文章编号：2095-2163(2019)05-0045-05 中图分类号：TP391.41 文献标志码：A

图像描述生成方法研究文献综述

张 姣，杨振宇
（齐鲁工业大学（山东省科学院），济南 250353）

摘 要： 随着人工智能技术的兴起，图像特征提取技术和文本自动生成技术都得到了长足的进步，将两者结合的图像描述生成技术也越来越受到学术界和工业界的重视。图像到文本生成是一个综合性问题，涉及自然语言处理和计算机视觉等领域。本文介绍了图像描述生成技术的研究背景及国内外研究现状，概述了目前研究者评估生成图像描述质量的图像数据集，对现有模型进行了详细的分类概括：基于模板的图像描述生成方法、基于检索的图像描述生成方法、基于深度学习的图像描述生成方法。与此同时一并总结阐述了该领域面临的问题和挑战。

关键词： 图像描述；文本生成；特征提取；计算机视觉

Research on image caption generation method based on deep learning: A literature review

ZHANG Jiao, YANG Zhenyu
(Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China)

【Abstract】 Image caption generation technology is used in many fields such as news communication, smart transportation, smart home and smart medical. Therefore, this technology has important academic and practical value. Image-to-text generation is a comprehensive problem involving areas such as natural language processing and computer vision. This paper introduces the research background of image caption generation technology and the research status at home and abroad, and summarizes the current image datasets that researchers evaluate to generate quality of the image caption. The existing models are classified and summarized in detail: template-based image caption generation method, retrieval-based image caption generation method and deep-learning-based image caption generation method. It also summarizes the problems and challenges which the field is facing.

【Key words】 image caption; text generation; feature extraction; computer vision

0 引 言

0.1 研究背景

大数据时代的到来使人工智能产品不断进入人们的视野。图像描述生成技术的产生为计算机从图像中快速获取信息带来了新的发展和应用前景。

图像描述生成技术与图像语义分析、图像标注和图像高级语义提取等技术紧密相关。图像描述生成技术是计算机自动为图像生成一个完整、通顺的描述语句。大数据背景下的图像描述生成技术在商业领域有着广泛的应用。如购物软件中用户输入关键字快速地搜索出符合要求的商品；用户在搜索引擎中进行的图片搜索；视频中多事物目标的识别、医学图像专业的自动语义标注以及自动驾驶中目标物体的识别等。如何在计算机中更有效、准确、快速地实现这一过程即是本文的研发课题。

从图像描述生成的发展过程[1]来看，可以分为3个主要发展阶段：基于模板的图像描述生成方法；基于检索的图像描述生成方法；基于深度学习的图像描述生成方法。

0.2 国内外研究现状

结合国内外研究人员对图像描述生成方法的研究以及各个阶段所采用的不同关键技术，可将图像描述的方法分为3类。对此可做分析阐述如下。

(1)基于模板的图像描述生成方法。该方法[2]利用图像标注技术为物体、物体场景以及组成部分进行标注[3]。选择与图像内容描述场景相关的句子作为表达模板，将提取的图像特征填入模板，继而得到图像的描述句子。概率图模型方法[4]对文本信息和图像信息建立模型，可从文本数据集中挑选合适的关键词，将其作为体现图像描述内容的关键词，利用语言模型技术[5-7]将选取的内容关键词组合为合乎语法规则习惯的英文句子。该方法的研究虽然能够描述图像内容，但是在一定程度上限制了

基金项目： 山东省自然科学基金(ZR2017LF021)；山东省重点研究发展计划(2017XCGC0605)。

作者简介： 张 姣(1993-)，女，硕士研究生，主要研究方向：深度学习、大数据智能制造与分析；杨振宇(1980-)，男，博士，副教授，主要研究方向：深度学习、强化学习、人工智能与大数据。

收稿日期： 2019-07-10

描述语句的多样性,使生成的描述不够灵活、新颖。

(2)基于检索的图像描述生成方法。该方法探寻文本与图像之间的关联[8-9],把文本和图像映射到一个共同语义空间。结合相似度[10-11]的计算方法,对图像内容和文本意义的关系程度进行排名,检索出和测试图像关系最接近的文本作为测试图像的最终文本描述。该方法把生成图像描述看作是一种检索任务,但检索前都需要调整和泛化过程,这无疑给描述任务又增加了处理过程和复杂度。

