标签增强的中文看图造句

蓝玮毓 王晓旭 杨 刚 李锡荣

(中国人民大学数据工程与知识工程教育部重点实验室 北京 100872)

(中国人民大学信息学院多媒体计算实验室 北京 100872)

摘 要 图像数据飞速增多,而计算机与人对图像数据的理解间存在语义鸿沟.如何有效地理解、管理和组织图像数据是学术界和工业界面临的一个重大挑战.利用计算机自动生成能够描述图像内容的自然语言描述有助于弥合语义鸿沟,从而提升对图像数据的理解.现有工作致力于英文句子生成.与之不同的是,该文实现了一个面向中文的看图造句系统.通过大规模机器翻译克服了中文训练数据缺乏的问题,同时提出了结合中文标签自动预测,对深度模型预测句子进行重排序的增强方法,改善句子生成质量.在两个中文图像句子数据集 Flickr8k-cn 和 Flickr30k-cn 上的实验表明,该文提出的标签增强方法可以有效改善现有两种看图造句模型(Google 模型和 Attention 模型)所生成句子的质量.标签增强使得 Google 模型在 Flickr8k-cn 测试集上的 CIDEr 指标从 0.474 提高到 0.503,Flickr30k-cn 测试集上的 CIDEr 指标从 0.325 提高到 0.356.通过标签增强,Attention 模型在这两个数据集上的 CIDEr 分别从 0.510 提高到 0.536,从 0.392 提高到到 0.411.

关键词 图像句子生成;中文模型;中文标签预测;深度学习;机器翻译

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2019.00136

《计算机学报》

Improving Chinese Image Captioning by Tag Prediction

LAN Wei-Yu WANG Xiao-Xu YANG Gang LI Xi-Rong

(*MOE Key Lab of Data Engineering and Knowledge Engineering, Renmin University of China, Beijing* 100872)

(*Multimedia Computing Laboratory, School of Information, Renmin University of China, Beijing* 100872)

Abstract How to effectively annotate, organize and consequently manage the increasing amounts of images has been a grand challenge for both academia and industry. On one hand, the rapid growth of the images makes it impossible to manually annotate the data. On the other hand, the semantic gap, i. e., the lack of correspondence between low-level information a computer extract directly from the visual content and the interpretation that the same content has for a given user, makes automated annotation by computer challenging. Image captioning, which aims to automatically generate a natural-language description for a given image, helps bridge the semantic gap and consequently improves the management and accessibility of image data at a semantic level. With few exceptions, the task of image captioning has so far been explored only in the English language since publicly available datasets are mostly in this language. The application of image captioning, however, should not be bounded by language. Extending the study of image captioning in the dimension of language is essential for a large population in the planet who cannot speak English. Different from the mainstreams that focus on English sentence generation, this paper develops an image captioning system that describes images in the Chinese language. Only

收稿日期:2017-07-17;在线出版日期:2018-05-15.本课题得到国家自然科学基金项目"面向中文的看图造句若干关键问题研究"(61672523)、"基于社会网络计算的企业舆情管理新理论新方法"(71531012)资助.**蓝玮毓**,女,1993 年生,硕士研究生,主要研究方向为图像理解.E-mail:bbluey@126.com.**王晓旭**,女,1993 年生,硕士研究生,主要研究方向为图像检索.**杨 刚**,男,1979 年生,博士,讲师,主要研究领域为人工智能.**李锡荣**(通信作者),男,1983 年生,博士,副教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为多媒体内容分析与检索.E-mail:xirong.li@gmail.com.

few studies have been conducted for image captioning in a cross-lingual setting, most of which tackle this problem by constructing new datasets in the target languages. Such an approach is constrained by the availability of manual annotation and thus difficult to scale up. Instead of building large datasets in a new language manually, we make the best use of existing English datasets to construct Chinese captioning datasets by machine translation. Moreover, to improve the quality of generated Chinese sentences, we propose to exploit automated tagging of Chinese words to provide external information for sentence generation. From English datasets we train a Multilayer Perceptron (MLP) model that can directly predict Chinese tags for a test image. Per training image we obtain its Chinese tags by first extracting noun, verb and adjective words from its original English descriptions and then performing dictionary-based translation. For the caption generation model, we follow a popular CNN+RNN approach. During the beam search for generating a sentence, the captioning model can choose several candidate words and therefore construct several candidate sentences. Using the predicted tags, we propose to improve the quality of the final generated Chinese sentence by reranking the candidate words or sentences in terms of their matches with the tags predicted by the MLP model. To evaluate the effectiveness of our proposed approach, we conduct experiments on two Chinese captioning datasets, namely Flickr8k-cn and Flickr30k-cn, using two models, which are (1) the Google model that directly uses full-image representation; (2) an Attention model that learns to focus on specific regions of the image that are deemed to be salient according to the attention mechanism. The experimental results show that, while reranking candidate words does not bring extra improvement, reranking candidate sentences using predicted tags improves the performance of both the Google model and the Attention model. Specifically, our sentence-reranking strategy improves the CIDEr score from 0.474 to 0.503 on the test set of Flickr8k-cn, from 0.325 to 0.356 on the test set of Flickr30k-cn for the Google model. As for the Attention model, the strategy improves the CIDEr score from 0.510 to 0.536 on Flickr8k-cn, from 0.392 to 0.411 on Flickr30k-cn. For state-of-the-art performance on Chinese image captioning, we suggest the use of the Attention model with sentence reranking using predicted tags.

Keywords　image captioning; Chinese model; Chinese image tagging; deep learning; machine translation

1　引　言

图像数据正处于爆炸式增长的阶段.由于计算机直接从图像像素中获取的信息与特定用户对于图像内容的解读之间存在着天然的语义鸿沟,处理图像较处理文本更具挑战性.如何有效地理解、管理和组织不断增加的图像数据已是学术界和工业界面临的一个重要挑战.

根据图像内容自动生成文字描述被认为是弥合语义鸿沟的一种有效途径[1].通过在大量带标注图像数据上进行(深度)学习,计算机可以对图像进行词语级别标签的标注,或者给出一句话的文字描述.

看图造句①(Image Captioning)是一个连接了多媒体分析、计算机视觉与自然语言处理领域的研究课题.句子有更丰富的语义信息,比起孤立的标签能够更形象、准确地描述图片内容,有助于对图像数据进行更好的理解和管理.如何自动生成高质量的描述图像的句子是一个吸引了许多研究者关注的、具有挑战性的问题.看图造句模型需要同时具备理解图像内容,检测物体、场景等概念的能力,以及将抽象的视觉信息组织为完整的自然语言句子的能力.目前主流的构建看图造句模型的方法是在大型的图像句子对应数据集上利用反向回馈训练深度神

① 目前国内并没有对 Image Captioning 的中文翻译达成共识.本文认为"看图造句"是一个能够清楚表述该问题(即计算机根据图像内容自动生成自然语言句子)的翻译,比起"图像描述"等宽泛的翻译更合适.

