软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2019,30(4):867−883 [doi: 10.13328/j.cnki.jos.005665]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

视觉场景描述及其效果评价*

马 苗[1,2], 王伯龙[2], 吴 琦[3], 武 杰[2], 郭 敏[2]

[1](现代教学技术教育部重点实验室(陕西师范大学),陕西 西安 710062)

[2](陕西师范大学 计算机科学学院,陕西 西安 710119)

[3](School of Computer Science, The University of Adelaide, Adelaide SA5005, Australia)

通讯作者: 马苗, E-mail: mmthp@snnu.edu.cn

摘 要: 作为计算机视觉、多媒体、人工智能和自然语言处理等领域的交叉性研究课题,视觉场景描述的研究内容是自动生成一个或多个语句用于描述图像或视频中呈现的视觉场景信息.视觉场景中内容的丰富性和自然语言表达的多样性使得视觉场景描述成为一项充满挑战的任务,综述了现有视觉场景描述方法及其效果评价.首先,论述了视觉场景描述的定义、研究任务及方法分类,简要分析了视觉场景描述与多模态检索、跨模态学习、场景分类、视觉关系检测等相关技术的关系;然后分类讨论视觉场景描述的主要方法、模型及研究进展,归纳日渐增多的基准数据集;接下来,梳理客观评价视觉场景描述效果的主要指标和视觉场景描述技术面临的问题与挑战,最后讨论未来的应用前景.

关键词: 深度学习;图像描述;视频描述;基准数据集;性能评价

中图法分类号: TP37

中文引用格式: 马苗,王伯龙,吴琦,武杰,郭敏.视觉场景描述及其效果评价.软件学报,2019,30(4):867−883. http://www.jos.org.cn/1000-9825/5665.htm

英文引用格式: Ma M, Wang BL, Wu Q, Wu J, Guo M. Visual scene description and its performance evaluation. Ruan Jian Xue Bao/Journal of Software, 2019,30(4):867−883 (in Chinese). http://www.jos.org.cn/1000-9825/5665.htm

Visual Scene Description and Its Performance Evaluation

MA Miao[1,2], WANG Bo-Long[2], WU Qi[3], WU Jie[2], GUO Min[2]

[1](Key Laboratory of Modern Teaching Technology of Ministry of Education (Shaanxi Normal University), Xi'an 710062, China)

[2](School of Computer Science, Shaanxi Normal University, Xi'an 710119, China)

[3](School of Computer Science, The University of Adelaide, Adelaide SA5005, Australia)

Abstract: As a cross-domain research topic related to Computer Vision, Multimedia, Artificial Intelligence and Natural Language Processing, the task of visual scene description is to produce automatically one or more sentences to describe the content of visual scene from an image or a video snippet. The richness of the content in the visual scene and the diversity of the expression of natural language make visual scene description a challenging task. This paper gives a review about the generation methods and performance evaluation on the recently developed visual scene description methods. Specifically, the research object and main tasks of visual scene description are firstly defined; the relationships between visual scene description and multi-modal retrieval, cross-modal learning, scene classification, visual relationship detection and other related technologies are discussed sequentially. And then, main methods and research progress of

* 基金项目: 国家自然科学基金(61877038, 61801282, 61601274); 陕西省自然科学基金(2018JM6068); 中央高校基本科研业务经费(GK 201703054, GK201703058)

Foundation item: National Natural Science Foundation of China (61877038, 61801282, 61601274); Natural Science Foundation of Shaanxi Province, China (2018JM6068); Fundamental Research Funds for the Central Universities of Shaanxi Normal University (GK201703054, GK201703058)

本文由"多媒体数据的知识关联与理解专题"特约编辑蒋树强研究员、刘青山教授、孙立峰教授、李波教授推荐.

收稿时间: 2018-04-15; 修改时间: 2018-06-13; 采用时间: 2018-09-30

visual scene description are summarized in three categories, while the increasing benchmark datasets are discussed. Besides, some widely-used evaluation metrics and the corresponding challenges on the visual scene description are discussed. Finally, some potential applications in future are suggested.

Key words: deep learning; image captioning; video captioning; benchmark dataset; performance evaluation

视觉场景描述技术通过对输入图像或视频的内容分析,自动生成一个语句或若干语句的形式对视觉场景中的内容进行描述,属于计算机视觉、多媒体、人工智能和自然语言处理等领域的交叉性研究课题.视觉场景描述问题可归结为视觉语义理解、多媒体语义学习、场景理解等领域中的子问题,其历史可追溯到多模态检索、跨模态学习等问题的研究.

近年来,得益于深度学习相关模型、方法的突破性进展和大样本数据集的出现,尤其是随着 MS COCO、Flickr 等基准数据集的出现和深度学习框架下卷积神经网络(convolution neural network,简称 CNN)、循环神经网络(recurrent neural network,简称 RNN)、长短时记忆网络(long short-term memory,简称 LSTM)等深度网络模型研究的日益成熟,视觉场景描述技术再度掀起研究高潮,并正在变为现实.然而,由于视觉场景中呈现内容的丰富性和自然语言表达的形式多样性,使得视觉场景描述成为一项复杂而富于变化的挑战性任务.

视觉场景描述问题在业界和学术界均引起了高度重视,国内外相关研究机构包括 Google 实验室、Baidu 研究院、微软研究院、中国科学院、斯坦福大学、伯克利大学、加利福尼亚大学等.在国际知名的学术论文图书馆 ACM、IEEE、Elsevier、Springer 和国内外学术论文搜索引擎 Google Scholar 和百度学术中,以“image description、video description、image captioning 或 video captioning”等为关键字,检索论文,其结果表明:近年来,与视觉场景描述有关的学术论文发表数量一直呈增长趋势,反映最新成果的一系列论文在许多知名国际会议中如雨后春笋般产生.例如,“计算机视觉与模式识别国际会议(IEEE Conf. on Computer Vision and Pattern Recognition,简称 CVPR)”[1−22]、“计算机视觉国际会议(IEEE Int’l Conf. on Computer Vision,简称 ICCV)”[23−33]、“欧洲计算机视觉会议(European Conf. on Computer Vision,简称 ECCV)”[34−40]、“神经信息处理系统国际会议(Int’l Conf. on Neural Information Processing Systems,简称 NIPS)”[41−49]和“自然语言处理国际会议(Int’l Joint Conf. on Natural Language Processing,简称 NLP)”[50,51]等.

图 1 所示为近年来关于“视觉场景描述”在计算机视觉领域中三大会议上论文发表数量的统计图,直观地展现了该技术的研究趋势.这些研究成果不仅推动了计算机视觉、自然语言处理等相关学科的融合发展,而且展现了其在视觉信息相关的跨模态检索、智能监控、海量数据压缩、帮助视觉障碍人士感知与理解周围环境等众多领域的潜在应用.

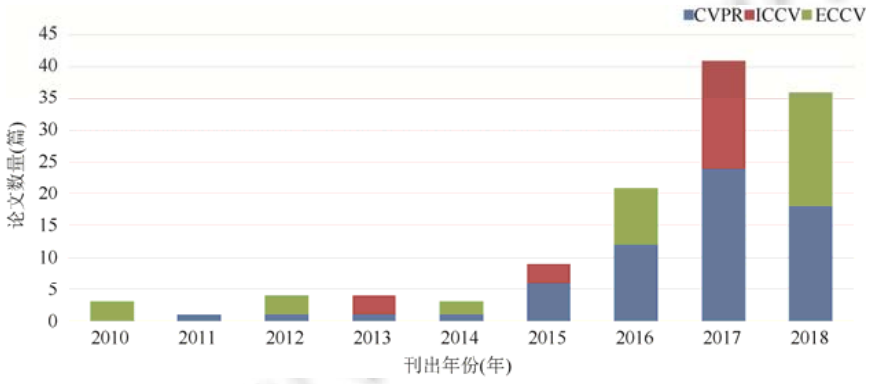


Fig.1 Papers on visual scene description published in the proceedings of three top conferences

图 1 三大顶级会议关于“视觉场景描述”论文的发表情况

本文综述视觉场景描述及其效果评价的研究现状和进展.具体来说,第 1 节论述视觉场景描述的定义、研究任务,简要分析视觉场景描述与跨模态学习、场景理解等相关技术的关系.第 2 节分类总结视觉场景描述的

主要方法、模型及研究成果.第 3 节整理可用于图像描述和视频描述研究与竞技的基准数据集.第 4 节讨论客观评价视觉场景描述效果的主要指标、方法和存在的问题.最后,第 5 节展望视觉场景描述的应用前景.

