基于梯度核特征及N-gram模型的商品图像句子标注

张红斌 姬东鸿 尹 兰 任亚峰
(武汉大学计算机学院 武汉 430072)

摘 要 提出为商品图像标注句子,以便更准确地刻画图像内容。首先,执行图像特征学习,选出标注性能最优的梯度核特征完成图像分类和图像检索,该特征能客观描绘商品图像中形状和纹理这两类关键视觉特性。然后,基于语义相关度计算结果从训练图像的文本描述中摘取关键单词,并采用N-gram模型把单词组装为蕴涵丰富语义信息且满足句法模式兼容性的修饰性短语,基于句子模板和修饰性短语生成句子。最后,构建Boosting模型,从若干标注结果中选取BLEU-3评分最优的句子标注商品图像。结果表明,Boosting模型的标注性能优于各基线。

关键词 梯度核特征,N-gram模型,商品图像,句子标注,语义相关度计算,修饰性短语
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.5.051

Product Image Sentence Annotation Based on Gradient Kernel Feature and N-gram Model

ZHANG Hong-bin JI Dong-hong YIN Lan REN Ya-feng
(Computer School, Wuhan University, Wuhan 430072, China)

Abstract Product image sentence annotation was presented because sentence describes online products more accurately than single words. Firstly, image feature learning was executed. Gradient kernel feature that achieves the best annotation performance was chosen because the feature describes the key visual characteristics of product image such as shape and texture better than other features. Therefore, the gradient kernel feature was selected to complete image classification and image retrieval. Secondly, several key words were summarized from training images' captions based on semantic correlation computing. Thirdly, a modified sequence that not only contains rich semantic information but also satisfies syntactic mode compatibility was created based on these key words by N-gram model. Sentence was generated according to predefined sentence template and the modified sequence. Finally, a Boosting model was designed to choose those sentences that obtain the best BLEU-3 scores to annotate product images. Experiments show sentences generated by the boosting model achieve the state of art annotation performances.

Keywords Gradient kernel feature, N-gram model, Product image, Sentence annotation, Semantic correlation computing, Modified sequence

1 引言

传统图像标注[1]采用单词描绘图像,标注结果能较好地反映图像中蕴涵的语义信息。但单词之间缺少语义关联,实际标注中噪声和歧义等问题较严重。相比单词,句子基于语义相关性和句法模式兼容性把独立的语义单元(单词)组装起来,它是一种粒度适中、语义信息丰富、句法结构准确的组合语义。因此,句子标注比单词标注能更准确、全面地刻画图像内容。此外,图像句子标注还具备广阔的应用前景,主要应用包括:基于语义的图像检索[2]、盲人视觉感知辅助系统[3]、旅行伴侣机器人[4]和安全巡逻机器人[4]等,这些系统的研发和应用定会为人们的工作和生活带来极大的便利,并创造巨大的经济和社会效益。

相关工作较多,如Yang[2]识别图像中的目标,基于语料库推理目标间的交互关系,根据推理结果采用HMM(Hidden Markov Model)生成句子。Kulkarni[3]识别图像中的目标、场景和属性,他将图像句子标注转换为基于CRF(Conditional Random Field)的"语义标签标注",并采用句子模板或语言模型生成句子。Nwogu[4]基于场景本体对已识别的场景属性做本体演化,结合演化结果和句子模板生成句子。Hodosh[5]、Li[6]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句子。Feng[7]挖掘描述图像和文本的共享隐含主题,基于隐含主题摘取单词或短语,最后运用Beam Search算法生成句子。近年,电子商务的快速发展使商品图像的句子标注成为研究热点。Tamara[8]基于MIL(Multiple Instance Learning)摘取描述商品视觉属性的文本片段。Kiapour[9]用

到稿日期:2015-04-22 返修日期:2015-09-19 本文受国家自然科学基金重点项目(61133012),国家社科重大招标项目(11&ZD189),教育部人文社会科学研究青年项目(12YJCZH274),江西省科技厅科技攻关项目(20142BBG70011,20121BBG70050),江西省高校人文社科基金项目(XW1502,TQ1503)资助。

张红斌(1979—),男,博士生,副教授,CCF会员,主要研究领域为图像标注、自然语言处理及机器学习,E-mail:zhanghongbin@whu.edu.cn;姬东鸿(1967—),男,博士,教授,主要研究方向为自然语言处理、机器学习;尹 兰(1979—),女,博士生,副教授,主要研究方向为自然语言处理、机器学习;任亚峰(1986—),男,博士生,主要研究方向为自然语言处理、机器学习。

文本片段标注服装中蕴含的时尚元素。文本片段[8,9]显然无法完整、准确地刻画商品图像内容。Rebecca[10]基于 Gist 特征检索训练图像，摘取图像标题中的关键文本生成句子标注图像。仅用 Gist 特征做图像检索不能全面、细致地刻画商品图像中的形状和纹理等关键视觉特性。Kiros[11]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)内分析词向量与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。众所周知，CNN 参数极多，易使 MLBL 在反向传播调制模型参数时陷入过拟合，最终影响句子标注性能。

