doi:10.3772/j.issn.1002-0470.2021.01.009

基于级联网络和语义层次结构的图像自动标注方法①

翟 晴②*** 顾广华③*** 孙雅倩*** 任贤龙***

(*燕山大学信息科学与工程学院 秦皇岛 066004)

(**河北省信息传输与信号处理重点实验室 秦皇岛 066004)

摘 要 针对大多数的图像自动标注结果中含有冗余标签、信息量不够丰富的问题,本文提出了一种基于级联网络和语义层次结构的图像自动标注方法(CNSH)。首先,输入数据集的图片和标签列表,采用级联的 VGG 网络提取图像特征,训练条件行列式点过程(DPP)算法模型,计算标签的质量分数确定候选标签列表;其次,利用 WordNet 检索数据集标签得到语义层次结构和同义词,进而构建加权语义路径;最后,利用 DPP 算法在候选标签集中采样,得到最终的标注结果。与传统的图像标注任务相比,本文方法得到的标注结果能准确描述图片内容,且不含冗余标签。许多评估指标已用于图像标注和多标签学习,但是它们只专注于评估代表性,忽略了多样性。为了解决上述问题,本文采用了基于语义层次结构的语义指标来共同评估代表性和多样性。在 IAPRTC-12 和 ESP Game 2 个基准数据集上的实验表明,与现有方法相比本文方法能够产生更具代表性和多样性的标签。

关键词 图像自动标注;级联网络;行列式点过程(DPP);语义层次结构(SH);语义指标

0 引 言

随着互联网技术的快速发展和移动终端的普及,互联网上的图片数据快速增长。为方便管理和利用图片数据,人们会为每幅图片添加一些相应的标注词,然而人工标注成本越来越高,并且人工标注的结果也具有一定主观性,不能很客观地表示图片内容,所以实现图像的自动标注成为越来越迫切的需求。近年来,人工智能的一些方法和思想应用于图像标注领域,并取得了一定的研究成果。

图像自动标注就是让计算机系统依据图像内容为图像自动添加几个恰当的标签(关键词)的过程。现有的图像自动标注方法可以大致分为两类:基于分类的方法[1]和基于概率模型的方法[2]。基于分类的标注方法是以图像分类的思想出发的标注算法,数据集中存在的每个语义类均需要利用语义目标图像训练分类器,主要方法有基于最近邻(k-nearest neighbor, KNN)的方法[3-4]、基于支持向量机(support vector machine, SVM)的方法[5-7]、基于决策树(decision tree, DT)的方法[8]以及基于深度学习的方法等。基于概率的方法主要是通过提取图像(或者图像区域)的视觉信息(如颜色、形状、纹理、空间关系等),然后计算图像的视觉特征与图像标签之间的联合概率分布,最后利用该概率分布对未标注图像(图像区域)进行标注,主要方法有基于主题相关模型的方法[9]、基于混合模型的方法[10]以及基于稀疏编码的方法等[11-12]。众所周知,不同的标签可以传递相似的语义信息,但是上述绝大多数的图像自动标注方法,并没有过多考虑图像标签之间的关系,忽视了标签之间具有相互依赖性的问题,导致最终的标注结果含有冗余标签。2017 年 Wu 等

① 国家自然科学基金面上(62072394)和河北省自然科学基金面上(F2017203169)资助项目。

② 女,1992 年生,硕士生;研究方向:图像自动标注;E-mail: zq0312612611@163.com

③ 通信作者,E-mail: guguanghua@ysu.edu.cn

(收稿日期:2020-01-30)

人[13]利用基于语义层次结构的条件行列式点过程(determinantal point process, DPP)采样算法,得到了不含冗余标签的标注结果。但由于其利用单个网络为图像提取特征,提取的特征不够全面,导致标注的准确率和召回率较低。

为了解决这一问题,本文提出了一种基于级联网络和语义层次结构的图像自动标注方法(automatic image annotation method based on cascade network and semantic hierarchy, CNSH)。级联的融合网络能扩充网络的整体宽度,减少图像特征的信息损失,提取的特征更丰富;语义层次结构能够描述标签之间的语义依赖关系,对基于此结构的加权语义路径和候选标签集进行采样,能够得到更准确、更紧凑的标签列表。本文方法不仅抑制了信息冗余,还能得到不含标签冗余的标注结果,同时提升了标注的准确率和召回率。

1　基于级联网络和语义层次结构的图像自动标注

本文算法框图如图1所示。首先对图像进行预处理,即利用级联的VGG16和VGG19预训练模型提取图像的特征。然后利用提取的图像特征训练条件DPP模型[14],通过标签的质量分数计算得到候选标签集。同时输入数据集的标签列表,确定标签之间的相互依赖关系,找到同义词对、父子对和部分整体对,利用WordNet[15]网络为数据集构建语义层次结构,进而得到数据集的加权语义路径。最后利用DPP算法在候选标签集中进行采样,在每个采样过程结束时,计算采样子集中的标签权重。众所周

