JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2021
04
VOL.26

ISSN1006-8961
CN11-3758/TB

细粒度图像分类 *P847*

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第26卷第4期（总第300期）
2021年4月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 吴一戎
编辑出版 《中国图象图形学报》编辑部
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@radi.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial Office of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@radi.ac.cn
Telephone 010-58887035
Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

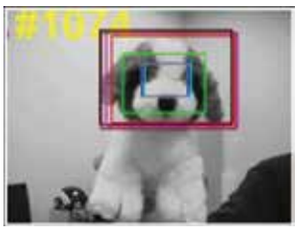
CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

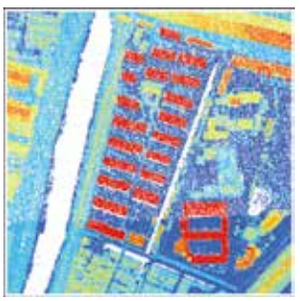
2021年4月 第26卷 第4期（总第300期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

目次

YOLOv3剪枝模型的多人姿态估计(第0837页)



融入时序和速度信息的自适应更新目标跟踪(第0883页)

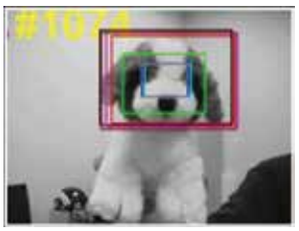


可变半径Alpha Shapes提取机载LiDAR点云建筑物轮廓(第0910页)

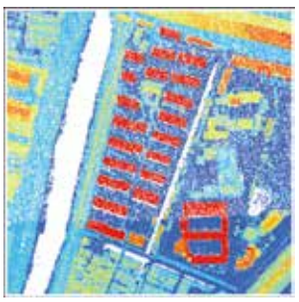
Volume 26, Number 4
Published April 16, 2021

CONTENTS
JOURNAL OF IMAGE AND GRAPHICS

Research on multiperson pose estimation combined with YOLOv3 pruning model(P0837)



Adaptive update object tracking algorithm incorporating timing and speed information(P0883)



Extraction of building contours from airborne LiDAR point cloud using variable radius Alpha Shapes method(P0910)

中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2021)04-0727-24

论文引用格式: Tan Y L, Tang P J, Zhang L and Luo Y P. 2021. From image to language: image captioning and description. Journal of Image and Graphics,26(04):0727-0750(谭云兰,汤鹏杰,张丽,罗玉盘. 2021. 从图像到语言:图像标题生成与描述. 中国图象图形学报,26(04):0727-0750)[DOI:10.11834/jig.200177]

从图像到语言:图像标题生成与描述

谭云兰[1,2,3],汤鹏杰[1,2*],张丽[1],罗玉盘[3]

1. 井冈山大学电子与信息工程学院, 吉安　343009; 2. 江西省农作物生长物联网技术工程实验室, 吉安　343009;
3. 井冈山大学网络信息中心, 吉安　343009

摘　要:　图像标题生成与描述的任务是通过计算机将图像自动翻译成自然语言的形式重新表达出来,该研究在人类视觉辅助、智能人机环境开发等领域具有广阔的应用前景,同时也为图像检索、高层视觉语义推理和个性化描述等任务的研究提供支撑。图像数据具有高度非线性和繁杂性,而人类自然语言较为抽象且逻辑严谨,因此让计算机自动地对图像内容进行抽象和总结,具有很大的挑战性。本文对图像简单标题生成与描述任务进行了阐述,分析了基于手工特征的图像简单描述生成方法,并对包括基于全局视觉特征、视觉特征选择与优化以及面向优化策略等基于深度特征的图像简单描述生成方法进行了梳理与总结。针对图像的精细化描述任务,分析了当前主要的图像"密集描述"与结构化描述模型与方法。此外,本文还分析了融合情感信息与个性化表达的图像描述方法。在分析与总结的过程中,指出了当前各类图像标题生成与描述方法存在的不足,提出了下一步可能的研究趋势与解决思路。对该领域常用的 MS COCO2014(Microsoft common objects in context)、Flickr30K 等数据集进行了详细介绍,对图像简单描述、图像密集描述与段落描述和图像情感描述等代表性模型在数据集上的性能进行了对比分析。由于视觉数据的复杂性与自然语言的抽象性,尤其是融合情感与个性化表达的图像描述任务,在相关特征提取与表征、语义词汇的选择与嵌入、数据集构建及描述评价等方面尚存在大量问题亟待解决。

关键词:图像标题生成;深度特征;视觉描述;语段生成;图像情感;逻辑语义

From image to language: image captioning and description

Tan Yunlan[1,2,3], Tang Pengjie[1,2*], Zhang Li[1], Luo Yupan[3]

1. School of Electronics and Information Engineering, Jinggangshan University, Ji'an　343009, China;
2. Jiangxi Engineering Laboratory of IoT Technologies for Crop Growth, Ji'an　343009, China;
3. Network Information Center, Jinggangshan University, Ji'an　343009, China

Abstract: Image captioning and description belong to high-level visual understanding. They translate an image into natural language with decent words, appropriate sentence patterns, and correct grammars. The task is interesting and has wide application prospects on early education, visually impaired aid, automatic explanation, auto-reminding, development of intelligent interactive environment, and even designing of intelligent robots. They also provide support for studying image

收稿日期:2020-06-02;**修回日期**:2020-07-27;**预印本日期**:2020-08-03

*通信作者:汤鹏杰　tangpengjie@jgsu.edu.cn

基金项目:国家自然科学基金项目(62062041);井冈山大学博士科研启动项目(JZB1923, JZB1807);江西省艺术科学规划项目(YG2017283);江西省高校人文社科基地招标项目(JD17082);江西省高校信息化学会一般项目(GJJ191662,GJJ191663)

Supported by: National Natural Science Foundation of China (62062041); Art Planning Project of Jiangxi Province (YG2017283); Bidding Project for the Foundation of College's Key Research on Humanities and Social Science of Jiangxi Province (JD17082); General Project of Jiangxi University Information Chemistry Association(GJJ191662,GJJ191663)

retrieval, object detection, visual semantic reasoning, and personalized description. At present, the task has attracted the attention of several researchers, and a large number of effective models have been proposed and developed. However, the task is difficult and challenging because the model has to bridge the visual information and natural language and close the semantic gap between the data with different modalities. In this work, the development timeline, popular frameworks and models, frequently used datasets, and corresponding performance of image captioning and description are surveyed comprehensively. Additionally, the remaining questions and limitations of current works are investigated and analyzed in depth. Overall, there are four parts for image captioning and description illustration in this study: 1) the image simple captioning and description (one sentence is generated for an image generally), including handcraft feature-based methods and deep feature-based approaches; 2) image dense captioning (multiple but relatively independent sentences are generated in general) and refined paragraph description (paragraph with a certain structure and logic is generated generally); 3) image personalized and sentimental captioning and description (sentence with personalized style and sentimental words is generated in general); and 4) corresponding evaluation datasets, metrics, and performances of the popular models. For the first part, the research history of image captioning and description is first introduced, including template-based framework and visual semantic retrieval-based framework based on handcraft visual feature. The classical and significant works such as semantic space sharing model and visual semantic component reorganization model are described in detail. Then, the current popular works based on deep learning techniques are sorted out carefully and elaborated in great detail. According to the usage of visual information, the models for image captioning and description based on deep feature can be mainly classified into three categories: 1) global visual feature-based model, 2) visual feature selection and optimization-based model, and 3) optimization strategy-oriented model. For each kind of model, the current popular works including the proposed models, superiority, and possible problems are analyzed and discussed. The models based on selected or optimized visual features such as visual attention region, attributes, and concepts as prior knowledge are usually more intuitive and show better performance, especially when advanced optimization strategies such as reinforcement learning are employed and the quality of generated sentences frequently possesses more accurate words and richer semantics, although a few methods based on global visual feature perform as good as them. Besides the models for image simple captioning and description, popular works on dense captioning and refinement description for images are presented and sorted out in the second part. The models for dense captioning generate more sentences for images and offer more detailed description. However, the semantic relevance among different visual objects, scenes, and actions is usually ignored and not embedded into the sentences although a few approaches take advantage of the possible visual relation to predict more accurate words. With regard to refined paragraph description for images, the hierarchical architecture with multiple recurrent neural network layers is the most employed basic framework, where hierarchical attention mechanism, visual attributes, and reinforcement learning strategy are also introduced into the related models to further improve the performance. However, the semantic relevance and logic among different visual objects remain to be further explored and represented, and the coherence and logicality of the generated description paragraph for images need to be further polished and refined. Additionally, in consideration of human habit of describing an image, personal experience is usually embedded into the description, and then the generated sentences often contain personalized and sentimental information. Therefore, a few significant works for personalized image captioning and sentimental description are also introduced and discussed in this paper. In particular, the discovery, representation, and embedding of personalized information and sentiment in the models are surveyed and analyzed in depth. Moreover, the limitations and problems about the task including the granularity and intensity of sentiments, the evaluation metrics of personalized and sentimental description, are worthy of further research and exploration. In addition to classical frameworks and popular models, the related public evaluation datasets and metrics are also summarized and presented. First of all, the image simple captioning and description datasets, including Microsoft common objects in context (MS COCO2014), Flickr30K, and Flickr8K, and the performances of a few popular models on these datasets are briefly introduced. Afterward, the datasets, including Visual Genome and VG-P (Paragraph) for image dense captioning and paragraph description, and the performances of certain current works on the datasets are described and provided. Next, the datasets for the task of image description with personalized and sentimental expression, including SentiCap and FlickrStyle10K, are briefly introduced. Moreover, the performances of the main models are reported and discussed. Additionally, the frequently used

evaluation methods, including traditional metrics and special targeted metrics, are described and compared. In conclusion, breakthrough has been made on image captioning and description in recent years, and the quality of generated sentences has been greatly improved. However, more efforts are still needed to generate more coherent and accurate sentences with richer semantics for images. The possible trends and solutions about image captioning and description are reconsidered and put forward in this study. To further promote the task to practical applications, the semantic gap between visual dataset and natural language should be narrowed by generating a structured paragraph with sentiment and logical semantics for images. However, several problems, including visual feature refinement and usage, sentiment and logic mining and embedding, corresponding training dataset collecting and metric designing for personalization, and sentiment and paragraph description evaluation, remain to be addressed.

Key words: image captioning; the depth feature; visual description; paragraph generation; image sentiment; logical semantic

0 引 言

对给定一幅图像的内容进行抽象和总结，并使用自然语言的形式对其进行重新表达与描述属于视觉高层理解的范畴。该任务对于具有一定生活经验和相关知识结构的人类而言比较简单，能够轻易地完成，但对于计算机来说，由于图像数据的高度非线性和繁杂性，以及自然语言高度的抽象性以及严谨的逻辑性，完成这一任务具有很大的挑战性。但由于该任务具有极大的应用前景和学术研究价值，如婴幼儿教育、视觉障碍者日常辅助、自动解说、智能人机交互环境，以及机器人设计等主题，已吸引了越来越多的研究者投入这一领域，提出了一系列效果显著的算法和模型框架，为图像生成的标题或描述句子的质量也越来越高。

对图像标题生成与描述领域的研究已开展了 20 多年，研究水平随着计算机视觉的技术发展而不断进步。从基于模板（Farhadi 等，2010）和基于语义迁移（Kuznetsova 等，2013）的方法，到主要基于神经机器翻译的模型，其视觉技术的发展水平直接决定了性能的提升程度和生成句子的质量。在深度学习技术出现之前，主要采用手工特征对图像进行特征编码，然后利用图像分割、目标检测或视觉检索等技术识别其中的视觉语义对象或相关的视觉语义描述，最后使用模板填充或视觉语义迁移等方法将视觉语义对象填充到相应的位置或将相似的语义片段组合成新的标题与描述语句。手工特征虽然设计严谨，可解释性强，但由于其多为分步设计，特征非线性变换次数较少，抽象性不高，表达能力有限，导致生成的标题和描述句子在准确性、灵活性和语义丰富程度等方面都难以达到人们的要求。

自 2012 年以来，得益于深度卷积神经网络（deep convolutional neural networks, DCNN）（Krizhevsky 等，2012）的设计及在多种视觉任务上的突破性表现，学者开始将其应用在视觉描述任务上（Mao 等，2015；Vinyals 等，2015），使用 DCNN 实现对图像的特征编码，将其转换为“临时中间语言”，然后使用循环神经网络（recurrent neural network, RNN）对其进行解码，逐个生成词汇，最终组成具有一定结构的句子。这种方法使用的视觉特征抽象程度高，泛化能力强，具有手工特征难以比拟的优势，同时在生成句子时较为灵活，不受句子长度及固定句式的影响，其过程较为贴近人类的表达习惯。目前，DCNN + RNN 已成为视觉描述任务的主流框架，其后续的研究工作多是在此基础上融合更多的视觉先验知识（如结合注意力、视觉属性/概念等），进一步提升词汇预测的准确性，改善句子的语义丰富程度。除此之外，也有研究者从模型优化（包括视觉特征提取、语言解码和联合优化等）的角度设计功能更为强大的记忆单元或模型，提高生成句子的质量。从任务的角度而言，为图像生成单条简单描述是当前研究的主流，但对于很多复杂图像，单条语句难以对其中的细节进行详细的表达。针对该问题，研究人员提出了密集描述任务，为各视觉区域生成句子，最终实现图像内容的精细化表达（Kulkarni 等，2013）。为了进一步使得描述更加生动，学者还参照人类日常的对话与交互习惯，提出嵌入情感与个性化表达的视觉描述任务（Mathews 等，2016），并取得了一定的进展，但鉴于情感与个性化信息的复杂性，其效果仍然有待于进一步改善。