研究动机为：1)图像特征学习是关键，应选择理论基础完备、运行效率优良的特征学习模型；2)商品图像多用修饰性短语刻画，为了充分利用语言学统计信息，选取兼顾语义相关性和句法模式兼容性的 N-gram 模型来构建修饰性短语最为合适。本文采用生成式[2,3,10,11]方法实现商品图像句子标注，主要工作有：1)引入核描述子[12] KDES(Kernel Descriptors)抽取图像局部特征；2)基于高效匹配核 EMK[13] (Efficient Match Kernels)将图像局部特征转换为更紧凑的核特征，完成图像特征学习；3)设计语义相关度计算模型 SCCM(Semantic Correlation Computing Model)摘取刻画图像内容的关键单词；4)基于 N-gram 模型及语言学统计信息将关键单词组装为既蕴涵丰富语义信息又满足句法模式兼容性的修饰性短语，最后根据句子模板和修饰性短语生成句子，完成商品图像句子标注。

2 商品图像句子标注方法

2.1 图像特征学习

相比深度学习[14]和稀疏编码[15]，核特征模型[12,13]的主要优势如下：1)基于匹配核理论，理论基础更完备；2)无需构建完整核矩阵，特征生成方式更简单，特征提取速度也更快；3)能获取有竞争力的判别性能[12]。核特征提取分两步：抽取图像核描述子[12]及基于匹配核[13]把核描述子转换为高效、紧凑的核特征。首先，抽取图像的梯度核描述子：

$$F_{grad}(P)=\sum_{z\in P}\tilde{m}(z)\phi_{orien}(\tilde{\theta}(z))\otimes\phi_{pos}(z)$$
$$\text{subject to } \tilde{\theta}(z)=[\sin(\theta(z))\cos(\theta(z))] \quad (1)$$
$$\tilde{m}(z)=m(z)/\sqrt{\sum_{z\in P}m(z)^2+\varepsilon_g}$$

其中，$\tilde{m}(z)$是像素点 z 的梯度强度 $m(z)$的 L2 规范化值，$\tilde{\theta}(z)$是梯度方向 $\theta(z)$的规范化值，P 指图像块，$\otimes$表示张量积。像素之间梯度方向的相似度及二维空间位置的相似度定义如下：

$$k_{orien}(\tilde{\theta}(z),\tilde{\theta}(z'))=\exp(-\gamma_{orien}\|\tilde{\theta}(z)-\tilde{\theta}(z')\|^2) \quad (2)$$
$$k_{pos}(z,z')=\exp(-\gamma_{pos}\|z-z'\|^2) \quad (3)$$

其中，$\phi_{orien}(\tilde{\theta}(z))$是 $\tilde{\theta}(z)$的核映射，$\phi_{pos}(z)$是对 z 空间位置的核映射。

梯度核描述子从梯度强度、梯度方向和像素点在二维空间中的相对位置这 3 个角度度量像素点之间的相似度，它蕴涵了针对图像中关键纹理特征和形状特征的重要判别信息。同理，导入颜色核描述子和形状核描述子：

$$F_{color}(P)=\sum_{z\in P}\phi_{color}(c(z))\otimes\phi_{pos}(z) \quad (4)$$
$$F_{shape}(P)=\sum_{z\in P}\tilde{s}(z)\phi_{shape}(b(z))\otimes\phi_{pos}(z)$$
$$\text{subject to } \tilde{s}(z)=s(z)/\sqrt{\sum_{z\in P}s(z)^2+\varepsilon_s} \quad (5)$$

$k_{color}(c(z),c(z'))=\exp(-\gamma_{color}\|c(z)-c(z')\|^2)$，$c(z)$是像素点 z 在的颜色值，$\phi_{color}(c(z))$是 $c(z)$的核映射，像素点 z 二维空间位置的核映射 $\phi_{pos}(z)$同梯度核描述子中的定义。颜色核描述子从颜色值和像素点在二维空间中的相对位置这两个角度度量像素之间的相似度，它蕴涵了针对图像中关键颜色特征的重要判别信息。同理，还得到像素点形状的相似度度量：$k_{shape}(b(z),b(z'))=\exp(-\gamma_{shape}\|b(z)-b(z')\|^2)$，$b(z)$是像素点 z 的 LBP(Local Binary Patterns)值，$\phi_{shape}(b(z))$是 $b(z)$的核映射，像素点 z 二维空间位置的核映射 $\phi_{pos}(z)$同梯度核描述子中的定义，$\tilde{s}(z)$是像素点 z 的 3×3 邻域内颜色值的规范化值。形状核描述子从 LBP 值和像素点在二维空间中的相对位置这两个角度度量像素之间的相似度，它蕴涵了针对图像中关键形状特征的重要判别信息。

由于 $\phi_{orien}(\tilde{\theta}(z))$、$\phi_{color}(c(z))$、$\phi_{shape}(b(z))$及 $\phi_{pos}(z)$均为高维空间中的特征描述，考虑到模型运行效率，需把它们映射为低维、高效的核特征。以 $\phi_{orien}(\tilde{\theta}(z))$为例，基于 EMK[13]设计基特征$\{\phi_{orien}(x_i)\}_{i=1}^{dim_{orien}}$映射 $\phi_{orien}(\tilde{\theta}(z))$的低维描述$\tilde{\phi}(\tilde{\theta}(z))$。最终，式(1)的梯度核描述子 $F_{grad}(P)$ 被转换为梯度核特征(Grad-KDES)：

$$\tilde{F}_{grad}(P)=\sum_{z\in P}\tilde{m}(z)\tilde{\phi}_{orien}(\tilde{\theta}(z))\otimes\tilde{\phi}_{pos}(z) \quad (6)$$

