软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(5):2152−2169 [doi: 10.13328/j.cnki.jos.006773]

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

基于视觉区域聚合与双向协作的端到端图像描述生成*

宋井宽, 曾鹏鹏, 顾嘉扬, 朱晋宽, 高联丽

(电子科技大学 计算机科学与工程学院, 四川 成都 611731)

通信作者: 宋井宽, E-mail: jingkuan.song@gmail.com

摘　要: 近几年, 基于 Transformer 的预训练模型展现了强大的模态表征能力, 促使了多模态的下游任务 (如图像描述生成任务) 正朝着完全端到端范式的趋势所转变, 并且能够使得模型获得更好的性能以及更快的推理速度. 然而, 该技术所提取的网格型视觉特征中缺乏区域型的视觉信息, 从而导致模型对对象内容的描述不精确. 因此, 预训练模型在图像描述生成任务上的适用性在很大程度上仍有待探索. 针对这一问题, 提出一种基于视觉区域聚合与双向协作学习的端到端图像描述生成方法 (visual region aggregation and dual-level collaboration, VRADC). 为了学习到区域型的视觉信息, 设计了一种视觉区域聚合模块, 将有相似语义的网格特征聚合在一起形成紧凑的视觉区域表征. 接着, 双向协作模块利用交叉注意力机制从两种视觉特征中学习到更加有代表性的语义信息, 进而指导模型生成更加细粒度的图像描述文本. 基于 MSCOCO 和 Flickr30k 两个数据集的实验结果表明, 所提的 VRADC 方法能够大幅度地提升图像描述生成的质量, 实现了最先进的性能.

关键词: 图像描述; 端到端训练; 预训练模型; 视觉区域聚合; 双向协作

中图法分类号: TP391

中文引用格式: 宋井宽, 曾鹏鹏, 顾嘉扬, 朱晋宽, 高联丽. 基于视觉区域聚合与双向协作的端到端图像描述生成. 软件学报, 2023, 34(5): 2152–2169. http://www.jos.org.cn/1000-9825/6773.htm

英文引用格式: Song JK, Zeng PP, Gu JY, Zhu JK, Gao LL. End-to-end Image Captioning via Visual Region Aggregation and Dual-level Collaboration. Ruan Jian Xue Bao/Journal of Software, 2023, 34(5): 2152–2169 (in Chinese). http://www.jos.org.cn/1000-9825/6773.htm

End-to-end Image Captioning via Visual Region Aggregation and Dual-level Collaboration

SONG Jing-Kuan, ZENG Peng-Peng, GU Jia-Yang, ZHU Jin-Kuan, GAO Lian-Li

(School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China)

Abstract: In recent years, Transformer-based pre-trained models have demonstrated powerful capabilities of modality representation, which leads to a shift towards a fully end-to-end paradigm for multimodal downstream tasks, such as image captioning tasks, and enables better performance and faster inference speed of models. However, the grid visual features extracted with such pre-trained models lack regional visual information, which results in inaccurate descriptions of the object content. Thus, the applicability of pre-trained models in image captioning remains largely unexplored. Therefore, this study proposes a novel end-to-end image captioning method based on visual region aggregation and dual-level collaboration (VRADC). Specifically, to learn regional visual information, this study designs a visual region aggregation module that aggregates grid features with similar semantics to obtain a compact visual region representation. Next, the dual-level collaboration module uses the cross-attention mechanism to learn more representative semantic information from the two visual features, which guides the model to generate more fine-grained image captions. The experimental results on the MSCOCO dataset and Flickr30k dataset show that the proposed VRADC-based method can significantly improve the quality of image captioning and achieves state-of-the-

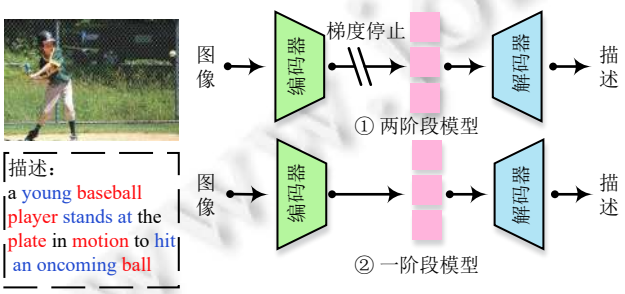
* 基金项目: 国家自然科技支撑计划 (2022YFC2009900/2022YFC2009903); 国家自然科学基金 (62122018, 62020106008, 61772116, 61872064)
本文由“融合预训练技术的多模态学习研究”专题特约编辑宋雪萌副教授、聂礼强教授、申恒涛教授、田奇教授、黄华教授推荐.
收稿时间: 2022-04-18; 修改时间: 2022-05-29, 2022-08-03; 采用时间: 2022-08-24; jos 在线出版时间: 2022-09-20
CNKI 网络首发时间: 2023-03-17

art performance.

Key words: image captioning; end-to-end training; pre-trained model; visual region aggregation; dual-level collaboration

随着智能终端和多媒体设备的广泛普及, 每天有数以万计的图像、视频等多媒体数据被上传到一些分享网站平台, 如 Instagram、Twitter、抖音、YouTube 等. 这些多媒体数据蕴含着丰富且有价值的信息, 因此如何使计算机能够自动地理解和分析多媒体数据内容为用户提供更好的交互体验, 已经成为当计算机学科中重要的研究热点. 图像描述生成[1−3]作为该研究热点中一项重要的基础任务, 其目的是根据所给定的一幅图像, 能够自动地生成一个与图像内容相符的自然语言描述的句子. 该任务有着广泛的应用前景: 基于内容的图像检索和推荐[4]、智能盲导系统[5]和人机交互[6,7]等. 尽管现有的计算机视觉和自然语言处理技术分别在视觉内容理解和语言语义分析方面取得了巨大的进步, 但由于图像描述生成既要理解丰富的视觉信息又要生成复杂的语义描述, 这一特性使得该任务极其具有挑战, 并且可被视为判别人工智能是否完成了从感知到认知的巨大飞跃的试金石.

图像描述生成的方法基本上都来源于神经机器翻译的编码器-解码器框架 (encoder-decoder)[8], 其中编码器利用视觉特征提取器 (如卷积神经网络 (CNN)[9]、物体检测网络[10]等) 来分析和提取图像的上下文视觉信息, 而解码器利用语言生成器 (如循环神经网络 (RNN)[11]、Transformer[12]等) 来生成自然语言描述. 目前大部分主流的图像描述生成方法属于两阶段的方法, 如图 1(a)①所示, 即采用预先训练好的视觉特征提取器来得到离线的视觉特征, 然后单独训练所提出的网络框架, 例如 RSTNet[3]、DLCT[13]等. 虽然上述的方法已经取得了显著的进步, 但是预先训练好的视觉特征提取器与下游的图像描述生成任务之间存在数据域和任务形式上的差异, 使得模型性能的提升存在很大的瓶颈. 除此之外, 提取离线的视觉特征相对费时, 难以应用于实时的图像描述生成场景.



(a) 两阶段与一阶段模型结构比较

训练方式	方法	特征提取器	CIDEr
两阶段模型	RSTNet	ResNeXt101	133.3
	RSTNet	ResNeXt152	135.6
	DLCT	ResNeXt101	133.8
一阶段模型	Transformer	Swin Transformer	136.4
	PureT	Swin Transformer	138.2
	VRADC (本文)	Swin Transformer	143.3

(b) 两阶段与一阶段模型性能比较

图 1 图像描述生成任务中两阶段与一阶段模型结构和性能的比较

最近, 由于基于 Transformer 的视觉预训练模型 (如 Vision Transformer[14]、Swin Transformer[15]等) 在图像分类、物体检测等视觉任务上取得了显著的发展, 完全基于 Transformer 的端到端图像描述生成模型 (即单阶段的方法, 如图 1(a)②所示) 成为一个新的研究趋势, 并且拥有更强大的性能. 该模型的优势有两点, 一是能够保证编码器和解码器模型架构上的统一, 二是能够同时对编码器和解码器进行参数优化. 从图 1(b) 中可以看出, 相比于两阶段模型, 一阶段模型的性能总体上有较大的提高, 在 CIDEr 指标[16]上平均提高了 3.7 个百分点左右. 尽管完全基于 Transformer 端到端的模型展现了巨大的潜力, 但是如何将预训练模型所得到的强大的视觉特征与语言单词进行对齐仍有待充分挖掘. 例如, Fang 等人[17]通过从给定图像中提取实体概念或属性表示来辅助文本生成, 或者 Wang 等人[18]在生成的单词嵌入和图像全局特征之间添加一个预融合操作来增加多模态特征之间的交互, 以提高从图像到字幕的推理能力.

现有的视觉特征表示主要分为两种类型, 即网格型视觉特征和区域型视觉特征. 其中, 网格型视觉特征是只以相同大小均分网格的形式将图像通过视觉特征提取器提取特征, 该特征能够覆盖整个图像的信息; 而区域型视觉特征是将图像输入到物体检测模型中得到图像中有代表性的对象区域特征. 上述端到端的图像描述生成方法采用 Vision Transformer 所提取的视觉特征都是网格型视觉特征, 缺少图像中明确的区域对象信息. 完整描述一幅图像的内容所涉及的语义信息比较复杂, 如图 1 所示. 该条描述语句的语义信息比较复杂, 其中细节性的单词信息

(如“young”“stands at”和“hit an oncoming”) 往往需要整个图像的上下文信息进行推理, 这个是网格型特征的优势; 而句子中的一些名词 (如“baseball player”“plate”“motion”和 “ball”) 往往对应于图像中某些物体对象信息, 这是区域型特征的优势. 一种解决方案是将端到端的物体检测技术也融入到模型, 但其模型设计复杂、计算成本过高等因素, 并不有利于端到端的模型训练及测试.

