基于Sentence-Rank的图像句子标注

徐守坤[1],徐 坚[1],李 宁[1,2],周 佳[1],刘楚秋[3]
1.常州大学 信息科学与工程学院 数理学院,江苏 常州 213164
2.福建省信息处理与智能控制重点实验室(闽江学院),福州 350108
3.常州工学院 电气与光电工程学院,江苏 常州 213032

摘 要:传统的图像语义句子标注是利用句子模板完成对图像内容描述,但其标注句子很难做到符合语言逻辑。针对这一问题,提出基于统计思想从语料库中选出一条最优的句子来描述图像内容,设计以 N -gram算法为主要思想的Sentence-Rank算法生成标注句子。首先执行机器视觉特征学习,选择标注性能最好的HSV-LBP-HOG融合特征完成图像分类,获得图像标注关键词。然后,利用字符串匹配算法从语料库中列出包含所有标注关键词的句子,并将得到的句子通过Sentence-Rank算法进行价值排序,选取评分最高的句子描述图像。实验结果表明,该方法得到的标注句子具有较低的困惑度,较好地解决了句子的语言逻辑问题。

关键词:机器学习;自然语言处理;特征融合;Sentence-Rank; N -gram
文献标志码:A **中图分类号:**TP391 doi:10.3778/j.issn.1002-8331.1709-0422

徐守坤,徐坚,李宁,等.基于Sentence-Rank的图像句子标注.计算机工程与应用,2019,55(2):121-127.
XU Shoukun, XU Jian, LI Ning, et al. Image sentence annotation based on Sentence-Rank algorithm. Computer Engineering and Applications, 2019, 55(2):121-127.

Image Sentence Annotation Based on Sentence-Rank Algorithm

XU Shoukun[1], XU Jian[1], LI Ning[1,2], ZHOU Jia[1], LIU Chuqiu[3]
1.College of Information Science and Engineering, School of Mathematics and Physics, Changzhou University, Changzhou, Jiangsu 213164, China
2.Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou 350108, China
3.School of Electrical and Optoelectronic Engineering, Changzhou Institute of Technology, Changzhou, Jiangsu 213032, China

Abstract: In the traditional image semantic sentence annotation, sentence templates are used to describe the content of image. However, it is hard to meet the logic of language using the traditional method. Aiming at this problem, this paper proposes to describe the image content by selecting an optimal sentence from the corpus , and design the Sentence-Rank algorithm with the N -gram algorithm to generate the annotated sentence. Firstly, the HSV-LBP-HOG fusion feature with the best performance is used for image classification, the image markings are obtained. Then, it uses the string matching algorithm to list all the sentences with the marked keywords from the corpus and sorts the obtained sentences by Sentence-Rank algorithm, and selects the highest rated sentence to describe the image. The experimental results show that the annotation sentence obtained by this method has lower perplexity, and solves the linguistic logic problem of sentences better.

Key words: machine learning; natural language processing; feature fusion; Sentence-Rank; N -gram

基金项目:闽江学院福建省信息处理与智能控制重点实验室开放课题(No.MJUKF201740)。

作者简介:徐守坤(1972—),通讯作者,男,博士,教授,CCF高级会员,研究方向为人工智能、普适计算等;徐坚(1993—),男,硕士生,研究方向为自然语言处理与图像处理;李宁(1974—),男,博士,副教授,研究方向为数据与信息处理;周佳(1991—),女,硕士生,研究方向为自然语言处理与图像处理;刘楚秋(1999—),男,研究方向为自动化。

收稿日期:2017-09-29 **修回日期:**2017-11-28 **文章编号:**1002-8331(2019)02-0121-07
CNKI **网络出版:**2018-04-08, http://kns.cnki.net/kcms/detail/11.2127.TP.20180404.1652.018.html

1 引言

传统图像标注[1]采用单词标注图像,其标注结果能够较好描述图像内容的语义信息,能够满足图像检索和分类的基本需求。但是单词之间缺少语义关联,容易造成标注词歧义。例如,词语“老虎”有多种含义,即食肉动物老虎、高尔夫球手老虎伍兹和武装直升机老虎等。相比单词,句子具有语义信息更为丰富、准确的优点。因此,句子标注比单词标注能更准确、更全面地描述图像内容,并且能够有效降低标注歧义。此外,图像句子标注应用领域比起单词标注更加广泛,不仅仅是图像检索分类,还有盲人视觉辅助[2]、自动驾驶[3]等,这些系统的研发和应用定会为未来人类生活造成极大的便利,因此句子标注已成为图像标注的研究热点。

