文章编号：1003-0077(2018)04-0001-12

图像的文本描述方法研究综述

马龙龙，韩先培，孙 乐

（中国科学院 软件研究所 中文信息处理实验室，北京 100190）

摘 要：随着深度学习技术的兴起，自然语言处理与计算机视觉领域呈现相结合的趋势。作为融合视觉和语言的多模态研究任务，图像的文本描述可应用于基于文本内容的图像检索、网络图像分析等众多场景中，从而受到了研究界和企业界的广泛关注。图像的文本描述方法可归纳为三大类：基于生成的方法、基于检索的方法和基于编码—解码的方法。该文详细介绍了这三类方法各自具有代表性的工作，并进一步分析了各方法的优劣；然后对图像文本描述方法的相关数据集、评测标准和主要开源工具包进行了阐述；最后，分析了图像的文本描述中需要解决的关键技术问题。

关键词：图像的文本描述；生成；检索；编码—解码

中图分类号：TP391 **文献标识码**：A

A Survey of Image Captioning

MA Longlong, HAN Xianpei, SUN Le

(Chinese Information Processing Laboratory, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: As a new multimodal task which connects vision and language, image captioning can be applied to text-based image retrieval and network image analysis etc., thereby has drawn wide attention from the research and business community. Generally, existing image captioning methods fall into three categories: generation-based method, retrieval-based method and encoder-decoder method. In this paper, we first present the representative work of three methods with analysis of the advantages and disadvantages of these methods. Then we give the datasets, evaluation metrics and several open-source toolkits of image captioning. Finally we reveal the key technical problems in image captioning task.

Key words: image captioning; generation; retrieval; encoder decoder

0 引言

随着可拍照移动智能终端的广泛使用和互联网的快速发展，融合视觉和文本信息的多模态数据在急剧增加，例如，带文本标注的照片、报纸文章中的图文对照内容、带标题的视频以及社交媒体出现的多模态交互数据。多模态机器学习（multi modal machine learning）为机器提供了处理多模态数据的能力，多模态学习的长远目标是使机器充分感知环境，更智能地和环境进行交互。当前多模态处理包括图像/视频的文本描述、基于视觉的问答和看图讲故事等任务。本文聚焦于多模态学习中的图像文本描述（image captioning）[1]方法。使用图像文本描述方法可以有效组织图像数据，结合文本信息检索技术方便地对海量图像数据进行搜索，能够从幻灯片中的图片读懂演讲者所讲的内容。此外，使用图像文本描述方法可以帮助视觉障碍者理解图像。

图像的文本描述也是计算机视觉和自然语言处理领域的交叉任务，能够完成从图像到文本的多模态转换，最早由 Farhadi[2]等人提出。该任务可具体形式化描述为：给定二元组(I,S)，其中I表示图像，S表示图像的文本描述句子，模型完成从图像I到描述句子S的多模态映射$I\rightarrow S$。该任务对于人

收稿日期：2017-12-06 **定稿日期**：2018-02-28

基金项目：国家自然科学基金(61772505)

类来说非常容易，但是却给机器带来了巨大挑战，因为机器不仅要理解图像的内容，还要产生人类可读的描述性句子。

图像的文本描述方法可用来分析图像中的视觉内容并产生文本描述。典型任务是用一句话描述图像中出现的视觉对象、对象属性及对象之间的关系；给出图像中描述情境的特征，提供图像的情境背景知识，如室内还是户外；描述图像中出现的对象之间的相互关系，甚至推理出图像中未出现的内容。例如，图像内容为火车站候车室，人们在等候火车，虽然图像中并未出现火车，但是自动生成的文本描述中可能出现火车字样。而传统的图像理解任务主要集中于发现并分割出图像中的对象、确定对象的属性、计算图像情境的属性和识别出现在图像中的人与对象的相互关系。图像理解的结果为无结构的标签列表，无法直接用于图像的文本描述。

图1给出了MS COCO[3]数据集中图像的英文文本描述实例。在该实例中，要生成图像的文本描述句子，首先需要模型能够分析图像，理解图像中出现的对象、动作、属性和场景等信息，通过选择并执行一定的语义和语法规则，生成概括性的描述句子。

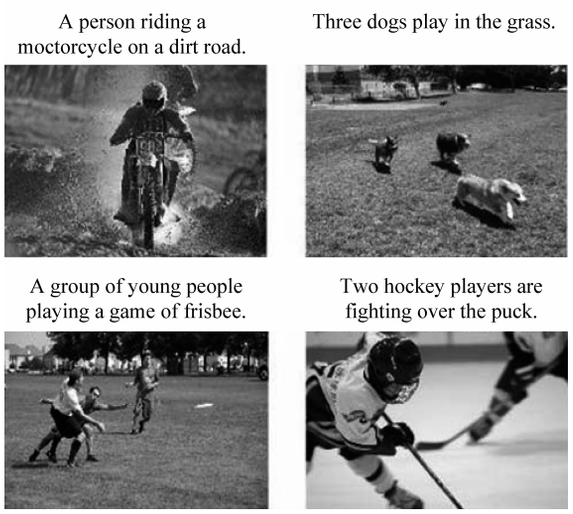

图1 图像的文本描述实例

图像的文本描述不但需要理解图像内容，而且需要实现内容选择、内容组织，以及用语言生动表现出所要表达内容的自然语言产生过程，因此图像的文本描述句子需要满足以下三个条件。

(1) 正确描述图像内容。

(2) 产生的文本描述必须类似于人类的描述，而且能够较好地描述个性化的特征，例如，对于同一幅图片，艺术评论显然不同于一般的娱乐性杂志撰稿。

(3) 产生的图像文本描述能够尽可能地包含不同视角、人类对图像各个层次的理解。

纵观国内外研究人员关于图像的文本描述方法的研究，根据所处阶段的关键技术不同及文本描述方法的不同，我们将图像的文本描述方法分为以下三大类。

(1) 基于生成的方法(generation-based method)。该方法分为检测过程和生成过程。检测过程基于图像特征检测图像中出现的对象、对象属性、图像表达内容的场景和行为等信息；生成过程使用这些信息驱动自然语言产生系统输出图像的文本描述。

