计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200500039

图像描述技术综述

苗　益[1]　赵增顺[1,2,3]　杨雨露[1]　徐　宁[1]　杨皓然[1]　孙　骞[1]
1 山东科技大学电子信息工程学院　山东 青岛 266590
2 山东大学控制科学与工程学院　济南 250061
3 佛罗里达大学电子与计算机工程系　佛罗里达州 盖恩斯维尔 32611
(617544375@qq.com)

摘　要　图像描述技术,就是以图像为输入,通过数学模型和计算使计算机输出对应图像的自然语言描述文字,使计算机拥有“看图说话”的能力,是图像处理领域中继图像识别、图像分割和目标跟踪之后的又一新型任务。文中以图像描述技术的发展历程为主线,对图像描述任务的方法、评价指标和常用数据集进行了详细的综述。针对图像描述任务的技术方法,总结了基于模板、检索和深度学习的图像描述生成方法,重点介绍了基于深度学习的图像描述的多种方法,并对不同方法的实验结果进行了总结和讨论;详细介绍了图像描述任务的实验结果评价指标及其计算方法和该任务中常用的数据集;最后提出了该任务现有的问题和未来的发展方向。

关键词:图像处理;图像描述;深度学习;计算机视觉;自然语言处理
中图法分类号　TP301

Survey of Image Captioning Methods

MIAO Yi[1],ZHAO Zeng-shun[1,2,3],YANG Yu-lu[1],XU Ning[1],YANG Hao-ran[1] and SUN Qian[1]
1 College of Electronic and Information Engineering,Shandong University of Science and Technology,Qingdao,Shandong 266590,China
2 School of Control Science and Engineering,Shandong University,Jinan 250061,China
3 Department of Electrical & Computer Engineering,University of Florida,Gainesville,Florida 32611,USA

Abstract　Image captioning is a task that uses an image as input to generate the natural language description of this image by modeling and calculation,so that computers have the ability to “talk about the pictures”. It is another new type of computer vision task after image recognition,image segmentation and target tracking. This paper focuses on the development of image captioning and gives a detailed survey of the image captioning methods based on template,retrieval and deep learning. And this paper especially focuses on the deep learning-based methods and discusses the experimental results of various methods. Experimental evaluation indexes and the common datasets used in this field are introduced in detail. Finally,this paper points out the problems and research directions in the future.

Keywords　Image processing,Image captioning,Deep learning,Computer vision,Natural language processing

1　引言

随着“互联网 + ”技术的发展,网络成为人们日常生活不可或缺的一部分,网络信息包含了生活的方方面面,人们在网络中对情感的表达形式表现出多样化趋势。如今我们每天都在网络中共享大量的图像数据,一个大型的社交网站一天可以产生数亿级规模的图像数据,且这些数据不包含标注信息[1-3],如何管理数量如此庞大的数据,使人们快速地检索图像信息,以及如何整合网络上庞大的图像资源并为人们所用,发挥其巨大的价值,成为亟待解决的问题。

在日常生活中,人们可以将图像中的场景、色彩、逻辑关系等低层视觉特征信息自动建立关系,从而感知图像的高层语义信息,但是计算机作为工具只能提取到数字图像的低层数据特征[4-11],而无法像人类大脑一样生成高层语义信息,这就是计算机视觉中的“语义鸿沟”问题[12-13]。图像描述(字幕)技术(Image Caption Generation)的本质就是将计算机提取的图像视觉特征转化为高层语义信息,即解决“语义鸿沟”问题,使计算机生成与人类大脑理解相近的对图像的文字描述,从而可以对图像进行分类、检索、分析等处理任务[14-16](如图 1 所示)。图像描述技术结合了计算机视觉(Computer Vision,

到稿日期:2020-05-11　返修日期:2020-08-13　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61403281);中国博士后科学基金(2015T80717);山东省自然科学基金(ZR2014FM002)
This work was supported by the National Natural Science Foundation of China(61403281),China Postdoctoral Science Foundation(2015T80717) and Natural Science Foundation of Shandong Province,China(ZR2014FM002).
通信作者:赵增顺(zhaozengshun@163.com)

CV)和自然语言处理(Nature Language Process,NLP)两大人工智能领域,计算机不仅需要识别输入图像中的物体和物体的属性,还要识别出不同物体之间的相互关系,并用正确的自然语言表达出来,这也是该技术的难点所在。近两年,随着大型图像数据集的产生和深度学习的发展,图像描述成为计算机视觉和自然语言处理领域的热点[17-21]。

图 1 图像描述任务实例

Fig. 1 Examples of image captioning

深度学习(Deep Learning,DL)是机器学习领域中一个重要的分支,由 Hinton 等于 2006 年提出[22],它能够使机器像人一样具有学习和分析能力,对语音、图像和声音等数据的处理远超当前其他相关的技术。近年来,计算机运算能力的提高和深度学习算法的发展,使得舆情分析[23-24]、图像分类[5,7-9,25-27]、目标检测[10,28]、图像描述[29-33]等任务占据了人们的视野,并且成为计算机视觉和自然语言处理领域的热点。目前,国内外有关图像描述的综述文章并不多[34],并且大多没有对评价指标的计算方法进行详细总结。本文以图像描述技术的发展历程为主线,主要对以下几个方面进行了综述:1)图像描述任务的发展历程以及图像描述关键技术的类别和发展现状;2)图像描述的评价指标及其计算方法;3)图像描述任务中常用的数据集。

2 图像描述方法

在计算机视觉的发展初期,研究者们尝试利用计算机来模拟人类的视觉系统,并且让计算机告知人们它所看到的内容,这就是最初的图像识别任务[35-37]。在此之后,研究者们提出了更高的要求:让计算机既要识别出图像中的物体,又要对其进行分割并确定目标属性,甚至确定识别对象之间的关系,并且以自然语言的形式来描述图像内容。为此,图像描述任务应运而生,图像描述的相关方法也逐渐产生,且不断完善。

在最初的图像描述任务中,研究者们使用基于模板和检索的方法来使计算机生成图像描述。

2.1 基于模板的图像描述方法

在基于模板的图像描述方法中,生成的句子有固定的模板。通常使用语法决策树算法来构建数据模型,并基于视觉依存表来检测图像中的物体(object)、动作(action)、场景(scene)等相关元素,然后往模板中填充相应单词以组成完整的句子,如图 2 所示。Farhadi 等[38]于 2010 年提出了一种基于模板的图像描述方法,该方法通过支持向量机(Support Vector Machine,SVM)[39-40]来构建节点特征,检测图像中的对象、动作、场景 3 种元素,然后填充既定的模板来输出句子描述。由于当时并没有包含图像和对应描述以及表示语义空间标签的数据集,作者随机选用了 PASCAL 2008 数据集中[41]的 1 000 张图片,人工添加描述和标签,制作了自己的数据集。由于数据集的限制和基于模板算法的局限性,最终的效果并不是很好。

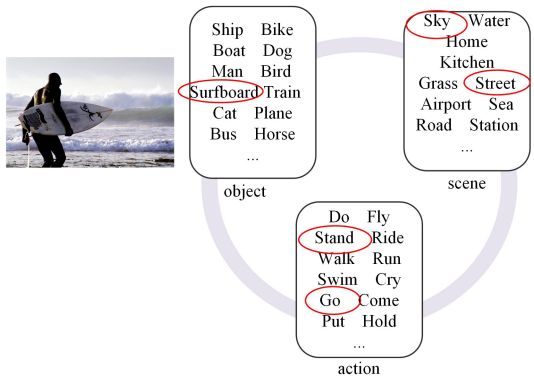


图 2 基于模板的图像描述

Fig. 2 Image captioning based on template

Li 等[42]在 2011 年使用全网域语言模型(Web-scale N-grams)提取与检测到的对象、属性、动作、关系等相关的短语来填充模板。Kulkarni 等[43]利用条件随机场(Conditional Random Field,CRF)[44-45]从大量视觉描述性文本池中提取的统计数据来平滑基于计算机视觉的检测和识别算法的输出,以确定用于描述图像内容的最佳单词,生成对特定图像内容更真实的描述。Mitchell 等[46]通过利用常用语法中的单词统计信息来处理模板填充过程,使用生成器对视觉系统输出的噪声进行过滤和约束来构建语法决策树[47],从而生成计算机视觉系统所见内容的详细描述。这种方法优于当时其他基于模板的图像描述方法,生成了较为自然的图像描述。基于模板的图像描述生成方法可以生成语法正确的描述,但是由于模板是固定的,生成的描述内容较为单一且长度不可变,还需要对图像的目标、属性、关系等进行大量标注,因此不能灵活地处理大规模图像数据。

