文章编号：1003-0077(2020)07-0019-11

从视觉到文本：图像描述生成的研究进展综述

魏忠钰[1]，范智昊[1]，王瑞泽[2]，承怡菁[1]，赵王榕[1]，黄萱菁[3]

(1. 复旦大学 大数据学院，上海 200433；
2. 复旦大学 工程与应用技术研究院，上海 200433；
3. 复旦大学 计算机科学与技术学院，上海 200433)

摘 要：近年来，跨模态研究吸引了越来越多学者的关注，尤其是连接视觉和语言的相关课题。该文针对跨视觉和语言模态研究中的核心任务——图像描述生成，进行文献综述。该文从基于视觉的文本生成框架、基于视觉的文本生成研究中的关键问题、图像描述生成模型的性能评价和图像描述生成模型的主要发展过程四个方面对相关文献进行介绍和总结。最后，该文给出了几个未来的重点研究方向，包括跨视觉和语言模态的特征对齐、自动化评价指标的设计以及多样化图像描述生成。

关键词：图像描述生成；跨模态特征对齐；文献综述

中图分类号：TP391 **文献标识码**：A

From Vision to Text: A Brief Survey for Image Captioning

WEI Zhongyu[1], FAN Zhihao[1], WANG Ruize[2], CHENG Yijing[1],
ZHAO Wangrong[1], HUANG Xuanjing[3]

(1. School of Data Science, Fudan University, Shanghai 200433, China;
2. Academy for Engineering and Technology, Fudan University, Shanghai 200433, China;
3. School of Computer Science and Technology, Fudan University, Shanghai 200433, China)

Abstract: In recent years, increasing attention has been attracted to the research field related to cross-modality, especially vision and language. This survey focuses on the task of image captioning and summarizes literatures from four aspects, including the overall architecture, some key questions for cross-modality research, the evaluation of image captioning and the state-of-the-art approaches to image captioning. In conclusion, we suggest three directions for future research, i.e., cross-modality representation, automatic evaluation metrics and diverse text generation.

Keywords: image captioning; cross-modality alignment; literature review

0 引言

近年来，结合图像和文本的跨模态研究越来越多地引起自然语言处理和机器视觉领域学者的关注[1-3]。主流的任务包括图像描述生成[4-5]、机器视觉问答[6]、相册故事生成[7]、视觉对话生成[8]、视觉推理[9]、视觉导航[10]，以及从文本到图像的自动生成[11]等。相关的技术在很多实际应用场景中有很大的价值，包括构建能够对视觉信息进行处理的智能聊天机器人；在社交媒体上，为图片或者相册自动产生描述；在幼儿教育中扮演讲师的角色；帮助视觉障碍人士感知周边环境的视觉内容等。得益于深度学习的最新进展，视觉—文本的跨模态研究在很多应用中取得了在自动化评价指标上的大幅度进步。本文将针对基于视觉的文本生成领域的最新研究进展进行综述。鉴于图像描述生成是目前受到关注最多的应用任务，同时包括了基于视觉的文本生成任

收稿日期：2019-03-16 **定稿日期**：2019-04-26

基金项目：国家自然科学基金(71991471)；国家社会科学基金(20ZDA060)；上海市科学技术委员会(18DZ1201000，17JC1420200)

务主要的技术模块，本文以图像描述生成作为切入点，从基于视觉的文本生成框架、基于视觉的文本生成的关键问题、图像描述生成模型的性能评价和图像描述生成模型的主要发展过程四个方面对相关的文献进行梳理。

1 基于视觉的文本生成框架

基于视觉的文本生成框架主要可以分成三类：早期的基于模板匹配和检索的框架、基于神经网络的端到端自动生成框架以及融合神经网络的组合式生成框架。

1.1 基于模板匹配和检索的框架

基于视觉的文本生成的早期方法大致可以分为两类。第一类是基于模板匹配的方法[12-14]。该类方法先从图像中检测出相关的实体信息（物体、属性、动作），然后将这些实体填充到手工设计的句子模板中，存在的问题是生成的文本往往欠缺流利性，在表达的多样性方面也不能满足实际需求。第二类是基于检索的方法[15-16]。该类方法预先准备一批与目标图像视觉上相似的图像—文本匹配语料集合，然后在该语料集合上搜索与目标图片最相近的样本，直接采用查询到的图片所对应的文本作为需要生成的描述。该类方法的问题是依赖预先准备的图像—文本匹配的语料集合，对于内容上差异大的图片往往很难找到相关的匹配对象，因此在生成精度上存在很大的缺陷。