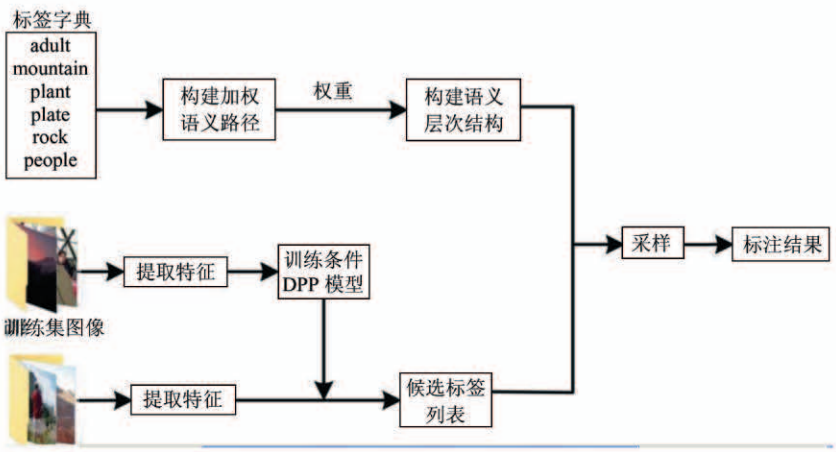

图1　基于级联网络和语义层次结构的图像自动标注算法框图

知,具有较大权重和的标签子集能够表示更多内容,所以本文选取了具有最大标签权重和的子集作为最终输出,从而得到最终的标注结果。

1.1　特征提取

卷积神经网络可以为许多计算机视觉任务学习非常有效的特征,近年来深度学习的方法在许多领域具有良好的表现,例如人脸识别[16]和目标检测[17]。VGG卷积神经网络[18]是牛津大学计算机视觉实验室在2014年提出来的模型,并在当年的LSVRC比赛中取得佳绩。与其他的深度卷积神经网络相比,VGG网络的优势在于采用了多层的卷积层组合并配以小尺寸的滤波器,在实现大尺寸滤波器感受野的同时,还能使参数数量减少,表示层深度增加,网络性能得到有效提升,模型融合的性能优于单模型的性能。训练期间分阶段设置了梯次下降的不同学习率,有助于模型的快速收敛。

大多数图像处理领域内的学习任务,通常采用单一的网络结构为图像提取特征,提取的特征不够充分,网络性能不够好,最终结果也差强人意。本文分别用VGG16和VGG19两种网络对图像提取4096维特征,然后利用主成分分析法(principal component analysis, PCA)[19]对特征进行降维,分别得到520维特征和620维特征,再将两种网络级联在一起,得到1140维的级联特征。这种级联网络能有效

地改善网络的学习效果。

特征提取的具体过程如图 2 所示。

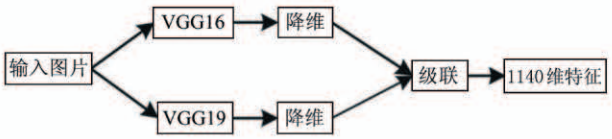

图 2 特征提取示意图

1.2 条件 DPP 模型

令 $X = \{x_1, \cdots, x_n\}$ 表示训练图像集，对于每个图像 x_i，提供一个标签集 $y_i \subset \tau = \{1,2,\cdots,m\}$，$\tau$ 是包含 m 个候选标签的整个标签集。关于图像和标签子集对 (x, y) 的条件 DPP 公式如下：

$$P_\delta(y \mid x) = \frac{\det(\boldsymbol{L}_y(x;\delta))}{\det(\boldsymbol{L}_\tau(x;\delta) + \boldsymbol{I})} \tag{1}$$

其中，m 个标签的核矩阵 $\boldsymbol{L}_\tau(x;\delta) \in \boldsymbol{R}^{m\times m}$ 是半正定的，δ 是其中的参数，$\boldsymbol{I}$ 为单位矩阵，$\boldsymbol{L}_y(x;\delta) \in \boldsymbol{R}^{|y|\times|y|}$ 表示从 $\boldsymbol{L}_\tau(x;\delta)$ 中提取与 $y \subset \tau$ 中的标签索引相对应的行和列而生成的子核矩阵。