目前图像自动句子标注的方法分为两大类:一是基于统计分类的图像句子标注方法;二是基于概率模型的图像句子标注方法。基于统计分类的图像句子标注方法是将图像中的每一个语义标签看作一个分类,即图像语义标注问题转化为图像分类问题。代表方法有基于支持向量机(Support Vector Machine,SVM)的方法[4-5]、K 最近邻(K-Nearest Neighbor,KNN)分类方法[6-7]、基于决策树的方法[8-9]和基于Adaboost的方法[10]等,Zhang[11]通过梯度核特征识别出图像中的标注词,利用句子模板将标注词拼装成描述商品图像的句子,在商品图像句子标注方面取得了较好的效果。基于概率模型的图像句子标注方法尝试推断图像和标注句子之间的相关性或联合概率分布。代表方法有Duygulu[12]等和Ballan[13]等提出的机器翻译模型以及主题相关模型等,Hodosh[14]、Li[15]在中间语义空间分析图像和文本的相关性,为图像检索最匹配的标注句子。Kiros[16]采用CNN(Convolutional Neural Network)对图像做深度学习,抽取其图像特征,然后在深度学习模型MLBL(Modality-Biased Log-Bilinear)内分析词向量和图像特征的跨模态相关性,基于相关性优选匹配单词生成句子。

本文采用统计分类方法获得图像标注关键词,将标注关键词通过语料库查询匹配生成最优的标注句子。在关键词生成阶段,利用特征融合方法来保证识别精度,并识别图像中用于消歧的关键词。例如图片中有“老虎”关键词,“老虎”有动物老虎、老虎伍兹和老虎直升机等多重含义,如果还能识别出关键词“高尔夫”,那么老虎就能确定为高尔夫球手老虎伍兹。在句子生成阶段,将语料库中最优的句子作为图像的句子标注,保证了图像句子标注的可读性。本文标注方法的主要工作有:(1)抽取图像的轮廓、纹理和颜色特征;(2)将图像的轮廓、纹理和颜色特征融合,作为新的特征进行学习;(3)通过语料库筛选预备句子,并根据两种情况分别讨论;(4)设计Sentence-Rank[17]评价方法,进一步生成评价最优的句子作为图像描述,完成图像句子标注。

2 特征融合

机器学习的优点有:基于机器学习的方法理论更为成熟;对数据量和实验硬件无太高要求。机器学习中特征学习分为两步:先抽取图像某个特征的特征向量,再利用特征向量训练出分类器。现阶段有着各种不同的图像特征可以表征图像,其中颜色、形状、纹理三大特征最为常用,并且在反映图像时各有优势,但只使用某一特征来反映图像内容过于局限,造成识别效果较差。本文方法是利用最佳句子标注图像,而句子的优劣与关键词息息相关,关键词的获得取决于识别精度,因此识别效果好坏会影响到句子标注性能,本文提出用线性融合特征来表征图像以提高识别精度。

2.1 提取HSV特征

首先提取图像颜色特征的特征向量,本文使用颜色直方图(Hue-Saturation-Value,HSV)方法来提取图像的颜色特征,Swain和Ballard[18]最先使用颜色直方图作为图像特征,颜色直方图具有旋转不变性等优势。

HSV的三个分量分别代表色彩(Hue)、饱和度(Saturation)和值(Value)。通过公式(1)计算可以得到相应的 H、S、V:

$$\begin{cases} H=\begin{cases} 0, max=min \\ 60\times\dfrac{G-B}{max-min}, max=R且G\geqslant B \\ 60\times\dfrac{G-B}{max-min}+360, max=R且G<B \\ 60\times\dfrac{B-R}{max-min}+120, max=G \\ 60\times\dfrac{R-G}{max-min}+240, max=B \end{cases} \\ S=\begin{cases} 0, max=0 \\ 1-\dfrac{min}{max}, max\neq 0 \end{cases} \\ V=max \end{cases} \tag{1}$$