(2) 基于检索的方法(retrieval-based method)。为了生成图像的文本描述，该方法检索数据库中与输入图像相似的图像集，基于检索到的相似图像集的文本描述，用最相似的检索结果合理组织生成图像的文本描述。

(3) 基于编码—解码的方法(encoder-decoder method)。该方法以深度学习为基础，采用编码—解码的方式直接生成文本描述。这种方法需要大规模的训练语料支撑，生成的文本描述形式多种多样，不受限于固定的语言模板。

本文其余章节结构如下：第一～三节分别详细介绍了基于生成的方法、基于检索的方法和基于编码—解码的方法；第四节阐述了图像文本描述的数据集和评测标准；第五节对主要开源工具包进行简介；第六节分析目前图像的文本描述所要解决的关键问题及研究难点；第七节为结束语。

1 基于生成的方法

基于生成的方法用计算机视觉技术检测出图像中的对象，预测对象的属性和相互关系，识别图像中可能发生的行为，然后用特定的模板、语言模型或句法模型生成图像的文本描述句子。

该方法依赖于预先设定的场景对象、对象属性以及行为等语义类别，根据句子生成方法的不同又可分为基于模板的方法、基于句法分析的方法和基于语言模型的方法。

1.1 基于模板的方法

基于模板的方法需要预先设置包含多个需要用对象关系和属性标签去填充的模板，这些对象关系

和属性标签形成空槽，对空槽进行填充，形成图像的文本描述句子。

Kulkarni[4]等人提出 Baby Talk 模型，该模型使用检测器识别对象、属性和相互关系，采用 CRF 算法预测标签，最后使用模板生成文本描述。Kuznetsova[5]等人学习训练集已有的句子描述产生树形句子片段，测试时与新生成的文本描述再组合，产生最终的图像文本描述。Yang[6]等人用隐马尔科夫模型选择可能的对象、动词、介词及场景类型填充句子模板。

1.2　基于句法分析的方法

基于句法分析的方法首先检测对象、对象属性、对象之间空间关系、图像场景类型、对象行为等，然后使用依存句法树/图驱动句子的各个部件逐步生成完整的描述句子。

Elliott[7]等人提出首个基于句法分析的方法 VDR(visual dependency representation)，该方法用依存图表示对象之间的关系，将图像解析为 VDR，然后遍历 VDR 并考虑 VDR 与依存句法树的约束关系填充句子模板的空槽，从而生成图像的文本描述。Elliott[8]等人进一步改进了 VDR 方法，提出了从数据自动生成依存图的方法，该方法通过使用图像和文本数据自动学习图像中对象的颜色、纹理和形状等属性，并对各属性按打分进行排序。该方法的优势是解决了 VDR 方法对大量人工标注数据的依赖问题。Mitchell[9]等人把图像文本描述问题看作是 VDR 句子对的机器翻译问题，执行显式的图像内容选择和语法约束，用带约束的整数规划方法得到图像的文本描述。

1.3　基于语言模型的方法

基于语言模型的方法首先生成若干句子中可能出现的短语，然后依赖语言模型对这些短语片段进行组织，从而生成图像的文本描述。

Kulkarni[4]等人首先确定图像中的对象、属性和介词等相关信息，将其表示成元组，然后使用预先训练好的 N-gram 语言模型生成流畅的文本描述句子。同样，Li[10]等人先产生多个句法合理的句子片段并用维基百科数据训练 N-gram 语言模型，然后组合这些句子片段产生最终的图像文本描述。Fang[11]等人提出基于最大熵语言模型生成图像文本描述的方法，该方法首先使用多实例学习的方法生成若干单词，然后使用最大熵语言模型确定已知若干单词的条件下最可能产生的文本描述句子。

最近得益于深度神经网络的快速发展，越来越多的方法采用 RNN 作为语言模型，RNN 是基于时序的神经网络结构，相比于传统的 N-gram 语言模型，RNN 能够捕获任意长度的上下文信息，而不仅仅局限于前后 n 个上下文单词。关于 RNN 语言模型的方法我们将在第三节详细描述。

1.4　小结

基于生成的方法在检测过程中依赖于概念检测的质量，在生成过程中受限于人工设计的模板、不完备的语言模型以及有限的句法模型，因而，该方法生成的文本描述句子单一，不具有多样性。

2　基于检索的方法

基于检索的方法将图像的文本描述问题看作信息检索问题，即在数据集 C 中寻找查询图像 I_q 的相似子集 $M=(I_m,S_m)$，其中 I_m 表示图像集，S_m 表示图像对应的文本描述集，通过合理地组织 S_m 输出查询图像 I_q 的文本描述结果 S_q。

根据图像表示方法和相似度计算方法，基于检索的方法进一步分为基于视觉空间的检索方法和基于多模态空间的检索方法。

2.1　基于视觉空间的检索方法

基于视觉空间的检索方法利用图像视觉特征的相似性，从训练图像集中查询，得到候选图像集，然后利用候选图像集中的图像和文本信息生成图像的文本描述，具体步骤如下：

(1) 用特定视觉特征表示输入图像；

(2) 从训练图像集中基于视觉特征空间相似性度量标准检索得到候选图像集；

(3) 利用包含在候选集的图像和文本信息，根据一定规则或方法组合生成图像的候选文本描述，最后对图像的候选文本描述进行排序，选取最优结果。

Torralba[12]等人构建了 Tiny Image 数据库，该数据库使用 WordNet 中的单词为每张图像建立多个标签。Kuznetsova[13]等人基于 Tiny Image 数据库来描述查询图像，检索视觉相似性图像集。大多数基于视觉空间的检索方法以这个步骤为基准，然后用对象行为检测及场景分类器对候选图像进行处理，将视觉和短语识别结果作为特征，根据排序算法