本文回顾并讨论了当前图像标题生成与描述的主流方法和模型。此外，还对一些如密集描述、情感描述等特殊的图像描述任务进行了总结；然后对各

模型的性能表现进行了对比分析;最后对本文内容进行总结,指出下一步可能的研究方向。

1 图像简单标题生成与描述

对图像中的视觉内容进行归纳和总结,并使用合适的词汇与合理的语法结构将其重新组织并表达出来,是图像标题生成与描述的主要研究内容。如图 1 所示,首先对图像中的视觉内容进行解析,将其转换成视觉语义编码,然后根据编码内容进行解码,将其映射到语言空间中,生成相关词汇,并组合成用词准确、结构合理的自然语言。本节围绕该基本框架,从视觉特征提取、视觉语义选择和模型设计与优化等方面,介绍当前流行的方法和模型架构。

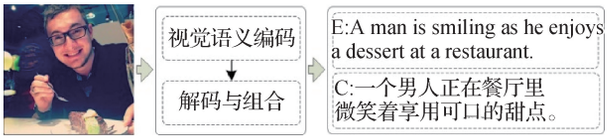


图 1 图像标题生成与描述框架

Fig. 1 General pipeline of image captioning and description

1.1 基于手工特征的图像简单描述生成方法

手工特征是相对于当前流行的深度特征而言的,一般根据视觉内容的特点,如颜色、纹理和边缘信息等,统计其像素出现的次数、梯度方向等,并通过投影和变换形成固定维度的特征向量。常用的图像手工特征提取方法主要包括 LBP(local binary pattern)(Ojala 等,1996)、SIFT(scale invariant feature transform)(Lowe,2004)、HOG(histogram of oriented gradient)(Dalal 和 Triggs,2005)等。手工特征在设计时一般具有较为明确的动机,每个步骤都具有清晰的设计思路,因此其可解释性强,能够对其进行严密的推理和控制;且由于其特征提取多采用统计学习的方式,使用的参数较少,算法可直接集成在硬件之上,运算速度较快。但也由于其设计时多关注于某个方面的视觉特性,且非线性变换次数较少,使得特征表达能力和抽象能力有限,因而使用手工特征的模型性能受到极大限制。

早期研究中一般利用手工特征来完成图像标题的生成与描述任务。Farhadi 等人(2010)首先提出了“视觉—语言”共享语义空间的方法,通过检测图像中的视觉对象,并将其映射到预先设计的语义空间元组上,建立该图像与元组所关联的描述句子之间的对应关系;通过这种关系,不仅能够为图像生成句子,也能够根据句子内容找到相应的图像。这种方法在视觉语义编码阶段采用了 HOG 特征,其中间语义空间的设计采用了固定的模板(〈object(对象),action(动作),scene(场景)〉),其视觉对象检测的准确性难以保证,造成语义空间映射的不确定性,从而造成通过中间语义空间元组所找到的对应句子也可能与图像实际内容存在较大偏差;且由于模板和句子空间都是事先设定,其生成的句子不够灵活,尤其是在对视觉语义对象的细节描述上,其偏差可能更大。

Yang 等人(2011)也借鉴了其共享语义空间的做法,但其语义空间设计更为复杂,其共享模板的结构也更加多样化,增强了句子的灵活性。Mitchell 等人(2012)则将检测到的视觉语义对象进行筛选,并通过构造语法树生成更为准确和灵活的句子。这种方法中所使用的语法树虽然也可归为模板,但由于其句子结构可根据语义内容进行灵活变化,已经非常贴近人们的表达习惯。这些工作虽然由于手工特征的局限性造成生成的句子质量还有所欠缺,但其思想动机、模型设计与开发的合理性、可解释性及有效性都为后续工作的进一步推进提供了良好的借鉴。

使用手工特征的方法除基于模板的框架外,研究人员还提出了一种基于语义迁移与重组的模型框架。Yao 等人(2010)首先使用图像分割与 SIFT 特征等将图像解析为视觉语义组件,然后将其转换为 Web 本体语言,实现其与通用知识库的衔接,并通过检索技术与语义解析图,将视觉概念转换成自然语言。这种方法首先依赖于特征的表达能力,用以支撑将图像解析成准确的视觉语义概念;其次,需要构建较为完善的 Web 语义库,使得能够查询到置信度较高的语义本体,并组合成新的描述语句。由于取消了模板的限制,其生成的句子在灵活性和语义性方面都有较大改善。但当 Web 语义库不完整时,其生成的句子与图像实际内容之间也会存在一定的偏差,影响了句子的整体质量。

Kuznetsova 等人(2012)则简化了这一过程,重点关注于句子的重组与生成。他们首先根据视觉的相似性,在检索库中搜索近似的视觉内容及其相关的词汇或词组,将检索到的语义片段组合成新的句

子。在后续的工作中，Kuznetsova 等人（2013）还采用了启发式的集束搜索算法（beam search）扩展词汇或词组的搜索空间。此外，Kuznetsova 等人（2014）提出了另一种基于随机树合成的图像描述生成方法，首先检测出待描述图像中的语义片段，然后从检索库中寻找携带类似语义的图像及其描述，并将其视觉片段和对应描述单独抽取出来，然后将其组合成新的描述句子。这种方法使用的视觉及语言颗粒更小，组成的句子也更加灵活多变。Mason 和 Charniak（2014）则根据待描述图像中视觉内容所对应的标签词频，将描述生成问题转化为文本摘要提取问题，使用更成熟的自然语言处理技术实现生成质量更高的标题或描述的目标。基于语义迁移与重组的方法整体上更符合人们的表达习惯，但同样囿于手工特征有限的表达能力，以及检索库的完整度，在句子的准确性和语义性方面同样不能满足人们的需求。

1.2　基于深度特征的图像简单描述生成方法

深度学习的概念由 Hinton 和 Salakhutdinov（2006）首次提出，为了使得图像特征更加抽象，使用逐层优化和全局微调的方式构建了一个深度神经置信网络，并应用在数字和人脸识别任务上。其实在此之前，LeCun 等人（1998）已经设计了一种卷积神经网络（convolutional neural networks，CNN）模型，其权重层数超过其他神经网络模型，但限于当时的软硬件水平，其进一步的开发与应用受到了限制。直到 2012 年，Krizhevsky 等人（2012）将深度学习的思想与 CNN 融合起来，开发了更深的 AlexNet 模型，由于视觉数据经过了更多层次的线性与非线性变换，其特征抽象能力更高，泛化能力更强，使得该模型在 ImageNet 分类数据集（Russakovsky 等，2015）上获得了极大的性能突破。

AlexNet 的成功促进了深度网络的研究热潮，先后开发了如 VGG-Net（Visual Geometry Group）（Simonyan 和 Zisserman，2014）、GoogLeNet（Szegedy 等，2015）、ResNet（residual net）（He 等，2016）等深度模型，从模型结构和优化策略上都进行了大量探索，同时也获得了性能上的不断提升；研究人员还将深度特征应用于其他各种视觉任务，包括目标检测（Girshick 等，2014；Girshick，2015；Ren 等，2017a）、动作识别（Wang 等，2015，2018b；Wu 等，2019a）和人脸识别（Sun 等，2013；Han 等，2018；Wu 等，2019b），以及其他如视觉关系检测与推理（Dai 等，2017b；Yang 等，2018）、视觉描述等高层理解任务。

在图像描述领域，其一般框架如图 2 所示，其中 w_i 是指第 i 个单词，t_i 是指时间序列。首先使用 DCNN 提取图像特征，对图像进行特征编码，然后使用序列模型对特征进行解码，逐个生成单词并将其组合为具有一定语法结构的描述句子。该过程中，提取图像特征时，可以直接采用预训练完毕的深度模型，也可以在此基础上，对其中的参数进行微调更新；在解码时，可使用 RNN 或 CRF（condition random field）建立语言模型，在生成词汇的同时，也隐式地将语法结构嵌入其中。这种方法的流程较基于模板和语义迁移的方法更为简洁，省去了大量手工调参的步骤，更为接近人类处理类似问题的思路。但需注意到的是深度模型一般包含大规模的参数，其优化时间更长，所需的计算资源更多，且由于深度模型的“黑盒”特性，也导致其可解释性不强，模型设计的动机及合理性能都受到一定质疑。但不可否认的是，使用深度特征的图像描述模型，其性能已远超过使用手工特征的模型，生成的句子在准确性、连贯性和语义丰富程度等方面都得到巨大改善，缩小了视觉与自然语言之间的语义鸿沟。

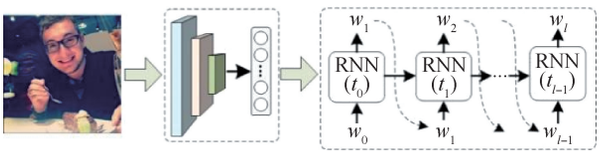


图 2　基于深度特征的图像标题描述一般框架

Fig. 2　General deep feature based framework of image captioning and description

根据对视觉信息的处理方式不同，可以将基于深度特征的图像描述模型分为 3 类：1）基于全局视觉特征的描述框架；2）基于视觉特征选择与优化的描述框架；3）面向优化策略的描述框架。以下将对其进行详细论述。

1.2.1　基于全局视觉特征的描述模型

基于全局视觉特征的描述模型是将图像特征提取出来之后直接送入语言模型中，语言模型根据记忆对不同的特征进行解码，生成句子，其模型框架如图 3 所示（其中 *vf* 表示图像的全局视觉特征）。

Mao 等人（2015）首先提出应用多模特征的方式将图像整体特征与 RNN 网络输出的语言特征结

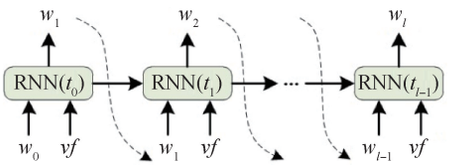


图 3　基于多模特征的图像标题与描述生成

Fig. 3　Multi-modal feature based architecture for image captioning and description

合在一起,送入多模混合单元,对视觉与语言特征进行记忆训练,而在生成句子时,当前时间步 RNN 单元则根据视觉特征和已生成的前续词汇决定当前输出。他们的工作极大地促进了人们对图像描述研究的兴趣。Donahue 等人(2017)研究了多模特征在不同的语言模型中的性能表现,指出使用单层 RNN 的多模特征不利于语言特征的建模,语言特征不够抽象,且容易引入更多的语言噪声,影响词汇预测的准确性;为此,设计了因子分解结构的语言模型,避免了噪声干扰。受此启发,汤鹏杰等人(2018)认为当前的语言模型深度不够,限制了语言特征的表达能力。为此,结合深度学习中的逐层优化和深度监督方法,构建了更深的语言模型,并借鉴深度融合思想,将不同深度 RNN 网络的概率输出进行融合。此外,汤鹏杰等人(2017)还认为由于当前模型多使用 ImageNet 对特征提取器进行预训练,但该数据集主要是面向物体分类,可能会导致部分描述场景的词汇预测不准确。Tang 等人(2017)提出了一种融合场景与物体先验知识的图像描述模型,使用物体识别数据集 ImageNet(Russakovsky 等,2015)和场景识别数据集 Place205(Zhou 等,2014)分别对两个 CNN 模型进行预训练,然后将其应用在图像描述任务上,提升词汇预测的准确性。

在上述几个模型中,RNN 网络的每个时间步上都会输入图像的全局特征,这种方法能够防止 RNN 网络中可能的视觉信息丢失,但 Vinyals 等人(2015)与 Karpathy 和 Li(2015)则认为在每个时间步上都输入视觉信息会引入额外的视觉噪声,干扰词汇预测的准确性。因此,Vinyals 等人(2015)提出将图像特征经过降维变换后直接输入 RNN 网络的第一个时间步中,后续的时间步将不再接收视觉信息,其生成句子时只依赖于先前时间步上的词汇特征与隐层输出。但这种方式只将视觉特征输入语言模型一次,虽然克服了可能的视觉干扰,但在时间步较长时,可能会产生影响整个模型的“长期依赖”问题,进而破坏句子的整体语义性。

1.2.2　基于视觉特征选择与优化的描述模型

视觉特征的表达能力与使用方法对描述句子质量具有巨大的影响,一般而言,直接使用 CNN 模型提取的是全局特征,是多层多个大小不同的卷积核进行多次滤波并进行融合而得到的,其对于整幅图像的表达能力已得到实验证明。但对于图像描述而言,无论是在语言模型的每个时间步上还是在第一个时间步上输入全局特征,都可能会引起新的问题,其根本原因在于语言模型中的视觉语义与语言词汇没有进行有效而合理的对应与校准。为解决这一问题,受机器翻译中“注意力机制”与视觉显著性的启发,将注意力机制与视觉概念/属性引入到图像标题生成与描述任务中,在不同的时间步上,赋予不同的视觉区域(特征)以不同的权值,或者结合不同的视觉概念/属性,以此缓解视觉噪声的干扰。其一般模型框架如图 4 所示。