其中,max 为图像像素中的最大值,min 为图像像素的最小值。其次根据人眼视觉对颜色的感知特性,对HSV三个分量进行非等间隔量化,H 量化成16级,S 和 V 均量化为4级,然后依据公式(2)将三个颜色分量表示成一维特征矢量:

$$L=Q_S Q_V H+Q_V S+V \tag{2}$$

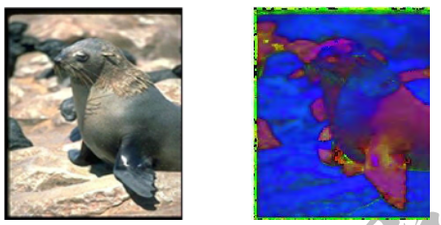
式中 Q_S、Q_V 分别表示 S 和 V 的量化级,将对应值代入,即表示成公式(3),L 即HSV的特征向量:

$$L=16H+4S+V \tag{3}$$

利用matlab可将待检测图像转换为HSV图像,如图1所示。

2.2 提取LBP特征

本文使用局部二值模式[19](Local Binary Patterns,LBP)来提取图像的纹理特征,该方法的优点是它具有旋转不变性和灰度不变性,但当图像的像素分辨率变化明显时,得到的纹理特征偏差就会明显增大。



(a)原始图像　(b)HSV空间图像

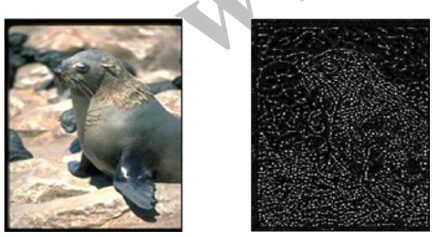
图1　原始图像与HSV空间图像

其中，对LBP特征的计算如公式(4)所示：

$$LBP_{P,R}(x_c,y_c)=\sum_{p=0}^{P-1}s(g_p-g_c)2^p$$

$$s(x)=\begin{cases}1, x\geqslant 0\\ 0, x<0\end{cases} \tag{4}$$

式中，P 代表领域个数，R 代表领域大小，g_p 为领域像素点，g_c 为中心像素点。利用matlab可将待检测图像转换为LBP特征图像，如图2所示。



(a)原始图像　(b)LBP图像

图2　原始图像与LBP图像

2.3　提取HOG特征

将本文使用方向梯度直方图(Histogram of Oriented Gradient，HOG)[20]方法来提取图像的形状特征。HOG算法是法国研究人员Dalal[20]在2005的CVPR上提出的，其优点是保持了几何和光学转化不变性，但其一较严重缺点就是不能解决遮挡问题。HOG的梯度计算公式如公式(5)所示：

$$\begin{cases}G_x(x,y)=H(x+1,y)-H(x-1,y)\\ G_y(x,y)=H(x,y+1)-H(x,y-1)\end{cases} \tag{5}$$

式中 G_x、G_y、H 分别表示输入图像中像素点 (x,y) 处的水平方向梯度、垂直方向梯度和像素值。像素点 (x,y) 处的梯度幅值和梯度方向计算公式如公式(6)所示：

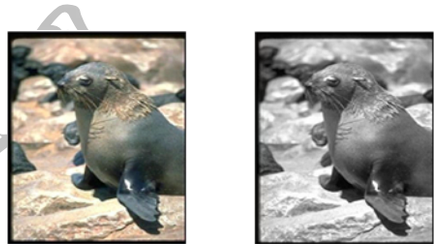
$$G(x,y)=\sqrt{G_x(x,y)^2+G_y(x,y)^2}$$

$$\alpha(x,y)=\tan^{-1}\left(\frac{G_y(x,y)}{G_x(x,y)}\right) \tag{6}$$

利用matlab可将待检测图像转换为HOG灰度图像，如图3所示。

2.4　特征归一化并融合

由于HSV特征提取的是图像的全局信息，而LBP特征提取的是图像的局部信息，HOG特征对图像几何的和光学的形变都能保持很好的不变性，这样本文提出的HSV-LBP-HOG融合特征更具鲁棒性。由于融合后的特征向量维数过高，使得计算难度较大，计算时间较久，因此在特征融合前将三种图像特征进行归一化。其表达式如公式(7)所示。式中 x 为当前变量，y 为 x 归一化后的变量，Min$Value$ 和 Max$Value$ 分别为最小变量和最大变量。



(a)原始图像　(b)HOG灰度图像

图3　原始图像与HOG灰度图像

$$y=(x-\mathrm{Min}Value)/(\mathrm{Max}Value-\mathrm{Min}Value) \tag{7}$$

再将归一化后的三类特征进行线性级联，得到新的融合特征。首先，提取图像的HSV特征直方图，再求取该图像中所有像素点的LBP特征和HOG特征，然后串接成一个直方图，即得到该图的LBP特征和HOG特征；最后，将HSV特征、LBP特征和HOG特征串接在一起形成HSV-LBP-HOG融合特征。其特征式如式(8)所示：

$$F_{\mathrm{HSV\text{-}LBP\text{-}HOG}}(I)=F_{\mathrm{HSV}}(I)+F_{\mathrm{LBP}}(I)+F_{\mathrm{HOG}}(I) \tag{8}$$