得到最优文本描述。

Verma[14]等人使用RGB、HSV颜色直方图、Gabor和Haar描述、GIST和SIFT描述作为图像视觉特征，利用这些图像视觉特征的相似性得到图像的文本描述信息。候选图像的文本描述划分为一定类型的短语，如主语、介词、宾语等，查询图像的最优描述，由图像相似性、谷歌搜索计数值以及图像三元组构成的联合概率分布确定。

Ordonez[15]等人提出了Im2Text模型，并在规模为一百万的图像文本描述数据库中进行检索。Patterson[16]等人构造了大规模场景属性数据集，在该数据集上训练属性分类器作为图像文本描述的全局属性特征，通过扩展Im2Text模型，可产生更好的图像检索和文本描述结果。Mason[17]等人使用该场景属性描述方法，先从训练集中找出视觉相似的图像，基于相似图像集的文本描述采用概率密度估计的方法预测描述句子中单词的条件概率。最终查询图像的文本描述使用两种方法得到，一种方法基于SumBasic模型[18]，另一种方法由查询图像的单词条件概率分布与候选图像集描述概率分布的K-L散度最小化得到。

Yagcioglu[19]等人提出组合分布语义平均查询扩展方法，图像特征表示由卷积神经网络VGG-CNN(visual geometry group convolutional neural network)[20]得到，图像特征为在ImageNet数据集上训练的深度神经网络最后一层计算激活函数值得到，查询图像的文本描述由相似性检索得到图像集的分布式表示得到，权值为查询图像与检索训练图像之间的相似性。Devlin[21]等人使用VGG-CNN最后一层激活函数作为全局图像描述特征，用K近邻方法确定查询图像的视觉相似图像集。计算相似度时，用训练集中图像和查询图像的N-gram重叠F测度作为度量距离标准，查询图像的文本描述由具有最高平均n元重叠F测度得到，也就是K近邻中心描述。

2.2 基于多模态空间的检索方法

基于多模态空间的检索方法分为两步：

(1) 用训练集上的图像和对应的文本描述学习多模态空间表示；

(2) 给定查询图像，在图像和对应文本描述的联合表示空间进行图像和文本模态的交叉检索，即查询图像得到图像的文本描述和查询句子可得到对应的图像内容。

Hodosh[22]等人提出KCCA(kernel canonical correlation analysis)方法学习多模态空间表示，该方法使用核函数提取高维特征，并将图像的文本描述问题看作检索问题，使用最近邻方法进行检索，最后对候选文本综合排序，产生图像的文本描述结果。该方法需要保存核矩阵，只适用于小规模数据集。Socher[23]等人用深度神经网络学习图像—句子联合隐嵌入空间，分别学习图像和文本模态表示，然后再映射到多模态空间。Socher[24]等人进一步提出一种基于KCCA的半监督视觉语义对齐模型，该模型能够使用少量的标注数据和大量的无标注数据训练，完成单词和图像区域的对齐。单词和图像区域被映射到多模态空间，根据EM算法估计模型参数，多模态特征相似的单词和图像区域显式地对齐。

Karpathy[25]等人考虑嵌入细粒度单元，即图像中对象对应的依存树嵌入共有子空间，最终模型集成了全局图像—句子特征和对象—部分句子依存树局部特征。Kiros[26]等人基于深度神经网络产生文本描述，使用LSTM递归神经网络计算句子特征，用卷积神经网络提取图像特征，将图像特征投影到LSTM隐状态空间，神经网络语言模型从多模态空间产生查询图像的文本描述。

2.3 小结

基于检索的方法能够很好地利用训练数据集，当训练集与测试集相关性较高时效果显著。该方法依赖于大规模的训练语料，产生的文本描述局限于训练集的描述文本。

3 基于编码—解码的方法

近几年，基于编码—解码的方法在计算机视觉和自然语言处理等领域都有广泛的应用。基于编码—解码方法的图像文本描述过程分为两步。

(1) 编码阶段：用深度卷积神经网络CNN提取图像的视觉特征；

(2) 解码阶段：基于提取的图像视觉特征作为解码阶段的输入，利用RNN/LSTM输出图像的文本描述句子。

Vinyals[27]等人提出了谷歌NIC模型，该模型将图像和单词投影到多模态空间，并使用长短时记忆LSTM网络生成文本描述。Xu[28]等人提出模型gLSTM，该模型使用语义信息引导长短时记忆

LSTM 网络生成文本描述。Li[30] 等人构建了首个中文图像文本描述数据集 Flickr8kCN，并提出中文文本描述生成模型 CS-NIC，该方法使用 GoogLeNet[19] 对图像进行编码，并使用长短时记忆 LSTM 网络对图像生成过程建模。Donahue[31] 等人提出的学习模型把静态图像和图像文本描述单词输入到四层 LSTM 网络。Gan[32] 等人提出基于语义组合网络的图像文本描述方法，在文本描述生成过程中引入高层语义概念。Rennie[33] 等人提出分两步生成段落长度的图像文本描述方法。第一步，LSTM 沿时间展开的每个时刻，图像特征向量都输入到 LSTM，生成表示图像文本描述句子的单词向量序列；第二步，将第一步生成的单词向量序列作为另一个用来生成图像描述句子的 LSTM 输入，这个 LSTM 通过在序列模型的输入中加入句子向量来预测图像描述句子中的下一个单词。

根据编码和解码方法不同，基于编码—解码的图像文本描述方法又可分为三种：基于融合的方法、基于注意力的方法以及基于强化学习的方法。

3.1　基于融合的方法

基于融合的方法主要是将图像特征向量和文本描述过程中产生的文本特征向量相融合。融合操作分为三种方式。

(1) 叠加融合：将图像特征向量和图像文本描述过程中产生的文本特征向量叠加在一起，形成增广向量，增广向量长度是两个向量长度之和。叠加融合方法直观、简便，易于实现，但是如果深度学习的层数较多的话，叠加融合方法使得神经网络的参数个数增加。

