型指标对比(2)

Dataset&Model	NUS_WIDE				MSCOCO			
	P/%	*R*/%	*F*1/%	*N*+	*P*/%	*R*/%	*F*1/%	*N*+
KSVMMN[55]	84.2	61.1	70.8					
CNN+RNN[98]	41	31	35		66	56	61	
CNN+WARP[109]	32	36	34	97	53	60	56	

通过分析表 1 中的数据以及对比表 3 中各种模型的优缺点可知,早期的图像标注模型如相关模型、隐 Markov 模型等,有较强的可解释性,但这类模型比较依赖于其他算法,图像的标注性能和泛化能力往往一般;近些年取得广泛研究成果的矩阵分解模型、近邻模型、CCA(KCCA)模型等,其泛化性能有所提高,但可解释性有所降低;深度学习模型可以处理超大规模的数据集,性能相比其他模型也更高,但可解释性最差;在如何与其他的计算机方法相结合来解决标注的一般性问题,以及对"I-I","W-W"关系的利用程度等方面,每一类模型各有不同;此外,由于图像标注模型存在的一些固有特征(如使用的数据集过大、优化求解算法复杂等),几乎所有的模型运算速度都有待提升.

从各类模型出现的相关年份可反映出,标注模型的性能随着时间往越来越高的方向演进,所能处理的数据集也由早期的小型逐渐过渡到大型与超大型,深度学习模型以其优异的标注性能成为近年来标注模型的主流;同时,近邻模型以及矩阵分解模型等也有相当的发展.不难理解,造成这种趋势的原因

是随着时代的发展,硬件处理、运算能力不断提升,以及大型超大型数据集的出现对深度学习等依赖于运算能力和数据集的模型有显著的促进作用,而其他类型标注模型根据其自身特点也可被运用于一些仅需要处理中小型数据集、运算需求较小的应用场景.

Table 3 Comparison of Advantages and Disadvantages of Different Type Models

表 3 各方法类型的标注模型优缺点对比

Model Types & Characteristics	Interpretability	Performance	Speed	Scale of Adapted Dataset	Dependence on Previous Steps	Flexibility in Combination with Other Methods	Generalization	Utilization of "I-I", "W-W"	Relevant Literature	
									Start Year	Active Year
Relevance Model	Very High	Very Low	Very Low	Small	High	Very Low	Very Low	Very Low	2002	2003
Hidden Markov Model	Very High	Low	Very High	Small	High	Very Low	Very Low	Very Low	2005	2008
Topic Model	High	Medium	Very Low	Small	None	Low	Low	Very High	2003	2007
Matrix Decomposition Model	High	High	Low	Small and Medium	None	Medium	High	Very High	2008	2014
Neighborhood Model	High	High	Medium	Small and Medium	None	Very High	High	Very High	2010	2015
SVM-based Model	Very High	Medium	Low	Small and Medium	None	Very Low	Low	Low	2007	2013
Graph Model	Medium	Medium	Depend on Graph Structure	Small and Medium	None	Medium	Low	High	2004	2010
CCA(KCCA) Model	High	High	Low	Small and Medium	None	Medium	High	Medium	2006	2013
Deep Learning Model	Low	Very High	Very High in Training Phase	Big and Super Big	None	Low	Very High	High	2013	2016

4 图像标注数据集与评测指标

图像数据集常用来对各种任务模型进行训练,并可以结合各种评测指标对不同的模型进行对比.由于任务的不同,图像数据集的具体针对情况可能会存在偏差,但多数都可用于图像标注任务.因此,本文对已有标注工作中常用的部分图像数据集以及各种标注模型评测指标进行了归纳总结.

4.1 图像标注数据集

1) Corel5K.Corel5K 数据集包含科雷尔(Corel)公司收集整理的 5 000 幅图片,因此命名为 Corel5K.自从第 1 次被提出用于图像目标识别[5]实验后,已经成为图像实验的标准数据集,被广泛应用于标注算法性能的比较,也常用于图像的分类、检索等任务.图像集涵盖 50 个语义主题,如公共汽车、恐龙、海滩等,每个主题包含 100 张大小相等的图像,可以转换成多种格式.5 000 幅图片常被分为 3 个部分,其中训练集包含 4 000 幅,验证集和测试集各 500 幅.集合共包含 260 个标签,每幅图像包含 1～5 个标签.