1.2 端到端的自动生成框架

针对早期框架所产生的文字存在流畅性欠缺和不能产生新内容的缺点，基于深度神经网络的端到端模型可以潜在地解决这两个问题。端到端的学习框架[17]包括编码器和解码器两个基本部件，其中编码器将输入的信息转换成低维稠密的隐式向量，再通过解码器将该隐式向量转换成目标输出。该学习范式最早应用在机器翻译的场景，并取得了显著的性能提升。研究者于是将端到端的学习框架引入到针对视觉信息的自动化描述任务中[4]。在该框架中，编码器由在大规模图像分类中被证明有效的多层卷积神经网络（CNN）构成，以实现对图像进行视觉信息的特征提取[18-19]。解码器则由在语言模型构建方面表现良好的循环神经网络（RNN）构成，包括长短时记忆网络[20]（LSTM）和带门控机制的循环神经网络[21]（GRU），RNN对于句子的生成依循逐字的方式进行，以图像特征作为初始状态，每一个时间片使用前一个时间片产生的隐藏状态和生成单词作为输入，产生这个时间片的单词。最近，也有学者在编码器和解码器端使用transformer作为文本生成的基本部件[22]。当前，端到端的生成方式是基于视觉的文本生成任务的主流框架，不同模型的差别在于采用不同的视觉特征抽取方式，以及采用RNN的不同变种进行自动化的文本生成。基于端到端的生成框架在生成文本的流畅性方面有很大的提高，但是视觉和文本的关联仅仅通过中间的隐藏表示构建，这使得生成的文本包含一些不可控的结果，如生成与图片信息无关的文字。

1.3 融合神经网络的组合式的框架

针对端到端框架会产生不确定性生成结果的缺点，学者研究组合式的框架进行图像描述的自动化生成。该框架主要包含两个部件，视觉语义提取和结合视觉语义的描述生成。文献[23]的研究首先从图像中发现一组语义概念，以名词、动词和形容词的形式表征。基于语义概念，语言模型被用来生成多个候选描述。最后，多模态的相似度计算模型为候选描述进行重要性排序，并从中选择得分最高的作为输出。文献[24]沿用了早期基于模板匹配的图像文本生成的思路，采用神经网络模型对两个基础模块进行替换，以达到对传统的基于模板匹配的方法和基于神经网络的端到端方法的调和作用。该方法首先从整体的视觉信息中自动化地构建句子“模板”，模板的每一个空格可以关联到图像中的一个局部区域。在第二个步骤中，从槽位相关联的局部图像中进行物体识别，并将检测的物体填充到模板中作为最后的描述。文献[25]进一步修正文本生成模块，抛弃了基于RNN的文本生成组件。该框架包含两个步骤，首先，从视觉信息中发现一些明确的语义表示单元，以短语的形式表示。其次，用短语拼接的方法来形成最后的图像描述。组合式的框架在生成文本的准确性上相较单纯的端到端模型有一定程度的性能提升，但是非端到端模型在生成文本的流畅性方面则有不可避免的缺陷。这也是目前针对该类方法进行提升的主要方向。

2 基于视觉的文本生成的关键问题

基于视觉信息的文本生成主要包含四个核心的

研究问题：视觉端的特征表示、视觉与文本的特征对齐、加入强化学习的图像描述生成，以及多样化图像描述生成。

2.1　视觉端的特征表示

当前研究在视觉端的特征提取，往往以两种基本形式存在：视觉表示和文本概念。视觉表示代表从图像中直接提取的特征信息。早期的工作，将图像划分为大小相等的视觉区域，再基于 CNN[26] 对图像区域进行顺次的卷积处理，过程中不编码更细粒度的信息。按照指定大小划分得到的图像区域，往往难以灵活捕捉图像中包含的实体信息，这限制了对于图像端的语义理解。为了更好地进行图片信息的特征提取，研究者开始采用 R-CNN[27] 进行视觉端的信息处理，该方法采用物体检测模型作为特征提取器，在图片中划定大小不等的边框，并从中发现实体信息，作为视觉信息的表示。该方法依赖于物体检测模型的实体识别能力，同时也受限于物体检测模型所使用的语料集合中包含的标注标签个数。基于 R-CNN，研究者进一步设计了在实际应用中更高效的 Faster R-CNN[28]，并得到更广泛的使用。仅仅依靠图像处理的视觉信息提取方法建构视觉信息特征无法解决视觉信息与文本之间存在的语义鸿沟。为了在视觉端的特征提取中考虑语义信息，相关研究[29-31]将图像的语义概念识别转换成多标签分类问题，采用单词、短语作为语义的表示单元，并通过不同的方式（注意力机制，嵌入到循环神经网络的解码单元中）将这一组语义单元作用到文本解码的过程中。最近，有研究者[32]引入场景图的方法将视觉特征和语义信息相结合作为图片端的特征表示。场景图中的节点代表视觉信息中发现的视觉实体，而场景图的边则是基于语义信息的实体关系。然而，为了构造场景图，需要复杂的流水线，并且不能避免错误传播。这在某种程度上限制了场景图在更多场景进行推广。

2.2　视觉与文本的特征对齐

在跨模态的相关研究中，核心部件是不同模态信息的联合表示学习。在端到端的学习框架中[4]，基于卷积神经网络的视觉特征抽取模块将图像信息表征成低维稠密向量，而基于循环神经网络的文本生成部件则从该低维稠密向量中逐字生成图像相关的描述。这个过程假设了整张图片的信息和待生成文本的信息共享了一个隐空间，以低维稠密向量表示。卷积神经网络和循环神经网络的参数在一个联合训练的框架下完成。文献[5]使用两个任务分别针对图像特征提取和循环神经网络的参数进行训练。在卷积神经网络部分，一个图像特征和句子特征对齐的任务被构造用来进行参数学习。在循环神经网络部分，图像特征提取器的参数被固定，句子生成任务被用来进行参数学习。在这种句子生成框架中，图像特征以隐状态的形式仅仅直接影响首个单词的生成，对于句子中其他单词生成的影响则是间接的。