(2) 加融合：假定图像特征向量和图像文本描述过程中产生的文本特征向量的维数一样，将这两种模态特征向量的相同下标的元素相加，产生一个相同维数的向量。

(3) 乘融合：假定图像特征向量和图像文本描述过程中产生的文本特征向量的维数一样，将这两种模态特征向量的相同下标的元素相乘(element-wise product)，产生一个相同维数的向量。

融合过程把描述图像的句子用 RNN 处理，再与 CNN 计算得到的图像特征向量按上述三种融合策略合并，合并后的向量输入到 softmax，最终输出图像的文本描述句子。

Kiros[34] 等人通过求解图像特征向量和图像文本描述过程中产生的文本特征向量最大相似性，把图像特征向量和图像文本描述过程中产生的文本特征向量投影到多模态共有子空间，将对数双线性语言模型的输出、图像特征向量或者文本特征向量进行融合，以便预测图像文本描述句子的下一个单词。Mao[35] 等人提出首个基于神经网络的图像文本描述生成模型 m-RNN，该模型使用 CNN 对图像建模，用 RNN 对句子建模，并使用多模态空间为图像和文本建立关联。Hendricks[36] 等人也使用了把图像特征向量和 LSTM 生成的文本嵌入向量融合形成多模态空间向量的方法。Tanti[37] 等人提出的图像文本描述方法采用两种不同融合方式：(1)图像特征和文本特征融合后作为 RNN 的输入；(2)RNN 仅处理文本序列，RNN 的输出与图像特征融合后送入前馈神经网络产生输出结果。实验结果表明 RNN 仅处理文本序列效果较好。

3.2　基于注意力的方法

Xu[38] 等人最早将基于注意力的方法引入到图像的文本描述中，使用卷积层提取基于位置的空间特征，在图像多个局部区域和文本句子之间建立关联。文中介绍了两种基于注意力的方法：Hard attention 和 Soft attention 注意力机制。解码使用整个图像特征向量来初始化 LSTM 单元。用融合的方式把区域图像(指整个图像中的一块区域)输入到 LSTM，使 LSTM 产生新的状态；然后把这个状态和区域图像融合在一起，以此来预测图像文本描述句子中的下一个单词。图像特征向量经过加权平均被融合到 LSTM 的解码过程中，使得文本描述生成网络能够捕捉图像的局部信息，提升了图像文本描述方法的性能。Andrej[39] 等人提出对图像中的多个局部区域和文本描述片段进行显式对齐，使用 RCNN(region convolutional neural network)的方法选取可能的图像区域进行排序，选择概率最大的 19 个作为候选区域，经过仿射变换得到图像区域特征，与单词特征进行相似度匹配，使用注意力的思想为每个单词找到最匹配的图像区域。图像文本描述的生成过程用 RNN 完成，首先将第一个单词和图像特征向量一同输入 RNN 中，在其后的输入中，图像特征被看作是一个全零向量。

Zhou[40] 等人提出一种基于 text-conditional 注意力机制的方法，传统的注意力机制方法关注于图像的局部区域，而该方法强调关注于文本描述句子的某个单词，使用文本信息改善局部注意力。模型采用 td-gLSTM(time-dependent gLSTM)方法，该

方法对句子中各单词的嵌入表示求平均，并与图像嵌入表示相融合，生成 text-conditional guidance 信号，该信号用于引导 LSTM 产生文本描述序列。

Yang[41]等人描述了一种通用的基于注意力机制的编码—解码模型，这个模型可以用来生成图像描述句子。该方法在编码—解码结构中添加了评价网络(review network)，评价网络基于注意力机制设计，每个步骤输出一个思考向量(thought vector)，思考向量用来作为注意力网络的输入。注意力机制模型中的图像子区域和整个图像融合后，参与到图像文本描述生成过程中。

You[42]等人提出一种注意力机制的图像文本描述方法，该方法将自底向上和自顶向下的方法相融合。基于语义注意(semantic attention)的思想，整个方法有选择地聚焦于单词，分别提取图像整体特征和若干概念的局部特征，将若干局部特征加权与图像的整体特征在单词级执行融合，并参与到 RNN 运算过程中。Chen[43]等人基于注意力机制的编码-解码框架提出 StructCap 模型，通过联合训练视觉分析树、结构语义注意和基于 RNN 的文本描述生成模块来改进图像文本描述的性能。Li[44]等人提出一种全局-局部注意的图像文本描述方法，通过注意力机制集成图像层的全局表示和对象层的局部表示。Mun[45]等人提出基于文本引导的注意力模型来生成图像的文本描述，采用基于实例的学习方法获取相似图像的文本描述句子集，并通过相似图像的文本描述句子来学习图像相关区域的注意力。

3.3 基于强化学习的方法

强化学习是近年来机器学习和智能控制领域的热点方法，它关注于智能体如何在环境中采取一系列行为，从而获得最大的累积(reward)。

Zhang[46]等人将强化学习应用在图像的文本描述生成中，该过程被看作有限马尔科夫决策过程(Markov decision process，MDP)，决策过程的状态值由 CNN 提取的图像特征和已经生成的文本序列构成。训练过程采用 actor-critic 方法，包括策略网络(policy network)和值网络(value network)，策略网络根据状态值生成一系列决策，值网络根据当前的状态给出策略的 reward。图像文本描述模型首先采用最大似然估计的方法进行预训练，然后使用强化学习再优化。训练过程使用蒙特卡洛抽样，根据采样序列的 CIDEr 或 BLEU 作为 reward 更新目标函数。

Liu[47]等人提出基于强化学习的神经网络结构用于图像的文本描述，训练过程采用策略梯度(policy gradient)的方法，策略梯度方法根据值函数对策略进行改进，从而选取最优策略。实验结果表明，使用 BLEU-4、METEOR、CIDEr 和 SPICE 评测标准组合指导最优化过程，生成的图像文本描述质量优于传统方法。

Ren[48]等人提出基于决策框架的图像文本描述方法，利用强化学习中的策略网络和值网络共同来确定执行每次决策的下一个单词的输出。策略网络根据当前状态预测下一个单词的概率，值网络根据预测值给出 reward，reward 函数采用视觉语义嵌入(visual semantic embedding)的形式，这种形式能够评判图像和句子的相似度，可以作为最终优化的全局目标，这两种网络的参数通过基于 actor-critic 的强化学习算法训练得到。

