计算机科学
COMPUTER SCIENCE

基于深度学习的图像描述优化策略

周子懿, 熊海灵

引用本文

周子懿, 熊海灵. 基于深度学习的图像描述优化策略[J]. 计算机科学, 2023, 50(8): 99-110.

ZHOU Ziyi, XIONG Hailing. Image Captioning Optimization Strategy Based on Deep Learning[J]. Computer Science, 2023, 50(8): 99-110.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230200091

基于深度学习的图像描述优化策略

周子懿[1]　熊海灵[2]
1 西南大学计算机与信息科学学院　重庆 400715
2 西南大学电子信息工程学院　重庆 400715
(zzy2671@163.com)

摘　要　图像描述旨在用语法正确的自然语句描述图像内容,自动地生成文本。图像描述涉及计算机视觉与自然语言处理,是多模态领域的经典任务。近年来,大量的研究开始关注图像描述这类联合了视觉和语言的多模态任务,并取得了许多突破性成果。目前已有的关于图像描述的综述大多以技术为核心,从分类的角度来进行分析。考虑到基于深度学习的图像描述已成为当前的主流研究方法,而且其实质就是一种图像到序列的问题,因此,文中以视觉输入子任务和语言输出子任务为主题,以优化策略为核心,对比分析这两项子任务的优化逻辑与技术发展趋势;同时就图像描述的现有挑战与任务变体等关键共性问题进行讨论,最后期望进一步厘清基于深度学习图像描述的优化策略与发展方向。

关键词: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

Image Captioning Optimization Strategy Based on Deep Learning

ZHOU Ziyi[1] and XIONG Hailing[2]
1 College of Computer and Information Science,Southwest University,Chongqing 400715,China
2 College of Electronic and Information Engineering,Southwest University,Chongqing 400715,China

Abstract　Image captioning aims to describe image content with grammatically correct sentences and automatically generate text. Image captioning involves computer vision and natural language processing,which is a classic task in multimodal field. In recent years,a large number of studies have begun to focus on image captioning,a multimodal task that combines vision and language,and has achieved many breakthrough results. Most of the existing surveys on image captioning take technology as the core and analyze from the perspective of classification. Considering that image captioning based on deep learning has become the mainstream research method at present,and its essence is an image-to-sequence problem,this paper takes visual input subtasks and language output subtasks as the theme,takes optimization strategy as the core. The optimization logic and technical development trend of these two subtasks are compared and analyzed. The existing challenges and task variations of image captioning are discussed. Finally,the optimization strategy and development direction of image captioning based on deep learning are expected to be further clarified.

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

1　引言

图像描述(ImageCaptioning,IC)需要生成与图像内容相关的自然语言描述,如图1所示,图像附有两句图像描述。早期的图像描述任务依赖固定的句子模板[1]或图像检索池[2],缺乏灵活性与泛用性。Kiros等[3]提出的多模态神经语言模型首次将神经网络应用于图像描述领域,开启了基于深度学习的图像描述时代。基于深度学习的图像描述已成为当前图像描述的主流研究方法。基于深度学习的图像描述模型可分为两个子任务:1)视觉输入子任务,也可称为图像处理任务,旨在优化模型获取图像信息的能力;2)语言输出子任务,也可称为序列生成任务,即根据视觉输入子任务提取出的图像内容生成对应描述语句。

目前已有的图像描述综述详细总结了图像描述的早期研究方法[4-7]、常见模型结构[4-10]、训练策略[6-7,10]、数据集与评估方法[5-8,10]等,然而针对图像描述优化策略的综述较少。

到稿日期:2023-02-14　返修日期:2023-05-25
基金项目:国家自然科学基金(41271292);重庆市科技局项目(cstc2019jscxgksbX0103,cstc2020ngzx0010);中央高校基本科研业务费专项(SWU2009107)
This work was supported by the National Natural Science Foundation of China(41271292),Project of Chongqing Science and Technology Bureau(cstc2019jscxgksbX0103, cstc2020ngzx0010) and Fundamental Research Funds for the Central Universities of China(SWU2009107).
通信作者:熊海灵(xionghl@swu.edu.cn)

因此本文详细分析了自 2015 年以来国内外计算机领域主流期刊及国际会议论文中基于深度学习图像描述的优化策略与发展方向，以期为后续研究提供参考。

A group is sistting around a snowy crevasse.
Five people are sitting together in the snow.