假设 $\boldsymbol{L}_\tau(x;\delta) = [a_{ij}]$，$ij = (1,2,3,4)$，$y = \{1,2\}$，那么 $\boldsymbol{L}_y(x;\delta) = [a_{11}, a_{12}; a_{21}, a_{22}]$，$a_{11}$ 和 a_{22} 分别表示两个标签各自的分数，a_{12} 和 a_{21} 表示标签之间的相关性。$\boldsymbol{L}_y$ 的行列式可以表示为

$$\det(\boldsymbol{L}_y) = a_{11}a_{22} - a_{12}a_{21} \tag{2}$$

$\det(\boldsymbol{L}_y)$ 编码子集 y 中的标签之间的负相关。如果 $\det(\boldsymbol{L}_y)$ 很小，表明这两个标签是高度相关的，那么概率 $P_\delta(y \mid x)$ 的值也很小；如果 $\det(\boldsymbol{L}_y) = 0$，表明这两个标签是完全相关的，$P_\delta(y \mid x)$ 的值为 0。可以看出，若将 $\boldsymbol{L}_y$ 推广为一般的行列式，该结论也是成立的。显然，式(1)不鼓励使用带有冗余标签的标签子集。

当标签子集很大时，直接学习核矩阵 $\boldsymbol{L}_\tau(x;\delta)$ 会特别困难。由于一个葛朗姆矩阵(Gram-matrix)可以表示任何一个实对称矩阵[20]，所以分解式(1)有：

$$\boldsymbol{L}_{ij}(x;\delta) = q_i(x)\boldsymbol{\phi}_i(x)^{\mathrm{T}}\boldsymbol{\phi}_j(x)q_j(x) \tag{3}$$

其中 $\delta = [\delta_1, \delta_2, \cdots, \delta_m]$ 表示与每个标签相对应的一组质量参数。$q_i(x)$ 衡量的是第 i 个标签的质量分数，表式如下：

$$q_i(x) = \exp(0.5\delta_i^{\mathrm{T}}x) \tag{4}$$

$\boldsymbol{\phi}_i(x) \in \boldsymbol{R}^m$ 是一个正则化的特征向量，其中 $\|\boldsymbol{\phi}_i(x)\| = 1$。

$$\boldsymbol{\varphi}(x) = \boldsymbol{\phi}_i^{\mathrm{T}}\boldsymbol{\phi}_j(x) = \frac{\boldsymbol{L}_{ij}}{\sqrt{\boldsymbol{L}_{ii}\boldsymbol{L}_{jj}}} \in \boldsymbol{R}^{m\times m} \tag{5}$$

$\boldsymbol{\varphi}(x)$ 是衡量标签之间相似度的矩阵，它编码标签之间的负相关，使标注结果不含冗余信息。为了清楚起见，将 $\boldsymbol{\varphi}(x)$ 表示为 $\boldsymbol{S}$，本文采用了与 x 无关的余弦相似性矩阵，即：

$$\boldsymbol{S}(i,j) = \frac{1}{2} + \frac{< \boldsymbol{t}_i, \boldsymbol{t}_j >}{2\|t_i\|_2\|t_j\|_2} \in [0,1] \quad \forall i,j \in \tau \tag{6}$$

其中 $t_i \in R^{50}$ 表示标签，用 GloVe 算法[21]推导得到。$< \boldsymbol{t}_i, \boldsymbol{t}_j >$ 表示两个向量 $\boldsymbol{t}_i$ 和 $\boldsymbol{t}_j$ 的内积，而 $\|\boldsymbol{t}_i\|_2$ 和 $\|\boldsymbol{t}_j\|_2$ 分别表示向量 $\boldsymbol{t}_i$ 和 $\boldsymbol{t}_j$ 的 2 范数。则式(1)可以被重新表示为

$$P_\delta(y \mid x) = \frac{\prod_{i\in y}[\exp(\delta_i^{\mathrm{T}}x)]\det(\boldsymbol{S}_y)}{\sum_{y'\subset\tau}\prod_{i\in y'}[\exp(\delta_i^{\mathrm{T}}x)]\det(\boldsymbol{S}_{y'})} \tag{7}$$

其中 $\boldsymbol{S}_y \in \boldsymbol{R}^{|y|\times|y|}$ 和 $\boldsymbol{S}_{y'} \in \boldsymbol{R}^{|y'|\times|y'|}$ 是对应于标签子集 $y \subset \tau$ 和 $y' \subset \tau$ 的 $\boldsymbol{S}$ 的子矩阵。

根据 DPP 的相关定义，得到多样性核 $\boldsymbol{S}$ 之后，本文通过使用 l_2 正则化最小化负对数似然[22]学习参数 δ，然后给定梯度，采用反向传播和随机梯度下降算法[23]优化 δ。最后利用参数 δ 计算标签的质量分数 q，根据 q 确定候选标签集。