随着句子长度的增加，图像特征对于单词生成的影响慢慢淡化，导致句子的生成更多地受到语言模型的影响，而不能很好地描绘图像中的具体信息。为了进一步关联局部图像特征和句子中字词的生成过程，采用注意力机制来进行基于视觉信息的文本生成任务。注意力机制最早在机器翻译领域被提出作为编码器—解码器框架的一个补充部件[33]，在解码器生成单词时，用来在输入序列中寻找最能提供辅助信息的序列单元部件。文献[34]引入注意力机制连接文本生成过程和图片中的局部区域特征，在解码某个单词时，解码器计算图片局部区域对于该单词的决策权重，并采用加权平均的方法引入图片区域特征来计算单词的生成概率分布。学者将这种关联图像区域特征和文本生成的注意力框架称为自顶向下的方案，而将关联图像中的实体特征和文本生成的方法称为自底向上的方案。文献[35]结合了自顶向下和自底向上两种模式，自底向上的机制采用 Faster R-CNN[28] 从图像中选取有显著意义的区域，在解码每个单词的时候，使用其关联的局部图像特征的重要性权重，重新调整 Faster R-CNN 发现的实体的重要性，用来计算单词的生成概率向量。

随着预训练模型在视觉和文本单一模态场景中的成功应用，如 BERT[36]、ResNet[37] 等，学者开始研究结合视觉和文本的预训练模型。基本的研究思路借鉴 BERT 等的预训练模型，将视觉与语言的混合表示以序列的方式输入到基于 transformer 的框架中，然后依照自监督的方式进行优化。到目前为止，出现了 VisualBert[38]、Unicoder-VL[39]、VL-BERT[40]、ViLBERT[41]、LXMERT[42] 和 UNITER[43] 等研究工作。依据处理文本和图片的方式，相关工作可以分成两大类：单流编码（VisualBert、Unicoder-VL、VL-BERT 和 UNITER）和双流编码（ViLBERT 和 LXMERT）。单流编码将图片和句子拼接成一个序

列,输入到同一个编码器中,同时对两种模态的信息进行编码。双流编码则认为图片和文本的底层表示有着不同的特性,所以先采用不同的编码器对图片和文本进行单模态编码,之后再通过互注意力机制对两种模态进行联合编码。这些预训练模型采用的自监督训练任务包括,遮盖语言模型、遮盖区域分类/回归、视觉文本匹配以及视觉问答。遮盖语言模型和BERT的训练任务基本类似,但在推断被遮盖的单词时,更希望模型学会从视觉信息中寻找线索。遮盖区域和遮盖语言的本质是相同的,只不过遮盖的部件从字符变成了某个图片区域。视觉文本匹配任务和BERT中的后续句子推断任务相似,通过构造负样例的方式来判断文本和图片是否匹配。虽然跨模态的预训练模型越来越多地引起学者的关注,但是在跨模态文本生成方面的应用还没有起步。鉴于跨模态的联合表示学习在当前的图像描述生成研究中并未引起足够关注,未来如何将跨模态的预训练模型融入到图像描述生成中会是一个研究的重点。

2.3 引入强化学习的图像描述生成

在图像描述生成任务中,基于神经网络的文本生成解码器将每一个单词的产生看成一个分类问题,并使用交叉熵损失函数来进行文本生成模型的训练。这样的模型训练方式存在两个问题,一个被称为暴露偏置(exposure bias),即在训练时,模型当前时刻的输入是来自训练集的真实单词,而在测试时,输入的却是上一时刻的预测结果,一旦模型单步表现不佳,就会导致误差累积,从而影响整体的生成效果;另一个问题被称为损失评估失配(loss-evaluation mismatching),即在训练时,模型采用交叉熵损失函数来评估生成结果的好坏,而在测试时,却采用其他的自动化评价指标(见3.2节),存在训练和测试评估方式不一致的问题,从而影响模型在测试时的表现。

为了解决上述问题,文献[44]引入强化学习进行模型训练,以解决模型训练和测试过程的不一致问题。一方面,采用计划采样(scheduled sampling)的方法在训练时也采用上一时刻产生的词语,从而解决暴露偏置的问题;另一方面,强化学习可以在训练时通过最大化一些测试时采用的评价指标来完成模型参数的更新,从而解决损失评估失配问题。

强化学习的方法在图像描述任务中的应用需要构造三个关键要素,即状态(state)、动作(action)与奖励(reward)。状态就是解码过程中每个时间片的隐藏状态表示,而动作是对当前时刻生成单词的选择,奖励一般采用自动化的评价指标。基于强化学习的基本方法存在模型训练不稳定的问题。该问题产生的原因是模型执行过程中计算期望梯度时会产生较高的方差,其中一种解决办法是加上基线模型的约束。比如在MIXER[44](mixed incremental cross-entropy reinforce)中,基线就是一个简单的多个采样样本的奖励均值。在SCST[45](self-critical sequence training)中,这个基线采用固定策略采样(贪婪搜索或者束搜索)进行奖励计算。还有一些方法比如actor-critic[46],则训练一个评论(critic)网络来估算奖励。