(3)基于深度学习的图像描述生成方法。目前主流的深度学习模型是端到端的训练方法。一方面采用多层深度卷积神经网络技术对图像中的物体特征概念建立模型;另一方面采用循环神经网络对文本建立模型。运用循环神经网络[12-15]进行建模,将文本信息与图像信息映射在同一个循环神经网络中,利用图像信息指导文本句子的生成。随着深度学习的研究进展,基于注意力机制和强化学习改进的研究方法[16-20]相继涌现,并不断推动图像描述生成模型的发展。该方法没有任何模板、规则的约束,能自动推断出测试图像和其相应的文本,自动地从大量的训练集中去学习图像和文本信息,生成更灵活、更新颖的文本描述,还能描述从未见过的图像内容特征。

1 数据集

大量免费公开的数据集用于图像描述研究,这些数据集中的图像与文本描述相关联,某些方面彼此不同,例如大小、描述的格式和描述词的长短。多种数据集信息汇总见表1。

表1 中,Flickr8K 数据集及其扩展版本 Flickr30K 数据集包含来自 Flickr 的图像,分别包含约 8 000 和 30 000 幅图像。这2个数据集中的图像是针对特定对象和动作的。这些数据集包含5个描述句子,每个图像是工作人员采用类似于 Pascal1K 数据集的策略收集的。

MSCOCO 数据集包括 123 287 幅图像,每幅图像均可给出5个不同的描述。此数据集中的图像包括80个对象类别,所有图像都可以使用这些类别中的所有实例。该数据集已被广泛用于图像描述,目前有研究者正在开发 MSCOCO 的扩展,包括增加问题和答案。

Flickr30K 和 MSCOCO 数据集举例如图1所示。

表1 多种数据集汇总信息

Tab. 1 Summary of multiple datasets

	Images	Texts	Judgments	Object
Pascal1K (Rashtchian et al., 2010)	1 000	5	No	Partial
VLT2K(Elliott & Keller,2013)	2 424	3	Partial	Partial
Flick8K(hodosh & Hockenmaier, 2013)	8 108	5	Yes	No
Flick30K(Young et al.,2014)	31 783	5	No	No
Abstract Scenes(Zitnick & parikh, 2013)	10 000	6	No	Complete
IAPR-TC12 (Grubinger et al., 2006)	20 000	1~5	No	Segmented
MSCOCO(Lin et al.,2014)	164 062	5	Soon	Partial
BBCNews(Feng & Lapata,2008)	3 361	1	No	No
SBU1M Captions(Ordonez et al., 2011)	1 000 000	1	Possibly	No
Deja-Image Captions (Chen et al.,2015)	4 000 000	Varies	No	No

1. A man is skateboard over a structure on the seaside.
2. A skateboarder jumps through the air on the seaside.
3. A skateboarder wearing blackpans doing a trick on the seaside.
4. Someone in black pants is on a ramp on the seaside.
5. The man is performing a trick on a skateboard high in the air.

(a) Flickr30K

1. A blue smart car parked in a parking lot.
2. Some vehicles on a very wet wide city street.
3. Serveral cars and a motorcycle are on a snow convered street.
4. Many vehicles drive down an icy street.
5. A small smart car driving in the city.

(b) MSCOCO

图1 Flickr30K 和 MSCOCO 数据集举例

Fig. 1 Examples of Flickr30K and MSCOCO datasets

2　图像描述生成方法

2.1　基于模板的图像描述

基于模板的方法需要包含多个填充模板，由对象关系和属性标签形成空槽，通过对空槽进行填充，生成图像描述句子。

Yao 等人提出使用场景三元组方法生成图像描述。利用某种特定的机器学习算法-邻相似度计算 2 个三元组之间的匹配程度。匹配程度越高，得分也就越高。提出模型可写作如下数学形式：

$$\begin{aligned}&\min_{w}\frac{\lambda}{2}\|w\|^{2}+\frac{1}{n}\sum_{i\in example}\xi_{i},\\&w\Phi(x_{i},y_{i})+\xi_{i}\geqslant\\&\qquad\max_{y\in meaningspace}w\Phi(x_{i},y)+L(y_{i},y),\\&\forall i\in examples,\\&\xi_{i}\geqslant 0\,\forall i\in examples.\end{aligned}\tag{1}$$

其中，λ 是正则化变量和松弛变量 ξ_i 之间的权衡因子；Φ 是本次研发的功能函数；x_i 对应着第 i 幅图像；y_i 对应着第 i 幅图像的结构化标签，使用随机梯度下降法来解决这个极小化问题。