1 视觉场景描述

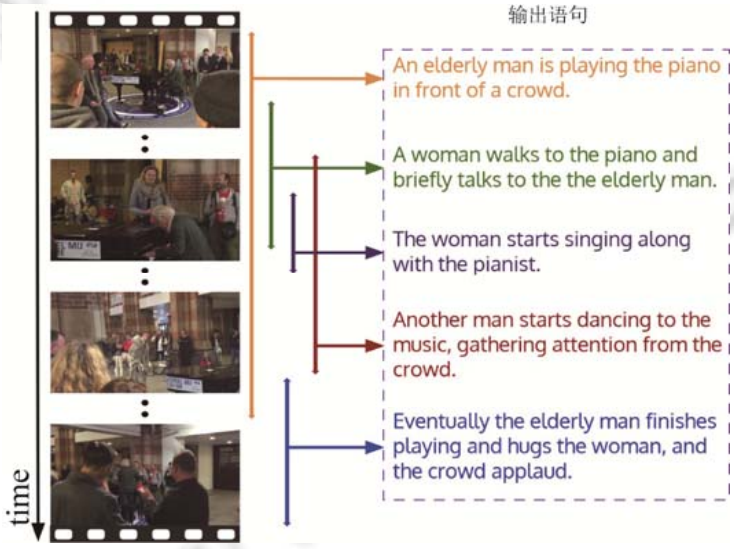
1.1 定义与研究内容

视觉场景描述是指用计算机视觉技术模拟人眼观察到一幅静态图像或观看了一段视频片段后,用自然语言的形式描述观察到的视觉场景内容的方法与技术.由于视觉场景主要源于图像和视频,故视觉场景描述的研究主要针对图像和视频两类输入信息展开.前者用自然语言形式的文本语句描述图像的场景内容,称为图像字幕(image captioning)或图像描述(image description);后者用自然语言形式的文本语句描述视频片段提供的场景内容,称为视频字幕(video captioning)或视频描述(video description).

视觉场景描述的研究任务是自动生成一个或多个句子来描述输入图像或视频中呈现的视觉场景内容,最终目标是用自然语言准确、快速、详细地重述人眼可以观察到的场景,内容涉及场景中存在的目标检测、跟踪(如所在区域、目标属性、目标状态)及各目标或相应事件之间关系的生成与表达等.

图 2 所示的 3 个例子给出了通过视觉场景描述技术自动生成自然语言形式描述 1 幅图像和 2 段视频片段内容的语句.

(a) 图像描述[18] (b) 视频描述[52]



(c) 密集视频描述[8]

Fig.2 Examples of visual scene description

图 2 视觉场景描述的一组例子

获得理想视觉场景描述效果的前提是计算机具有和人类类似的视觉感知能力,能够对静态或动态的场景进行感知、分析和理解,并能得出符合人类习惯的语义描述.因此,从这个角度来看,视觉场景描述是场景语义分

析和视觉场景理解任务的重要组成部分,也是对场景语义理解和分析结果的进一步呈现方式之一.

1.2 相关技术

鉴于视觉场景描述技术的多学科交叉性质,下面我们分别简要论述与之密切相关的多模态检索、跨模态学习、场景理解、场景分类、场景解析、视觉关系检测、场景图生成、视觉问答、指示表达生成等技术.

(1) 多模态检索、跨模态检索和跨模态学习

模态是指数据的存在形式.现实世界中,人们可以用文本、音频、图像、视频等不同模态的数据描述同一对象或事件,得到同步数据.因此,计算机也可以利用这些同步数据学习同一对象或事件的视觉、声音或文本等不同模态的特征.

多模态检索(multimodal retrieval):这是指融合不同模态的检索方法和技术.其特点在于,它不对各模态信息间的关系建模.查询和待检索的文档不止包含 1 个模态,但至少有 1 个模态是相同的.显然,对多媒体数据进行多模态检索可有效提高单模态检索的准确度[53].

跨模态检索(cross-modal retrieval):这是指通过寻找不同模态样本之间的关系,实现利用某一种模态样本搜索近似语义的其他模态样本的方法和技术.其特点在于,检索结果的模态和查询的模态不同.例如用图像检索文本、视频或音频,其关键在于对不同模态的关系进行建模,难点在于需要跨越不同模态间的语义鸿沟[54].

跨模态学习(cross-modal learning):这是指通过对已有多模态训练样本的学习,努力学习到无标记数据的单一模态的更好表示的方法和技术.其特点在于,多模态数据仅在特征学习期间可用,在监督训练和测试阶段,只有单一模态数据可用.

视觉场景描述可看作是一种跨模态学习,即通过大样本视觉场景及其对应的文本形式的描述语句的学习,掌握如何用自然语言去描述未标记的场景内容,包括场景中的对象、对象属性或状态,以及对象之间的关系.在此基础上,可以完成跨模态检索、视觉问答等更高级的场景分析及理解任务.

(2) 场景理解、场景识别/分类、场景解析[55−59]

场景理解(scene understanding):这是指以图像及视频为研究对象,分析什么场景(场景分类或场景识别)、场景中有什么目标(目标检测、目标识别、场景解析)、目标之间的相互关系(场景图、视觉关系)以及如何表达场景(场景描述)的方法和技术.该领域中的大规模场景理解挑战赛 LSUN(large-scale scene understanding)主要聚焦于场景分类、显著预测、房间布置估计和字幕生成这 4 类任务.

场景识别(visual place recognition 或 scene recognition):这是指将一幅图像或一段视频片段中的场景标记为不同类别的方法和技术.若事先给出待识别场景的类别标签,则场景识别问题可归结为一个分类问题,即场景分类(scene classification)[55−59].

场景解析(scene parsing):这是指对场景图像进行分割,并进一步解析为与语义类别相关的不同区域的方法和技术.其特点在于,它预测场景中每个像素的类别标签、位置以及形状,提供了对场景的完全理解,是自动驾驶、机器人感知等应用的前提和基础.

显然,场景理解涵盖了场景识别、场景解析与场景描述.场景识别与场景解析的结果可以作为场景描述的基础和前提,而场景描述是场景理解、场景识别和场景解析的一种自然语言形式的表达和呈现.

(3) 视觉关系检测、场景图生成和指示表达生成

视觉关系检测(visual relation detection):这是指将对象置于一个上下文语义环境中,研究如何提取不同对象的位置和对象间的空间逻辑关系等内容的方法和技术.不同于视觉内容与自然语言之间的关系,视觉关系检测研究的是各对象之间交互的直接关系,可以为图像注释、问答系统等应用提供深层语义信息[60].

视觉问答(visual question and answer)[26−29,34,41,61]:这是指让计算机根据输入的图像(视频)和问题,研究如何输出符合人类表达习惯且内容合理的答案的方法和技术.目前,该研究多集中在看图问答方面,相关技术涉及目标识别、行为识别和问题解析等.

场景图生成(scene graphs generation):这是指通过显式建模对象、对象属性和对象之间的关系来捕获视觉场景的详细语义的方法和技术.该技术可以为视觉场景描述和视觉问答等应用提供深层次的语义信息,有助于

发现和利用场景中各对象之间的关系[62].

指示表达生成(referring expression generation)[63,64]:这是指研究如何明确、清晰地描述特定对象的方法和技术.该技术常使用属性来描述特定对象,进而能够在给定的上下文中辅助识别相应对象.理解和生成是与指示表达相关的两个任务:理解任务要求系统选择给定表达所描述的对象;生成任务是为图像内的指定对象生成表达.

从场景内容分析角度,视觉关系检测、视觉场景图和指示表达生成的相关研究致力于场景中存在的对象、关系及属性、状态,因此,其研究结论均可引入到场景描述中来深入发掘场景构成、对象属性与状态等信息,这均有利于提高视觉场景描述的准确度.

2 主要方法和研究进展

如第 1.1 节所述,按照场景载体的不同,视觉场景描述从图像描述和视频描述两个维度展开.根据研究思路的不同,视觉场景描述方法可细分为基于模板的方法(template based approaches)、基于检索的方法(retrieval based approaches)以及目前主流的基于序列学习的方法(sequence learning based approaches).根据生成语句的数目不同,视觉场景描述也可分为基于单一语句的视觉场景描述(用一句话描述场景内容)、基于多语句的视觉场景描述(用一段话去描述场景内容)和基于密集描述的视觉场景描述(以不同区域、不同对象或不同事件为单位,详细地描述场景内容),如图 3 所示.

Fig.3 Categories of visual scene description methods

图 3 视觉场景描述方法的分类

下面以视觉场景描述的原理为主线,分别讨论基于模板、检索以及序列学习的视觉场景描述方法、原理和代表性成果.