目前,引入强化学习进行图像描述生成的模型训练方法一般遵循以下流程:先采用交叉熵损失函数进行模型训练,当性能达到一定程度之后,再以自动化评价指标为奖励,使用强化学习进行模型训练。经过强化学习训练过的模型通常可以具有更好的性能,因此当前针对图像描述的模型往往会配置使用强化学习的版本进行有针对的性能比较。

2.4 多样化图像描述生成

不同的人在对同一张图像进行描述时,往往会产生不同的描述语言,为了使机器产生的文本能够具有多样性和创新性,学者开始研究多样化的图像描述生成方法。文献[47]在句子生成阶段产生多个样本,并引入惩罚因子,对生成样本之间重复的词进行惩罚以激励产生更多样化的句子。文献[48]尝试在互相不重叠的数据集分割中单独训练文本生成器,以通过不同的文本生成器来产生多样性的句子。文献[49]在训练的过程中生成多个候选句子,并使用核方法来计算句子之间的相似性,随后将相似性计算模型加入到对抗生成网络中,以激励生成器达成生成的多样化。文献[50]在问题生成这个场景中,将问题类型和图片信息一起建模,以达到产生多样化问题的目的。文献[51]将图片中的物体作为先验信息加入到变分自编码器的隐空间中,引导面向不同物体的描述生成,文献[52]更精细地在变分自编码器的隐空间当中去建模词汇和语法结构。文献[53]认为生成模型之所以缺乏多样性,是因为生成模型会倾向于选择在数据集合中出现频率更高(相对来

说更安全）的词。生成内容的多样性以及图片和内容的相关性其实是一体两面。他们因此通过负采样的方法，训练模型辨识图片和句子的相关性，并通过对抗生成网络来强化生成器在这方面的性能。文献[54]也讨论了相同的问题，并引入了RankGAN，该方法虽然在自动化指标上稍有逊色，但在人工评价上显示出更优越的效果。

文本生成的多样性评估可以从数据集和单个样例两个方面进行。在数据集方面：①计算生成的描述没有出现在训练数据集中的比例，②计算基于图片生成的描述中包含的词汇数量。前一个指标越高，表示生成的描述在创新性方面的得分越低；后者的指标越高，表示生成的描述的多样化程度越高。在单个样例方面，学者提出 Dist-n[55] 和 mBLEU[48] 来度量生成的多个句子之间的相似（不相似）性，从而考量描述的多样化程度。Dist-n 统计针对单个图像生成的多个描述中不重复的 n 元组个数。mBLEU 在为目标图像生成的多个描述中每次选取一个描述，并计算它与其他描述之间的 BLEU 值。因为 BLEU 是基于 n 元组的相似性度量指标，所以 mBLEU 越低说明生成的句子之间的差异性越大，表明生成的描述多样性越好。

3　图像描述生成模型的性能评价

3.1　图像描述生成的评测语料集

适用于图像描述生成任务的数据集主要有：Flickr8k[15]、Flickr30k [56-57] 和 MS COCO [58]。在预训练图像识别或者特征对齐模块时，常用到的数据集有：ImageNet[59] 和 Visual Genome[60]。除此以外，经常被使用的语料集还包括 IAPR TC-12[61-62]、ReferIt[63]、Instagram[64-65]、Stock3M[66]、MIT-Adobe FiveK[67]、FlickrStyle10k[68]等。表 1 呈现了各个语料集的详细介绍。Flickr8k 是第一个公开的大规模图像和描述匹配的语料集，扩充版本 Flickr30k 一共包含了 31 783 张图片，每张图片有 5 个人工产生的描述。在扩充版本中，Flickr30k 还包含了实体标注。MS COCO 在图片规模上有很大的提升，包含了超过 16 万张图片，并且每张图片中有 7.7 个实体标注，一共包括了 80 个实体类别，因此可以针对图像物体检测和描述生成两个应用同时开展研究。Visual Genome 的语料集则有更加细粒度的标注，包含与图片中局部区域相关联的较短的描述，因此可以满足多样化描述生成的研究需求；同时，该语料集包含了场景图的标注信息，一般用来预训练场景图的构建模型。

表 1　视觉文本生成相关语料集

名称	图片数量	描述个数	实体种类	实体个数	场景图
Flickr8k[15]	8 000	5	—	—	—
Flickr30k[56,67]	31 783	5	—	8.9	—
MS COCO[58]	164 062	5	80	7.7	—
ImageNet[59]	14 197 122	—	21 841	1	—
Visual Genome[60]	108 077	50	76 340	16	是
IAPR TC-12[61-62]	20 000	2	255	5.0	—
ReferIt[63]	19 894	6.6	—	4.8	—
Instagram[64-65]	1 100 000	1	—	—	—
Stock3M[66]	3 217 654	1	—	—	—
MIT-Adobe Fivek [67]	5 000	—	—	—	—
Flickr-Style10k[68]	10 000	7	—	8.9	—