针对上述问题, 本文提出一种基于视觉区域聚合与双向协作的端到端图像描述生成方法, 图 2 是模型的整体架构. 首先, 本文采用 Vision Transformer 中从原始的图像中提取视觉网格特征作为初始的特征. 为了学习到紧凑的视觉区域表征, 本文提出了视觉区域聚合模块, 该模块采用了一些可学习的语义聚类将网格特征中有相同语义信息的特征聚合在一起, 每一个聚类中心隐式地代表了一种有代表性的视觉区域特征. 此外, 本文提出一个双向协作模块来学习视觉网格特征和区域特征之间的联系. 具体来讲, 本文采用交叉注意力机制, 在区域特征的引导下, 最终得到的视觉语义特征将拥有更高质量的视觉信息, 进而来提升生成描述的质量.

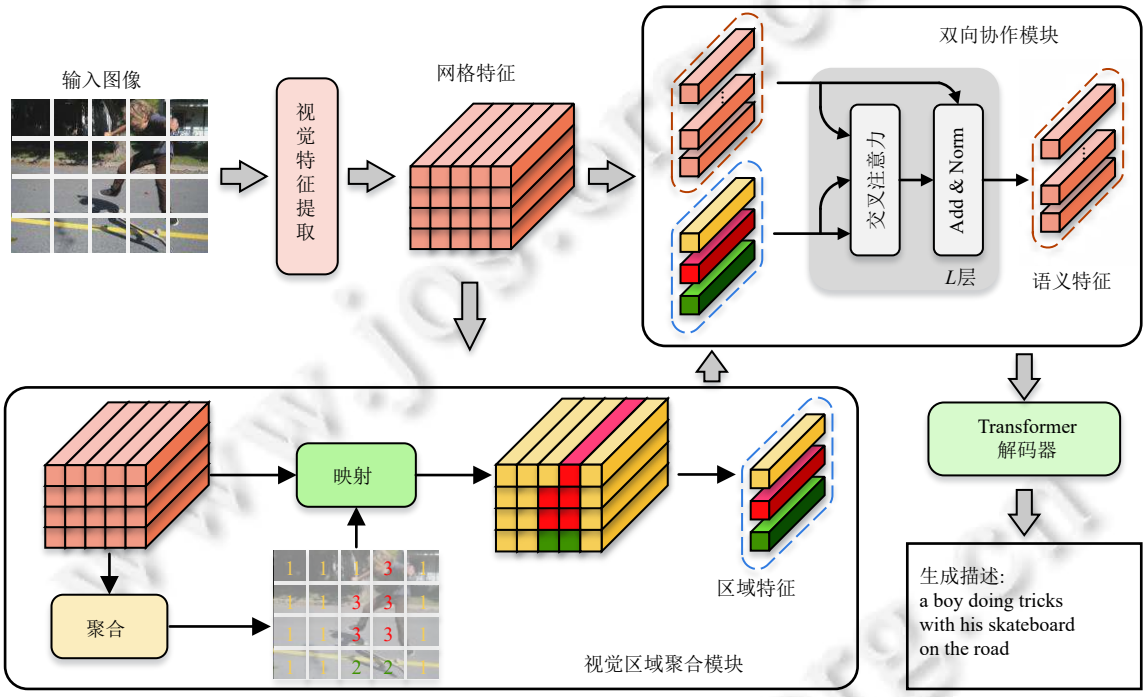


图 2 VRADC 方法概览

本文的主要贡献包括以下 3 个方面.

(1) 据本文所知, 本文是第 1 个在端到端的图像描述生成任务上考虑图像中区域对象信息的工作. 在没有显式监督的条件下, 本文提出一种视觉区域聚合模块, 从网格特征中隐式地学习到有区别、有代表性的区域特征.

(2) 本文提出了一个双向协作模块, 将所学习到的区域视觉特征整合到网格视觉特征中, 从而获得更具代表性的视觉语义特征去生成更加细粒度的描述.

(3) 本文所提出的 VRADC 方法在 MSCOCO 和 Flickr30k 两个公开数据集上进行了大量的实验验证, 实验结果表明本文的方法取得了最先进的结果, 并极大地超过了其他主流方法, 甚至优于一些大规模预训练的视觉与语言模型.

本文第 1 节介绍图像描述生成、视觉预训练模型和视觉与语言预训练模型的相关方法和研究现状. 第 2 节详细介绍基于视觉区域聚合与双向协作的图像描述生成方法. 第 3 节将介绍本文的实验设置, 包括数据集、训练细节等, 并将开展定量和定性的实验分析, 以证明本文所提出方法的有效性. 第 4 节总结全文, 并对未来研究方向做出展望.

1 相关工作

1.1 图像描述生成

图像描述生成是一个活跃的研究领域, 近年来该领域的工作层出不穷. 主流的图像描述生成方法大多基于编码器-解码器框架: 编码器用来提取图像特征, 解码器则利用图像编码器所提取好的图像特征来生成单词描述. 在本节, 我们根据图像描述生成模型的结构, 将图像描述生成的方法分为两大类, 即基于 RNN 的编码器-解码器框架和基于 Transformer 的编码器-解码器框架.

● 基于 RNN 的编码器-解码器框架. 早期由于基于 RNN 的编辑器-解码器架构在机器翻译领域显著改善了翻译的质量, 因此该架构被借鉴到图像描述生成任务上. 最初的工作[19−21]采用 CNN 来提取图像视觉信息, 利用 RNN 来解码生成最终的描述. 然而这些方法大多采用全局的视觉信息而丢失了图像中的局部特征, 导致解码器无法对图像细节进行精准解析. 人类在描述图像时会将注意力放在感兴趣的区域上, 受此启发, 研究者将注意力机制应用到基于 RNN 的编解码框架. Xu 等人[22]提出了基于 LSTM 的注意力机制模块, 该模块计算 LSTM 当前隐藏特征对图像特征中各个区域的注意力权重, 动态地选择与当前单词相关的图像区域作为解码器的视觉输入, 从而使得生成的文本能更准确地关注到显著的目标和细节. Chen 等人[23]提出了一种基于层级注意力的图像描述生成方法, 使用层级注意力自适应地选择 CNN 的卷积特征图来指导单词生成. Chen 等人[20]还提出了基于 RNN 的属性注意力模块, 帮助模型对图片中的属性信息进行预测, 并利用预测得到的属性信息辅助描述生成, 进一步提升了模型对高频率属性词汇的预测准确度. 后来, Anderson 等人[1]提出使用目标检测器 Faster R-CNN[10]作为编码器, 并用提取好的目标特征取代了以往 CNN 所提取的区域特征, 并提出自下而上和自上而下的注意力实现了区域对象层面的关注, 而不是传统的视觉注意力在等大小网格特征的关注. 此外, Yang 等人[24]将细粒度语义的场景图映入到基于 RNN 的编码器-解码器中, 旨在利用语言归纳偏差来提升图像描述生成的质量.

● 基于 Transformer 的编码器和解码器架构. 由于近期 Transformer 在自然语言领域取得了重大突破, 利用基于 Transformer 的编码器-解码器架构的图像描述生成方法也开始大量出现. 该架构的核心思想是采用 Transformer 中自我注意或交叉注意机制来加强视觉信息编码和视觉与语言之间信息交互. 为了使得 Transformer 的编码器和解码器结构能够更好地适应图像描述生成任务, 研究者们提出了各种改进方案. 其中, Pan 等人[25]提出了一个双线性注意力模块, 其利用空间和通道双线性注意分布来捕捉对象特征之间的交互信息. Song 等人[26]通过将相对位置信息嵌入到多头注意力机制中, 增强了视觉特征之间的方向感知能力. Cornia 等人[27]将不同层级对象特征之间的关系存储为先验知识, 并在解码阶段使用先验知识辅助生成不同层级的语义描述. 后来, Jiang 等人[28]重新审视了网格视觉特征的优势, 发现该特征在性能和时间成本方面都优于区域型视觉特征. Zhang 等人[3]在网格型视觉特征的基础上提出了网格增强模块来考虑网格位置之间的相对关系, 并且提出的适应性注意力模块在 Transformer 解码器的基础上会自适应地衡量视觉线索和语言线索的贡献以预测最终的单词.

然而, 上述方法都是属于两阶段的方法, 即先提取好图像的视觉特征, 然后再训练所设计的模型. 最近, 基于 Transformer 的视觉预训练模型在视觉分类任务上表现优异, 甚至超过了基于卷积网络的视觉预训练模型. 因此, 基于完全 Transformer 架构的图像描述生成方法被提出 (属于一阶段的方法). 这类方法能够很好地被用于实时场景, 也能够生成更加准确的图像描述句子. 其中, Wang 等人[18]首先将 Swin Transformer 应用到该领域, 并通过将全局视觉特征和当前预测单词的预融合操作, 增加了两个模态之间的交互. Fang 等人[16]则运用 Vision Transformer 作为编码器, 并且引入一个额外的模块来预测高频的概念单词, 以提高编码器对与概念相关的视觉网格特征的感知能力. 然后将所提取的概念单词特征合并到视觉网格特征中, 进一步改善视觉特征. 这些方法都能证实视觉 Transformer 所编码的网格特征蕴含丰富的语义信息, 适合图像描述生成等视觉语言下游任务. 此外完全基于 Transformer 编解码网络也实现了网络结构的统一.