1.3 构建语义层次结构

为了描述标签之间的语义依赖关系，本文利用 WordNet 分别为数据集 IAPRTC-12[24] 和 ESP Game[25] 构建语义层次结构(semantic hierarchy，SH)。WordNet 是一个英语词库，收录了超过 10 万个实词。在 WordNet 中，意义相近的单词组成一个同义词组，而同义词组之间则以上-下义，同义-反义，整体-部分以及蕴含等语义关系连接在一起，构成一个由同义词组作为结点、语义关系作为边的网状结构。

首先在 WordNet 中搜索数据集中的每个类，然后提取一个或多个有向路径(即从父类到子类的长序列有向边)。在每个路径中，将标识词汇表中最

近的类(即该数据集的所有类的集合($\{c_1,\cdots,c_m\}$)标识为父类。对数据集中所有 m 个类重复该过程,以形成语义层次结构。在构建语义层次结构的过程中,将标签视为一个节点,做如下规定。

(1) 既无父类标签也无子类标签为类标签;

(2) 有父类标签无子类标签为叶节点;

(3) 无父类标签有子类标签为根节点;

(4) 既有父类标签又有子类标签为中间节点。

在构建语义层次结构的过程中,本文考虑了两种不同类型的语义依赖,即子类与父类、部分与整体之间的标签依赖。图 3 显示了同义词和子类父类关系的语义层次结构。图 4 显示了部分与整体关系的语义层次结构。其中,第 1 行标签为语义层次结构的根节点,第 2 行标签为语义层次结构的中间节点,第 3 行标签为语义层次结构的叶节点。

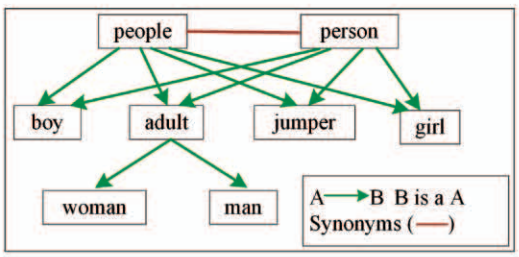

图 3　同义词和子类父类关系的语义层次结构

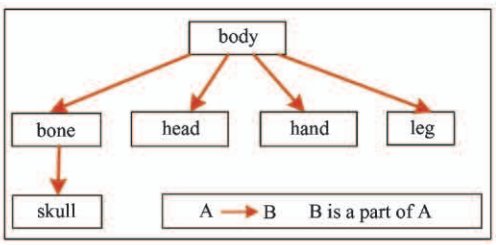

图 4　部分与整体关系的语义层次结构

在 ML-MG 中[26],如果后代标签存在(例如“woman”),则其所有祖先标签(例如“adult”和“people”,“adult”和“person”)也必须存在,导致最终的标注结果中含有大量冗余信息。在本文的采样算法中,为了减少冗余,语义层次结构的使用规则与 ML-MG 完全不同,本文不鼓励同时选择具有语义依赖性的两个标签,即如果后代标签存在,则其祖先标签一定不存在。例如,若存在后代标签“woman”,则其所有祖先标签(“adult”和“people”,“adult”和“person”)都将不存在。

同义词表示两个标签具有相同或相似含义的状态,例如“people”和“person”,“street”和“road”。通过对 WordNet 的研究发现,在许多基准图像数据集中,例如 IAPRTC-12 和 ESP Game,同义词标签对普遍存在。Weston 等人[27]曾使用同义词来修改评估度量,在本文中,同义词不仅用于定义语义度量,还用于阻止选择同义标签,从而减少冗余信息。

1.4　加权语义路径

加权语义路径(weighted semantic paths, WSP)是在所有候选标签中的同义词和语义层次结构的基础上构建而成的。构建加权语义路径的具体过程如下。首先,将每个标签看成一个节点,如果两个有向路径仅在同义标签上不同,则将同义词合并为一个节点。然后,在语义层次结构中找到所有叶节点,连接其直接父节点,重复此过程,直至连接至全部的根节点。最后,给每个标签加上不同的权重,则连接的所有节点就构成了叶节点的加权语义路径。对应于整个标签集的所有语义路径表示为

$$WSP = \{WSP_1,\cdots,WSP_\theta\} \tag{8}$$

其中,θ 表示整个标签集的语义路径个数。

为了得到标签权重的计算模型,本文从两个方面进行考量:(1)后代标签比其祖先标签含有更加具体的信息(例如,“boy”比“people”更具信息性)。因此,后代标签的权重应高于其祖先标签的权重。(2)在语义层次结构中,每个祖先标签的后代个数各不相同,说明每个祖先标签含有的信息量不同,后代越多,祖先标签含有的信息量越少,反之含有的信息量越多,所以每个标签的后代个数是另一个值得考虑的因素。因此,将标签权重建模为与其后代数成反比。结合以上观察结果,将 WSP_j 中标签 y_i 的权重定义为

$$\boldsymbol{W}_{ij} = \alpha^{l_{ij}}/|d_i| \tag{9}$$

其中,$|d_i|$ 表示标签 y_i 的后代数,l_{ij} 代表 WSP_j 中的标签 y_i 的层数,$\alpha \in (0,1)$ 表示层与层之间的衰变因子,设置 $\alpha = 0.7$。

