基于深度学习的图像描述综述

石义乐[1],杨文忠[2],杜慧祥[1],王丽花[1],王 婷[1],理珊珊[1]
(1. 新疆大学软件工程技术重点实验室,新疆乌鲁木齐 830000; 2. 新疆大学信息科学与工程学院,新疆乌鲁木齐 830000)

摘 要: 图像描述旨在通过提取图像的特征输入到语言生成模型中最后输出图像对应的描述,来解决人工智能中自然语言处理与计算机视觉的交叉领域问题——智能图像理解.现对2015—2020年间图像描述方向有代表性的论文进行汇总与分析,以不同核心技术作为分类标准将图像描述大致划分为基于Encoder-Decoder框架的图像描述、基于注意力机制的图像描述、基于强化学习的图像描述、基于生成对抗网络的图像描述和基于新融合数据集的图像描述五大类.使用NIC、Hard-Attention和Neural Talk三个模型在真实数据集MS-COCO数据集上进行实验,并从BLEU1、BLEU2、BLEU3、BLEU4四处平均评分对比分析,展示三个模型效果.本文点明了未来图像描述的发展趋势,并指出了图像描述将要面临的挑战和可深入挖掘的研究方向.

关键词: 智能图像理解; Encoder-Decoder框架; 注意力机制; 强化学习
中图分类号: TP391.2 **文献标识码:** A **文章编号:** 0372-2112(2021)10-2048-13
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.12263/DZXB.20200669

Overview of Image Captions Based on Deep Learning

SHI Yi-le[1], YANG Wen-zhong[2], DU Hui-xiang[1], WANG Li-hua[1], WANG Ting[1], LI Shan-shan[1]
(1. *Key Laboratory of Software Engineering Technology, Xinjiang University, Urumqi, Xinjiang* 830000, *China*;
2. *School of Information Science and Engineering, Xinjiang University, Urumqi, Xinjiang* 830000, *China*)

Abstract: Image caption aims to extract the features of the image and input the description of the final output image into the language generation model, which solves the intersection of natural language processing and computer vision in artificial intelligence-image understanding. Summarize and analyze representative thesis of image description orientation from 2015 to 2020, different core technologies as classification criteria, it can be roughly divided into: image caption based on Encoder-Decoder framework, image caption based on attention mechanism, image caption based on reinforcement learning, image caption based on Generative Adversarial Networks, and based on new fusion data set these five categories. Use three models of NIC, Hard-Attention and Neural Talk to conduct experiments on the real data set MS-COCO data set, and compare the average scores of BLEU1, BLEU2, BLEU3, and BLEU4 to show the effects of the three models. This article points out the development trend of image caption in the future, and the challenges that image caption will face and the research directions that can be digged in.

Key words: intelligence-image understanding; encoder-decoder framework; attention mechanism; reinforcement learning

1 引言

人工智能上升到国家战略地位,体现出人工智能在国家大环境中的受重视程度.近几年人工智能的两大分支——计算机视觉和自然语言处理非常火热,伴随而出的计算机视觉与自然语言处理交叉领域——智能图像理解,属于多模态问题,已出现图像描述、智能问答系统等方向的应用研究.

在2015年以前图像描述的研究发展很艰难,但在此之后图像描述正式进入深度学习时代,如图1所示.在国外,图像描述的主要研究机构有谷歌、微软和斯坦福大学等,在国内,主要有阿里巴巴、腾讯和上海交通大学等,虽然图像描述是由国外带头发起研究的,但目前国内的研究在技术上与国外处于同一水平.当前图像描述只能达到一定程度,还远达不到深层次的语义

收稿日期:2020-07-08;修回日期:2020-09-13;责任编辑:孙瑶
基金项目:国家自然科学基金(No.U1603115);新疆维吾尔自治区自然科学基金(No.2017D01C042)

理解,不足以支撑将图像描述应用到现实复杂环境中,这才是形成本文的最终原因.

图像描述通常在图像处理和语言生成方向进行研究. 图像处理研究希望更大程度提取出足够详细的对象或属性特征,使图像特征最大程度对应到生成描述,利用目标检测法提取特征. 2020 年权宇等人[1]通过融合深度扩张网络和轻量化网络优化目标检测,提高图像检测有效性与可扩展性. 刘颖等人[2]从小目标处做目标检测研究,解决小目标难以检测的技术问题,同时提高图像描述对应细粒度图像特征. 目标检测研究可以促进提取图像特征发展.

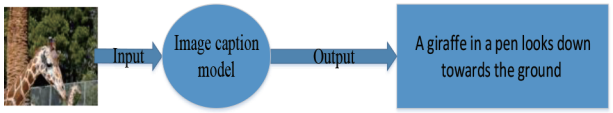

图 1　图像描述实现图

2　图像描述算法研究现状

通过对关于 image caption 在 2010—2018 年内简要综述的博客[3]做出补充,按照热门核心技术应用的思路对截至 2020 年在图像描述上具有重大贡献的研究进行总结和分析. 2015 年 Encoder-Decoder 框架在机器翻译上取得了巨大成功,这使得基于深度学习技术的图像描述进入 Encoder-Decoder 框架时期. 随后注意力机制在自然语言处理上的强势表现使图像描述进入注意力机制时期. 在强化学习克服了神经网络梯度消失和梯度爆炸问题后,图像描述又进入了强化学习时期. 随着 2017 年 GAN 网络的突然爆红,出现了序列 GAN 网络,使图像描述进入 GAN 网络时期. 从 2020 年开始,在保证多样性的前提下追求独特性,近两年涌现出大量利用新融合数据提高生成描述独特性的论文,图像描述自此进入新融合数据集时期.

2.1　基于 Encoder-Decoder 框架的研究

2015 年谷歌 Show and Tell 论文[4]受机器翻译模型[5]影响将 Inception v3 作为 CNN 提取图像特征,用 LSTM 获得更好的全局语义[6]. 同年斯坦福大学 Karpathy 等人[7]提出 Neural Talk,基本框架与谷歌模型相似,区别是图像特征提取器为 VGGnet[8]. 图像描述近五年的研究基本框架几乎都是 Encoder-Decoder 框架,在此基础上进行创新,如图 2 所示.

2.1.1　在 Encoder 端的研究

首先介绍编码端创新论文,框架如图 3 所示. Fang 等人[9]通过多实例学习训练一个词探测器,用于产生图像描述中可能出现的词,将得到的词输入生成模型并获得描述,最后从中选出最优描述. 这种利用提取关键词作为输入的方法为未来结合图像-语义编码提供了经验. 2018 年 Li 等人[10]提出一种新图像特征提取法,通