2) ESP-Game.ESP-Game 数据集[112]源自一款双人标注游戏,游戏的内容是让 2 个人在不进行交流的情况下对同一图像进行标注,然后选取相同的词作为图像的标注.常用作图像标注的数据集共包含 20 770 幅图像和 268 个标签,其中训练集包含 18 689 幅图像,测试集包含 2 081 幅图像,每幅图像包含 3～5 个标签.

3) IAPR TC-12.IAPR TC-12 数据集[113]包含 19 627 幅从世界各地拍摄的自然图像,其中包括各种语义场景,如体育运动、人物、动物、城市、海滩等.训练集图像为 17 665 幅,测试集图像为 1 962 幅,并包含 291 个语义标签.

4) NUS-WIDE.NUS-WIDE 数据集[114]是由新加坡国立大学 NUS 实验室收集整理的网络图像数据集,图像内容侧重于人们的日常生活和事件.数据集被划分为 3 个集合:第 1 个集合包含 81 个从 Flickr 网站中获取到的基本标签,包括通用标签(如动物、

植物)和特殊标签(如狗、花),而且标签大部分是由高校学生提供,因此标签噪声相对其他从网络上搜集的图像要少;第 2 个和第 3 个集合中的图像也来自于 Flickr 网站,分别包含 1 000 和 5 000 个原始标签.图像集共包含 269 648 幅图像,训练集和测试集根据任务需求可自行设置.

5) MS-COCO.MS-COCO 数据集[115]是微软团队发布的一个可以用于图像识别、分割和标注的多任务用途数据集,其内容主要从复杂的日常场景中截取.数据集共包含 91 个类别,平均每张图片包含 3.5 个类别和 7.7 个实例目标,仅有不到 20%的图片只包含 1 个类别,仅有 10%的图片包含 1 个实例目标,也即每一类所包含的图像较多,有利于获得更多的每类中位于某种特定场景的能力.MS-COCO 数据集分 2 部分发布,前部分发布于 2014 年,后部分发布于 2015 年,其中 2014 年的版本中包含 82 783 训练集图像、40 504 验证集图像和 40 775 测试集图像,还有 27 万的被分割出来的人物目标和 88.6 万的实物目标;2015 年的版本其训练集、验证集和测试集图像数量分别为 165 482,81 208,81 434.

6) PASCAL VOC.PASCAL VOC 数据集是由每年一度的 PASCAL VOC 挑战赛所发布的数据集,其主要任务包括图像分类、目标识别、目标分割、人物定位以及行为识别等.以 PASCAL VOC 2012 为例,数据集总共分 4 个大类 20 个小类.数据集共包含 23 080 幅图像,其中训练集图像数量为 5 717 幅,验证集图像数量为 5 823 幅,测试集图像数量为 11 540 幅,共包含 54 900 个目标.

7) ImageNet.ImageNet 图像数据集始于 2009 年文献[116]中的计算机视觉系统识别项目工作,而后从 2010 年开始每年基于 ImageNet 数据集会举办大规模视觉识别挑战赛,到 2017 年后截止.比赛项目包括:图像分类、目标定位、目标检测、视频目标检测、场景分类、场景解析等.ImageNet 数据集共包含 14 197 122 幅图像,常用于竞赛 ISLVRC 使用的公开数据集是 ImageNet 的子集,以 2012 年 ILSVRC 分类数据集为例,其训练集为 1 281 167 张图像,验证集为 50 000 张图像,测试集为 100 000 张图像在每次竞赛中单独发布,共包含 1 000 个不同的类别.

此外,还有其他如 MIRFlickr,Flickr25k,Flickr30k 等从 Flickr 网站获取的带有描述性标题的图像集合,也常被用于图像的标注任务.

4.2 常用评测指标

图像标注算法的评测指标常用来检验各种模型的优劣,其中包括精确度 P、召回率 R、$F1$、N_+、平均精度(average precision, AP)、平均准确率(mean average precision, MAP)等,下文将对这些指标进行简单介绍.

精确度用来表明预测为正的样本中有多少是真正的正样本,它表示的结果为:某一标签预测正确的图像数量对该标签的总预测图像数量的占比.P 值可表示为:

$$P=\frac{1}{M}\sum_{k=1}^{M}\frac{Correct(w^k)}{Predicted(w^k)}. \tag{85}$$

召回率用来表明样本中的正例有多少被正确预测,它表示的结果为:某一标签预测正确的图像数量对该标签的总的真实图像数量的占比.R 值可表示为:

$$R=\frac{1}{M}\sum_{k=1}^{M}\frac{Correct(w^k)}{GroundTruth(w^k)}. \tag{86}$$

其中,$Correct(w^k)$表示针对第 k 个标签预测正确的图像数量,$Predicted(w^k)$表示针对第 k 个标签预测图像的总数量,$GroundTruth(w^k)$表示所有标注有第 k 个标签的总的图像数量.