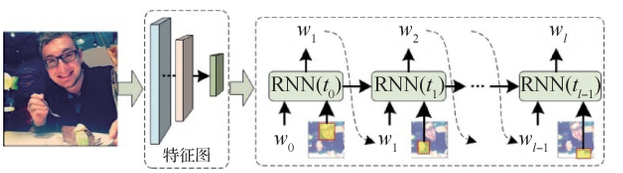


图 4　基于注意力机制的图像描述框架

Fig. 4　The attention mechanism based framework for image captioning and description

Xu 等人(2015)首先将注意力机制应用在图像描述任务上,将经过 CNN 变换后的最后一层特征图(非特征向量)的不同位置作为注意力关注对象,将多幅特征图的相同位置组合在一起作为该位置的特征片段,在不同的时间步上,根据记忆单元和先前生成词汇,决定该特征片段的权值。这种方法通过设置注意力单元,并将其概率输出与 RNN 单元的输入进行结合,能够自适应地学习视觉区域与语句中不同词汇的对应规律,有效地缓解了在不同时间步上无关视觉信息的干扰,提升了词汇预测的准确性,同时,该方法也具有较好的可解释性。后续的研究者以此为基础,提出了多种改进或扩展的注意力模型。Fu 等人(2017)则直接使用视觉区域作为注意力关注的对象,结合主题场景信息,为不同的视觉区域分配不同的权重。Pedersoli 等人(2017)也采用了同样的思路,使用视觉区域作为注意力单元关注的对象,

但他们构建了一个联合概率分布函数，将 LSTM（long short term memory）中的隐层状态、视觉区域与已生成的词汇联合起来实现注意力的精准定位。这种方法相较于使用特征图作为关注区域更加侧重于视觉语义对象，提供了另外一种较为有效的注意力机制。

Chen 等人（2017）认为最后一层特征图并不能完全反映图像中的真实内容，由于经过多层非线性变换，其大量细节信息被丢弃，不同视觉语义区域的相关性难以充分利用。为此，在不同层次的特征图上以及不同通道的特征图上也应用了注意力机制，充分挖掘有效视觉信息。但这种方法对计算资源要求更高，且由于参数增加，模型容易陷入过拟合状态。Lu 等人（2017）从视觉语义概念与语言词汇的关联性出发，认为并不是每个时间步上都应关注于具体的视觉区域，对于部分虚词，难以确定其对应的视觉信息，因此，在注意力单元上设置了一个哨兵单元，自适应地学习需要重点关注的时间步。Jiang 等人（2018）从视觉特征互补的角度出发，使用多个 CNN 模型提取图像的视觉特征，然后将其送入多个 RNN 网络，结合多注意力机制，在不同的时间步上关注更为丰富的视觉信息。Mun 等人（2017）认为在视觉注意力定位过程中，可以引入与其相似的相关描述作为辅助，提高模型对于视觉区域关注的准确性，为此，提出一种基于文本引导注意力机制的图像描述模型。首先根据图像内容使用相似度与标题共识分值，从训练集中检索出相关的描述句子，然后使用文本引导注意力单元计算词汇与视觉区域的相关度，并据此提取图像的上下文特征。

以上应用注意力的方法都是基于整幅图像的特征图，且赋予注意力权重的视觉区域大小也是固定的，这些都可能导致注意力所关注的视觉特征仍然携带有一定的干扰信息。为解决这一问题，部分研究人员尝试对注意力进行再次精炼，如 Anderson 等人（2018）提出一种自底向上与自顶向下两向结合的注意力模型，首先使用快速目标检测算法，框定视觉显著区域，然后使用自顶向下的注意力机制赋予该区域更多的权重。Fang 等人（2018）则采用一条单独的 RNN 网络对注意力关注区域进行精炼，然后将其送入语言模型中生成词汇，既考虑了空域上的视觉信息，也兼顾了句子时域上的时序关联关系。Gu 等人（2018）采用融合两层堆叠注意力机制的 LSTM 网络，对视觉信息进行过滤，实现由粗到细的图像描述。此外，Shi 等人（2019）以 R-CNN（region based CNN）为基础生成图像视觉语义区域，并结合全局视觉特征，采用一种级联的注意力单元，增强权重分配的差异性，进一步去除噪声干扰。Zhang 等人（2019b）认为传统的使用特征图作为注意力关注的母版，其各区域是固定的，不能反映出图像中真实的语义内容，因此其注意力权重分配可能并不公平。为此，他们使用全卷积网络（fully convolutional network，FCN）（Johnson 等，2016）生成细粒度网格的特征图，在此基础上提取特征图中的语义对象区域并表达为特征向量，以此引导注意力在特征图上的权重分配。

基于注意力机制的模型主要是为语言模型提供时序上的观测位置以及具体特征，实现视觉特征的选择与优化，在序列建模时有效地抑制了视觉无关信息的干扰。除此之外，研究人员还引入更多的先验知识，如视觉概念、视觉属性等，使用学习到的语义概念，引导或辅助描述语句的生成。Fang 等人（2015）提出利用视觉概念的方法完成这一任务，首先使用多示例学习技术（multi-instance learning，MIL）构建视觉概念检测器，在图像中检测可能的视觉语义概念，然后根据这些视觉概念词汇使用统计方法为图像搜索可用的相关词汇，并组成多条候选句子，最后根据图像与句子之间的距离，确定最终的描述语句。这种方法从视觉概念检测、生成句子到句子排序几个步骤之间是离散的，没有使用端到端的优化技术，从而也可能使得整个模型陷入局部最优状态，性能受到限制。但这种使用弱监督 MIL 技术检测视觉概念的方法启发了很多研究者，并提出了多种改进的模型。

You 等人（2016b）提出一种将视觉属性与注意力相结合的图像描述模型，同样使用弱监督的方法训练视觉属性检测器，但与 Fang 等人（2015）方法不同的是，其不使用目标检测的方式指定视觉区域，而是直接从参考句子中获取与对应图像相关的语义属性，避免了可能的视觉噪声干扰；然后使用注意力机制在每个时间步上关注特定的视觉属性，为不同的视觉属性分配不同的权值。这种方法绕过了为固定位置或区域分配权重的弊端，同时也引入了更多视觉和语言的先验知识，是一次在视觉特征优化与选择方面极为有益的尝试。Wu 等人（2016）为了使

用更高层次的视觉语义信息,也提出一种基于视觉属性的图像描述框架,但并未使用 MIL 方法来生成视觉属性,而是直接从参考语句中按照出现次数对属性进行选择;同时通过一种多尺度组合分组的技术(Wu 等,2018)检测视觉语义区域,并采用参数迁移的方式,使用优化完毕的模型对各视觉区域进行特征提取与融合,并结合图像的全局特征,将其送入语言模型进行解码。Yao 等人(2017)提出了一种更为简洁的模型,首先使用 MIL 方法检测出常用的视觉属性,然后将视觉属性与图像全局特征一起送入 RNN 网络中,辅助整条描述语句的生成。这种方法中语言模型的每个时间步上都补充了相应的视觉属性,本质上它并不属于对视觉特征的选择与优化,而是对图像特定区域特征进行了增强,在一定程度上抑制了视觉噪声的干扰,但并不能完全消除。Gan 等人(2017b)扩展了 You 等人(2016b)的工作,认为将视觉属性通过注意力的方式直接输入传统的 RNN 网络会导致其参数规模过于巨大,为此,他们采用矩阵分解的方式对 RNN 单元进行了改进。Yao 等人(2018)认为各视觉对象之间具有一定的语义关系,尤其在生成句子时,需要将这种视觉语义关系映射到句子结构中。为此,Yao 等人(2018)使用有向图为图像中的各目标区域构建语义关系图,通过图卷积网络(graph convolutional network, GCN)对视觉关系进行特征提取,并结合注意力机制生成描述语句。Yang 等人(2019)认为基于视觉信息的语言再表达与机器翻译不同,而是与视觉高层抽象的语义符号相关。他们使用目标检测技术得到图像中的各语义对象,并结合其属性、关系等特征,生成图像的场景图(scene graph),然后利用 GCN 提取其特征,结合在大规模文本库上预训练的字典,为图像生成描述语义信息更为丰富的语句。这些方法提出的视觉关系也可以看做视觉属性的一种,其采用 GCN 的方式对关系进行提取与建模,在视觉推理与可解释性分析等方面具有很好的启发作用,值得进一步研究与探讨。Zhang 等人(2019a)则认为直接从参考句子中获取的视觉概念或属性并不完整,还需要从集外选取更多的视觉先验,补充其可能由于正负样本不均衡导致的不准确或训练样本中缺失的概念。Yao 等人(2018)与 Zhang 等人(2019a)的方法在本质上突破了视觉特征选择与优化的范畴,在策略上进一步拓展了视觉属性/概念的内涵和外延。

1.2.3 面向优化策略的图像描述模型

与人类的学习过程一样,图像描述模型同样需要对其中的参数进行充分训练和优化,使其模型函数逼近于训练数据的整体分布,而在测试时则只需将数据代入模型,并给出相应结果。在该过程中,其不同的优化方式将直接影响模型的最终性能。在图像描述任务中,其模型的优化方式一般可以分为 3 种:1)局部优化策略;2)全局优化策略;3)局部与全局相结合的优化策略。

在局部优化策略中,一般为图像对应参考句子设置虚标签,并将其与模型预测结果进行对比,使用交叉熵的方式计算两者误差,以此对语言模型中的参数进行迭代更新。这种方式可以直接使用预训练之后的 DCNN 模型提取图像特征,然后直接送入语言模型对其进行训练,其使用较为灵活。但这种方法也割裂了图像特征提取模型(视觉模型)与语言模型之间的固有联系,使整个模型易陷入局部最优状态。为解决这一问题,研究将视觉模型与语言模型进行联合优化,误差回传到语言模型的尽头时,继续向后回传至 DCNN 模型中,对视觉模型中的参数进行微调更新。其典型工作包括 Donahue 等人(2017)提出的 LRCN(long-term recurrent convolutional network)模型、Vinyals 等人(2015)设计的 NIC(neural image caption)模型,以及其他多数使用注意力机制的模型等。

使用基于局部或全局联合优化策略的模型较为简洁,但大多是为了更有效地利用视觉语义信息,对语言模型改进的较少。因此,研究人员还通过改进语言模型的内部或外部结构,从数据的流向上对模型进行改进,通过优化记忆单元或模型架构充分利用模型训练时的局部与全局信息,改善生成句子的质量。如 Jia 等人(2015)对 LSTM 单元进行了扩展,设计了 gLSTM 结构,在每个时间步上既使用模型产生的局部信息对词汇进行预测,同时又使用全局信息指导整条句子的生成,兼顾了词汇的准确性与句子的语义性。为解决 RNN 网络难以并行计算且易发生梯度消失的问题,Aneja 等人(2018)提出一种基于卷积网络的图像描述框架,通过构建与注意力机制相结合的 CNN 网络,在每个时间步上根据已生成词汇与视觉注意区域的融合特征预测当前词汇输出。李勇等人(2019)也使用了相似的思路,提出一种基于文本 CNN 的建模方法。而 Rennie 等人

(2017) 则根据 CIDEr (consensus-based image description evaluation)评价指标,提出一种基于强化学习思想的图像描述框架,认为当前的模型在训练与测试时的评价指标是不统一的,在逐个生成词汇并产生误差的过程中,无法根据如 BLEU(bilingual evaluation understudy)(Papineni 等,2002),METEOR(metric for evaluation of translation with explicit ordering)(Banerjee 和 Lavie,2005)、CIDEr(Vedantam 等,2015)等评价指标对模型进行直接优化,造成优化方向上可能的偏差与事实上的局部最优。为此,以当前迭代模型生成的句子为基础,根据评价分值为模型计算奖励,引导模型的优化方向。除此之外,Liu 等人(2017a)还将 CIDEr 和 SPICE(semantic propositional image caption evaluation)(Anderson 等,2016)两种指标结合在一起作为优化目标,对模型进行更为充分的优化。这种方法解决了训练和测试不一致的问题,在模型优化策略上进行了非常有益的探索,实验结果也表明了该方法的有效性。Ren 等人(2017b)采用了强化学习的方式,将图像描述中生成词汇的过程作为决策过程,通过策略网络和价值网络之间的互动,对目标进行优化。但这种方法整体上采用以逼近真实句子为目标的优化策略,其训练和测试仍然是分离的。Dai 和 Lin(2017)、Dai 等人(2017a)认为当前很多模型生成的句子较为相似,尤其是针对一些内容较为相近的图像,其生成的句子重复率较高。为此,使用一种条件生成对抗网络(generative adversarial network, GAN)(Goodfellow 等,2014)以代替传统的交叉熵函数对模型进行优化。在判别过程中,该模型更为关注句子表达的自然性(逼真程度)及其与视觉内容的相关性,避免描述趋向重复。Yan 等人(2020)提出了一种更加复杂的强化学习模型,将 GAN 与强化学习策略相结合,并结合了双层注意力机制,将图像全局特征与局部特征结合起来,提高词汇预测的准确性。这种方法集合了注意力机制、GAN 和强化学习等现有技术,对于设计更高性能的模型具有很好的借鉴价值。