3.2　图像描述生成的评价方法

对图像描述生成模型的评价，指的是依据给定的图像判断模型所生成描述的质量。当前的主流评价方法主要包括以下三种类型。

第一，人工评价。人工设计一些评价指标，比如，表达的流畅度、与图片的相关度、表达的多样化程度等[69-70]。人工的评分结果一般通过网上众包的形式进行收集。人工评价的方法可以准确反映图像描述模型的性能，但是操作的过程中需要引入大量的标注者，从而影响了评价的灵活性。

第二，基于规则的自动化评价方法。该方法提前为图片收集固定数量的人工撰写的参考描述，并采用关键词匹配的技术，计算模型生成描述与参考描述之间的重叠程度作为模型的性能评价。主流的指标包括 ROUGE[71]、BLEU[72]、CIDEr[73]、METEOR[74] 和 SPICE [75]。BLEU 被广泛地使用在机器翻译中，其主要计算模型生成的描述与参考描述之间的 n 元组重合程度，重合程度越高，生成描述的

质量就越高。不同元组的选择可以从不同的侧面反映生成描述的质量，单元组(uni-gram)的准确率可以用于评估单词级别的生成准确性，更高阶的元组准确率可以用来判断句子的整体流畅性。该指标更关注生成的准确率，即更关心生成描述里有多少n元组是对的。ROUGE[71]最早被用来评价文本摘要模型的质量，它与BLEU的计算公式非常相似，但它只计算召回率，考虑参考描述中有多少元组被机器生成的描述覆盖。为了在评价过程中考虑相同语义不同表达的句子，学者提出了METEOR[74]，它引入一个外部资源库(WordNet)对词语的同义词进行考虑，同时也考虑单词的词形(stemming)。在评价句子流畅性时，用词块(chunk)作为基本评价单元，考虑了调和召回率和准确率的F值作为最终评价指标。相比BLEU和ROUGE，METEOR引入了外部资源和额外的句子分块算法，这也给其评价结果带来一些不稳定性。CIDEr[73]是针对图像描述任务提出的，它同样采用n元组作为基本评价单元，并采用词频和倒排文档频率(TF-IDF)作为n元组的权重，这样可以降低高频n元组对于结果的影响。最近，学者提出SPICE[75]来考察图像中的实体和实体间关系是否被图像描述生成模型考虑。该指标为机器生成描述构建场景图，并与基于图像构建的场景图进行相关性计算，从而达到衡量生成描述对于图像中包含的实体和实体间关系的覆盖度的目的。

第三，基于学习的自动化评价方法。该方法构建一个机器学习模型直接计算图像描述生成模型产生的描述和给定图片之间的相关度。文献[76]引入对抗生成的方法，训练一个判别器，在给定图片和一个候选描述的情况下，判定有多大概率这个候选描述是人工产生的，分数越高则该描述的质量越高。在训练过程中，他们自动化地为给定图片产生一些不相关的描述作为负样本，用以训练判别器。文献[77]结合了基于学习和规则的方法，综合考量机器生成的描述与图片的相关度，以及机器生成的描述与参考描述的相关度。在生成描述与图片相关度方面，他们采用预训练的视觉和文本的对齐模块，计算机器生成的描述与图片区域之间的关联分布向量。分布向量的集中度越高，则相关度越高。在生成描述与参考描述的相关度计算方面，他们通过图片区域将二者进行关联，从而细粒度地评价生成描述和参考描述之间的相关度。基于学习的自动化评价虽然增加了灵活性，但是评价模型本身是参数化的，因此也有被图像描述模型攻击和欺骗的风险。

4 图像描述生成模型的主要发展过程

近年来，基于神经网络的端到端模型作为主流的图像描述生成模型，在公开评测语料集上不断刷新各种评测指标的记录。因此，在本节中，我们主要针对这一类别的方法进行发展进程的介绍。早期的研究工作主要集中于通过基于卷积神经网络(CNN)和循环神经网络(RNN)的基础端到端框架来生成图像描述[4-5, 23, 78-79]；随后，针对多模态特征对齐问题，研究人员提出了不同的注意力机制，从图像特征和语言特征方面对图像标注进行改进[31-32,34-35,46,80-82]；同时，不同于使用交叉熵作为优化目标的方法，一些研究人员采用强化学习方法，将自动评价指标(一般选用CIDEr或METEOR)作为优化目标来训练模型[16,32,35,46,81-84]。将自动评价指标作为优化目标已经成为目前图像描述生成工作中主流的实验设置。最近一些研究工作通过目标检测算法(如Faster R-CNN)提取图像中的物体区域，来引入更为丰富多样的信息，如物体、属性和关系等[24,32,35]。随着预训练语言模型在自然语言处理领域的兴起，也出现了跨模态预训练模型的工作。比如，Zhou等人[85]提出了一种视觉—语言预训练模型，可以应用到视觉—文本生成和理解任务当中；此外，最近也出现了一些从其他方面进行探究的开创性工作：Feng等人[86]使用了视觉概念(concept)作为连接图像和文本的桥梁，将无监督学习应用到图片标注任务当中；Sammani等人[84]提出了一种自适应“编辑”网络，可以对生成的描述进行迭代地润色。