$\tilde{\phi}_{orien}(\tilde{\theta}(z))=GK_{orien}(\tilde{\theta}(z),H)$，$H^TH$ 被记作 K_{orien}，$\{K_{orien}\}_{ij}=\kappa(x_i,x_j)$，对 K_{orien} 执行乔里斯基分解：$G^TG=K_{orien}^{-1}$，H 是由 k-means 聚类得到的梯度核描述子词典，$H=[\phi_{orien}(x_1),\cdots,\phi_{orien}(x_{dim_{orien}})]$。同理，得到 $\phi_{pos}(z)$的 EMK 特征描述$\tilde{\phi}_{pos}(z)$。此外，颜色核特征(Color-KDES)和形状核特征(Shape-KDES)定义如下：

$$\tilde{F}_{color}(P)=\sum_{z\in P}\tilde{\phi}_{color}(c(z))\otimes\tilde{\phi}_{pos}(z) \quad (7)$$
$$\tilde{F}_{shape}(P)=\sum_{z\in P}\tilde{s}(z)\tilde{\phi}_{shape}(b(z))\otimes\tilde{\phi}_{pos}(z) \quad (8)$$

2.2 生成句子

生成式句子标注方法需设计 NLG(Natural Language Generation)算法来完成句子生成。NLG 算法包括内容选择(Content Selection)和表层实现(Surface Realization)。内容选择摘取刻画商品图像内容的关键单词，表层实现根据语言学统计信息把单词组装成语义信息丰富、句法模式兼容的修饰性短语，最后基于模板和修饰性短语生成句子，标注商品图像。

2.2.1 内容选择

基于核特征检索出 n'个训练样本 $D_i=\langle I_i,T_i,Y_i\rangle$，$I_i$、$T_i$、$Y_i$ 分别表示商品图像、文本描述和类别标签。训练样本的文本描述构成文档集 W，对 W 中各句子做分词、词性标注、去除停留词，生成新文档集 W'。内容选择指从 W'中摘取与测试图像 I_q 语义相关的 K 个单词$\{wrd_1,\cdots,wrd_K\}$。因此，定义语义相关度计算模型 SCCM 如下：

$$p(wrd_j|I_q)=\log_2(\sum_i p(wrd_j|I_i)p(I_i|I_q)) \quad (9)$$

$p(wrd_j|I_q)$度量单词 wrd_j 与 I_q 的语义相关度，其值越大则单词 wrd_j 与 I_q 的图像内容越相关。式(9)中 $p(wrd_j|I_i)$

计算单词 wrd_j 与训练图像 I_i 的语义相关度，它被归一化为：

$$p(wrd_j \mid I_i)=\frac{tf\text{-}idf(wrd_j)}{\sum_{wrd_j' \in W'} tf\text{-}idf(wrd_j')} \tag{10}$$

$tf\text{-}idf$ 函数计算单词 wrd_j 在 W' 中的 $tf\text{-}idf$ 值，$p(wrd_j \mid I_i)$ 与 $tf\text{-}idf$ 值成正比，$tf\text{-}idf$ 值越大，单词 wrd_j 越重要，$p(wrd_j \mid I_i)$ 确保单词 wrd_j 与训练图像 I_i 语义相关。式(9)中 $p(I_i \mid I_q)$ 计算测试图像 I_q 与训练图像 I_i 的视觉相似度，它被归一化为：

$$p(I_i \mid I_q)=\frac{\exp(-1\times dist(I_i \mid I_q))}{\sum_{\ddot{u}} \exp(-1\times dist(I_{\ddot{u}} \mid I_q))} \tag{11}$$

$dist$ 函数基于核特征计算图像间的视觉相似度，$p(I_i \mid I_q)$ 与 $dist$ 函数值成反比，$dist$ 函数值越小，两幅图像越相似。$p(I_i \mid I_q)$ 确保单词 wrd_j 尽可能出现在与 I_q 视觉相似的训练图像中。

2.2.2 表层实现

电子商务网站多用修饰性短语刻画商品图像内容，例如“silk satin”、“magnetic closure”、“removable adjustable strap”、“fully bead vintage bag”等，这些短语由多个形容词或名词根据语义相关性及句法模式兼容性组合在一起。单词间的语义相关性依赖于它们之间的共现统计关系，而句法模式兼容性依赖于单词间的词序先后关系，它们都是完成商品图像句子标注的关键因素。显然，N-gram 模型非常适合构建描述商品图像内容的修饰性短语，因为它兼顾了语义相关性和句法模式兼容性。此外，由于无需大规模的参数调制，就复杂性而言，N-gram 模型必优于深度学习模型 MLBL。基于上述分析，选取 N-gram 模型完成修饰性短语生成。以 3-gram 模型为例，设计构建修饰性短语的表层实现模型：

$$p(wrd_1, wrd_2, \cdots, wrd_n)=\prod_{j=1}^{n} p(wrd_j \in seq) \cdot \prod_{j=3}^{n} p(wrd_j \mid wrd_{j-1}, wrd_{j-2}) \tag{12}$$

$$\text{subject to } p(wrd_j \in seq)=p(wrd_j \mid I_q)$$

$$p(wrd_j \mid wrd_{j-1}, wrd_{j-2})=\frac{count(wrd_j, wrd_{j-1}, wrd_{j-2})}{count(wrd_{j-1}, wrd_{j-2})}$$