图 1 图像描述示例

Fig. 1 Example of image captioning

2 视觉输入子任务的优化策略

视觉输入子任务的主要目标是获取图像蕴含的信息，分析并整合图像内容，最后用于语言输出子任务。该任务能得到的图像信息越优质，就越能有效提升模型的整体性能。本节根据视觉输入子任务包含的 3 个目标——图像特征获取、图像特征分析与图像特征整合学习的优化方向进行分类，得到以下 3 种优化策略：1)从卷积全局特征到区域特征；2)从注意力机制到自注意力机制；3)从卷积表征学习到图表征学习。

2.1 从卷积全局特征到区域特征

综合分析相关文献发现：优化视觉编码器以获取更优质的图像特征为模型优化策略，在图像描述领域中具体表现为从图像级特征到区域级特征的转变。其中，区域级特征的优化方向又包括从图像网格区域特征到图像显著区域特征的转变。

2.1.1 卷积全局特征

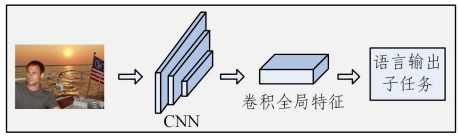
卷积神经网络(Convolutional Neural Networks，CNN)在图像分类、目标检测和特征提取等任务上表现优异，因此在基于深度学习的图像描述任务研究前期，大部分模型都采用卷积全局特征。图 2(a)展示了常见的基于卷积全局特征的视觉输入子任务流程。

Kiros 等[3]将 CNN 应用到图像描述任务中，提出多模态神经语言模型，在对数双线性模型的基础上，利用神经网络学习高层图像特征。随后，得益于在机器翻译领域表现优异的编码器-解码器结构，Vinyals 等[11]在 2015 年首次将编码器-解码器结构应用于图像描述任务中，构建了神经图像描述模型(Neural Image Captioning，NIC)。NIC 模型对图像描述任务的后续发展有着突破性的贡献，它将图像描述任务看作序列训练任务，将图像作为输入序列，将描述语句作为输出序列，并使用 GoogLeNet[12]作为图像特征提取器。随后，Mao 等[13]提出了更为灵活的多模态循环神经网络模型，该模型中深层卷积神经网络和深层循环神经网络互相作用于多模态层中，整合图像特征与词特征，使图像与描述之间更容易互相调度。综上所述，随着 CNN 技术的发展，从 AlexNet[14]到 VGG[15]与 GoogLeNet，再到 ResNet[16]等，视觉输入子任务的性能不断提升，研究者更倾向于根据模型的不同需求，使用对应的更高级的 CNN[17]。

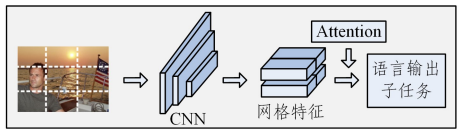
包括上述模型在内，早期大部分模型将卷积全局特征直接用于语言输出子任务。Wu 等[18]对此提出新观点，认为拥有高层语义信息的特征能更好地作用于语言输出子任务。Wu 等提出的模型包含一个预训练单标签 VGG 模型，通过多标签数据集 MSCOCO[19]的微调，模型可输出多标签分类结果。将该方法推广到多个图像区域中获得最终的属性预测向量，这样视觉输入子任务的输出将包含高层次语义属性。在后续研究中，许多模型也在 CNN 部分添加了用于判断高层次语义的模块[20-22]。

2.1.2 网格区域特征

基于卷积全局特征的图像描述模型结构较为简洁，且能较为有效地提取整体上下文信息。然而，卷积全局特征对图像中的所有区域一视同仁，可能会忽略图像中的部分细节，且获取到的特征不包含空间信息，因此此类模型很难再产生更精准具体的描述，对于图像中重点区域的把控仍有待改进。针对此问题，研究者开始关注图像网格区域特征。Xu 等[23]提出的“Show，Attend and Tell”模型在 VGG 网络中引入加性注意力，以计算图像网格特征的注意力权重。自此，许多研究引入加性注意力联合图像网格特征进行图像描述生成[24-26]。图 2(b)给出了常见的基于网格区域特征的视觉输入子任务流程。



(a)卷积全局特征



(b)网格区域特征

(c)显著区域特征

图 2 从卷积全局特征到区域特征优化策略的 3 种常见流程

Fig. 2 Three common processes used for optimization strategy from convolution global feature to regional feature

以“Show，Attend and Tell”模型为代表的基于网格特征的模型，在一定程度上解决了基于卷积全局特征的模型缺乏细粒度描述能力的问题。然而基于网格特征的图像描述模型仍存在一定缺陷，如模型注意力仅作用于每个固定大小的网格区域，此类网格区域的设定与图像的实质内容并无关联，因此此类特征在一定程度上不利于对图像内容的深入检测与分析。有相关研究指出，注意力机制作用于目标物体和其他图像显著区域时更自然且有效[27-28]。鉴于此，一部分研究者从使用网格区域特征转变为使用显著区域特征，具体表现为向

模型中加入目标检测等技术。

2.1.3 显著区域特征

目标检测是一种图像分类技术，它能判断图像中对象类别以及对象位置关系等。作为目标检测的经典算法之一，R-CNN[29]有别于卷积全局特征和网格区域特征的一视同仁，它能更细致地摘取目标图像的重点内容。Karpathy 等[30]关注区域和短语的语义对齐，提出了一种结合 R-CNN 与双向循环神经网络的模型，实现图像特征和文本特征的联合。自此，一部分研究者开始向图像描述模型中加入目标检测模块，优化模型对实体的判别与分析能力。图 2(c)给出了常见的基于目标检测的视觉输入子任务流程。