1.2 视觉预训练模型

视觉语言任务的发展与视觉预训练模型密不可分, 视觉预训练模型为视觉语言的下游任务提供了语义丰富的

视觉特征. 早期的视觉语言工作采用传统的卷积神经网络 (例如 ResNet[29], VGG[30]) 作为视觉特征提取器. 通常而言, 这些 CNN 是在图像分类数据集 (如 ImageNet) 上采用分类损失进行预先训练而获得的, 但是由于分类数据集进对每张图片只有一个粗略的类别标注, 所以在此基础上所训练得到的图像特征信息并不丰富. 目标检测网络 (例如 Faster-RCNN[10]、YOLO[31]) 的出现解决了这一问题. 与分类任务不同, 目标检测网络的预训练任务和预训练数据集更加适合视觉语言的下游任务. 该任务在 Visual Genome 数据集[32]上进行训练, 原因是 Visual Genome 数据集使用了大量属性、目标对象等更细粒度信息的标注数据, 而这样的属性信息和类别信息正好是图像描述生成任务所关注的. 除此之外, 目标检测器处理图像的分辨率高于上述 CNN (例如 448×448 vs. 256×256), 因此获得的目标区域特征质量高于 CNN 所提取的网格特征. 至于最近提出的视觉 Transformer[14,15], 其将图像编码为一系列视觉令牌, 与传统的 CNN 结构相比, 基于 Transformer 模块的视觉 Transformer 具有更大的感受野, 并且没有归纳偏置. 因此, 它们在大规模多标签分类视觉数据集 (如 ImageNet-21k) 中取得了领先的性能. 此外, 基于 Transformer 的结构更加适合与视觉语言的下游任务相结合, 因此用视觉 Transformer 预训练模型作为图像特征提取器成为当前视觉语言任务的主流做法.

1.3 视觉与语言预训练模型

受预训练方法在视觉或者语言领域中大获成功的启发, 大规模的视觉与语言预训练模型也逐渐成为研究的热点. 一般而言, 视觉与语言预训练模型也是基于 Transformer 结构, 并为此设计一系列代理任务 (如视觉掩码、文本掩码、视觉文本匹配等) 来训练模型. 预训练所用到的数据集通常是大规模的视觉文本对的数据集, 如 Conceptual Captions[33]和 HowTo100M[34]等. 现有的视觉与语言预训练模型可以从视觉与文本信息融合的角度, 可分为单流和双流两种架构. 单流架构将文本和视觉特征拼接在一起, 然后传入一个 Transformer 块, 完成特征的融合, 代表的模型有 VisualBERT[35]、VideoBERT[36]、VinVL[37]、MDETR[38]等. 而双流架构则是将两种模态的特征分别传入各自分支的 Transformer 块, 随后使用各自分支的交叉注意力模块完成特征的融合, 代表的模型有 LXMERT[39]、ALIGN[40]、ALBEF[41]、Frozen[42]等.

除此之外, 视觉与语言预训练模型还可以根据整体架构设计的角度, 分为仅有编码器架构和编码器-解码器架构. 现有的视觉语言预训练模型大部分采用仅编码器架构, 将跨模态表示直接馈入输出层以生成最终输出, 如 ALIGN、VinVL、ALBEF. 相比之下, 采用使用转换器编码器-解码器架构的视觉语言预训练模型, 将跨模态的表示信息先输入到编码器然后传输到解码器进行解码工作, 如 MDETR、UniVL.

2 方 法

本文提出了一种基于视觉区域聚合和双向协作学习的端到端图像描述生成方法 VRADC, 该方法是完全基于 Transformer 的端到端编码器-解码器框架来构建模型. 具体而言, 给定一张原始图像 I, 首先采用预训练好的 Swin Transformer 作为视觉特征提取来提取图像的网格特征, 表示为 $G=\{g_1,g_2,\ldots,g_M\}$, 其中 M 表示图像中网格块的数量, g_m 表示第 m 个的网格特征; 为了学习到图像中的区域信息, 本文设计了一个视觉区域聚合模块, 将语义相似的网格特征聚合为紧凑的区域特征, 表示为 $R=\{r_1,r_2,\ldots,r_N\}$, 其中 N 表示聚合区域的数量, r_n 表示第 n 个的区域特征; 接着, 采用一个双向协作模块来学习网格特征和区域特征之间的关系, 并生成语义更加丰富完整的视觉语义特征, 表示为 $V=\{v_1,v_2,\ldots,v_M\}$; 最后, 将特征 V 输入到 Transformer 解码器中生成最终的图像描述 $Y=\{y_1,y_2,\ldots,y_T\}$, 其中 T 为生成描述的最长长度. 图 2 展示了本文所提出方法 VRADC 的整体框架图, 本节将对 VRADC 的具体实现细节进行详尽介绍.

2.1 视觉特征提取

现有的算法已经证明有代表性的视觉特征更有利于图像描述生成任务. 基于 Transformer 架构的模型以及其变体 (如 BERT[43]、GPT[44]等) 在自然语言处理任务上取得了显著的性能提升, 一些先驱工作也将其引入到计算机视觉任务中, 相比于基于 CNN 架构的模型, 获得了更优的性能以及更强的表征能力, 如 Vision Transformer、Swin Transformer 等. 本文采用性能强大的 Swin Transformer 作为视觉特征提取器, 它能够为本文建立起一个完全基于

Transformer 的框架, 并能从原始图像进行端到端的图像描述生成训练.

在本模块中, 首先将输入图像 $I \in R^{H\times W\times 3}$ 分割为 B 个不相交的补丁区域 (patch), 将局部区域记为 $I_p \in R^{p\times p\times 3}$. 其中 $\{H, W\}$ 和 $\{P, P\}$ 分别表示输入图像和补丁区域的大小, 并且 $N=(H\times W)/P^2$ 是补丁区域的数据, 也是输入到 Swin Transformer 中的有效序列长度. 然后将这些补丁进行平铺操作并输入到一个可训练的嵌入层得到补丁嵌入向量. 为了保留位置信息, 位置嵌入也被融合到补丁嵌入向量中. 接着, 补丁嵌入向量经过 4 个编码阶段, 每个阶段包含一个补丁特征融合层和多个相邻的 Swin Transformer 核心单元, 用于获得分层次的视觉表征. 其中, 补丁特征融合层通过将 2×2 个补丁区域特征进行拼接, 将局部特征的总规模缩小到原来的 1/4; 每个核心单元由基于移位窗口的多头自注意力模块、多层感知器模块、GELU 非线性层和归一化模块组成, 多个核心单元能在保持原本不重叠窗口有效计算的同时引入跨窗口的连接, 显著增强了整体模型的表征能力. 最终, 本文将 Swin Transformer 最后一个阶段输出的特征作为网格特征, 用 G 来表示, 并将其输入到下一个模块中.

2.2 视觉区域聚合

如前所述, 通过 Swin Transformer 所提取的网格特征存在区域空间信息的丢失, 从而导致模型对一些物体对象信息描述不准确. 一种可行的方案是将端到端的物体检测技术融入到模型中来弥补其不足. 然而, 物体检测技术由于计算过程复杂以及计算成本过高等因素, 并不利于端到端的模型训练及测试. 直观上, 如果能够隐式地将相同语义的网格特征选取并聚合成多个子空间中以获得伪区域特征, 操作将变得更加灵活. 为此, 本文设计一种视觉区域聚合模块, 在没有明确监督的情况下, 将网格特征聚合到多个聚类中心, 其中每一个聚类中心整合了网格特征中某种相似的语义信息, 表示一种可能的有代表性的区域.

对于该模块, 本文通过计算网格特征与聚类中心的距离来得到紧凑的区域特征 R. 具体而言, 首先设置了 N 个可学习的聚类中心, 表示为 $C=\{c_1, c_2, \ldots, c_N\}$. 接着, 给定第 m 个网格特征 g_m, 将计算该网格特征与第 n 个聚类中心 c_n 的相关性:

$$r_{m,n}=\frac{\exp(g_m c_n^{\mathrm{T}}+b_n)}{\sum_{k=1}^{N}\exp(g_m c_k^{\mathrm{T}}+b_k)} \tag{1}$$

其中, b_n 和 b_k 是可学习的参数. 第 n 个区域的特征 r_n 将通过对所有网格特征加权和得到:

$$r_n=Norm\left(\sum_{m=1}^{M} r_{m,n}(g_m-\tilde{c}_n)\right) \tag{2}$$

其中, $Norm$ 表示 l_2 范数归一化操作, $\tilde{c}_n$ 是一个与 c_n 大小相同的可学习参数. 因此, 我们将最终的特征 R 定义为区域特征.

2.3 双向协作

本文采用双向协作模块来建立网格特征和区域特征之间的联系, 从而得到更加丰富的视觉表征, 该模块主要采用了交叉注意力机制. 如图 2 所示, 首先采用不同全连接层将网格特征 G 映射查询 Q_G 以及将区域特征 R 映射为键 K_R 和值 V_R. 采用点积注意函数计算网格特征 G 与区域特征 R 之间的相似度, 用矩阵 S 表示:

$$S=Similarity(Q_G, K_R)=Softmax\left(\frac{Q_G K_R^{\mathrm{T}}}{\sqrt{d}}\right) \tag{3}$$

其中, $S(m, n)$ 表示 n 个区域特征 r_n 对第 m 个网格特征 g_m 的重视程度. 为了同时关注语义相关的区域信息, 我们进一步采用多头注意机制 (multi-head attention, MHA) 重新计算公式 (3) 中的相似度 S.

在多头注意机制中, Q_G, K_R 和 V_R 将会映射到多个不同的子空间, 然后通过公式 (3) 计算不同子空间的相似度 S^h, 将 MHA 所有的相似性头的输出连接在一起并通过非线性层得到一个特征 $\overline{P}$:

$$\begin{cases} P^h=S^h V_R^h,\ h=1,2,\ldots,H \\ \overline{P}=\left[P^h; P^h; \ldots; P^h\right]W^R \end{cases} \tag{4}$$

其中, H 为子空间的数量. 最后, 通过 Layer Normalization 被归一化, 并添加到一个原网格特征, 以产生一个与区域相关的网格特征 $\tilde{G}$:

$$\tilde{G} = G + LayerNorm\left(\overline{P}\right) \tag{5}$$

通过堆叠 L 层的 MHA 模块, 可以获得一个更精细化的视觉特征. 我们将最后一个 MHA 块的输出作为最终的视觉语义特征 V.