$p(wrd_j \in seq)$ 计算修饰性短语 seq 中包含单词 wrd_j 的概率，该值由语义相关度计算模型 $p(wrd_j \mid I_q)$ 决定。$\prod_{j=3}^{n} p(wrd_j \mid wrd_{j-1}, wrd_{j-2})$ 是当前短语中的 3 元单词序列 wrd_j，wrd_{j-1}，wrd_{j-2} 的 3-gram 分布，即 3 元语言模型，n 是修饰性短语的长度。$\prod_{j=3}^{n} p(wrd_j \mid wrd_{j-1}, wrd_{j-2})$ 通过挖掘训练语料库 W' 获取，其中 $count(wrd_j, wrd_{j-1}, wrd_{j-2})$ 计算 2 元单词序列 wrd_j，wrd_{j-1}，wrd_{j-2} 中关键词的共现统计值(频率)，$count(wrd_{j-1}, wrd_{j-2})$ 计算 2 元单词序列 wrd_{j-1}，wrd_{j-2} 中关键词的共现统计值(频率)。同理，构造 1-gram、2-gram、4-gram 等多个不同的表层实现模型。若表层实现模型输出多条修饰性短语，则需获取语义相关度最高的修饰性短语。定义短语 seq 与图像 I_q 的语义相关度计算方法如下：

$$p(seq \mid I_q)=\log_2\left(\prod_{j=1}^{n} p(wrd_j \mid I_q)\right) \tag{13}$$

具体实现时，需借助 Beam Search 算法快速筛选出合适的单词，以构建修饰性短语。在得到修饰性短语后，将短语输入到预先定义的句子模板中，生成句子标注商品图像。句子模板由“前缀短语＋修饰性短语＋后置短语”构成，每个部分的具体描述如下。

• 前缀短语：“This is a picture of”，它的作用是保证句子的完整性。

• 修饰性短语：由关键单词组装生成的修饰性短语，它的作用是描述商品图像的核心内容。

• 后置短语：商品图像分类获取的类别标签“Bag label”，它的作用是明确商品所属的类别。

显然，高元 N-gram 模型(3-gram 或 4-gram)的引入能更好地兼顾单词间的语义相关性和句法模式兼容性，从而获取更优的标注性能。因此，综合多个表层实现模型的标注结果，挑选性能最优的句子完成商品图像句子标注更为合理。设计 Boosting 模型：在所有标注结果中选取 BLEU-3[16] 评分最优的句子对商品图像做标注，因为 BLEU-3 能综合评价 NLG 算法的内容选择和表层实现能力(参见第 2.3 节)。

2.3 句子标注评估模型

参考相关工作[5,10,11]的评价方法，对商品图像标注结果做 BLEU(BiLingual Evaluation Understudy)[16] 评分：

$$BLEU=BP\times\exp\left(\sum_{n=1}^{N'} w_n \log_e p_n\right)$$

$$\text{subject to } BP=\begin{cases}1, & c>r\\ e^{(1-r/c)}, & c\leqslant r\end{cases} \tag{14}$$

$$p_n=\frac{\sum_{seq\in s}\min(count_s(seq), \max_r(count_r(seq)))}{\sum_{seq\in s} count_s(seq)}$$

式中，BP(Brevity Penalty)是长度惩罚因子，BLEU 评分高的句子需在长度、语义相关性(“选词”)和句法模式兼容性(词序先后关系)3 方面都与原标注有较好匹配。p_n 表示标注句子 S 的 n 元语法精度，$n=1$ 表示 1 元语法，对应 BLEU 评分称为 BLEU-1；BLEU-2 和 BLEU-3 评分以此类推。$count_s(seq)$ 统计短语 seq 在句子 S 中的频率，$count_r(seq)$ 统计短语 seq 在原标注 r 中的频率，N' 表示 n 元语法的最大基元数，w_n 是加权平均值，$BLEU\in[0,1]$。显然，1 元模型仅需“命中”单词(语义相关性)，而不考虑单词间的词序先后关系(句法模式兼容性)，故 BLEU-1 仅评价 NLG 算法的内容选择能力。相反，2 元模型或 3 元模型既考虑“命中”单词，又兼顾单词间的词序先后关系，它们同时评价 NLG 算法的内容选择和表层实现能力。因此，BLEU-2 和 BLEU-3 评分的实际价值较 BLEU-1 评分的更大。