2.1 基于模板的场景描述方法

该类方法预先定义生成语句的一些特定语法规则,如将句子分为主语、动词和宾语等组成成分,然后检测给定场景的内容、属性,使用概率图模型将状态对齐到属性,并用预定义的句子模板推导出句子结构.

在图像描述方面的代表性工作中,Yang 等人(2011 年)从 Gigaword 语料库训练的语言模型获得动作的估计以及名词、场景和介词共同定位的概率,然后将其作为隐马尔可夫模型(hidden Markov model,简称 HMM)的参数,模拟句子生成过程[51].Mitchell 等人(2012 年)给出计算机视觉检测中产生图像描述的 Midge 系统,它过滤不可能的属性,并将对象放置到有序的句法结构来生成场景内容的语句描述[65].Krishnamoorthy 等人(2013 年)利用 SVO 语言模型来选择“主语、动作、对象”三元组,并生成语句[66].Kulkarni 等人(2013 年)通过检测图像中的对象和属性及它们的介词关系,使用条件随机场来预测包含这些对象、修饰符和关系的最佳结构[67].Lebret 等人(2015 年)从图像中预测短语,然后将它们与一个简单的语言模型结合起来,生成关于图像内容的场景描述[68].

在视频描述方面的代表性工作中,Kojima 等人(2002 年)引入动作的概念层次来描述人类活动[69].Rohrbach 等人(2013 年)采用条件随机场(conditional random field,简称 CRF)算法模拟对象和视觉输入的活动之间的连接,并生成描述的语义特征[33].Guadarrama 等人(2013 年)定义语义层次以学习不同句子成分之间的语义关系[32].此外,Xu 等人(2015 年)提出了一个由语义语言模型、深度视频模型和联合嵌入模型组成的统一框架,来学习视

频和自然语句之间的关联[70].

显然,基于模板的场景描述方法总是能够在预定义的语句模板中直接生成具有检测关键字且语法正确的句子,其不足在于,该类方法高度依赖于预定义的语句模板,生成语句受到固定句法结构的限制,句子描述的内容和形式失去了新颖性和灵活性.

2.2 基于检索的场景描述方法

该类方法的主要思路是通过在数据库中搜索视觉上与输入图像相似的图像,并从检索到的图像标题中利用最近邻法找到最佳描述语句.因此,该类方法本质上是通过从数据库中的句子池中选择语义最相似的句子来生成输出图像的视觉场景描述.

该类方法主要出现在图像描述应用中.Farhadi 等人(2010 年)使用近邻法则选出候选的图像描述语句,将这些语句和对应图像映射到 Meaning Space,并用 Tree-F1 法则进行匹配,得到 5 个最佳描述语句[40].Ordonez 等人(2011 年)提出 Web 图像字幕生成方法,该方法依赖于从互联网收集的大量图像数据,使用全局检索或结合内容估计检索这两种策略产生新的图像标题[49].Kuznetsova 等人(2014 年)提出基于树结构的语句生成方法,其主要思想是从现有的图像描述中收集表达短语,然后选择性地组合所提取的片段来产生新的描述语句[71].Hodosh 等人(2015 年)提出基于 KCCA 的基准系统来进行图像描述和搜索,通过构建序列核及能够捕获语义相似性的核来建立图像与文本间的联合空间,进而描述图像内容[72].Devlin 等人(2015 年)利用 CNN 获得图像的候选词袋,然后用 k 邻近检索模型获得该图像的共识描述,在 COCO 基准数据集上性能优良[73].

易知,该类方法产生的视觉场景描述语句与人工标注的描述语句在表达方式和风格上较为一致,不足在于生成效果受检索数据库中句子池里人工标注的样本数量、样本描述精细粒度以及输出图像与检索图像的相似程度的约束和影响.

2.3 基于序列学习的场景描述方法

基于序列学习的场景描述方法是深度网络模型获得突破性进展以来主流的视觉场景描述方法.“编码器-解码器(encoder-decoder)”框架下的“CNN(或 3D CNN)+RNN”和“CNN(或 3D CNN)+LSTM”是该类方法的常见组合.其中,RNN 在传统神经网络中引入时序概念,将上一时刻的输出作为下一时刻的输入重新进入到网络,可分为单向 RNN、Bi-RNN 和 m-RNN;LSTM 模型可视为 RNN 的改进版本,又可细分为单向 LSTM 模型、双向 LSTM 模型、深层结构的双向 LSTM 模型以及 GRU 模型等[74−76].该类方法的一般过程如图 4 所示.

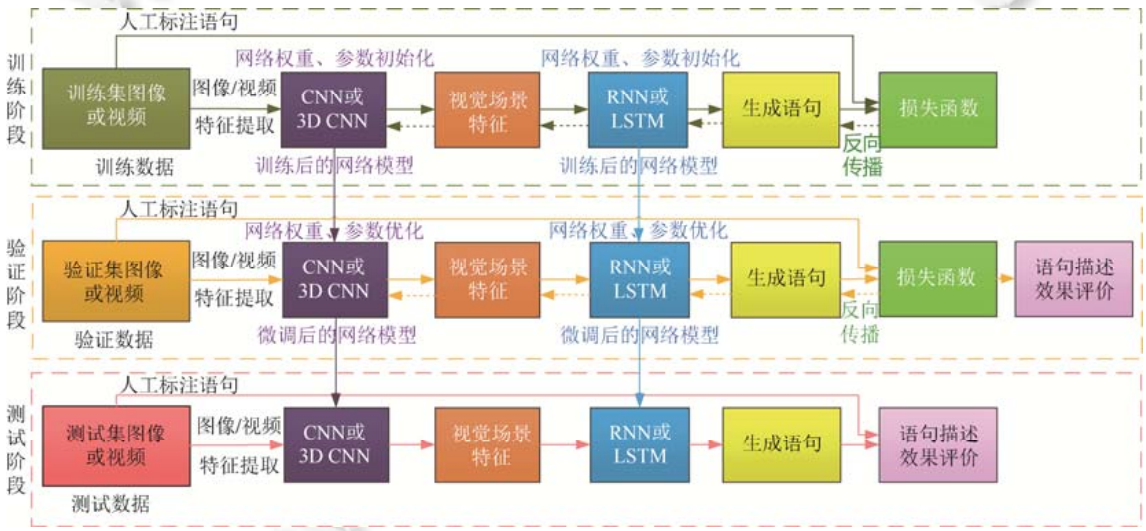


Fig.4 General framework of visual scene description based on sequence learning

图 4 基于序列学习的视觉场景描述方法的一般框架

在图像描述方面的代表性工作包括:(1) 在“CNN+RNN”方法研究中,Vinyals 等人(2015 年)从图像中提取特征并与人工标注语句输入到 RNN 中训练,得到图像内容描述[20].Karpathy 等人(2015 年)以 RCNN(regions with CNN features)为 Encoder 提取图像中各个目标区域,再以 BRNN(bidirectional recurrent neural network)作为 Decoder,并参考上下文来生成语句,最终得到图像中各个区域的描述[17].(2) 在“CNN+LSTM”方法研究中,Donahue 等人(2015 年)利用 LSTM 模型生成内容描述[18].Huang 等人(2016 年)提出具有选择性的多通道 LSTM 模型,以改进局部图像信息与生成文本语句之间的匹配效果,提升图像描述的合理性[9].Ren 等人(2017 年)利用局部预测模型“政策网络”和全局评估模型“价值网络”共同协作生成图像描述[2].以上方法均未考虑场景中的感兴趣区域.(3) 在引入注意机制的方法研究中,Xu 等人(2015 年)将 LSTM 模型与人类视觉中的注意机制相结合,在生成对应的单词时自动聚焦于显著对象[77].Lu 等人(2017 年)引入视觉“哨兵”策略,设计自适应视觉注意模型[3].You 等人(2016 年)使用预生成的语义概念建议来指导描述生成,并学习在不同时刻选择性地关注这些概念[13].Wang 等人(2017 年)则利用基于视觉注意机制的 CNN 提取图像特征,设计了 Skel-LSTM 模型和 Atrr-LSTM 模型,分别用来产生文本语句中的“主、谓、宾”和“定、状、补”[4].(4) 在引入外部知识场景和属性方法的研究中,Wu 等人(2016 年)用高层次的概念(属性),显著改进了 RNN 的图像描述质量[14].该属性进一步被 You 等人(2016 年)用来增强图像描述性能[13].