其中，I 表示图像，$F_{\mathrm{HSV}}(I)$ 表示图像 I 的HSV特征，$F_{\mathrm{LBP}}(I)$ 表示图像 I 的LBP特征，$F_{\mathrm{HOG}}(I)$ 表示图像 I 的HOG特征，将融合后的特征用于图像识别，大大缩小了特征量的值，从而减少了分类时间，并得到图像语义关键词集Lable。特征提取示意图如图4所示。

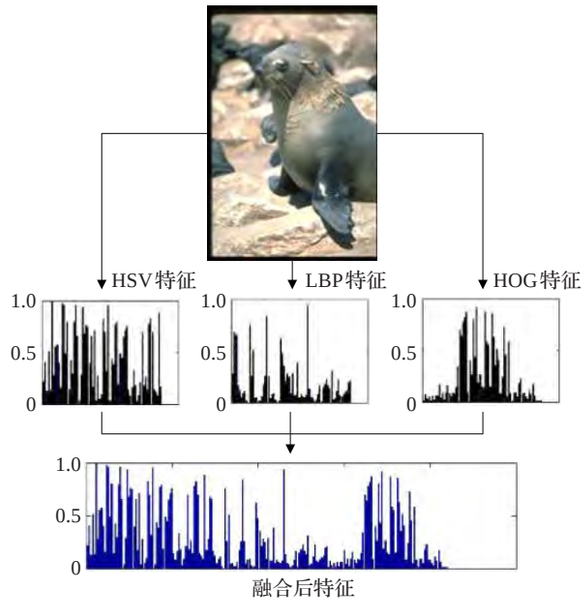


图4　特征提取示意图

3　生成句子

本文提出Sentence-Rank算法对语料库中的句子进

行排序,并将评分最高的句子作为图像标注。Sentence-Rank算法包括句子筛选(Sentence Selection)和句子评分(Sentence Scoring)两部分。句子筛选获取包含图像关键词描述的句子,即初步筛选出与标注关键词相关的句子,句子评分根据自然语言处理方法对筛选出的句子进行评分排序,最后选择评分最高的句子标注图像。本文从语料库中抽取标注句子有效地避免了句子模板方法造成的句子逻辑错误,并最高效地描述图像内容。

3.1 句子筛选

基于特征融合识别出图像语义关键词集存入Lable文件中并作为匹配目标,应用字符串匹配算法(KMP)对语料库中句子进行匹配查询。查询结果分为匹配到包含所有关键词的句子和匹配不到包含所有关键词的句子两种情况。句子筛选阶段的框架图如图5所示。

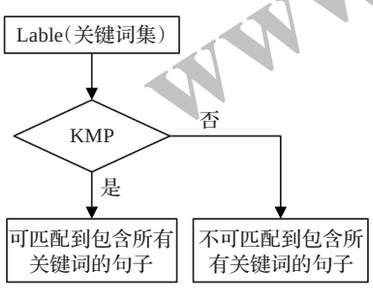


图5 句子筛选阶段流程图

本文利用基于自然语言处理的方法处理匹配不到包含所有关键词句子情况,首先对语料库中各句子进行词频统计,然后去除Lable中词频最低的关键词,生成新的关键词集文本Lable-1,并作为匹配目标对语料库中句子再查询,若依旧无法匹配到句子,则再去除新文本Lable-1中词频最低的关键词,再进行句子匹配查询,直到匹配到句子。去除词频最低关键词的方法是自然语言处理中常用方法,旨在将这一不常见的词替换成与之相似的一个常见词,以利于句子的查询匹配。

3.2 句子评分

在句子评分阶段,对筛选出的句子进行句子评分,再将评分最高的句子作为图像的标注。首次对语料库中句子匹配查询时会出现两种情况,即匹配到包含所有关键词的句子和匹配不到包含所有关键词的句子。Sentence-Rank算法根据以上两种情况分为两部分,针对第一种情况运用基于 N -gram算法中的句子价值评估方法进行句子评分,并组成Sentence-Rank算法的第一部分;针对第二种情况运用词向量余弦值加权方法进行句子评分,并组成Sentence-Rank算法的第二部分。