3 实验结果和分析

3.1 数据集及基线

“Bag”是电商网站的代表性商品，故选取 Attribute Data[8] 中的“包”数据集来评价各模型的标注性能。该数据集包括 5 种商品：“Clutch”、“Hobo”、“Evening”、“Shoulder”和“Totes”，对应样本数分别为 1643、1630、1681、1596 和 1577，每个样本包含一张商品图像和一段文本描述。随机选择 5689(70%)个样本构成训练集，剩余 2438(30%)个样本即为测试集。在图像特征抽取中，纹理特征选取 RGB-Gist[10]，形状特征选取基于 SIFT 的 SP-BoW 和 SIFT-EMK[13]。核特征选取 Grad-KDES[12]、Shape-KDES[12] 和 Color-KDES[12]。

基线包括：基于 1-gram、2-gram、3-gram 和 4-gram 等 N-gram 模型分别构建的 4 个标注模型，分析不同 N-gram 模型的标注性能差异。此外，选取已有工作 Gist-Based[10] 和 ML-BL[11] 作为重要的对比基线。

3.2 图像特征学习对 BLEU 评分的影响

本实验针对不同图像特征，分别执行图像分类、图像检索和句子生成，以判断图像特征选取对标注性能的影响。由于聚焦了图像特征选择，因此在 NLG 阶段未基于 N-gram 模型生成修饰性短语，而是直接将 SCCM 模型输出的 Top K 个单词无序地拼装为一个“简单短语”，并将“简单短语”输入句子模板生成句子，$K=1,\cdots,10$。为每个测试样本分别标注 10 个句子，计算各个 K 位置标注句子的 BLEU-1 和 BLEU-2 评分均值，得到如图 1 和图 2 所示的 BLEU 评分柱状图。

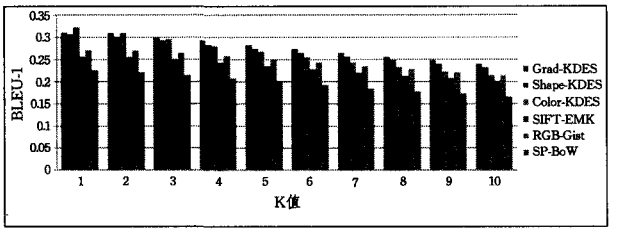

图 1 不同图像特征的 BLEU-1 评分

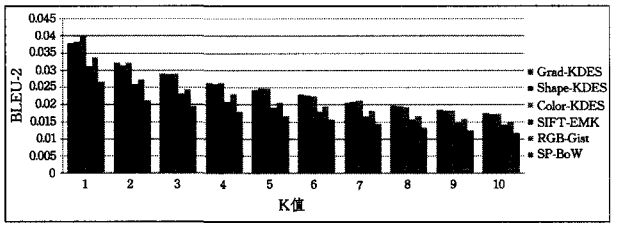

图 2 不同图像特征的 BLEU-2 评分

图 1 中各核特征的 BLEU-1 评分排序：Grad-KDES＞Shape-KDES＞Color-KDES，若采用多个单词（$K>2$）生成“简单短语”，Grad-KDES 的 BLEU-1 评分优势更明显。关键原因有：Grad-KDES 识别图像中的梯度变化，梯度变化能较好地反映商品图像中形状和纹理这两类关键视觉特性。相反，Color-KDES 识别图像中的颜色变化，由于它在同类图像中变化剧烈，因此 Color-KDES 无法准确地刻画图像中最具判别性的视觉特性。此外，人们也倾向于用文字描绘商品的形状和纹理，而 Grad-KDES 与文本单词共现频率高，使得标注句子中有关形状和纹理的文本描述比重更大，BLEU-1 评分自然更优。

图 2 中各核特征的 BLEU-2 评分排序：Grad-KDES＞Shape-KDES＞Color-KDES，这意味着基于 Grad-KDES 所生成的“简单短语”无论是语义相关性还是句法模式兼容性都更胜一筹。相似结果也出现在 BLEU-3 评分柱状图中（略）。其它发现：BLEU 评分随单词数（K 值）的增加而不断衰减。以 Grad-KDES 为例，其 $K=10$ 的 BLEU-1 评分较 $K=1$ 时下降 $(0.3113-0.2396)/0.3113\approx23.03\%$，对应 BLEU-2 评分下降 $(0.0379-0.0176)/0.0379\approx53.57\%$。关键原因有：在“简单短语”生成过程中会引入噪声，从而抑制了 BLEU 评分。此外，未妥善处置单词间的词序先后关系，忽略句法模式兼容性也是制约 BLEU 评分（尤其是 BLEU-2、BLEU-3 评分）的另一关键因素。

综合上述分析，选取 Grad-KDES 特征完成商品图像句子标注。

3.3 语义相关度计算对 BLEU 评分的影响

内容选择的目的是从训练样本中摘取与图像内容语义相关的关键单词（“选词”），为句子生成奠定重要基础。本实验评价 NLG 算法的“选词”能力，设计对比基线 Random：基于 Grad-KDES 特征完成图像分类、图像检索，但不考虑语义相关度计算和表层实现模型，随机抽取训练样本中的 K 个单词生成修饰性短语，$K=1,\cdots,10$。为每个测试样本分别标注 10 个句子，计算各个 K 位置标注句子的 BLEU-1 和 BLEU-2 评分均值，结果如图 3 和图 4 所示。