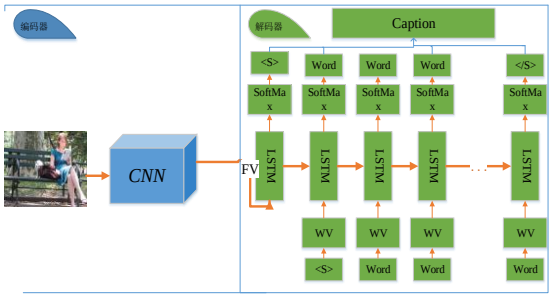

图 2　Encoder-Decoder 框架图

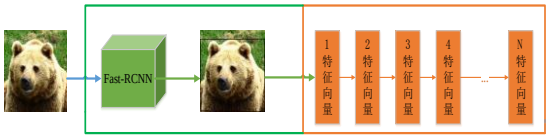

图 3　Encoder 框架图

过对象检测法获取很多对象检测框作为图像特征,同时利用图像特征作为输入来训练一个属性检测器,此属性属于高层语义特征,它与图像特征一起输入到已特殊设计的 Visual-Semantic LSTM 进行解码,而对象检测会使输入特征变得紧促,而不像输入整张图像那样获得视觉注意力. 同年 Anderson 等人[11]的研究也使用了类似编码器. 2019 年 Fei 等人[12]提出在编码端使用层次视觉关系法,将语义和空间视觉关系嵌入图像编码器. 同年 Lee 等人[13]提出视觉相关关系的场景图生成器,可以提取有效视觉翻译特征来促进语言-视觉关系. Yao 等人[14]提出层次解析架构 HIP,构建一个利用树状结构的 LSTM 对层次结构进行交互的层次树,通过增强实例级、区域级和图像级特征来提高 CIDEr-D 评分. 2020 年 He 等人[15]提出图像转换器,由图像区域之间相对空间关系驱动来修改编码转换器和隐式解码转换器,拓宽了原始转换器层的内部架构来适应图像结构. 张红斌等人[16]在改进有效区域选择与语义挖掘的图像属性标注处做出创新,通过有效图像区域特征融合迁移学习策略来实现属性内容.

随着对编码端的不断研究,单个图像特征不足以提高生成描述的质量,未来编码端需要多图像联合训练,例如,Chen 等人[17]提出的图像描述模型 Groupcap,受编码多张图像启发,可同时提高相似性、多样性. 模型第一部分是一个视觉语法分析树,作用是将得到图像特征建模到树节点内. 模型第二部分由结构化相关性与多样性限制模块组成,输入的图像三元组之间的相似性和多样性由其解析树叶节点实体决定. 除了训练图像三元组目标图像之外,将另外两个图像标签定义为正性或负性,用来表示测试图像与目标图像之间的相近性,训练的最终目的是最大化同组图像相似性并最小化非同组图像相似性,而多样性则相反. 模型第

三部分是描述生成过程. 多模型联合训练后拥有分辨图像相似程度的能力. 在视频描述领域,Pasunuru 等人[18]将视频描述任务与无监督视频重建任务在一个限定生成任务中进行联合训练,得到质量更好的字幕. 2019 年 Zhou 等人[19]提出视觉语言预训练模型 VLP,为共享生成器提供特定自注意掩码.

2.1.2 在 Decoder 端的研究

有大量对于编码端的研究,解码端的研究也多种多样,框架如图 4 所示. Wang 等人[20]提出一种新型解码方法,Skeleton-Attribute decoder 由 Skel-LSTM 和 Attr-LSTM 组成,前者使用 CNN 提取出图像特征得到一个主干描述,后者为每一个主干描述中的核心词匹配出相应属性词,再将两部分合成最终描述. 此类研究还有 Neural baby talk. 受句子填充模板 baby talk 影响,2018 年 Lu 等人[21]提出基于模板生成和填充的图像描述,先将生成描述中的词分为实体词与非实体词两类词表,再利用语言生成模型得到句子模板,该词来源于非实体词表,后用对象检测法从图像中得到实体词来填充句子模板空处,形成目的描述. 这种开创性的用深度学习技术提取句子模板的方法,成功解决了在第二部分传统模板填充时缺乏多样性的问题. 在分离解码器的研究中,Mathews 等人[22]为了得到风格化图像描述,使用两个解码器来完成,第一个解码器是 term generator,利用 CNN 得到图像特征再输入 GRU 获得一系列基本语义对,即词语-属性,再将它输入生成模型得到描述,生成模型使用双向 GRU 编码得到按顺序排列基本语义,最后新 GRU 进行解码. 2020 年 Wang 等人[23]提出语义指南、回忆词槽和单独回忆词奖励的组合法,最大程度利用召回单词配合单独回忆词奖励来增强训练. Nguyen 等人[24]提出持续学习图像描述模型 ContCap,是以简单微调模式为基础克服遗忘困境的技术.

对解码器进一步的深入研究中,Gu 等人[25]提出了更加精细的 stack caption,由一个粗粒度解码器和多个细粒度解码器组成,先把提取到的图像特征输入粗粒度解码器得到结果,再将输出结果和图像特征由一个细粒度解码器进行更精细解码,同时使用 attention 机制使细粒度解码器在每一阶段对粗粒度解码器生成多方面描述. Fei 等人[26]提出改进的非自回归预测模型,解码包括位置对齐、对图像中检测到单词排序和一个精细非自回归解码器. 2020 年 Thapliyal 等人[27]在解码端提出枢轴语言生成标准转换法,把生成英文描述转换成目标语言描述,却无意间提高了原始英文描述质量. Xia 等人[28]提出了更复杂的 XGPT,通过图像掩蔽语言建模条件、图像降噪自编码条件和文本图像特征生成条件对生成器预处理,得到最新的效果.

在解码端利用层级解码思路提高图像描述解释性. 2017 年 Yao 等人[29]首次将复制机制引入解码端,在图像特征上训练对象检测器,通过 LSTM 解码器的输出层 ht 与检测到实体对象的相似性共同决定是否复制. 2018 年一项关于 Convolutional image caption 的研究[30],在每一时刻都将词和提取到的图像特征输入卷积编码器,最后使用卷积解码得到词概率,使用 CNN 克服 RNN 的时序限制问题,其优点是参数数量相同但训练速度快. 2019 年 Nikolaus 等人[31]提出了多任务模型,使用解码机制根据标题与图像的相似性对其重新排序,解决了其他模型泛化性能差的问题.

仅仅利用图像特征描述图像是不够的. Lu 等人[32]使用类似于 Neural Baby Talk 的模型,利用端对端框架得到一个含有实体空槽的语言生成模板,再用实体填充,该研究先将与训练数据标签相近的图像描述作为上下文,从中抽取命名实体输入知识图谱中,再从知识图谱中选出概率最高的实体组合输入插槽. 如何通过引入外部语料来提高生成描述的语义丰富度是值得深入研究的问题. 2020 年 Sammani 等人[33]提出修改学习,对残差信息进行建模来学习修改现有描述,在每个时间步上保留学习到的模型,删除或添加到现有描述中,从而使模型可以专注修改而不是预测.

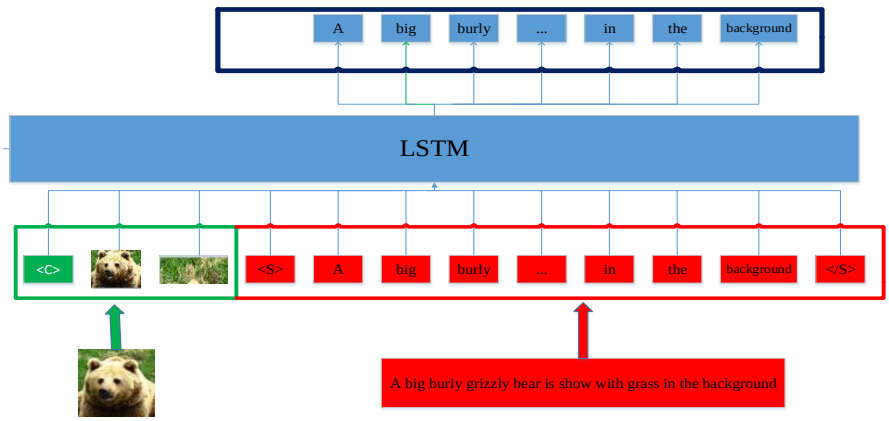

图 4 Decoder 框架图

在编码端使用基于CNN的已预训练图像检测模型,例如inception v4等,在解码端使用LSTM. 具体实现如下:

$$x_{-1}=\mathrm{CNN}(I) \tag{1}$$

$$\boldsymbol{x}_t=W_e S_t,\quad (t\in\{0,\cdots,N-1\}) \tag{2}$$

$$p_{t+1}=\mathrm{LSTM}(\boldsymbol{x}_t),\quad (t\in\{1,\cdots,N-1\}) \tag{3}$$

其中,I是输入的图像,x_{-1}是提取到的图像特征,$\boldsymbol{x}_t$是词向量,p_{t+1}是词预测概率.