《计算机学报》

经网络.现有的看图造句工作以研究如何为图像生成英文句子为主[2-6].显然,此项研究不应该受限于语言.将看图造句扩展到母语使用人口最多的中文环境,不仅能够拓宽该项研究的边际,也将产生更大的社会价值.

1.1 本文贡献

本文研究面向中文的看图造句.截止本文投稿时,文献[7]是仅有的尝试用机器翻译得到的中文句子,在现有英文模型框架下训练中文看图造句模型.但是,现有深度学习模型的优化目标主要是基于当前图像和已经生成的词语,使预测的下一个单词是正确单词的概率最大化,一定程度上忽略了最终生成句子整体语义与图像内容的关联度.针对此问题,本文提出了一种标签增强方法,通过自动预测的中文标签优化预测句子语义与图像内容关联度,改进最终生成的句子质量.

之前的工作如文献[8-9]对利用自动预测的标签改进句子生成展开了有益的探索,但还是针对英语.现有的图像分类模型也都是预测英语标签,不能直接用于中文看图造句任务.就本文作者知识所及,目前尚无公开的可用于训练多标签图像分类模型的中文数据集.为此,本文提供了一个解决方案,基于英文数据集训练一个可为一张图预测多个中文标签的图像分类模型.此外,本文进行了比之前的工作更全面的实验,验证了标签增强方法对于不同看图造句模型均能有效提升生成句子的质量.

本文第 2 节将回顾看图造句方面的相关工作;第 3 节描述如何获得大量的中文图像句子训练数据;第 4 节介绍面向中文的句子生成模型、标签模型以及利用标签信息改进句子的方法,并在第 5 节中展示实验结果;第 6 节中将给出结论.

2 相关工作

现有的看图造句的方法主要有以下三个类别.

(1)基于搜索的方法.比如,Hodosh 等人[2]提出查找与当前测试图像最相似的训练图像,把训练集中的描述句子作为候选答案;文献[10-12]则提出在视觉和文本的共同映射空间中对训练集的描述句子进行排序.这种方法对于不在训练集中的概念组合无法给出新的描述.

(2)基于模板或语言模型的方法.得益于物体检测方面的进展,也有一些研究[13-14]通过使用人为定义的模板或者语法,根据测试图像中物体的属性和位置等信息生成描述.在最近的研究中[15],研究者使用 CNN(卷积神经网络)预测一组可能出现在图像描述中的词语,并利用最大熵语言模型生成描述图像的句子.由于这种方法需要依赖于事先定义的物体、属性、场景等概念,限制了所产生的描述句子的丰富性.

(3)深度学习相关的方法.受到 RNN(循环神经网络)在机器翻译领域的成功应用的启发,在最近的一些研究工作中[3-4],结合 CNN 和 RNN 的模型被用于自动为图像生成描述性句子,取得了很好的效果.图像首先通过在 ImageNet 上预先训练好的 CNN 提取视觉特征,该特征经过基于仿射变换的语义映射,作为 RNN 的初始输入.之后,RNN 依次输出单词序列,组成一个描述图像内容的自然语言句子.

基于这种 CNN+RNN 的框架,Xu 等人[5]提出了 Attention(注意)机制,进一步优化了模型的效果.Attention 机制使用 CNN 低层的保留图像空间信息的卷积层特征,在预测单词序列过程中始终作为视觉信息输入到 RNN 中.在每一轮预测单词后,模型重新聚焦于图像某些位置,得到新的视觉信息作为 RNN 的输入,以预测下一个单词.相似的,把语义相关注意机制[16]、场景信息[17]或高层次语义信息[18]作为图像的视觉特征来预测句子的方法也使模型的效果得到了优化.

最新工作尝试将强化学习(Reinforcement Learning)应用于训练看图造句的模型[6].考虑到现有模型的训练目标是使正确的词语序列被预测的概率最大,而这个目标与预测时的评价指标并不是一致的.基于强化学习的方法在训练时直接优化评价指标,获得了更好的效果.

仅有少数的工作研究了面向非英语语种的图像句子生成.文献[19]提出通过结合图像以及图像原有的英语描述句子来生成新的目标语言句子.但是在实际的跨语言看图造句时,对于任意测试图像,并没有正确的源语言的描述句子可以用于辅助产生新的目标语言句子.因此该方法的应用范围相当受限.文献[7]提出了利用机器翻译得到的中文数据集可以用于构建面向中文的图像自动描述系统.但该工作只是用中文数据重新训练了 Google 模型,并没有针对中文改进算法.图像标签级别的信息比起句子更容易获取,并且标签比句子更容易翻译成其他的语言,因此本文提出了结合标签级信息来改进跨语言看图造句的方法.此前利用自动预测的标签改进句子生成的一些工作[8-9]则只是针对英语的,没有考虑跨语言的问题.这些工作自然也未考虑如何构建

中文标签预测模型.并且,上述工作都是以 Google 模型做为句子生成模块.目前尚不清楚自动预测的标签对于更先进的引入了 Attention 机制的模型是否同样有效.以上这些问题在本文中都进行了探索.

3　中文句子集的获取

深度学习相关的图像看图造句方法的发展得益于现有的包含成对图像与自然语言句子的公开数据集.这些数据集含有大量来自图像共享网站 Flickr 的用户照片,每张图片有对应的多个由英语母语人士给出的描述图像内容的句子.其中有较多研究的一些数据集包括:包含 8 千张图像、4 万个英语句子的 Flickr8k[2];基于 Flickr8k 扩展的含有 3 万张图像、15 万个英语句子的 Flickr30k[20];包含约 12 万张图像、60 万个英语句子的 MSCOCO[21] 数据集.然而这些数据集都是单语言的数据集,无法得到跨语言的看图造句模型,限制了看图造句应用的语言环境.

为了把英语的数据集扩展为中文的数据集,文献[7]中采用了两种方法,一是采用众包的方式,对于数据集中的图像,请中文母语者写出描述图像内容的中文语句,每张图像有五个人写出的描述性句子.另一种方法是采用机器翻译,将 Flickr8k 图像句子数据集中的英语描述语句直接翻译成中文语句,并且人工翻译测试集中的描述句子作为标准答案.通过两种方式扩展得到了 Flickr8k 的中文版,称为 Flickr8k-cn.文献[22]中对 Flickr30k 进行了汉化,通过机器翻译将 Flickr30k 训练集和验证集中的句子从英文变为中文,通过人工翻译将 Flickr30k 测试集的句子从英文变为中文,构建了 Flickr30k-cn.

用众包的方式重新构建中文的图像句子数据集,需要花费较多的时间和资源,并且很难扩展到更大的数据集,比如包含约 60 万个句子的 MSCOCO 数据集.而通过机器翻译,从现有的英语图像句子数据集扩展为中文的数据集,虽然翻译质量并不总是很理想,但扩展性更好,更容易生成大量的中文训练数据.之前的实验[7]也表明,使用机器翻译将现有的数据集的英语句子翻译成中文,在此基础上训练中文的图像句子生成模型,直接生成的中文句子,比将英语模型的预测结果翻译得到的中文句子的更好,并且表现出了与原来英语模型的句子生成效果相似的水平.因此,本文也采用机器翻译得到的中文数据训练看图造句模型.