3.4 小结

基于编码—解码的方法生成的句子具有多样性，不依赖于单一的语言模板，有时甚至可以推理出图像中未出现的内容，例如，火车站候车室中的人们正在等候火车，虽然图像中并未出现火车，但模型能够基于图像的情境信息进行推理。

4 数据集和评测标准

公开的数据集和评测标准对于推动图像的文本描述方法研究起着至关重要的作用。本节将对现有比较有影响力的数据集、评测标准和评测组织进行小结。

4.1 数据集

当前图像的文本描述数据集主要包括英文、德文、日文和中文数据集。英文数据集包括 IAPR-TC12[49]、PASCAL[50]、Flickr8k[22]、SBU[15]、MS COCO[3]、Flickr30k[51]、Visual Genome[52]和 Multi30k[53]；德文数据集包括 IAPR-TC12[49]和 Multi30k[53]；日文数据集有 STAIR[54]；中文数据集有 Flickr8kCN[29]和 AIC-ICC[55]。数据集的发表年份如图 2 所示，从发表年份来看，首先出现英文数据集，然后其他研究者逐渐开始构建德文数据集、日文数据集以及中文数据集。数据集的具体统计情况如表 1 所示。

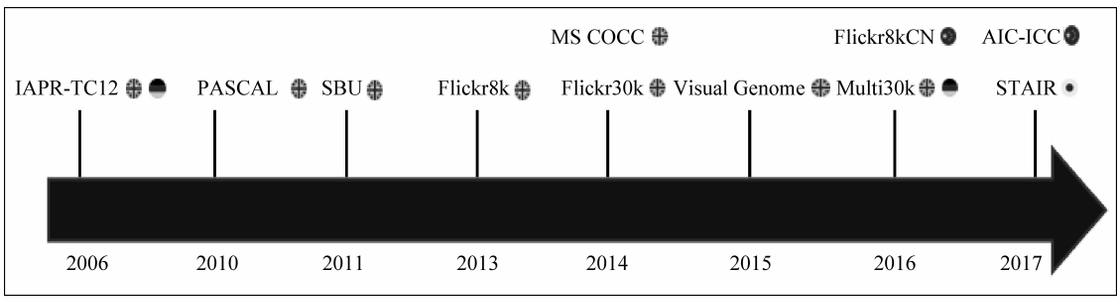

图 2　各图像文本描述数据集发表年份

表 1　图像文本描述数据集的统计信息

数据集	规模	语言	标准划分
Flickr8k	8 000	英	有
Flickr30k	30 000	英	有
MS COCO	82 783	英	有
SBU	1 000	英	无
Multi30k	31 014	英、德	有
PASCAL	1 000	英	无
IAPR-TC12	20 000	英、德	无
Flickr8kCN	8 000	中	有
AIC-ICC	300 000	中	有
STAIR	82 783	日	有
Visual Genome	108 077	英	无

4.2　评测标准

面向图像文本描述方法的评测标准主要包括四大类，分别是主流评测标准、概率评测标准、检索评测标准以及多样性评测标准(图 3)。下面将对这四种评测标准分别进行介绍。

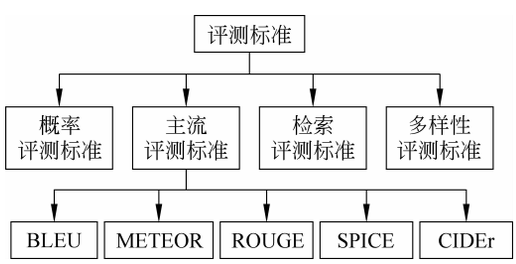

图 3　图像文本描述的评测标准分类

4.2.1　主流评测标准

目前的研究多采用生成的文本描述句子和参考句子之间的匹配程度来评价图像文本描述结果的优劣，我们将采用这类方法的评测标准归为主流评测标准。包括 BLEU(bilingual evaluation understudy)[56]、METEOR(metric for evaluation of translation with explicit ordering)[57]、ROUGE(recall-oriented understudy for gisting evaluation)[58]、CIDEr(consensus-based image description evaluation)[59] 和 SPICE(semantic propositional image caption evaluation)[60]五种衡量指标。其中 BLEU 和 METEOR 两种指标来源于机器翻译，ROUGE 来源于文本摘要，而 CIDEr 和 SPICE 是依据图像文本描述任务定制的指标。

BLEU 是基于 N-gram 共现统计的评测标准。给定生成的图像文本描述 s 和多个人工标注的参考文本描述 R_i，图像—文本描述对(i,s)的 BLEU 值是指在 n 元模型下，图像文本描述 s 在参考文本描述 R_i 上的查准率。

ROUGE 与 BLEU 类似，它是基于查全率的相似度衡量方法，根据计算方法的不同又可分为 ROUGE-N、ROUGE-L、ROUGE-W、ROUGE-S。其中 ROUGE-N 基于 N-gram 计算查全率；ROUGE-L 基于最大公共序列(longest common subsequence，LCS)计算查全率；ROUGE-W 与 ROUGE-L 类似，基于带权重的最大公共序列计算查全率；ROUGE-S 基于 skip-bigram 度量参考文本描述与预测文本描述的共现统计来计算查全率。

CIDEr 是基于共识的评测标准，计算 n 元语言模型(N-gram)在参考描述句子和模型生成待评测句子的共现概率。其目标是计算图像 I 的生成的测评句子 c_i 与 m 个参考描述 $S_i=\{s_{i,1},\cdots,s_{i,m}\}$ 的一致性。研究证明，CIDEr 与人的共识的匹配度好于其他评测标准。

METEOR 用于计算图像描述句子和参考描述句子的相似程度，考虑了单词精确匹配、词干、同义词和释义等因素，其计算基于单精度加权调和平均和单字查全率，相比于基于查全率的 BLEU 评测标准，METEOR 结果与人工判别结果更具有相关性。