2.2 基于检索的图像描述方法

基于检索的图像描述生成方法是将大量的图像描述储存在一个描述集合中,进行图像描述时,将待描述的图像与训练集中的图像进行比较,搜寻相似的图像,以该相似图像的图像描述为候选描述并进行适当修改,作为新图像的描述。Vicente 等[48]从网络中搜集大量图像并标注标题、描述等作为数据库,通过计算待描述图像与网络数据库图像的全局相似度,找到最相似的匹配图像,然后将描述从匹配图像转移到待描述图像。Socher 等[49]提出一种 Dependency Trees-RNN(DT-RNN)模型,该模型使用树结构将句子嵌入向量空间中,从单词顺序和句法表达的细节着眼于动作和主体,以检索由这些句子描述的图像,最终得到不错的效果。Kuznetsova

等[50]提出一种基于树的方法来构成图像描述，该方法是从网络上搜寻带有标题的图像，并从现有图像描述中获取表达性短语作为树片段，然后通过选择性地组合、提取的树片段来组成新的描述。Mason 等[51]提出一种用于图像标题生成的非参数密度估计技术（Nonparametric Density Estimation Technique），通过估计待描述图像的视觉内容的词频表示，将字幕生成转换为提取摘要问题。Sun 等[52]提出一种使用并行文本和视觉语料库的视觉概念自动发现算法，其根据待描述图像和图像库中相似图像的视觉特性过滤文本术语，使用双向图像和句子检索来为图像生成描述。基于检索的图像描述生成方法生成的图像描述能够贴近自然语言，但其过度依赖数据库，不易为特定图像生成描述。

以上就是早期生成图像描述的各种方法，这些方法过多地依赖前期的视觉处理过程，对于描述生成的模型优化有限，因此难以生成高质量的图像描述。

2.3　基于深度学习的图像描述方法

近年来，随着深度学习技术的不断发展，神经网络在计算机视觉和自然语言处理领域得到了广泛应用。受机器翻译领域中编码器-解码器（Encoder-Decoder）模型[53]的启发，图像描述可以通过端到端的学习方法直接实现图像和描述句子之间的映射，将图像描述过程转化成为图像到描述的“翻译”过程[54]。深度学习方法可以直接从大量数据中学习图像到描述语句的映射，生成更加准确的描述，其性能远远超过传统方法。

Kiros 等[55]在 2014 年首次使用神经网络来处理图像描述任务，打开了深度学习在图像描述领域的大门。他们将图像的不同区域以及其对应的文本映射到同一向量空间，使用深度神经网络和序列建模递归神经网络 LSTM 构建两种不同的多模态神经网络模型，结合单词和图像的语义信息来实现文本和图像的双向映射，不断融合语义信息来生成当前的单词，如图 3 所示。

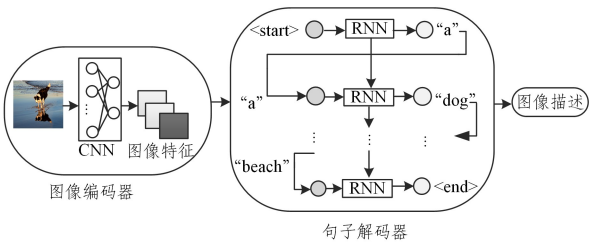


图 3　基于编码器-解码器的图像描述模型

Fig. 3　Encoder-decoder based image captioning model

基于深度学习的图像描述生成方法大多采用以 CNN-RNN 为基本模型的编码器-解码器框架[56]。CNN[5,7-9,11]通常包含卷积层、池化层和全连接层，它可以提取图像特征并通过矩阵转化将图像内容编码成固定长度的向量。CNN 决定了整个模型的图像识别能力，其最后的隐藏层的输出被用作解码器的输入。RNN[57-59]是用来读取编码后的图像并生成文本描述的网络模型，它能够训练输入序列在时间范围内的复杂动态，并能够使用内部存储单元记住或使用输入序列中的信息。Vinyals 等[31]推出的 NIC 模型使用 LSTM[60]代替 Mao 等[56]的模型中的 RNN。Jia 等[61]进一步改进了 NIC 模型，解决了其在生成描述句的过程中出现梯度消失的问题。此后，Mao 等[62]还使用基于候选框的 Region-CNN 方法为图像的特定区域生成描述。

2.3.1　基于注意力机制的方法

随着深度学习的发展，Bahdanau 等[63]在机器翻译领域提出注意力机制。注意力机制被广泛应用于计算机视觉领域[64-67]，其本质是为了解决编码器-解码器在处理固定长度向量时的局限性。注意力机制并不是将输入序列编码成一个固定向量，而是通过增加一个上下文向量来对每个时间步的输入进行解码，以增强图像区域和单词的相关性，从而获取更多的图像语义细节。Xu 等[32]将注意力机制融入编码器-解码器框架（见图 4）进行图像描述，使模型效果得到很大的提升。

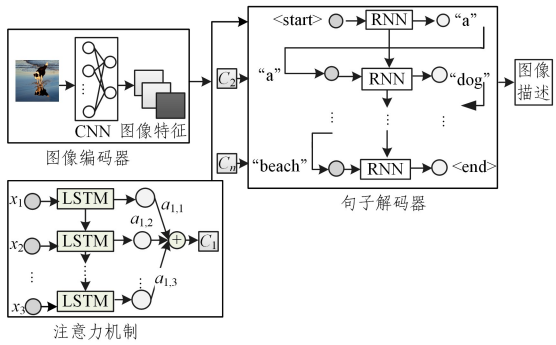


图 4　融入注意力机制的编码器-解码器图像描述模型

Fig. 4　Encoder-decoder image captioning model combined with attention mechanism

在注意力机制的基础上，Lu 等[15]提出了自适应注意力机制模型（Adaptive Attention），该方法提出了“视觉哨兵”（Visual Sentinel）的概念，即在 LSTM 的隐藏层加入一个视觉哨兵向量，用来控制对非视觉词的生成，如介词、量词等。该方法使模型不仅依赖于图像信息，还依赖于句子的语义信息，从而生成更加详细的描述句。Anderson 等[29]在注意力机制模型的基础上提出自上而下与自下而上相结合的注意力模型（Bottom-Up and Top-Down Attention），该方法使用 Faster R-CNN[68]来代替原有的 CNN 部分，过滤掉无用的图像特征，从而提高检测效率；在句子生成部分结合 2 个 LSTM 层来有选择性地提取图像局部特征。Yang 等[14]于 2019 年提出在注意力机制中嵌入场景图结构（Scene Graph Auto-Encoder，SGAE）作为句子重建网络，将图像通过检测网络得到的物体、属性、关系等生成场景图，再通过图卷积网络处理作为输入，送到已经利用 SGAE 结构预训练好的与编码器-解码器模型共享的字典当中，对产生的词向量进行转换重建，从而利用语料库实现了更加接近人类语言的图像描述。Zhou 等[69]在 Up-Down 注意力模型[29]的基础上融入图文匹配模型（Stacked Cross Attention Network，SCAN）[70]，对注意力机制的训练过程进行弱监督，并且使用自关键序列训练算法（Self-Critical sequence training，SCST）[71]对图像和文本的匹配程度进行强化学习，增强了注意力机制对单词和图像区域的对应能力，从而生成更加合理的描述。在风格化描述方面，Cornia 等[33]在自适应注意力模型的基础上，通过加入控制序列（或图像区域集合）来增强模型对输出的图像描述的可控性和

多样性。Chen 等[16]于 2020 年提出了一种抽象场景图(Scene Graph, ASG)的方法来控制图像不同细粒度的细节描述程度,如控制模型描述什么物体、是否描述属性和关系等,该方法在生成风格化图像描述方面获得了更好的效果。计算机视觉发展至今,注意力机制得到了广泛的研究和应用,在处理图像描述任务时,注意力机制可以直接通过上下文向量获取局部图像与全局信息的相关性,弥补了解码器端在处理长序列方面的不足,使模型获取到更多重要的图像信息。基于注意力机制的图像描述方法的准确率普遍高于其他方法。表 1 所列为上述方法在不同数据集上的效果对比。