本文使用的这两个中文数据集相关的统计信息见表 1.

表 1　本文实验使用的两个中文数据集

	Flickr8k-cn[6]			Flickr30k-cn[22]		
	训练	验证	测试	训练	验证	测试
图像	6000	1000	1000	29783	1000	1000
机器翻译句子	30000	5000	—	148915	5000	—
人工翻译句子	—	—	5000	—	—	5000

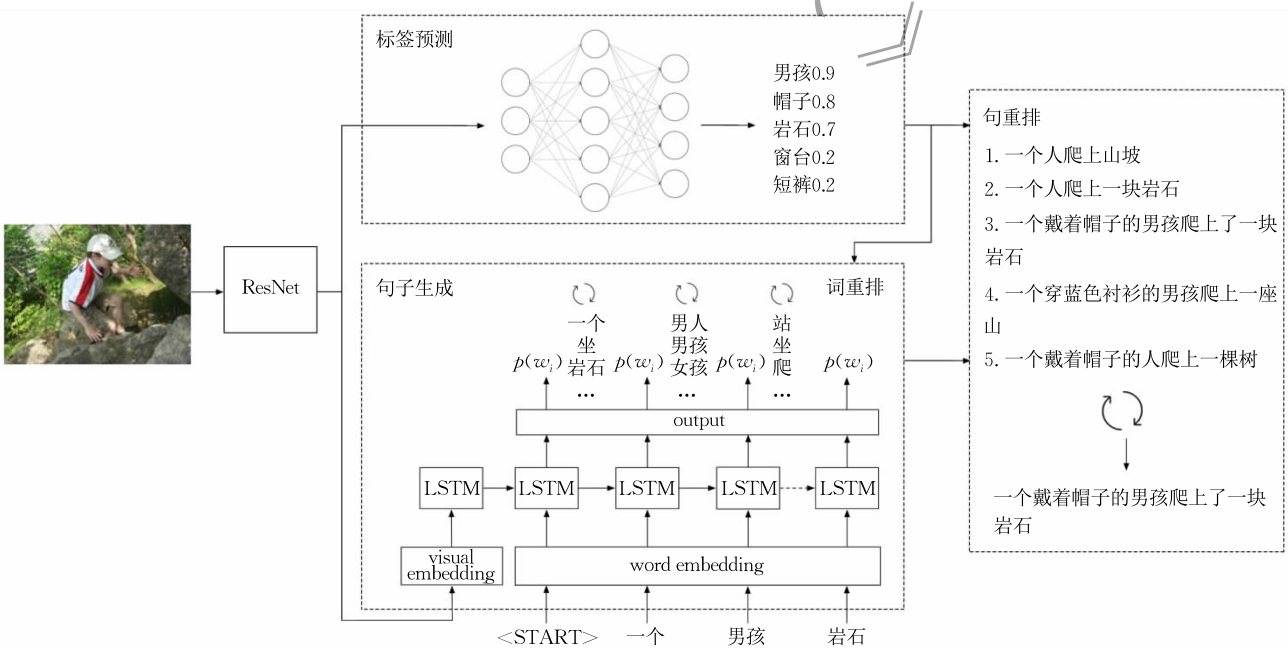


图 1　标签增强的中文看图造句框架(本系统为测试图片预测一组标签,对句子生成模块预测的候选词语或候选句子进行重排序,改进生成句子的质量)

4 面向中文的句子生成

为了描述一张给定的测试图片 I,本文拟为其预测一个中文标签集合$\{T_i\}$,并生成一个中文句子 S. 令中文词表为 V. 句子 S 由一个词语序列$\{w_1,w_2,\cdots,w_n\}$组成,$w_j\in V$;标签 T_i 与 V 中的单词对应,即 $T_i\in V$. 如图 1 所示,本文提出的标签增强方法是结合自动预测的标签,对候选词语或句子进行重新排序,得到最终的预测句子.

由于中文语句缺乏自然分隔符,调用了 BosonNLP[23] 的服务对中文句子进行分词,得到词语序列. 本文在接下来将分别介绍中文句子生成和中文标签预测的模型,以及结合两者进行重排序的算法.

4.1 中文句子生成

本文采用了基于深度学习的结合 CNN 与 RNN 的两种生成式模型为图片自动生成一句话的中文描述. 这两种模型是(1)直接利用图像全局特征的 Google 模型[4];(2)加入 Attention 机制的利用图像局部特征的 Attention 模型[6].

4.1.1 Google 模型

Google 模型由文献[4]提出. 模型结构如图 1 句子生成模块所示. 对于给定的图片,概率模型根据图像全局视觉特征给出词语序列的后验概率. 令模型的参数为 θ,则句子的概率表示为 $p(S|I)$,将句子展开表示为词语序列,并且为了计算方便对概率取对数,可以得到以下的表达式

$$\log p(S|I;\theta)=\sum_{t=1}^{n+1}\log p(w_t|I,w_0,\cdots,w_{t-1};\theta) \tag{1}$$

其中 $w_0=w_{n+1}=START/END$,是一个表示句子的开始或结束的特殊符号.

概率的估计由 RNN 计算得出. 图像 I 经过 CNN 得到特征表示,与映射矩阵 $\boldsymbol{E}_I$ 相乘后作为 RNN 的初始输入提供视觉信息,RNN 再生成对应的句子. RNN 是一种循环的具有记忆功能的神经网络. 它在每一轮接收一个输入向量,给出一个输出向量,并通过更新自身状态向量以记忆此前的信息. 输入的单词由一位有效向量(one-hot vector)表示,经过映射矩阵 $\boldsymbol{E}_w$ 映射到与视觉信息相同的向量空间中,从而可以作为 RNN 的输入. 更具体地,RNN 首先接收图像视觉信息,忽略这一轮的输出;然后输入一个特殊的开始符号 START 后,RNN 给出单词表 V 中的每个词语被预测的概率的估计,每次选取被生成的概率最大的词语作为本轮输出;再把这一个词语作为模型下一轮的输入,继续预测下一个可能的单词,直到预测出结束符号 END. RNN 模型训练时的目标是在给定图像的情况下,使正确描述的句子,即正确的词语序列被预测的概率最大. 模型使用随机梯度下降的算法进行训练.

本文中 RNN 的结构选择了 LSTM(Long Short Term Memory)网络[24],LSTM 的单元中含有控制是否忘记此前信息(forget gate f)、是否接受新的输入(input gate i)以及是否输出当前信息(output gate o)的"闸门". 闸门是通过 sigmoid 层来实现的,数值在[0,1]之间,将闸门与目标数据对应相乘达到控制数据流通的效果. 这些闸门有效地解决了普通 RNN 的梯度消失的问题.