除了使用强化学习方法外,针对用于图像描述的数据集构建困难、质量难以保证等问题,还提出了使用小样本学习技术和无监督学习技术对模型进行优化。Dong 等人(2018)首先提出一种使用参数快速自适应学习的小样本图像描述模型。以元学习(Meta-learning)思想为基础,将图像和文本结合在一起作为学习目标,通过动态地学习文本中的少量先验知识,进而影响视觉模型中的参数更新,并实现视觉模型与语言模型的参数共享。该方法是解决图像描述任务中海量训练数据难以获得这一问题的有益尝试。但该模型在图像特征的利用方面还停留在初级阶段,性能也有待于进一步提升。Feng 等人(2019)提出了使用无监督的方法训练图像描述模型,其在思想上结合了强化学习、视觉属性和迁移学习等方法,通过计算生成句子与外部句子之间的对抗奖励、语义概念与图像重构奖励,以及句子重构损失,对生成器进行强化与更新,并采用对抗损失与图像重构损失对判别器进行更新优化。该方法针对性地解决缺乏大量可用的“图像—句子对”样本问题,不仅在方法上进行了创新,使图像描述任务更接近真实场景,且生成的句子也更为准确,语义更加丰富。

2 图像精细化描述

通过对图像简单描述常用模型与框架进行阐述,可以看出,这些模型为图像所生成的都是单句描述,对图像中物体、动作及位置关系等视觉语义的描述粒度较为粗糙,忽略了大量的细节信息,难以真实再现图像中的语义内容。在早期基于手工特征与预设模板的描述框架中,Kulkarni 等人(2013)设计了一种更为复杂的框架,检测到视觉语义对象后,通过属性映射和关系推理,实现对象的精细化描述。然后使用条件随机场(CRF),生成具有一定逻辑的描述语段,对图像中的细节信息及逻辑信息进行了较为完整的表达。只是由于手工特征的弱表现力与模板的欠灵活性,其模型的鲁棒性不强,句子的语义性也受到限制。

随着深度学习技术的发展,学者也投入越来越多的精力开发相关模型,为图像生成更为准确、语义性更强的精细化描述。Krishna 等人(2017)首先提出了视觉基因组计划 (visual genome),认为视觉信息是由多种视觉语义(“基因”)所构成,而不同的视觉语义相互组合,最终形成人们可以理解的视觉语义对象及相关视觉描述,为用户提供更为多样化的精细化表达(密集描述(dense captioning))。其示例如图 5 所示。

Johnson 等人(2016)提出了一种全卷积定位网

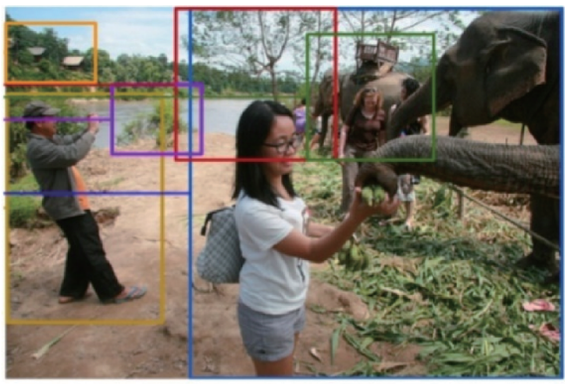


图 5 图像密集描述示例

Fig. 5 Example of image dense captioning

络,通过在模型中添加一个视觉语义区域定位模块,对候选区域进行插值与回归运算,实现区域定位与语言模型的联合优化。该工作启发了人们对于图像更为丰富表达的探索。但 Yang 等人(2017)认为,仅使用目标检测所生成的视觉区域因重叠可能导致位置不准确,且单个区域难以生成具有真正意义的描述句子。为此,提出一种多区域联合推理模型,将各视觉区域结合起来实现上下文语境融合,解决区域定位不准确及生成无关场景的视觉描述问题。

Li 等人(2019)采取了一种更为简洁的方法,采用 gLSTM 为检测到的视觉对象进行上下文编码,并以此指导每个区域的句子生成及视觉区域边界框的位置微调,实现更为精准的密集描述。Yin 等人(2019)同样为解决上下文引用的问题,引入了局部信息、邻居信息与全局信息,从多个尺度上辅助每条语句的生成。Kim 等人(2019)设计了一种更为复杂的模型的三重数据记忆流模型,根据人们在描述图像中视觉对象关系时常使用<主—介—宾>形式的习惯,为每种视觉对象区域(主/介/宾)设计一条专用的编码网络,然后使用多任务学习(multi-task learning)的方法为每个区域生成更具价值的视觉对象关系描述。这些方法都对直接使用视觉区域生成描述进行了改进,虽然密集描述中对视觉对象的内在逻辑关注较少,但研究者们也不约而同地认为上下文语境对于生成可用且更为准确的区域描述的重要性,认识到图像各语义对象之间并不是完全孤立的,必须结合其周围的具体环境生成相应的细节描述,使各语句之间既相互独立,又互有联系。

图像的密集描述是为图像生成更多条语句,各语句之间虽然可能具有一定的关联性,但其相互之间是离散且杂乱的,不能形成流畅、合理的结构化描述语段。这违背了人们在日常生活中对一幅图像的描述表达习惯,且削弱了图像描述在具体实践中的应用价值。因此,除通过为视觉语义区域生成密集描述实现图像的精细化表达之外,人们也开始着力于研究为图像生成具有一定逻辑层次的描述语段,进一步丰富视觉高层理解的内涵。Huang 等人(2016)提出了一种基于图像流的视觉描述任务,将具有相互关联的多幅图像作为研究对象,根据其中的视觉内容变化,生成具有一定故事情节的段落描述。Liu 等人(2017b)引入每幅图像与上下文描述之间的语义相关性,并使用双向注意力机制对视觉语义进行区别关注,使得生成的语段更加准确,表达更为严谨。Wang 等人(2018a)将强化学习与对抗学习的思想引入图像流的描述任务中,使用层次化 RNN 作为句子生成器,并设计多模判别器与语言风格判别器计算奖励,整个模型通过对抗训练的方式进行优化。

Krause 等人(2017)则重点关注单幅图像的段落描述任务,与图像流的描述相比,这也是人们日常生活中更为常见的问题,论文设计了一个层次式的 RNN 网络,通过语句级 RNN 确定生成句子的数量及其相关主题,然后使用词汇级 RNN 逐个生成词汇组成句子,最后将所有句子组合在一起形成描述段落。其模型基本框架如图 6 所示。

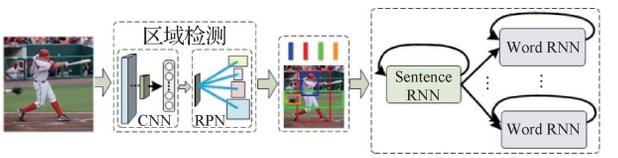


图 6 层次式 RNN 的图像段落描述生成框架

Fig. 6 Hierarchical RNN based framework for image paragraph description

通过构建多级 RNN 网络,模拟段落的层次化结构,为图像生成段落描述提供了直观思路,是其他一些相关工作的思想基础。Liang 等人(2017)也使用了语句级和词汇级两级 RNN 结构为图像生成段落表达,但其采用了对抗优化的方法,通过结构化段落

生成器与多级句子判别器之间的相互对抗迭代，实现模型的最终优化。Che 等人(2018)提出一种较为直接的方法，首先检测图像中较重要的视觉对象区域，然后判断各主要区域之间的关系，并采用 one-hot 方式将其各视觉关系表示成特征向量，最后将各关系特征与区域特征结合在一起输入语言模型(多条 RNN 网络)，生成具有一定结构的逻辑语段。Chatterjee 和 Schwing(2018)为了生成连贯、结构化及多样化的图像段落描述，通过句子级 RNN 生成全局主题特征向量，使多条句子之间具有一致性，并在此基础上结合词汇级 RNN 生成连贯性特征向量，保证句子之间的平滑过渡，最后采用变分自动编码器挖掘图像中的固有隐藏信息，保证生成的段落描述具有多样化与个性化特征。Wang 等人(2019)也采用了使用图像主题特征指导语段生成的思路，将视觉语义对象的特征通过卷积网络生成多个主题特征，然后使用反卷积网络对特征进行重构，进而形成全局主题特征，引导语言模型生成段落描述。在这些方法中，生成各区域的单条描述句子已经不再是关注的重点，其各视觉对象之间的内在关联挖掘，以及在描述中如何体现这种关联关系，生成具有严谨逻辑结构的语段成为模型设计时的核心。但鉴于视觉语义与自然语言中逻辑表达的复杂性，当前模型生成的语段在语言结构合理性、逻辑准确性等方面尚存在大量问题需要解决。

3 图像情感与个性化描述

除了为图像生成更为详细的精细化描述外，人们也注意到在日常交流中其语言常蕴含多种个性化和情感信息。在描述一幅图像时，常常根据个人经验和观感在句子中掺杂多种情感信息。由此，研究者提出了融合个性化或情感信息的图像描述任务，旨在进一步缩小视觉数据与自然语言表达之间的差异性，增强描述句子的灵活性和生动性。在早期研究中，人们曾经对视觉数据中的情感信息进行分析，按照视觉内容，对图像内部的情感进行分类识别，或者分析图像内容对观看者可能产生的情感影响(Machajdik 和 Hanbury，2010；Chen 和 Jin，2016；You 等，2016a；Zhao 等，2018)。这种形式的视觉数据情感分析其实与分类识别、检索等任务类似，属于视觉低层或中层理解范畴。而在图像描述中，不仅需要分析情感，还需要在如何选择合适的词汇、风格及与其他视觉对象的关系等方面进行研究，将情感与事实有机地融合在一起，形成更具吸引力的图像标题与描述。

Mathews 等人(2016)从情感表达的个性化特点出发，通过设计一种开关式 RNN 单元，为图像生成具有“积极(positive)”或“消极(negative)”情感的描述句子。这种方法本质上是在传统的 RNN 单元上增加了一个“情感门”函数，通过主动控制或自主学习参考句子中的情感极性，优化其中的参数，实现为相同图像生成具有不同情感极性的句子。该工作在思想上认为情感可以通过具体的描述进行表现，而不是局限于某些选定的情感词汇，其在方法上通过情感极性阈值将情感与其他语义表达无缝地衔接在一起，与人们的日常表达具有相似的效果。但这种方法对于情感的定义粒度较粗，难以适应更为复杂的视觉情感表达。Shin 等人(2016)使用多标签学习机制训练包括物体与情感的两个 CNN 模型，通过融合两个模型的视觉特征，为语言模型提供情感与事实描述信息，生成具有情感极性的描述句子。Karayil 等人(2019)将“积极”与“消极”表示为二值情感特征向量，结合对抗学习的思想，通过计算基于情感与基于关联的判别奖励，实现生成模型的优化。Gan 等人(2017a)则从风格表达的角度出发，提出为图像生成具有不同风格类型的描述，通过矩阵分解的方式将 RNN 单元中的权值矩阵拆分为三种不同的风格因子，并使用多任务学习机制，为图像生成“幽默(humor)”、“浪漫(roman)”和“事实(factual)”3 种风格的描述语句。这种方式模拟了不同的人对于相同事物的不同理解与描述风格，但风格的类型不仅仅包含以上 3 种，采用矩阵分解与多任务学习的方法可能难以应对描述风格的多样化。Chen 等人(2018)综合了 Mathews 等人(2016)和 Gan 等人(2017a)的思想，扩展了情感类型，提出了一种自适应情感注意力机制，在确保事实描述准确性的基础上，使得生成的描述能够表现出更多的情感风格(“积极”、“消极”、“幽默”、“浪漫”)。Guo 等人(2019)采用了对抗学习的思想，首先根据图像特征与输入的风格标志生成相应的风格化句子，然后将其与真实的描述句子进行对比分析，使用判别器判断句子真假，并对句子风格进行分类，以此实现模型的优化目标。同样地，Zhao 等人(2020)也采用了同

样的情感分类思路,但认为无论是不同的情感极性还是多样的情感风格,其实都属于描述风格问题,论文设计了一种句子分解算法,将反映描述风格的部分与视觉事实的描述部分进行分离,然后结合风格标志(即情感极性或风格)对特定的风格记忆单元进行训练,在生成句子时,则根据不同的风格标志生成相应的句子。

除生成蕴含一定情感信息的描述句子外,人们也寻求为图像生成具有一定个性化色彩的描述。Park 等人(2017)为了生成更为个性化的图像描述,提出一种上下文序列记忆网络,当给定一幅图像后,用户输入个性化信息,使用不同的 CNN 网络分别对图像特征与文本特征进行编码,并在每个时间步上结合注意力机制对记忆单元进行更新,最终生成可定制的个性化图像描述句子。Shuster 等人(2019)统计了 200 余种常用的个性化特征(如“sweet”、“anxious”等),在生成图像描述时,分别使用 CNN 和转换器(transformer)对图像和词汇进行编码,同时将输入的个性化特征编码为嵌入式特征向量,并将三者映射到同一特征空间中,然后结合注意力机制,为其生成具有相应色彩的描述句子。