在判断模型优劣的时候往往希望 P,R 值都是越高越好,但这两者之间有时会存在矛盾.$F1$ 可以综合两者之间的关系,也即对两者进行调和:

$$F1=\frac{2PR}{P+R}. \tag{87}$$

$F1$ 综合了 P,R 两者的结果,当 $F1$ 的值较高时说明模型方法比较理想.

N_+值表示至少被正确标注过一次的标签数量.对于具有标签不平衡问题的测试数据,该指标具有很大参考意义.

P,R,$F1$,N_+即为图像标注最常用的评测指标.此外,还有相对复杂的指标如 AP 和 MAP,也是综合考虑了 P,R 的值.AP 的计算方法相对复杂,该指标表示针对某一标签,在一组设定的 R 阈值之下,每个 R 值对应最大的 P 值的平均值.MAP 指标表示对所有标签 AP 求均值.

5 总 结

本文通过对近年来公开发表的图像标注文献的研究,总结了图像标注模型的一般性框架,并通过该框架结合各种具体工作,分析出在图像标注问题中需要解决的一般性问题.此外,在对各种标注模型的归类方面,本文通过其主要使用的方法类型对各种模型进行了归类.首先介绍了每种方法类型的基本

原理,然后具体分析了各种图像标注模型之间的差异,最后简单总结了每一类方法类型的标注模型,并结合了图像标注模型常用的一些数据集、评测指标以及比较著名的标注模型的性能和实验数据,对各种方法类型的标注模型做了优缺点分析.

总而言之,图像标注技术仍然是一个广泛、开放且具有挑战性的研究领域,其最主要的目标仍然是缩短图像的高级语义信息同低层视觉特征之间的语义鸿沟问题.本文结合标注领域的一般性问题以及各种方法类型的标注模型,有针对性地提出 5 个改善图像标注性能的方向:

1) 由于近年来社交网站的快速发展,用于图像标注领域的数据集往往是由不同的用户进行标注,标签里面夹杂着大量不相关或者错误的标注词汇.这些主观性的问题会导致生成的数据集产生标签不平衡、弱标签等问题.因此,如何采取自动化的方法剔除不相关标签仍然是标注模型亟待解决的问题之一.

2) 一些专业或特殊领域的图像集(如医学图像数据集、商品图像数据集等)拥有很多复杂特性,如:①图像中需要标注的目标之间具有环绕、遮挡情况;②关注的目标是细微的物品,在图像占比太小,不够显著;③关注的目标与图像中其他目标非常相似.因此,在特定的应用场景下还需要解决这类特定问题.

3) 图像的标注模型性能和速度往往会受到所采用的图像视觉特征的干扰,因此,如何使得选取的特征具有强大的泛化性能且低维是需要改进的方向.

4) 充分利用"图像-图像"、"标签-标签"之间的关联信息有助于提升标注的性能,而高层的语义关联信息在其他研究领域如自然语言处理(NLP)中也得到了广泛研究.因此,如何结合其他领域的研究方法也可作为提升图像标注性能的方向之一.

5) 由于近年来深度学习模型在图像标注领域表现出了良好的性能但可解释性低的特性,而较低的可解释性意味着深度学习模型的架构往往存在可复现及调参的问题.因此,深度学习模型在图像标注领域中的模型构建依据以及调参技巧也是需要研究的课题.此外,由于迁移学习表现出的可将已学习知识应用到新领域的强大性能,研发新的迁移学习理论和模型方法也成为改善图像标注性能的新思路.

参 考 文 献

[1] Datta R, Joshi D, Li Jia, et al. Image retrieval: Ideas, influences, and trends of the new age [J]. ACM Computing Surveys, 2008, 40(2): 35-94

[2] Long Fuhui, Zhang Hongjiang, Feng D D. Fundamentals of Content-based Image Retrieval [M]. Berlin: Springer, 2003

[3] Manaf S, Nordin M. Review on statistical approaches for automatic image annotation [C] //Proc of 2009 Int Conf on Electrical Engineering and Informatics. Piscataway, NJ: IEEE, 2009: 56-61

[4] Cheng Qimin, Zhang Qian, Fu Peng, et al. A survey and analysis on automatic image annotation [J]. Pattern Recognition, 2018, 79(1): 242-259