2.2　基于注意力机制的研究

注意力机制在自然语言处理上取得的巨大成功引起研究者对图像描述的兴趣. 2016年Xu等人[34]首次将注意力用到图像描述中,如图5所示. 编码端注意力加权图像特征后,把它输入LSTM中解码出描述. 该文提出两类注意力:软注意力和硬注意力. 软注意力是在每一个图像区域中训练一个介于0与1之间和为1的注意力权重,再将各图像区域进行加权求和. 而硬注意力则将1作为最大权重、其他区域权重设为0,以实现仅侧重描述一个区域. 在现实研究中软注意力应用更有意义.

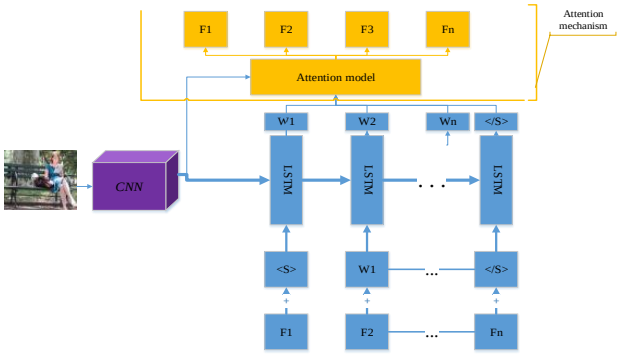

图5　注意力框架图

Lu等人[35],考虑传统注意力在某时段从图像特征中获得的新特征并不具有良好的灵活性,提出视觉岗哨法,其岗哨向量表示解码器记忆中已经得到知识,而岗哨门控是一个用于控制岗哨向量和经attend之后图像特征所占权重比的门控机制,将其加权相加当作该时间戳的解码向量. 该方法使得模型能够决定是在语言生成模型中得到语义关注更多还是图像特征关注更多. 2020年Huang等人[36]提出自适应时间注意模型AAT,它可以使一个图像区域映射到任意数量的描述词,同样地,一个描述词也可以处理任意数量图像区域.

之前注意力都类似硬注意力,从2018年开始了对自适应注意力的研究. Anderson等人[11]用Faster R-CNN进行目标检测得到目标与标签,实现Bottom-Up attention描述结果. 解码端根据输出描述使用注意力LSTM层对输入图像特征随时进行注意力权重调整. 2019年Goel等人[37]在编码端提出改进标准软注意力,使用两流网络自动学习每个图像关联部分潜在类别. 2020年Guo等人[38]在上个标准软注意力基础上提出几何感知自我注意力,它有效地考虑了图像中对象之间的相关几何关系,此方法比较超前. Pan等人[39]提出一种统一的X-Linear注意力网络,它充分利用双线性池对视觉信息进行集中利用或多模态推理. 通过获取语句过程对这些概念进行注意,代表有Wu等人[40]. Sun等人[41]在解码端提出解释注意力模型,将注意力热图属性与解释法计算出的属性进行比较.

将语义和图像注意力结合是新的研究点. 2018年Liu等人[42]提出Sim Net图像描述模型,通过结合视觉注意力和语义注意力,使用多实例学习从图像特征中提取描述词汇作为语义注意力的对象,将同时刻已注意过的图像特征向量与标注描述向量一起编码为混合向量,根据输出的注意力对描述向量和混合向量再次进行加权决定语义和图像模块的重要性. 以下几个公式就是处理过程.

$$x_i^1=\left[h_{t-1}^2,\bar{v},W_c\prod\nolimits_t\right] \tag{4}$$

$$\alpha_{i,t}=w_a^T\tanh\left(W_{va}v_i+W_{ha}h_t^1\right) \tag{5}$$

$$\hat{v}_i=\sum_{i=1}^{K}\alpha_{i,t}v_i \tag{6}$$

$$x_t^2=\left[\hat{v}_t h_t^1\right] \tag{7}$$

其中,W_c是词嵌入权值,$\prod_t$是one-hot编码. 在每个时间戳Attention LSTM模型都会输出一个h_t^1且还会为每一个区域的图像特征向量提供一个已经标准化的注意力强度$\alpha_{i,t}$,得到Attend后的$\hat{v}_i$.

2.3　基于强化学习的研究

人工智能领域强化学习已经很成熟,使反向梯度训练带来的问题迎刃而解. 强化学习是策略梯度优化模型正向训练,如图6所示. 在评价指标处的研究有:Ranzato等人[43]在解码端使用强化学习,通过直接优化BLEU、ROUGE等评价指标训练模型,解决模型训练过程中对数似然与评价指标相关性不强的问题;Liu等人[44]提出直接将评价指标作为奖励通过策略梯度优化参数,针对生成非完整描述而无法评价的问题,该文使用蒙特卡洛搜索法对每步非完整描述扩展到完整描述.

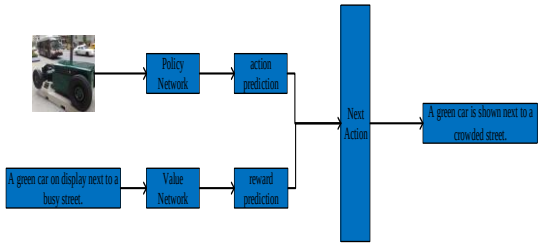

图6　强化学习框架图

另一类复杂强化学习研究有：2017年Ren等人[45]，基于强化学习的图像描述，将完整强化学习引入描述生成，把图像描述看作决策生成，将输入图像和当前产生词汇作为status，将词表中词汇作为action. 策略网络是典型端对端结构，先由CNN将图像编码，再由RNN解码生成词汇来输出奖励，奖励由视觉-语义编码决定，通过联合训练视觉描述编码，最终奖励由欧氏距离决定；2019年Seo等人[46]提出利用实例级人类描述评分作为离线强化学习奖励；2020年Bujimalla等人[47]提出基于政策梯度的贝叶斯强化学习模型B-SCST，通过使用贝叶斯DNN模型获得策略梯度基准奖励；在视频字幕任务上Pasunuru等人[48]使用强化学习解决了上述所说的对数似然法局限性问题，该方法还提出对CIDEr评价标准升级后的CIDEnt奖励机制. 正是强化学习的特点决定了对文本生成任务的优势，现有研究证明了强化学习在提高生成序列多样性和合理化上确实比传统方法有优势. 以下是策略网络过程：

$$x_0 = w^{x,v}\mathrm{CNN}_p(I) \tag{8}$$

$$h_t = \mathrm{RNN}_p\left(h_{t-1}, x_t\right) \tag{9}$$

$$p_\pi\left(a_t s_t\right) = \varphi\left(h_t\right) \tag{10}$$

其中，$w^{x,v}$是线性嵌入模型权重，ϕ函数和φ函数表示RNNp模型的输入和输出. 价值网络作用是价值评估，过程如下.

$$V^p(s) = E\left[\Upsilon | s_t = s, a_{t,\cdots,T} \sim p\right] \tag{11}$$

2.4 基于生成对抗网络的研究

传统端对端框架在训练时采用交叉熵损失函数，使用似然最大化训练生成模型生成描述，但得到的描述与标注描述还存在很大差异. 基于传统对抗网络提出序列对抗网络并应用到图像描述中，如图7所示. 对抗网络的思路是在博弈对抗中向优发展. 对抗网络由生成器和鉴别器组成，其生成器最大化拟合真实数据使假数据能够以假乱真，而鉴别器分类真假数据.