3.2.1 匹配到包含所有关键词的句子

N -gram算法模型是基于这样一种假设,句子中第 N 个词的出现只与前 $N-1$ 个词相关,而与其他词都不相关,而句子的概率就是各个词出现概率的乘积。这些词概率可以通过统计语料库中 N 个词同时出现的次数得到。本文提出通过 N -gram算法得到的句子概率作为标注句子的评分标准,旨在利用最常用最合常理的句子来标注图像。

具体方法分为以下几步。首先,对语料库和匹配到的句子分词,假设匹配到的句子 (T) 分词后为 <s> $W_1\ W_2\ W_3\cdots W_n$ </s>,其中<s>为句子开始标识,$W_1\ W_2\ W_3\cdots W_n$ 为组成句子 T 的单词,</s>为句子结束标识。句子 T 的概率值推导如式(9)所示:

$$\begin{aligned}&P(W_2|W_1)=P(W_1W_2)/P(W_1)\\&P(W_1W_2)=P(W_1)P(W_2|W_1)\\&P(T)=P(W_1W_2W_3\cdots W_n)=\\&\quad P(W_1)P(W_2|W_1)P(W_3|W_1W_2)\cdots P(W_n|\\&\quad W_1W_2\cdots W_{n-1})\end{aligned}\tag{9}$$

这里引入一种简单的估计方法(最大似然估计法)来算出 $P(W_n|W_1W_2\cdots W_{n-1})$,即如公式(10)所示:

$$\begin{aligned}&P(W_n|W_1W_2\cdots W_{n-1})=\\&\quad C(W_1W_2W_3\cdots W_n)/C(W_1W_2W_3\cdots W_{n-1})\end{aligned}\tag{10}$$

$C(W_1W_2W_3\cdots W_n)$ 为语料库中 $W_1W_2W_3\cdots W_n$ 同时出现的概率,这样就可以通过计算词序列同时出现次数算出句子 T 的概率。在实验中,可能会出现数据稀疏情况,即此序列出现次数为0,导致整条句子概率为0,本文通过"平滑技术"(Smoothing)方法来解决这一问题。其主要策略是把在训练样本中出现过的事件的概率适当减小,然后把减小得到的概率密度分配给训练语料库中没有出现的事件。Add-one(Laplace) Smoothing是最简单、最直观的一种平滑算法。本文规定任何一个 N -gram在训练语料时至少出现一次,即 N -gram的概率不为0,则第 n 个词出现概率为式(11)所示:

$$\begin{aligned}&Padd_1(W_n|W_1W_2\cdots W_{n-1})=\\&\quad C(W_1W_2\cdots W_n)+\frac{1}{C(W_1W_2\cdots W_{n-1})}+|V|\end{aligned}\tag{11}$$

其中 V 是所有可能的不同的 N -gram的数量,由此可得到句子 T 的评分。

那么,句子 T 的Sentence-Rank值为式(12)所示:

$$\begin{aligned}&R(T)=P(T)=\\&\quad Padd_1(W_1)Padd_1(W_2|W_1)\cdots Padd_1(W_n|\\&\quad W_1W_2\cdots W_{n-1})\end{aligned}\tag{12}$$

3.2.2 匹配不到包含所有关键词的句子

本文方法的原理是从语料库中抽取句子并评分排序,将评分最高的句子作为图像的标注。因此,在出现匹配不到包含所有关键词的句子情况时,本文采用自然语言处理中常用的词替换方法,原理是首先对标注关键词进行词频统计,将词频最低的关键词从Lable中去除,并生成新的关键词集Lable-1,再将新关键词集作为输入对语料库中句子匹配查询,如果还是不能匹配到句子,则继续去除新关键词集中词频最低的关键词,再

作为输入进行匹配查询，直到匹配到句子，否则输出ERROR。流程框图如图6所示。

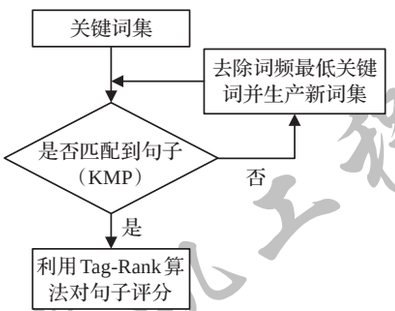


图6　句子匹配阶段流程图

首先，基于word2vec[21]工具通过语料库训练出词向量，词向量间的余弦值代表词之间的相似度。假设匹配到的句子为 T，被去除的关键词为 $a_1, a_2, \cdots, a_n$，可获得句子 T 中与被去除关键词最相似的词 $b_1, b_2, \cdots, b_n$，则余弦值相似度为 $\cos(a_1, b_1), \cos(a_2, b_2), \cdots, \cos(a_n, b_n)$。词频作为单词重要性与常用性的判断，词频高的单词越重要越常用。本文将词频作为词向量余弦的权值，作为理想标注句子相似度判断标准旨在将最常用最合常理的句子作为图像的描述，因此计算出词 $b_1, b_2, \cdots, b_n$ 的词频为 $f_1, f_2, \cdots, f_n$，则匹配句子与理想标注句子的相似度计算公式如公式(13)所示：