表 1 基于编码器-解码器模型的图像描述方法在各数据集上的实验结果对比

Table 1 Performance of encoder-decoder based image captioning methods on different datasets

方法	数据集	评价指标					
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
Mao[56]	Flickr 8K	0.57	0.39	0.257	0.17	—	—
	Flickr 30K	0.60	0.41	0.28	0.19	—	—
	MS COCO	0.67	0.49	0.35	0.25	—	—
Jia[61]	Flickr 8K	0.65	0.46	0.32	0.22	0.20	—
	Flickr 30K	0.65	0.47	0.31	0.21	0.18	—
	MS COCO	0.47	0.49	0.36	0.26	0.23	—
Xu[32]	Flickr 8K	0.67	0.46	0.31	0.21	0.20	—
	Flickr 30K	0.67	0.44	0.30	0.20	0.18	—
	MS COCO	0.72	0.50	0.36	0.25	0.23	—
Lu[15]	Flickr 8K	—	—	—	—	—	—
	Flickr 30K	0.68	0.49	0.35	0.25	0.20	0.53
	MS COCO	0.75	0.58	0.44	0.34	0.26	1.04
Anderson[29]	Flickr 8K	—	—	—	—	—	—
	Flickr 30K	—	—	—	—	—	—
	MS COCO	0.80	0.64	0.49	0.37	0.28	1.18
Yang[14]	Flickr 8K	—	—	—	—	—	—
	Flickr 30K	—	—	—	—	—	—
	MS COCO	0.81	—	—	0.39	0.28	1.29
Zhou[69]	Flickr 8K	—	—	—	—	—	—
	Flickr 30K	0.734	—	—	0.301	0.226	0.693
	MS COCO	0.802	—	—	0.380	0.285	1.261
Chen[16]	Flickr 8K	—	—	—	—	—	—
	Flickr 30K	—	—	—	—	—	—
	MS COCO	—	—	—	0.230	0.245	2.042

表 1 中的数据均引自原文,虽然存在实验平台差异对实验结果的影响,但从表中可以看出,随着研究者们对编码器-解码器框架的不断改进以及注意力机制的融入,模型的各项评价指标在不断地提高,意味着模型生成的图像描述更加接近于自然人工描述。虽然目前基于深度学习的图像描述方法以注意力机制模型为主,但其他方法依然可以生成质量较好的图像描述。

2.3.2 基于生成对抗网络的方法

生成对抗网络(Generative Adversarial Networks, GANs)[72-73]是一种无监督的深度学习模型,近年来被广泛应用于人工智能领域,是目前最具有研究前景的方法之一。生成对抗网络模型中至少有两个模块:生成网络和判别网络。在训练过程中,生成网络生成尽量真实的数据以“欺骗”判别网络,并且通过判别网络的损失不断进行学习;而判别网络的任务就是区分生成的数据和真实数据。这两个网络通过动态的博弈学习,可以从无标签的数据中学习特征,从而生成新的数据。Dai 等[74]于 2017 年使用生成对抗网络通过控制随机噪声向量来生成多样化的描述。该模型分为两部分(如图 5 所示):第一部分是句子生成部分,在该部分中依然使用 CNN 来提取图像特征,使用 LSTM 来生成句子,区别是在生成单词时加入了随机噪声,并在描述句生成完成后将其输入到第二部分的判别器进行评估。第二部分用来做句子评估,使用 LSTM 对句子进行编码,与图像特征一起处理获得一个概率值,评估该描述句是否与人类描述相似,是否符合图像内容,最后使用策略梯度方法反向传播更新参数,使其获得最大的概率值,直到输出理想的描述句。

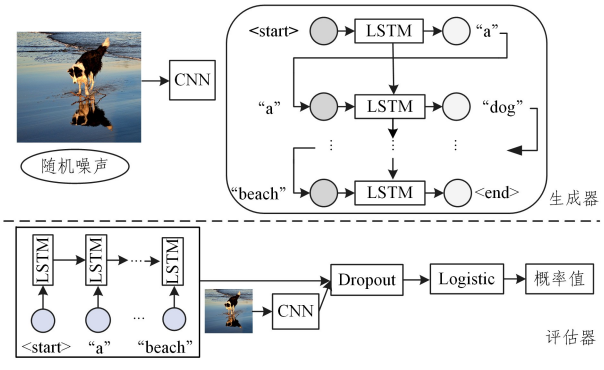


图 5 基于生成对抗网络的图像描述

Fig. 5 Image captioning based on generative adversarial network

Deshpande 等[30]于 2018 年指出,基于 GAN 的图像描述方法虽然可以实现描述语句的多样性,但准确率不足,并提出使用词性模板的方法来生成描述。该方法以词性作为标签并约束描述句,以推动描述句的生成。实验表明该方法在保证描述语句多样性的同时,可以提高模型的运行速度和准确率。Chen 等[75]提出一种称为 GroupCap 的基于组的图像描述方法,该方法通过一种视觉树解析器来构造单个图像的结构化语义相关性,并且利用树结构来计算图像之间的相关性和多

样性,最终将相关信息发送到 LSTM 生成器中以生成图像描述。Dognin 等[76]将图像描述看作条件式的对抗生成训练任务,同时提出了基于上下文的 LSTM 识别器和注意力判别器,并且在对抗网络的生成和判别过程中分别融入注意力机制,以增强图像与句子之间的语义对齐,另外还使用自关键序列训练算法对 GAN 进行优化,以解决因文本的离散性而导致模型难以训练的问题。Feng 等[77]首次提出使用完全无监督的训练方法来生成图像描述,该方法使用 MS COCO 数据集的图像和一个由 200 多万个句子组成的语料库进行模型训练,图像和句子之间没有任何配对集合。该模型首先使用语料库训练一个对抗生成网络,从而使模型能够生成一个完整的句子,然后通过预先训练好的视觉编码器对图像进行编码,并且通过图像特征和句子的双重映射对生成的句子不断进行重建,直到生成的描述句和图像互相匹配。该方法不需要成对的图像和描述句,对图像数据集完全没有依赖性。Guo 等[78]提出一种基于生成对抗网络的多风格图像描述模型(Multi-Style Image Captioning,MSCap),在生成器和判别器对抗训练生成描述句的同时,使用一个风格分类器将描述句分类为不同的风格,通过反向翻译模块来保证句子和描述的互相匹配,最终使用 softmax 对整个模型进行端到端的优化。Zhao 等[79]于 2020 年提出 MemCap(Memorizing Style Knowledge for Image Captioning)模型,由于描述语言风格无法从图像中获取,该模型通过设置一个包含语义风格的存储记忆模块,在生成描述时检索对应的描述风格,能够在保证句子准确的前提下生成带有明显语言风格的图像描述。实验表明,在多风格评价准则下,MemCap 模型的效果优于 MSCap 模型。基于生成对抗网络的图像描述方法在生成风格化描述方面有着广阔的前景,并且对数据集的精度没有过多的依赖,这也是图像描述任务未来的发展趋势。生成对抗网络虽然由于训练自由度高等问题导致准确率相对较低,但是因为具有无监督训练和反向传播机制等优点,依然成为人们研究的热点[80]。

2.3.3　基于强化学习的方法

强化学习[81]也是机器学习领域中重要的方法之一,也称为鼓励学习、增强学习。在强化学习中,智能体(Agent)以尝试的方式与环境之间不断交互,如图 6 所示。在交互过程中,环境的状态由于智能体的动作而发生改变,并且环境将奖赏和当前时间的状态作为强化信号反馈到智能体,智能体在强化信号的作用下改变其在环境中的动作,可以针对具体的问题实施特定的动作策略,旨在获取最大的奖赏。在图像描述任务中,强化学习可以解决在训练和预测过程中解码器的不同参数带来的解码(曝光)偏差的问题,并且在训练时通过反向传播算法对模型进行训练优化,从而解决训练和测评指标不匹配的问题。