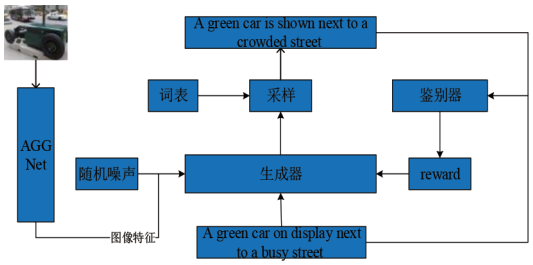

图7 生成对抗网络框架图

在图像描述中引入生成对抗网络的研究起步很晚. 2017年Dai等人[49]提出Conditional GAN模型，目的是能够生成丰富、多样的图像描述. 通过控制随机初始化生成器LSTM隐藏层向量方差，为同一个图像生成不同描述，而鉴别器在每一时间戳随机接收标注描述和生成描述，同时还接受图像特征作为真假描述差异值，此研究引入经典GAN结构，提高了描述的多样性. 同年Shetty等人[50]也引入生成对抗网络，其与前文对抗网络结构一样，主要创新点在于针对一个图像有多个标注特点，引入一种新判别器验证结构，它除了使用传统图像-文字距离判定图像-语义相似度外，还引入对同一个图像不同描述之间距离测量方法，在语义多样性上对不同描述之间进行判别，进一步提高生成器去关注语义多样性能力. 2020年Liu等人[51]提出文本与图像互译的对抗网络TIME，采用Transformer对图像特征与词嵌入之间的交叉模式连接进行训练，并设计一个铰链式和退火条件式损失，动态平衡对抗性学习在性能上有很强的竞争力. Li等人[52]提出改进的对抗逆增强学习法rAIRL，通过解开句子中每个单词奖励来处理奖励歧义问题，通过完善损失函数使生成器向Nash转移来实现稳定对抗训练，实验表明可以学习到图像描述紧凑奖励. 还有一些使用对抗样本对图像描述进行攻击来检测鲁棒性研究，如Chen等人[53]使用图像对抗样本进行攻击研究和Shekher等人[54]通过使用语义对抗样本来评估图像描述模型鲁棒性. 而Dai等人[55]使用对抗样本训练生成更多样描述.

采用蒙特卡洛法将句子补充完整，后交给鉴别器进行打分获得奖励，根据得到的奖励进行生成器梯度调整完成生成器G的优化.

$$V_{\theta,\eta}\left(I, Z, S_{1:t}\right) = E_{S_{t+1:T}\sim G_\theta(I,Z)}\left[\gamma_\eta\left(I, S_{1:t} \oplus S_{1+t:T}\right)\right] \tag{12}$$

$$\tilde{E}\left[\sum_{t=1}^{T_{\max}}\sum_{w_t \in V}\nabla_\theta \pi_\theta\left(w_t | I, Z, S_{1:t-1}\right).V_{\theta',\psi}\left(I, Z, S_{1:t} \oplus W_t\right)\right] \tag{13}$$

其中，γ_η函数就是用来返回奖励的鉴别器，π函数是强化学习中的策略，而I、Z、$S_{1:t-1}$作为条件输入到生成器预测词的出现概率.

2.5 基于新融合数据集的研究

无监督学习法首次出现在图像描述是2019年由腾讯论文[56]提出的. 数据集由传统数据集和网上爬取200万文本数据集组成，利用语义对齐预处理得到新融合数据集. 该文编码端使用经过微调的inception v4提取图像特征，解码端均采用LSTM的CGAN网络. 同年Bhargava等人[57]提出在数据上消除偏见的技术，训练了一个性别中立的图片配图模型，它基于性别产生语言模型的偏见，并给出高质量配图. 为了在标题中注入性别，只用包含人的裁剪部分图像训练性别分类器，利用RPN模型得到三元组，输入到LSTM预测词.

通过融合其他领域的文本数据集来提高图像描述独特性，如图8所示. 2019年Shuster等人[58]提出在理解图像内容的同时生成具有吸引力的描述，解决其他模

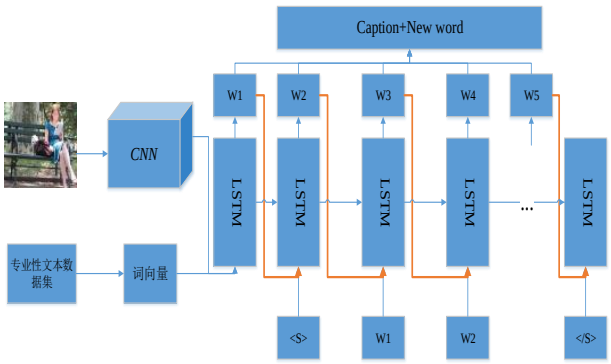

图 8　新融合数据集框架图

型无吸引力的问题,新数据集融入了个性化、有吸引力的词组,从源头提高生成描述质量. 而 Kim 等人[59]提出多任务三元组网络模型,解决了通过寻找每对 object 之间关系时因 object 独立导致生成描述语义独特性丧失的问题. Biten 等人[60]提出在传统数据集上融合新闻数据集的三等级描述模型,从图像的目标枚举、基本描述和描述解释三处构建,解决了专业词汇缺乏的问题. Guo 等人[61]提出多风格描述模型,通过对抗网络解决无法同时生成多风格描述的问题. 2020 年 Sidorov 等人[62]为研究如何在图像背景下理解文本,收集了一个新数据集 TextCaps,解决了无法将书面文本出现在图像描述的问题,它向旧数据集提出许多新的技术问题. 同年 Tran 等人[63]为新闻文章中嵌入图像描述,提出通过多模式、多头注意将新闻描述中的词与图像中的对象相关联来解决依赖现实世界中有关命名实体的知识.

2.6　本章小结

本文总结与分析了在图像描述任务中图像处理模型和文本生成模型的各类研究. 对近五年在 Encoder-Decoder 架构、注意力机制、强化学习、生成对抗网络和新融合数据集方面的研究进行了详细归纳和分析. 针对 1.1~1.5 小节中不同分类下各类图像描述模型,在三类数据集上对各类模型评分进行比较,,如表 1 所示. 以上所述就是图像描述近五年顶会上有代表性的模型,而所有模型的最终目的都是希望生成多样化、独特性接近甚至超越人类的描述.

表 1 说明端对端框架中效果最好的是 Stack-Cap 模型,在注意力中效果最好的是 Up-Down Attention 模型,在强化学习中效果最好的是 PG-SPIDEr 模型,在生成对抗网络中效果最好的是 base+CL 模型,而新融合数据中效果最好的是 M4C-Captioner 模型. 未来图像描述是在端对端框架下用自适应注意力和多条件对抗网络做出研究. 目前自适应注意力没有源码,但未来应用意义巨大. 表 2 展示了表 1 中不同模型的优缺点,未来研究者可以针对优缺点继续深入研究,通过表 1 和表 2 说明近两年大部分研究者向融合专业数据集方向进行研究,本文作者也会关注融合数据集技术.