在视频描述方面的代表性工作包括:(1) 在“3D CNN+RNN”方法研究中,Socher 等人(2014 年)利用 RNN 和 C3D 从视频帧序列中提取出来的三维特征进行时序上的编码并进行建模,最后融合音频特征完成视频分类与单句视频内容描述[78];为了产生更多的句子来详细描述视频场景中的内容,Yu 等人(2016 年)利用分层递归神经网络结合视觉注意机制建模句子间的依赖性,从而生成视频的多句描述[12];(2) 在“CNN+LSTM”方法研究中,Subhashini 等人(2014 年)利用 CNN 提取每个视频帧序列特征图并将它们进行平均池化,再利用 LSTM 模型生成描述语句[52].Torabi 等人(2016 年)用 CNN 提取待描述视频的 C3D 矩阵作为视频信息的三维特征,再通过 LSTM 模型生成描述语句[79];Pan 等人(2017 年)利用基于 COCO 数据集的弱监督多实例学习的语义检测器,分别提取图像和视频的语义属性,将整合后的语义属性送入 LSTM-TSA 网络实现视频场景内容的语义描述[5].同年,Zhang 等人(2017 年)提出任务驱动的动态融合机制来降低视频描述中的模糊度,细化对视频内容的刻画程度[6];Shen 等人(2017 年)利用弱监督的多事例多标记学习方法建立视频区域与词标注的全卷积网络,实现视频内容的多样化密集描述[7].(3) 在引入事件概念的方法研究中,Krishna 等人(2017 年)以事件为单位,通过检测事件、分析事件间的时序关系,建立基于事件驱动的视频描述模型[8].在此基础上,Wang 等人(2018 年)将“只利用过去上下文来进行建议预测”改进为“用双向建议模块编码过去和将来的上下文”,提出双向视觉融合的密集视频描述方法.该方法能够区分和描述时间上高度重叠的事件,进一步提高对视频内容进行密集描述的能力[1].

该类方法的特点在于,利用深度网络模型在视觉内容和文本句子的联合空间中学习概率分布,来生成句法结构灵活的句子,能够提供较为准确的场景描述效果.其优点体现在,通过“CNN+RNN”等深度网络结构自动获取场景内容的特征表达能力,去掉了繁杂的人工特征提取过程,属于端到端的解决问题方式,但是该类方法依赖于大样本基准数据集的支撑,其在应用中的性能取决于实际场景与大量样本场景间的相似性.相似度高的场景内容描述质量高,反之,场景内容描述结果可能与实际情况不符.

3 视觉场景描述的基准数据集

在视觉场景描述的研究中,尤其是 Encoder-Decoder 框架下基于序列学习的方法及模型构建大多属于有监督的机器学习方法,因此离不开人工标注的基准数据集的发展.这些基准数据集不仅提供了大量的图像和视频等资源,而且提供了对数据集中图像、视频对应的人工标注语句.它们一方面供研究人员对所提出模型或方法的正确性与有效性进行检验,另一方面也为不同场景描述方法或模型的性能对比提供了开放的平台.

下面给出人工标注的产生和视觉场景描述的常用基准数据集.

3.1 人工标注的产生

近年来,人工智能技术被引入各类复杂应用,如语音理解、物体识别、环境感知等.然而,这些智能系统的构

建往往需要含有标注的大量数据样本作为训练资源,而提供这些符合分类规则和人类认知规律的标注还不能完全由计算机生成.实际上,绝大多数图像、视频的标注还是人工完成的.

随着机器学习应用的不断普及,人工主导、监督学习、半监督学习和无监督学习的混合训练方法将是未来人工智能系统的主要学习方式.这意味着越来越多的数据需要被正确标记.实际上,针对此任务,亚马逊、苹果、谷歌、微软等均有自己的劳务众包平台或直接使用第三方服务.其中,始于 2005 年的亚马逊劳务众包平台(Amazon mechanical turk,简称 AMT)是最有影响的在线劳务众包平台之一.目前 AMT 注册工作人员累计超过 50 万.这些工作人员被称为 Turker,他们通过互联网可以全天候地完成数据标定任务.例如,在计算机视觉领域产生重要影响的 ImageNet 数据集中大部分标注工作是在 AMT 上由 50 000 名 Turker 历时约 2 年完成.

3.2 视觉场景描述的基准数据集

目前,国际上可用于视觉场景描述研究与竞技的公开基准数据集有 10 余种.其中,图像描述基准数据集包括 Pascal VOC[80]、Flickr 系列[72,81,82]、MS COCO[83]、YFCC100M[84]、Visual Genome[85]和 ICC[86]等,见表 1.

Table 1 Datasets on image captioning

表 1 图像描述的基准数据集

数据集	产生时间(年)	主要内容	主题或来源	备注
Pascal VOC[80]	2005	20 个类别,每类别含 50 幅图像的随机样本与 5 个人工标注语句	含跳、跑等人类动作,头、手、脚等人体部位	有 2005、2007 和 2012 这 3 个版本
Flickr8K[72]、Flickr30[81]	2010	分别包括 8 000 幅和 31 783 幅图像.Flickr8K 对于每个图像提供 5 个人工标注语句	人类活动	–
MS COCO[83]	2014	82 783 幅训练图像,40 504 幅验证图像和 40 775 幅测试图像.包括 91 类目标.每幅图像均有 5 个人工标注语句	日常场景	有 2014、2015 和 2017 这 3 个版本
YFCC100M[84]	2014	9 920 万幅图像、80 万视频;76%、20%和 4%的图像分别具有标题、自动标题和没有标题.每个标题平均 3.08 个单词;32%的图像有描述语句,每个语句平均 22.52 个单词	公共多媒体	69%的图像有标签,平均每幅图像有 7.07 个标签
Flickr30k Entities[82]	2015	244k 共指链,158k 人工标注,共指链连接了同一图像不同标题的同一描述实体	–	276k 人工标注的区域及实体描述
Visual Genome[85]	2017	108 077 幅图像,540 万区域描述.结构化的图像概念与语言连接起来的数据集/知识库	–	170 万视觉问答,380 万对象实例,280 万属性,230 万关系,支持 WordNet SynSets
ICC[86]	2017	210 000 幅训练图像,30 000 幅验证图像,60 000 幅测试图像等分为测试集 1 和测试集 2.包含 200 多种场景和 150 多类动作.每个图像提供 5 个标注语句	日常场景	标注语句为中文描述

与之类似,现有的国际上通用的视频描述基准数据集包括 MSVD[87]、YouCook[22]、TACoS multilevel dataset[88]、YouTube2Text[32]、MPII-MD[89]、M-VAD[90]、MSR-VTT[11]、ActivityNet Captions[8]和 YouCook2[91]等数据集,见表 2.

Table 2 Datasets on video captioning

表 2 视频描述的基准数据集

数据集	产生时间(年)	主要内容	主题或来源	备注
MSVD[87]	2010	2 89 个视频,85 50 个语句,每个视频约 40 个描述语句	YouTube 视频 Microsoft Research 提供	
YouCook[22]	2013	89 种烹饪,200 个视频	烹饪	各烹饪步骤均有起止时间标记和对应语句描述

Table 2 Datasets on video captioning (Continued)
表 2 视频描述的基准数据集(续)

数据集	产生时间(年)	主要内容	主题或来源	备注
TACoS multilevel dataset[88]	2013	185 个视频,用 3 种方式来描述每个视频:最多 15 句的详细描述、3~5 句的简短描述和单一句子描述	烹饪,源于 MPII Cooking2 数据集	有起止时间标记、对应语句描述和行为类别标签
YouTube2Text[32]	2014	80 000 视频/语句描述对,包含目标的活动行为及与相关对象的关系	体育、动物和音乐	词汇表含有 16 000 个词条
MPII-MD[89]	2015	68 000 个视频,每个视频都有 1 个描述语句(电影脚本)	94 部电影(好莱坞)	–
M-VAD[90]	2015	84.6 小时,含 48 986 个视频和 55 904 个描述语句	92 部电影	视频平均时长 6.2s
MSR-VTT[11]	2016	41.2 小时,10k 网络视频和对应的 200k 描述语句	音乐、游戏、运动、新闻、教育等 20 类	每个视频对应 20 个语句描述
ActivityNet Captions[8]	2017	849 小时,2 万视频,10 万事件描述语句(含起止时间标注).每个句子平均 13.48 个单词,描述 36s 的事件	200 种日常活动	包含修剪和未修剪视频
YouCook2[91]	2017	176 小时,2 000 个视频,含起止时间和对应的描述语句	89 种烹饪食谱,每个食谱有 22 个视频	视频平均时长 5.26 分钟

4 视觉场景描述的效果评价

随着视觉场景描述生成方法及模型日渐增多和基准数据集的不断丰富,研究人员希望能够通过设计一些客观指标自动判断视觉场景描述生成的深度网络模型及方法的有效性,由此提出了一些客观的性能评价指标[9,19,92−105].这些指标的本质是对人工标注语句和自动生成语句的相似度比较.