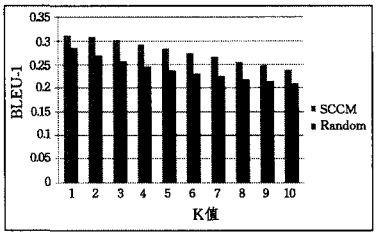

图 3 SCCM 对 BLEU-1 评分的影响

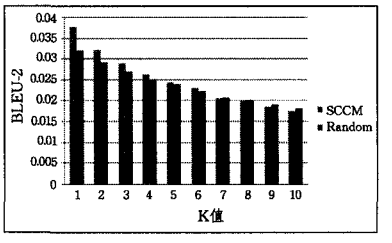

图 4 SCCM 对 BLEU-2 评分的影响

图 3 中 SCCM 模型的 BLEU-1 评分优于 Random 模型，当 $K=1$ 时，SCCM 模型的 BLEU-1 评分比 Random 模型有提升 $(0.3113-0.2845)/0.2845\approx9.42\%$，这充分说明：SCCM 模型的引入有助于改善 NLG 算法的“选词”能力。图 4 中，当 $K\leqslant6$ 时，SCCM 模型的 BLEU-2 评分更优，但优势并不明显；相反当 $K>6$ 时，SCCM 模型的标注性能还略低于 Random 模型，关键原因有：1）未考虑具体的表层实现模型，无法有效地利用语言学统计信息生成有实际意义的修饰性短语；2）在修饰性短语中组装新单词会带来噪声，从而抑制了 BLEU-2 评分（同图 2）。因此，有必要引入 N-gram 模型来组装关键单词，生成有实际价值的修饰性短语，最终提升标注性能。

3.4 与标注基线的比较

选取 Grad-KDES 特征完成图像分类、图像检索，基于 SCCM 模型优选与图像内容语义相关的关键单词（语义相关度 Top 10 的单词），分别采用 1-gram、2-gram、3-gram 和 4-gram 等表层实现模型生成修饰性短语，基于修饰性短语和句子模板构造句子，标注商品图像。当得到多个标注模型之后，设计 Boosting 模型：从上述 4 个标注模型的标注结果中选取 BLEU-3 评分最优的句子标注商品图像。最终，各模型的 BLEU 评分如表 1 所列。

表 1 各模型的 BLEU 评分（最优指标如0.3313等所示）

评分	Gist-Based	MLBL	1-gram	2-gram	3-gram	4-gram	Boosting
BLEU-1	0.1254	0.1890	0.3151	0.3079	0.3016	0.3057	0.3313
BLEU-2	0.0360	0.0480	0.0370	0.0516	0.0524	0.0445	0.0590
BLEU-3	0.0092	0.0170	0.0036	0.0049	0.0115	0.0086	0.0188

首先，各 N-gram 模型的 BLEU-1 评分均优于基线。这得益于两个方面：1)Grad-KDES 特征识别商品图像中的梯度变化，它能客观反映图像中形状和纹理这两类关键视觉特性，且 Grad-KDES 与文本单词共现频率高；2)SCCM 模型摘取出了准确刻画图像内容的关键单词，增强了 NLG 算法的“选词”能力。其次，2-gram、3-gram 及 4-gram 模型的 BLEU-2 评分值优于图 4 中的最佳值，这表明 N-gram 模型的引入能生成有实际意义的修饰性短语。再次，3-gram 模型无论是 BLEU-2 评分还是 BLEU-3 评分均优于其它 N-gram 模型，这表明结合 Grad-KDES 及 3-gram 模型能更好地挖掘出刻画商品图像内容的 3 元修饰性短语；相比之下，虽然 2-gram 模型的 BLEU-2 评分尚可，但由于仅聚焦 2 元短语抽取，因此它的 BLEU-3 评分较差。

与已有工作对比，3-gram 模型仅在 BLEU-3 评估中稍逊于 MLBL 模型，这说明充分挖掘并利用训练样本中的语言学统计信息有助于构造正确的修饰性短语。相比 N-gram 模型，MLBL 的优势在于：基于词向量刻画单词，精确度量单词间的语义相关性，且 MLBL 的本质是语言模型，故它在词序先后关系生成中更具优势。相比 MLBL，本文方法无需大规模的图像特征学习、词向量抽取及大量参数的调制，而仅需挖掘训练样本所构成的语料库，故模型的复杂性远低于 MLBL。最后，Boosting 模型的各 BLEU 评分均优于基线，其中它的 BLEU-1 评分比次优的 1-gram 模型有提升(0.3313－0.3151)/0.3151≈5.14%，BLEU-2 评分比次优的 3-gram 模型有提升(0.0590－0.0524)/0.0524≈12.6%，BLEU-3 评分比次优的 MLBL 模型有提升(0.0188－0.017)/0.017≈10.6%。因此，Boosting 模型是切实有效的。

3.5 标注结果展示

部分句子标注结果如表 2 所列，在 Boosting 模型的标注结果中每个句子均由 3 部分组成：前缀短语、修饰性短语和后缀短语；此外，3-gram 或 4-gram 模型所生成句子的语义信息更丰富、更完整，这与表 1 中 Boosting 模型的 BLEU-2、BLEU-3 评分结果也是完全吻合的。