MS COCO数据集目前已经成为研究人员在图像标注任务上进行性能评测的首选。由于官方测试集的真实标签没有公布，大多数研究人员常使用Karpathy等人[5]的数据集分割方式，进行离线验证和测试，不同的图像标注模型在MS COCO Karpathy测试集[5]上的性能如表2所示。为了和最先进的模型进行在线性能比较，部分工作会进一步使用MS COCO官方测试集进行测试，并将结果上传到评估服务器进行评测，不同的图像标注模型在MS COCO评估服务器上的性能如表3所示。

表 2　不同的图像标注模型在 MS COCO 测试集（Karpathy[5]）上的性能

方法	MS COCO							
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Google NIC[4]†	66.6	46.1	32.9	24.6	—	—	—	—
BRNN[5]	62.5	45.0	32.1	23.0	19.5	—	66.0	—
m-RNN[78]	67.0	49.0	35.0	25.0	—	—	—	—
Hard-Att [34]	71.8	50.4	35.7	25.0	23.0	—	—	—
Adaptive[80]	74.2	58.0	43.9	33.2	26.6	—	108.5	20.4
SCST：Att2all[46]	—	—	—	34.2	26.7	55.7	114.0	—
StackCap[81]	78.6	62.5	47.9	36.1	27.4	56.9	120.4	20.9
Up-Down[35]	79.8	—	—	36.3	27.7	56.9	120.1	21.4
NBT[24]	75.5	—	—	34.7	27.1	—	107.2	20.1
GCN-LSTM[32]	80.9	—	—	38.3	28.6	58.5	128.7	22.1
UIC[86]	41.0	22.5	11.2	5.6	12.4	28.7	28.6	8.1
SGAE[83]	80.8	—	—	38.4	28.4	58.6	127.8	22.1
AoANet[82]	80.2	—	—	38.9	29.2	—	129.8	22.4
Unified VL[85]	—	—	—	39.5	29.3	—	129.3	23.2
ETN[84]	80.6	65.3	51.1	39.2	—	58.9	128.9	22.6

注：† 代表使用了不同的测试集分割。

表 3　不同的图像描述生成模型在 MS COCO 评估服务器上的性能

方法	MS COCO															
	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr		SPICE	
	C5	C40	C5	C40	C5	C40	C5	C40	C5	C40	C5	C40	C5	C40	C5	C40
Google NIC[4]	71.3	89.5	54.2	80.2	40.7	69.4	30.9	58.7	25.4	34.6	53.0	68.2	94.3	94.6	18.2	63.6
MS Captivator[23]	71.5	90.7	54.3	81.9	40.7	71.0	30.8	60.1	24.8	33.9	52.6	68.0	93.1	93.7	18.0	60.9
m-RNN[78]	71.6	89.0	54.5	79.8	40.4	68.7	29.9	57.5	24.2	32.5	52.1	66.6	91.7	93.5	17.4	60.0
LRCN[79]	71.8	89.5	54.8	80.4	40.9	69.5	30.6	58.5	24.7	33.5	52.8	67.8	92.1	93.4	17.7	59.9
Hard-Att[34]	70.5	88.1	52.8	77.9	38.3	65.8	27.7	53.7	24.1	32.2	51.6	65.4	86.5	89.3	17.2	59.8
ATT-FCN[31]	73.1	90.0	56.5	81.5	42.4	70.9	31.6	59.9	25.0	33.5	53.5	68.2	94.3	95.8	18.2	63.1
Adaptive[80]	74.8	92.0	58.4	84.5	44.4	74.4	33.6	63.7	26.4	35.9	55.0	70.5	104.2	105.9	19.7	67.3
SCST：Att2all[46]	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7	20.7	68.9
StackCap[81]	77.8	93.2	61.6	86.1	46.8	76.0	34.9	64.6	27.0	35.6	56.2	70.6	114.8	118.3	—	—
Up-Down[35]	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5	21.5	71.5
GCN-LSTM[32]	—	—	65.5	89.3	50.8	80.3	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5	—	—
AoANet[82]	81.0	95.0	65.8	89.6	51.4	81.3	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6	—	—

5 结论

本文回顾了近年来研究领域在图像描述生成模型方面的研究进展。虽然相关研究在自动化评价指标方面有长足发展,但受限于真实场景的复杂性,对于图像的自动化描述离实际应用依然存在不小的差距。以下三个方面或许是未来的研究重点,包括基于跨模态预训练模型的图像描述生成框架研究、基于视觉的文本生成评价方法、面向应用的多样化文本生成框架研究。

参考文献

[1] He X, Deng L. Deeplearning for image-to-text generation: A technical overview[J]. IEEE Signal Processing Magazine, 2017, 34(6): 109-116.

[2] Hossain M Z, Sohel F, Shiratuddin M F, et al. Acomprehensive survey of deep learning for image captioning [J]. ACM Computing Surveys, 2019, 51(6): 1-36.

[3] Zhang C, Yang Z, He X, et al. Multimodal intelligence: Representation learning, information fusion, and applications[C]//Proceedings of IEEE Journal of Selected Topics in Signal Processing, 2020.