视觉语义特征将输入到 Transformer 的解码器中生成最终的描述 Y:

$$Y = TransformerDecoder(V) \tag{6}$$

2.4 模型训练

为了训练本文所提出的模型, 与主流的算法相同[1,3], 采用两阶段的训练策略: 1) 监督学习阶段, 使用交叉熵损失来学习生成单词和参考单词之间的差异; 2) 强化学习阶段, 使用句子的 CIDEr 指标作为强化学习的奖励来优化模型的生成效果.

具体来说, 在第 1 阶段, 通过给定图像的参考描述的句子 $Y^* = \left\{y_1^*, y_2^*, \ldots, y_T^*\right\}$, 最小化模型的交叉熵损失:

$$L_{CE} = -\sum_{t=1}^{T} \log(p_\theta(y_t^* | y_{1:t-1}^*)) \tag{7}$$

其中, θ 表示模型的参数. 然后, 对于一阶段的训练, 模型预测当前时间步的单词在很大程度上依赖于之前生成的单词, 一旦模型在前面时间步生成的单词不准确, 会导致误差累积, 从而导致模型生成的描述准确性不高. 此外, 一阶段的训练损失与评价标准不相关, 即模型的训练和测试目标不一致.

为解决上述问题, 在第 2 阶段, 采用 SCST[45]的思想, 将强化学习引入到模型训练过程中, 对 CIDEr 指标作为奖励进行优化, 目标是最大限度地减少负奖励期望:

$$L_{RL} = -E_{y_{1:T} \sim p_\theta}\left[r(y_{1:T})\right] \tag{8}$$

其中, $r(\cdot)$ 表示生成描述 $y_{1:T}$ 的 CIDEr-D 得分. 通过蒙特卡洛方法来近似估计梯度 $\nabla_\theta L_{RL}$:

$$\begin{cases} \nabla_\theta L_{RL} \approx -\dfrac{1}{k}\displaystyle\sum_{i=1}^{k}\left(\left(r\left(y_{1:T}^i\right) - b\right)\right)\nabla_\theta \log p_\theta\left(y_{1:T}^i\right) \\ b = \dfrac{1}{k}\left(\displaystyle\sum_{i}^{k} r(y_i)\right) \end{cases} \tag{9}$$

其中, k 表示采样的序列数量, $y_{1:T}^i$ 表示第 i 个采样得到的描述序列, b 表示通过采样的序列得到的平均奖励值. 通过强化学习的策略, 能直接从指标出发, 优化描述的生成, 避免了只用一个交叉熵函数带来的可能陷入局部最优等问题.

3 实 验

为了充分验证所提出 VRADC 方法的有效性, 本文在 MSCOCO[46]和 Flickr30k[47]两个公开数据集上开展了大量的实验, 从定量和定性两个方面进行了实验结果分析, 并将 VRADC 与目前主流图像描述算法进行对比实验. 本节首先介绍实验数据集、评价标准以及实现细节, 其次对实验结果进行详细的介绍与分析.

3.1 数据集和评价标准

3.1.1 数据集

与主流的方法一样, 本文在 MSCOCO 和 Flickr30k 两个数据集上进行实验来评估所提出模型的有效性.

MSCOCO 数据集是当前图像描述生成任务中通用的大规模英文数据集, 总共有 164 062 张从在线网站上收集的图像. 为了方便离线训练和测试, 其中有 123 287 张图像公开提供了人类标注的句子描述, 每张图像至少有 5 条以上的参考描述. 本文采用 Karpathy 等人[48]所提供的划分方式, 将 MSCOCO 数据划分为训练集、验证集和测试集, 其中每个集合的图像数量分别为 113 287、5 000 和 5 000 张. 除此之外, MSCOCO 还提供了在线测试平台, 有 40 775 幅图像的人类标注的句子描述没有公开提供, 用于进一步衡量模型的性能.

Flickr30k 数据集是从 Flickr 网站上收集 31 783 张图像, 图像涵盖了多种人类活动、不同事件以及多种场景等视觉内容, 并且每张图片都会由网络用户生成 5 个英文描述的句子相匹配. 为了公平, 本文采用 Young 等人[47]

所提供的划分方式, 将 28 783 张图片用于训练, 1 000 张图片用于验证以及 1 000 张图片用于测试.

3.1.2 评价标准

为了公平评价, 本文采用了 5 种在图像描述任务中广泛使用的评价指标来衡量模型所生成描述的质量, 包括 BLEU@N[49]、METEOR[50]、ROUGE-L[51]、CIDEr[16] 和 SPICE[52]. 其中, BLEU@N 常用于衡量机器生成描述的准确性, 做法是比较参考描述与生成描述之间的 n-gram 重合程度, 重合程度越高, 则生成的描述越准确; METEOR 是 BLEU 的改进版, 是计算 unigram 加权的召回率和准确率, 用以评估句子间的序列、同义词、词根、词缀和释义等匹配关系; ROUGE-L 常用于评估文本摘要的质量, 利用句子之间的最长公共子串来计算准确率和召回率; CIDEr 对句子之间的每个 n-gram 执行 TF-IDF 加权, 并计算它们的 TF-IDF 权重向量之间的余弦相似度, 来衡量参考描述和生成描述之间的一致性, 评测图像描述一致性和丰富度; SPICE 是一种基于场景图结构的语义分析指标, 将参考描述和生成描述转化为句子场景图, 进而去分析句子之间的对象、属性以及它们之间的关系. 为了方便起见，BLEU@N、METEOR、ROUGE-L、CIDEr、SPICE 在实验表格中简写为 BN、M、R、C、S, 其中 N 可以取 1、2、3、4. 本文所提出的代码现已开发布在: https://github.com/jkdxg/VRADC.

3.2 实验设置

对于视觉特征提取, 本文实验使用 Swin Transformer 作为编码器的骨架, 由于该模型大小和图像的精度有不同的配置, 本文在模型规模上选择了 Swin-B 和 Swin-L, 两种规模 在图像分辨率上选择了 224×224 和 384×384 两种配置. Swin Transformer 输出的视觉特征将会映射到 512 维. 对于描述语句, 首先, 本文将所有的句子通过分词工具删除标点符号, 并划分成一个个单词, 然后将单词全部转换为小写字母, 最后去掉出现次数少于 5 次的单词, 最终得到 10 201 个单词作为最终的单词库. 双向协作模块和基于 Transformer 解码器的特征维度都是 512, 每层有 8 个自注意头, 层数内部的全连接层维度为 2 048, 双向协作和解码器的层数分别为 5 和 3.

对于模型训练, 使用 Adam 优化器进行参数优化, 分为监督学习和强化学习两个阶段. 在监督学习阶段, 本文设置总批量大小和迭代周期分别为 50 和 20. 对模型视觉特征提取的网络参数和其他部分参数分别设置初始学习率为 4×10^{-5} 和 4×10^{-4}, 在经过 10 轮迭代过后, 学习率每 3 轮衰减一次, 衰减率为 0.8. 在强化学习阶段, 设置总批量大小和迭代周期分别为 10 和 30, 固定模型的学习率为 2×10^{-5}. 在每轮训练结束后, 在验证集上进行评估模型的性能, 最终选择在验证集上具有最高 CIDEr 值的模型用于测试. 在测试阶段, 模型使用 Beam Search 方法[53]进行描述生成, Beam size 大小设置为 5.

3.3 实验结果定量分析

3.3.1 框架探索和比较

本文方法的核心是通过引入视觉区域聚合和双向协作学习来得到语义更加丰富的视觉信息已生成高质量的图像描述. 在本节中, 将开展充分的对比实验来证明本文方法的有效性, 主要研究以下几个问题.

- RQ1: 在视觉区域聚合模块中, 不同的聚类数量是否会影响实验结果?
- RQ2: 在双向协作模块中, 不同层数的交叉注意是否会影响实验结果?
- RQ3: 采用不同的视觉特征提取骨架是否会影响实验结果?
- RQ4: 在双向协作模块中, 采用不同注意力方式是否会影响实验结果?
- RQ5: 采用完全端到端的训练策略是否会影响实验结果?

RQ1: 在视觉区域聚合模块中, 不同的聚类数量是否会影响实验结果?

为了验证这一个问题的结果, 该实验设置不同聚类中心的数量 (N=5, 7, 9, 11, 13), 其中双向协作模块中的层数为 3, 视觉特征提取器为 Swin-B, 图像大小为 224×224, 实验结果见图 3.

注意的是, 所提出的视觉区域聚合模块是为了获得有代表性的视觉区域信息. 从图 3 中可以看出, 聚类数量为 9 时性能表现最好. 如果聚类数量太大, 可能会导致模型难以找到有代表性的区域, 从而影响模型的性能. 相反, 如果聚类数量太小, 很多弱语义视觉信息又会被大量丢弃, 导致结果不佳. 因此, 选择一个合适的参数对模型取得最优结果很重要, 最终的模型选择 9 作为聚类中心的数量.

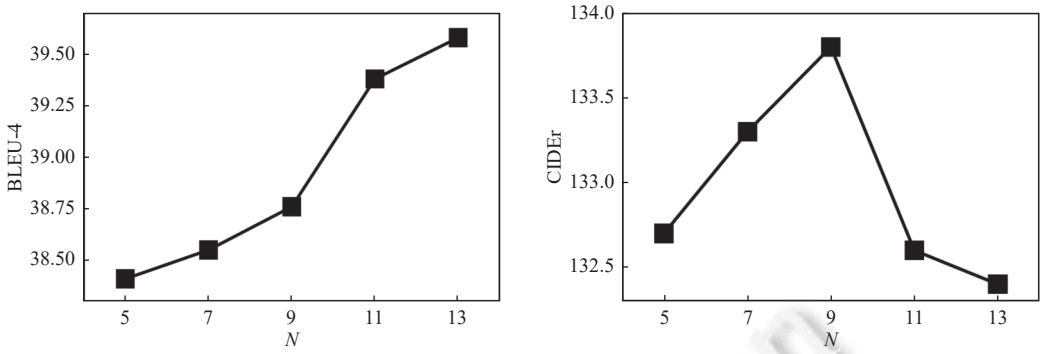


图 3 视觉区域聚合模块中不同聚类中心的影响

RQ2: 在双向协作模块中, 不同层数的交叉注意是否会影响实验结果?