对于图像描述任务的研究目前仍在发展之中,人们不仅在模型和方法上不断做出尝试与改进,同时也根据实际需求,提出了更多的任务分支。在方法层面,借鉴机器翻译的工作流程,采用“编码—解码”框架,将图像作为“源语言”,将待生成的句子描述作为“目标语言”。在多种改进模型中,也多是参考了机器翻译中的概念和思路,如注意力机制、因子分解架构等。从总体上看,学者还是更多地关注于视觉信息的合理使用,以及如何使得视觉信息与语言的有效对应,但对于自然语言处理,如句子的后续优化、语言模型的预训练等,还未获得足够的重视。目前,谷歌公司(Devlin 等,2019)开发了一个用于自然语言理解的 BERT(bidirectional encoder representation from transformers)框架,通过双向转换器(bidirectional transformer)对自然语言进行编码,并设计了遮挡语言模型和后续语句预测模型分别捕捉词汇级别和句子级别的特征表达。将该框架与图像描述任务有机融合是值得进一步研究的方向,尤其是其使用转换器(Vaswani 等,2017)的思路,能够进一步克服 RNN 网络中由于梯度消失而导致的训练困难的问题,同时也能够实现模型的并行部署与运算。而对于延伸出的任务分支,包括密集描述、语段描述和情感描述等各有其应用场景,尤其是生成具有合理逻辑关系的段落描述,研究价值巨大,应用前景广阔,但由于逻辑本身的复杂性,目前还存在可解释性的模型构建困难、评价指标与数据集缺失等大量问题亟待解决。

4 相关数据集与评价方法

模型与方法不仅需要经过理论上的严格证明或详细论述,也需要通过大量的实验对模型的有效性和优越性进行对比验证。在该过程中,需要同时考虑生成描述的评价指标与实验选择的评测数据集两个方面。在评价指标方面,不同于简单的分类、识别等视觉任务,它不仅需要对生成句子的准确性进行统计,还要对句子的连贯性、语义性等方面进行衡量,因此,其指标设计较为复杂。对于实验评测,目前针对图像的标题生成与描述多为数据集内测试,即在同一个封闭的数据集内进行模型训练、参数寻优与最终测试。

4.1 生成描述的评价方法与具体指标

目前,对于图像与视频描述的评价机制多是参考机器翻译中的方法,即将视觉内容对应的参考句子与生成句子进行比较分析,按照不同的标准对用词、短语等的准确率进行统计,计算参考句子与生成句子之间的距离等。常用的评价指标主要包括 BLEU、METEOR、ROUGE-L(recall-oriented understudy for gisting evaluation)(Lin 和 Och, 2004)等。其中 BLEU 指标主要用于衡量生成句子的准确性与连贯性,通过统计生成句子中与参考句子中“n-元组”的匹配程度对生成句子进行打分(通常使用 B-n 表示其在不同“n-元组”下的分值),其中 n 一般取 {1, 2, 3, 4},在 n 确定的情况下,BLEU 值越高,说明句子的连贯性越高。在该指标中,对句子长度也有一定的要求,若生成句子的长度(词汇数)比参考句子短,则通过惩罚因子降低相应的分值。该方法在计算时主要关注了词汇和短语的准确率,忽视了召回率。相对而言,METEOR 和 ROUGE-L 指标对句子的评价更为全面。对于 METEOR 指标,总体思想是使用匹配对齐的方式,通过定义惩罚因子、调和均值因子,兼顾待评价句子中所用词汇与短语的准确率和召回率。首先将待评价句子与参考句子中的

词汇或短语按照精准匹配、同义匹配与词根匹配的方式，按顺序搜索并匹配其各自的最大值。如果3种匹配方式所得到的最大值相同，则将其两两匹配中交叉次数最少的匹配作为对齐集合中的可用元素。按照这种方式进行迭代，得到最终的对齐集合。然后将该集合的元素个数与待评价句子词汇数的比值作为准确率，与参考句子中词汇数的比值作为召回率。而对于ROUGE-L指标而言，虽然也同时考虑了生成词汇与短语的准确率与召回率，但它是以最长子串匹配方法为基础，更为关注句子的连贯性。以上3种方法对句子的质量评价各有侧重，但整体而言，METEOR方法考虑的因素更为全面，其通过多种匹配方式（尤其是同义匹配或词根匹配），不仅能够对句子的准确性与连贯性进行衡量，同时也能够在一定程度上反映出句子的语义性。因此，人们在图像的密集描述工作中，更倾向于使用该指标对生成的多条语句或语段进行评价。

以上方法都是针对机器翻译任务设计的，因此其只从自然语言的角度衡量生成句子的质量，这对于视觉描述任务来说，则割裂了视觉信息与语言之间的联系。为此，学者也开始针对视觉描述问题而设计新的评价方法，以期建立能够联系视觉语义的生成句子评价机制。目前，这类较为流行的指标主要包括CIDEr和SPICE。其中CIDEr指标主要是基于图像参考句子集合共同语义的思想（或称为“共识”），通过计算生成句子与相应参考句子集合的余弦距离，评价两者的语义契合程度；在该过程中，主要通过统计参考句子中“n-元组”的分布规律，为生成句子中不同的“n-元组”分配不同的TF-IDF（term frequency-inverse document frequency）值，克服其他方法对所有“n-元组”一视同仁的弊端，体现了对生成句子中语义性的重视程度。由于该方法更符合人们对句子的评价习惯，与BLEU、METEOR等指标一样，已成为视觉描述的主要评价标准之一。SPICE指标的计算方法更为关注图像中视觉语义对象（包括物体、属性、关系）的准确程度。首先使用一种概率上下文无关语法依赖（probabilistic context-free grammar，PCFG）（Klein和Manning，2003）的分析方法，将句子解析为句法依赖树，然后将其映射为各语义对象的有向场景图，并根据场景图的匹配程度，衡量生成句子的质量；在具体计算时，参考句子集合与生成句子的场景图转换为“n-元组”的集合，然后借鉴METEOR方法中的匹配方法统计匹配集合，并计算生成句子中视觉语义对象的准确率与召回率，最后使用调和均值的方式得到最终分值。该指标与CIDEr指标一样，其设计目的都是为了更加合理地衡量句子的语义性，但其更侧重于名词性的视觉语义对象，对于动态性的语义判断可能不够准确。因此，在具体评判时，一般需要结合多个指标，从不同侧面对句子质量进行综合评价。

在视觉密集描述任务中，还常使用平均准确率均值（mean average precision，mAP）指标同时对定位准确性与描述准确性进行衡量。该指标最初用于目标检测任务，后由Johnson等人（2016）进行改进并用于图像密集描述任务的评价。它使用联合交叉（intersection over union，IoU）机制，将区域重叠阈值在$\{0.2, 0.3, 0.4, 0.5, 0.6\}$上的精度平均值作为对定位准确性的衡量，同时使用METEOR分值在$\{0, 0.05, 0.10, 0.15, 0.20, 0.25\}$上的平均值衡量参考句子与生成句子的相似度，最后计算两者均值作为整体评价指标。

4.2　图像标题生成与描述数据集

4.2.1　图像简单描述数据集及模型性能对比

针对图像标题生成与描述，目前已有多个面向不同任务的常用数据集。在传统的单条句子描述方面，较为常用的数据集包括MS COCO2014（Microsoft common objects in context）（Lin等，2014）、Flickr30K（Young等，2014）、Flickr8K（Hodosh等，2013）等。其中MS COCO2014数据集由微软研究院收集并发布，该数据集不仅用于图像描述，还可以用于目标检测、图像分割等任务。在图像描述部分，可以公开使用的图像共包含123 287幅，每幅图像对应5条人工标注的描述语句，其中训练集包含82 783幅图像及其描述句子，验证集包含40 504幅图像及其对应描述。在具体使用时，目前一般按照Karpathy和Li（2015）的划分标准，即从验证集中选取10 000幅图像及其描述句子，其中5 000幅图像及其描述用于验证，另外5 000幅图像用于最终测试。该数据集目前已成为事实上的图像描述评测标准数据集，不同的方法和模型在该数据集上的表现是评价其性能优劣的重要指标之一。表1—表3中分别列出了当前主流方法对于该数据集在各评价指标上的主要性能表现。需要指出的是，由于各模型使用的DCNN模型可能不同（如NIC模型使用的是GoogLeNet，

LRCN 使用的是 AlexNet,而 OPR-MCM(online positive recall and missing concepts mining)则使用的是 ResNet101,因此其结果只能表示该工作在 MS COCO2014 数据集上获得了相应的性能,并不能完全反映出各方法的优劣。从整体上看,使用以评价指标为目标函数的强化学习框架能够更容易地获得比其他方法更优越的性能表现,这也反映出以交叉熵作为目标函数与最终评价指标的分离,造成了另外一种局部最优问题。基于无监督训练方式的模型,其性能与其他方法在各个指标上相比,仍有较大差距,但作为一种极具应用价值的研究思路,值得人们继续探索。而基于全局特征与注意力机制(或视觉属性、视觉概念)的方法在结合诸如深度监督、深度融合和 MIL 等技术的情况下,也可以获得更好的性能表现,但整体而言,其模型对数据的处理过程较为复杂,模型的鲁棒性与环境适应性可能会受到很大约束。

表 1　基于全局特征的模型在 MS COCO2014 数据集上的性能表现

Table 1　Performance of some global visual feature based models on MS COCO2014 dataset

方法	B1/%	B2/%	B3/%	B4/%	METEOR/%	CIDEr
multi-modal RNN(Karpathy 和 Li,2015)	62.5	45.0	32.1	23.0	19.5	66.0
Google NIC(Vinyals 等,2015)	–	–	–	27.7	23.7	85.5
LRCN(Donahue 等,2017)	62.8	44.2	30.4	21.0	–	–
m-RNN(Mao 等,2015)	67.0	49.0	35.0	25.0	–	–
Tang 等人(2018)	74.1	57.8	43.8	33.0	25.8	101.8

注:“–”表示相应方法没有提供对应指标实验结果。

表 2　基于视觉特征选择与优化的模型在 MS COCO2014 数据集上的性能表现

Table 2　Performance of some visual feature selection and optimization based models on MS COCO2014 dataset

方法	B1/%	B2/%	B3/%	B4/%	METEOR/%	ROUGE-L/%	CIDEr
soft-attention(Xu 等,2015)	70.7	49.2	34.4	24.3	23.9	–	–
hard-attention(Xu 等,2015)	71.8	50.4	35.7	25.0	23.0	–	–
Att + CNN + LSTM(Wu 等,2016)	74.0	56.0	42.0	31.0	26.0	–	94.0
SCA-CNN-ResNet(Chen 等,2017)	71.9	54.8	41.1	31.1	25.0	53.1	95.2
adaptive attention(Lu 等,2017)	74.2	58.0	43.9	33.2	26.6	–	108.5
seq-attention(Fang 等,2018)	–	–	45.0	34.9	26.7	–	108.1
Att-FCN(You 等,2016b)	70.9	53.7	40.2	30.4	24.3	–	–
LSTM-A5(Yao 等,2017)	73.4	56.7	43.0	32.6	25.4	54.0	100.2
SCN-LSTM(Gan 等,2017b)	72.8	56.6	43.3	33.0	25.7	–	101.2
OPR-MCM(Zhang 等,2019a)	75.8	59.6	46.0	35.6	27.3	56.0	110.5
semantic-guided(Zhang 等,2019b)	71.2	51.4	36.8	26.5	24.7	–	88.2
cascade-attention(Shi 等,2019)	79.4	63.7	48.9	36.9	27.9	57.7	122.7
SGAEfuse(Yang 等,2019)	81.0	–	–	39.0	28.4	58.9	129.1

注:“–”表示相应方法没有提供对应指标实验结果。

对于 Flickr30K 数据集,其数据量相对较小,共包含反映日常动作、事件和场景的 31 783 幅图像。与 MS COCO2014 一样,其中每幅图像对应 5 条人工标注的描述句子。按照 Karpathy 和 Li(2015)的划分标准,使用其中的 29 000 幅图像及其参考句子作为训练集,1 000 幅图像及其参考句子作为验证集,其余样本作为测试集。该数据集一般也作为图像描述模型测试的重要标准,与 MS COCO2014 数据集一

起,验证模型的有效性。在该数据集上,不同的模型方法性能对比如表 4 所示。此外,对于 Flickr8K 数据集,其样本量更少,共包含 8 091 幅图像。同样地,每幅图像对应 5 条参考句子。在具体使用时,一般选取其中的 6 000 幅图像及其参考句子用于模型训练,另外 1 000 幅图像及其参考句子用于模型验证,其余 1 091 幅图像用于最终的模型测试。在该数据集上,各模型的性能表现如表 5 所示。由结果可以看出,在 Flickr30K 和 Flickr8K 两个数据集上参与测试的模型较少,且提供的可供参考的性能指标结果也不全面。多个模型在这两个数据集上的性能也并未表现出很大差异,因此,一般是将在该数据集上的结果作为模型验证的补充,验证模型在较小数据集上的通用性和适用性。

表 3　面向优化策略的模型在 MS COCO2014 数据集上的性能表现

Table 3　Performance of some optimization strategy oriented models on MS COCO2014 dataset