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

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

[6] Antol S, Agrawal A, Lu J, et al. Vqa: visual question answering[C]//Proceedings of the International Conference on Computer Vision,2015: 2425-2433.

[7] Ting-Hao Huang, Ferraro F, et al. Visual storytelling [C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016: 1233-1239.

[8] Das A, Kottur S, Gupta K, et al. Visual dialog[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2017: 326-335.

[9] Johnson J, Hariharan B, van der Maaten L, et al. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2017: 2901-2910.

[10] Zhu Y, Mottaghi R, Kolve E, et al. Target-driven visual navigation in indoor scenes using deep reinforcement learning [C]//Proceedings of the 2017 IEEE International Conference on Robotics and Automation,2017: 3357-3364.

[11] Reed S, Akata Z, Yan X, et al. Generative adversarial text to image synthesis[J]. arXiv preprint: 1605.05396, 2016.

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

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

[14] Li S, Kulkarni G, Berg T L, et al. Composing simple image descriptions using web-scale n-grams [C]// Proceedings of the 15th Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2011: 220-228.

[15] 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.

[16] Ordonez V, Kulkarni G, Berg T L. Im2text: Describing images using 1 million captioned photographs [C]// Proceedings of the Advances in Neural Information Processing Systems,2011: 1143-1151.

[17] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Proceedings of the Advances in Neural Information Processing Systems, 2014: 3104-3112.

[18] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks [C]// Proceedings of the Advances in Neural Information Processing Systems,2012: 1097-1105.

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

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

[21] ChungJ, Gulcehre C, Cho K, et al. Gated feedback recurrent neural networks[C]//Proceedings of the International Conference on Machine Learning, 2015: 2067-2075.

[22] Herdade S, Kappeler A, Boakye K, et al. Imagecaptioning: Transforming objects into words[C]//Proceedings of the Advances in Neural Information Pro-

cessing Systems，2019：11135-11145.

[23] 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.

[24] Lu J，Yang J，Batra D，et al. Neural baby talk[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2018：7219-7228.

[25] Dai B，Fidler S，Lin D. A neural compositional paradigm for image captioning[C]//Proceedings of the Advances in Neural Information Processing Systems，2018：658-668.

[26] Lecun Y，Boser B，Denker J，et al. Backpropagation-applied to handwritten zip code recognition[J]. Neural Computation，1989，1(4)：541-551.

[27] Girshick R，Donahue J，Darrell T，et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2014：580-587.

[28] Ren S，He K，Girshick R，et al. Faster r-cnn：Towards real-time object detection with region proposal networks[C]//Proceedings of the Advances in Neural Information Processing Systems，2015：91-99.

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

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

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

[32] Yao T，Pan Y，Li Y，et al. Exploring visual relationship for image captioning[C]//Proceedings of the European Conference on Computer Vision（ECCV），2018：684-699.

[33] Bahdanau D，Cho K，Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint：1409.0473，2014.

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

[35] Anderson P，He X，Buehler C，et al. Bottom-up and top-down attention for image captioning and visual question answering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2018：6077-6086.

[36] Devlin J，Chang M W，Lee K，et al. Bert：Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint：1810.04805，2018.

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

[38] Li L H，Yatskar M，Yin D，et al. Visualbert：A simple and performant baseline for vision and language[J]. arXiv preprint：1908.03557，2019.

[39] Li G，Duan N，Fang Y，et al. Unicoder-vl：A universal encoder for vision and language by cross-modal pre-training[J]. arXiv preprint：1908.06066，2019.

[40] Su W，Zhu X，Cao Y，et al. Vl-bert：Pre-training of generic visual-linguistic representations[J]. arXiv preprint：1908.08530，2019.

[41] Lu J，Batra D，Parikh D，et al. Vilbert：Ptretraining task-agnostic visiolinguistic representations for vision-and-language tasks[C]//Proceedings of the Advances in Neural Information Processing Systems，2019：13-23.

[42] Tan H，Bansal M. Lxmert：learning cross-modality encoder representations from transformers[J]. arXiv preprint：1908.07490，2019.

[43] Chen Y C，Li L，Yu L，et al. Uniter：Learning universal image-text representations[J]. arXiv preprint：1909.11740，2019.

[44] Ranzato M A，Chopra S，Auli M，et al. Sequence level training with recurrent neural networks[J]. arXiv preprint：1511.06732，2015.

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

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

[47] Vijayakumar A K，Cogswell M，Selvaraju R R，et al. Diverse beam search：Decoding diverse solutions from neural sequence models[J]. arXiv preprint：1610.02424，2016.

[48] Wang Z，Wu F，Lu W，et al. Diverse image captioning via grouptalk[C]//Proceedings of the IJCAI，2016：2957-2964.

[49] Shetty R，Rohrbach M，Anne Hendricks L，et al. Speaking the same language：Matching machine to human captions by adversarial training[C]//Proceedings of the IEEE International Conference on Com-

puter Vision,2017: 4135-4144.

[50] Fan Z, Wei Z, Li P, et al. A question type driven framework to diversify visual question generation [C]//Proceedings of the IJCAI,2018: 4048-4054.