表 2 部分标注结果(Boosting 生成的修饰性短语如“**detachable chain strap**”等所示)

商品图像	Boosting 模型标注	原标注
	〈商品 1〉This is a picture of **detachable chain strap** clutch bag.	〈商品 1〉Ruffles add a feminine touch to this satin clutch with detachable chain strap.
	〈商品 2〉This is a picture of **lisa david designs foldover** clutch bag.	〈商品 2〉Lisa David Designs Foldover Clutch: Woven Lamb Brown-Boutique Designer.
	〈商品 3〉This is a picture of **cadillac croco embossed small** shoulder bag.	〈商品 3〉Cadillac croco embossed small shoulder bag.
	〈商品 4〉This is a picture of **vanessa satin matte handheld** evening bag.	〈商品 4〉Vanessa Satin Matte Handheld Ivory-Evening Bags.
	〈商品 5〉This is a picture of **resist audrey will win** evening bag.	〈商品 5〉Almost impossible to resist, Audrey will win you over with non-stop charm!
	〈商品 6〉This is a picture of **rhinestone snap closure** clutch bag.	〈商品 6〉This satin clutch has two pleats criss crossing the front exterior, and a rhinestone snap closure.

在表 2 中，Boosting 模型所标注的句子能准确刻画商品图像内容，例如：“vanessa satin matte handheld”、“cadillac croco embossed small”等主要描述商品的纹理特性，“detachable chain strap”、“rhinestone snap closure”等主要描述商品的形状特性，这得益于 Grad-KDES 特征识别商品图像中的梯度变化以及 N-gram 模型对关键单词的正确组装。当然，有部分视觉特性被 Boosting 模型所忽略，例如：商品 1 的“ruffles”、商品 4 的“satin”等。主要原因：文本描述中存在噪声，导致 Grad-KDES 与某些文本单词共现频率低，故 SCCM 模型很难准确地摘要出相关单词生成修饰性短语。

结束语 在 Grad-KDES 特征抽取的基础上，基于 SCCM 模型摘取刻画商品图像内容的关键单词，采用 N-gram 模型构造修饰性短语，并借助句子模板生成句子。实验表明：1)Grad-KDES 全面、细致地刻画了商品图像的纹理和形状特性，获取了最优的标注性能；2)N-gram 模型兼顾句中语义相关性和句法模式兼容性，它获取了十分有竞争力的 BLEU 评分，而设计 Boosting 模型还能进一步提升句子标注性能；3)3-gram 和 4-gram 模型所生成的修饰性短语蕴涵更丰富、更完整的语义信息，能更准确地刻画商品图像内容；4)标注结果也会受到文本单词噪声干扰，如何抑制这些噪声及更好地摘取关键单词值得深入研究。

未来工作：尝试在核化稀疏表示 KSR[17](Kernel Sparse Representation)的特征学习基础上深度挖掘图像的关键视觉特性，更全面、客观地描绘图像内容。此外，尝试在词向量[18]学习基础上分析图像与文本之间的跨模态相关性，期望能抑制噪声文本对标注的影响。

参 考 文 献

[1] Makadia A, Pavlovic V, Kumar S. A New Baseline for Image Annotation[C]// Proceedings of European Conference on Computer Vision. 2008:316-329

[2] Yang Y, Teo C L, Daume H, et al. Corpus-guided sentence generation of natural images[C]// Proceedings of Conference on Empirical Methods on Natural Language Processing. 2011:444-454

[3] Kulkarni G, Premraj V, Dhar S, et al. Baby talk: Understanding and generating simple image descriptions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2891-2903

[4] Nwogu I, Zhou Ying-bo, Brown C. DISCO: Describing Images Using Scene Contexts and Objects[C]// Proceedings of American Association for Artificial Intelligence. 2011:1487-1493

[5] Hodosh M, Young P, Hockenmaier J. Framing image description as a ranking task: Data, models and evaluation metrics[J]. J. Artif. Intell. Res. (JAIR), 2013(47):853-899

[6] Li Pi-ji, Ma Jun, Gao Shuai. Learning to Summarize Web Image and Text Mutually[C]// Proceedings of International Conference on Multimedia Retrieval. 2012

[7] Feng Y S, Lapata M. Automatic Caption Generation for News Images [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(4):797-812

(下转第 287 页)

tributes[J]. University of Michigan,2011,42(7):3337-3344

[5] Han Lei,Li Jun-feng,JiaYun-de. Human Interaction Recognition Using Spatio—Temporal Words[J]. Chinese Journal of Computers,2010,33(4):776-784(in Chinese)
韩磊,李君峰,贾云得. 基于时空单词的两人交互行为识别方法[J]. 计算机学报,2010,33(4):776-784

[6] Niebles J C, Feifei L. A hierarchical model of shape and appearance for human action classification[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2007:1-8

[7] Yuan He-jin, Wang Cui-ru. Human action recognition using Markov random walk based scmi — supervised learning[J]. Journal of Computer-Aided Design & Compueter Graphics, 2011,23(10):1749-1757(in Chinese)
袁和金,王翠茹. 人体行为识别的Markov随机游走半监督学习方法[J]. 计算机辅助设计与图形学学报,2011,23(10):1749-1757