模型是利用贪婪算法最大化这个三元组的得分：$\arg\max_y w^{\mathrm{T}}\Phi(x_i,y)$。一旦预测的图像和句子是一个非常相似的三元组，在意义空间的映射就会有很高的得分。最后，寻找具有很好的匹配分数的句子来生成最终的描述句子。Kulkarni 等人提出使用条件随机场算法（Conditional Random Field Algorithm，CRF）算法预测标签，并结合模板生成图像描述。Yang 等人利用隐马尔科夫模型选择可能的对象、动词、介词以及场景类型填充句子模板。

该方法较为直观，但却需要为图像中的物体、图像、场景和属性等指定准确的类别信息，工作量较大。生成过程受限于人工设计的有限模板，句子模板固定，生成的句子在自然、多样的特性上表现欠佳。因此，该方法在图像描述领域已逐渐退出人们的视野。

2.2　基于检索的图像描述

基于检索的图像描述方法类似于信息检索问题。在数据集中寻找查询图像 I 的相似子集，通过合理地组织图像对应的文本描述集，计算相似度，输出查询图像 I 的文本描述集。根据研究发现，基于检索的图像描述可以进一步细分为基于视觉空间的检索方法和基于多模态空间的检索方法。这里拟展开研究论述如下。

（1）基于视觉空间的检索方法，利用图像特征的相似性得到文本描述信息。如使用尺度不变特征转化（Scale Invariant Feature Transform，SIFT）描述作为图像特征，候选图像的文本描述划分为某类短语，如主语、宾语等，查询图像的最优描述，最后由图像相似性、谷歌搜索计数值以及图像三元组构成的联合概率确定。文献[21]用检索方法在一个相关文本集中搜索符合图像内容的文本描述。基于全局图像相似性的比较，该研究提出几种图像内容的比较：Object 和 Stuff。每种类型的内容都用于计算匹配图像和查询图像之间的相似性，根据相似度对匹配的图像进行排序。对于 Object 的匹配概率可表示为：

$$P(O_q,O_m)=e^{-D_o(O_q,O_m)}.\tag{2}$$

其中，$-D_o(O_q,O_m)$ 是查询目标 O_q 和匹配目标 O_m 之间的欧几里得距离。

对于在将查询图像区域 S_q 和匹配图像区域 S_m 探索到的类别，其 Stuff 的匹配度表示如下：

$$P(S_q,S_m)=P(S_q=s)*P(S_m=s).\tag{3}$$

其中，$P(S_q=s)$ 是查询图像区域 SVM 概率，$P(S_m=s)$ 是匹配图像区域 SVM 概率。

前文已经计算了查询图像内容和匹配图像内容的相似性，而后利用相关排序算法对相似度进行排序，从而得出描述最佳的句子。

（2）基于多模态空间的检索方法。是由 Hodosh 等人提出 KCCA 核函数提取高维特征的方法，使用最近邻方法进行检索，对候选文本加以排序，产生图像描述句子。该方法在图像语义特征提取方面也取得不错的效果，有很大的研究价值。

基于检索的图像描述方法相比于基于模板的方法具有更好的语言表现力、灵活性和创造性。对于一个图像，若要检索相似的图像，从对应的图像描述中提取可能有用的片段，使用检索到的文本片段做出新的图像描述。但是检索过程依赖于大规模的训练文本，增加复杂度，且产生的描述文本局限于训练文本。

2.3　基于深度学习的图像描述

基于深度学习的图像描述方法，典型的方法主要有基于注意力机制的方法和基于强化学习的方法。应用最广泛的模型是编-解码的图像描述生成模型。文中将给出其研究综述如下。

基于注意力机制[22]的方法的提出，极大地促进了图像描述生成的研究。Xu 等人提出 Hard 和 Soft 注意力机制，通过特征向量的加权平均输入到 LSTM[23] 的解码过程中，使得文本描述能够利用局部图像信息，提升图像描述性能。后来提出使用文

本信息改善局部注意力,采用 time-dependent 的 gLSTM 方法对语句中各单词的嵌入表示求平均并与图像嵌入相融合,用于 LSTM 生成文本描述序列。研究中,将图像当作源语言,把生成的句子当作目标语言,通过提取图像特征,将图像表示为维度固定的特征向量,再使用注意力 LSTM[23]将其翻译成句子。提出的最大化概率模型为:

$$\theta^* = \arg\max_{\theta} \sum_{(I,S)} \log p(S \mid I;\theta). \qquad (4)$$