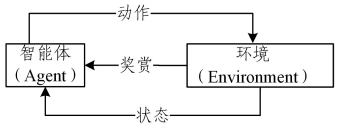


图 6　强化学习图示

Fig. 6　Reinforcement learning setting

Ranzato 等[82]于 2015 年使用强化学习来解决自然语言处理领域的问题。由于强化学习可以解决基于马尔可夫的动态规划问题[83],而 RNN 的隐藏层的传递符合马尔可夫过程,因此 Ranzato 等在模型的解码端引入强化学习,解决了因在模型的训练和预测过程中解码部分存在不同参数依赖而导致的解码误差传递问题。Liu 等[84]提出基于强化学习的图像描述方法,该方法同样是以编码器-解码器为基础,使用 CIDEr[85]和 SPICE[86]两个指标的组合作为模型的奖励函数,分别用于衡量句子语法和句子与图像的相似程度,并用策略梯度方法进行优化。Rennie 等[71]以注意力机制模型为基础,把序列问题看作强化学习问题,提出 SCST(Self-Critical sequence training)强化学习方法,并且对注意力机制中的 LSTM 进行了改进,大大提高了实验的准确率。Gao 等[87]提出了一种新的优势函数,并且在强化学习过程中使用 n 个时间步的累计奖赏代替交叉熵损失函数来评价智能体的动作,取得了不错的效果。同样地,Chen 等[88]为了消除解码偏差问题,提出在交叉熵损失函数中使用序列级监督代替单词级监督,实验表明该方法对模型的准确率和召回率均有所提升。基于强化学习的方法能够促使模型在特定环境下实现自身的调整与升级,使模型考虑长期的高回报,而不是一次性的匹配问题,从而使计算机的训练过程更加接近人类学习的过程。强化学习算法的关键在于其奖励和反馈机制,近年来,随着研究者们对奖励函数不断进行改进,强化学习在图像描述任务中取得了良好的表现。

2.3.4　基于密集描述的方法

基于密集描述的图像描述方法就是将图像描述分解为多个图像区域描述,当描述一个物体时,可以看作目标识别,当描述很多物体或一幅图像时,就是图像描述,如图 7 所示。

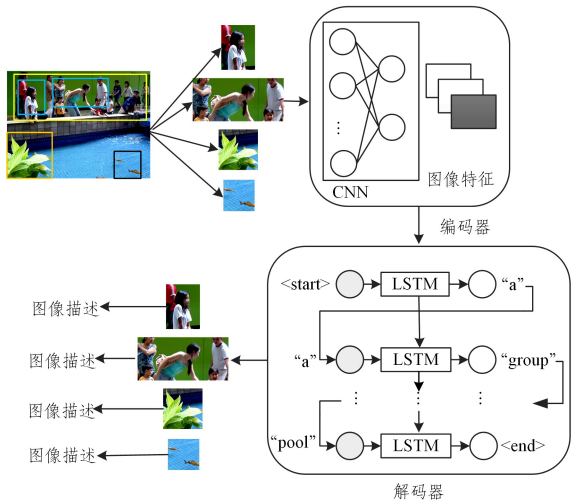


图 7　基于密集描述的图像描述

Fig. 7　Image captioning based on dense captioning

Johnson 等[89]于 2015 年提出了一种全卷积网络结构,包括 CNN、密集定位层和 LSTM 语言层。该网络利用双线性插值定位层的可导性,实现端到端式的训练,无需生成候选区域,只需进行一次优化和前馈计算就可以得到输出结果。Yang 等[90]提出一种基于推理和上下文融合的密集描述方法。推理机制依赖于区域的图像特征和预测描述,以便定位

区域边界，从而解决因区域密集而产生的区域重叠问题。上下文融合机制将文本特征与图像特征相结合，提供更加丰富的语义描述。Kim 等[91]基于密集描述提出一个多任务三流网络(Multi-task Triple-stream Network，MTTSNet)，该网络由区域生成网络(Region Proposal Network，RPN)和 3 个不同词性标签的循环单元组成，在 RPN 生成对应图像区域后，不同的循环单元共同作用于单词的预测和生成。该方法在不同词性标签之间建立不同的语义关系，生成更加密集、信息量更大的图像描述。Yin 等[92]指出在密集描述中不同图像区域之间缺少语义关联，并基于此提出了多尺度特征融合模型和语义属性监督机制，使模型在生成更加人性化的描述句的同时，增强不同图像区域之间的相关性，保证了不同图像区域之间的上下文关系。密集描述可以依靠图像的全局信息结合多个区域生成具有上下文相关性的图像描述，但是由于密集描述是对图像的不同区域内容进行描述，因此当选框中的内容不是图像的主要内容时，会导致图像的整体描述出现偏差。

表 2 基于生成对抗网络、强化学习的图像描述方法的实验结果对比

Table 2 Performance of GAN based and reinforcement learning based image captioning methods

方法	数据集	评价指标						
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	ROUGEL
Dai[74]	MS COCO	—	—	0.305	0.207	0.224	0.795	0.527
Deshpande[30]	MS COCO	0.739	0.569	0.425	0.316	0.255	1.045	0.532
Dognin[76]	MS COCO	—	—	—	—	0.260	1.027	—
Chen[75]	MS COCO	0.744	0.581	0.443	0.338	0.262	—	—
Liu[84]	MS COCO	0.754	0.591	0.445	0.332	0.257	1.013	0.550
Rennie[71]	MS COCO	—	—	—	0.354	0.271	1.175	0.566
Gao[87]	MS COCO	0.776	0.613	0.465	0.348	0.269	1.126	0.561
Chen[88]	MS COCO	—	—	—	—	0.270	1.172	—

从表 2 可以看出，不同方法的结果的准确率差别较小，但是这些方法在图像描述领域为研究者们提供了更广阔的思路，并且有着广阔的发展前景。

3 评价指标

目前图像描述领域常用的评价指标有 BLEU[93]，METEOR[94]，ROUGE[95]，CIDEr[85] 和 SPICE[86]，其中 BLEU 和 Meteor 起初适用于机器翻译，ROUGE 适用于文本自动摘要，CIDEr 和 SPICE 才是为图像描述专门定制的。由于图像描述借鉴并融合了机器翻译、自动摘要领域的注意力机制、生成对抗网络等深度学习方法，以上 5 种评价指标通常都可以用于图像描述。下文将分别进行详细描述。

3.1 BLEU

双语互评辅助工具(Bilingual Evaluation Understudy，BLEU)是由 IBM 公司于 2002 年提出[93]的。该工具的算法核心在于研究中提出的 N-grams 和 BP 惩罚因子，其中 N-grams 为 N 单位片段，N 一般取值为 4，通过对待评价句和参考句不同单词数目片段进行比较来计算句子的准确度，公式如下：

$$P_N=\frac{\sum_i\sum_k \min(h_k(c_i),\max_{j\in m} h_k(s_{ij}))}{\sum_i\sum_k \min(h_k(c_i))} \tag{1}$$

其中，P_N 为各阶 N-gram 的精度，$h_k(c_i)$ 表示第 k 个 N-gram 在待评句 c_i 中出现的次数，$h_k(s_{ij})$ 表示第 k 个 N-gram 在标准参考句 s_{ij} 中出现的次数，$\max_{j\in m} h_k(s_{ij})$ 表示某 N-gram 在多条标准参考句中出现最多的次数，分子整体表示取 N-gram 在待评句和参考句中出现的最小次数。由于阶数升高会导致 N-gram 统计量的精度减小，为了平衡各阶统计量的作用，取其加权平均值，再乘以惩罚因子，得到最后的评价公式：

$$BLEU=BP*\exp(\sum_{n=1}^{N} W_n \log P_N) \tag{2}$$

其中，W_n 为各阶 N-grams 的精度的权重，当待评句长度大于参考句长度时，BP 为 1，否则 BP 为 $e^{1-\frac{l_s}{l_c}}$，其表达式如下：

$$BP=\begin{cases}1, & l_c>l_s\\ e^{1-\frac{l_s}{l_c}}, & l_c\leqslant l_s\end{cases} \tag{3}$$

BLEU 是较为常用的评价指标，使用方便、快捷，评价结果接近人类评价，但其不考虑待评句在语法上的准确性，也没有涉及同义词或相似表达，可能会出现合理描述分值低的情况。