[5] Duygulu P, Barnard K, De-freitas J, et al. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary [C] //Proc of European Conf on Computer Vision. Berlin: Springer, 2002: 97-112

[6] Jeon J, Lavrenko V, Manmatha R. Automatic image annotation and retrieval using cross-media relevance models [C] //Proc of the 26th Annual Int ACM SIGIR Conf on Research and Development in Informaion Retrieval. New York: ACM, 2003: 119-126

[7] Lavrenko V, Manmatha R, Jeon J. A model for learning the semantics of pictures [C] //Proc of Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2003: 553-560

[8] Feng S, Manmatha R, Lavrenko V. Multiple Bernoulli relevance models for image and video annotation [C] //Proc of the 2004 IEEE Computer Society Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2004: 1002-1009

[9] Ghoshal A, Ircing P, Khudanpur S. Hidden Markov models for automatic annotation and content-based retrieval of images and video [C] //Proc of the 28th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2005: 544-551

[10] Zhao Yufeng, Zhao Yao, Zhu Zhenfeng. TSVM-HMM: Transductive SVM based hidden Markov model for automatic image annotation [J]. Expert Systems with Applications, 2009, 36(6): 9813-9818

[11] Yu Feiyang, Ip H S. Automatic semantic annotation of images using spatial hidden Markov model [C] //Proc of 2006 IEEE Int Conf on Multimedia and Expo. Piscataway, NJ: IEEE, 2006: 305-308

[12] Lu Zhiwu, Ip H S. Spatial Markov kernels for image categorization and annotation [J]. IEEE Transactions on Systems Man and Cybernetics, 2011, 41(4): 976-989

[13] Lei Yinjie, Wong W, Liu Wei, et al. An HMM-SVM-based automatic image annotation approach [C] //Proc of Asian Conf on Computer Vision. Berlin: Springer, 2010: 115-126

[14] Landauer T, Foltz P, Laham D. Introduction to latent semantic indexing [J]. Discourse Processes, 1998, 25(5): 259-284

[15] Monay F, Gatica-perez D. On image auto-annotation with latent space models [C] //Proc of the 11th ACM Int Conf Multimedia. New York: ACM, 2003: 275-278

[16] Pham T, Maillot N, Lim J, et al. Latent semantic fusion model for image retrieval and annotation [C] //Proc of the 16th ACM Conf on Information and Knowledge Management. New York: ACM, 2007: 439-444

[17] Lowe D. Object recognition from local scale-invariant features [C] //Proc of Int Conf on Computer Vision. Piscataway, NJ: IEEE, 1999: 1150-1157

[18] Hofmann T. Unsupervised learning by probabilistic latent semantic analysis [J]. Machine Learning, 2001, 42(1/2): 177-196

[19] Monay F, Gatica-perez D. PLSA-based image auto-annotation: Constraining the latent space [C] //Proc of the 12th Annual ACM Int Conf on Multimedia. New York: ACM, 2004: 348-351

[20] Monay F, Gatica-perez D. Modeling semantic aspects for cross-media image indexing [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(10): 1802-1817

[21] Li Zhixin, Shi Zhiping, Li Zhiqing, et al. Automatic image annotation by fusing semantic topics [J]. Journal of Software, 2011, 22(4): 801-812 (in Chinese)
(李志欣, 施智平, 李志清, 等. 融合语义主题的图像自动标注 [J]. 软件学报, 2011, 22(4): 801-812)

[22] Lienhart R, Romberg S, Hörster E. Multilayer plsa for multimodal image retrieval [C] //Proc of ACM Int Conf on Image and Video Retrieval. New York: ACM, 2009: 1-8

[23] Zhang Rui, Zhang Lei, Wang Xinjing, et al. Multi-feature plsa for combining visual features in image annotation [C] //Proc of the 19th ACM Int Conf on Multimedia. New York: ACM, 2011: 1513-1516

[24] Lee D, Seung H. Learning the parts of objects by non-negative matrix factorization [J]. Nature, 1999, 401(6755): 788-791

[25] Tang Jiayu, Lewis P H. Non-negative matrix factorisation for object class discovery and image auto-annotation [C] //Proc of the 2008 Int Conf on Content-based Image and Video Retrieval. New York: ACM, 2008: 105-112

[26] Benabdallah J, Caicedo J, Gonzalez F, et al. Multimodal image annotation using non-negative matrix factorization [C] //Proc of 2010 IEEE/WIC/ACM Int Conf on Web Intelligence and Intelligent Agent Technology. Piscataway, NJ: IEEE, 2010: 128-135