常见的客观评价指标见表 3.

Table 3 Performance evaluation on visual scene description
表 3 视觉场景描述的性能评价

评价指标	产生背景	主要特点
BLEU 系列(2002,2006)[92,93]	机器翻译	*n* 元组精度
ROUGE 系列(2004)[94,95]	文档摘要	*n* 元组精度与召回率
METEOR(2005)[96]	语言学,机器翻译	WordNet 同义词库 Stemmed tokens 相似单词
ATEC(2009)[97]	机器翻译	增加词语权重
CIDEr(2015)[19]	图像描述	TF-IDF 权重的 *n* 元组相似度 平均余弦相似度、精度与召回率
WMD(2015)[98]	文档相似性	word2vec 的 EMD 距离
SPICE(2016)[99].	图像描述	场景图的同义词匹配
GRAO(2017)[100]	图像描述,视频描述	灰色关联分析,综合性能评价
Discriminative evaluation (2018)[101]	图像描述	学习的方法

4.1 基于*n*元组匹配的客观评价

早期的研究工作主要集中在基于 *n* 元组的匹配情况来评价生成语句与人工标注语句之间的相似程度.然而,由于此类方法未考虑语义信息的一致性,有时这些方法的评价结果与人类感知不符.

(1) BLEU 指标系列[92,93].包括 BLEU-1、BLEU-2、BLEU-3、BLEU-4,主要思想是基于人工标注语句与生成语句之间 *n* 个连续字符的严格匹配情况进行评价.它的计算过程是对生成语句与人工标注语句的 *n* 元组进行比较,并计算出匹配片段的个数.这些匹配片段与它们在文字中的位置无关.匹配片段数越多,该指标取值越大,说明生成语句与人工标注语句相似度越高,场景描述效果越好.因该系列指标计算简单,故广泛用于机器翻译的效果评价.不足之处在于,计算过程中人工标注语句的单词会被重复利用,易引起评价结果出现偏差.

(2) ROUGE 指标系列[94].包括 ROUGE-L、ROUGE-N、ROUGE-W 和 ROUGE-S[95].其中,ROUGE-L 用于计

算一个生成语句与一个人工标注语句之间的相似度,主要思想是对比系统生成语句与人工标注语句,通过统计二者之间基本单元的重叠数目来评价生成语句的质量;ROUGE-N 用于计算一个生成语句与多个人工标注语句之间的相似度,当单一生成语句与多个人工标注语句计算评分时,ROUGE-N 最终取值为生成语句与各人工标注语句 ROUGE-L 评分中的最高分.该指标取值越大,说明生成语句与人工标注语句相似度越高.不足在于,其计算过程只是简单地采用人工标注语句与生成语句间的公共子序列长度进行计算,未考虑生成语句与人工标注语句之间的语句关联度.

(3) CIDEr-D 指标[19].主要思想是将每个句子都看作“文档”,将其表示成 TF-IDF 向量的形式来计算每个 *n* 元组的权重,将句子表示成向量形式,每个人工标注语句和待评价语句之间通过 TF-IDF 向量的余弦距离来度量其相似性,在 *n* 元组的计算过程中同时考虑了精度与召回率,提高了以往计算指标在度量共识方面的准确性.当单一生成语句与多个人工标注语句计算评分时,CIDEr-D 最终取值为生成语句与各人工标注语句 CIDEr-D 评分中的最高分.该指标常用于图像描述的语句评价,取值越大,说明生成语句与人工标注语句相似度越高.

(4) GRAO 指标[100].主要思想是先用单一性能指标对源于不同描述生成算法得到的语句给出评分,再对这些评分结果进行带权值的灰色关联分析,实现对各种描述生成算法的性能优劣排序.该评价指标的特点在于把人们主观评价时的先验知识映射为权值,与多个客观评价指标相结合进行综合性能评价.不足在于,其计算结果依赖于各单一指标的取值.

4.2 基于语义信息匹配的指标

基于 *n* 元组匹配的度量指标在“因单词不同而语义相同”或“句子中的 *n* 元组相同但语义不同”两类场景描述语句评价时,结果往往与人类感知不符,难以合理地度量和反映视觉场景内容生成的形式多变的语句与内容的一致性,严重时可能会得到与人类感知相反的结果.为解决此类问题,研究人员提出了基于语义信息匹配的度量指标[6].

(1) WMD 指标[98].主要思想是在计算人工标注语句与生成语句的相似度时,把其中一个语句的多个单词映射到多个隐层向量里,分别计算各单词间的距离,再通过加上单词的权重来计算两个语句间的距离.该指标取值越大,说明生成语句与人工标注语句的相似度越低.

(2) METEOR 指标[96].将“准确匹配的单词”扩展到基于 WordNet 同义词库或“Stemmed Tokens”的“语义相似单词”,计算最佳生成语句与人工标注语句之间的精度与召回率的调和均值.当单一生成语句与多个人工标注语句计算评分时,METEOR 最终取值为生成语句与各人工标注语句 METEOR 评分中的最高分.该指标考虑了人工标注语句与生成语句的单词或词组的前后顺序,但因其依赖语句间 *n* 元组的相似性,无法评估待评价语句的语义相关度.该指标取值越大,说明生成语句与人工标注语句相似度越高.

(3) ATEC 指标[97].将选择的单词及其语序视为句子表达中的两个关键要素,根据多匹配模板和单词信息量化评价选择的单词,通过对单词的位置距离及词序的差异性评价单词的语序,并通过训练的方式来确定两者的最佳权重.该指标取值越大,说明生成语句与人工标注语句相似度越高.

(4) SPICE 指标[99].考虑了同义词现象,并运用 WordNet 模块的 Synset 功能来进行同义词合并与匹配.该指标计算语句间的单词相似度,也参考了语句间的关联度,与人类判断有很好的相关性,其不足在于未参考句子的句法结构,仍依赖 *n* 元组的匹配情况.该指标取值越大,说明生成语句与人工标注语句相似度越高.

(5) SM LSTM 指标[9].主要思想是用全局“视觉-语义”相似度度量图像和句子之间的匹配关系.全局相似性可看作由图像(对象)和语句(词)成对实例之间的多个局部相似性的复合聚集.因此,Huang 等人(2016 年)提出了一个选择性多模态的长短时记忆网络,用来计算图像和句子间的匹配程度.

综上所述,人们提出了很多客观指标或评价方法来判断视觉场景描述方法的性能优劣.但是,合理、有效、快速地评价视觉场景描述结果仍然充满挑战,主要原因包括:

① 用不同方法或模型对同一场景进行描述时,场景内容与生成语句之间的关系为“一对多”映射,即生成语句具有非唯一性.

② 同一场景或视频序列的生成描述已经可以由一个语句扩展到多个语句组成的一段语句.但是,如何用

现有数据集提供的一个人工标注语句去匹配若干语句形成的段落还有待进一步研究.

③ 人类语言表达方式的多样性使得即使在语义相同的情况下,对同一场景的描述语句也会千差万别.例如,生成语句和人工标注语句之间由于表述问题可能存在主谓倒装、一义多词的现象,这使得生成语句与人工标注语句间的主、谓、宾匹配变得更加复杂,因此有必要研究基于语义的性能评价指标.

④ 已有文献表明,注意力机制、概念(属性)等策略可以用来有效提升语句的描述能力,但是现有的评价指标并不支持基于感兴趣区域或关注对象的描述效果评价.

5 视觉场景描述面临的问题与挑战

尽管视觉场景描述的研究得到了国内外计算机视觉、自然语言处理、多媒体等相关领域研究人员的普遍重视,但其真正走向实际应用还有很多关键问题需要解决,包括:

(1) 从场景描述内容角度来看,现阶段最先进的视觉场景描述模型都是有监督方法,即公开的基准数据集提供了人工标注语句作为理想输出,而实际应用中的场景数据往往是特定场合的,如记录公安侦查过程、描述学生课堂行为等.这些特殊应用中的词汇往往不能被现有公开基准数据集所涵盖,因此没有现成的语句可供参考,无法生成符合真实场景的词汇和描述语句.

(2) 从描述准确性的角度来看,生成语句的精细度取决于训练阶段和验证阶段选用的训练样本和验证样本的人工标注语句的精细程度.现实中的视觉场景可能千变万化、转瞬即逝,是否能够准确地捕捉到各个事件及其起止时间,给出相应细微变化的内容描述非常困难,如人脸微表情变化的判断与精细描述.