表 1　不同模型、数据集下的评分对比

			不同数据集下的评价得分		
技术分类	模型称		Flickr 8k 数据集	Flickr 30k 数据集	MSCOCO
端对端框架	首用	NIC	R@1:20 R@10:61 Med r:6	R@1:17 R@10:56 Med r:7	B4:27.7 M:23.7 C:85.5
		BRNN	B1:57.9 B2:38 B3:24.5 B4:16	B1:57 B2:37 B3:24 B4:15.7	B1:62 B2:45 B3:32 B4:23 M:19 C:66
	编码端	Up-Down	Flickr 8k 和 Flickr 8k 无评价	B1:80.2 B2:64.1 B3:49.1 B4:36.9 M:27.6 R:57.1 C:117.9 S:21.5	
		VS-LSTM	B1:78.9 B2:63.4 B3:48.1 B4:36.3 M:27.3 C:120.8		B1:79 B2:63 B3:48 B4:35.9 M:27 R:56.5 C:116
	解码端	CNN+Attn	Flickr 8k 和 Flickr 8k 无评价	B1:71.5 B2:54.5 B3:40.8 B4:30.4 M:24.6 R:52.5 C:91	
		NBT	B1:69.0 B4:27.1 M:21.7 C:57.5 S:15.6		B1:75.5 B4:34.7 M:27.1 C:107.2 S:20.1
		Entity-aware	Flickr 8k 和 Flickr 8k 无评价	B1:25.5 B2:14.9 B3:8.0 B4:4.7 M:11.0 R:21.1 C:29.9 F:39.7 H:0.87	
		GroupCap	Flickr 8k 和 Flickr 8k 无评价	B1:72.9 B2:56.5 B3:42.5 B4:31.6 M:25.8 R:54.5 C:101.9	
		one hot+Glove	Flickr 8k 和 Flickr 8k 无评价	F(bottle):29.6 F(bus):74 F(couch):38 F(pizza):68 F(average):55.66 M:23	
		SemStyle-coco	Flickr 8k 和 Flickr 8k 无评价	B1:0.653 B4:0.238 M:0.219 C:0.769 S:0.157 C:0.003 L:6.905 G:6.691	
		Stack-Cap	Flickr 8k 和 Flickr 8k 无评价	B1:93.2 B2:86.1 B3:76.0 B4:64.6 M:35.6 R:70.6 C:118.3 S:20.9	
注意力机制	Up-Down Attention		Flickr 8k 和 Flickr 8k 无评价	B1:95.2 B2:88.8 B3:79.4 B4:68.5 M:36.7 R:72.4 C:120.5 S:71.5	
	Adaptive		B1:0.677 B2:0.494 B3:0.354 B4:0.251 M:0.204 C:0.531		B1:0.74 B2:0.58 B3:0.44 B4:0.33 M:0.27 C:1.085
	Hard-Attention		B1:67 B2:45 B3:31 B4:21 M:20	B1:66 B2:43 B3:29 B4:19 M:18.4	B1:71.8 B2:50.4 B3:35.7 B4:25.0 M:23.04

续表

		不同数据集下的评价得分		
技术分类	模型称	Flickr 8k数据集	Flickr 30k数据集	MSCOCO
	SimNet	S:0.160 C:0.585 M:0.221 R:0.489 B4:0.251		S:0.220 C:1.135 M:0.283 R:0.564 B4:0.332
	Att-KB+LSTM	COCO-QA数据集:Acc:67.66 WUPS@0.9:75.76 WUPS@0.0:93.63		B1:0.8 B2:0.64 B3:0.5 B4:0.4 M:0.28 C:1.07 P:9.6
对抗网络	G+MLE	B3:37 B4:30 M:21 R:47 C:76 S: E-NGAN:46 E-GAN:43		B3:0.4 B4:0.3 M:0.2 R:0.5 C:1.0 S:0.2 E-NGAN:0.4 E-GAN:0.42
	Tgt	Flickr 8k和Flickr 8k无评价		B1:53.4 B2:39.8 B3:30.7 B4:24.5 R:50.7 M:23.2
	Base+CL	Flickr 8k和Flickr 8k无评价		B1:75.5 B2:59.8 B3:46.0 B4:35.3 R:55.9 M:27.1 C:114.2
	Adv-samp	Flickr 8k和Flickr 8k无评价	M:27.2 S:18.7 S(color):10.1 S(Attribute):8.5 S(object):34.5 S(relaton):4.9 S(count):2.5	
强化学习	Full-model	Flickr 8k和Flickr 8k无评价		B1:71.3 B2:53.9 B3:40.3 B4:30.4 R:52.5 M:25.1 C:93.7
	B-SCST	Flickr 8k和Flickr 8k无评价		B1:72.7 B4:29.6 M:22.6 R:50.6 C:67.0 S:16.4
	PG-SPIDEr	Flickr 8k和Flickr 8k无评价		B1:74.3 B2:57.8 B3:43.3 B4:32.2 R:54.4 M:25.1 C:100.0
新融合数据	MTTSNet	Flickr 8k和Flickr 8k无评价		R@1:0.29 R@1:0.60 R@1:0.73 Med:4
	M4C-Captioner	Flickr 8k和Flickr 8k无评价		B4:18.9 M:19.8 R:43.2 S:12.8 C:81.0 H:3.0
	Transformer +OA	Flickr 8k和Flickr 8k无评价		B4:6.30 R:21.7 C:54.4 Named entities(P:24.6 R:22.2)
	Unsupervised	Flickr 8k和Flickr 8k无评价		B1:58.9 B2:40.3 B3:27 B4:18.6 R:43.1 M:17.9 C:54.9 S:11.1

其中评分字符表示：B:BLEU,M:METEOR,C:CIDER,S:SPICE,R:ROUGE,A:Accuracy,H:Human,F:F1,C:CLF,L:LM,G:GRULM

表2 不同模型优缺点对比

		算法	年份	优点	缺点
端对端框架	首用	NIC	2016	模型设计简单但取得还不错的结果	没考虑图像细微特征之间的关系
		BRNN	2016	在全帧和区域级别情况下多峰RNN都优于检索基准	生成的主题只是区域的而不是整个图像
	编码端	Up-Down	2018	能够一次考虑与一个对象有关的所有信息	训练参数增加了导致训练量明显变大了
		VS-LSTM	2018	引入状态扰动探索在频繁和不太频繁单词中适当的词汇	统一网格输出的均匀网格模糊了边界导致对物体的误识别
	解码端	one hot+Glove	2017	不需要巨量的数据集一样可以到一样评分的结果	—
		CNN+Attn	2018	训练每个参数所需要的时间比LSTM的方法要好	无法考虑到语句上下文信息
		NBT	2018	对文本赋予不同权重抽取出更关键信息而不会增加模型计算量	训练集主题长度小于17
		Entity-aware	2018	使用细粒度来命名插槽可以在生成主题中体现特殊信息	填充实体的关系错误、生成的Template 错误
		GroupCap	2018	更好的捕获并准确的解释群组间图像的关联性	每次参数的优化都需要过一遍整个数据集
		SemStyle-coco	2018	模型可以生成具有语言风格的主题	外部信息量和寻找与图像相关文本处理工作量比较大
		Stack-Cap	2018	模型产生出越来越精细的描述	通过传统贪心解码获得描述增加计算量