$$S(T)=f_1\times\cos(a_1,b_1)+f_2\times\cos(a_2,b_2)+\cdots+f_n\times\cos(a_n,b_n) \tag{13}$$

最后，将句子的 N -gram评分作为句子 T 的权值，得到句子的Sentence-Rank值为公式(14)所示：

$$R(T)=S(T)\times P(T) \tag{14}$$

4　实验及结果分析

4.1　图像数据集和语料库

为了验证本文算法的标注性能，采用文献[22]使用的Corel 5k图像库进行测试。Corel 5k数据集中包含5 000幅图像，其中包括4 500个训练样本和500个测试样本。每一幅图像平均有3.5个关键词。Corel 5k共有371个关键词，将至少标注了8幅图像的关键词选入词汇表，合计260个关键词。

本文采用搜狗实验室互联网语料库(SogouT)[23]为图像生成标注句子。SogouT包含了来自互联网各种类型的1.3亿个原始网页，语句数量足以满足实验需求。

本文实验标注方法分为两部分，即关键词生成阶段和句子生成阶段。

4.2　关键词生成阶段实验

4.2.1　实验设置

关键词生成过程属于传统图像识别过程，而识别效果好坏直接影响到句子标注的性能，因此关键词生成阶段的识别效果尤为重要，识别精度越高，标注性能越好。首先采用召回率和查准率两种方法作为关键词生成阶段的评估标准。假设一个给定的标签的图像数量为 N_1，N_2 为有正确标注词的图像数量，N_3 为由图像标注方法得到标签的图像的数量。召回率(*recall*)和查准率(*precision*)的计算公式如公式(15)所示。

$$recall=\frac{|N_2|}{|N_1|},precision=\frac{|N_2|}{|N_3|} \tag{15}$$

在本文实验中，将本文多特征融合标注方法与单特征标注方法以及其他特征融合方法进行比较，将召回率和查准率作为评价标准，验证本文方法的可行性。

4.2.2　实验结果分析

为了说明本文特征(HSV-LBP-HOG)融合方法的优越性，本文设计了四组对比实验分别是LBP特征、HOG特征、HSV-LBP特征和本文方法。在相同的实验环境条件下，将各个特征经过Adaboost进行训练识别，其实验结果如表1所示。

表1　不同特征算法的实验对比

方法	召回率R	查准率P	单幅样本提取时间/s
LBP	0.27	0.31	0.010 0
HOG	0.41	0.35	0.095 0
HSV-LBP	0.46	0.41	0.041 0
CNN	0.65	0.42	0.320 0
本文方法	0.59	0.48	0.116 0

从表1中可以看出，本文算法的召回率和查准率优于其他四种算法，比较HSV-LBP算法，召回率提高了28.3%，查准率提高了17.1%。比较CNN算法，召回率稍低，但查准率与样本提取时间均优于CNN。本文算法识别时间稍多于前三种算法，其原因是融合后的特征维数增加，但是这种融合后的特征能更好地对图像进行表征，且改善了单一特征的局限性。

4.3　句子生成阶段实验

4.3.1　实验设置

N -gram模型是Sentence-Rank算法的重要组成部分，选取不同元的 N -gram模型会获得不同的标注性能。在句子生成阶段，应用BLEU[24](BiLingual Evaluation Uderstudy)一元模型评分方法来比较Sentence-Rank算法中的不同 N -gram模型对标注性能的影响。BLEU算法常用于判断机器翻译结果的优劣，本文引入BLEU算法，旨在分析各 N -gram模型对标注结果的影响，并分析出最佳的 N -gram模型。首先抽取出50组图像作为实验对象，每幅图像做两条人工句子标注作为原标注。抽取50组图像的关键词集，在相同的实验环境下，每组关键词集再通过4种不同元的 N -gram算法来生成句子标注，并通过两条人工句子标注获得BLEU的平均得分。