其中,θ 是模型参数;I 是图像;S 代表句子;$S_0,\cdots,S_{T-1}$ 表示每个句子的单词。该模型为所有的单词产生一个概率分布。

基于强化学习的研究方法是近期智能控制领域应用最广的方法。Liu 等人提出的基于强化学习的策略梯度的图像描述方法,根据值函数对策略进行改进,选取最优策略。经过实验证明该方法生成的描述质量优于传统方法。深度强化学习[24-26]的融合极大地推动了图像描述生成的效果。将强化学习的奖惩机制[27]引入图像字幕任务中,可以通过抽取字幕来优化句子级评价标准,利用"策略网络"和"价值网络"[28]来共同预测每个时间步中的下一个单词。

基于深度学习的图像描述生成的主流是端到端的训练方法,生成的描述语句具有多样性,不依赖于单一的语言模板。不仅结构清晰明确、容易理解,而且训练速度和生成效果相当突出。

3 图像描述的挑战与难点

图像描述生成技术的研究经历了多个发展阶段并渐趋成熟,而且也已取得突破性的进步。深度学习技术的发展为图像描述领域打开一个新的局面。虽然图像描述生成技术表现出了强大的研发能力,但仍存在一定问题亟待解决,对此可做分述如下。

(1)描述文本信息的不完整。视觉特征的提取是生成图像文本描述的重要基础,包括图像类别、场景、对象及对象关系等。这些都依赖于目前还不成熟的计算机视觉技术。所以图像的视觉特征提取关键技术的提高是有待解决的关键问题和难点。

(2)复杂图像关注点的选取。图像中常存在多义和不确定的事物、隐式和显式的信息,如何充分利用图像特征和文本信息的融合特征,有效进行图像关注点的选取是图像描述中仍待解决的关键问题和难点。

(3)图像描述的泛化能力较低。从以往的研究中可以看出,对于同一个图像数据集中的图片进行测试时,效果往往是令人满意的。但是当采用随机的图片进行测试时,效果并不尽如人意。所以图像描述的泛化能力的提高是尚待解决的难题。

4 结束语

图像描述生成技术已广泛应用于新闻传播、智慧交通、智能家居、智能医疗等众多领域,现已成为各大顶尖科研机构综合研究实力的较量方式之一。

本文简述了图像描述生成任务的研究背景以及国内外研究现状;讨论了基于模板的图像描述生成方法、基于检索的图像描述生成方法和基于深度学习的图像描述生成方法。综前论述可以发现,图像描述生成技术正在向着更复杂、更灵活、更智能的方向发展。

针对图像描述面临的挑战与问题,未来可考虑结合更复杂的多任务或注意力机制,充分融合图像特征和语言特征向量。在图像描述文本信息不完整的问题上可考虑 3D 建模的方式对原 2D 数据进行映射处理,图像描述技术还可融入深度强化学习,使用无监督自主学习模型,在减少耗费资源的情况下,提升图像描述的性能。

参考文献

[1] HELMUT H. Building natural language generation systems[J]. Artificial Intelligence in Medicine,2001,22(3):277-280.

[2] YAO B Z, YANG Xiong, LIN Liang, et al. Image2text: Image parsing to text description[J]. Proceedings of the IEEE,2010,98(8): 1485-1508.

[3] 郭乔进,丁轶,李宁. 基于关键词的图像标注综述[J]. 计算机工程与应用,2011,47(30):155-158.

[4] FENG Yansong, LAPATA M. How many words is a picture worth? Automatic caption generation for news images [C]// Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden:ACL, 2010: 1239-1249.

[5] 康莹莹. 新闻图像内容与字幕文本协同识别与检索方法研究[D]. 哈尔滨:哈尔滨工业大学,2012.

[6] LITĂ L, PELICAN E. A low-rank tensor-based algorithm for face recognition[J]. Applied Mathematical Modelling, 2015, 39(3):1266-1274.

[7] KULKARNI G, PREMRAJ V, DHAR S, et al. Babytalk: Understanding and generating simple image descriptions[C]// 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Colorado Springs, Co, USA:IEEE,2011, 35(12):1601-1608.