续表

	算法	年份	优点	缺点
注意力机制	Hard-Attention	2015	首次在图像主题领域使用注意力提高生成主题的质量	给图像区域确定权重时比较死板
	Att-KB+LSTM	2016	能够获取到图像的高层次的抽象信息	词表量不大导致一些词无法表达
	Adaptive	2017	通过改变空间注意力从时间的维度来决定什么时间看、看多少	相近意义的同源词的概率却相差很大
	Up-Down Attention	2018	候选注意区域与对象相关视觉概念同一位置一起处理	模型会过分依赖句子前后信息误导图像识别
	SimNet	2018	引入逐步合并机制使生成的字幕既详尽又全面	许多对象和模型无法掌握前景和背景的关系
生成对抗网络	G+MLE	2017	首次用条件对抗网络提高生成主题的多样性	对抗网络生成器和鉴别器训练循序和次数难控制
	Base+CL	2017	同一类的图片能够生成有差异但语义相似的描述	只有正样本模型表现只提升了一点,只有负样本模型表现大幅下降
	Adv-samp	2017	能够降低字幕任务固有的含糊性	—
	Tgt	2018	不同的生成模型能够生成语义相似的句子,具有很强的迁移性	无法生成被动语态的目标主题
强化学习	state-of-the-art	2017	模型能够更加强大对政策错误保持稳定	顺序单词生成器提供的置信度仅考虑本地信息
	PG-SPIDEr	2017	融入了强化学习的专门用于评价图像主题的方法	训练收敛有点慢
	Full-model	2017	政策网络做本地指南价值网络做全局和前瞻性指南性能最优	光束尺寸对SL的影响相对较大,模型对光束大小不太敏感
	B-SCST	2020	首个使用策略梯度的强化学习来训练技术的贝叶斯变体	SoftMax概率可能是过度自信而不是很好的预测信心的衡量
新融合数据	MTTSNet	2019	建立的关系字幕在多样性和信息量方面具有优势	对场景图生成或VRD模型无法执行
	Unsupervised	2019	首次用无监督学习来解决图像主题的问题	使用图像区域和重文本作为条件条件对抗网络结果改变不大
	M4C-Captioner	2020	无论在新数据集还是COCO图像上都能够生成令人印象深刻字幕	数据需生成较长的句子且OCR和词汇标记之间进行的许多切换
	Transformer +OA	2020	在任何标题中都没有出现但出现在描述上下文中的专有名词	模型生成字幕和人工字幕之间在TTR和长度上仍然存在差距

3　部分图像描述模型实验

本实验模型来自NIC模型、Hard-Attention模型和Neural Talk模型,三类模型在MS-COCO、Flickr30k和Flickr8k数据集上分别训练,从BLEU1、BLEU2、BLEU3、BLEU4四处平均评分来展现三个模型实际对比效果,如图9所示.

近五年的论文中程序几乎是以NIC模型为基础的,本文也做了实验.实验结果具有语句单一且多样性不足等缺点,在2018—2020年间的研究是追求生成描述多样性、独特性.目前生成描述的长度都控制在20字符内,不然效果会变差.本实验中NIC模型采用多线程训练数据集,但在评分上Hard-Attention模型效果更好.本实验采用CPU版TensorFlow框架训练各类数据集.

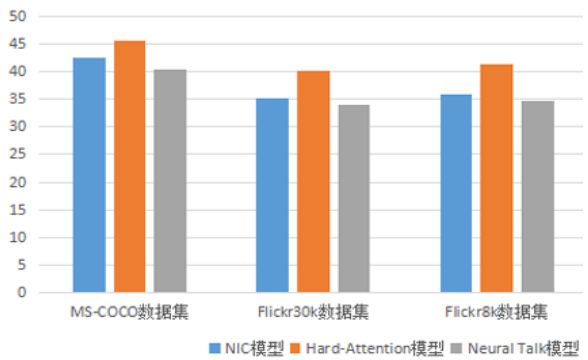

图9　实验评分对比图

本实验用两处图像验证模型,其验证图的实验效果如图10所示,其网图实验效果如图11所示,说明数据集

不能涵盖现实所有的语句,才导致此结果.

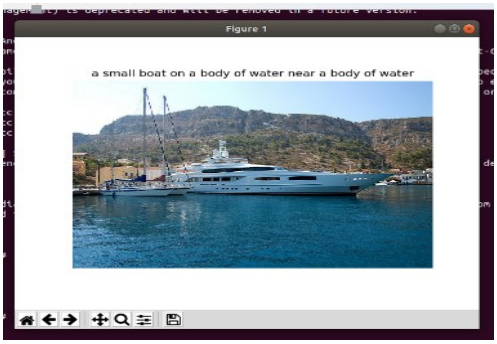

图10 验证集结果图

图11 随机网图结果图

4 图像描述数据集和评价指标

4.1 数据集

目前图像描述领域使用的公共数据集有MS-COCO2017、MS-COCO2014、Flickr30k和Flickr8k四类数据集,如表3所示. 其中MS-COCO2014数据集不仅包含大量的图像还有足够科学的标注,它是微软花了巨大的人力和财力而获得的数据集,到2017年在MS-COCO2014版数据集基础上出现了MS-COCO2017版数据集,它在保持2014版结构的基础上只是扩大了数据集的数量,此数据集一直沿用到2020年,一直作为图像描述的主训练数据集,MS-COCO数据集的制作目的是服务于整个计算机视觉领域. 图像描述领域以前使用Flickr8k、Flickr30k两个数据集,由于这两个数据集的数量不足,导致模型训练不好,所以目前它们主要用于数据集对比.

表3 不同数据集分类

数据集名称	数量/张		
	训练集	测试集	验证集
MS-COCO2017	118287	50694	5000
MS-COCO2014	82783	40775	5000
Flickr30k	28000	1000	1000
Flickr8k	6000	1000	1000

4.2 评价指标

目前图像描述常用的评价指标有BLEU、ROUGE、Meteor、CIDEr和SPICE,如表4所示. 图像描述最初没有自己评价方法,引用机器翻译常用的评价方法BLEU和Meteor,BLEU是计算精确率,而Meteor是计算序列匹配、同义词、词根词缀之间的匹配度指标. 看到自动文摘取得了巨大的成功,有学者将自动文摘评价方法ROUGE引入图像描述,ROUGE是计算召回率. 为了解决语句语义相似性,出现了图像描述自己评价方法CIDEr,它是计算语句语义的余弦相似性. SPICE是基于图的语义表示,利用PCFG和规则专门评价图像描述的方法.

表4 常用评价方法

用途分类	评价方法	功能
机器翻译	BLEU	计算候选译文与参考译文中N元组共同出现的程度
	Meteor	计算特定序列匹配,释义同义词、词根和词缀之间的关系
自动文摘	ROUGE	计算生成结果的召回率
图像主题	CIDER	计算参考主题与生成主题的余弦相似度
	SPICE	计算生成主题中对象、属性和语义的F-score值

在2019年Levinboim等人[64]提出图像配图质量估计模型QE,它在不依赖基本事实情况下对描述质量进行建模,实验表明此方法可行. 2020年Xie等人[65]提出直接评估生成字幕的语法性、真实性和多样性模型GTD,它提供了对模型能力和限制洞察的方法,以补充目前匮乏的标准评估. 在未来还需要继续研究更适合、更优秀的图像描述评价方法,使生成描述更公正、更合理.

5 总结与展望

图像描述是一个非常复杂但又非常重要的领域,也是自然语言处理与计算机视觉交叉的领域,它在现实中具有很好的应用前景. 本文将全球近五年基于深度学习的大部分顶会上的研究划分为基于Encoder-Decoder框架、基于注意力机制、基于强化学习、基于生成对抗网络和基于新融合数据集五大类,对这五大类不同模型的具体创新研究进行详细介绍与深入分析. 本文还介绍了目前图像描述领域几类数据集存在的优缺点和用途,同时也介绍了图像描述领域几大常用评价方法和新出评价指标的研究. 最后本文也挑选了三个比较基础但有影响的模型进行了实验并进行了生成描

述评分对比. 目前图像描述领域技术应用到现实中还比较有难度,例如现有模型在公共数据集Blue4上的评分还没超过45分,像一些特定专业领域的名词、动词等在描述中还体现不出,并且还存在庞大的模型训练、大数据存储等硬件问题. 未来图像描述模型在以下几个方面值得进一步深入研究.