在具体实现中,本文使用的 LSTM 结构与文献[4]中使用的 LSTM 结构基本相同,使用 TensorFlow 中 BasicLSTMCell 的实现,增加了 forget gate 的 bias 以减少模型训练初期的忘记部分. LSTM 单元包含有单元向量 $\boldsymbol{c}$(cell vector)和隐藏状态向量 $\boldsymbol{h}$(hidden state vector),可以处理并记忆当前输入的信息. 在每一个时间节点 t,LSTM 单元计算三个闸门的取值,再根据闸门对单元自身的向量进行更新:

$$\begin{aligned}
x_t&=E_w w_{t-1}\ \text{for}\ t>0,\ x_0=E_I CNN(I)\\
i_t&=\sigma(W_i x_t+U_i h_{t-1})\\
f_t&=\sigma(W_f x_t+U_f h_{t-1}+b_f)\\
o_t&=\sigma(W_o x_t+U_o h_{t-1})\\
c_t&=f_t\odot c_{t-1}+i_t\odot\tanh(W_c x_t+U_c h_{t-1})\\
h_t&=o_t\odot c_t
\end{aligned} \tag{2}$$

其中 x_t 是 LSTM 的输入,σ 是 sigmoid 激活函数,$\odot$ 表示对应向量与闸门取值的乘积,W、U 和 b 是 LSTM 结构中需要训练的模型参数.

4.1.2 Attention 模型

本文使用的第二种模型是基于 Attention 机制的模型. 与 Google 模型相比,该模型使用了卷积神经网络低层的保留了图像空间信息的卷积层特征,并且在预测单词序列过程中始终作为视觉信息输入到 RNN. 在每一轮预测单词后,Attention 模型重新计算图像不同位置的权重以聚焦图像的某些区域并作为 RNN 的输入. 具体地,本文使用的 Attention 模型是文献[6]中提出的 Att2in 模型. 与最初的 Attention 模型[5]相比,主要的区别是,此模型把由 Attention 机制计算得出的加权图像特征只作为 LSTM 的单元向量 $\boldsymbol{c}$(cell vector)的输入,而不是同时作为 LSTM 的三个门和单元向量的输入.

上述的模型更新过程用公式表示如下:

《计算机学报》

$$
\begin{aligned}
x_t &= E_w w_{t-1} \\
i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\
f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\
o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \Phi(W_c x_t + W_z I_t + U_c h_{t-1} + b_c) \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned} \tag{3}
$$

其中，ϕ 是 maxout(最大输出)非线性激活函数，I_t 为 Attention 模型计算得出的 t 时刻的权重加权得到的图像特征. 给定 N 个位置的卷积层特征 $\{I_1, I_2, \cdots, I_N\}$，$I_t$ 的计算方式如下：

$$
\begin{aligned}
I_t &= \sum_{j=1}^{N} \alpha_t^j I_j \\
\alpha_t &= softmax(a_t + b_a) \\
a_t^j &= W tanh(W_{aI} I_j + W_{ah} h_{t-1} + b_a)
\end{aligned} \tag{4}
$$

式(3)中的单元向量和隐藏状态向量的初始值，即 $\boldsymbol{c}_0$ 和 $\boldsymbol{h}_0$ 设为零向量.

4.2　中文标签预测

为预测图像的中文标签，本文训练了一个多标签图像分类模型. 如图 2 所示，本文从描述图像的英文句子中为图像获取中文标签作为训练数据.

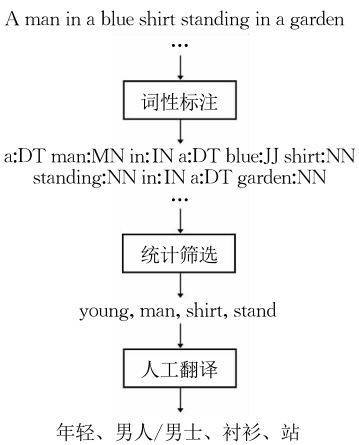


图 2　从描述图像的英语句子中获取中文标签

考虑到机器翻译对英文句子的翻译可能会有歧义或漏翻译的部分，本文选择先从原始英语句子通过自动词性标注筛选出具有描述能力的英文标签. 再通过人工翻译将英文标签替换为相应的中文标签.

在 Flickr8k 和 Flickr30k 数据集中，每个图像对应了 5 个英文句子. 为了保证标签能够有效地描述图像，本文仅保留词性为名词、形容词或者动词的词汇作为图像的标签；同时为了排除人工标注句子主观性导致的与图像内容无关的词汇，仅保留在同一张图片对应的 5 个句子中出现了至少 2 次的词语. 本文用斯坦福大学自然语言研究小组提供的语法解析工具[25]对英文句子进行词性标注. 被保留的英文词汇的词性分布统计见图 3.

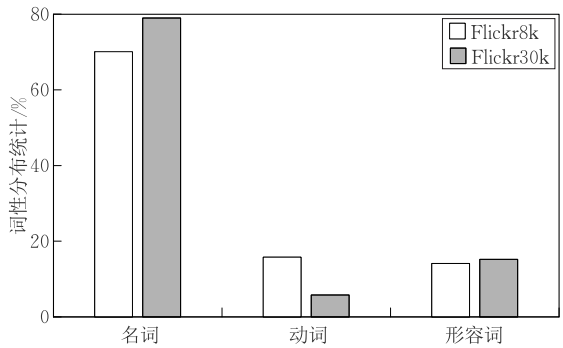


图 3　被保留英文词汇的词性分布统计

为了保证每个标签有足够的训练样本，对于每个数据集，本文经验地保留出现次数不小于 20 次的词汇，并对其进行人工翻译. 经过筛选，Flickr8k-cn 和 Flickr30k-cn 中无中文标签的图像数目分别为 9 和 14. 无标签的图像不能用来训练和测试，所以在后续实验中将这些图像去除. 表 2 给出了两个数据集每张图像获得的中文标签的数目统计.

表 2　数据集中每张图像对应的中文标签个数统计

数据集	平均值	最大值	最小值
Flickr8k-cn	4.64	12	1
Flickr30k-cn	5.72	19	1

4.2.1　多标签预测模型

本文的多标签预测模型由特征提取模块和分类模块构成. 其中，特征提取模块采用预训练的 CNN 网络提取图像特征. 因此，该部分与之前的句子生成模块共享，如图 1 所示.

分类模块由一个多层感知器(MLP)实现. 其输入层的神经元个数等同于 CNN 图像特征的维度，输出层的神经元个数与中文标签数目相同，中间层的神经元个数经验地取为 1024. 注意到这是个多标签问题，即一个图像可以同时被多个不同标签所描述. 因此，输出层的激活函数选用 sigmoid 而非图像分类中常用的 softmax. 同时相关的损失函数的计算方式也变为对应的多标签的计算方式，即二元交叉熵损失函数(binary cross-entropy loss).

为了用相对少的时间获得更好的模型，本文采用的训练策略如下. 在用 Flickr8k-cn(或 Flickr30-cn)训练模型时，每个 epoch 之后都在相应的验证集上评测其性能. 若连续 5 个 epoch 在验证集上的准确率都没有提高，则将学习速率减半. 若调整后的学习速率低于预先指定的阈值(本文取 1e-8)则结束训练.