(3) 从场景描述的时长角度来看,现有基准数据集提供的视频多是几秒或几分钟的短视频,而在实际应用中,各类视频文件历时较长,需要能够支撑更长序列预测的模型来完成,例如在标准化考试场景中,潜在的作弊行为的关注需要持续更长时间才能捕获有用信息,这不仅涉及由短时间视频向长时间视频方法转换的问题,而且还包括了“微弱动作”的时序检测等问题.

(4) 从场景描述的语言呈现角度来看,绝大多数基准数据集提供的人工标注是英文的,少数有其他语言的标注.尽管从技术环节来看,不同语言的描述转换可以通过机器翻译的手段完成,但是不同语言间的转换结果受各国文化背景、生活习俗及表达风格等因素的影响.

(5) 从场景载体的质量角度来看,真实应用中的实际场景与训练样本集中图像、视频的质量匹配情况,以及训练资源的丰富程度(数量、质量)是决定描述语句质量的关键因素.此外,场景载体文件的低分辨率、低对比度、复杂背景和其中可能存在的不同方向、样式、颜色、对齐方式的文字信息也使场景内容的理解与描述变得复杂.

(6) 从视觉场景描述的多学科交叉角度来看,根据第 1.2 节和第 2 节,现有的场景描述技术与场景图、视觉关系检测、指示表达生成等相关技术的最新结果并未被充分应用在改善视觉场景描述语句的生成质量上.如何以管道(pipeline)方式将其集成在场景描述模型中,以及如何优化和完善深度网络的体系结构,使之能够以更少的参数、更小的内存,更快地加以训练,是未来值得深入研究的又一问题.

6 未来应用前景

视觉场景描述技术利用计算机模仿人眼的“视觉功能”和大脑的“语言功能”,以自然语言的形式自动描述视觉场景内容,有效连接了视觉信息和语言信息,是集计算机视觉、人工智能、多媒体、自然语言处理等领域的交叉性研究课题.随着更多特定场景数据集的出现,我们相信,该技术在未来 10 年内会在许多行业和领域中有力地推动视频内容分析与理解的研究进程,并加速跨模态检索、视觉问答技术相关应用的发展,具有重要的应用价值,例如:

(1) 个性化教育中的学生行为分析:各类视频监控系统为代表的现代化教育设施迅速普及到传统教室、图书馆、报告厅、标准化考场等,由此产生了海量的与学习者行为、活动及状态相关的学习场景原始数据.通过视觉场景描述技术可将这些海量数据转换为文字表达的描述语句,利用跨模态检索技术准确地捕获学习者的

个性化特征并综合分析不同学习者的共性特征,进而提供有针对性的评估、引导与干预.例如,在智慧课堂教学中,利用计算机实时分析统计学生行为,帮助老师及时掌握学生的学习特征和状态;在军训等集体活动中,预判学生可能发生的危险行为,提高安全防范能力;在中小学生课堂纪律维持方面,通过行为分析对学生的不良行为予以及时警告,避免其因课堂注意力不集中而导致学业警示等.

(2) 智能服务中的人机交互应用:有效的人机交互在任何服务型机器人应用中都至关重要.视觉场景描述技术提供了人机交互的自然语言交互接口.通过该技术,智能机器人能够以人类易于理解的自然语言方式来实现视觉场景内容信息的表达.另一方面,视频场景内容的自然语言描述也可以作为机器人内部场景的表现形式,为基于自然语言问答的智能环境感知提供了良好基础[76].使这些机器人可以像人一样有“感情”地进行语言表达,提供高质量的服务和陪伴是未来的研究重点之一.

(3) 视力障碍人员的辅助视听:该类应用旨在对人类活动场所中的视觉感知物体进行检测、识别、分析和判断,并给视力障碍人员予以提示,以辅助视力障碍人员顺利完成行为活动.其中,如何有效地将感知到的信息正确地传递给视力障碍人员是辅助视听应用技术的关键问题之一.如何快速、有效地感知人类活动场景中与活动相关的环境信息,通过视觉问答,并以友好的方式将相关信息传递给视力障碍人员是视觉场景描述应用中需解决的重要问题.

References:

[1] Wang J, Jiang W, Ma L, Liu W, Xu Y. Bidirectional attentive fusion with context gating for dense video captioning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 7190−7198.

[2] Ren Z, Wang X, Zhang N. Deep reinforcement learning-based image captioning with embedding reward. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 1151−1159. [doi: 10.1109/CVPR.2017.128]

[3] Lu J, Xiong C, Parikh D, Socher R. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 3242−3250. [doi: 10.1109/CVPR.2017.345]

[4] Wang Y, Lin Z, Shen X, Cohen S, Cottrell GW. Skeleton key: Image captioning by skeleton-attribute decomposition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 7272−7281. [doi: 10.1109/CVPR.2017.780]

[5] Pan Y, Yao T, Li H, Mei T. Video captioning with transferred semantic attributes. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 6504−6512. [doi: 10.1109/CVPR.2017.111]

[6] Zhang X, Gao K, Zhang Y, Zhang D, Li J, Tian Q. Task-driven dynamic fusion: Reducing ambiguity in video description. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 3713−3721. [doi: 10.1109/CVPR.2017.662]

[7] Shen Z, Li J, Su Z, Li M, Chen Y, Jiang YG, Xue XY. Weakly supervised dense video captioning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 1916−1924. [doi: 10.1109/CVPR.2017.548]

[8] Krishna R, Hata K, Ren F, Niebles JC. Dense-captioning events in videos. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition 2017. 706−715. [doi: 10.1109/ICCV.2017.83]

[9] Huang Y, Wang W, Wang L. Instance-aware image and sentence matching with selective multimodal LSTM. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 7254−7262. [doi: 10.1109/CVPR.2017.767]

[10] Johnson J, Karpathy A, Li FF. DenseCap: Fully convolutional localization networks for dense captioning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 4565−4574. [doi: 10.1109/CVPR.2016.494]

[11] Xu J, Mei T, Yao T, Rui Y. MSR-VTT: A large video description dataset for bridging video and language. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 5288−5296. [doi: 10.1109/CVPR.2016.571]

[12] Yu H, Wang J, Huang Z, Yang Y, Xu W. Video paragraph captioning using hierarchical recurrent neural networks. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 4584−4593. [doi: 10.1109/CVPR.2016.496]

[13] You QZ, Jin HL, Wang ZW, Fang C, Luo JB. Image captioning with semantic attention. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 4651−4659. [doi: 10.1109/CVPR.2016.503]

[14] Wu Q, Shen C, Liu L, Dick A, Hengel AVD. What value do explicit high level concepts have in vision to language problems. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 203−212. [doi: 10.1109/CVPR.2016.29]

[15] Devlin J, Cheng H, Fang H, Gupta S, Deng L, He XD, Zweig G, Mitchell Z. Language models for image captioning: The quirks and what works. In: Proc. of the Int'l Joint conf. on Natural Language Processing. 2015. 100–105. [doi: 10.3115/v1/P15-2017]

[16] Chen X, Zitnick CL. Mind's eye: A recurrent visual representation for image caption generation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 2422–2431. [doi: 10.1109/CVPR.2015.7298856]

[17] Karpathy A, Li FF. Deep visual-semantic alignments for generating image descriptions. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 3128–3137.

[18] Donahue J, Hendricks LA, Rohrbach M, Venugopalan S, Saenko K, Darrell T. Long-term recurrent convolutional networks for visual recognition and description. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 2625–2634. [doi: 10.1109/CVPR.2015.7298878]

[19] Vedantam R, Zitnick CL, Parikh D. CIDEr: Consensus-based image description evaluation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 4566–4575. [doi: 10.1109/CVPR.2015.7299087]

[20] Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: A neural image caption generator. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 3156–3164. [doi: 10.1109/CVPR.2015.7298935]

[21] Yao L, Torabi A, Cho K, Ballas N, Pal C, Larochelle H, Courville AC. Describing videos by exploiting temporal structure. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 4507–4515. [doi: 10.1109/ICCV.2015.512]

[22] Das P, Xu C, Doell RF, Corso JJ. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2013. 2634–2641. [doi: 10.1109/CVPR.2013.340]

[23] Yao T, Pan Y, Li Y, Qiu Z, Mei T. Boosting image captioning with attributes. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 22–29. [doi: 10.1109/ICCV.2017.524]

[24] Chen TH, Liao YH, Chuang CY, Hsu WT, Fu JL, Sun M. Show, adapt and tell: Adversarial training of cross-domain image captioner. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 521–530. [doi: 10.1109/ICCV.2017.64]

[25] Li Y, Ouyang W, Zhou B. Scene graph generation from objects, phrases and region captions. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 1261–1270. [doi: 10.1109/ICCV.2017.142]

[26] Na S, Lee S, Kim J, Kim G. A read-write memory network for movie story understanding. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 677–685. [doi: 10.1109/ICCV.2017.80]

[27] Hu R, Andreas J, Rohrbach M, Darrell T, Saenko K. Learning to reason: End-to-end module networks for visual question answering. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 804–813. [doi: 10.1109/ICCV.2017.93]

[28] Teney D, Liu L, Den Hengel AV. Graph-structured representations for visual question answering. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 1536–1544. [doi: 10.1109/ICCV.2017.93]

[29] Zhu C, Zhao Y, Huang S. Structured attentions for visual question answering. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2017. 1291–1300. [doi: 10.1109/ICCV.2017.145]

[30] Xu H, Venugopalan S, Ramanishka V, Rohrbach M, Saenko K. A multi-scale multiple instance video description network. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2015. 272–279.