(1)充分利用好图像的标注描述. 在提取到图像基本对象特征后,通过生成描述不断与标注描述对抗训练来降低损失使趋于收敛,促使生成描述与标注描述之间最大化评分. 因为标注描述是需要庞大MS-COCO数据集来训练没见过或者无覆盖的词,应充分利用好该数据集来之不易的庞大标注描述.

(2)研究最适合图像描述的评价算法. 2020年图像描述领域的顶级论文中出现几篇评价图像描述结果的模型,说明对此领域的评价方法一直在研究. 评价结果时参考标注描述和重要区域图像特征的语义做综合对比,这样评价结果才能从图像-文本语义处真正衡量生成描述与图像之间的差异.

(3)图像描述的终极目标是生成描述能够完全表达图像中各个对象在特定背景下的相互逻辑语义关系. 比如一个男人和一个女人带着一个小孩在走路,目标是能够识别男人、女人和小孩,接着通过语义逻辑输出短语是“幸福的一家三口在散步”,而不是输出短语是“街边站着一群人”,只能简单语义的表述,缺少进一步高层次语义归纳、提升. 高层次逻辑语义能够实现的话可以真正做到超越标注描述.

(4)能够利用外部专业性句子的数据集去修复目前生成的描述. 使用多种类型的专业性句子的数据集做成专业性大型数据集,生成描述需要具体某个专业词汇时就调用此类型的辅助数据集来做修复训练,这个方面的研究不仅可以提高生成描述的独特性,而且还可以为生成的描述组成提供新词汇.

(5)未来图像描述模型希望做到自学习方式,利用少量数据集训练出能够生出非常简单的描述,再把此模型使用不同方面、不同时段的在线网络文本中出现句子来训练生成模型,直到生成模型的损失值趋于收敛到极小停止.

参考文献

[1] 权宇, 李志欣, 张灿龙, 等. 融合深度扩张网络和轻量化网络的目标检测模型[J]. 电子学报, 2020, 48(2): 390 – 397.
Quan Y, Li Z X, Zhang C L, et al. Fusing deep dilated convolutions network and light-weight network for object detection[J]. Acta Electronica Sinica, 2020, 48(2): 390 – 397. (in Chinese)

[2] 刘颖, 刘红燕, 范九伦, 等. 基于深度学习的小目标检测研究与应用综述[J]. 电子学报, 2020, 48(3): 590 – 601.
Liu Y, Liu H Y, Fan J L, et al. A survey of research and application of small object detection based on deep learning [J]. Acta Electronica Sinica, 2020, 48(3): 590 – 601. (in Chinese)

[3] 杨KL. Image caption的发展历程和最新工作的简要综述(2010−2018) [EB/OL]. https://blog. csdn. net/qq_41533506/article/details/84671195, 2018-12-01.

[4] Vinyals O, Toshev A, Bengio S, et al. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 652 – 663.

[5] Tan X, Ren Y, He D, et al. Multilingual neural machine translation with knowledge distillation[EB/OL]. https://arxiv.org/abs/1902.10461v3, 2019.

[6] 亲历者. show_and_tell代码实现及测试——批量训练[EB/OL]. https://blog.csdn.net/m0_38073193/article/details/82502063, 2018-09-07.

[7] Karpathy A, Li F F. Deep visual-semantic alignments for generating image descriptions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 664 – 676.

[8] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. https://arxiv.org/abs/1409.1556v4, 2014.

[9] Fang H, Gupta S, Iandola F, et al. From captions to visual concepts and back[A]. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [C]. Boston, MA, USA: IEEE, 2015. 1473 – 1482.

[10] Li N, Chen Z. Image cationing with visual-semantic LSTM[A]. Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence[C]. Shanghai, China: IJCAI, 2018. 793 – 799.

[11] Anderson P, He X D, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering[A]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition[C]. Salt Lake City, UT, USA: IEEE, 2018. 6077 – 6086.

[12] Fei Z C. Better understanding hierarchical visual relationship for image caption[EB/OL]. https://www.researchgate.net/publication/337756448_Better_Understanding_Hierarchical_Visual_Relationship_for_Image_Caption, 2019.

[13] Lee K H, Palangi H, Chen X, et al. Learning visual relation priors for image-text matching and image captioning with neural scene graph generators[EB/OL]. https://arxiv.

org/abs/1909.09953v1, 2019.

[14] Yao T, Pan Y W, Li Y H, et al. Hierarchy parsing for image captioning[A]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) [C]. Seoul, Korea (South): IEEE, 2019. 2621 – 2629.

[15] He S, Liao W T, Tavakoli H R, et al. Image captioning through image transformer[A]. Computer Vision – ACCV 2020[M]. Cham, Germany: Springer International Publishing, 2021. 153 – 169.

[16] 张红斌, 蒋子良, 熊其鹏, 等. 基于改进的有效区域基因选择与跨模态语义挖掘的图像属性标注[J]. 电子学报, 2020, 48(4): 790 – 799.
Zhang H B, Jiang Z L, Xiong Q P, et al. Image attribute annotation via a modified effective range based gene selection and cross-modal semantics mining[J]. Acta Electronica Sinica, 2020, 48(4): 790 – 799. (in Chinese)

[17] Chen F H, Ji R R, Sun X S, et al. GroupCap: group-based image captioning with structured relevance and diversity constraints[A]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition[C]. Salt Lake City, UT, USA: IEEE, 2018. 1345 – 1353.

[18] Pasunuru R, Bansal M. Multi-task video captioning with video and entailment generation[A]. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)[C]. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017. 1273 – 1283.

[19] Zhou L W, Palangi H, Zhang L, et al. Unified vision-language pre-training for image captioning and VQA[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 13041 – 13049.

[20] Wang Y F, Lin Z, Shen X H, et al. Skeleton key: Image captioning by skeleton-attribute decomposition[A]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [C]. Honolulu, HI, USA: IEEE, 2017. 7378 – 7387.

[21] Lu J S, Yang J W, Batra D, et al. Neural baby talk[A]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition[C]. Salt Lake City, UT, USA: IEEE, 2018. 7219 – 7228.

[22] Mathews A, Xie L X, He X M. SemStyle: learning to generate stylised image captions using unaligned text[A]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition[C]. Salt Lake City, UT, USA: IEEE, 2018. 8591 – 8600.

[23] Wang L, Bai Z C, Zhang Y H, et al. Show, recall, and tell: Image captioning with recall mechanism[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12176 – 12183.

[24] Nguyen G, Jun T J, Tran T, et al. ContCap: A comprehensive framework for continual image captioning[EB/OL]. https://onikle.com/articles/14900, 2019.

[25] Gu J X, Cai J F, Wang G, et al. Stack-captioning: Coarse-to-fine learning for image captioning[EB/OL]. https://arxiv.org/abs/1709.03376, 2018.

[26] Fei Z C. Fast image caption generation with position alignment[EB/OL]. https://arxiv.org/abs/1912.06365, 2019.

[27] Thapliyal A V, Soricut R. Cross-modal language generation using pivot stabilization for Web-scale language coverage[A]. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics[C]. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020. 160 – 170.

[28] Xia Q L, Huang H Y, Duan N, et al. XGPT: cross-modal generative pre-training for image captioning[EB/OL]. https://arxiv.org/abs/2003.01473, 2020.