为了回答这个问题, 在保持聚类中心为 9 的前提下, 该实验采用不同的交叉注意层数来验证 (L=1, 3, 5, 7, 9), 同样也采用视觉特征提取器为 Swin-B, 图像大小为 224×224, 为了更方便地进行实验证明, 实验结果见图 4. 从图中可以观察到, 不同层数对模型性能的影响比较小, 越大的层数也会引入更多的参数量, 为此, 双向协作模块的层数最终选择为 5.

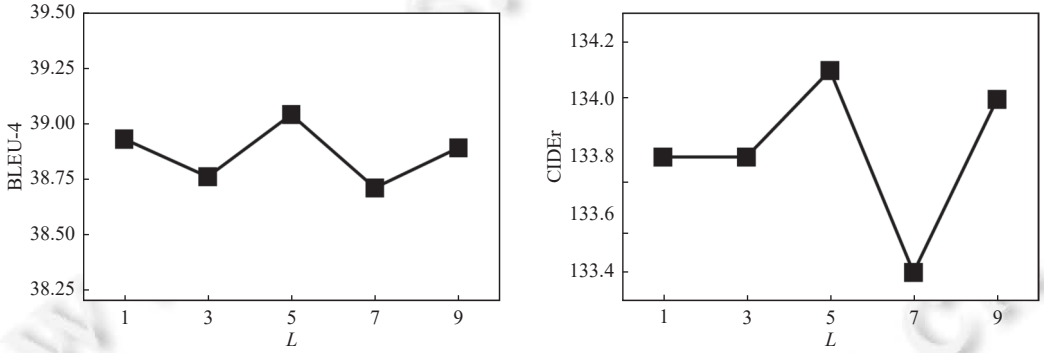


图 4 双向协作模块中不同交叉注意层数的影响

RQ3: 采用不同的视觉特征提取骨架是否会影响实验结果?

为了证实这一问题, 该实验采用了 Swin Transformer 模型中不同预训练设置来初始化视觉特征提取器, 主要有两个方面: 预训练模型的大小 (Swin-B 和 Swin-L) 和输入图像的分辨率 (224×224 和 384×384), 实验结果如表 1 所示, 除了呈现图像描述生成任务的性能外, 表 1 还列出了预训练模型在图像分类上原始性能的准确率 (Acc@1 和 Acc@5) 及参数量 (Params).

表 1 不同视觉特征提取器的影响

视觉特征提取	图像大小	图像描述任务					图像分类		
		B1	B4	M	R	C	Acc@1 (%)	Acc@5 (%)	Params (M)
Swin-B	224×224	81.02	39.04	29.16	58.55	134.1	85.2	97.5	88
Swin-B	384×384	82.75	41.30	30.28	59.82	141.5	86.4	98.0	88
Swin-L	224×224	82.48	40.64	29.73	59.43	139.5	86.3	97.9	197
Swin-L	384×384	**83.09**	**41.33**	**30.56**	**60.34**	**143.3**	**87.3**	**98.2**	197

从表 1 可以得出以下几个结论. (1) 预训练模型的性能对于本文所提出的方法是至关重要的, 当使用 Swin-L 作为视觉特征提取器, 模型能够带来更加显著的性能提升, 一个主要的原因是规模更大的模型在图像分类任务上

表现更佳, 具备更好的模态表征能力. (2) 关于输入图像的大小, 可以发现在相同预训练模型条件上, 图像分辨率从 224×224 增加到 384×384, 可以带来巨大的性能提升, 如 Swin-B 从 134.1 CIDEr 值到 141.5 CIDEr 值、Swin-L 从 139.5 CIDEr 值到 143.3 CIDEr 值.

RQ4. 在双向协作模块中, 采用不同注意力方式是否会影响实验结果?

为了在验证 VRADC 中不同的注意力方式在双向协作模块有效性, 对双向协作进行了消融实验, 如表 2 所示, 有两种注意方式: 并行注意力机制和交叉注意力机制, 其中并行注意力机制是指将网格型特征和学习到的区域特征一起输入到自注意力模块得到有代表性的特征, 而本文所采用的交叉注意力机制是将一种特征融入到另一种特征得到有代表性的特征, 有两种方式: 1) 交叉注意 (区域): 将网格特征融合到学习到的区域特征中得到最终有代表性的视觉特征; 和 2) 交叉注意 (网格): 将学习到的区域特征融合到网格特征中得到最终有代表性的视觉特征. 从表 2 中可以看出, 交叉注意 (网格) 的方式相比于其中两种方式在 CIDEr 指标上获得了最优, 因此, 本文最终的模型采用交叉注意 (网格) 的方式.

RQ5: 采用完全端到端的训练策略是否会影响实验结果?

本文一个核心观点是完全基于 Transformer 架构的端到端训练策略有助于图像描述生成任务. 为了验证该观点, 本文构造了对应的消融实验, 通过是否固定 Swin Transformer 的参数来训练所提出的模型, 实验结果如表 3 所示. 从表 3 中可以看出, 端到端的训练策略相比于非端到端的训练策略除了 B1 指标, 其他指标上均有所提升, 特别是在 CIDEr 指标上, 从 130.3 提升到 134.1, 表明使用端到端的训练策略的有效性.

表 2 不同注意力方式的影响

注意方式	B1	B4	M	R	C
并行注意	81.33	39.00	29.18	58.66	133.9
交叉注意(区域)	81.07	38.53	28.83	58.19	131.9
交叉注意(网格)	81.02	39.04	29.16	58.55	134.1

表 3 端到端训练策略的影响

训练策略	B1	B4	M	R	C
非端到端(加载预训练参数)	81.19	38.46	28.38	57.81	130.3
端到端(加载预训练参数)	81.02	39.04	29.16	58.55	134.1
端到端(未加载预训练参数)	64.7	21.89	20.24	46.19	65.9

除此之外, 表 3 还呈现了在端到端的训练策略下, 加载预训练好的模型参数对最终实验结果影响巨大. 其主要原因是使用预先训练好的 Swin Transformer 在大规模 ImageNet 图像数据集中进行过预训练, 从而该视觉特征提取器有着很好的图像表征能力. 然而, 若未加载预训练好的参数, 从头开始训练整个模型, 由于参数量过大而数据太少, 模型很难拟合, 从而导致最终模型性能表现差.

3.3.2 在 MSCOCO 数据集上的性能比较

在本实验中, 本文将所提出的方法 VRADC 与主流算法在 MSCOCO 数据集上进行比较来验证方法的有效性. 所比较的主流算法大致可以分为 3 类: 第 1 类为两阶段模型, 该类模型采用离线的视觉特征训练模型并生成描述, 包括有 Adaptive[54]、SCST[45]、Up-down[1]、VRCDA[55]、GCN-LSTM[56]、AoANet[57]、M^2 Transformer[27]、GET[58]、X-Transformer[25]、DRT[26]、RSTNet[3]、DLCT[13]. 第 2 类为一阶段模型, 该类模型输入为原始图像, 同时优化特征提取器和描述生成解码器, 包括有 ViTCAP[17]、PureT[18]. 第 3 类为视觉与语言预训练模型, 该类模型使用大规模的数据来学习通用的多模态模型表征, 包括有 Oscar[59]、VinVL[37]、SimVLM[60]. 本文的模型 VRADC 属于第 2 类, 即一阶段模型. 为了与上述方法进行公平比较, 本文在两种设置下进行实验对比, 包括离线测试以及在线测试.

1) 离线测试比较. 表 4 展现了本文的方法和上述方法在离线测试设置下的结果比较, 该表仅呈现单模型的比较结果. 从表中可以观察到以下几点: 第一, 与两阶段的模型相比, 一阶段的模型都实现了更优的性能, 特别是本文的方法 VRADC. 与 RSTNet (ResNeXt152) 相比, VRADC (Swin-L 384) 在 CIDEr 指标上提高了 7.7 个点; 第二, 与一阶段的模型相比, 本文的方法 VRADC 在所有评价指标方面均领先于其他算法; 第三, 与使用大规模多模态数据的视觉与语言预训练模型相比, 本文的方法仍然取得了优异的性能, 与 SimVLM 相比, VRADC 在 CIDEr 上提高了 0.8 个点. 总的来说, 上述实验结果充分显示了本文方法的潜力.

2) 在线测试比较. 为了进一步验证方法的有效性, 本文所提出的方法也提交到官方在线测试服务器与主流的算法进行比较. 与其他方法做法类似, 本文用相同的参数训练了 4 个 VRADC 模型, 然后将 4 个模型所生成的描述

集成在一起生成最终描述的句子. 本文测试了两种配置: Swin-B 224 和 Swin-L 384, 实验结果见表 5. 与所有其他方法相比, VRADC 采用 Swin-L 384 配置的模型在所有指标上仍然保持着最好的性能. 特别是, 在 CIDEr 的 c5 和 c40 指标上, VRADC (Swin-L 384) 比 PureT (Swin-L 384) 高出 4.6 和 5.1 个点. 除此, 本文采用 Swin-B 224 配置的模型也取得了与大多数主流方法可比较的性能.