于是,整个语义路径集合中的标签 y_i 的权重被定义为其在所有语义路径中的权重之和,即 $W_i = \sum_j^{|WSP|} W_{ij}$,所有标签的权重可以连接成一个向量:

$$\boldsymbol{W} = (W_1, W_2, \cdots, W_m) \tag{10}$$

本文还定义了一个图像的标签子集 Y 的语义路径 WSP_y，如图 5 所示，完整标签列表为 Y = {“people”，“person”，“child”，“boy”}。首先，从完整的语义路径中，剪裁出与 Y 相对应的部分标签路径，即[(“person”，“people”)→“child”]；然后，调整标签权重，“child”的权重从 0.7 变为 1，变化因子为 1/0.7。最后，使用相同的变化因子，将“people”和“person”的权重从 0.245 调整为 0.35 = 0.245 × (1/0.7)，如图 6 所示。

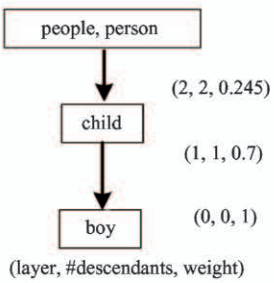

图 5 完整的加权语义路径

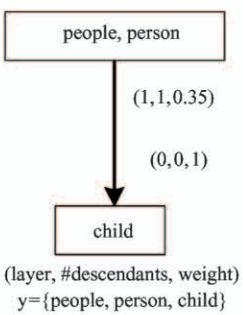

图 6 从完整语义路径中重新构造的语义路径

1.5 具有加权语义路径的 k-DPP 采样

根据候选标签集和加权语义路径(WSP)，获得具有至多 k 个标签的标签子集 Y，这就是 k-DPP 采样的顺序采样过程，主要操作步骤如图 7 所示。为了避免标注结果中含有冗余标签，本文不鼓励同时选择具有语义依赖性的两个标签。具体地，在每个采样过程中，将检查新采样的标签是否来自与任何先前采样的标签相同的语义路径。如果不是，它将被包含在标签子集中；如果是，它将会被放弃，继续对下一个标签进行采样，直至获得 k 个标签。整个采样过程重复多次，以获得不同的标签子集 Y。由于标签子集的权重和越大，含有的冗余信息越少，有效信息越多，越能够更好地表示图像内容，所以本文选取具有最大标签权重和的子集作为最终输出。

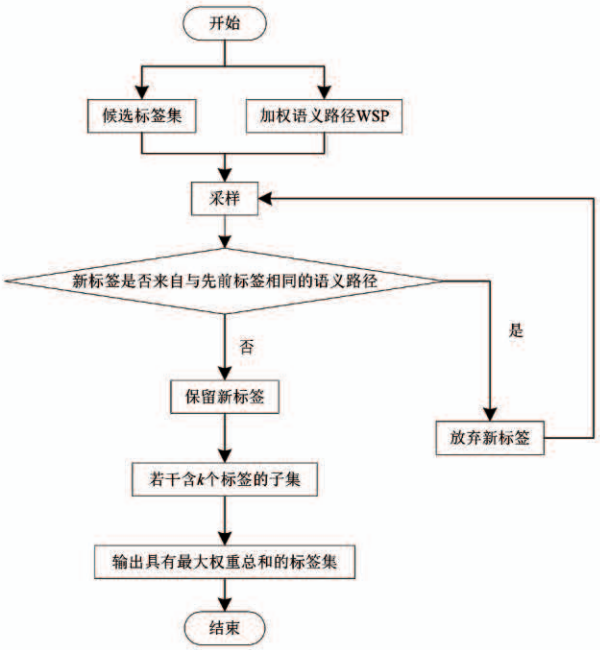

图 7 具有加权语义路径的 k-DPP 采样框图

2 实验结果与分析

2.1 实验设置

数据集 本文在图像标注的两个基准数据集上进行实验，即 IAPRTC-12(19 628 幅图像，291 个标签)和 ESP Game(20 770 幅图像，268 个标签)。在构建语义层次结构过程中，发现在 IAPRTC-12 中存在许多重复图像，本文删除了这些冗余图像(170 个训练图像和 5 个测试图像)。

参数 学习 $\boldsymbol{W}$ 的随机梯度下降参数设置如下。初始学习率为 20，衰减为 0.02。学习率每 50 次迭代更新一次，动量为 0.9，批量大小为 1024。最大历元数为 5，l_2 正则化的参数为 $\eta = 0.0001$。