方法	B1/%	B2/%	B3/%	B4/%	METEOR/%	ROUGE-L/%	CIDEr	SPICE/%
emb-gLSTM(Jia 等,2015)	67.0	49.1	35.8	26.4	22.7	–	81.3	–
up-down (Anderson 等,2018)	79.8	–	–	36.3	27.7	56.9	120.1	21.4
SCST (Rennie 等,2017)	–	–	–	31.9	25.5	54.3	106.3	–
Ren 等人(2017b)	71.3	53.9	40.3	30.4	25.1	52.5	93.7	–
hierarchical attention(Yan 等,2020)	72.6	52.8	37.8	27.2	24.7	56.0	88.1	–
CNN + attention(Aneja 等,2018)	72.2	55.3	41.8	31.6	25.0	53.1	95.2	17.9
unsupervised(Feng 等,2019)	58.9	40.3	27	18.6	17.9	43.1	54.9	11.1

注:“–”表示相应方法没有提供对应指标实验结果,SCST 为 self-critical sequence training。

表 4　不同模型在 Flickr30K 数据集上的性能表现

Table 4　Performance of some models on Flickr30K dataset

	方法	B1/%	B2/%	B3/%	B4/%	METEOR/%	CIDEr
基于全局特征	LogBilinear(Kiros 等,2014)	60.0	38.0	25.4	17.1	16.9	–
	multi-modal RNN(Karpathy 和 Li,2015)	57.3	36.9	24.0	15.7	15.3	–
	Google NIC(Vinyals 等,2015)	66.3	42.3	27.7	18.3	–	–
	LRCN(Donahue 等,2017)	58.7	39.1	25.1	16.5	–	–
	m-RNN(Mao 等,2015)	60.0	41.0	28.0	19.0	–	–
	Tang 等人(2018)	68.1	50.1	36.0	25.8	20.6	–
基于视觉特征选择与优化	soft-attention(Xu 等,2015)	66.7	43.4	28.8	19.1	18.5	–
	hard-attention(Xu 等,2015)	66.9	43.9	29.6	19.9	18.5	–
	SCA-CNN-ResNet(Chen 等,2017)	66.2	46.8	32.5	22.3	19.5	–
	adaptive attention(Lu 等,2017)	67.7	49.4	35.4	25.1	20.4	53.1
	seq-attention(Fang 等,2018)	–	–	34.5	24.8	19.4	47.8
	Att-FCN(You 等,2016b)	64.7	46.0	32.4	23	18.9	–
	SCN-LSTM(Gan 等,2017b)	73.5	53.0	37.7	26.5	21.8	–
面向优化策略	emb-Glstm (Jia 等,2015)	64.6	44.6	30.5	20.6	17.9	–

注:“–”表示相应方法没有提供对应指标实验结果。

表 5　不同模型在 Flickr8K 数据集上的性能表现

Table 5　Performance of some popular models on Flickr8K dataset

方法		B1/%	B2/%	B3/%	B4/%	METEOR/%
基于全局特征	LogBilinear(Kiros 等,2014)	65.6	42.4	27.7	17.7	17.3
	multi-modal RNN(Karpathy 和 Li,2015)	57.9	38.3	24.5	16.0	16.7
	Google NIC(Vinyals 等,2015)	63.0	41.0	27.0	–	–
基于视觉特征选择与优化	soft-attention(Xu 等,2015)	67.0	44.8	29.9	19.5	18.9
	hard-attention(Xu 等,2015)	67.0	45.7	31.4	21.3	20.3
	SCA-CNN-ResNet(Chen 等,2017)	68.2	49.6	35.9	25.8	22.4
面向优化策略	emb-gLSTM(Jia 等,2015)	64.7	45.9	31.8	21.6	20.2

注:“–”表示相应方法没有提供对应指标实验结果。

除基于英语的图像描述数据集外,目前来自中国的一些企业也在积极推进基于中文的图像描述数据集建设。由创新工场、搜狗等公司联合举办的全球 AI 挑战赛(AI Challenger)中,专门设置了图像中文描述赛道。其构建的数据集更为庞大,共包含图像 300 000 幅,图像内容更为复杂,每幅图像对应 5 条参考句子,且句子中引入了更多的形容词、成语等,其表达更为灵活,语义更为丰富。由于其较高的挑战性,吸引了众多的高校与研究机构参与其中,推动了图像描述技术在实际场景中的应用进程。

4.2.2　图像密集描述与段落描述数据集及模型性能

以上图像描述数据集中,其描述语句一般为单条语句。对于图像密集描述任务,目前常用的数据集主要是 Visual Genome (VG)(Krishna 等,2017)。该数据集已发布多个版本(包括 VG1.0,VG1.1 和 VG1.2),但人们在使用时,为了对比的公平性,仍多使用 VG1.0 对模型进行验证。该版本数据集中共包含了 94 313 幅图像,主要来源于 MS COCO 与 YFCC100M(Yahoo Flickr Creative Commons 100 Million),每幅图像对应的描述句子条数不等,总数超过 410 万条。为控制数据集质量,多于 10 个词汇的标注句子被丢弃,同时标注句子少于 20 条或多于 50 条的图像被去除,最终得到 87 398 幅可用图像及其标注。按照常用的使用方法,各取 5 000 对样本作为验证集与测试集,剩余的 77 398 幅图像及其标注作为训练集。在该版本数据集上,目前常用方法的性能表现如表 6 所示。

表 6　密集描述模型在 Visual Genome(VG1.0)数据集上的性能表现

Table 6　Performance of some image dense captioning models on Visual Genome(VG1.0) dataset

方法	mAP/%
FCLN(Johnson 等,2016)	5.39
JIVC(Yang 等,2017)	9.31
COCG(Li 等,2019)	9.82
CAG-Net(Yin 等,2019)	10.51

注:FCLN 为 fully convolutional localization network, JIVC 为 joint inference and visual context fusion, COCG 为 captioning based on object context as guidance, CAG-Net 为 context and attribute grounded network。

VG数据集为每幅图像圈定了更为密集的视觉语义对象,并对其进行描述。部分研究人员认为,对图像进行过多的圈定并对其进行标注可能是没有必要的。Li 等人(2019)对比了 VG 数据集与 MS COCO 和 ImageNet 中目标检测数据集,发现 VG 训练集中每幅图像的圈定框数量为其他数据集的 5 倍左右。为了平衡圈定框的数量,为图像生成更为合理的多条描述,对 VG 进行了改进,将其与 MS COCO 目标检测数据集进行交叉取样,形成数据量更小的 VG-COCO 数据集。按照其使用标准,最终的训练集包含了 38 080 幅图像,验证集与测试集则分别包含 2 489 幅和 2 476 幅图像。该数据集由于其包含样本量更少,可能会导致模型产生过拟合现象,因此,并未得到大规模的使用;但其缩减冗余圈定框及其标注的思路使得图像密集描述任务更符合现实需求,值得人们进一步探索,构建更大更实用的数据集。

VG 数据集主要用于图像的密集描述模型验

证,与图像对应的多条描述句子通常是离散的,各条句子之间缺乏联系与逻辑性。Krause 等人(2017)在 VG 数据集的基础上,构建了一个用于图像段落描述验证的数据集。从 VG 数据集中选出部分图像,为其进行段落标注,将其中的视觉内容作为一个整体进行细粒度刻画,其段落的平均长度为 MS COCO 数据集中标注句子的 6 倍左右,描述更为详细、具体,语义更加丰富。该数据集最终共包含 19 551 幅图像及其段落描述,按照 Krause 等人(2017)的划分标准,训练集样本量为 14 575 幅,验证与测试样本分别为 2 487 幅与 2 489 幅。目前该数据集已被广泛引用,成为图像段落描述的通用验证数据集。部分模型与方法在该数据集上的性能表现如表 7 所示(为便于表示,本文将该数据集标记为 VG-P(Paragraph))。

表 7　段落描述模型在 VG-P 数据集上的性能表现

Table 7　Performance of some image paragraph description models on VG-P dataset

方法	B1/%	B2/%	B3/%	B4/%	METEOR/%	CIDEr
Human(Krause 等,2017)	42.88	25.68	15.55	9.66	19.22	28.55
Image-Flat(Karpathy 和 Li,2015)	34.04	19.95	12.20	7.71	12.82	11.06
Regions-Hierarchical(Krause 等,2017)	41.90	24.11	14.23	8.69	15.95	13.52
RTT-GAN (Semi + Fully)(Liang 等,2017)	42.06	25.35	14.92	9.21	18.39	20.36
Che 等人(2018)	41.74	24.94	14.94	9.34	17.32	14.55
CapG-RevG(Chatterjee 和 Schwing,2018)	42.38	25.52	15.15	9.43	18.62	20.93
CAE-LSTM(Wang 等,2019)	–	–	–	9.67	18.82	25.15

注:“–”表示相应方法没有提供对应指标实验结果。

4.2.3　图像情感描述数据集及各模型性能

在融合情感的图像描述领域,也已构建出多个相应的验证数据集。由于情感的多样性,其相关数据集的构建过程也更为复杂。目前较为流行的数据集包括 SentiCap(Mathews 等,2016)、FlickrStyle10K(Gan 等,2017a)等。其中 SentiCap 数据集中的图像主要取自于 MS COCO;在标注时,其主要从情感的极性出发,借鉴 Visual SentiBank 数据集(Borth 等,2013)中的“形容词—名词对(adjective noun pair,ANP)”,将其嵌入到描述句子中,为每幅图像形成“正面(positive)”和“负面(negative)”的图像描述。该数据集共选用了 1 027 个正面的 ANPs 与 436 个负面的 ANPs,其最终共包含了 3 171 幅图像,每幅图像对应至少 3 条正面、3 条负面的语句描述。按照常用的使用标准,998 幅图像及其对应的 2 873 条含有正面情感的句子用于正面训练,673 幅图像与对应的 2 019 条句子用于测试;997 幅图像及其对应的 2 468 条含有负面情感的句子用于负面训练,503 幅图像及其 1 509 条句子用于模型测试。在该数据集上,部分方法与模型的性能表现如表 8 所示。

FlickrStyle10K 数据集(Gan 等,2017a)中的图像主要来源于 Flickr30K。在具体标注时,根据给定的事实描述及指定风格,为每幅图像重新书写具有不同风格的描述语句。在原数据集中,共包含了约 10 000 幅图像,每幅图像对应 5 条事实描述句子,1 条幽默风格的句子,1 条浪漫风格的句子。按照 Gan 等人(2017a)的使用标准,选取其中的7 000 幅图像及其描述用于训练,2 000 幅图像及其描述用于验证,其余的样本则作为测试集。但该数据集并未完全公开,目前可用的部分主要是训练集,因此,使用时选取其中的 6 000 幅图像及其描述用于训练与验证,剩余的 1 000 幅图像与描述则作为测试集。目前常用的模型与方法在该数据集上的性能表现如表 9 所示。

5　结　语

图像标题生成与描述将具象的视觉数据转换为更为抽象的自然语言,其与机器翻译具有相似的处理流程,但相比于语言,视觉数据更为复杂,信息量更大,且融合了多种拍摄者与观看者的个人体验,将其内容编码为计算机可理解的中间特征更为困难。在将中间特征“翻译”成自然语言时还需要考虑各视觉语义对象之间的关系,以及整条句子或

表 8　情感描述模型在 SentiCap 数据集上的性能表现

Table 8　Performance of some image sentimental description models on SentiCap dataset

情感极性	方法	B1/%	B2/%	B3/%	B4/%	METEOR/%	ROUGE-L/%	CIDEr
正面	NIC(CNN + RNN)(Vinyals 等,2015)	48.7	28.1	17.0	10.7	15.3	36.6	55.6
	RNN-Transfer(Mathews 等,2016)	49.3	29.5	17.9	10.9	17.0	37.2	54.1
	SentiCap(Mathews 等,2016)	49.1	29.1	17.5	10.8	16.8	36.5	54.4
	SF-LSTM + Adap(Chen 等,2018)	50.5	30.8	19.1	12.1	16.6	38.0	60.0
	MSCap(Guo 等,2019)	46.9	–	16.2	–	16.8	–	55.3
	MeMCap(Zhao 等,2020)	50.8	–	17.1	–	16.6	–	54.4
负面	NIC(CNN + RNN)(Vinyals 等,2015)	47.6	27.5	16.3	9.8	15.0	36.1	54.6
	RNN-Transfer(Mathews 等,2016)	47.8	29.0	18.7	12.1	16.2	36.7	55.9
	SentiCap(Mathews 等,2016)	50.0	31.2	20.3	13.1	16.8	37.9	61.8
	SF-LSTM + Adap(Chen 等,2018)	50.3	31.0	20.1	13.3	16.2	38.0	59.7
	MSCap(Guo 等,2019)	45.5	–	15.4	–	16.2	–	51.6
	MemCap(Zhao 等,2020)	48.7	–	19.6	–	15.8	–	60.6

注:“–”表示相应方法没有提供对应指标实验结果。

表 9　情感描述模型在 FlickrStyle10K 数据集上的性能表现

Table 9　Performance of some image sentimental description models on FlickrStyle10K dataset