Faster R-CNN 算法[31]极大地优化了当时目标检测的性能。经典的基于 FasterR-CNN 的图像描述模型之一是由 Anderson 等[26]提出的结合了自下而上和自上而下注意力的 Up-Down 模型。该模型视觉输入部分拥有一个自下而上的注意力模块，利用基于 ResNet-101[16]的 Faster R-CNN 获取区域特征。有别于卷积全局特征，该模型中的 ROIPooling 层获取到的是蕴含更完整信息的特征向量。Up-Down 模型还利用属性类获取图像中主要目标的属性-目标二元描述，主要实现手段是使模型预先学习 Visual Genome 数据集[32]中的属性类和目标类知识。

在此基础上，基于 FasterR-CNN 的图像描述模型不断涌现[33-38]，此类模型的主要优化方向是将显著区域特征作用于不同的模型中以发挥其优势。Yao 等[34]使用两个连续的 FasterR-CNN 分别提取图像区域特征和语义分割实体特征，为后续构建层次结构剖析模型(HIerarchy Parsing，HIP)奠定基础。Datta 等[35]将重点放在感兴趣区域与短语的对齐问题上。Lu 等[36]的模型使用属性分类的方法将区域的细节属性填入模型实时生成的句子模板中。随着目标检测技术的不断进步，图像描述的细节越来越具体，模型准确率随之上升。除了改变显著区域特征的使用方法，Chen 等[17]也对注意力机制进行了改进。Chen 等认为当前大多数注意力模型仅关注空间注意力，由此提出空间和通道注意力模型，通过向注意力模块中加入权重计算公式，使其拥有关注图像空间注意力和通道注意力的能力。值得注意的是，除了上述 HIP 模型使用了语义分割技术外，Li 等[39]也提出了一种基于语义分割的模型。语义分割技术能获取到像素级特征，有助于对图像中实例的判断，能辅助模型生成更丰富的描述。但获取过于精细的分类所带来的准确性收益与多样性收益能否与它所需要付出的代价平衡是一个值得讨论的问题。Li 等使用 UNet 提取语义特征的做法也在一定程度上平衡了这一问题。

在上述 3 种使用不同视觉特征的优化策略中各选取一个代表性模型，并列举相关模型的各项标准评估指标，如表 1 所列，所有结果均来自 MSCOCO 数据集。表 1 中第 3－8 列表示模型的评估指标，是用于评价模型生成图像描述质量的方式之一。其中 B1 和 B4 分别代表 BLEU-1 分数和 BLEU-4 分数[40]，M 代表 METEOR 分数[41-42]，R 代表 ROUGE 分数[43]，C 代表 CIDEr 分数[44]，S 代表 SPICE 分数[45]。对比这 3 个代表性模型的指标可以发现，随着视觉特征的改变，模型性能有了一定提升。从仅使用卷积全局特征作为视觉编码，到结合注意力机制使用网格区域特征，直至使用当前图像描述领域主流的显著区域特征作为视觉编码，此类视觉输入子任务优化策略的优化逻辑是图像的处理方式从简单编码到更具指向性检测的演化。

表 1 使用不同视觉特征的代表性图像描述模型的评估指标

Table 1 Evaluation metrics of classical image captioning models with different visual features

模型名称	视觉特征	B1	B4	M	R	C	S
NIC[11]	卷积全局特征	72.4	31.4	25	53.1	97.2	18.1
Show，Attend and Tell[23]	网格区域特征	74.1	33.4	26.2	54.6	104.6	19.3
Up-Down[26]	显著区域特征	79.4	36.7	27.9	57.6	122.7	21.5

2.2 从注意力机制到自注意力机制

以“Show，Attend and Tell”模型为开端，注意力机制一直作为“向导”穿插在视觉输入子任务与语言输出子任务中，为模型提供信息对齐和增强信息间的关联性等服务，并致力于联合视觉和语言，捕捉两者间的关系。为获取更丰富的上下文信息，自注意力机制与 Transformer 结构[46]被提出，随后在图像描述领域初露锋芒。本节以自注意力机制为核心，根据深化特征关联和简化模型结构这两个思路的区别，将从注意力机制到自注意力机制的优化策略分为以下两种优化方向。