句子标注性能的判定主要分为句子准确性和句子

可读性两方面。标注句子的主要组成部分是关键词生成阶段识别出的关键词,而关键词生成阶段的实验结果表明本文方法识别效果很好,即关键词的准确性很好,因此也保证了标注句子的准确性。困惑度[25]是统计机器学习翻译系统中的一种评价指标,采用 N -gram 语言模型,判断机器翻译给出的译文是否是一个合理的句子。本文利用困惑度作为评价指标检测 Sentence-Rank 算法模型,旨在检验标注句子是否有很好的可读性。本文抽取 50 组图像作为实验对象比较本文方法与句子模板方法[2]的性能,算出两种方法的平均困惑度值,困惑度的公式如式(16)所示。式中 K 为句子 T 的字数,$P(W_i|W_{i-2}W_{i-1})$ 为 W_i 的 3-gram 值。

$$PP(T)=\left(\prod_{i=1}^{K}P(W_i|W_{i-2}W_{i-1})\right)^{-\frac{1}{K}} \tag{16}$$

4.3.2 实验结果分析

本文抽取 50 组图像作为实验对象,并对每组图像计算 4 种 N -gram 算法的 BLEU-1 评分。BLEU 算法对各个 N -gram 算法的评分如表 2 所示。

表 2 各 *N*-gram 模型的平均 BLEU-1 评分

评分	1-gram	2-gram	3-gram	4-gram
BLEU-1	0.037 0	0.050 6	0.052 4	0.051 5

由表 2 可看出,当 N -gram 为三元模型(3-gram)时,标注的 BLEU 评分最高即标注性能最好。本文选取 3-gram 模型来构成 Sentence-Rank 算法。

本文方法与句子模板方法的平均困惑度值对比如表 3 所示。其中 TOP10 为 50 组实验对象困惑度值中最小 10 组的平均值,MIDDLE10 为困惑度值中间 10 组的平均值,BOTTOM10 为困惑度值最大 10 组的平均值。

表 3 平均困惑度值对比

方法	TOP10	MIDDLE10	BOTTOM10
句子模板方法	127.4	158.2	161.2
本文方法	91.3	107.5	133.8

从表中可看出,本文方法的各阶段平均困惑度值都比句子模板法的低,而困惑度值越低,句子越合理,说明了本文方法生成的标注句子更合理、更具可读性。

4.4 标注结果展示

部分句子标注结果与关键词标注结果如表 4 所示。

在表 4 中,图像 1、2、3、5 和 6 的本文预测较为准确,基本可以正确描述出图像的内容,证明了本文方法的可行性。其中图像 3 首次匹配时没有查询到句子,所以去除了词频最低的关键词“冲浪板”,并用“冲浪”替换“冲浪板”。图像 4 的正确描述为“一只麋鹿在河上前行”,本文预测的句子描述为“一只麋鹿在河边喝水”,其预测不准确的原因是没有识别出图像各目标间的空间关系和交互动作,如马嘴与水的空间关系,马与河的交互动作。

表 4 部分标注结果

图像	本文预测(关键词)	原标注(关键词)	本文预测(句子)	人工标注(句子)
	飞机 天空	飞机 天空 云	飞机飞过 天空	一架飞机翱翔在天空
	人 笛子	喋嘛 笛子 蓝天 草地	一个人在 吹笛子	一名喇嘛在 吹笛子
	人 冲浪板 海	男人 冲浪板 大海	一个人在 海上冲浪	一名男子在 海上冲浪
	麋鹿 河	麋鹿 河 水草	一只麋鹿在 河边喝水	一只麋鹿在 河上前行
	人 泳池	运动员 泳池	一个人在泳 池游泳	运动员在进 行游泳比赛
	马 草地	马 草地	马在吃草	两匹马在草 地上吃草

5 结语

在机器视觉特征学习的基础上,基于特征融合提取描述图像内容的关键词,利用 Sentence-Rank 算法从语料库中抽取最优的标注句子作为图像标注。本文在经典、常用的图像数据集 Corel 5k 上进行了定量的验证,实验表明:(1)融合后的图像特征(HSV-LBP-HOG)比单一图像特征有更好的识别效果,并使得标注性能更优;(2)通过 Sentence-Rank 算法所生成的标注句子可读性好,能够准确地刻画图像内容;(3)标注结果受到识别目标间的空间关系和交互动作影响,值得深入研究来获得更完善的句子标注模型。

未来工作:尝试使用更大规模的数据集进行测试,并选择具备更多噪声的图像数据集,因为图像是可能存在遮挡、变形等噪声的,根据所得到的结果进行相应的改进,进一步提高标注模型的稳定性。对识别目标间的空间关系和交互动作做深入研究,使得标注模型更全面、更客观地描述图像。此外,尝试使用深度学习方法进行识别实验,并与本文 Sentence-Rank 算法相结合,因为深度学习方法是图像识别领域的当前研究热点,有很大的发展空间。

参考文献:

[1] Makadia A,Pavlovic V,Kumar S.A new baseline for image annotation[C]//European Conference on Computer Vision,2008:316-329.