[29] Yao T, Pan Y, Li Y,et al. Incorporating copying mechanism in image captioning for learning novel objects[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. New York, USA: IEEE, 2017. 6580 – 6588.

[30] Aneja J, Deshpande A, Schwing A G. Convolutional image captioning[A]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition[C]. Salt Lake City, UT, USA, IEEE, 2018: 5561 – 5570.

[31] Nikolaus M, Abdou M, Lamm M, et al. Compositional generalization in image captioning[A]. Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)[C]. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019. 87 – 98.

[32] Lu D, Whitehead S, Huang L F, et al. Entity-aware image caption generation[A]. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing[C]. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018. 4013 – 4023.

[33] Sammani F, Elsayed M. Look and modify: Modification networks for image captioning[EB/OL]. https://arxiv.org/abs/1909.03169v1, 2019.

[34] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention[EB/OL]. https://arxiv.org/abs/1502.03044, 2015.

[35] Lu J S, Xiong C M, Parikh D, et al. Knowing when to

look: Adaptive attention via a visual sentinel for image captioning[A]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)[C]. Honolulu, HI, USA: IEEE, 2017. 3242 – 3250.

[36] Huang L, Wang W M, Xia Y X, et al. Adaptively aligned image captioning via adaptive attention time[EB/OL]. https://arxiv.org/abs/1909.09060, 2019.

[37] Goel A, Fernando B, Nguyen T S, et al. Learning to caption images with two-stream attention and sentence auto-encoder[EB/OL]. https://arxiv.org/abs/1911.10082, 2019.

[38] Guo L T, Liu J, Zhu X X, et al. Normalized and geometry-aware self-attention network for image captioning[A]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)[C]. Seattle, WA, USA: IEEE, 2020. 10324 – 10333.

[39] Pan Y W, Yao T, Li Y H, et al. X-linear attention networks for image captioning[A]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)[C]. Seattle, WA, USA: IEEE, 2020. 10968 – 10977.

[40] Wu Q, Shen C H, Liu L Q, et al. What value do explicit high level concepts have in vision to language problems? [A]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)[C]. Las Vegas, NV, USA: IEEE, 2016. 203 – 212.

[41] Sun J M, Lapuschkin S, Samek W, et al. Understanding image captioning models beyond visualizing attention[EB/OL]. https://arxiv.org/abs/2001.01037, 2020.

[42] Liu F L, Ren X C, Liu Y X, et al. simNet: stepwise image-topic merging network for generating detailed and comprehensive image captions[A]. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing[C]. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018. 137 – 149.

[43] Ranzato M, Chopra S, Auli M, et al. Sequence level training with recurrent neural networks[EB/OL]. https://arxiv.org/abs/1511.06732, 2015.

[44] Liu S Q, Zhu Z H, Ye N, et al. Improved image captioning via policy gradient optimization of SPIDEr[A]. 2017 IEEE International Conference on Computer Vision (ICCV)[C]. Venice, Italy: IEEE, 2017. 873 – 881.

[45] Ren Z, Wang X Y, Zhang N, et al. Deep reinforcement learning-based image captioning with embedding reward [A]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)[C]. Honolulu, HI, USA: IEEE, 2017. 1151 – 1159.

[46] Seo P H, Sharma P, Levinboim T, et al. Reinforcing an image caption generator using off-line human feedback [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(3): 2693 – 2700.

[47] Rennie S J, Marcheret E, Mroueh Y, et al. Self-critical sequence training for image captioning[A]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)[C]. Honolulu, HI, USA: IEEE, 2017. 1179 – 1195.

[48] Pasunuru R, Bansal M. Reinforced video captioning with entailment rewards[A]. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing[C]. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017. 979 – 985.

[49] Dai B, Fidler S, Urtasun R, et al. Towards diverse and natural image descriptions via a conditional GAN[A]. 2017 IEEE International Conference on Computer Vision (ICCV)[C]. Venice, Italy: IEEE, 2017. 2989 – 2998.

[50] Shetty R, Rohrbach M, Hendricks L A, et al. Speaking the same language: Matching machine to human captions by adversarial training[A]. 2017 IEEE International Conference on Computer Vision (ICCV)[C]. Venice, Italy: IEEE, 2017. 4155 – 4164.

[51] Liu B C, Song K P, Zhu Y Z, et al. TIME: text and image mutual-translation adversarial networks[EB/OL]. https://arxiv.org/abs/2005.13192, 2020.

[52] Li N N, Chen Z Z. Learning compact reward for image captioning[EB/OL]. https://arxiv.org/abs/2003.10925, 2020.

[53] Chen H G, Zhang H, Chen P Y, et al. Attacking visual language grounding with adversarial examples: A case study on neural image captioning[A]. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)[C]. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018. 2587 – 2597.

[54] Shekhar R, Pezzelle S, Klimovich Y, et al. FOIL it! Find One mismatch between Image and Language caption[A]. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)[C]. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017. 255 – 265.

[55] Dai B, Lin D H. Contrastive learning for image captioning [EB/OL]. https://arxiv.org/abs/1710.02534, 2017.

[56] Feng Y, Ma L, Liu W, et al. Unsupervised image captioning[A]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)[C]. Long Beach, CA,

USA: IEEE, 2019. 4120 – 4129.
[57] Bhargava S, Forsyth D. Exposing and correcting the gender bias in image captioning datasets and models[EB/OL]. https://arxiv.org/abs/1912.00578, 2019.
[58] Shuster K, Humeau S, Hu H X, et al. Engaging image captioning via personality[A]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)[C]. Long Beach, CA, USA: IEEE, 2019. 12508 – 12518.
[59] Kim D J, Choi J, Oh T H, et al. Dense relational captioning: Triple-stream networks for relationship-based captioning[A]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)[C]. Long Beach, CA, USA: IEEE, 2019. 6264 – 6273.
[60] Biten A F, Gomez L, Rusiñol M, et al. Good news, everyone! context driven entity-aware captioning for news images[A]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)[C]. Long Beach, CA, USA: IEEE, 2019. 12458 – 12467.
[61] Guo L T, Liu J, Yao P, et al. MSCap: multi-style image captioning with unpaired stylized text[A]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)[C]. Long Beach, CA, USA: IEEE, 2019. 4199 – 4208.
[62] Sidorov O, Hu R H, Rohrbach M, et al. TextCaps: A dataset for image captioning with reading comprehension[A]. Computer Vision – ECCV 2020[M]. Cham, Germany: Springer International Publishing, 2020. 742 – 758.
[63] Tran A, Mathews A, Xie L X. Transform and tell: Entity-aware news image captioning[EB/OL]. https://arxiv.org/abs/2004.08070, 2020.
[64] Levinboim T, Thapliyal A V, Sharma P, et al. Quality estimation for image captions based on large-scale human evaluations[A]. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies[C]. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021. 3157 – 3166.
[65] Xie H Y, Sherborne T, Kuhnle A, et al. Going beneath the surface: Evaluating image captioning for grammaticality, truthfulness and diversity[EB/OL]. https://arxiv.org/abs/1912.08960, 2019.

作者简介

石义乐 男,1994 年生,河南洛阳人.现为新疆大学软件学院研究生.主要研究方向为图像理解.
E-mail:2229842870@qq.com

杨文忠(通信作者) 男,1971 年生,新疆乌鲁木齐人.2011 年于武汉大学获得博士学位.现为新疆大学信息科学与工程学院研究生导师,副教授.主要研究方向为网络空间安全、机器学习和算法设计与分析.
E-mail:ywz_xy@163.com