[51] Wang L, Schwing A, Lazebnik S. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space[C]//Proceedings of the Advances in Neural Information Processing Systems, 2017: 5756-5766.

[52] Chen F, Ji R, Ji J, et al. Variational structured semantic inference for diverse image captioning[C]// Proceedings of the Advances in Neural Information Processing Systems,2019: 1929-1939.

[53] Dai B, Fidler S, Urtasun R, et al. Towards diverse and natural image descriptions via a conditional gan [C]// Proceedings of the IEEE International Conference on Computer Vision,2017: 2970-2979.

[54] Li D, Huang Q, He X, et al. Generating diverse and accurate visual captions by comparative adversarial learning[J]. arXiv preprint : 1804.00861, 2018.

[55] Li J, Galley M, Brockett C, et al. A diversity-promoting objective function for neural conversation models [J]. arXiv preprint : 1510.03055, 2015.

[56] 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.

[57] Plummer B A, Wang L, Cervantes C M, et al. Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models[C]// Proceedings of the IEEE International Conference on Computer Vision,2015: 2641-2649.

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

[59] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition,2009: 248-255.

[60] Krishna R, Zhu Y,Groth O, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations[J]. International Journal of Computer Vision, 2017, 123(1): 32-73.

[61] Grubinger M, Clough P, Müller H, et al. The iapr tc-12 benchmark: A new evaluation resource for visual information systems[C]//Proceedings of the International Workshop onto Image,2006: 13-22.

[62] Escalante H J, Hernández C A, Gonzalez J A, et al. The segmented and annotatediapr tc-12 benchmark [J]. Computer Vision and Image Understanding, 2010, 114(4): 419-428.

[63] Kazemzadeh S, Ordonez V, Matten M, et al. Referitgame: Referring to objects in photographs of natural scenes[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),2014: 787-798.

[64] Tran K, He X, Zhang L, et al. Rich image captioning in the wild[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops,2016: 49-56.

[65] Chunseong Park C, Kim B, Kim G. Attend to you: Personalized image captioning with context sequence memory networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2017: 895-903.

[66] Wang Y, Lin Z, Shen X, et al. Skeleton key: Image captioning by skeleton-attribute decomposition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2017: 7272-7281.

[67] Bychkovsky V, Paris S, Chan E, et al. Learning photographic global tonal adjustment with a database of input/output image pairs [C]//Proceedings of the CVPR. IEEE, 2011: 97-104.

[68] Gan C, Gan Z, He X, et al. Stylenet: Generating attractive visual captions with styles[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2017: 3137-3146.

[69] Wang X, Chen W, Wang Y F,et al. No metrics are perfect: Adversarial reward learning for visual storytelling[J]. arXiv preprint : 1804.09160, 2018.

[70] Fan Z, Wei Z, Wang S, et al. Bridging by word: Image grounded vocabulary construction for visual captioning[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 6514-6524.

[71] Lin C Y, Hovy E. Automatic evaluation of summaries using n-gram co-occurrence statistics[C]//Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, 2003: 150-157.

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

[73] Vedantam R, Lawrence Zitnick C, Parikh D. Cider: consensus-based image description evaluation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2015: 4566-4575.

[74] 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.

[75] Anderson P, Fernando B, Johnson M, et al. Spice: semantic propositional image caption evaluation[C]//Proceedings of the European Conference on Computer Vision. Springer, Cham, 2016: 382-398.

[76] Cui Y, Yang G, Veit A, et al. Learning to evaluate image captioning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 5804-5812.

[77] Jiang M, Huang Q, Zhang L, et al. Tiger: Text-to-image grounding for image caption evaluation[J]. arXiv preprint : 1909.02050, 2019.

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

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

[80] Lu J, Xiong C, Parikh D, et al. Knowing when to look: Adaptive attention via a visual sentinel for image captioning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 375-383.

[81] Gu J, Cai J, Wang G, et al. Stack-captioning: coarse-to-fine learning for image captioning[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018.

[82] Huang L, Wang W, Chen J, et al. Attention on attention for image captioning[C]//Proceedings of the IEEE International Conference on Computer Vision, 2019: 4634-4643.

[83] Yang X , Tang K , Zhang H , et al. Auto-encoding scene graphs for image captioning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 10685-10694.

[84] Sammani F, Melas-Kyriazi L. Show, edit and tell: a framework for editing image captions[J]. arXiv preprint : 2003.03107, 2020.

[85] Zhou L, Palangi H, Zhang L, et al. Unified vision-language pre-training for image captioning and vqa[J]. arXiv preprint : 1909.11059, 2019.

[86] Feng Y, Ma L, Liu W, et al. Unsupervised image captioning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 4125-4134.

魏忠钰(1987—)，博士，副教授，主要研究领域为跨视觉文本的相关研究、论辩挖掘、机器学习、强化学习。
E-mail: zywei@fudan.edu.cn

王瑞泽(1996—)，硕士研究生，主要研究领域为跨视觉文本的相关研究。
E-mail: rzwang18@fudan.edu.cn

范智昊(1996—)，硕士研究生，主要研究领域为跨视觉文本的相关研究。
E-mail: fanzh18@fudan.edu.cn