语义评估指标 已有许多评估指标用于图像标注和多标签学习，如准确度、召回率和 F 值[28]。但是，这些指标并不适合本文的标注任务，因为它们只专注于评估代表性，忽略了多样性。因此，本文采用了基于语义层次结构的语义指标来共同评估代表性和多样性。

本文规定每个标签的权重等于其标签分数，计算每个语义路径中每个标签的分数，那么这条路径

中最大的标签得分(即标注词的得分)就是路径得分,每个语义路径的预测分数在之间。具体做法如下。

(1) 统计预测标签子集的个数、预测语义路径的个数以及真实语义路径的个数,在预测标签某一子集内找到一个需要检索的标签。

(2) 在真实的语义路径中找到这个待检索的标签,计算这个真实语义路径的路径得分。

(3) 计算这个标签子集的路径总得分。

在实验中,本文将报告所有测试图像的平均分数。

本文采用 F 值评估实验效果,公式如下:

$$P = T/pred \tag{11}$$

$$R = T/gt \tag{12}$$

$$F = 2PR/(P + R) \tag{13}$$

其中,T 表示每个标签子集的路径总得分,$pred$ 表示预测语义路径的个数,gt 表示真实语义路径的个数,P 表示准确率,R 表示召回率。

2.2　实验结果与分析

首先,与现有的图像标注方法 DIA[13] 和 ML-MG[26] 进行比较。由于 DIA 标注 3 个、5 个标签时,分别在前 6 个、前 8 个候选标签中进行采样,为了统一比较标准,本文遵循与 DIA 方法一致的采样规则。其中,本文复现 DIA 方法时测得的评估结果与其原文报告的结果略有偏差。

其次,为了验证级联网络的性能,本文比较了所提方法的两种变体,即 CNSH-VGG16 和 CNSH-VGG19。CNSH-VGG16 表示仅使用 VGG16 预训练模型为图像提取特征;CNSH-VGG19 表示仅使用 VGG19 预训练模型为图像提取特征,然后训练 DPP 模型,构建语义层次结构以及加权语义路径,最后根据候选标签集和加权语义路径,获得具有至多 k 个标签的标签子集 Y 的方法。

最后,为了更直观地进行级联网络性能的实验对比,将本文提出的 CNSH 表示为 CNSH-VGG16-VGG19。

2.2.1　实验结果与分析

依据标注标签数量的不同(3 个、5 个标签)将语义指标评估结果分为两类,对本文与其他标注方法在数据集 IAPRTC-12 和 ESP Game 上的标注性能表现进行比较,结果如表 1 和表 2 所示。所有实验结果均在同一实验条件下测得。

为了更好地分析说明本文方法的优越性,本文把对比方法分为如下 3 类。

第 1 类是 ML-MG。它利用线性不等式约束来鼓励标签按语义层次的标签顺序排名,通过标签共现来鼓励类似标签具有相似的分数。因此,祖先标签总是排在其后代标签之前,如果从 ML-MG 的标签排名列表中选择前 3 个或前 5 个标签,则得到的标签大多是祖先标签(对应于加权语义路径中的具有较低权重的标签)和较少其他路径的标签。虽然通过这种方法能够得到准确度较高的标注结果,但是标注结果中含有的冗余标签比较多,不能够很好地代表图像的内容。

第 2 类是 DIA。它同样利用基于语义层次结构的条件 DPP 采样算法,得到准确且不含冗余标签的结果。不同之处在于它仅使用预训练的 VGGF 模型提取图像特征,本文采用的是级联的 VGG 网络提取

表 1　在 IAPRTC-12 基准数据集上的语义指标评估结果

方法	3 个标签			5 个标签		
	P/%	R/%	F/%	P/%	R/%	F/%
ML-MG	35.33	17.56	21.17	**41.78**	29.71	31.27
DIA*	42.99	24.86	29.81	38.37	34.73	34.27
CNSH-VGG16	41.35	25.03	29.27	37.71	34.16	33.71
CNSH-VGG19	40.89	24.46	28.72	37.12	33.83	33.31
CNSH-VGG16-VGG19	**44.26**	**26.17**	**30.93**	39.80	**35.90**	**35.56**

* 原文语义指标评估结果:3 个标签　P=44.01, R=25.16, F=30.13

5 个标签　P=38.91, R=34.21, F=34.23

表 2 在 ESP Game 基准数据集上的语义指标评估结果

方法	3 个标签			5 个标签		
	P/%	*R*/%	*F*/%	*P*/%	*R*/%	*F*/%
ML-MG	30.23	16.38	29.81	**36.42**	29.51	30.38
DIA*	40.20	29.42	32.21	34.70	39.73	35.04
CNSH-VGG16	38.25	27.62	30.21	33.81	38.25	33.77
CNSH-VGG19	37.96	27.25	29.86	33.28	37.19	33.08
CNSH-VGG16-VGG19	**41.38**	**30.23**	**32.26**	35.90	**40.44**	**35.83**