[27] Rad R, Jamzad M. Automatic image annotation by a loosely joint non-negative matrix factorisation [J]. IET Computer Vision, 2015, 9(6): 806-813

[28] Rad R, Jamzad M. Image annotation using multi-view non-negative matrix factorization with different number of basis vectors [J]. Journal of Visual Communication and Image Representation, 2017, 46: 1-12

[29] Jia Xu, Sun Fuming, Li Haojie, et al. Image multi-label annotation based on supervised nonnegative matrix factorization with new matching measurement [J]. Neurocomputing, 2017, 219(1): 518-525

[30] Li Zechao, Tang Jinhui, He Xiaofei. Robust structured nonnegative matrix factorization for image representation [J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(5): 1947-1960

[31] Li Zechao, Tang Jinhui. Deep matrix factorization for social image tag refinement and assignment [C] //Proc of the 17th IEEE Int Workshop on Multimedia Signal Processing. Piscataway, NJ: IEEE, 2015: 1-6

[32] Li Zechao, Tang Jinhui. Weakly supervised deep matrix factorization for social image understanding [J]. IEEE Transactions on Image Processing, 2017, 26(1): 276-288

[33] Li Zechao, Liu Jing, Zhu Xiaobin, et al. Image annotation using multi-correlation probabilistic matrix factorization [C] //Proc of the 18th ACM Int Conf on Multimedia. New York: ACM, 2010: 1187-1119

[34] Li Zechao, Tang Jinhui. Weakly supervised deep metric learning for community-contributed image retrieval [J]. IEEE Transactions on Multimedia, 2015, 17(11): 1989-1999

[35] Li Zechao, Tang Jinhui. Weakly-supervised deep nonnegative low-rank model for social image tag refinement and assignment [C] //Proc of the 31st AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2017: 4154-4160

[36] Makadia A, Pavlovic V, Kumar S. A new baselines for image annotation [C] //Proc of Int Journal of Computer Vision. Berlin: Springer, 2010: 88-105

[37] Guillaumin M, Mensink T, Verbeek J, et al. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation [C] //Proc of the 12th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2010: 309-316

[38] Verma Y, Jawahar C. Image annotation using metric learning in semantic neighbourhoods [C] //Proc of the European Conf on Computer Vision. Berlin: Springer, 2012: 836-849

[39] Ji Qian, Zhang Liyan, Li Zechao. KNN-based image annotation by collectively mining visual and semantic similarities [J]. KSII Transactions on Internet and Information Systems, 2017, 11(9): 4476-4490

[40] Tian Feng, Shen Xukun. Image annotation by semantic neighborhood learning from weakly labeled dataset [J]. Journal of Computer Research and Development, 2014, 51(8): 1821-1832 (in Chinese)
(田枫, 沈旭昆. 弱标签环境下基于语义邻域学习的图像标注 [J]. 计算机研究与发展, 2014, 51(8): 1821-1832)

[41] Ma Yanchun, Liu Yongjian, Xie Qing, et al. CNN-feature based automatic image annotation method [J]. Multimedia Tools and Applications, 2019, 78(3): 3767-3780

[42] Ma Yanchun, Xie Qing, Liu Yongjian, et al. A weighted KNN-based automatic image annotation method [J]. Neural Computing and Applications, 2020, 32(11): 6559-6570

[43] Li Jiancheng, Yuan Chun. Automatic image annotation using adaptive weighted distance in improved K nearest neighbors framework [C] //Proc of Pacific-Rim Conf on Advances in Multimedia Information Processing. Berlin: Springer, 2016: 345-354

[44] Kalayeh M, Idrees H, Shah M. NMF-KNN: Image annotation using weighted multi-view non-negative matrix factorization [C] //Proc of IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2014: 184-191

[45] Bakliwal P, Jawahar C. Active learning based image annotation [C/OL] //Proc of the 5th National Conf on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG). Piscataway, NJ: IEEE, 2015 [2020-10-10]. https://ieeexplore.ieee.org/abstract/document/7490061

[46] Verma Y, Jawahar C. Image annotation by propagating labels from semantic neighbourhoods [J]. International Journal of Computer Vision, 2017, 121(1): 126-148

[47] Duda R, Hart P, Stork D. Pattern Classification [M]. New York: Wiley-Interscience, 2001

[48] Liu Jialu, Wang Chi, Gao Jing, et al. Multi-view clustering via joint nonnegative matrix factorization [C] //Proc of the 2013 SIAM Int Conf on Data Mining. Philadelphia, PA: SIAM, 2013: 252-260