表 4 离线测试环境下, VRADC 与其他主流算法的性能比较

模型	方法	视觉特征提取	B1	B4	M	R	C	S
两阶段模型	Adaptive	ResNet101	74.2	33.2	26.6	—	108.5	—
	SCST	ResNet101	—	34.2	26.7	55.7	114.0	—
	Up-dowm	F-RCNN101	79.8	36.3	27.7	56.9	120.1	21.4
	VRCDA	F-RCNN101	80.6	37.9	28.4	58.2	123.7	21.8
	GCN-LSTM	F-RCNN101	80.9	38.3	28.6	58.5	128.7	22.1
	AoANet	F-RCNN101	80.2	38.9	29.2	58.8	129.8	22.4
	M^2 Transformer	F-RCNN101	80.8	39.1	29.2	58.6	131.2	22.6
	GET	F-RCNN101	81.5	39.5	29.3	58.9	131.6	22.8
	X-Transformer	F-RCNN101	80.9	39.7	29.5	59.1	132.8	23.4
	DRT	F-RCNN101	81.7	40.4	29.5	59.3	133.2	23.3
	RSTNet	ResNext101	81.1	39.3	29.4	58.8	133.3	23.0
	DLCT	ResNext101	81.4	39.8	29.4	59.1	133.8	23.0
	RSTNet	ResNext152	81.8	40.1	29.8	59.5	135.6	23.3
一阶段模型	VITCAP	ViT-B 384	—	40.3	29.4	59.5	133.6	23.3
	PURET	Swin-L 384	82.1	40.9	30.2	60.1	138.2	24.2
	VRADC (本文)	Swin-B 224	81.0	39.0	29.2	58.5	134.1	23.0
	VRADC (本文)	ViT-B 384	81.5	39.7	29.6	59.0	135.3	23.3
	VRADC (本文)	Swin-L 384	**83.09**	**41.3**	**30.6**	**60.3**	**143.3**	24.3
视觉与语言预训练模型	OSCAR	ResNeXt152	—	41.7	30.6	—	140.0	24.5
	VINVL	ResNeXt152	—	41.0	31.1	—	140.9	25.2
	SIMVLM	ViT-L	—	40.3	33.4	—	142.6	**24.7**

表 5 在线测试环境下, VRADC 与其他主流算法的性能比较

方法	B1		B2		B3		B4		M		R		C	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
SCST	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7
Up-down	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
GCN-LSTM	80.8	95.2	65.5	89.3	50.8	80.3	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
AoANet	81.0	95.0	65.8	89.6	51.4	81.3	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
M^2 Transformer	81.6	96.0	66.4	90.8	51.8	82.7	39.7	72.8	29.4	39.0	59.2	74.8	129.3	132.1
GET	81.6	96.1	66.5	90.9	51.9	82.8	39.7	72.9	29.4	38.8	59.1	74.4	130.3	132.5
X-Transformer	81.3	95.4	66.3	90.0	51.9	81.7	39.9	71.8	29.5	39.0	59.3	74.9	129.3	131.4
DRT	82.7	96.5	67.7	91.5	53.1	83.4	40.9	73.6	29.6	39.0	59.8	75.0	132.2	133.9
RSTNet (ResNeXt101)	81.7	96.2	66.5	90.0	51.8	82.7	39.7	72.5	29.3	38.7	59.2	74.2	130.1	132.4
RSTNet (ResNeXt152)	82.1	96.4	67.0	91.3	52.2	83.0	40.0	73.1	29.6	39.1	59.5	74.6	131.9	134.0
DLCT (ResNeXt101)	82.0	96.2	66.9	91.0	52.3	83.0	40.2	73.2	29.5	39.1	59.4	74.8	131.0	133.4
DLCT (ResNeXt152)	82.4	96.6	67.4	91.7	52.8	83.8	40.6	74.0	29.8	39.6	59.8	75.3	133.4	135.4
PureT (Swin-L 384)	82.8	96.5	68.1	91.8	53.6	83.9	41.4	74.1	30.1	39.9	60.4	75.9	136.0	138.3
VinVL*	81.9	96.9	66.9	92.4	52.6	84.7	40.4	74.9	30.6	**40.8**	60.4	**76.8**	134.7	138.7
VRADC (Swin-B 224)	81.9	96.1	66.5	90.7	51.7	82.1	39.4	71.8	29.0	38.3	58.7	73.4	130.4	132.7
VRADC (Swin-L 384)	**83.6**	**97.5**	**68.9**	**93.3**	**54.3**	**85.7**	**42.0**	**76.2**	**30.7**	40.6	**60.6**	76.3	**140.6**	**143.4**

3) 模型的参数量、计算复杂度以及推理速度比较. 除了对图像描述生成句子的质量评估外, 本文还开展了对计算量 (FLOPS)、推理速度以及模型参数量的比较. 本文方法 VRADC 与 VITCAP 同属于一阶段模型, 都是直接从原始图像生成最终图像描述的句子, 而 OSCAR、RSTNet、DLCT 属于两阶段的方法, 包括视觉特征提取和模型预测. 为了公平比较, 两阶段的方法将两个阶段的计算量、推理速度以及模型参数量进行累加, 并且上述的 4 个模型全部都在统一的环境下进行测试, 实验结果如图 5 所示.

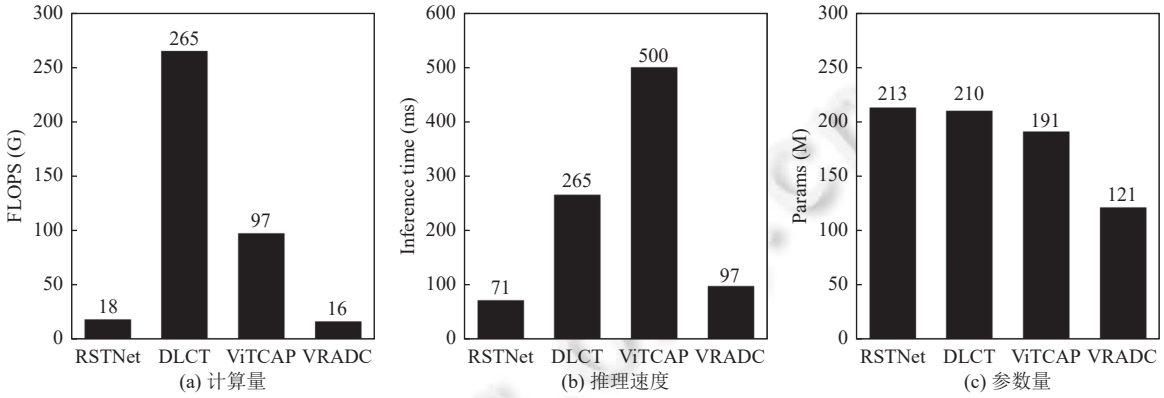


图 5 VRADC 与主流算法在计算量、推理速度、参数量上的比较

从图 5 可以观察到以下几点: 1) VRADC 在计算量 FLOPs 上是最少的 (如图 5(a)), 说明所提出的模型复杂度低; 2) VRADC 在推理速度上取得了第二的成绩 (如图 5(b)), 与第一名 RSTNet 相差了 26 ms, 但远低于另外两个方法, RSTNet 属于两阶段的模型在实际使用中会存在数据存取、转换等操作, 会有额外的耗时, 而本文的方法并不会存在上述问题; 3) VRADC 在参数量上是最小的 (如图 5(c)), 说明所提出的模型能够以较小的存储空间部署在设备上. 总体而言, 本文所提出的方法 VRADC 在计算量、推理速度以及参数量三者中权衡的最好, 该实验结果进一步地证明本文方法的优越性.

3.3.3 在 Flickr30k 数据集上的性能比较

在本节中, 本文将提出的方法与主流算法在 Flickr30k 数据集上进行比较来验证所提出方法的有效性, 所比较的方法主要有: Adaptive[54]、Soft-Attention[22]、Hard-Attention[22]、SCA-CNN[23]、Semantic-Attention[61]、NBT[62]、A_R_L[63]、IVAIC[64]、VRCDA[55], 实验结果如表 6 所示. 由表 6 可以看出, 本文所提出的方法 VRADC 相较于其他主流算法仍然展现了卓越的性能, 在各项指标上都达到最优的性能表现, 该实验结果进一步证明了 VRADC 方法的有效性.

表 6 VRADC 与其他主流算法在 Flickr30k 数据集上的性能比较

方法	B1	B4	M	R	C
Soft-Attention	66.7	19.1	18.5	—	—
Hard-Attention	66.9	19.9	18.5	—	—
Adaptive	67.7	25.1	20.4	—	53.1
SCA-CNN	66.2	22.3	19.5	—	—
Semantic-Attention	64.7	23.0	18.9	—	—
NBT	69.0	27.1	21.7	—	57.5
A_R_L	69.8	27.7	21.5	48.5	57.4
IVAIC	70.8	30.6	22.5	49.8	63.0
VRCDA	73.2	30.6	22.7	50.6	66.0
VRADC	**74.9**	**32.6**	**33.4**	**51.9**	**74.9**

3.4 实验结果定性分析

为了更好地定性评估所得到的视觉表征, 本实验可视化了视觉特征对最终输出单词的重要性. 视觉注意可视化的结果在图 6 所示, 每一个例子包括一张原始图像、基线模型和本文所提算法 VRADC 在每个时间步生成的单词以及注意机制的聚焦图. 从技术上讲, 本文选择了 Transformer 解码器最后一层的交叉注意权重来呈现.