* 原文语义指标评估结果：3 个标签 $P=42.37$，$R=30.48$，$F=33.43$

5 个标签 $P=36.15$，$R=40.10$，$F=35.90$

图像特征。通过实验结果可以看出，CNSH 能够得到更好的标注性能，初步证明级联的卷积神经网络能够有效改进实验性能。

第 3 类是 CNSH-VGG16、CNSH-VGG19 和 CNSH-VGG16-VGG19 之间的比较。CNSH-VGG16 和 CNSH-VGG19 分别表示只采用 VGG16 和 VGG19 为图像提取特征的实验结果，CNSH-VGG16-VGG19 则表示采用级联的 VGG 卷积神经网络为图像提取特征的实验结果，通过对实验结果的比较可以看出，无论标注结果是 3 个标签还是 5 个标签，CNSH-VGG16-VGG19（即本文提出的 CNSH）都表现出了不错的性能。

2.2.2 实际标注效果

图 8 显示了采用 CNSH 及人工标注得到的部分结果示例对比。人工标注结果具有一定的主观性，其标注词数量不确定。标注词数量过多，可能导致

图 8 在 IAPRTC-12 基准数据集上的标注结果对比

信息冗余,标注词过少,则易出现对图像内容描述不充分的情况。标签冗余包含几种情况,同义词关系冗余、父类子类关系冗余、部分和整体关系冗余。如图 8 所示,第 1 幅图的人工标注结果中,体现了同义词冗余的情况,比如同义词“street”和“road”。通过 CNSH 得到的标注结果则能适当选取同义词中的某一个单词,从而避免了冗余;第 2 幅图的人工标注结果中,展示了包含子类和父类关系的标签冗余,即“people”和“tourist”。使用 CNSH 则能得到一个更加紧凑的标签列表(即{“band”,“light”,“man”,“red”,“wheel”}),它不仅能准确描述图像内容,还避免了冗余,使得标注结果更加多样;第 3 幅图的人工标注结果中包含了部分和整体关系(“bed”和“bedside”)的标签冗余,利用 CNSH 能够得到既正确又多样化的标签列表。此外,当标注词过少,标注结果不能准确代表图像内容时,运用 CNSH 能适当增加标注词,完善对图像内容的描述,如图 8 的第 4 幅图所示。上述结果表明,使用 CNSH 得到的标签列表更加紧凑,能准确描述图像内容。

3　结　论

本文提出了一种基于级联网络和语义层次结构的图像自动标注方法(CNSH),它的任务就是要求少数被采样到的标签不仅能够代表图像,还更加多样化。本文将自动标注任务视为子集选择问题,并用条件 DPP 模型对其进行建模,根据候选标签集和加权语义路径(WSP),获得具有至多 k 个标签的标签子集。在 IAPRTC-12 和 ESP Game 2 个基准数据集上的实验表明,本文提出的方法优于当前的图像自动标注方法。同时,本文采用基于语义层次结构的语义指标来共同评估代表性和多样性,与传统评价指标相比更加符合人类标注的习惯。

参考文献

[1] Chu G L, Niu K, Tian B Y. Automatic image annotation combining SVMS and KNN algorithm[C]//The 3rd IEEE International Conference on Cloud Computing and Intelligence Systems, Shenzhen, China, 2014: 13-17

[2] Wang X Y, Feng S H, Lang C Y. Semi-supervised dual low-rank feature mapping for multi-label image annotation [J]. *Multimedia Tools and Applications*, 2019, 78(10): 13149-13168

[3] 柯逍, 周铭柯, 牛玉贞. 融合深度特征和语义邻域的自动图像标[J]. 模式识别与人工智能, 2017, 30(3): 193-203

[4] Lin Z J, Ding G G, Hu M Q, et al. Automatic image annotation using tag-related random search over visual neighbors[C]//21st ACM International Conference on Information and Knowledge Management, Maui, USA, 2012: 1784-1788

[5] Zhang L, Lin F Z, Zhang B. Support vector machine learning for image retrieval[C]//IEEE International Conference on Image Processing (ICIP), Thessaloniki, Greece, 2001: 721-724

[6] Murthy V N, Can E F, Manmatha R. A hybrid model for automatic image annotation[C]//2014 4th ACM International Conference on Multimedia Retrieval, Glasgow, UK, 2014: 369-376

[7] Yang C B, Dong M, Hua J. Region-based image annotation using asymmetrical support vector machine-based multiple instance learning[C]//2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, USA, 2016: 2057-2063