[31] Venugopalan S, Rohrbach M, Donahue J. Mooney RJ, Darrell T, Saenko K. Sequence to sequence-video to text. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2015. 4534–4542. [doi: 10.1109/ICCV.2015.515]

[32] Guadarrama S, Krishnamoorthy N, Malkarnenkar G, Venugopalan S, Mooney R, Darrell T, Saenko K. YouTube2Text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2013. 2712–2719. [doi: 10.1109/ICCV.2013.337]

[33] Rohrbach M, Qiu W, Titov I. Translating video content to natural language descriptions. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2013. 433–440. [doi: 10.1109/ICCV.2013.61]

[34] Mallya A, Lazebnik S. Learning models for actions and person-object interactions with transfer to question answering. In: Proc. of the European Conf. on Computer Vision. 2016. 414–428. [doi: 10.1007/978-3-319-46448-0_25]

[35] Rohrbach A, Rohrbach M, Hu R. Grounding of textual phrases in images by reconstruction. In: Proc. of the European Conf. on Computer Vision. 2016. 817–834. [doi: 10.1007/978-3-319-46448-0_49]

[36] Lu C, Krishna R, Bernstein M. Visual relationship detection with language priors. In: Proc. of the European Conf. on Computer Vision. 2016. 852−869. [doi: 10.1007/978-3-319-46448-0_51]

[37] Peter A, Basura F, Mark J, Stephen G. SPICE: Semantic propositional image caption evaluation. In: Proc. of the European Conf. on Computer Vision. 2016. 382−398.

[38] Lin X, Parikh D. Leveraging visual question answering for image-caption ranking. In: Proc. of the European Conf. on Computer Vision. 2016. 261−277. [doi: 10.1007/978-3-319-46475-6_17]

[39] Wu Q, Cai HP, Hall P. Learning graphs to model visual objects across different depictive styles. In: Proc. of the European Conf. on Computer Vision. 2014. 313−328. [doi: 10.1007/978-3-319-10584-0_21]

[40] Farhadi A, Hejrati M, Sadeghi MA. Every picture tells a story: Generating sentences from images. In: Proc. of the European Conf. on Computer Vision. 2010. 15−29. [doi: 10.1007/978-3-642-15561-1_2]

[41] Seo PH, Lehrmann A, Han B. Visual reference resolution using attention memory for visual dialog. In: Proc. of the 31st Annual Conf. on Neural Information Processing Systems. 2017. 3722−3732.

[42] Dai B, Lin D. Contrastive learning for image captioning. In: Proc. of the 31st Annual Conf. on Neural Information Processing Systems. 2017. 898−907.

[43] Wang L, Schwing A, Lazebnik S. Diverse and accurate image description using a variational auto-encoder with an additive Gaussian encoding space. In: Proc. of the 31st Annual Conf. on Neural Information Processing Systems. 2017. 5758−5768.

[44] Yeh R, Xiong J, Hwu WM. Interpretable and globally optimal prediction for textual grounding using image concepts. In: Proc. of the 31st Annual Conf. on Neural Information Processing Systems. 2017. 1909−1919.

[45] Fidler S. Teaching machines to describe images with natural language feedback. In: Proc. of the Neural Information Processing Systems. 2017. 5075−5085.

[46] Yang Z, Yuan Y, Wu Y. Review networks for caption generation. In: Proc. of the 30th Annual Conf. on Neural Information Processing Systems. 2016. 2361−2369.

[47] Fang H, Gupta S, Iandola F. From captions to visual concepts and back. In: Proc. of the 29th Annual Conf. on Neural Information Processing Systems. 2015. 1473−1482. [doi: 10.1109/CVPR.2015.7298754]

[48] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Proc. of the 26th Annual Conf. on Neural Information Processing Systems. 2012. 1097−1105. [doi: 10.1145/3065386]

[49] Ordonez V, Kulkarni G, Berg TL. Im2text: Describing images using 1 million captioned photographs. In: Proc. of the 25th Annual Conf. on Neural Information Processing Systems. 2011. 1143−1151.

[50] Cho K, Van MB, Gulcehre C. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proc. of the 3rd Int'l Symp. on Natural Language Processing. 2014. 1724−1734. [doi: 10.3115/v1/D14-1179]

[51] Yang Y, Teo CL, Aloimonos Y. Corpus-guided sentence generation of natural images. In: Proc. of the Int'l Symp. Natural Language Processing. 2011. 444−454.

[52] Subhashini V, Xu HJ, Donahue J, Rohrbach M, Mooney R, Saenko K. Translating videos to natural language using deep recurrent neural networks. In: Proc. of the 2015 Conf. of the North American Chapter of the Association for Computational Linguistics. 2015. 1494−1504. [doi: 10.3115/v1/N15-1173]

[53] Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY. Multimodal deep learning. In: Proc. of the Int'l Conf. on Machine Learning. 2011. 689−696.

[54] Wang K, He R, Wang L, Wang W, Tan T. Joint feature selection and subspace learning for cross-modal retrieval. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2016,38(10):2010−2023. [doi: 10.1109/TPAMI.2015.2505311]

[55] Li XL, Shi JH, Dong YS, Tao DC. A survey on scene image classication. Scientia Sinica Informationis, 2015,45(7):827−848 (in Chinese with English abstract).

[56] Lowry SM, Sunderhauf N, Newman P, Leonard JJ, Cox DD, Corke P, Milford M. Visual place recognition: A survey. IEEE Trans. on Robotics, 2016,32(1):1−19. [doi: 10.1109/TRO.2015.2496823]

[57] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556[cs.CV], 2014.

[58] Song X, Jiang S, Herranz L. Multi-scale multi-feature context modeling for scene recognition in the semantic manifold. IEEE Trans. on Image Processing, 2017,26(6):2721−2735. [doi: 10.1109/TIP.2017.2686017]

[59] Luis H, Jiang SQ, Li XY. Scene recognition with CNNs: Objects, scales and dataset bias. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2016. 571−579. [doi: 10.1109/CVPR.2016.68]

[60] Zhang H, Kyaw Z, Chang S, Chua T. Visual translation embedding network for visual relation detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2017. 3107−3115. [doi: 10.1109/CVPR.2017.331]

[61] Wu Q, Shen C, Wang P, Dick A, Hengel AVD. Image captioning and visual question answering based on attributes and external knowledge. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017,40(6):1367−1381. [doi: 10.1109/TPAMI.2017.2708709]

[62] Johnson J, Krishna R, Stark M, Li L, Shamma DA, Bernstein MS, Feifei L. Image retrieval using scene graphs. In: Proc. of the Computer Vision and Pattern Recognition. 2015. 3668−3678. [doi: 10.1109/CVPR.2015.7298990]

[63] Li XY, Jiang SQ. Bundled object context for referring expressions. IEEE Trans. on Multimedia, 2018,20(10):2749−2760.

[64] Krahmer E, Van Deemter K. Computational generation of referring expressions: A survey. Computational Linguistics, 2012,38(1):173−218. [doi: 10.1162/COLI_a_00088]

[65] Mitchell M, Han X, Dodge J, Mensch A, Goyal A, Berg A, Yamaguchi K, Berg T, Stratos K, Daume H III. Midge: Generating image descriptions from computer vision detections. In: Proc. of the European Association of Computational Linguistics. 2012. 747−756.

[66] Krishnamoorthy N, Malkarnenkar G, Mooney R, Saenko K, Guadarrama S. Generating natural-language video descriptions using text-mined knowledge. In: Proc. of the Association for the Advance of Artificial Intelligence. 2013. 541−547.