[49] Tian Feng, Shen Xukun. Learning label set relevance for search based image annotation [C] //Proc of 2014 Int Conf on Virtual Reality and Visualization. Piscataway, NJ: IEEE, 2014: 260-265

[50] Amato G, Falchi F. KNN based image classification relying on local feature similarity [C] //Proc of the 3rd Int Conf on Similarity Search and Applications. New York: ACM, 2010: 101-108

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

[52] Lin Zijia, Ding Guiguang, Hu Mingqing, et al. Automatic image annotation using tag-related random search over visual neighbors [C] //Proc of ACM Int Conf on Information and Knowledge Management. New York: ACM, 2012: 1784-1788

[53] Cusano C, Ciocca G, Schettini R. Image annotation using SVM [C] //Proc of Int Society for Optics and Photonics. Bellingham, WA: SPIE, 2003: 330-338

[54] Qi Xiaojun, Han Yutao. Incorporating multiple SVMs for automatic image annotation [J]. Pattern Recognition, 2007, 40(2): 728-741

[55] Liu Yang, Wen Kaiwen, Gao Quanxue, et al. SVM based multi-label learning with missing labels for image annotation [J]. Pattern Recognition, 2018, 78(1): 307-317

[56] Ji Ping, Gao Xianhe, Hu Xueyou. Automatic image annotation by combining generative and discriminant models [J]. Neurocomputing, 2017, 236(1): 48-55

[57] Zhang Lei, Ma Jun. Image annotation by incorporating word correlations into multi-class SVM [J]. Soft Computing, 2011, 15(5): 917-927

[58] Murthy V, Can E, Manmatha R. A hybrid model for automatic image annotation [C] //Proc of Int Conf on Multimedia Retrieval. New York: ACM, 2014: 369-376

[59] Jia Yupan, Hyung-jeong Y, Faloutsos C, et al. Gcap: Graph-based automatic image captioning [C/OL] //Proc of Conf on Computer Vision and Pattern Recognition Workshop. Piscataway, NJ: IEEE, 2004 [2020-10-10]. https://ieeexplore.ieee.org/abstract/document/1384943

[60] Liu Jing, Li Mingjing, Ma Weiying, et al. An adaptive graph model for automatic image annotation [C] //Proc of the 8th ACM Int Workshop on Multimedia Information Retrieval. New York: ACM, 2006: 61-70

[61] Chen Gang, Song Yangqiu, Wang Fei, et al. Semi-supervised multi-label learning by solving a Sylvester equation [C] //Proc of the 2008 SIAM Int Conf on Data Mining. Philadelphia, PA: SIAM, 2008: 410-419

[62] Liu Jing, Wang Bin, Lu Hanqing, et al. A graph-based image annotation framework [J]. Pattern Recognition Letters, 2008, 29(4): 407-415

[63] Tang Jinhui, Li Haojie, Qi Guojun, et al. Integrated graph-based semi-supervised multiple/single instance learning framework for imageannotation [C] //Proc of the 16th ACM Int Conf on Multimedia. New York: ACM, 2008: 631-634

[64] Liu Jing, Li Mingjing, Liu Qingshan, et al. Image annotation via graph learning [J]. Pattern Recognition, 2009, 42(2): 218-228

[65] Wang Hua, Hu Jian. Multi-label image annotation via maximum consistency [C] //Proc of 2010 IEEE Int Conf on Image Processing. Piscataway, NJ: IEEE, 2010: 2337-2340

[66] Tang Jinhui, Hong Richang, Yan Shuicheng, et al. Image annotation by KNN-sparse graph-based label propagation over noisily tagged Web images [J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(2): 1-15

[67] Wang Hua, Huang Heng, Ding C. Image annotation using bi-relational graph of images and semantic labels [C] //Proc of Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2011: 793-800

[68] Yang Yi, Wu Fei, Nie Feiping, et al. Web and personal image annotation by mining label correlation with relaxed visual graph embedding [J]. IEEE Transactions on Image Processing, 2011, 21(3): 1339-1351

[69] Liu Xi, Liu Rujie, Li Fei, et al. Graph-based dimensionality reduction for KNN-based image annotation [C] //Proc of the 21st Int Conf on Pattern Recognition. Piscataway, NJ: IEEE, 2012: 1253-1256

[70] Shi Caijuan, Ruan Qiuqi, An Gaoyun. Sparse feature selection based on graph laplacian for Web image annotation [J]. Image and Vision Computing, 2014, 32(3): 189-201