对于每一个测试图像，提取 CNN 特征并输入

《计算机学报》

到训练好的 MLP 中，通过前向计算获得每一个标签的概率. 按照概率值对标签降序排序，取排名前 10 的标签作为预测结果.

4.3 候选序列重排

生成句子时，基于 CNN+RNN 结构的看图造句模型，包括本文所采用的两种模型，均可以利用集束搜索(beam search)策略，根据预测概率选择多个不同的候选单词，并获得多个候选句子. 具体的集束搜索过程如下：在每个时间节点考虑 b 个目前的最优候选序列，并在此基础上对每个序列考虑 b 个候选单词以扩展该序列，保留所有得到的扩展序列中最优的 b 个候选序列，直到产生 b 个完整的候选句子. 该方法最终预测得出的句子是在集束搜索产生的候选句子中的最优序列，即由看图造句模型计算得出被预测概率最大的句子

而如前所述，这种深度学习模型的优化目标均是基于当前图像和已经生成的词语，使预测的下一个单词是正确单词的概率最大化，忽略了最终生成句子整体语义与图像内容的关联度. 本文通过自动预测得到标签信息，提出了两种利用标签提升句子整体语义与图像内容的关联度，从而改进句子生成质量的方法，分别是对预测候选词语和候选句子进行重排序.

4.3.1 重排序候选词语

第一种方法是重排序候选词语. 为一张测试图像产生预测句子时，在每个时刻，看图造句模型中的 LSTM 结构会对单词表 V 中的每个词语给出被预测的概率 $p_{\text{lstm}}(w_i)$ 的估计；标签预测模型则会给出标签集合中每个词语与图像相关的概率 $p_{\text{tag}}(w_i)$ 的估计.

利用预测图像标签重排序候选词语方法的具体步骤为，在每个时刻，对每个候选单词 w_i，同时考虑看图造句模型与标签预测模型的预测概率，对加权概率进行排序，将加权概率更高的词语作为候选词语加入集束搜索的序列中，以获得最终预测句子. 用公式表述为

$$p(w_i) = p_{\text{lstm}}(w_i) + \beta \cdot p_{\text{tag}}(w_i) \tag{5}$$

对于在句子单词表 V 中，而不在候选标签集合中的词语 w_j，$p_{\text{tag}}(w_j)=0$. 式(5)中，β 是一个可调整的权重值. 通过在验证集上进行搜索，选取在验证集上取得最好效果的 β，取 $\beta=0.1$.

4.3.2 重排序候选句子

第 2 种方法是重排序候选句子. 对于一张测试图像，看图造句模型直接利用集束搜索得到多个候选句子，模型本身评价候选句子优劣的方式是根据该模型计算得出的不同句子被预测概率的大小. 候选句子 S 由模型给出的预测分数用公式表述如下，为了计算方便对概率取对数：

$$\begin{aligned} score_{\text{lstm}}(S) &= \log p(S \mid I) \\ &= \sum_{t=1}^{n+1} \log p(w_t \mid I, w_0, \cdots, w_{t-1}) \end{aligned} \tag{6}$$

结合预测标签对候选句子进行重排序的步骤如下，统计候选句子中的词语与标签预测模型预测标签重合的词语，计算这些词语的标签预测分数之和，与看图造句模型本身计算得到的句子得分 $score_{\text{lstm}}(S)$ 相加，作为新的分数对句子进行排序，即

$$score(S) = \sum_{w_i \in \{T_j\} \cap S} p_{\text{tag}}(w_i) + score_{\text{lstm}}(S) \tag{7}$$

得分最高的候选句子作为最后的模型输出.

在本文接下来的实验部分，将会分析比较这两种方法是否能够改善生成的中文句子质量.

5 实 验

本文的实验使用单块 Titan X Pascal GPU，在深度学习框架 TensorFlow 及 PyTorch 上进行实验.

5.1 标签预测

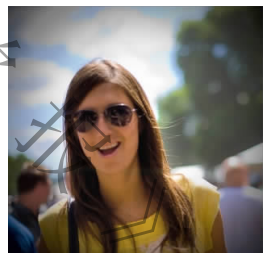
5.1.1 实验设置

为了验证标签预测模型的有效性，本文在两个测试集上对图像进行中文标签预测. 测试集上中文标签的答案是从 Flickr8k-cn 和 Flickr30k-cn 中的人工翻译的句子获取的. 其获取方式为，利用 BosonNLP 的中文语义 API 对人工翻译的句子进行分词和词性预测，仅保留其中的名词、动词、形容词. 表 3 给出了部分测试图像的中文标签答案(在表中用“GT 标签”标识).

本文采用了预先训练好的卷积神经网络 ResNet-152[26] 来抽取图像特征，这是一个在 ImageNet 和 MSCOCO 的图像分类、检测竞赛中获得了当前最好结果的网络结构. 残差网络(ResNet)中的残差结构可以根据网络每层的输入学习残差函数，这种残差函数有效解决了深层网络模型难以训练的问题，能大大增加网络深度的同时获得高精准度. 具体地，本文使用 ResNet-152 的 pool5 层，得到的图像特征是一个 2048 维的向量，并对特征进行了 L2 范数归一化.

训练时，利用验证集对 MLP 模型的一系列超参数进行调优，最终采用的设置是：batch 大小为 256，初始学习率为 0.001.

表 3　Google 模型利用预测标签重排序的实例(GT 标签、MLP 预测标签及其得分也展示在表格中)

Google 模型	一只黑色的狗跳进水里	两个人坐在岩石上俯瞰大海
GT 标签	水 狗 湖 跳 码头 黑色	山 人 顶部
MLP 标签	水 0.97 狗 0.96 跳 0.22 黑色 0.12 河 0.10 大 0.06 木 0.06 玩 0.06 湖 0.05	男人/男士 0.56 坐 0.30 山 0.30 岩石 0.18 女人/女士 0.18 人 0.16 看 0.15 顶部 0.11
词重排 MLP	水里的一只狗	坐在岩石上的人
句重排 MLP	一只黑色的狗跳进水里	两个人坐在一座山上的岩石上
Google 模型	两只黑色的狗在草地上玩耍	两只狗在草地上玩耍
GT 标签	狗 黑色	棕色 玩 球 狗 绿色 白色 草/草地
MLP 标签	狗 0.94 草/草地 0.74 黑色 0.64 玩 0.38	狗 0.99 草/草地 0.78 球 0.75 棕色 0.50
词重排 MLP	玩具 0.27 田野 0.25 跑 0.20 棕色 0.13 两只黑色的狗在草地上玩耍	白色 0.33 玩 0.33 跑 0.26 红色 0.18 一只狗在草地上跑
句重排 MLP	一只黑色的狗和一只棕色的狗在草地上玩耍	一只棕色和白色的狗在草地上玩球
		