[8] Hou J, Chen Z, Qin X, et al. Automatic image search based on improved feature descriptors and decision tree [J]. *Integrated Computer-Aided Engineering*, 2011, 18(2): 167-180

[9] Xu X, Shimada A, Taniguchi R. Correlated topic model for image annotation[C]//19th Korea-Japan Joint Workshop on Frontiers of Computer Vision, Incheon, Korea, 2013: 201-208

[10] Moran S, Lavrenko V. Sparse kernel learning for image annotation[C]//2014 4th ACM International Conference on Multimedia Retrieval, Glasgow, UK, 2014: 113-120

[11] Jing X Y, Wu F, Li Z Q, et al. Multi-label dictionary learning for image annotation[J]. *IEEE Transactions on Image Processing*, 2016, 25(6): 2712-2725

[12] Lu Z W, Han P, Wang L W, et al. Semantic sparse recoding of visual content for image applications[J]. *IEEE Transactions on Image Processing*, 2015, 24(1): 176-188

[13] Wu B Y, Jia F, Liu W, et al. Diverse image annotation [C]//The 30th IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 6194-6202

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

[15] Fellbaum C, Miller G. WordNet: An Electronic Lexical Database[M]. Cambridge, MA: MIT Press, 1998: 265-283

[16] Zheng T Y, Deng W H, Hu J N. Age estimation guided convolutional neural network for age-invariant face recognition[C]//The 30th IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, USA, 2017: 503-511

[17] Ren S Q, He K M, Girshick R, et al. CNN: towards real-time object detection with region proposal networks [J]. *IEEE Computer Society*, 2017, 39(6): 1137-1149

[18] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]//The 3rd International Conference on Learning Representations, San Diego, USA, 2015: 1-5

[19] Wold S. Principal component analysis[J]. *Chemometrics and Intelligent Laboratory Systems*, 1987, 2(1): 37-52

[20] 张亚召. 基于行列式点过程的多语言多文档摘要研究[D]. 北京:北京邮电大学计算机学院, 2018: 8-15

[21] Penning J, Socher R, Manning C D. Glove: global vectors for word representation[C]//2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 2014: 1532-1543

[22] Goodfellow I, Bengio Y, Courville A. Deep Learning [M]. Cambridge, MA: MIT Press, 2016: 224-236

[23] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors Cognitive modeling[J]. *Nature*, 1986, 323(6088): 533-536

[24] Grubinger M, Clough P, et al. The IAPR TC12 benchmark: a new evaluation resource for visual information systems[C]//2006 International Workshop onto Image, Genoa, Italy, 2006: 13-23

[25] Ahn V L, Dabbish L A. ESP: Labeling images with a computer game[C]//2005 AAAI Spring Symposium, Stanford, USA, 2005: 91-98

[26] Wu B Y, Lyu S W, Ghanem B. ML-MG: multi-label learning with missing labels using a mixed graph[C]//The 15th IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 4157-4165

[27] Weston J, Bengio S, Usunier N. Large scale image annotation: learning to rank with joint word-image embeddings [J]. *Machine Learning*, 2010, 81(1): 21-35

[28] Zhang M L, Zhou Z H. A review on multi-label learning algorithms[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2014, 26(8): 1819-1837

Automatic image annotation method based on cascade network and semantic hierarchy

Zhai Qing[*,**], Gu Guanghua[*,**], Sun Yaqian[*,**], Ren Xianlong[*,**]

([*]School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004)

([**]Hebei Key Laboratory of Information Transmission and Signal Processing, Qinhuangdao 066004)

Abstract

There are some problems in automatic image annotation, such as the redundant labels and the lack of sufficient information. Aiming at the problem, an automatic image annotation method based on cascade network and semantic hierarchy (CNSH) is proposed. Firstly, from the input images and label list of dataset, image features are extracted through a cascaded VGG network. The condition determinantal point process (DPP) model is trained to compute the quality score of labels that is used to determine the list of candidate labels. Secondly, the semantic hierarchy and synonyms are obtained via a label set, WordNet to build a weighted semantic path. Finally, this work samples the candidate label set by using the DPP algorithm to obtain the final annotation results. Compared with the traditional image annotation methods, the annotation results are able to accurately describe the image content without redundant labels. Though many evaluation indexes are applied for the image annotation and the multi-label learning, they only focus on evaluating the representativeness and ignore the diversity. To solve the above drawbacks, the semantic metrics based on semantic hierarchy structure are applied to jointly evaluate the representativeness and diversity in this paper. Experimental results on IAPRTC-12 and ESP Game benchmark datasets demonstrate that the proposed method produces the more representative and diversified labels than the existing methods.

Key words: automatic image annotation, cascade network, determinantal point process (DPP), semantic hierarchy (SH), semantic metrics