[71] Song Jingkuan, Gao Lianli, Nie Feiping, et al. Optimized graph learning using partial tags and multiple features for image and video annotation [J]. IEEE Transactions on Image Processing, 2016, 25(11): 4999-5011

[72] Liu Zheng, Ma Jun. Refining image annotation by graph partition and image search engine [J]. Journal of Computer Research and Development, 2011, 48(7): 1246-1254 (in Chinese)
(刘峥,马军.一种基于图划分和图像搜索引擎的图像标注改善算法[J].计算机研究与发展,2011,48(7):1246-1254)

[73] Hardoon D, Shawe-taylor J. KCCA for different level precision in content-based image retrieval [C/OL] //Proc of the 3rd Int Workshop on Content-Based Multimedia Indexing. Piscataway, NJ: IEEE, 2003 [2020-10-10]. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.13.6122

[74] Rasiwasia N, Costa P, Coviello E, et al. A new approach to cross-modal multimedia retrieval [C] //Proc of the 18th ACM Int Conf on Multimedia. New York: ACM, 2010: 251-260

[75] Hardoon D, Saunders C, Szedmak S, et al. A correlation approach for automatic image annotation [C] //Proc of Int Conf on Advanced Data Mining and Applications. Berlin: Springer, 2006: 681-692

[76] Zheng Yu, Takiguchi T, Ariki Y. Image annotation with concept level feature using PLSA+CCA [C] //Proc of Int Conf on Multimedia Modeling. Berlin: Springer, 2011: 454-464

[77] Hwang S, Grauman K. Learning the relative importance of objects from tagged images for retrieval and cross-modal search [J]. International Journal of Computer Vision, 2012, 100(2): 134-153

[78] Gong Yunchao, Ke Qifa, Isard M, et al. A multi-view embedding space for modeling Internet images, tags, and their semantics [J]. International Journal of Computer Vision, 2014, 106(2): 210-233

[79] Ballan L, Uricchio T, Seidenari L, et al.A cross-media model for automatic image annotation [C] //Proc of Int Conf on Multimedia Retrieval. New York: ACM, 2014: 73-80

[80] Murthy V, Maji S, Manmatha R. Automatic image annotation using deep learning representations [C] //Proc of the 5th ACM on Int Conf on Multimedia Retrieval. New York: ACM, 2015: 603-606

[81] Uricchio T, Ballan L, Seidenari L, et al. Automatic image annotation via label transfer in the semantic space [J]. Pattern Recognition, 2017, 71(1): 144-157

[82] Hardoon D, Szedmak S, Shawe-taylor J. Canonical correlation analysis: An overview with application to learning methods [J]. Neural Computation, 2004, 16(12): 2639-2664

[83] Mikolov T, Chen Kai, Corrado G, et al.Efficient estimation of word representations in vector space [J]. arXiv preprint, arXiv:1301.3781, 2013

[84] Gutierrez P, Sondag P, Butkovic P, et al. Deep learning for automated tagging of fashion images [C] //Proc of the European Conf on Computer Vision. Berlin: Springer, 2018

[85] Donahue J, Jia Yangqing, Vinyals O, et al. Decaf: A deep convolutional activation feature for generic visual recognition [C] //Proc of Int Conf on Machine Learning. Cambridge, MA: MIT, 2014: 647-655

[86] Oquab M, Bottou L, Laptev I, et al. Learning and transferring mid-level image representations using convolutional neural networks [C] //Proc of IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2014: 1717-1724

[87] Razavian A, Azizpour H, Sullivan J, et al. CNN features off-the-shelf: An astounding baseline for recognition [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2014: 512-519

[88] Zeiler M, Fergus R. Visualizing and understanding convolutional networks [C] //Proc of European Conf on Computer Vision. Berlin: Springer, 2014: 818-833

[89] Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks [C] //Proc of Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2012: 84-90

[90] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [J]. arXiv preprint, arXiv:1409.1556, 2014

[91] Szegedy C, Liu Wei, Jia Yangqing, et al. Going deeper with convolutions [C/OL] //Proc of the IEEE Conf Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015 [2020-10-10]. https://ieeexplore.ieee.org/document/7298594

[92] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition [C] //Proc of IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2016: 770-778

[93] Lecun Y, Boser B, Denker J, et al. Backpropagation applied to handwritten zip code recognition [J]. Neural Computation, 1989, 1(1): 541-551

[94] Lin Min, Chen Qiang, Yan Shuicheng. Network in network [J]. arXiv preprint, arXiv:1312.4400, 2014