[8] Wang Chuan-xu, Liu Yun, Li Wan-qing. Research of Unsupervised Posture Modeling and Action Recognition Based on Spatial-Temporal Interesting Points [J]. Acta Electronica Sinica, 2011,39(8):1751-1756(in Chinese)
王传旭,刘云,厉万庆. 基于时空特征点的非监督姿态建模和行为识别的算法研究[J]. 电子学报,2011,39(8):1751-1756

[9] Mohamed B,Kaaniche,Franios. Recognizing Gestures by Learning Local Motion Signatures of HOG Descriptors[J]. Arn Analy and Mahn Nllgn Ranaon on,2012,34(11):2247-2258

[10] Dollar P, Rabaud V, Cottrell G, et al. Behavior recognition via sparse spatio-temporal features[C]// 2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. 2005:65-72

[11] Du You-tian,Chen Feng,Xu Wen-li. Approach to Human Activity Multi-scale Analysis and Recognition Based on Multi-layer DynamicBayesian Network[J]. Acta Automatica Sinica,2009,35(3):225-232(in Chinese)
杜友田,陈峰,徐文立. 基于多层动态贝叶斯网络的人的行为多尺度分析及识别方法[J]. 自动化学报,2009,35(3):225-232

[12] Saad A, Mubarak S. Human action recognition in videos using kinematic features and multiple instance learning [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010,32(2):288-303

[13] Rodriguez M, Ali S, Kanade T. Tracking in unstructured crowded scenes[C]// 2009 IEEE 12th International Conference on Computer Vision. IEEE,2009:1389-1396

[14] Basri R,Hassner T,Zelnik-Manor L. Approximate nearest subspace search[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2011,33(2):266-278

[15] Ke Y, Sukthankar R, Hebert M. Event Detection in Crowded Videos[C]// IEEE 11th International Conference on Computer Vision,2007(ICCV 2007). IEEE,2007:1-8

[16] Kellokumpu V,Zhao G,Pietikäinen M. Human Activity Recognition Using a Dynamic Texture Based Method[J]. In BMVC, 2008

[17] Bermejo N E,Deniz S O,Bueno G G,et al. Violence detection in video using computer vision techniques[C]// International Conference on Computer Analysis of Images and Patterns. 2011: 332-339

[18] Orit K G,Tal H,Lior W,et al. The Action Similarity Labeling Challenge[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2012,34(3):615-621

[19] Yeffet L, Wolf L. Local Trinary Patterns for Human Action Recognition[C]// IEEE International Conference on Computer Vision,2009:492-497

[20] Hayashi K,Seki M,Hirai T,et al. Real-time violent action detector for elevator[J]. Optomechatronic Machine Vision,2005: 60510R-60510R-8

[21] Horn B K P,Schunck B G. Determining Optical Flow[J]. Artificial Intelligence,1980,17(81):185-203

[22] Kim H,Pang S,Je H,et al. Constructing support vector machine ensemble[J]. Pattern Recognition,2003,36(12):2757-2767

[23] Wang Yuan-yuan, Wang Bin. The Conditional Random Fields Methods for Human Action Recognition[J]. Journal of Chongqing University of Technology(Natural Science),2013,27(6):93-99, 105(in Chinese)
王媛媛,王斌. 人体行为识别的条件随机场方法[J]. 重庆理工大学学报(自然科学),2013,27(6):93-99,105

(上接第 273 页)

[8] Berg T L,Berg A C,Shih J. Automatic Attribute Discovery and Characterization from Noisy Web Data [C]//Proceedings of European Conference on Computer Vision. 2010:663-676

[9] Kiapour H, Yamaguchi K, Berg A C, et al. Hipster Wars: Discovering Elements of Fashion Styles[C]// Proceedings of European Conference on Computer Vision. 2014:472-488

[10] Rebecca. Domain-Independent Captioning of Domain-Specific Images [C]// Proceedings of North American Association for Computational Linguistics. 2013:69-76

[11] Kiros R, Zemel R S, Salakhutdinov R. Multimodal Neural Language Models[C]// Proceedings of International Conference on Machine Learning. 2014

[12] Bo L, Ren X, Fox D. Kernel Descriptors for Visual Recognition [C]// Proceedings of Advances in Neural Information Processing Systems. 2010:1734-1742

[13] Bo L, Ren X, Fox D. Efficient Match Kernels between Sets of Features for Visual Recognition [C]// Proceedings of Advances in Neural Information Processing Systems. 2009:135-143

[14] Sivaram G, Hermansky H. Sparse Multilayer Perceptron for Phoneme Recognition[J]. IEEE Trans. Audio, Speech, & Language Proc. ,2012,20(1):23-29

[15] Wang J, Yang J, Yu K, et al. Locality-constrained linear coding for image classification[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2010:3360-3367

[16] Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation[C]//Proceedings of the Annual Meeting on Association for Computational Linguistics. 2002:311-318

[17] Gao Sheng-hua, Tsang W-H, Chia L T. Sparse Representation With Kernels [J]. IEEE Transactions on Image Processing, 2013,22(2):423-434

[18] Maas A L, Daly R E, Pham P T, et al. Learning Word Vectors for Sentiment Analysis[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011:142-150