2.2.1 基于 Transformer 编码器的优化方向

基于自注意力机制模型主要针对模型中查询、键和值进行优化，以自注意力机制为核心，向模型中添加 Transformer 编码器结构。

Li 等[47]的模型利用两个独立的 Transformer 编码器，分别处理图像区域视觉信息和参考文本语义属性词信息，解决了部分语义鸿沟问题。Cornia 等[48]将重点放在优化模型记忆能力上，提出了一种带记忆的网状 Transformer 模型。该模型将多层编码器与多层解码器以网格结构相互连接。与常见的单特征、单模态模型相比，带记忆的网状 Transformer 采用全连接结构，可以更有效地利用各层次的视觉关系，该模型对自注意力机制的改进扩展了键和值的集合，可以更有效地利用模型先验知识以及图像低级特征和高级特征关系。Ji 等[49]的全局增强 Transformer 模型由全局自适应控制器将图像全局信息融入解码器中，增强了图像全局特征的影响力。此外还有许多利用自注意力及自注意力变体结构的模型[50-54]。

总结相关研究可知，该优化策略的优化逻辑是改善传统注意力机制在特征信息种类上的单一性。不管是深化图像特征和文本特征之间的关联，还是增强图像全局特征的影响力等优化方法，都表明模型仅依靠单一注意力生成描述是不够的。研究者越来越倾向于采用多种注意力来辅助生成图像描述，促进了从注意力机制到自注意力机制优化策略的发展。

2.2.2 基于网格区域特征的优化方向

一部分研究以简化模型结构为主要优化方向。Transformer 结构能直接作用于网格区域，减少了参数量，从而有效提升模型效率。这也使得网格区域特征再次受到研究者的重视。鉴于无检测器图像理解模型取得了较好成果，无目标检测的图像描述模型受到广泛关注[55]。与通过加入目标检测

提高模型性能的研究相反，已有部分基于网格特征和自注意力机制的无检测器图像描述模型被提出[55-56]。

Zhang 等[56]提出的 RSTNet 模型结合了自适应注意力模型以区分视觉词与非视觉词的优势，并使模型能在视觉信息和语言信息中进行合理取舍。同时该模型利用空间位置坐标的方法，巧妙地解决了 Transformer 处理网格特征时会丢失空间信息的问题。Fang 等[55]认为基于目标检测的图像描述模型计算代价高且受制于带信息标注的数据集，因此提出了一种无需目标检测、仅使用网格特征的基于 Transformer 的模型。该模型为了吸收基于目标检测模型的优势，结合了概念表征预测和知识蒸馏技术作为该模型中相较于目标检测更轻量的标签预测模块，有效地提高了模型准确性，减少了模型计算代价。由此可见，基于网格区域特征的优化方向取消目标检测的做法简化了模型结构，有效提升了模型推理速度。

综合分析以上两个优化方向，对比采用加性注意力的模型与上述采用自注意力和 Transformer 的模型，模型评估指标如表 2 所列，所有数据均来自 MSCOCO 数据集。其中，交叉熵优化训练指采用交叉熵策略训练模型，CIDEr 优化训练指采用强化学习策略训练模型。分析数据可知，从注意力到自注意力的优化策略使模型评估指标有较大提升。对比基于 Transformer 编码器的优化方向和基于网格特征的优化方向，两者的评估指标的提升幅度并不大，但后者的优势在于其模型相对轻量。

表 2 使用不同注意力机制的模型评估指标

Table 2 Evaluation metrics of image captioning models using different attention mechanisms

模型名称	注意力机制	交叉熵优化训练					CIDEr 优化训练				
		B4	M	R-L	C	S	B4	M	R-L	C	S
Up-Down[26]	加性注意力	36.2	27.0	56.4	113.5	20.3	36.3	27.7	56.9	120.1	21.4
Adaptive[25]	加性注意力	33.2	25.7	55.0	101.3	—	—	—	—	—	—
ETA[47]	自注意力	37.1	28.2	57.1	117.9	21.4	39.3	28.8	58.9	126.6	22.7
AoANet[51]	自注意力	37.2	28.4	57.5	119.8	21.3	38.9	29.2	58.8	129.8	22.4
$\mathcal{M}^2$-T[48]	自注意力	—	—	—	—	—	39.1	29.2	58.6	131.2	22.6
RSTNet[56]	自注意力	—	—	—	—	—	40.1	29.8	59.5	135.6	23.3
ViTCAP[55]	自注意力	36.3	29.3	58.1	125.2	22.6	41.2	30.1	60.1	138.1	24.1

综上所述，2.1 节中频繁提及的跨模态注意力是用于深化视觉输入和语言输出两者之间高级关系的机制，而此类基于自注意力机制的模态内注意力更利于探寻视觉输入和语言输出各自特征的内部相关性。自注意力机制使每个元素依次与其他元素相连，更关注每个通道和每个空间互相的联系，拥有比 CNN 更大的感受野，适用于建立全局依赖关系。

2.3 从卷积表征学习到图表征学习

上述两种优化策略已取得许多优秀成果，然而大多数模型以卷积表征学习为主，重点关注区域本身的内容，以各图像区域之间的关系作为切入点的研究相对较少。场景图是一种将数据排序到层次结构中的方法，以 n-tree 的形式存在，其父节点能影响多个子节点。越来越多的图像描述模型开始使用这种过渡数据结构，联合图卷积网络(Graph Convolutional Network，GCN)进行表征学习，将其作为视觉和语言信息交汇的桥梁。常见的基于图的优化策略结构如图 3 所示。根据模型引入“图”的目的可将其概括为两类：增加区域关联性和增加描述多样性。