3.2 METEOR

METEOR(Metric for Evaluation of Translation with Explicit ORdering)评价指标由 Satanjeev 于 2002 年提出[94]，该指标基于单字召回率和单精度的加权调和平均数，弥补了 BLEU 的一些缺陷，并且增加了基于 WordNet[96]的同义词库以解决同义词匹配的问题。基于同义词库给定一组校准值 m，通过最小化词片段 ch，计算待评句和参考句之间的准确率和召回率的调和平均值，得出 $METEOR$ 值，公式如下：

$$METEOR=(1-Pen)F_{\text{mean}} \tag{4}$$

其中，惩罚系数 $Pen=\gamma(\frac{ch}{m})^{\theta}$，$F_{\text{mean}}=\frac{P_mR_m}{\alpha P_m+(1-\alpha)R_m}$ 为准确率和召回率的调和平均值，准确率为 $P_m=\frac{|m|}{\sum_k h_k(c_i)}$，召回率为 $R_m=\frac{|m|}{\sum_k h_k(s_{ij})}$，$\gamma$，$\theta$ 和 α 为根据不同数据集确定的超参数。METEOR 指标计算待评句在整个语料库上的准确率和召回率，因此与人工评价更加相似，但是由于超参数较多且不适用于评价多个数据集，在过去使用较少，近年来才逐渐被人们接受。

3.3 ROUGE

ROUGE(Recall-oriented Understudy for Gisting Evaluation)由 Lin[95] 于 2004 年提出，起初用于自动摘要任务。ROUGE 的思想和 BLEU 一致，区别在于 BLEU 计算的是准

确率，而 ROUGE 计算的是召回率。该指标中定义了最长公共子句(Longest Common Subsequence，LCS)，即待评句和参考句的最长相同片段，因此 ROUGE 包含以下 4 种细分指标：ROUGE-N，ROUGE-L，ROUGE-W 和 ROUGE-S，其中 ROUGE-N 是基于 N-gram 的共现性统计，与 BLEU 类似，ROUGE-L 基于 LCS 来评价待评句和参考句之间的相似性，以待评句 X 和参考句 Y 为例，其公式如下：

$$ROUGE\text{-}L=\frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}} \tag{5}$$

其中，$R_{lcs}=\frac{LCS(X,Y)}{m}$，$P_{lcs}=\frac{LCS(X,Y)}{n}$，$m$ 和 n 分别代表 X 和 Y 的长度，$\beta=\frac{P_{lcs}}{R_{lcs}}$。

ROUGE-W 是带有权重的 LCS 统计，目的是为连续正确的片段打更高的分数。ROUGE-S 是对不连续二元组共现性的统计，即对于一个句子中两个有序单词，它们之间可以存在任意一个单词，公式如下：

$$ROUGE\text{-}S=\frac{(1+\beta^2)R_{skip2}P_{skip2}}{R_{skip2}+\beta^2 P_{skip2}} \tag{6}$$

其中，$\beta=\frac{P_{skip2}}{R_{skip2}}$，$R_{skip2}=\frac{SKIP2(X,Y)}{m}$，$P_{skip2}=\frac{SKIP2(X,Y)}{n}$。

3.4 CIDEr

CIDEr(Consensus-based Image Description Evaluation)评价标准由 Vedantam 于 2015 年在计算机视觉与模式识别大会上提出[85]。CIDEr 可以看作 BLEU 与向量空间模式的结合，其原理是将句子看作文档处理，使用 TF-IDF[97] 计算其单词权重，从而把句子表示成向量形式，通过计算待评句和参考句的 TF-IDF 向量的余弦距离来计算两者之间的相似性。具体公式如下：

$$CIDEr(c_i,S_i)=\frac{1}{N}\sum_{n=1}CIDEr_n(c_i,S_i) \tag{7}$$

其中，c_i 和 S_i 分别是待评句和参考句集合。对于不同 N 值，余弦相似度平均值为：

$$CIDEr_N(c_i,S_i)=\frac{1}{m}\sum_j\frac{g_n(c_i)*g_n(s_{ij})}{\|g_n(c_i)\|*\|g_n(s_{ij})\|} \tag{8}$$

TF-IDF 的权重 $g_n(\cdot)$的计算方式如下：

$$g_n(\cdot)=\frac{h_n(\cdot)}{\sum_{\omega_l\in\Omega}h_l(\cdot)}\log\left(\frac{|I|}{\sum_{I_p\in I}\min(1,\sum_q h_n(\cdot))}\right) \tag{9}$$

其中，ω 表示不同的 N-grams，$(\cdot)$ 表示 ω 在待评句 c_i 或参考句 s_{ij} 中出现的次数，Ω 是 N-grams 字表，I 是图像集合。CIDEr 指标的优点是通过 TF-IDF 给不同的关键词赋予不同的权重，能够判断待评句是否提取出图像的关键信息，并对提取到图像中较为关键的信息的描述给予更高的分数，而对提取到一些无关紧要的信息的描述给予较低的分数。

3.5 SPICE

SPICE (Semantic Propositional Image Caption Evaluation)由 Anderson 等于 2016 年提出[86]，专门用于评价图像描述任务。之前的评价方法都是基于 N-gram 进行计算，而该方法基于语义场景图(物体、属性、关系)定义，对图像的待评句和参考句用上下文无关文法 (Probabilistic Context-Free Grammar，PCFG)[98] 解析为语义决策树，进而映射为语义场景图，最后计算待评句中物体、属性及关系的 F 值来进行准确度评价。SPICE 评价指标与人类评价更吻合，但是不能准确判断语法结构错误。

4 常用数据集

4.1 MS COCO

MS COCO(Microsoft COCO(Common Objects in Context))数据集[99]是由微软于 2014 年出资，通过在 AMT(Advanced Manufacturing Technology)平台上进行数据标注构建的数据集。该数据集以场景理解为目标，主要用于目标检测、目标分割和图像语义分割等任务。该数据集有着丰富的数据资源，图像内容大多是自然图片或者背景复杂的日常生活图片。2014 版的 MS COCO 数据集共有 82 783 个训练图像、40 504 个验证图像和 40 775 个测试图像，其中包含 91 个物体类别(如汽车、自行车)，80 个对象类别(如交通工具)，平均每张图像包含 3.5 个类别、7.7 个实例目标和 5 个参考描述，只包含 1 个类别的图像不超过 20%，只包含 1 个实例目标的图像不超过 10%。另外，标签文件中标记了每个实例及其边界的精确坐标，精度为小数点后两位。由于 MS COCO 数据集图像的背景复杂，目标多且尺寸小，因此该数据集上的任务更加复杂，即便如此，由于其庞大且精准的数据，大多数人倾向于使用该数据集评估模型，MS COCO 数据集也成为目前计算机视觉领域最受欢迎的数据集。

4.2 Flickr 8K 和 Flickr 30K

Flickr 8K[100] 和 Flickr 30K[101] 数据集分别于 2013 年和 2015 年由美国伊利诺伊大学构建，数据集中的图片均从雅虎相册收集而来。Flickr 8K 数据集包含 8092 张 JPEG 格式的图像，其中训练图像 6 092 张，测试图像 1 000 张和开发图像 1 000 张，并且每张图像对显著实体和事件有 5 个不同的参考描述句，每张图像的描述句长度平均为 11.8 个单词。该数据集规模较小，适合初学者使用。

Flickr 30K 数据集包含 31 783 张图像，共有 513 644 个实体。该数据集中 94.2%的图像包含人类，12%的图像包含动物，包含衣物和肢体动作的分别占 69.9%和 28%，还有 18.1%的图像包含汽车及其他工具。与 Flickr 8K 一样，Flickr 30K 数据集中的每张图像均有 5 个参考描述句，并且研究者手动增加了图像中与每个实体对应的边界框。

上述数据集都有着丰富的实例、背景及标注，是近年来目标检测、图像分割、图像描述领域的研究者们最常用的数据集。此外，用于图像描述任务的数据集还有 Visual Genome[102]，Instagram[103]，IAPR TC-12[104]，MIT-Adobe FiveK[105] 等。

结束语　本文对图像描述技术的主要方法、评价指标的原理和计算方法以及常用数据集进行了总结。随着深度学习的发展，使用深度学习方法来解决图像描述任务使得准确率得到了本质上的提高，极大地促进了图像描述技术的发展，并且已经成为目前的主流趋势。目前，对于计算机而言，生成图