Google 模型	戴眼镜的人	一个戴着墨镜的女人看了看相机
GT 标签	衬衫 帽子 黑色 男人	衬衫 黄色 墨镜 女孩
MLP 标签	男人/男士 0.92 穿/戴 0.58 黑色 0.49 女人/女士 0.45 标志 0.31 拿着 0.26 衬衫 0.15 蓝色 0.13 帽子 0.10	女人/女士 0.88 女孩 0.83 微笑 0.61 墨镜 0.46 穿/戴 0.40 年轻 0.31 头发 0.19 相机 0.15 衬衫 0.11 眼镜 0.09
词重排 MLP	一个人戴着一顶黑色的帽子拿着一根烟	一个微笑的女人
句重排 MLP	一个男人穿着一件蓝色的衬衫和黑色的帽子	一个女人和一个年轻的女孩微笑

本文采用的评价指标是多标签分类中常用的三项指标：Precision@k、Recall@k 和 F1@k. 其中，Precision@k 表示前 k 个预测标签的准确率，即在前 k 个预测标签中正确的标签个数除以 k；Recall@k 表示前 k 个预测标签的召回率，即在前 k 个预测标签中正确的标签个数除以答案中的总标签个数；F1@k 表示对前 k 个预测标签的准确率和召回率计算加权调和平均.

5.1.2　实验结果

为了进一步验证 MLP 模型的有效性，本文加入 K 近邻(KNN)作为对比模型. 文献[1]的图像标签预测实验表明，基于深度特征的 KNN 虽然简单，仍是一个很强的基线模型. 参考文献[1]，本文选用余弦距离作为距离函数. 近邻个数取值为 25. KNN 使用的图像特征与 MLP 相同，即 ResNet-152 特征.

表 4 和表 5 中展示了在两个测试集上采用不同方法预测多标签的实验结果. 在两个数据集上，MLP 的准确率都比 KNN 要高. 在 Flickr8k-cn 上，KNN 的召回率高于 MLP. 注意到 MLP 需要学习的参数远多于 KNN，因此前者更适合大规模数据

集.这一点在 Flickr30k-cn 上得到了验证.在该数据集上,MLP 的三项指标均优于 KNN.

表 4 Flickr8k-cn 上中文标签预测性能 (单位:%)

k	模型	Precision	Recall	F1
1	KNN	68.5	**8.2**	**14.4**
	MLP	**70.5**	7.7	13.6
5	KNN	47.9	**27.5**	33.7
	MLP	**51.9**	27.2	**34.4**
10	KNN	33.2	37.2	33.8
	MLP	**40.0**	**40.6**	**38.8**

表 5 Flickr30k-cn 上中文标签预测性能 (单位:%)

k	模型	Precision	Recall	F1
1	KNN	68.5	**4.7**	**8.7**
	MLP	**72.6**	**4.7**	**8.7**
5	KNN	47.0	15.3	22.5
	MLP	**50.4**	**15.5**	**23.2**
10	KNN	36.0	23.1	27.3
	MLP	**41.0**	**24.8**	**30.0**

对于 MLP 模型,两个数据集的前 5 个预测标签的准确率分别达到了 51.9%和 50.4%.通过观察图像,可以发现有些预测错误是标准答案不完备导致的.以表 3 右侧第 1 张图像为例.前 5 个预测标签分别是"男人/男士"、"坐"、"山"、"岩石"、"女人/女士".而图像的英文句子中没有对应的词汇,因此提取的中文标签中没有包含"男人/男士"、和"女人/女士".因此,标签预测的实际性能要好于目前报告的结果.

总体看,无论是 MLP 还是 KNN,召回率较准确率明显偏低.需要在未来工作中继续探索改进.

5.2 句子生成

5.2.1 实验设置

在 Google 模型的实验中使用了与 5.1 节标签预测任务中相同的 ResNet-152 特征.Attention 模型实验中使用的是 ResNet-101 的最后一层卷积层特征,并且使用了平均池化(average pooling),得到维度为 14×14×2048 的特征.

两个模型的实验均采用了单层的 LSTM,图像和词语的映射向量的维度以及 LSTM 的隐藏层的大小都设置为 512.对于训练集中出现次数小于五次的词语,用一个特殊的"未知"符号代替.

实验采用了被广泛用于评价看图造句效果的评价指标来衡量中文看图造句模型预测句子的质量,即 BLEU-4、ROUGE-L 以及 CIDEr.BLEU 是设计用于评测机器翻译结果的评价方法,它的计算基于预测结果与多个参考答案中的 n-gram 准确度以及预测结果的长度罚分比(惩罚太短的预测句子).BLEU-4 即考虑了 1-gram 到 4-gram 的多元片段的重合度的得分.ROUGE-L 通过计算基于最长子序列的 F 值来评价生成句子的质量.CIDEr 是专门设计用于测评看图造句模型的评价指标,它考虑由句子的每个 n-gram 的 TF-IDF(Term Frequency Inverse Document Frequency)指标组成的向量,通过计算预测句子与参考答案的平均余弦距离来评价句子,对于句子中的重要的成分给予更大的权重.以上的指标都是得分越高表示效果越好.本文采用了 MSCOCO 比赛的评价计算代码来计算这些指标.

5.2.2 实验 1:不同排序方法的对比

本文提出了两种结合词语级标签信息,通过重排序改善生成句子质量的方法,即词重排与句重排.首先,本文在 Google 模型上进行了对比实验,以对比这两种不同排序方法的效果.

为了验证本文提出的标签重排序方法的有效性,与不进行重排序的基线方法进行了比较:基线方法的预测句子是在集束搜索产生的候选句子中由模型本身计算得出的被预测概率最大的句子.实验中使用了由两种方式得到的标签进行重排序:(1)使用本文提出的 MLP 模型预测的标签进行重排序,记为词重排-MLP 与句重排-MLP;(2)使用答案标签对预测候选词语、句子进行重排序,记为词重排-GT 与句重排-GT.

表 6、表 7 中分别展示了在两个中文数据集 Flickr8k-cn 和 Flickr30k-cn 上,两种方法,即重排序候选词语与重排序候选句子的效果以及与基线方法结果的对比.其中基线方法记为 Google 模型.与未重排的 Google 模型相比,不论是使用 MLP 模型的预测标签,还是使用答案标签,重排序候选词语的方法均没有带来改善.而重排序候选句子方法,在使用答案标签进行重排时,在两个数据集上、每个指标上都有了很大的提升.比如 CIDEr 指标在 Flickr8k-cn 上由基线方法的 47.4 提升到 59.5,在 Flickr30k-cn 上由基线方法的 32.5 提升到 42.1.这表明,当标签信息质量很高时,句重排方法可以有效地提高预测句子的质量.同时,如表格所示,使用本文提出的 MLP 模型预测标签的句重排方法,与基线方法对比,每个指标都有更高的得分,也有效地提升了预测句子的质量,是所有不使用答案的方法中效果最好的方法.