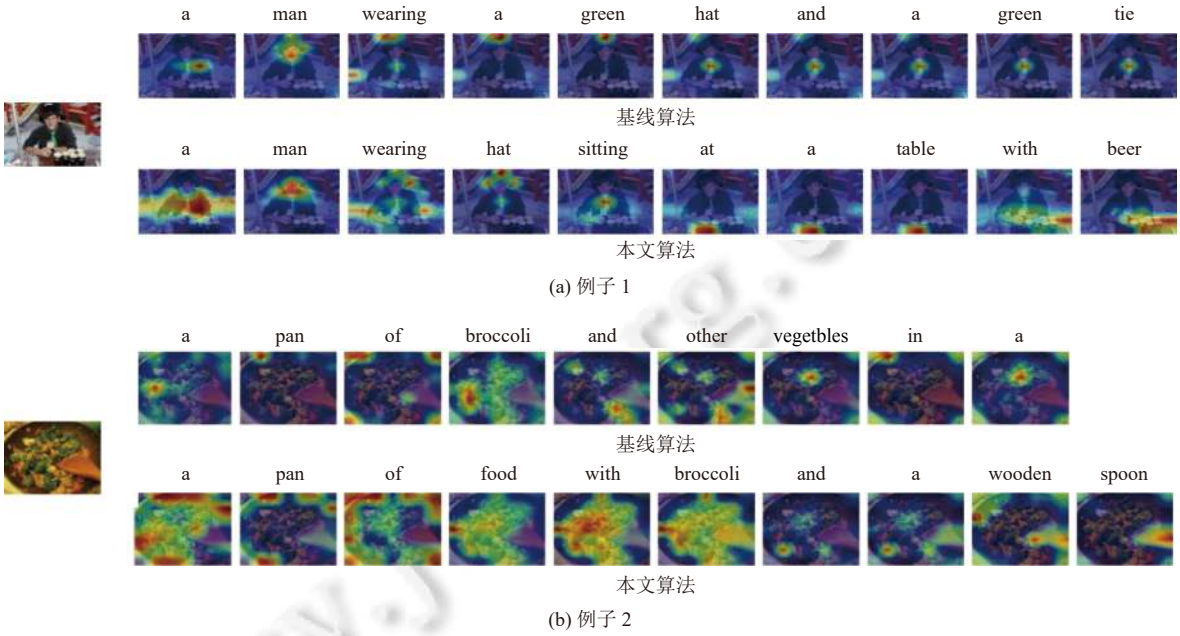


图 6 视觉注意的可视化

从图 6 可以看出, 基准模型和本文模型在生成单词时能够关注到对应的区域, 同时, 本文的生成描述在关注区域、对象内容、细节以及上下文连贯性上明显优于基准模型. 如图 6 的第 1 个例子, 本文算法生成的描述是“a man wearing hat sitting at a table with beer”, 该描述能精确地陈述人所处的环境, 而不是像基准模型生成的描述是“a man wearing a green hat and a green tie”, 该描述只是粗略地描述人的外貌信息. 同时, 本文的算法能够准确地关注“hat”的对应区域. 除此之外, 如图 6 的第 2 个例子可以看出本文算法 VRADC 生成的描述在语言层面更加平滑流畅, 符合人类表述习惯.

为了更直观地证明本文的模型能够生成更加细粒度的描述语句, 图 7 展示了本文方法 VRADC 额外生成的一些描述与参考描述的对比. 从图中可以看出, 本文的模型所生成的描述在细节、内容等方面都表现良好, 甚至一些句子的描述好于参考描述. 除此之外, 图 7 通过热力图来表示视觉区域聚合模型学习到的区域信息. 热图中, 不同颜色代表不同的指标值, 代表不同的区域信息. 正如图中所示, 视觉区域聚合模型不仅关注到了图像中特定前景的视觉区域, 而且也保留了有鉴别性的背景信息, 进一步证实了本文的方法能够有效地学习空间的区域信息.

除此之外, 图 8 还展示了 VRADC 与 Transformer 生成的一些描述示例对比, 从图 8 的实验结果中, 可以总结出以下几点结论.

1) 本文提出的 VRADC 模型相比于 Transformer 模型, 能够更加准确地对物体对象信息生成描述, 如第 1 幅图中的木质砧板 (“wooden cutting board”) 和第 2 幅图中的两个碗 (“two bowls”).

2) 本文提出的 VRADC 模型所生成的句子在主观上更加符合人类的描述习惯, 如第 3 幅图中的“一个人在有牛的田野里骑马” (“A person riding a horse in a field with cow”) 更加连贯.

3) 本文提出的 VRADC 模型所生成的句子多样性更加丰富, 如第 4 幅图, 表述了小男孩戴了一个帽子 (“wearing a hat”) 以及站在大巴前面 (“in front of a bus”).

图片				
热力图				
真实语句	A young girl inhales with the intent of blowing out a candle	A very cute brown dog with a disc in its mouth	A cat looking at his reflection in the mirror	A cat that is laying on a computer keyboard
VRADC	Two women sitting at a table with a bowl of food	A dog running with a purple frisbee in its mouths	A siamese cat looking at its reflection in a mirror	A cat laying on top of a computer keyboard

图 7 VRADC 生成描述与参考描述实例分析

图片				
真实语句	A square pizza on a wooden cutting board	Two bowls filled with broccoli soup on top of a table	A man on a horse in a field with cows and a dog	A young boy in a sweatshirt and baseball cap by a bus
Transformer	A pizza sitting in a box on a table	A bowl of soup and a plate of broccoli on a table	A group of cows grazing in a field of grass	A young boy holding a baseball bat
VRADC	A pizza sitting on top of a wooden cutting board	Two bowls of soup and broccoli on a wooden table	A person riding a horse in a field with cows	A young boy wearing a hat standing in front of a bus

图 8 VRADC 与 Transformer 生成描述实例分析

4 总结与展望

本文针对端到端的图像描述生成任务, 提出了一种新的完全基于 Transformer 的方法, 即基于视觉区域聚合和双向协作 VRADC. 该方法提出了一种简单有效的视觉区域聚合方法, 在没有任何显式监督的情况下, 隐式地从网格视觉特征中学习到有代表性的区域视觉信息. 然后, 为了将区域视觉信息融入到网格视觉特征中, 本文提出了一种双向协作的方式, 得到语义更加丰富的视觉信息, 以此来引导模型生成更加符合图像内容且质量更高的描述. 本文在 MSCOCO 和 Flickr30k 两个数据集上进行实验分析, 与主流的算法相比较. 实验结果表明本文所提出的方法 VRADC 在各项指标以及测试环境下实现了最先进的性能, 甚至超过了一些大规模预训练的视觉和语言模型. 定量和定性的实验分析也验证了本文所提方法的有效性.

未来计划将设计更加有效的视觉区域聚合模块来学习图像中的区域信息, 以及优化双向协作模块来提高两种特征融合方法, 以生成更加细粒度的图片描述. 此外, 还考虑加入一些预训练的语言模型或者知识图谱等, 以增强模型的可解释性, 以及引入端到端模型的轻量化设计以方便部署到移动智能终端设备.

References:

[1] Anderson P, He XD, Buehler C, Teney D, Johnson M, Gould S, Zhang L. Bottom-up and top-down attention for image captioning and visual question answering. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE,

2018. 6077–6086. [doi: 10.1109/CVPR.2018.00636]

[2] Gao LL, Li XP, Song JK, Shen HT. Hierarchical LSTMs with adaptive attention for visual captioning. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2020, 42(5): 1112–1131. [doi: 10.1109/TPAMI.2019.2894139]

[3] Zhang XY, Sun XS, Luo YP, Ji JY, Zhou YY, Wu YJ, Huang FY, Ji RR. RSTNet: Captioning with adaptive attention on visual and non-visual words. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 15460–15469. [doi: 10.1109/CVPR46437.2021.01521]

[4] Cui H, Zhu L, Li JJ, Yang Y, Nie LQ. Scalable deep hashing for large-scale social image retrieval. IEEE Trans. on Image Processing, 2019, 29: 1271–1284. [doi: 10.1109/TIP.2019.2940693]

[5] Wu SM, Wieland J, Farivar O, Schiller J. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In: Proc. of the 2017 ACM Conf. on Computer Supported Cooperative Work and Social Computing. Portland: ACM, 2017. 1180–1192. [doi: 10.1145/2998181.2998364]

[6] Das A, Kottur S, Gupta K, Singh A, Yadav D, Moura JMF, Parikh D, Batra D. Visual dialog. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 1080–1089. [doi: 10.1109/CVPR.2017.121]

[7] Jain U, Schwing AG, Lazebnik S. Two can play this game: Visual dialog with discriminative question generation and answering. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 5754–5763. [doi: 10.1109/CVPR.2018.00603]

[8] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Proc. of the 27th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2014. 3104–3112.

[9] Albawi S, Mohammed TA, Al-Zawi S. Understanding of a convolutional neural network. In: Proc. of the 2017 Int'l Conf. on Engineering and Technology. Antalya: IEEE, 2017. 1–6. [doi: 10.1109/ICEngTechnol.2017.8308186]

[10] Ren SQ, He KM, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proc. of the 28th Int'l Conf. on Neural Information Processing Systems. Montreal: NIPS, 2015. 91–99.

[11] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]

[12] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 5998–6008.

[13] Luo YP, Ji JY, Sun XS, Cao LJ, Wu YJ, Huang FY, Lin CW, Ji RR. Dual-level collaborative transformer for image captioning. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI Press, 2021. 2286–2293. [doi: 10.1609/aaai.v35i3.16328]

[14] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16×16 words: Transformers for image recognition at scale. arXiv:2010.11929, 2020.

[15] Liu Z, Lin YT, Cao Y, Hu H, Wei YX, Zhang Z, Lin S, Guo BN. Swin Transformer: Hierarchical vision transformer using shifted windows. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 9992–10002. [doi: 10.1109/ICCV48922.2021.00986]

[16] Fang ZY, Wang JF, Hu XW, Liang L, Gan Z, Wang LJ, Yang YZ, Liu ZC. Injecting semantic concepts into end-to-end image captioning. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 17988–17998. [doi: 10.1109/CVPR52688.2022.01748]

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

[18] Wang YY, Xu JG, Sun YF. End-to-end transformer based model for image captioning. In: Proc. of the 36th AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI Press, 2022. 2585–2594. [doi: 10.1609/aaai.v36i3.20160]

[19] Xue ZY, Guo PY, Zhu XB, Zhang NG. Image description method based on generative adversarial networks. Ruan Jian Xue Bao/Journal of Software, 2018, 29: 30–43 (in Chinese with English abstract). http://www.jos.org.cn/jos/article/abstract/18015?st=search

[20] Chen H, Ding GG, Lin ZJ, Zhao SC, Han JG. Show, observe and tell: Attribute-driven attention model for image captioning. In: Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence. Stockholm: IJCAI.org, 2018. 606–612. [doi: 10.24963/ijcai.2018/84]

[21] Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(4): 652–663. [doi: 10.1109/TPAMI.2016.2587640]

[22] Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel RS, Bengio Y. Show, attend and tell: Neural image caption generation with visual attention. In: Proc. of the 32nd Int'l Conf. on Machine Learning. Lille: JMLR.org, 2015. 2048–2057. [doi: 10.5555/3045118.3045336]

[23] Chen L, Zhang HW, Xiao J, Nie LQ, Shao J, Liu W, Chua TS. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6298–6306.