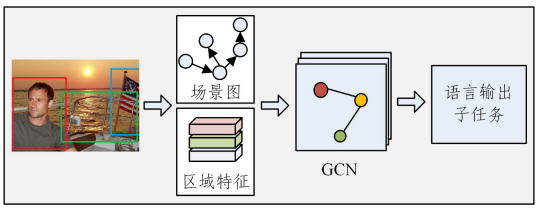


图 3 基于图的视觉输入子任务常见流程

Fig. 3 Common process of visual input subtask based on graph

2.3.1 增加区域关联性

场景图能有效整合不同区域之间的空间语义关系，在场景图的加持下，这类优化策略使图像区域之间的关联性更加清晰且可视。Yao 等[37]根据图像的语义和空间关系提出 GCN-LSTM 模型，模型包含两种场景图。首先根据图像显著区域之间的语义关系构建语义图；其次模型利用目标区域间的交并比、相对距离和角度来构建空间关系图。该模型改进了传统 GCN，使其拥有学习有向边和边标签信息的能力，并利用 GCN 将语义图和空间关系图融入图像特征中，使得图像中各区域之间的语义信息和关系信息能更轻易地被语言输出子任务模块所接收并利用，有效地提高了视觉输入子任务的特征输出质量。Yao 等[34]在 2019 年改进 GCN-LSTM 模型，提出了 HIerarchy Parsing(HIP)模型。Yao 等认为每个检测出的区域总会包含某种从属关系，因此引入 n-tree 结构，令蕴含更丰富信息的细节区域为粗略区域的子节点。HIP 结构将图像信息划分为 3 层树状结构，其中包括了全图和图像细节实例在内的多方面图像特征，优化了普通模型在视觉输入内容方面的单一性，且该模型与其他语言输出模型的适配性很强。Yang 等[57]提出场景图自动编码模块，利用 GCN 编码结合场景图信息训练共享字典。共享字典是一个具有先验知识的模块，它通过文本语料库预训练对上下文词嵌入向量进行转换重建。当该模型进行正式训练时，目标图像的信息被转换为场景图，共享字典辅助模型修正场景图。修正后的场景图拥有更具体且准确的区域关系，最终用于生成描述。

综合分析相关案例可知，大部分模型利用“图”能有效整合信息、挖掘区域关系的优势，使视觉输入子任务的图像特征输出更加丰富。该类优化策略对缩小语义鸿沟有一定辅助作用。

2.3.2 增加描述多样性

大多数图像描述模型通常缺少可控性和可解释性[58]，生成的自然语言描述较为单一。为了让图像描述模型的输出能够像人类一样可控且多样，许多研究通过引入额外控制信号作为模型生成描述的约束，这类研究被称为可控图像描述

(Controllable Image Captioning，CIC)。可控图像描述旨在根据输入图像和指定控制信号生成符合图像内容且符合语法的自然语言描述[59]。

场景图在可控图像描述研究中常被用于增加描述多样性。Chen 等[60]利用抽象场景图将图像中的对象、属性和关系进行结构化整合，作为控制信号输入到模型中，从而生成可根据用户意图控制的多样化描述。该模型利用关系分类器判断区域间是否存在有意义的关系，推理区域在全图中的整体表现、视觉特征以及不同目标间的位置关系。

根据以上案例总结，“图”在图像描述中总是扮演一个辅助模型整合信息、构建框架的角色。基于图编码优化策略的模型能够有效利用区域间的关系，且利于相邻节点的信息交流。尽管这些优势在一部分基于自注意力机制的模型中也有所体现，但不可否认的是，基于图编码优化策略的模型对图像信息的逻辑构建与拓扑关系构建都有促进作用，且对数据规模的依赖相对更小。

3 语言输出子任务的优化策略

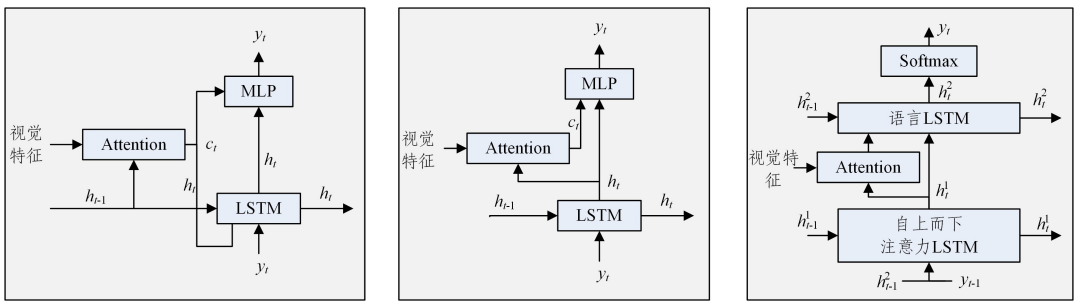
语言输出子任务旨在根据视觉输入子任务获取到的信息，预测单词出现的概率，并生成符合图像内容且语法合理的描述语句。本节先从语言输出子任务中经典的循环模型 LSTM(Long-Short Term Memory)和新兴的非循环模型 Transformer 的优化方向进行分类，得到两种优化策略，分别是从单 LSTM 到多 LSTM 的优化策略和从异质架构到同质架构的优化策略。随后，以实用性为导向，得到从简单语句到新颖语句的优化策略。