SPICE 考虑语义命题内容(semantic proposi-

tional content),图像的文本描述应包含图像中存在的各个语义命题。SPICE 通过将生成的描述句子和参考句子均转换为基于图的语义表示,即场景图,来评价图像文本描述的质量。场景图提取自然语言中词法和句法信息,显式地表示出图像中包含的对象、属性和关系。场景图的计算过程包含两个阶段:使用预先训练的依存语法器建立依存句法树;采用基于规则的方法将依存句法树映射为场景图。

4.2.2 概率评测标准

概率评测标准采用困惑度来评价图像文本描述的生成质量,困惑度也是语言模型常见的评测标准,计算困惑度的公式定义如式(1) 所示。

$$\text{perplexity}(P,C,I) = 2^{H(P,C,I)}$$

$$H(P,C,I) = \frac{1}{|C|}\sum_{n=1}^{|C|}\log_2 P(C_n \mid C_{0,\cdots,n-1},I) \quad (1)$$

这里,P 是已知前 $n-1$ 个单词得到下一单词的概率,C 为包含 $|C|$ 个单词的图像文本描述句子,I 是 C 所描述的图像,H 是熵函数。C_n 是 C 中的第 n 个单词,$C_{0,\cdots,n-1}$ 是从句子起始标识符开始的 $n-1$ 个单词。为了得到整个测试集的困惑度,可以取测试集中所有图像描述句子的算术均值、几何均值和所有图像描述句子的困惑度的中值。

4.2.3 多样性评测标准

生成图像文本描述时,多样性评测标准使用了词汇的多样性。如果图像文本描述方法每次产生的文本描述都是一样的,则这个图像文本描述方法具有最低多样性。多样性评测标准定义为式(2)所示。

$$\text{diversity}(P,F) = -\sum_{i=1}^{|F|} P_i(F)\log_2 P_i(F)$$

$$P_i(F) = \frac{F_i}{\sum_{j=1}^{|F|} F_j} \quad (2)$$

F 是 1-gram 或 2-gram 的极大似然概率估计,$|F|$ 是 1-gram 或 2-gram 的个数,F_n 是第 n 个 1-gram 或 2-gram 的频率,熵度量频率分布的均匀程度,熵越高,分布越均匀。分布越均匀,1-gram 或 2-gram 更可能等比例出现,而在大多数时候不会只使用很少的几个单词,此时,图像文本描述中出现的单词的变化会更大,从而使得文本描述具有更大的多样性。

4.2.4 检索评测标准

许多模型采用基于检索的方法生成图像的文本描述,检索评测标准能够很好地衡量基于视觉空间的检索方法和多模态空间的检索方法的性能。检索评测标准常用的指标是正确率和召回率。正确率是衡量某一检索方法信号噪声比的指标,即相关结果占全部结果的比率。召回率是衡量检索方法检出相关结果成功度的一项指标,即检出相关结果占所有相关结果的百分比。

4.3 评测组织

图像的中文文本描述评测是“AI challenger 全球挑战赛”的五项评测内容之一,由创新工场、搜狗、今日头条三方于 2017 年联合首次主办①。该评测的主要任务是针对给定的每一张测试图片输出一句话的描述,要求描述句子符合自然语言习惯,涵盖图像中的重要信息,如主要人物、场景、动作等内容。对参加评测的系统从客观指标(BLEU, METEOR, ROUGE-L 和 CIDEr)和主观指标(Coherence, Relevance, Helpful for Blind)进行评价。来自清华大学的胡晓林团队获得 2017 年该竞赛任务的冠军,在 AIC-ICC 的测试数据集 B 上取得 BLEU-4、CIDEr、METEOR 和 ROUGE-L 值分别为 0.746 57、2.145 95、0.431 9 和 0.721 72。

Microsoft COCO Image Captioning Challenge② 是微软于 2015 年推出的图像英文文本描述评测,迄今共有 103 个队伍参加。参加评测的系统通过评测 API 平台提交图像在 MS COCO 测试数据集的英文文本描述结果。该平台将实时展示提交系统的排名。截至 2018 年 2 月底,来自腾讯的 TencentAI 团队暂排系统的第一名,在 C5 数据集上取得 BLEU-1、BLEU-2、BLEU-3、BLEU-4、METEOR、ROUGE-L 和 CIDEr-D 值分别为 0.811、0.657、0.508、0.386、0.286、0.587 和 1.254。

5 主要开源工具包简介

基于图像文本描述方法的介绍,对目前的主要开源工具包进行简介,如表 2 所示。

表 2 图像文本描述的主要开源工具包简介

名称	简介
Google NIC[27]	基于 GoogleNet 和 LSTM 构建的图像文本描述方法

① https://challenger.ai/datasets/caption

② https://competitions.codalab.org/competitions/3221#results

续表

名称	简介
Show Attend and Tell[38]	基于注意力机制的方法，能够对图像和摘要进行显式对齐
Neural-image-captioning[61]	基于场景上下文和图像区域的图像文本描述方法
m-RNN[35]	基于 CNN＋RNN 多模态图像文本描述方法
Review Net[41]	基于 Review Network 的图像文本描述方法
NeuralTalk[39]	基于 CNN＋RNN 实现的图像文本描述方法
e2e-gLSTM-sc[40]	基于语义层的注意融合单词和图像信息的文本描述方法
SCA-CNN[62]	基于空间和通道注意力的图像文本描述方法

6　关键问题及研究难点

综上所述，虽然图像的文本描述研究已经取得显著效果，但对于诸如图像的视觉概念提取、图像与文本模态融合、图像的跨语言文本描述等子任务的性能仍有待改进。本节针对现有的图像文本描述尚存的关键问题和研究难点予以介绍。