像描述需要大量带有标注的数据，并且受到图像参考句的限制，计算机生成的描述只是客观平淡、不带感情色彩的句子。完全无监督的训练方式和风格化描述是图像描述未来的发展趋势。在无监督的训练方式下，不再需要图像和句子严格匹配的数据集，大大降低了模型对数据集的依赖性。多风格化的描述使生成的图像描述富有多样性、可解释性，不仅可以针对人们的需求生成侧重点不同的描述句，而且能够针对不同环境生成不同情感倾向的图像描述。在注意力机制的框架下，虽然计算机能够针对不同重点区域生成特定内容的图像描述，但还是摆脱不了数据集中参考描述句的限制[16,33]。而无监督的训练方式由于不受数据集的限制，能够在保证生成完整句子的同时，很好地融入风格化模块，例如在生成对抗网络中加入风格分类器[78]或风格储存单元[79]，生成包含多类情感倾向的图像描述。目前图像描述任务的数据资源和技术方法日渐成熟，但是对于深度学习框架的探索和提升无监督环境下图像描述的准确率还需要进一步的研究。图像描述技术已被广泛应用于智能信息传播、智慧家居和智慧交通等领域，对人们的日常生活有着重要的实际意义，将来图像描述任务在深度学习和人工智能领域仍是一个重要的研究方向。

参 考 文 献

[1] XU H J, HUANG C Q, HUANG X D, et al. Multi-modal multi-concept-based deep neural network for automatic image annotation[J]. Multimedia Tools and Applications, 2019, 78(21): 30651-30675.

[2] ROYA R, MANSOUR J. Image annotation using multi-view non-negative matrix factorization with different number of basis vectors[J]. Journal of Visual Communication and Image Representation, 2017, 46(1): 1-12.

[3] ZHANG Z, ZHAO Y X, LI D, et al. A novel image annotation model based on content representation with multi-layer segmentation[J]. Neural Computing and Applications, 2015, 26(6): 1407-1422.

[4] REN Y M, CHENG X Y, LI X Y, et al. Description and Recognition of Image Based on Concept Semantics[J]. Computer Science, 2008, 35(7): 206-212.

[5] XIE D N, ROSS G, PIOTR D, et al. Aggregated residual transformations for deep neural networks[C]// 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. Honolulu, HI, United states: Institute of Electrical and Electronics Engineers Inc., 2017: 5987-5995.

[6] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]// 29th IEEE Conference on Computer Vision and Pattern Recognition(CVPR 2016). Las Vegas, NV, United States: IEEE Computer Society, 2016: 770-778.

[7] CHRISTIAN S, LIU W, YANG Q J, et al. Going deeper with convolutions[C]// IEEE Conference on Computer Vision and Pattern Recognition(CVPR 2015). Boston, MA, United States: IEEE Computer Society, 2015: 1-9.

[8] CHRISTIAN S, SERGEY I, VINCENT V, et al. Inception-v4, inception-ResNet and the impact of residual connections on learning[C]// 31st AAAI Conference on Artificial Intelligence (AAAI 2017). San Francisco, CA, United States: AAAI Press, 2017: 4278-4284.

[9] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the Inception Architecture for Computer Vision[C]// 29th IEEE Conference on Computer Vision and Pattern Recognition(CVPR 2016). Las Vegas, NV, United States: IEEE Computer Society, 2016: 2818-2826.

[10] ZHANG W W, ZHOU H, SUN S Y, et al. Robust multi-modality multi-object tracking[C]// 17th IEEE/CVF International Conference on Computer Vision(ICCV 2019). Seoul, Korea, Republic of Institute of Electrical and Electronics Engineers Inc., 2019: 2365-2374.

[11] KAREN S, ANDREW Z. Very deep convolutional networks for large-scale image recognition[C]// 3rd International Conference on Learning Representations(ICLR 2015). San Diego, CA, United States: International Conference on Learning Representations, ICLR, 2015.

[12] STEFANO M. Knowledge enhanced representations to reduce the semantic gap in clinical decision support[C]// 9th PhD Symposium on Future Directions in Information Access (FDIA 2019). Milan, Italy: CEUR-WS, 2019: 4-9.

[13] TANG J H, ZHA Z J, TAO D C, et al. Semantic-Gap-Oriented Active Learning for Multilabel Image Annotation[J]. Ieee Transactions on Image Processing, 2012, 21(4): 2354-2360.

[14] YANG X, TANG K H, ZHANG H W, et al. Auto-encoding scene graphs for image captioning[C]// 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019). Long Beach, CA, United States: IEEE Computer Society, 2019: 10677-10686.

[15] LU J S, XIONG C M, DEVI P, et al. Knowing when to look: Adaptive attention via a visual sentinel for image captioning [C]// 30th IEEE Conference on Computer Vision and Pattern Recognition(CVPR 2017). Honolulu, HI, United states: Institute of Electrical and Electronics Engineers Inc., 2017: 3242-3250.

[16] CHEN S Z, JIN Q, WANG P, et al. Say As You Wish: Fine-grained Control of Image Caption Generation with Scene Graphs [J]. arXiv: 2003.00387.

[17] CHEN T S, LIN L, ZUO W M, et al. Learning a wavelet-like auto-encoder to accelerate deep neural networks[C]// 32nd AAAI Conference on Artificial Intelligence(AAAI 2018). New Orleans, LA, United states: AAAI press, 2018: 6722-6729.

[18] HYUN K, HYUNSOO Y, KI-WOONG P. Multi-targeted backdoor: Indentifying backdoor attack for multiple deep neural networks[J]. IEICE Transactions on Information and Systems, 2020, E103D(4): 883-887.

[19] OORD A V D, LI Y Z, BABUSCHKIN I, et al. Parallel WaveNet: Fast high-fidelity speech synthesis[C]// 35th International Conference on Machine Learning, ICML. Stockholm, Sweden: International Machine Learning Society (IMLS), 2018: 6270-6278.

[20] ZHU J Y, ZHANG R, PATHAK D, et al. Toward multimodal image-to-image translation[C]// 31st Annual Conference on Neural Information Processing Systems (NIPS 2017). Long Beach, CA, United States: Neural Information Processing Systems Foundation, 2017: 466-477.

[21] WANG Q, MAO Z D, WANG B, et al. Knowledge graph embedding: A survey of approaches and applications[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(12): 2724-2743.

[22] HINTON G, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural computation, 2006, 18(7): 1527-1554.

[23] WANG Y Y, WANG L, QI J, et al. Improved text clustering algorithm and application in microblogging public opinion analysis [C]// 2013 4th World Congress on Software Engineering (WCSE 2013). Hong Kong, China: IEEE Computer Society, 2013: 27-31.

[24] YANG Y X. Research and Realization of Internet Public Opinion Analysis Based on Improved TF-IDF Algorithm[C]// 16th International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES 2017). An-Yang, He Nan, China: Institute of Electrical and Electronics Engineers Inc., 2017: 80-83.

[25] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]// 32nd International Conference on Machine Learning (ICML 2015). Lile, France: International Machine Learning Society (IMLS), 2015: 448-456.

[26] HE K M, ZHANG X Y, REN S Q, et al. Identity mappings in deep residual networks[C]// 21st ACM Conference on Computer and Communications Security (CCS 2014). Scottsdale, AZ, United states: Springer Verlag, 2016: 630-645.

[27] WU Z F, SHEN C H, HENGEL A V D. Wider or Deeper: Revisiting the ResNet Model for Visual Recognition[J]. Pattern Recognition, 2019, 90: 119-133.

[28] WANG N, SONG Y B, MA C, et al. Unsupervised deep tracking [C]// 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019). Long Beach, CA, United States: IEEE Computer Society, 2019: 1308-1317.

[29] ANDERSON P, HE X D, BUEHLER C, et al. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering[C]// 31st Meeting of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018). Salt Lake City, UT, United States: IEEE Computer Society, 2018: 6077-6086.

[30] DESHPANDE A, ANEJA J, WANG L W, et al. Fast, diverse and accurate image captioning guided by part-of-speech[C]// 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019). Long Beach, CA, United states: IEEE Computer Society, 2019: 10687-10696.