《计算机学报》

表 6　Flickr8k-cn 上标签增强对于 Google 模型的影响

方法	BLEU4	ROUGE-L	CIDEr
Google 模型	23.6	44.8	47.4
词重排-MLP	23.4	44.9	47.2
词重排-GT	23.6	45.0	47.3
句重排-MLP	**24.7**	**45.8**	**50.3**
句重排-GT	**28.6**	**48.9**	**59.5**

表 7　Flickr30k-cn 上标签增强对于 Google 模型的影响

方法	BLEU4	ROUGE-L	CIDEr
Google 模型	18.2	40.0	32.5
词重排-MLP	18.0	39.9	32.5
词重排-GT	18.2	40.0	32.8
句重排-MLP	**20.0**	**41.9**	**35.6**
句重排-GT	**23.3**	**45.0**	**42.1**

表 3 中展示了利用 MLP 模型预测标签对 Google 模型预测句子进行重排序的一些实例. 通过观察预测句子可以发现，词重排的方法使某些预测标签词语更早地出现在句子中，限制了有更丰富语义信息的句子的生成. 比如表格中的第一张图片，预测标签中概率最大的词语是“水”，重排序候选词语时选中了“水”作为预测序列中的句首词语，使得“跳进”等动词无法被组织进预测句子中，最终生成的句子“水里的一只狗”，相对其他方法的“一只黑色的狗跳进水里”而言，描述句子的完整度和丰富度受到了影响. 而句重排的方法，既保证了句子的完整度，也使得预测标签更多的出现在预测句子中，生成了更丰富、准确的预测句子.

5.2.3　实验 2：句重排方法对于不同模型的提升

利用标签对候选句子进行重排序是一种对于可以进行集束搜索的看图造句模型的通用改善方法. 因此，为了进一步验证句重排方法的有效性，本文在不同模型，即 Google 模型与 Attention模型上进行了对比实验. 重排序 Google 模型的候选句子的实验结果展示在表 6、表 7 中. 与实验 1 相同，实验 2 中也使用了由两种方式得到的标签，即答案标签与 MLP 模型预测标签进行重排序，记号同上.

表 8、表 9 中展示了在中文数据集 Flickr8k-cn 和 Flickr30k-cn 上，对 Attention 看图造句模型的候选句子，利用不同的标签进行重排序得到的结果的评价指标得分.

表 8　Flickr8k-cn 上标签增强对于 Attention 模型的影响

方法	BLEU4	ROUGE-L	CIDEr
Attention 模型	25.8	46.8	51.0
句重排-MLP	25.3	48.0	53.6
句重排-GT	30.4	50.1	63.0

表 9　Flickr30k-cn 上标签增强对于 Attention 模型的影响

方法	BLEU4	ROUGE-L	CIDEr
Attention 模型	20.2	41.9	39.2
句重排-MLP	20.5	43.8	41.1
句重排-GT	26.7	47.7	53.2

表 6、表 7 中的结果表明，使用 Google 模型预测看图造句时，利用标签信息对候选句子进行重排序，有效地提升了预测句子的质量.

从表 8 和表 9 中的数据可以看出，当使用更复杂、预测句子质量更高的 Attention 模型来产生图像描述时，结合预测标签对候选句子进行重排序同样可以改善最终生成句子的质量. 具体表现为，利用 MLP 模型预测的标签对句子重排序后，与未排序的句子相比，在 Flickr8k-cn 数据集上，CIDEr 指标由 51.0 提升至 53.6，在 Flickr30k-cn 数据集上，CIDEr 指标由 39.2 提升至 41.1.

综合以上实验，不论是使用 Google 模型还是性能更好的 Attention 模型，重排序候选句子的方法均可以改善最终生成的描述句子的质量.

5.3　讨　论

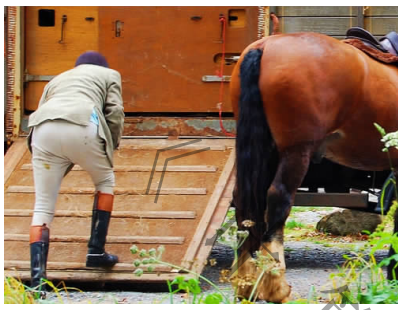
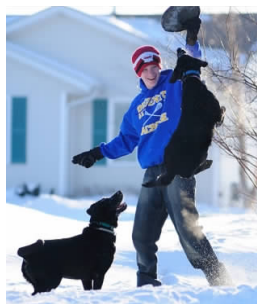
理论上，物体检测(object detection)也可用于给图像打上标签. 但其词汇限于提供实体类名词. 本文的标签集合中还包括动词和形容词，描述更为丰富，因此更适合用于看图造句任务.

表 10 中展示了利用 MLP 预测标签对 Attention 模型预测句子进行重排序的实例，预测标签为最终生成的句子提供了更多语义信息，使最终得到的句子更准确、丰富. 但是从表 10 中最后一排的例子中可以看出，当标签预测不准确时，也会引入一定的噪音，影响最终预测句子的准确性. 同时，如表 8、表 9 所示，当使用答案标签对 Attention 模型预测的候选句子进行重排序，句子质量可以获得更进一步的改善. 这表明，本文提出的句重排方法仍有很大的提升潜力，采用性能更好的标签预测模块可以进一步提升重排序方法带来的效果改善.

6　总结与展望

本文使用机器翻译得到的大规模中文数据集建立了一个面向中文的看图造句系统，并且针对跨语言看图造句数据不足的问题，提出了结合使用中文标签级别信息改善生成句子质量的方法. 在两个中文数据集 Flickr8k-cn 和 Flickr30k-cn 上的实验表

表 10 Attention 模型利用预测标签重排序候选句子的实例(GT 标签、MLP 预测标签及其得分也展示在表格中)

Attention 模型	一个滑板的人	一个骑自行车的人
GT 标签	坡道 滑板 滑板者 男人/男士	自行车 技巧 男人/男士
MLP 标签	黄色 0.40 红色 0.30 男人/男士 0.27 滑板 0.15 穿/戴 0.14 男孩 0.14 衬衫 0.13 白色 0.11 跳 0.11	自行车 0.62 男人/男士 0.55 山 0.38 跳/跳跃 0.31 空中 0.21 人 0.12 泥土 0.10 山丘 0.10 技巧 0.07
句重排 MLP	一个穿着红色衬衫的人在一个黄色的滑板上跳下来	一个骑自行车的人在一座山上跳跃
		
Attention 模型	一个戴着帽子的人站在一头奶牛旁边	两只狗在雪地里玩耍
GT 标签	马 坡 背/后背 男人/男士	玩 狗 雪 黑色 男人/男士
MLP 标签	男人/男士 0.77 马 0.69 衬衫 0.26 棕色 0.22 黑色 0.21 帽子 0.20 墙 0.13 蓝色 0.11 穿/戴 0.09	雪 0.76 狗 0.57 黑色 0.55 男人/男士 0.30 穿/戴 0.15 蓝色 0.12 山 0.08 外套 0.07 红色 0.06
句重排 MLP	一个戴着帽子的人站在一匹棕色的马旁边	一个男人和一只黑色的狗在雪地里玩耍
Attention 模型	新娘和新郎在婚礼上跳舞	一个人坐在长椅上
GT 标签	结婚 教堂 气球 拿 情侣 新郎 新娘	商店 咖啡 女人/女士 年轻 帽子
MLP 标签	人 0.43 白色 0.27 群 0.23 男人/男士 0.22 女人/女士 0.20 婚礼 0.19 站 0.194 走 0.18 新郎 0.14 穿/戴 0.12	男人/男士 0.62 坐 0.61 女人/女士 0.54 白色 0.36 海滩 0.32 衬衫 0.13 公园 0.13 看 0.11 年轻 0.11 女孩 0.10
句重排 MLP	一群穿着白色衣服的男人和一个穿着白衣服的女人	一个男人和一个女人坐在一张桌子上看书