[2] Kulkarni G,Premraj V,Ordonez V,et al.Babytalk:understanding and generating simple image descriptions[C]//

Computer Vision and Pattern Recognition,2011:1601-1608.
[3] Nwogu I,Zhou Y,Brown C.DISCO:describing images using scene contexts and objects[C]//AAAI Conference on Artificial Intelligence,2011:1487-1493.
[4] Yang C,Dong M,Hua J.Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning[J].Journal of Computer Research & Development,2009,46(5):2057-2063.
[5] Murthy V N,Can E F,Manmatha R.A hybrid model for automatic image annotation[C]//International Conference on Multimedia Retrieval,2014:369-376.
[6] Moran S,Lavrenko V.Sparse kernel learning for image annotation[C]//International Conference on Multimedia Retrieval,2014:113.
[7] Verma Y,Jawahar C V.Image annotation using metric learning in semantic neighbourhoods[C]//European Conference on Computer Vision,2012:836-849.
[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] 蒋黎星.基于集成分类算法的自动图像标注[J].自动化学报,2012,38(8):1257-1262.
[10] Li R,Lu J,Zhang Y,et al.Dynamic Adaboost learning with feature selection based on parallel genetic algorithm for image annotation[J].Knowledge-Based Systems,2010,23(3):195-201.
[11] 张红斌,姬东鸿,尹兰,等.基于梯度核特征及 N -gram模型的商品图像句子标注[J].计算机科学,2016,43(5):269-273.
[12] Duygulu P,Barnard K,Freitas J F G D,et al.Object recognition as machine translation:learning a lexicon for a fixed image vocabulary[C]//European Conference on Computer Vision,2002:97-112.
[13] Ballan L,Uricchio T,Seidenari L,et al.A cross-media model for automatic image annotation[C]//International Conference on Multimedia Retrieval,2014:73.
[14] Hodosh M,Young P,Hockenmaier J.Framing image description as a ranking task:data,models and evaluation metrics[J].Journal of Artificial Intelligence Research,2013,47(1):853-899.
[15] Li P,Ma J,Gao S.Learning to summarize web image and text mutually[C]//ACM International Conference on Multimedia Retrieval,2012:28.
[16] Kiros R,Salakhutdinov R,Zemel R.Multimodal neural language models[C]//International Conference on International Conference on Machine Learning,2014.
[17] Zhang H,Ji D,Lan Y,et al.Product image sentence annotation based on kernel descriptors and tag-rank[J].Journal of Southeast University(English Edition),2016.
[18] Swain M,Ballard D H.Color indexing[J].International Journal of Computer Vision,1991,7(1):11-32.
[19] Ojala T,Harwood I.A comparative study of texture measures with classification based on feature distributions[J]. Pattern Recognition,1996,29(1):51-59.
[20] Dalal N,Triggs B.Histograms of oriented gradients for human detection[C]//CVPR,2005:886-893.
[21] Goldberg Y,Levy O.word2vec Explained:deriving Mikolov et al.'s negative-sampling word-embedding method[J]. Eprint Arxiv,2014.
[22] Duygulu P,Barnard K,Freitas J F G D,et al.Object recognition as machine translation:learning a lexicon for a fixed image vocabulary[C]//European Conference on Computer Vision,2002:97-112.
[23] Tahara T,Sogou T,Suezawa C,et al.Filtered QRS duration on signal-averaged electrocardiography correlates with ventricular dyssynchrony assessed by tissue Doppler imaging in patients with reduced ventricular ejection fraction[J].Journal of Electrocardiology,2010,43(1):48-53.
[24] Papineni K,Roukos S,Ward T,et al.BLEU:a method for automatic evaluation of machine translation[C]//Meeting on Association for Computational Linguistics,2002:311-318.
[25] Wang M,Lu Z,Li H,et al.genCNN:a convolutional architecture for word sequence prediction[J].Computer Science,2015.