3.1 从单 LSTM 到多 LSTM

语言输出子任务可以被看作是序列任务。循环神经网络在处理序列任务方面拥有优异表现，被广泛应用于机器翻译领域。然而循环神经网络难以维持长期依赖关系，且存在梯度消失等问题。长短期记忆网络(LSTM)[61]的提出缓解了这一问题，因此语言输出子任务的其中一种优化策略以 LSTM 为基础，主要表现为从单 LSTM 到多 LSTM。其中，使用多 LSTM 的优化方向又包括从多层 LSTM 到其他 LSTM 变体的转变。

3.1.1 单 LSTM

Vinyals 等[11]在 NIC 模型中首次使用 LSTM 作为语言解码器。此后，单 LSTM 被频繁应用于图像描述模型[17,62]。为了解决 NIC 模型可能忽略某些特定区域细节的问题，Xu 等[23]的“Show，Attend and Tell”模型在 LSTM 中加入加性注意力，使模型在生成单词时能更明确当前时间步应该关注的重点区域，同时使描述的生成更加可视化。此类模型的语言输出子任务常见结构如图 4(a)所示。



(a)单 LSTM 与加性注意力 (b)单 LSTM 与空间注意力 (c)双层 LSTM

图 4 从单 LSTM 到多 LSTM 优化策略使用的 3 种常见结构

Fig. 4 Three common structures used for optimization strategies from single LSTM to multi-LSTM

为了更好地发挥 LSTM 的优势，一部分图像描述模型对单 LSTM 基础架构和注意力机制进行了优化。Lu 等[25]构建了自适应注意力模型，改进了传统软注意力模型，引入空间注意力机制，并添加了视觉哨兵向量。其中，空间注意力机制关注“在哪里查看”，主要思路来自 ResNet 的残差连接，其模型结构如图 4(b)所示；视觉哨兵向量则关注“何时查看”，旨在确定模型是否需要通过关注图像本身来预测下一个单词，主要实现方法是在空间注意力模块中添加存储了长时和短时的视觉-语义信息的视觉哨兵向量。当模型判定不依赖图像本身时，模型会选择依赖视觉哨兵。自适应注意力模型将视觉词与非视觉词区分开，解决了“Show，Attend and Tell”模型[23]中部分注意力关注的位置无意义的问题。该研究思路为其他许多研究提供了基础，基于自适应注意力模型的图像描述研究不断涌现[36,58-59]。

综合分析相关案例可知，单 LSTM 虽在一定程度上缓解了长期依赖关系的匮乏，但仍存在问题。如图 5 所示，单 LSTM 的图像视觉特征仅在模型初始阶段输入，对后续单词的生成几乎无贡献。为了能够更好地结合多方面特征，获取更优质的上下文信息与语义信息，多 LSTM 结构自然地出现在图像描述模型之中。

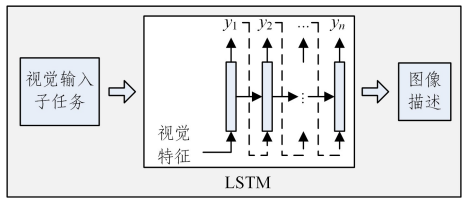


图 5 基于单 LSTM 的语言输出子任务的常见流程

Fig. 5 Common process of language output subtask based on single LSTM

3.1.2 多 LSTM

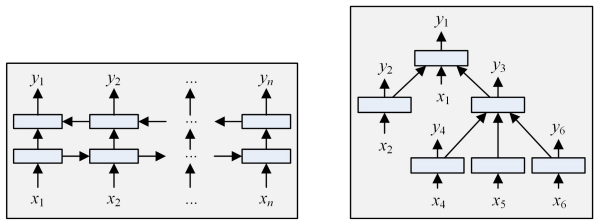
多 LSTM 结构最初以双层 LSTM 的形式出现[26,36,63-64]。Anderson 等在 Up-Down 模型[26]中首次使用双层 LSTM 作为语言解码器，分别是自上而下注意力 LSTM 与语言 LSTM。双层 LSTM 的结构能更好地确定图像区域特征权重，其结构如图 4(c)所示。两层 LSTM 互相利用对方的隐状态，用于特征的对齐与解码。该结构也是后续许多基于双层