(1) 图像的视觉概念提取

图像的文本描述是视觉与语言结合的新任务，其性能的提升离不开视觉与语言本身的技术突破。图像的视觉概念包括图像类别、场景信息、检测对象、对象属性和对象关系等，视觉概念的提取依赖于计算机视觉技术，目前还不十分成熟。而视觉概念的提取是生成图像文本描述的重要基础，直接决定图像文本描述的性能。因此，图像的视觉概念提取是图像的文本描述中待解决的关键问题及研究难点。

(2) 图像与文本模态融合

图像的文本描述首先要解决的是语义鸿沟问题，即用单纯的图像视觉特征信息在图像内容的表达上存在多义性和不确定性问题。图像中常常隐式或显式包含文本信息，充分利用与图像数据共现的文本信息，进行多模态的语义分析和相似性度量，是克服语义鸿沟的有效方法。目前已有基于深度神经网络的多种融合方法(见 3.1)，但并未真正深入到图像与文本在高层语义的融合问题，因此如何对图像和文本模态信息进行多模态高层语义融合是图像的文本描述中待解决的关键问题及研究难点。

(3) 图像的跨语言文本描述

现有的图像文本描述方法通常采用基于深度学习或机器学习的方法，然而，当有标记的训练样本非常少时，这种方法的效果往往较差。而在实际应用中，要求针对图像能够给出多种语言文字的文本描述来满足不同母语的用户需求。目前图像英文和中文文本描述的训练样本较多，其他语言文本描述对应的标记训练样本较少，若对图像的每一种语言文本描述进行人工标记将需要耗费大量的人力和时间。因此，如何实现图像的跨语言文本描述是图像的文本描述中待解决的关键问题及研究难点。

7　结束语

图像的文本描述近几年得到研究界和企业界的广泛关注，它借助深度学习技术为视觉和语言搭建的桥梁获得了突飞猛进的发展，其跨越了视觉和语言的领域界限，把直观上的感知提升到了认知的概念范畴。图像的文本描述能够提高基于内容的图像检索效率，扩大在医学、安全、军事等领域的可视化理解应用范围，具有广阔的应用前景。同时，图像文本描述的理论框架和研究方法可以推动图像标注和视觉问答的理论和应用的发展，具有重要的学术和实践应用价值。

图像的文本描述，不仅需要理解视觉，也需要知道如何对语言进行建模。当前的主要解决方案是端到端的黑盒子式深度学习，并未真正深入到视觉与语言的本质问题。如何进行视觉与语言的深度语义融合，将有助于提升图像文本描述的性能，这也是多模态智能交互的关键步骤，是未来的主要发展方向。

参考文献

[1] Bernardi R, Cakici R, Elliott D, et al. Automatic description generation from images: A survey of models, datasets, and evaluation measures[J]. J. Artif. Intell. Res. (JAIR), 2016(55): 409-442.

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

[3] Lin T, Maire M, Belongie S, et al. Microsoft Coco: Common objects in context[C]//Proceedings of European Conference on Computer Vision, 2014: 740-755.

[4] Kulkarni G, Premraj V, Dhar S, et al. Baby talk: Understanding and generating simple image descrip-

tions[C]//Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition, 2011: 1601-1608.

[5] Kuznetsova P, Ordonez V, Berg T, et al. TREETALK: Composition and compression of trees for image descriptions [J]. TACL, 2014, (2): 351-362.

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

[7] Elliott D, Vries A. Describing images using inferred visual dependency representations[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, 2015: 42-52.

[8] Elliott D, Keller F. Image description using visual dependency representations[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013: 1292-1302.

[9] Mitchell M, Dodge J, Goyal A, et al. Midge: Generating image descriptions from computer vision detections [C]//Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012: 747-756.

[10] Li S, Kulkarni G, Berg T, et al. Composing simple image descriptions using Web-scale N-grams[C]// Proceedings of the 15th Conference on Computational Natural Language Learning, CfoNLL 2011. Portland, Oregon, USA, 2011: 220-228.

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

[12] Torralba A, Fergus R, Freeman W. 80 million tiny images: A large data set for nonparametric object and scene recognition[J]. IEEE TPAMI, 2008, 30(11): 1958-1970.

[13] Kuznetsova P, Ordonez V, Berg A, et al. Collective generation of natural image descriptions[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju Island, Korea, 2012: 359-368.

[14] Verma Y, Gupta A, Mannem P, et al. Generating image descriptions using semantic similarities in the output space[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2013. Portland, OR, USA, 2013: 288-293.

[15] Ordonez V, Kulkarni G, Berg T. Im2Text: Describing images using 1 million captioned photographs [C]//Proceedings of Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Granada, Spain: NIPS, 2011: 1143-1151.

[16] Patterson G, Xu C, Su H, et al. The SUN Attribute Database: Beyond Categories for Deeper Scene Understanding[J]. International Journal of Computer Vision, 2014,108 (1-2): 59-81.

[17] Mason R, Charniak E. Nonparametric method for 143 image captioning[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014. Baltimore, MD, USA, 2014: 592-598.

[18] A Nenkova A,L Vanderwende L. The impact of frequency on summarization[R]. Microsoft Research, 2005.

[19] Yagcioglu S, Erdem E, Erdem A, et al. A distributed representation based query expansion approach for image captioning[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015. Beijing, China, 2015: 106-111.

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

[21] Devlin J, Cheng H, Fang H, et al. Language models for image captioning: The quirks and what works [C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015. 2015: 100-105.

[22] 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): 853-899.

[23] Socher R, Li F. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora[C]//Proceedings of the 23rd IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010. San Francisco, CA, USA, 2010: 966-973.

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

[25] Karpathy A, Joulin A, Li F. Deep fragment embeddings for bidirectional image sentence mapping[C]//

Proceedings of Advances in Neural Information Processing Systems 27：Annual Conference on Neural Information Processing Systems. Montreal，Quebec，Canada，2014：1889-1897.

[26] Kiros R，Salakhutdinov R，Zemel R. Unifying visual-semantic embeddings with multimodal neural language models[C]//Proceedings of Advances in Neural Information Processing Systems Deep Learning Workshop，2015.

[27] Vinyals O，Toshev A，Bengio S，et al. Show and tell：A neural image caption generator[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition，2015：3156-3164.