[8] MITCHELL M, HAN Xufeng, DODGE J, et al. Midge: Generating image descriptions from computer vision detections [C]//Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon,

France：ACL，2012：747-756.
[9] ELLIOTT D，KELLER F. Image description using visual dependency representations [C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle，Washington，USA：ACL，2013：1292-1302.
[10] 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.
[11] KARPATHY A，LI Feifei. Deep visual-semantic alignments for generating image descriptions [C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston，MA，USA：IEEE，2015：3128-3137.
[12] 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 (TACL)，2014，2：207-218.
[13] CHEN X，ZITNICK C L. Mind's eye：A recurrent visual representation for image caption generation [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston，MA，USA：IEEE，2015：2422-2431.
[14] MAO Junhua，XU Wei，YANG Yi，et al. Deep captioning with multimodal recurrent neural networks (m-RNN)[J]. arXiv preprint arXiv：1412.6632，2014.
[15] XU Hongteng，WANG Wenlin，LIU Wei，et al. Distilled Wasserstein learning for word embedding and topic modeling [C]// 32nd Conference on Neural Information Processing Systems (NIPS) 31. Montréal，Canada：[s.n.]，2018：1-10.
[16] XU K，BA J，KIROS R，et al. Show，attend and tell：Neural image caption generation with visual attention [C]//32nd International Conference on Machine Learning. Lille，France：dblp，2015：2048-2057.
[17] 陈强普. 面向图像描述的深度神经网络模型研究[D]. 重庆：重庆大学，2017.
[18] 申永飞. 图像描述文本自动生成方法研究[D]. 重庆：重庆大学，2017.
[19] 陈龙杰，张钰，张玉梅，等. 基于多注意力多尺度特征融合的图像描述生成算法[J]. 计算机应用，2017，39(2)：354-359.
[20] 陈晨. 基于深度学习及知识挖掘的零样本图像分类[D]. 北京：中国矿业大学，2016.
[21] SILVER D，HUANG A，MADDISON C J，et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature，2016，529(7587)：484-489.
[22] XU K，BA J，COURVILLE R，et al. Show，attend and tell：Neural image caption generation with visual attention[J]. arXiv preprint arXiv：1502.03044v1，2015.
[23] 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)，Boston，MA，USA：IEEE，2015，1：3156-3164.
[24] LEIBFRIED F，TUTUNOV R，VRANCX P，et al. Model-based stabilisation of deep reinforcement learning[J]. arXiv preprint arXiv：1809.01906v1，2018.
[25] WANG Pin，CHAN C Y，LI Hanhan. Maneuver control based on reinforcement learning for automated vehicles in an interactive environment[J]. arXiv preprint arXiv：1803.09200，2018.
[26] WANG Jing，FU Jianlong，TANG Jinhui，et al. Show，reward and tell：Automatic generation of narrative paragraph from photo stream by adversarial training[C]// Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans，Louisiana，USA：AAAI，2018：7396-7403.
[27] LIU Xihui，LI Hongsheng，SHAO Jing，et al. Show，tell and discriminate：Image captioning by self-retrieval with partially Labeled data [M]//FERRARI V，HEBERT M，SMININCHISESCU C，et al. Computer Vision-ECCV 2018. Lecture Notes in Computer Science. Cham：Springer，2018，11219：353-369.
[28] REN Zhou，WANG Xiaoyu，ZHANG Ning，et al. Deep reinforcement learning-based image captioning with embedding reward[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu，HI，USA：IEEE，2017：1151-1159.

(上接第 44 页)

[5] 王伟. CAN FD 突破 CAN 总线应用局限[J]. 电子技术应用，2015，41(5)：3，13.
[6] ZAGO G M，FREITAS E P D. A quantitative performance study on CAN and CAN FD vehicular networks[J]. IEEE Transactions on Industrial Electronics，2018，65(5)：4413-4422.
[7] CENA G，BERTOLOTTI I C，HU Tingting，et al. Improving compatibility between CAN FD and legacy CAN devices[C]// 2015 IEEE 1st International Forum on Research & Technologies for Society & Industry Leveraging A Better Tomorrow (RTSI). Turin，Italy：IEEE，2015：1-8.
[8] 唐文俊，李维波，贺洪，等. 一种基于 ARM 的远程监控系统的设计与实现[J]. 船电技术，2011，31(11)：1-5.
[9] 孙乐鸣，江来，代鑫. 嵌入式 TCP/IP 协议栈 LWIP 的内部结构探索与研究[J]. 电子元器件应用，2008，10(3)：79-82.