情感风格	方法	B1/%	B2/%	B3/%	B4/%	METEOR/%	ROUGE-L/%	CIDEr
浪漫	CaptionBot*(Tran 等,2016)	40.4	20.2	12.7	7.6	13.3	36.0	26.0
	NIC(CNN + RNN)*(Vinyals 等,2015)	42.0	21.4	12.5	7.8	13.4	36.0	28.0
	Multi-task*(Luong 等,2016)	44.1	23.7	14.3	9.5	14.5	36.0	29.0
	StyleNet(R)*(Gan 等,2017a)	46.1	24.8	15.2	10.4	15.4	38.0	31.0
	SF-LSTM + Adap(R)(Chen 等,2018)	48.2	31.5	20.6	13.5	18.7	40.2	26.7
幽默	CaptionBot*(Tran 等,2016)	43.4	21.4	12.2	7.1	13.4	35.0	21.0
	NIC(CNN + RNN)*(Vinyals 等,2015)	43.1	22.8	13.2	7.9	13.6	36.0	23.0
	Multi-task*(Luong 等,2016)	47.1	23.9	13.9	8.8	14.8	37.0	25.0
	StyleNet(R)*(Gan 等,2017a)	48.7	25.4	14.6	10.1	15.2	38.0	27.0
	SF-LSTM + Adap(R)(Chen 等,2018)	51.5	34.6	23.1	15.4	19.3	41.7	34.2

注:“*”表示结果来自 Gan 等人(2017a)文献。

语段的逻辑结构,因此,该任务具有很大的挑战性。目前,人们借助深度模型、注意力机制和视觉属性/概念等技术,已能够为图像生成较为连贯且语义丰富的描述语句。在此基础上,人们也提出了多种更有针对性的描述任务,如密集描述、结构化描述和情感/风格化描述等,并提出了可行的解决方案,丰富了图像描述的内涵,推动了该领域进一步走向应用。

本文系统性地阐述了图像描述的发展历史、当前研究现状及其动态,对其主流工作进行了梳理,并对面临的问题与挑战进行了分析,指出研究构建图像的情感与逻辑性精细化表达的有效模型,提高深度学习模型的可解释性,以及构建更优秀的数据集和有效的评价方法,是图像标题生成与描述任务未来待解决的问题和研究方向。

参考文献(References)

Anderson P, Fernando B, Johnson M and Gould S. 2016. SPICE: semantic propositional image caption evaluation//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer: 382-398 [DOI: 10.1007/978-3-319-46454-1_24]

Anderson P, He X D, Buehler C, Teney D, Johnson M, Gould S and Zhang L. 2018. Bottom-up and top-down attention for image captioning and visual question answering//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake, USA: IEEE: 6077-6086 [DOI: 10.1109/CVPR.2018.00636]

Aneja J, Deshpande A and Schwing A G. 2018. Convolutional image captioning//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 5561-5570 [DOI: 10.1109/CVPR.2018.00583]

Banerjee S and Lavie A. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments//Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Ann Arbor, USA: Association for Computational Linguistics: 65-75

Borth D, Ji R R, Chen T, Breuel T and Chang S F. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs//Proceedings of the 21st ACM International Conference on Multimedia. Barcelona, Spain: ACM: 223-232 [DOI: 10.1145/2502081.2502282]

Chatterjee M and Schwing A G. 2018. Diverse and coherent paragraph generation from images//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 747-763 [DOI: 10.1007/978-3-030-01216-8_45]

Che W B, Fan X P, Xiong R Q and Zhao D B. 2018. Paragraph generation network with visual relationship detection//Proceedings of the 26th ACM International Conference on Multimedia. Seoul, Korea (South): ACM: 1435-1443 [DOI: 10.1145/3240508.3240695]

Chen L, Zhang H W, Xiao J, Nie L Q, Shao J, Liu W and Chua T S. 2017. SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6298-6306 [DOI: 10.1109/CVPR.2017.667]

Chen S Z and Jin Q. 2016. Multi-modal conditional attention fusion for dimensional emotion prediction//Proceedings of the 24th ACM International Conference on Multimedia. Amsterdam, the Netherlands: ACM: 571-575 [DOI: 10.1145/2964284.2967286]

Chen T L, Zhang Z P, You Q Z, Fang C, Wang Z W, Jin H L and Luo J B. 2018. "Factual" or "Emotional": stylized image captioning with adaptive learning and attention//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 527-543 [DOI: 10.1007/978-3-030-01249-6_32]

Dai B, Fidler S, Urtasun R and Lin D H. 2017a. Towards diverse and natural image descriptions via a conditional GAN//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2989-2998 [DOI: 10.1109/ICCV.2017.323]

Dai B and Lin D H. 2017. Contrastive learning for image captioning//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: NIPS: 898-907[DOI: 10.5555/3294771.3294857]

Dai B, Zhang Y Q and Lin D H. 2017b. Detecting visual relationships with deep relational networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 3298-3308 [DOI: 10.1109/CVPR.2017.352]

Dalal N and Triggs B. 2005. Histograms of oriented gradients for human detection//Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA: IEEE: 886-893 [DOI: 10.1109/CVPR.2005.177]

Devlin J, Chang M W, Lee K and Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2019-05-24]. https://arxiv.org/pdf/1810.04805.pdf

Donahue J, Hendricks L A, Rohrbach M, Venugopalan S, Guadarrama S, Saenko K and Darrell T. 2017. Long-term recurrent convolutional networks for visual recognition and description. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(4): 677-691 [DOI: 10.1109/TPAMI.2016.2599174]

Dong X Y, Zhu L C, Zhang D, Yang Y and Wu F. 2018. Fast parameter adaptation for few-shot image captioning and visual question answering//Proceedings of the 26th ACM International Conference on Multimedia. Seoul, Korea (South): ACM: 54-62 [DOI: 10.1145/3240508.3240527]

Fang F, Li Q Y, Wang H L and Tang P J. 2018. Refining attention: a sequential attention model for image captioning//Proceedings of 2018 IEEE International Conference on Multimedia and Expo. San Diego, USA: IEEE: 1-6 [DOI: 10.1109/ICME.2018.8486437]

Fang H, Gupta S, Iandola F, Srivastava R K, Deng L, Dollár P, Gao J F, He X D, Mitchell M, Platt J C, Zitnick C L and Zweig G. 2015. From captions to visual concepts and back//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 1473-1482 [DOI: 10.1109/CVPR.2015.7298754]

Farhadi A, Hejrati M, Sadeghi M A, Young P, Rashtchian C, Hockenmaier J and Forsyth D. 2010. Every picture tells a story: generating sentences from images//Proceedings of the 11th European Conference on Computer Vision. Crete, Greece: Springer: 15-29 [DOI: 10.1007/978-3-642-15561-1_2]

Feng Y, Ma L, Liu W and Luo J B. 2019. Unsupervised image captioning//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4120-

4129 [DOI: 10.1109/CVPR.2019.00425]

Fu K, Jin J Q, Cui R P, Sha F and Zhang C S. 2017. Aligning where to see and what to tell: image captioning with region-based attention and scene-specific contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(12): 2321-2334 [DOI: 10.1109/TPAMI.2016.2642953]

Gan C, Gan Z, He X D, Gao J F and Deng L. 2017a. StyleNet: generating attractive visual captions with styles//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 955-964 [DOI: 10.1109/CVPR.2017.108]

Gan Z, Gan C, He X D, Pu Y C, Tran K, Gao J F, Carin L and Deng L. 2017b. Semantic compositional networks for visual captioning//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1141-1150 [DOI: 10.1109/CVPR.2017.127]

Girshick R. 2015. Fast R-CNN//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 1440-1448 [DOI: 10.1109/ICCV.2015.169]

Girshick R, Donahue J, Darrell T and Malik J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 580-587 [DOI: 10.1109/CVPR.2014.81]

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2014. Generative adversarial nets//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT: 2672-2680 [DOI: 10.5555/2969033.2969125]

Gu J X, Cai J F, Wang G and Chen T. 2018. Stack-captioning: coarse-to-fine learning for image captioning//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI: 6837-6844

Guo L T, Liu J, Yao P, Li J W and Lu H Q. 2019. MSCap: multi-style image captioning with unpaired stylized text//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4199-4208 [DOI: 10.1109/CVPR.2019.00433]

Han H, Jain A K, Wang F, Shan S G and Chen X L. 2018. Heterogeneous face attribute estimation: a deep multi-task learning approach. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(11): 2597-2609 [DOI: 10.1109/TPAMI.2017.2738004]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Hinton G E and Salakhutdinov R R. 2006. Reducing the dimensionality of data with neural networks. Science, 313(5786): 504-507 [DOI: 10.1126/science.1127647]

Hodosh M, Young P and Hockenmaier J. 2013. Framing image description as a ranking task: data, models and evaluation metrics. Journal of Artificial Intelligence Research, 47(1): 853-899 [DOI: 10.5555/2566972.2566993]

Huang K T H, Ferraro F, Mostafazadeh N, Misra I, Agrawal A, Devlin J, Girshick R, He X D, Kohli P, Batra D, Zitnick C L, Parikh D, Vanderwende L, Galley M and Mitchell M. 2016. Visual storytelling//Proceedings of 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, USA: Association for Computational Linguistics: 1233-1239 [DOI: 10.18653/v1/N16-1147]

Jia X, Gavves E, Fernando B and Tuytelaars T. 2015. Guiding the long-short term memory model for image caption generation//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 2407-2415 [DOI: 10.1109/ICCV.2015.277]

Jiang W H, Ma L, Jiang Y G, Liu W and Zhang T. 2018. Recurrent fusion network for image captioning//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 510-526 [DOI: 10.1007/978-3-030-01216-8_31]

Johnson J, Karpathy A and Li F F. 2016. DenseCap: fully convolutional localization networks for dense captioning//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 4565-4574 [DOI: 10.1109/CVPR.2016.494]

Karayil T, Irfan A, Raue F, Hees J and Dengel A. 2019. Conditional GANs for image captioning with sentiments//Proceedings of the 28th International Conference on Artificial Neural Networks. Munich, Germany: Springer: 300-312 [DOI: 10.1007/978-3-030-30490-4_25]

Karpathy A and Li F F. 2015. Deep visual-semantic alignments for generating image descriptions//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 3128-3137 [DOI: 10.1109/CVPR.2015.7298932]

Kim D J, Choi J, Oh T H and Kweon S O. 2019. Dense relational captioning: Triple-stream networks for relationship-based captioning//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 6264-6273 [DOI: 10.1109/CVPR.2019.00643]

Kiros R, Salakhutdinov R and Zemel R. 2014. Multimodal neural language models//Proceedings of the 31st International Conference on Machine Learning. Beijing, China: ICML: 595-603 [DOI: 10.5555/3044805.3044959]

Klein D and Manning C D. 2003. Accurate unlexicalized parsing//Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. Sapporo, Japan: ACL: 423-430 [DOI: 10.3115/1075096.1075150]

Krause J, Johnson J, Krishna R and Li F F. 2017. A hierarchical approach for generating descriptive image paragraphs//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 3337-3345 [DOI: 10.1109/CVPR.

2017.356］

Krishna R, Zhu Y K, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li L J, Shamma D A, Bernstein M S and Li F F. 2017. Visual genome: connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision, 123(1): 32-73［DOI: 10.1007/s11263-016-0981-7］

Krizhevsky A, Sutskever I and Hinton G E. 2012. ImageNet classification with deep convolutional neural networks//Proceedings of the 25th International Conference on Neural Information Processing Systems. Red Hook, USA: 1097-1105

Kulkarni G, Premraj V, Ordonez V, Dhar S, Li S M, Choi Y, Berg A C and Berg T L. 2013. BabyTalk: understanding and generating simple image descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(12): 2891-2903［DOI: 10.1109/TPAMI.2012.162］

Kuznetsova P, Ordonez V, Berg A, Berg T and Choi Y. 2013. Generalizing image captions for image-text parallel corpus//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria: Association for Computational Linguistics: 790-796

Kuznetsova P, Ordonez V, Berg A C, Berg T L and Choi Y. 2012. Collective generation of natural image descriptions//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju Island, Korea(South): ACL: 359-368［DOI: 10.5555/2390524.2390575］

Kuznetsova P, Ordonez V, Berg T L and Choi Y. 2014. TREETALK: composition and compression of trees for image descriptions. Transactions of the Association for Computational Linguistics, 2: 351-362［DOI: 10.1162/tacl_a_00188］

LeCun Y, Bottou L, Bengio Y and Haffner P. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11): 2278-2324［DOI: 10.1109/5.726791］

Li X Y, Jiang S Q and Han J G. 2019. Learning object context for dense captioning//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Hawaii, USA: AAAI: 8650-8657

Li Y, Cheng H H, Liang X Y, Guo Q and Qian Y H. 2019. CNN image caption generation. Journal of Xidian University, 46(2): 152-157(李勇, 成红红, 梁新彦, 郭倩, 钱宇华. 2019. CNN 图像标题生成. 西安电子科技大学学报, 46(2): 152-157)［DOI: 10.19665/j.issn1001-2400.2019.02.025

Liang X D, Hu Z T, Zhang H, Gan C and Xing E P. 2017. Recurrent topic-transition GAN for visual paragraph generation//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 3382-3391［DOI: 10.1109/ICCV.2017.364］

Lin C Y and Och F J. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics//Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Barcelona, Spain: ACL: 605-612［DOI: 10.3115/1218955.1219032］

Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick C L. 2014. Microsoft COCO: common objects in context//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer: 740-755［DOI: 10.1007/978-3-319-10602-1_48］