[28] Xu J，Gawes E，Fernando B，et al. Guiding the long-short term memory model for image caption generation[C]//Proceedings of 2015 IEEE International Conference on Computer Vision，ICCV 2015. Santiago，Chile，2015：2407-2415.

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

[30] Szegedy C，Liu W，Jia Y，et al. Going deeper with convolutions[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition，CVPR 2015. Boston，MA，USA：IEEE Computer Society，2015：1-9.

[31] Donahue J，Hendricks L，Guadarrama S，et al. Long-term recurrent convolutional networks for visual recognition and description[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition，2015：2625-2634.

[32] Gan Z，Gan C，He X，et al. Semantic compositional networks for visual captioning[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu，Hawaii，2017：5630-5639.

[33] Rennie S，Cui X，Goel V. Efficient non-linear feature adaptation using maxout networks[C]//Proceedings of 2016 IEEE International Conference on Acoustics，Speech and Signal Processing. Shanghai，China，2016：5310-5314.

[34] Kiros R，Zemel R，Salakhutdinov R. A multiplicative model for learning distributed text-based attribute representations [C]//Proceedings of Advances in Neural Information Processing Systems 27：Annual Conference on Neural Information Processing Systems. Montreal，Quebec，Canada，2014：2348-2356.

[35] Mao J，Xu W，Yang Y，et al. Deep captioning with multimodal recurrent neural networks（m-rnn）[J]. arXiv preprint arXiv：1412.6632，2014.

[36] Hendricks L，Venugopalan S，Rohrbach M，et al. Deep compositional captioning：Describing novel object categories without paired training data[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas，NV，USA，2016：1-10.

[37] Tanti M，Gatt A，Camilleri K. What is the role of recurrent neural networks（RNNs）in an image caption generator[J]. arXiv preprint arXiv：1708.02043，2017.

[38] Xu K，Ba J，Kiros R，et al. Show，attend and tell：Neural image caption generation with visual attention [C]//Proceedings of the 32nd International Conference on Machine Learning，2015：2048-2057.

[39] Andrej K，Li F. Deep visual-semantic alignments for generating image descriptions [C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston，MA，USA，2015：3128-3137.

[40] Zhou L，Xu C，Koch P，et al. Watch what you just said：Image captioning with text-conditional attention [J]. arXiv preprint arXiv：1606.04621，2016.

[41] Yang Z，Yuan Y，Wu Y，et al. Review networks for caption generation[C]//Proceedings of Advances in Neural Information Processing Systems 29：Annual Conference on Neural Information Processing Systems，2016：2361-2369.

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

[43] Chen F，Ji R，Su J，et al. StructCap：structured semantic embedding for image captioning[C]//Proceedings of the ACM Multimedia，Mountain View，CA USA，2017：46-54.

[44] Li L，Tang S，Deng L，et al. Image caption with global-local attention[C]//Proceedings of the 31st AAAI Conference on Artificial Inteligence. San Francisco，Califonrial USA，2017：4133-4139.

[45] Mun J，Cho M，Han B. Text-guided attention model for image captioning[C]//Proceedings of the 31st AAAI Conference on Artificial Inteligence. San Francisco，Califonrial USA，2017：4233-4239.

[46] Zhang L，Sung F，Liu F，et al. Actor-critic sequence training for image captioning[J]. arXiv preprint arXiv：1706.09601，2017.

[47] Liu S，Zhu Z，Ye N，et al. Improved image captioning via policy gradient optimization of Spider[C]//Proceedings of the International Conference on Computer Vision，2017：873-881.

[48] Ren Z，Wang X，Zhang N. Deep reinforcement

learning-based image captioning with embedding reward [J]. arXiv preprint arXiv:1704.03899, 2017.

[49] Grubinger M, Clough P, et al. The IAPR TC-12 benchmark: A new evaluation resource for visual information systems[C]//Proceedings of the International Conference on Language Resources and Evaluation, 2006: 13-23.

[50] Rashtchian C, Young P, Hodosh M, et al. Collecting image annotations using amazon's mechanical turk[C]//Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 2010: 139-147.

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

[52] Krishna R, Zhu Y, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations [J]. International Journal of Computer Vision, 2016, 123(1): 32-73.

[53] Elliott D, Frank S, Sima'an K, Multi30K: Multilingual English-German image descriptions [C]//Proceedings of the 5th Workshop on Vision and Language, 2016: 70-74.

[54] Yoshikawa Y, Shigeto Y, Takeuchi A, STAIR captions: Constructing a large-scale Japanese image caption dataset [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 417-421.

[55] Wu J, Zheng H, et al. AI challenger: A large-scale dataset for going deeper in image understanding. arXiv preprint arXiv:1711.06475, 2017.

[56] Papineni K, Roukos S, Ward T, et al. BLEU: A method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002: 311-318.

[57] Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments [C]//Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005: 65-72.

[58] Lin C. ROUGE: A package for automatic evaluation of summaries[C]//Proceedings of the ACL Workshop, 2004: 25-26.

[59] Vedantam R, Zitnick C, Parikh D. CIDEr: Consensus-based image description evaluation[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition, 2015: 4566-4575.

[60] Anderson P, Fernando B, Johnson M, et al. SPICE: Semantic propositional image caption evaluation[C]//Proceedings of European Conference on Computer Vision. Springer International Publishing, 2016: 382-398.

[61] Fu K, Jin J, Cui R, et al. Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts [J]. IEEE TPAMI, 2017, 39(12): 2321-2334.

[62] Chen L, Zhang H, Xiao J, et al. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 6298-6306.

马龙龙(1980－),博士,副研究员,主要研究领域为多模态信息处理与自然语言处理。
E-mail: longlong@iscas.ac.cn

孙乐(1971－),博士,研究员,主要研究领域为信息检索与自然语言处理。
E-mail: lesunle@163.com

韩先培(1984－),博士,副研究员,主要研究领域为信息抽取、知识库构建以及自然语言处理。
E-mail: hanxianpei@qq.com