LSTM 的 IC 模型的基础。Lu 等[36]在 2018 年提出的模型也利用双层 LSTM 生成视觉词的细粒度描述。该模型是对自适应注意力模型[25]的改进，与早期基于模板的图像描述模型的区别在于其加入了基于目标检测的神经网络模板填充模型。该模型可理解为利用视觉哨兵为视觉词产生一个插槽，最后直接由图像区域特征生成的视觉词填充插槽。Ke 等[63]提出了一种同时利用视觉和文本注意力的反射解码网络。Ke 等的模型主要分为 4 个模块，第一个模块是基于目标检测的图像编码模块，第二至第四个模块用于语言输出，由反射解码网络、反射注意力模块和反射定位模块组成。其中，反射注意力模块包含在反射解码网络中，拥有该模型的第二层 LSTM，主要作用是建立当前与过去单词的注意力权重关系。反射定位模块根据句法结构知识，通过缩小单词实际位置与预测位置之间的差距，以达到准确定位单词、规范句法的目的。该模型有利于长序列建模，使句法更规范、更具体，且将模型决策的过程可视化。

除了主流的双层 LSTM，部分研究使用了 3 层 LSTM 结构。Qin 等[33]提出的回溯预测模型包含正向预测模块和回溯注意力模块。该模型在 Up-down 模型原有的双层 LSTM 末尾添加了第三层语言 LSTM，与第二层语言 LSTM 一起组成正向预测模块。该模型以这两个语言 LSTM 获得的概率之和作为单词预测新依据，有效减少了暴露偏差。Li 等[65]的模型在使用视觉注意获取视觉信息的基础上，增加了场景语义信息作为描述生成的指导。该模型在语言输出子任务中，使用了 3 个 LSTM 互相协同。其中，视觉 LSTM 和场景 LSTM 分别为语言 LSTM 提供指导。该模型与自适应注意力模型[25]有异曲同工之妙，它们都将每个时间步中不同视觉信息的重要程度进行了区分。这在一定程度上使得模型具有像人类一样的思维方式，能够推理出此时此刻图像中更为重要的部分在哪里。该模型比自适应注意力模型更有利的一点在于，语言输出子任务中使用的控制门能过滤部分无用信息与干扰项，且结合了场景语义先验知识的模型拥有更优秀的全局建模能力。

3.1.3 LSTM 变体

除了多 LSTM，LSTM 的其他变体结构也作为一个新的优化方向存在，并受到广泛关注。Wang 等[66]的模型使用双向 LSTM 结构，用于获取额外的上下文信息且使信息来源更加充足。双向 LSTM 结构如图 6(a)所示。Zheng 等[38]提出带目标引导的描述模型，利用双向 LSTM 使模型能灵活地从用户指定单词的两侧的文本开始生成描述。Feng 等[67]提出的无监督模型将语言输出部分分为生成器 LSTM 和鉴别器 LSTM，利用 MSCOCO 数据集和从网络抓取的文本语料库进行无监督学习，采用双向反复重构的方法，使参数在生成器与鉴别器中周转并优化。



(a)双向 LSTM (b)Tree-LSTM

图 6 两种 LSTM 变体的结构

Fig. 6 Structure of two LSTM variants

HIP 模型[34]的语言输出子任务则是利用 Tree-LSTM[68]对 3 个层次的图像特征进行编码，增强了所有实例级、区域级和图像级特征。Tree-LSTM 的结构如图 6(b)所示，它可以促进层次结构内上下文信息的挖掘，从而增强图像特征。传统的 LSTM 仅通过上一个时刻隐状态更新记忆单元，而 Tree-LSTM 的更新依赖于多个子节点隐状态，能更好地判断每个子节点与求解任务的相关程度，使无关的信息趋近于 0，有关的信息趋近于 1。Dai 等[69]更改了 LSTM 隐状态的形式，将其从一维向量扩展成二维向量，并通过卷积来连接。这种二维特征表达方法能保存空间位置信息。Mathews 等[70]的 SentiCap 模型利用两个 LSTM 分别进行粗调与微调，生成积极或消极的风格化图像描述，使该模型的输出语句在对图像进行客观描述的基础上拥有了语言风格。

LSTM 作为语言输出子任务最常用的解码器，其技术发展在一定程度上影响了图像描述的发展。表 3 列出了上述几个基于不同 LSTM 结构的图像描述模型的评估指标，模型分数来源于相关文章，所有数据均来自 MSCOCO 数据集。对比分析表 3 可以发现，随着 LSTM 结构的更替，模型性能不断提升。然而随着更复杂的 LSTM 结构的加入，这类优化策略开始进入瓶颈期，模型性能的提升幅度变得不尽人意。总结相关研究可知，从简单的单 LSTM，到后续的多层 LSTM、双向 LSTM 等众多变体结构，基于 LSTM 的优化策略一直是以构建更具针对性、更具上下文语境利用能力的 LSTM 结构为目标展开的。到目前为止，大多数基于 LSTM 的图像描述模型仍使用以 Up-Down 模型为基础的双层 LSTM 结构以及与其思路接近的 LSTM 变体结构。