[67] Kulkarni G, Premraj V, Dhar S, Berg AC, Berg TL. Babytalk: Understanding and generating simple image descriptions. IEEE Trans.on Pattern Analysis and Machine Intelligence, 2013,35(12):2891−2903. [doi: 10.1109/TPAMI.2012.162]

[68] Lebret R, Pinheiro PHO, Collobert R. Phrase-based image captioning. arXiv: 1502.03671[cs.CV], 2015.

[69] Kojima A, Tamura T, Fukunaga K. Natural language description of human activities from video images based on concept hierarchy of actions. Int'l Journal of Computer Vision, 2002,50(2):171−184.

[70] Xu R, Xiong C, Chen W, Corso JJ. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In: Proc. of the Association for the Advance of Artificial Intelligence. 2015. 2346−2352.

[71] Kuznetsova P, Ordonez V, Berg T, Choi Y. Treetalk: Composition and compression of trees for image descriptions. Trans. of the Association of Computational Linguistics, 2014,2(10):351−362.

[72] Hodosh M, Young P, Hockenmaier J. Framing image description as a ranking task: Data, models and evaluation metrics. In: Proc. of the Association for the Advance of Artificial Intelligence. 2015. 4188−4192. [doi: 10.1613/jair.3994]

[73] Devlin J, Gupta S, Girshick R, Mitchell M, Zitnick CL. Exploring nearest neighbor approaches for image captioning. arXiv: 1505.04467[cs.CV], 2015.

[74] Graves A. Supervised Sequence Labeling with Recurrent Neural Networks. Berlin, Heidelberg: Springer-Verlag, 2012.

[75] Lipton Z, Berkowitz J, Elkan C. A critical review of recurrent neural networks for sequence learning. arXiv: 1506.00019[cs.CV], 2015.

[76] Cascianelli S, Costante G, Ciarfuglia TA, Valigi P, Fravolini ML. Full-GRU natural language video description for service robotics applications. IEEE Robotics & Automation Letters, 2018,3(2):841−848. [doi: 10.1109/LRA.2018.2793345]

[77] Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y. Show, attend and tell: Neural image caption generation with visual attention. In: Proc. of the Int'l Conf. on Machine Learning. 2015. 2048−2057.

[78] Socher R, Karpathy A, Le QV, Manning CD, Ng AY. Grounded compositional semantics for finding and describing images with sentences. Trans. of the Association for Computational Linguistics, 2014,2:207−218.

[79] Torabi A, Tandon N, Sigal L. Learning language-visual embedding for movie understanding with natural-language. arXiv: 1609.08124[cs.CV],2016.

[80] Everingham M, Zisserman A, Williams CKI, *et al*. The 2005 PASCAL visual object classes challenge. In: Proc. of the Int'l Conf. on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment. 2005. 117−176. [doi: 10.1007/11736790_8]

[81] Young P, Lai A, Hodosh M, Micah H, Julia H. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Trans. of the Association for Computational Linguistics, 2014,2(1):67−78.

[82] Plummer BA, Wang L, Cervantes CM, Caicedo JC, Hockenmaier J, Lazebnik S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. Int'l Journal of Computer Vision, 2015,123(1):74−93. [doi: 10.1007/s11263-016-0965-7]

[83] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollar P, Zitnick CL. Microsoft COCO: Common objects in context. In: Proc. of the European Conf. on Computer Vision. 2014. 740−755. [doi: 10.1007/978-3-319-10602-1_48]

[84] Ni K, Pearce R, Boakye K, Van Essen B, Borth D, Chen B, Wang EX. Large-scale deep learning on the YFCC100M dataset. arXiv: 1502.03409[cs.CV], 2015.

[85] Krishna R, Zhu Y, Groth O, *et al*. Visual Genome: Connecting language and vision using crowd sourced dense image annotations. Int'l Journal of Computer Vision, 2017,123(1):32−73. [doi: 10.1007/s11263-016-0981-7]

[86] Wu J, Zheng H, Zhao B, Li YX, Yan BM, Liang R, *et al*. AI challenger: A large-scale dataset for going deeper in image understanding. arXiv: 1711.06475v1 [cs.CV], 2017.

[87] Chen DL, Dolan WB. Collecting highly parallel data for paraphrase evaluation. In: Proc. of the Association for Computational Linguistics: Human Language Technologies. 2011. 190−200.

[88] Rohrbach A, Rohrbach M, Qiu W, Friedrich A, Pinkal M, Schiele B. Coherent multi-sentence video description with variable level of detail. In: Proc. of the German Conf. on Pattern Recognition. 2014. 184−195. [doi: 10.1007/978-3-319-11752-2_15]

[89] Rohrbach A, Rohrbach M, Tandon N, Schiele B. A dataset for movie description. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2015. 3202−3212. [doi: 10.1109/CVPR.2015.7298940]

[90] Torabi A, Pal C, Larochelle H, Courville A. Using descriptive video services to create a large data source for video annotation research. arXiv: 1503.01070v1[cs.CV], 2015.

[91] Zhou L, Xu C, Corso JJ. Towards automatic learning of procedures from Web instructional videos. In: Proc. of the Association for the Advance of Artificial Intelligence. 2018. 7591−7598.

[92] Papineni K, Roukos S, Ward T, Zhu W. BLEU: A method for automatic evaluation of machine translation. In: Proc. of the 40th Annual Meeting on Association for Computational Linguistics. 2002. 311−318. [doi: 10.3115/1073083.1073135]

[93] Callison-Burch C, Osborne M, Koehn P. Re-evaluation the role of Bleu in machine translation research. In: Proc. of the European Association of Computational Linguistics. 2006. 249−256.

[94] Mahathir F. Sistem pendeteksi plagiat pada dokumen teks berbahasa Indonesia menggunakan metode Rouge-N, Rouge-L dan Rouge-W. 2011. http://repository.ipb.ac.id/handle/123456789/50046

[95] Lin CY, Och FJ. Automatic evaluation of machine translation quality using longest common subsequence and Skip-bigram statistics. In: Proc. of the Association for Computational Linguistics. 2004. 605−612. [doi: 10.3115/1218955.1219032]

[96] Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proc. of the 43rd Annual Meeting on Association for Computational Linguistics. 2005,(29):65−72.

[97] Wong B, Kit C. ATEC: Automatic evaluation of machine translation via word choice and word order. Machine Translation, 2009, 23(2-3):141−155. [doi: 10.1007/s10590-009-9061-x]

[98] Kusner MJ, Sun Y, Kolkin NI, Weinberger KQ. From word embeddings to document distances. In: Proc. of the Int'l Conf. on Machine Learning. 2015. 957−966.

[99] Anderson P, Fernando B, Johnson M. SPICE: Semantic propositional image caption evaluation. In: Proc. of the European Conf. on Computer Vision. 2016. 382−398. [doi: 10.1007/978-3-319-46454-1_24]

[100] Ma M, Wang B. A grey relational analysis based evaluation metric for image captioning and video captioning. In: Proc. of the Grey Systems and Intelligent Services. 2017. 76−81. [doi: 10.1109/GSIS.2017.8077673]

[101] Cui Y, Yang G, Veit A, Huang X, Belongie SJ. Learning to evaluate image captioning. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 5804−5812.

[102] Kilickaya M, Erdem A, Ikizler-Cinbis N, Erdem E. Re-evaluating automatic metrics for image captioning. In: Proc. of the European Chapter of the Association for Computational Linguistics. 2016. 199−209. [doi: 10.18653/v1/E17-1019]

[103] Giménez J, Màrquez L. Linguistic features for automatic evaluation of heterogenous MT systems. In: Proc. of the 2nd Workshop on Statistical Machine Translation. 2007. 256−264.

[104] ShafieiBavani E, Ebrahimi M, Wong R, Chen F. A semantically motivated approach to compute ROUGE scores. arXiv: 1710.07441 [cs.CV], 2017.

[105] Koehn P, Monz C. Manual and automatic evaluation of machine translation between European languages. In: Proc. of the Workshop on Statistical Machine Translation. 2006. 102−121. [doi: 10.3115/1654650.1654666]

附中文参考文献:

[55] 李学龙,史建华,董永生,陶大程.场景图像分类技术综述,中国科学:信息科学,2015,45(7):827−848.

马苗(1977－),女,山东聊城人,博士,教授,CCF 高级会员,主要研究领域为图像处理,模式识别,视频分析.

王伯龙(1993－),男,硕士,主要研究领域为视频分析与描述.

吴琦(1987－),男,博士,助理教授,博士生导师,主要研究领域为计算机视觉,机器学习,视觉问答.

武杰(1985－),男,博士,讲师,主要研究领域为遥感影像处理.

郭敏(1964－),女,博士,教授,博士生导师,主要研究领域为图像处理,模式识别,智能信息处理.