[95] Johnson J, Ballan L, Li Feifei. Love thy neighbors: Image annotation by exploiting image metadata [C] //Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2015: 4624-4632

[96] Rawat Y, Kankanhalli M. ConTagNet: Exploiting user context for image tag recommendation [C] //Proc of the 24th ACM Int Conf on Multimedia. New York: ACM, 2016: 1102-1106

[97] Yang Yang, Zhang Wensheng, Xie Yuan. Image automatic annotation via multi-view deep representation [J]. Journal of Visual Communication and Image Representation, 2015, 33(1): 368-377

[98] Wang Jiang, Yang Yi, Mao Junhua, et al. CNN-RNN: A unified framework for multi-label image classification [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 2285-2294

[99] Yeh C, Wu W, Ko W, et al. Learning deep latent space for multi-label classification [C] //Proc of the 31st AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2017: 2838-2844

[100] Wu Jiajun, Yu Yinan, Huang Chang, et al. Deep multiple instance learning for image classification and auto-annotation [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 3460-3469

[101] Wei Yunchao, Xia Wei, Lin Min, et al. HCP: A flexible CNN framework for multi-label image classification [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(9): 1901-1907

[102] Wu Baoyuan, Jia Fan, Liu Wei, et al. Diverse image annotation [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 2559-2567

[103] Wu Baoyuan, Chen Weidong, Sun Peng, et al. Tagging like humans: Diverse and distinct image annotation [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7967-7975

[104] Goodfellow I, Pouget-abadie J, Mirza M, et al. Generative adversarial nets [C] //Proc of Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2014: 2672-2680

[105] Kulesza A, Taskar B. Determinantal point processes for machine learning [J]. Foundations and Trends® in Machine Learning, 2012, 5(2/3): 123-286

[106] Zhang Junjie, Wu Qi, Zhang Jian, et al. Mind your neighbours: Image annotation with metadata neighbourhood graph co-attention networks [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 2956-2964

[107] Zhang Junjie, Wu Qi, Zhang Jian, et al. Kill two birds with one stone: Weakly-supervised neural network for image annotation and tag refinement [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2018

[108] Yosinski J, Clune J, Bengio Y, et al. How transferable are features in deep neural networks [C] //Proc of Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2014: 3320-3328

[109] Gong Yunchao, Jia Yangqing, Leung T, et al. Deep convolutional ranking for multilabel image annotation [J]. arXiv preprint, arXiv:1312.4897, 2013

[110] Tan Ben, Song Yangqiu, Zhong Erheng, et al. Transitive transfer learning [C] //Proc of the 21st ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2015: 1155-1164

[111] Tan Ben, Zhang Yu, Pan S, et al. Distant domain transfer learning [C] //Proc of the 31st AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2017: 2604-2610

[112] Von-ahn L, Dabbish L. Labeling images with a computer game [C] //Proc of the 2004 Conf on Human Factors in Computing Systems. New York: ACM, 2004: 319-326

[113] Grbinger M, Clough P, Mller H, et al. The IAPRTC-12 benchmark: A new evaluation resource for visual information systems [C] //Proc of LREC Workshop Ontoimage Language Resources for Content-Based Image Retrieval. Luxemburg: ELRA, 2006: 13-23

[114] Chua T, Tang Jinhui, Hong Richang, et al. NUS-WIDE: A real-world Web image database from National University of Singapore [C/OL] //Proc of the ACM Int Conf on Image and Video Retrieval. New York: ACM, 2009 [2020-10-10]. https://dl.acm.org/doi/10.1145/1646396.1646452

[115] Lin T, Maire M, Belongie S, et al. Microsoft coco: Common objects in context [C] //Proc of European Conf on Computer Vision. Berlin: Springer, 2014: 740-755

[116] Deng Jia, Dong Wei, Socher R, et al. ImageNet: A large-scale hierarchical image database [C] //Proc of 2009 IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2009: 248-255

Ma Yanchun, born in 1984. PhD candidate. His main research interests include computer vision, image processing and machine learning.

Liu Yongjian, born in 1962. Bachalor, professor, PhD supervisor. His main research interests include digital publication, digital communication and knowledge service.

Xie Qing, born in 1986. PhD, associate professor. Member of CCF. His main research interests include data mining, machine learning, and recommender system.

Xiong Shengwu, born in 1966. PhD, professor, PhD supervisor. His main research interests include machine learning and data mining.

Tang Lingli, born in 1989. PhD, lecturer, master supervisor. Her main research interests include knowledge management, digital communication engineering, and digital copyright.