Liu S Q, Zhu Z H, Ye N, Guadarrama S and Murphy K. 2017a. Improved image captioning via policy gradient optimization of SPIDEr//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 873-881［DOI: 10.1109/ICCV.2017.100］

Liu Y, Fu J L, Mei T and Chen C W. 2017b. Let your photos talk: generating narrative paragraph for photo stream via bidirectional attention recurrent neural networks//Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, USA: AAAI: 1445-1452［DOI: 10.5555/3298239.3298450］

Lowe D G. 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2): 91-110［DOI: 10.1023/b:visi.0000029664.99615.94］

Lu J S, Xiong C M, Parikh D and Socher R. 2017. Knowing when to look: adaptive attention via a visual sentinel for image captioning//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 3242-3250［DOI: 10.1109/CVPR.2017.345］

Luong M T, Le Q V, Sutskever I, Vinyals O and Kaiser L. 2016. Multi-task sequence to sequence learning//Proceedings of the 4th International Conference on Learning Representations. San Juan, Argentina: ICLR: 1-10

Machajdik J and Hanbury A. 2010. Affective image classification using features inspired by psychology and art theory//Proceedings of the 18th ACM International Conference on Multimedia. Firenze, Italy: ACM: 83-92［DOI: 10.1145/1873951.1873965］

Mao J H, Xu W, Yang Y, Wang J, Huang Z H and Yuille A L. 2015. Deep captioning with multimodal recurrent neural networks (m-RNN)//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: ICLR: 1-17［https://arxiv.org/abs/1412.6632］

Mason R and Charniak E. 2014. Nonparametric method for data-driven image captioning//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland, USA: ACL: 592-598［DOI: 10.3115/v1/P14-2097］

Mathews A, Xie L X and He X M. 2016. SentiCap: generating image descriptions with sentiments//Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix, USA: AAAI: 3574-3580［DOI: 10.5555/3016387.3016406］

Mitchell M, Han X F, Dodge J, Mensch A, Goyal A, Berg A, Yamaguchi K, Berg T, Stratos K and Daumé H. 2012. Midge: generating image descriptions from computer vision detections//Proceedings of the 13th Conference of the European Chapter of the Association for

Computational Linguistics. Avignon, France: ACL: 747-756 [DOI: 10.5555/2380816.2380907]

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

Ojala T, Pietikäinen M and Harwood D. 1996. A comparative study of texture measures with classification based on featured distributions. Pattern Recognition, 29(1): 51-59 [DOI: 10.1016/0031-3203(95)00067-4]

Papineni K, Roukos S, Ward T and Zhu W J. 2002. BLEU: a method for automatic evaluation of machine translation//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, USA: ACL: 311-318 [DOI: 10.3115/1073083.1073135]

Park C C, Kim B and Kim G. 2017. Attend to you: personalized image captioning with context sequence memory networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6432-6440 [DOI: 10.1109/CVPR.2017.681]

Pedersoli M, Lucas T, Schmid C and Verbeek J. 2017. Areas of attention for image captioning//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 1251-1259 [DOI: 10.1109/ICCV.2017.140]

Ren S Q, He K M, Girshick R and Sun J. 2017a. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149 [DOI: 10.1109/TPAMI.2016.2577031]

Ren Z, Wang X Y, Zhang N, Lyu X T and Li L J. 2017b. Deep reinforcement learning-based image captioning with embedding reward//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1151-1159 [DOI: 10.1109/CVPR.2017.128]

Rennie S J, Marcheret E, Mroueh Y, Ross J and Goel V. 2017. Self-critical sequence training for image captioning//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1179-1195 [DOI: 10.1109/CVPR.2017.131]

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S A, Huang Z H, Karpathy A, Khosla A, Bernstein M, Berg A C and Li F F. 2015. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 115(3): 211-252 [DOI: 10.1007/s11263-015-0816-y]

Shi J H, Li Y L and Wang S J. 2019. Cascade attention: multiple feature based learning for image captioning//Proceedings of 2019 IEEE International Conference on Image Processing. Taipei, China: IEEE: 1970-1974 [DOI: 10.1109/ICIP.2019.8803149]

Shin A, Ushiku Y and Harada T. 2016. Image captioning with sentiment terms via weakly-supervised sentiment dataset//Proceedings of British Machine Vision Conference. York, UK: BMVA Press: 53.1-53.12 [DOI: 10.5244/C.30.53]

Shuster K, Humeau S, Hu H X, Bordes A and Weston J. 2019. Engaging image captioning via personality//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 12508-12518 [DOI: 10.1109/CVPR.2019.01280]

Simonyan K and Zisserman A. 2014. Very deep convolutional networks for large-scale image recognition//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: ICLR: 1-14

Sun Y, Wang X G and Tang X O. 2013. Deep convolutional network cascade for facial point detection//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA: IEEE: 3476-3483 [DOI: 10.1109/CVPR.2013.446]

Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A. 2015. Going deeper with convolutions//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 1-9 [DOI: 10.1109/CVPR.2015.7298594]

Tang P J, Tan Y L and Li J Z. 2017. Image description based on the fusion of scene and object category prior knowledge. Journal of Image and Graphics, 22(9): 1251-1260 (汤鹏杰, 谭云兰, 李金忠. 2017. 融合图像场景及物体先验知识的图像描述生成模型. 中国图象图形学报, 22(9): 1251-1260) [DOI: 10.11834/jig.170052]

Tang P J, Wang H L and Kwong S. 2017. G-MS2F: GoogLeNet based multi-stage feature fusion of deep CNN for scene recognition. Neurocomputing, 225: 188-197 [DOI: 10.1016/j.neucom.2016.11.023]

Tang P J, Wang H L and Kwong S. 2018. Deep sequential fusion LSTM network for image description. Neurocomputing, 312: 154-164 [DOI: 10.1016/j.neucom.2018.05.086]

Tang P J, Wang H L and Xu K S. 2018. Multi-objective layer-wise optimization and multi-level probability fusion for image description generation using LSTM. Acta Automatica Sinica, 44(7): 1237-1249 (汤鹏杰, 王瀚漓, 许恺晟. 2018. LSTM 逐层多目标优化及多层概率融合的图像描述. 自动化学报, 44(7): 1237-1249) [DOI: 10.16383/j.aas.2017.c160733]

Tran K, He X D, Zhang L and Sun J. 2016. Rich image captioning in the wild//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Las Vegas, USA: IEEE: 434-441 [DOI: 10.1109/CVPRW.2016.61]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: NIPS: 5998-6008 [DOI: 10.5555/3295222.3295349]

Vedantam R, Zitnick C L and Parikh D. 2015. CIDEr: consensus-based image description evaluation//Proceedings of 2015 IEEE Conference

on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 4566-4575 [DOI: 10.1109/CVPR.2015.7299087]

Vinyals O, Toshev A, Bengio S and Erhan D. 2015. Show and tell: a neural image caption generator//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 3156-3164 [DOI: 10.1109/CVPR.2015.7298935]

Wang J, Fu J L, Tang J H, Li Z C and Mei T. 2018a. Show, reward and tell: automatic generation of narrative paragraph from photo stream by adversarial training//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI: 7396-7403

Wang J, Pan Y W, Yao T, Tang J H and Mei T. 2019. Convolutional auto-encoding of sentence topics for image paragraph generation//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China: IJCAI: 940-946 [DOI: 10.24963/ijcai.2019/132]

Wang L M, Qiao Y and Tang X O. 2015. Action recognition with trajectory-pooled deep-convolutional descriptors//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 4305-4314 [DOI: 10.1109/CVPR.2015.7299059]

Wang L M, Xiong Y J, Wang Z, Qiao Y, Lin D H, Tang X O and van Gool L. 2018b. Temporal segment networks for action recognition in videos. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(11): 2740-2755 [DOI: 10.1109/TPAMI.2018.2868668]

Wu J C, Wang L M, Wang L, Guo J and Wu G S. 2019a. Learning actor relation graphs for group activity recognition//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 9956-9966 [DOI: 10.1109/CVPR.2019.01020]

Wu Q, Shen C H, Liu L Q, Dick A and van den Hengel A. 2016. What value do explicit high level concepts have in vision to language problems? //Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 203-212 [DOI: 10.1109/CVPR.2016.29]

Wu Q, Shen C H, Wang P, Dick A and van den Hengel A. 2018. Image captioning and visual question answering based on attributes and external knowledge. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(6): 1367-1381 [DOI: 10.1109/TPAMI.2017.2708709]

Wu S Z, Kan M N, Shan S G and Chen X L. 2019b. Hierarchical attention for part-aware face detection. International Journal of Computer Vision, 127(6): 560-578 [DOI: 10.1007/s11263-019-01157-5]

Xu K, Le Ba J, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel R and Bengio Y. 2015. Show, attend and tell: neural image caption generation with visual attention//Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille, France: JMLR. org: 2048-2057 [DOI: 10.5555/3045118.3045336]

Yan S Y, Xie Y, Wu F Y, Smith J S, Lu W J and Zhang B L. 2020. Image captioning via hierarchical attention mechanism and policy gradient optimization. Signal Processing, 167: #107329 [DOI: 10.1016/j.sigpro.2019.107329]

Yang J W, Lu J S, Lee S, Batra D and Parikh D. 2018. Graph R-CNN for scene graph generation//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 690-706 [DOI: 10.1007/978-3-030-01246-5_41]

Yang L J, Tang K, Yang J C and Li L J. 2017. Dense captioning with joint inference and visual context//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1978-1987 [DOI: 10.1109/CVPR.2017.214]

Yang X, Tang K H, Zhang H W and Cai J F. 2019. Auto-encoding scene graphs for image captioning//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 10677-10686 [DOI: 10.1109/CVPR.2019.01094]

Yang Y Z, Teo C L, Daumé H and Aloimonos Y. 2011. Corpus-guided sentence generation of natural images//Proceedings of Conference on Empirical Methods in Natural Language Processing. Edinburgh, UK: ACL: 444-454 [DOI: 10.5555/2145432.2145484]

Yao B Z, Yang X, Lin L, Lee M W and Zhu S C. 2010. I2T: image parsing to text description. Proceedings of the IEEE, 98(8): 1485-1508 [DOI: 10.1109/jproc.2010.2050411]

Yao T, Pan Y W, Li Y H and Mei T. 2018. Exploring visual relationship for image captioning//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 711-727 [DOI: 10.1007/978-3-030-01264-9_42]

Yao T, Pan Y W, Li Y H, Qiu Z F and Mei T. 2017. Boosting image captioning with attributes. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 4904-4912 [DOI: 10.1109/ICCV.2017.524]

Yin G J, Sheng L, Liu B, Yu N H, Wang X G and Shao J. 2019. Context and attribute grounded dense captioning//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 6234-6243 [DOI: 10.1109/CVPR.2019.00640]

You Q Z, Cao L L, Jin H L and Luo J B. 2016a. Robust visual-textual sentiment analysis: when attention meets tree-structured recursive neural networks//Proceedings of the 24th ACM International Conference on Multimedia. Amsterdam, the Netherlands: ACM: 1008-1017 [DOI: 10.1145/2964284.2964288]

You Q Z, Jin H L, Wang Z W, Fang C and Luo J B. 2016b. Image captioning with semantic attention//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 4651-4659 [DOI: 10.1109/CVPR.2016.503]

Young P, Lai A, Hodosh M and Hockenmaier J. 2014. From image descriptions to visual denotations: new similarity metrics for semantic

inference over event descriptions. Transactions of the Association for Computational Linguistics, 2: 67-78 [DOI: 10.1162/tacl_a_00166]

Zhang M X, Yang Y, Zhang H W, Ji Y L, Shen H T and Chua T S. 2019a. More is better: precise and detailed image captioning using online positive recall and missing concepts mining. IEEE Transactions on Image Processing, 28(1): 32-44 [DOI: 10.1109/TIP.2018.2855415]

Zhang Z J, Wu Q, Wang Y and Chen F. 2019b. High-quality image captioning with fine-grained and semantic-guided visual attention. IEEE Transactions on Multimedia, 21(7): 1681-1693 [DOI: 10.1109/TMM.2018.2888822]

Zhao J M, Chen S Z and Jin Q. 2018. Multimodal dimensional and continuous emotion recognition in dyadic video interactions//Proceedings of the 19th Pacific-Rim Conference on Multimedia. Hefei, China: Springer: 301-312 [DOI: 10.1007/978-3-030-00776-8_28]

Zhao W T, Wu X X and Zhang X X. 2020. MemCap: memorizing style knowledge for image captioning//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI: 12984-12992

Zhou B L, Lapedriza A, Xiao J X, Torralba A and Oliva A. 2014. Learning deep features for scene recognition using places database//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: ACM: 487-495 [DOI: 10.5555/2968826.2968881]

作者简介

谭云兰,1972 年生,女,副教授,主要研究方向为深度学习、图像美学。
E-mail: tanyunlan@163.com

汤鹏杰,通信作者,男,高级实验师,主要研究方向为多媒体智能计算。
E-mail:tangpengjie@jgsu.edu.cn

张丽,女,高级经济师,主要研究方向为机器学习与图像处理。E-mail:hellen.zl@163.com
罗玉盘,女,实验师,主要研究方向为大数据分析处理。
E-mail:lyp@jgsu.edu.cn