[31] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: A neural image caption generator[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015). Boston, MA, United States: IEEE Computer Society, 2015: 3156-3164.

[32] XU K, BA J L, KIROS R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]// 32nd International Conference on Machine Learning. Lile, France: International Machine Learning Society (IMLS), 2015: 2048-2057.

[33] CORNIA M, BARALDI L, CUCCHIARA R. Show, control and tell: A framework for generating controllable and grounded captions[C]// 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019). Long Beach, CA, United States: IEEE Computer Society, 2019: 8299-8308.

[34] HOSSAIN M Z, SOHEL F, SHIRATUDDIN M F, et al. A comprehensive survey of deep learning for image captioning[J]. ACM Computing Surveys, 2019, 51(6): 118: 1-118: 36.

[35] YAGI M, SHIBATA T, TAKADA K. Human-perception-like image recognition system based on the Associative Processor architecture[C]// 11th European Signal Processing Conference, EUSIPCO. Toulouse, France: European Signal Processing Conference, EUSIPCO, 2002.

[36] ITO S, MITSUKURA Y, FUKUMI M, et al. The image recognition system by using the FA and SNN[C]// 7th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES 2003). Oxford, United Kingdom: Springer Verlag, 2003: 578-584.

[37] KEYSERS D, DESELAERS T, NEY H. Pixel-to-pixel matching for image recognition using Hungarian graph matching[C]// 26th DAGM Symposium on Pattern Recognition. Tubingen, Germany: Springer Verlag, 2004: 154-162.

[38] FARHADI A, HEJRATI S M M, SADEGHI M A, et al. Every Picture Tells a Story: Generating Sentences from Images[J]. lecture notes in computer science, 2010, 21(10): 15-29.

[39] DAVID V, SANCHEZ A. Advanced support vector machines and kernel methods[J]. Neurocomputing, 2003, 55(1/2): 5-20.

[40] CHEN P H, LIN C J, SCHOLKOPF B. A tutorial on v-support vector machines[J]. Applied Stochastic Models in Business and Industry, 2005, 21(2): 111-136.

[41] EVERINGHAM M, GOOL L V, WILLIAMS C K I, et al. The Pascal Visual Object Classes (VOC) challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303-338.

[42] LI S M, KULKARNI G, BERG T L, et al. Composing simple image descriptions using web-scale N-grams[C]// 15th Conference on Computational Natural Language Learning (CoNLL 2011). Portland, OR, United states: Association for Computational Linguistics (ACL), 2011: 220-228.

[43] KULKARNI G, PREMRAJ V, ORDONEZ V, et al. Baby talk: Understanding and generating simple image descriptions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2891-2903.

[44] QI Y, SZUMMER M, MINKA T P. Bayesian conditional random fields[C]// 10th International Workshop on Artificial Intelligence and Statistics (AISTATS 2005). Hastings, Christ

Church, Barbados: The Society for Artificial Intelligence and Statistics, 2005: 269-276.

[45] SUTTON C, MCCALLUM A, ROHANIMANESH K. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data[J]. Journal of Machine Learning Research, 2007, 8(2): 693-723.

[46] MITCHELL M, HAN X F, DODGE J, et al. Midge: Generating image descriptions from computer vision detections[C]// 13th Conference of the European Chapter of the Association for Computational Linguistics(EACL 2012). Avignon, France: Association for Computational Linguistics (ACL), 2012: 747-756.

[47] NOR W, MOHAMED H W, SALLEH M N M, et al. A comparative study of Reduced Error Pruning method in decision tree algorithms[C]// 2012 IEEE International Conference on Control System, Computing and Engineering (ICCSCE 2012). Penang, Malaysia: IEEE Computer Society, 2012: 392-397.

[48] ORDONEZ V, KULKARNI G, BERG T L. Im2Text: Describing images using 1 million captioned photographs[C]// 25th Annual Conference on Neural Information Processing Systems 2011 (NIPS 2011). Granada, Spain: Curran Associates Inc., 2011.

[49] SOCHER R, KARPATHY A, LE Q V, et al. Grounded Compositional Semantics for Finding and Describing Images with Sentences[J]. Transactions of the Association for Computational Linguistics, 2014, 2(Q14-1017): 207-218.

[50] KUZNETSOVA P, ORDONEZ V, BERG T L, et al. Tree Talk: Composition and Compression of Trees for Image Descriptions[J]. Transactions of the Association for Computational Linguistics, 2014, 2(Q14-1017): 351-362.

[51] MASON R, CHARNIAK R. Nonparametric method for data-driven image captioning[C]// 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014). Baltimore, MD, United states: Association for Computational Linguistics (ACL), 2014: 592-598.

[52] SUN C, GAN C, NEVATIA R. Automatic concept discovery from parallel text and visual corpora[C]// 15th IEEE International Conference on Computer Vision (ICCV 2015). Santiago, Chile: Institute of Electrical and Electronics Engineers Inc., 2015: 2596-2604.

[53] CHO K, MERRIENBOER B V, GULCEHRE C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[J]. arXiv: 1406.1078.

[54] LI M D, MU K, ZHONG P, et al. Generating steganographic image description by dynamic synonym substitution[J]. Signal Processing, 2019, 164: 193-201.

[55] KIROS R, SALAKHUTDINOV R, ZEMEL R. Multimodal neural language models[C]// 31st International Conference on Machine Learning (ICML 2014). Beijing, China: International Machine Learning Society (IMLS), 2014: 2012-2025.

[56] MAO J H, XU W, YANG Y, et al. Deep captioning with multimodal recurrent neural networks (m-RNN)[C]// 3rd International Conference on Learning Representations (ICLR 2015). San Diego, CA, United States, 2015.

[57] HERMANS M, SCHRAUWEN B. Memory in linear recurrent neural networks in continuous time[J]. Neural Networks, 2010, 23(3): 341-355.

[58] GREFF K, SRIVASTAVA R K, KOUTNIK J, et al. LSTM: A Search Space Odyssey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(10): 2222-2232.

[59] CHINEA A. Understanding the principles of recursive neural networks: A generative approach to tackle model complexity [C]// 19th International Conference on Artificial Neural Networks (ICANN 2009). Limassol, Cyprus: Springer Verlag, 2009: 952-963.

[60] SHEN Y K, TAN S, SORDONI A, et al. Ordered neurons: Integrating tree structures into recurrent neural networks[C]// 7th International Conference on Learning Representations (ICLR 2019). New Orleans, LA, United States, 2019.

[61] JIA X, GAVVES E, FERNANDO B, et al. Guiding the long-short term memory model for image caption generation[C]// 15th IEEE International Conference on Computer Vision (ICCV 2015). Santiago, Chile: Institute of Electrical and Electronics Engineers Inc., 2015: 2407-2415.

[62] MAO J H, HUANG J, TOSHEV A, et al. Generation and comprehension of unambiguous object descriptions[C]// 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016). Las Vegas, NV, United States: IEEE Computer Society, 2016: 11-20.

[63] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[C]// 3rd International Conference on Learning Representations (ICLR 2015). San Diego, CA, United States, 2015.

[64] XIAO T J, XU Y C, YANG K Y, et al. The application of two-level attention models in deep convolutional neural network for fine-grained image classification[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, United States: IEEE Computer Society, 2015: 842-850.

[65] STOLLENGA M F, MASCI J, GOMEZ F, et al. Deep networks with internal selective attention through feedback connections [C]// 28th Annual Conference on Neural Information Processing Systems 2014 (NIPS 2014). Montreal, QC, Canada, 2014: 3545-3553.

[66] CHU X, YANG W, OUYANG W, et al. Multi-context attention for human pose estimation[C]// 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017). Honolulu, HI, United States: Institute of Electrical and Electronics Engineers Inc., 2017: 5669-5678.

[67] ZHAO B, WU X, FENG J S, et al. Diversified Visual Attention Networks for Fine-Grained Object Classification [J]. IEEE Transactions on Multimedia, 2017, 19(6): 1245-1256.

[68] DENG Z P, SUN H, ZHOU S L, et al. Toward Fast and Accurate Vehicle Detection in Aerial Images Using Coupled Region-Based Convolutional Neural Networks[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2017, 10(8): 3652-3664.