[doi: 10.1109/CVPR.2017.667]

[24] Yang X, Tang KH, Zhang HW, Cai JF. Auto-encoding scene graphs for image captioning. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 10677–10686. [doi: 10.1109/CVPR.2019.01094]

[25] Pan YW, Yao T, Li YH, Mei T. X-linear attention networks for image captioning. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10968–10977. [doi: 10.1109/CVPR42600.2020.01098]

[26] Song ZL, Zhou XF, Dong LH, Tan JL, Guo L. Direction relation transformer for image captioning. In: Proc. of the 29th ACM Int'l Conf. on Multimedia. Virtual Event: ACM, 2021. 5056–5064. [doi: 10.1145/3474085.3475607]

[27] Cornia M, Stefanini M, Baraldi L, Cucchiara R. Meshed-memory transformer for image captioning. In: Proc. of the 2020 IEEE Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10575–10584. [doi: 10.1109/CVPR42600.2020.01059]

[28] Jiang HZ, Misra I, Rohrbach M, Learned-Miller E, Chen XL. In defense of grid features for visual question answering. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10264–10273. [doi: 10.1109/CVPR42600.2020.01028]

[29] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

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

[31] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 779–788. [doi: 10.1109/CVPR.2016.91]

[32] Krishna R, Zhu YK, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, LI LJ, Shamma DA, Bernstein MS, Li FF. Visual genome: Connecting language and vision using crowdsourced dense image annotations. Int'l Journal of Computer Vision, 2017, 123(1): 32–73. [doi: 10.1007/s11263-016-0981-7]

[33] Sharma P, Ding N, Goodman S, Soricut R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: ACL, 2018. 2556–2565. [doi: 10.18653/v1/P18-1238]

[34] Miech A, Zhukov D, Alayrac JB, Tapaswi M, Laptev I, Sivic J. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 2630–2640. [doi: 10.1109/ICCV.2019.00272]

[35] Li LH, Yatskar M, Yin D, Hsieh CJ, Chang KW. VisualBERT: A simple and performant baseline for vision and language. arXiv:1908.03557, 2019.

[36] Sun C, Myers A, Vondrick C, Murphy K, Schmid C. VideoBERT: A joint model for video and language representation learning. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 7464–7473. [doi: 10.1109/ICCV.2019.00756]

[37] Zhang PC, Li XJ, Hu XW, Yang JW, Zhang L, Wang LJ, Choi YJ, Gao JF. VinVL: Revisiting visual representations in vision-language models. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 5575–5584. [doi: 10.1109/CVPR46437.2021.00553]

[38] Kamath A, Singh M, LeCun Y, Synnaeve G, Misra I, Carion N. MDETR-modulated detection for end-to-end multi-modal understanding. In: Proc. of the 2021 IEEE Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 1760–1770. [doi: 10.1109/ICCV48922.2021.00180]

[39] Tan H, Bansal M. LXMERT: Learning cross-modality encoder representations from transformers. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019. 5100–5111. [doi: 10.18653/v1/D19-1514]

[40] Jia C, Yang YF, Xia Y, Chen YT, Parekh Z, Pham H, Le QV, Sung YH, Li Z, Duerig T. Scaling up visual and vision-language representation learning with noisy text supervision. In: Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 4904–4916.

[41] Li JN, Selvaraju R, Gotmare A, Joty S, Xiong CM, Hoi SCH. Align before fuse: Vision and language representation learning with momentum distillation. Advances in Neural Information Processing Systems, 2021, 34: 9694–9705.

[42] Bain M, Nagrani A, Varol G, Zisserman A. Frozen in time: A joint video and image encoder for end-to-end retrieval. In: Proc. of the 2021 IEEE Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 1728–1738. [doi: 10.1109/ICCV48922.2021.00175]

[43] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of the 2019 Annual Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019. 4171–4186. [doi: 10.18653/v1/N19-1423]

[44] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. Language models are few-shot learners. In: Proc. of the 34th Int'l

Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 1877–1901.

[45] Rennie SJ, Marcheret E, Mroueh Y, Ross J, Goel V. Self-critical sequence training for image captioning. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 1179–1195. [doi: 10.1109/CVPR.2017.131]

[46] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft COCO: Common objects in context. In: Proc. of the 13th European Conf. on Computer Vision. Zurich: Springer, 2014. 740–755. [doi: 10.1007/978-3-319-10602-1_48]

[47] Young P, Lai A, Hodosh M, Hockenmaier J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Trans. of the Association for Computational Linguistics, 2014, 2: 67–78. [doi: 10.1162/tacl_a_00166]

[48] Karpathy A, Li FF. Deep visual-semantic alignments for generating image descriptions. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015. 3128–3137. [doi: 10.1109/CVPR.2015.7298932]

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

[50] Denkowski M, Lavie A. Meteor universal: Language specific translation evaluation for any target language. In: Proc. of the 9th Workshop on Statistical Machine Translation. Baltimore: Association for Computational Linguistics, 2014. 376–380. [doi: 10.3115/v1/W14-3348]

[51] Lin CY. ROUGE: A package for automatic evaluation of summaries. In: Proc. of the 2004 Text Summarization Branches Out. Barcelona: Association for Computational Linguistics, 2004. 74–81.

[52] Anderson P, Fernando B, Johnson M, Gould S. SPICE: Semantic propositional image caption evaluation. In: Proc. of the 14th European Conf. on Computer Vision. Amsterdam: Springer, 2016. 382–398. [doi: 10.1007/978-3-319-46454-1_24]

[53] Wang PD, Ng HT. A beam-search decoder for normalization of social media text with application to machine translation. In: Proc. of the 2013 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta: Association for Computational Linguistics, 2013. 471–481.

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

[55] Liu MF, Shi Q, Nie LQ. Image captioning based on visual relevance and context dual attention. Ruan Jian Xue Bao/Journal of Software, 2022, 33(9): 3210–3222 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6623.htm [doi: 10.13328/j.cnki.jos.006623]

[56] Yao T, Pan YW, Li YH, Mei T. Exploring visual relationship for image captioning. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 711–727. [doi: 0.1007/978-3-030-01264-9_42]

[57] Huang L, Wang WM, Chen J, Wei XY. Attention on attention for image captioning. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 4633–4642. [doi: 10.1109/ICCV.2019.00473]

[58] Ji JY, Luo YP, Sun XS, Chen FH, Luo G, Wu YJ, Gao Y, Ji RR. Improving image captioning by leveraging intra- and inter-layer global representation in transformer network. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI Press, 2021. 1655–1663. [doi: 10.1609/aaai.v35i2.16258]

[59] Li XJ, Yin X, Li CY, Zhang PC, Hu, XW, Zhang L, Wang LJ, Hu HD, Dong L, Wei FR, Choi YJ, Gao JF. Oscar: Object-semantics aligned pre-training for vision-language tasks. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 121–137. [doi: 10.1007/978-3-030-58577-8_8]

[60] Wang ZR, Yu JH, Yu AW, Dai ZH, Tsvetkov YL, Cao Y. SimVLM: Simple visual language model pretraining with weak supervision. arXiv:2108.10904, 2021.

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

[62] Lu JS, Yang JW, Batra D, Parikh D. Neural baby talk. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7219–7228. [doi: 10.1109/CVPR.2018.00754]

[63] Wang JB, Wang W, Wang L, Wang ZY, Feng DD, Tan TN. Learning visual relationship and context-aware attention for image captioning. Pattern Recognition, 2020, 98: 107075. [doi: 10.1016/j.patcog.2019.107075]

[64] Li ZX, Wei HY, Huang FC, Zhang CL, Ma HF, Shi ZZ. Combine visual features and scene semantics for image captioning. Chinese Journal of Computers, 2020, 43(9): 1624–1640 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2020.01624]

附中文参考文献:

[19] 薛子育, 郭沛宇, 祝晓斌, 张乃光. 一种基于生成式对抗网络的图像描述方法. 软件学报, 2018, 29: 30–43. http://www.jos.org.cn/jos/article/abstract/18015?st=search

[55] 刘茂福, 施琦, 聂礼强. 基于视觉关联与上下文双注意力的图像描述生成方法. 软件学报, 2022, 33(9): 3210–3222. http://www.jos.org.cn/1000-9825/6623.htm [doi: 10.13328/j.cnki.jos.006623]

[64] 李志欣, 魏海洋, 黄飞成, 张灿龙, 马慧芳, 史忠植. 结合视觉特征和场景语义的图像描述生成. 计算机学报, 2020, 43(9): 1624–1640. [doi: 10.11897/SP.J.1016.2020.01624]

宋井宽(1986－), 男, 博士, CCF 专业会员, 主要研究领域为深度学习, 信息检索.

朱晋宽(1998－), 男, 硕士生, 主要研究领域为信息检索, 视频理解.

曾鹏鹏(1994－), 男, 博士生, 主要研究领域为多模态融合与分析, 计算机视觉.

高联丽(1987－), 女, 博士, 主要研究领域为计算机视觉, 深度学习, 语义网, 知识推理.

顾嘉扬(1999－), 男, 硕士生, 主要研究领域为深度学习, 多模态融合与分析.