明,不论是 Google 模型还是 Attention 模型,利用预测的标签对系统产生的多个候选句子进行重新排序有效地改善了最终预测句子的质量.在目前性能较好的 Attention 模型的基础上,Flickr8k-cn 测试集上的 CIDEr 指标从 51.0 提升到 53.6,Flickr30k-cn 测试集上 CIDEr 指标提升从 39.2 提升到 41.1.同时,实验显示利用答案标签对模型预测的候选句子进行重排序可以进一步改善最终预测句子的质量,表明本系统仍有很大的提升空间,因此在未来的工作中,将继续提升标签预测模型的效果.预测标签信息除了用于句子重排序,还可以通过其他方式结合到句子生成过程中改进效果,这也是未来的研究工作.此外,后续工作将尝试对目前在学界更流行的 COCO 数据集进行汉化.

参 考 文 献

[1] Li X, Uricchio T, Ballan L, et al. Socializing the semantic gap: A comparative survey on image tag assignment, refinement and retrieval. ACM Computing Surveys, 2016, 49(1): 14:1-14:39

[2] Hodosh M, Young P, Hockenmaier J. Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research, 2013, 47(1): 853-899

[3] Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 3128-3137

[4] Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: A neural image caption generator//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 3156-3164

[5] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention//Proceedings of the International Conference on Machine Learning. Lille, France, 2015: 2048-2057

[6] Rennie S, Marcheret E, Mroueh Y, et al. Self-critical sequence training for image captioning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA, 2017: 7008-7024

[7] Li X, Lan W, Dong J, Liu H. Adding Chinese captions to images//Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval. New York, USA, 2016: 271-275

[8] Dong J, Li X, Lan W. Early embedding and late reranking for video captioning//Proceedings of the 2016 ACM on Multimedia Conference. Amsterdam, The Netherlands, 2016: 1082-1086

[9] Li X, Jin Q. Improving image captioning by concept-based sentence reranking//Proceedings of the Advances in Multimedia Information Processing-Pacific Rim Conference on Multimedia. Xi'an, China, 2016: 231-240

[10] Frome A, Corrado G. S, Shlens J, et al. Devise: A deep visual-semantic embedding model//Proceedings of the Advances in Neural Information Processing Systems. Lake Tahoe, USA, 2013: 2121-2129

[11] Jia Y, Salzmann M, Darrell T. Learning cross-modality similarity for multinomial data//Proceedings of the 2011 IEEE International Conference on Computer Vision. Barcelona, Spain, 2011: 2407-2414

[12] Socher R, Karpathy A, Le Q V, et al. Grounded compositional semantics for finding and describing images with sentences. Transactions of the Association for Computational Linguistics, 2014, 2(1): 207-218

[13] Farhadi A, Hejrati M, Sadeghi M, et al. Every picture tells a story: Generating sentences from images//Proceedings of the 11th European Conference on Computer Vision. Berlin, German, 2010: 15-29

[14] Li S, Kulkarni G, Berg T L, et al. Composing simple image descriptions using web-scale *n*-grams//Proceedings of the Fifteenth Conference on Computational Natural Language Learning. Portland, USA, 2011: 220-228

[15] Fang H, Gupta S, Iandola F, et al. From captions to visual concepts and back//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1473-1482

[16] You Q, Jin H, Wang Z, et al. Image captioning with semantic attention//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 4651-4659

[17] Li X, Song X, Herranz L, et al. Image captioning with both object and scene information//Proceedings of the 2016 ACM on Multimedia Conference. Amsterdam, The Netherlands, 2016: 1107-1110

[18] Wu Q, Shen C, Liu L, et al. What value do explicit high level concepts have in vision to language problems?//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 203-212

[19] Elliott D, Frank S, Hasler E. Multilingual image description with neural sequence models. arXiv preprint arXiv: 1510.04709, 2015

[20] Young P, Lai A, Hodosh M, et al. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics, 2014, 2: 67-78

[21] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context//Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 740-755

[22] Lan W, Li X, Dong J. Fluency-guided cross-lingual image captioning//Proceedings of the 2017 ACM on Multimedia Conference. Mountain View, USA, 2017: 1549-1557

[23] Min K, Ma C, Zhao T, et al. BosonNLP: An ensemble approach for word segmentation and POS tagging//Proceedings of the National CCF Conference on Natural Language Processing and Chinese Computing. Nanchang, China, 2015: 520-526

[24] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780

[25] Chen D, Manning C D. A fast and accurate dependency parser using neural networks//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2014: 740-750

[26] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

《计算机学报》

LAN Wei-Yu, born in 1993, M. S. candidate. Her research interests focus on image content understanding.

WANG Xiao-Xu, born in 1993, M. S. candidate. Her research interests focus on image search.

YANG Gang, born in 1979, Ph. D., lecturer. His research interests focus on machine learning.

LI Xi-Rong, born in 1983, Ph. D., associate professor. His research interests include multimedia analysis and retrieval.

Background

Image captioning, which targets at automatically describing the pictorial content of an image by a natural-language sentence, is a fundamental problem that connects computer vision with natural language processing. The recent dominant approach in image captioning, inspired by the success of deep learning in image classification and machine translation, is to apply deep neural networks which contain a CNN and an RNN to automatically generate new captions for images. This has so far been explored mostly in English, as most available datasets are in this language. However, the application of image captioning should not be restricted by language. Only few studies have been conducted for image captioning in a cross-lingual setting. This paper extends research on automated image captioning in the dimension of language, learning a Chinese image captioning model from machine-translated sentences and enhancing the model by automated image tagging.

This work was supported by the National Natural Science Foundation of China (Nos. 61672523, 71531012). The project aims to develop novel algorithms that enable learning an image captioning model for a target language other than English, with minimal labeling effort.

The Multimedia Computing Lab (MMC) at Renmin University of China is doing cutting-edge research in the broad area of multimedia computation and applications. We publish papers in prestigious journals and conferences in the multimedia field, e. g., IEEE T-MM, T-ASLP, ACMMM, ICMR, SIGIR and ICASSP. The MMC Lab has been top performer in a number of international competitions and benchmark evaluations for image and video captioning, including TRECVID 2016 / 2017 Video-to-Text tasks, MSR VTT Challenge at ACM MM 2016 / 2017, and Image CLEF 2015 Image Sentence Generation task.

《计算机学报》