[69] ZHOU Y E,WANG M,LIU D Q,et al. More Grounded Image Captioning by Distilling Image-Text Matching Model[J]. arXiv:2004.00390.

[70] LEE K H,CHEN X,HUA G,et al. Stacked Cross Attention for Image-Text Matching[C]// 15th European Conference on Computer Vision(ECCV 2018). Munich,Germany:Springer Verlag,2018:212-228.

[71] RENNIE S J,MARCHERET E,MROUEH Y,et al. Self-critical sequence training for image captioning[C]// 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017). Honolulu,HI,United States:Institute of Electrical and Electronics Engineers Inc.,2017:1179-1195.

[72] ZHAO Z S,GAO H X,SUN Q,et al. Latest Development of the Theory Framework,Derivative Model and Application of Generative Adversarial Nets [J]. Journal of Chinese Mini-Micro Computer Systems,2018,39(12):44-48.

[73] ZHAO Z S,SUN Q,YANG H R,et al. Compression Artifacts Reduction by Improved Generative Adversarial Networks [J/OL]. Journal on Image and Video Processing,2019,https://doi.org/10.1186/s13640-019-0465-0.

[74] DAI B,SANJA F,RAQUEL U,et al. Towards Diverse and Natural Image Descriptions via a Conditional GAN[C]// 16th IEEE International Conference on Computer Vision (ICCV 2017). Venice,Italy:Institute of Electrical and Electronics Engineers Inc.,2017:2989-2998.

[75] CHEN F H,JI R R,SUN X S,et al. GroupCap:Group-Based Image Captioning with Structured Relevance and Diversity Constraints[C]// 31st Meeting of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR 2018). Salt Lake City,UT,United States:IEEE Computer Society,2018:1345-1353.

[76] DOGNIN P,MELNYK I,MROUEH Y,et al. Adversarial semantic alignment for improved image captions[C]// 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019). Long Beach,CA,United States:IEEE Computer Society,2019:10455-10463.

[77] FENG Y,MA L,LIU W,et al. Unsupervised image captioning [C]// 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019). Long Beach,CA,United States:IEEE Computer Society,2019:4120-4129.

[78] GUO L T,LIU J,YAO P,et al. MSCAP:Multi-style image captioning with unpaired stylized text[C]// 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019). Long Beach,CA,United States:IEEE Computer Society,2019:4199-4208.

[79] ZHAO W T,WU X X,ZHANG X X. MemCap:Memorizing Style Knowledge for Image Captioning[C]// The Thirty-Fourth AAAI Conference on Artificial Intelligence(AAAI 2020). New York,NY,USA,2020:12984-12992.

[80] SHETTY R,ROHRBACH M,HENDRICKS L A,et al. Speaking the Same Language:Matching Machine to Human Captions by Adversarial Training[C]// 16th IEEE International Conference on Computer Vision(ICCV 2017). Venice,Italy:Institute of Electrical and Electronics Engineers Inc.,2017:4155-4164.

[81] TESAURO G. Temporal difference learning and TD-gammon [J]. Communications of the ACM,1995,38(3):58-68.

[82] RANZATO M A,CHOPRA S,AULI M,et al. Sequence Level Training with Recurrent Neural Networks[J]. arXiv:1511.06732.

[83] LIM S H,XU H,MANNOR S. Reinforcement learning in robust markov decision processes[J]. Mathematics of Operations Research,2016,41(4):1325-1353.

[84] LIU S Q,ZHU Z H,YE N,et al. Improved Image Captioning via Policy Gradient optimization of SPIDEr[C]// 16th IEEE International Conference on Computer Vision(ICCV 2017). Venice,Italy:Institute of Electrical and Electronics Engineers Inc.,2017:873-881.

[85] VEDANTAM R,ZITNICK C L,PARIKH D. CIDEr:Consensus-based image description evaluation[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015). Boston,MA,United States:IEEE Computer Society,2015:4566-4575.

[86] ANDERSON P,FERNANDO B,JOHNSON M,et al. SPICE:Semantic Propositional Image Caption Evaluation[J]. Adaptive Behavior,2016,11(4):382-398.

[87] GAO J L,WANG S Q,WANG S S,et al. Self-critical n-step training for image captioning[C]// 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019). Long Beach,CA,United States:IEEE Computer Society,2019:6293-6301.

[88] CHEN J,JIN Q. Better Captioning with Sequence-Level Exploration[J]. arXiv:2003.03749.

[89] JOHNSON J,KARPATHY A,LI F F. DenseCap:Fully convolutional localization networks for dense captioning[C]// 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016). Las Vegas,NV,United States:IEEE Computer Society,2016:4565-4574.

[90] YANG L J,TANG K,YANG J C,et al. Dense captioning with joint inference and visual context[C]// 30th IEEE Conference on Computer Vision and Pattern Recognition(CVPR 2017). Honolulu,HI,United States:Institute of Electrical and Electronics Engineers Inc.,2017:1978-1987.

[91] KIM D J,CHOI J,OH T H,et al. Dense relational captioning:Triple-stream networks for relationship-based captioning[C]// 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019). Long Beach,CA,United States:IEEE Computer Society,2019:6264-6273.

[92] YIN G J,SHENG L,LIU B,et al. Context and attribute grounded dense captioning[C]// 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019). Long Beach,CA,United States:IEEE Computer Society,2019:6234-6243.

[93] PAPINENI K,ROUKOS S,WARD T,et al. BLEU:a Method for Automatic Evaluation of Machine Translation[C]// Procee-

dings of the 40th Annual Meeting of the Association for Computational Linguistics. Istanbul, Turkey: Association for Computational Linguistics, 2002: 311-318.

[94] BANERJEE S, LAVIE A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments[C]// Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Ann Arbor, Michigan: ACL, 2005: 65-72.

[95] LIN C Y. Automatic Evaluation of Summaries Using n-gram Co-occurrence Statistics[C]// Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. United States: Association for Computational Linguistics, 2003: 71-78.

[96] LEE Y Y, KE H, YEN T Y, et al. Combining and learning word embedding with WordNet for semantic relatedness and similarity measurement[J]. Journal of the Association for Information Science and Technology, 2020, 71(6): 657-670.

[97] ROBERTSON S. Understanding inverse document frequency: on theoretical arguments for IDF[J]. Journal of Documentation, 2004, 60(5): 503-520.

[98] MOHRI M, ROARK B. Probabilistic context-free grammar induction based on structural zeros[C]// 2006 Human Language Technology Conference-North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL 2006). New York, NY, United states: Association for Computational Linguistics (ACL), 2006: 312-319.

[99] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context[C]// 13th European Conference on Computer Vision (ECCV 2014). Zurich, Switzerland: Springer Verlag, 2014: 740-755.

[100] HODOSH M, YOUNG P, HOCKENMAIER J. Framing image description as a ranking task: Data, models and evaluation metrics[J]. Journal of Artificial Intelligence Research, 2013, 47(1): 853-899.

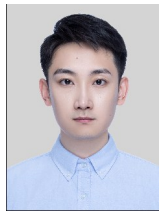
[101] PLUMMER B A, WANG L W, CERVANTES C M, et al. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models[C]// 15th IEEE International Conference on Computer Vision (ICCV 2015). Santiago, Chile: Institute of Electrical and Electronics Engineers Inc., 2015: 2641-2649.

[102] KRISHNA R, ZHU Y K, GROTH O, et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations[J]. International Journal of Computer Vision, 2017, 123(1): 32-73.

[103] TRAN K, HE X D, ZHANG L, et al. Rich Image Captioning in the Wild[C]// 29th IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW. Las Vegas, NV, United States: IEEE Computer Society, 2016: 434-441.

[104] GRUBINGER M, CLOUGH P, MÜLLER H, et al. The IAPR TC12 Benchmark: A New Evaluation Resource for Visual Information Systems[J]. Workshop Ontoimage, 2006, 5(10): 13-55.

[105] BYCHKOVSKY V, PARIS S, CHAN E, et al. Learning photographic global tonal adjustment with a database of input/output image pairs[C]// 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011. IEEE Computer Society, 2011: 97-104.



MIAO Yi, born in 1996, postgraduate. His main research interests include image processing and analysis.

ZHAO Zeng-shun, born in 1975, Ph.D, associate professor, Ph.D supervisor. His main research interests include computer vision, intelligent robots and machine learning.