表 3 使用不同 LSTM 结构的模型评估指标

Table 3 Evaluation metrics of image captioning models using different LSTM structures

模型名称	注意力机制	交叉熵优化训练					CIDEr 优化训练				
		B4	M	R-L	C	S	B4	M	R-L	C	S
NIC[11]	单 LSTM	29.6	25.2	52.6	94	—	31.9	25.5	54.3	106.3	—
Adaptive[25]	单 LSTM	33.2	26.6	—	108.5	—	—	—	—	—	—
Up-Down[26]	双层 LSTM	36.2	27.0	56.4	113.5	20.3	36.3	27.7	56.9	120.1	21.4
RDN[63]	双层 LSTM	36.8	27.2	56.8	115.3	20.5	—	—	—	—	—
LBPF[33]	三层 LSTM	37.4	28.1	57.5	116.4	21.2	38.3	28.5	58.4	127.6	22.0
HIP[34]	Tree-LSTM	37.0	28.1	57.1	116.6	21.2	38.2	28.4	58.3	127.2	21.9

3.2 从异质架构到同质架构

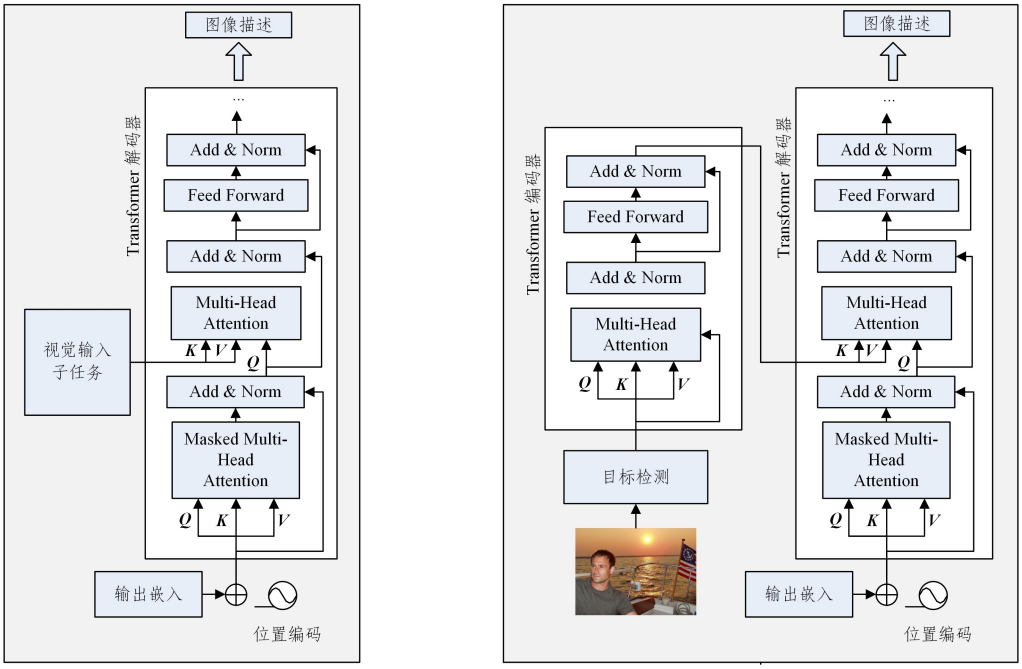
LSTM 是一种自回归模型，尽管多 LSTM 及 LSTM 其他变体结构能在一定程度上弥补单 LSTM 的不足，但仍存在训练速度慢、误差累积、语义贫乏和多样性匮乏等问题。后来，

Transformer 结构的应用大大改善了这些问题。Transformer 结构首先由 Vaswani 等[46]提出，其本质是编码器-解码器结构。它善于获取单词之间的相关性，在序列任务中起到关键作用。基于传统编码器-解码器结构的图像描述模型是分阶段训练的，因此无论解码器如何更新，编码器学习到的内容都不会有所更改。Transformer 作为一种同质架构[71]，能更完整地进行训练，因此编码器可学习到来自解码器的知识。自此，一部分图像描述模型开始从使用异质架构转向使用同质架构。

根据 Transformer 结构在模型中的使用方法进行分类，基于 Transformer 的语言输出子任务的优化方向主要分以下 3 种。

3.2.1 基于 Transformer 解码器的优化方向

基于 Transformer 解码器结构的优化核心思路是利用 Transformer 解码器替换模型原有的循环神经网络结构，具体结构如图 7(a)所示。Zhu 等[72]用 Transformer 解码器替换 LSTM 解码器，有效改善了序列依赖关系的记忆能力，并使模型拥有了并行训练的能力。Li 等[47]的语言输出模型组合了纠缠注意力和门控双边控制器、纠缠注意力融合编码器的视觉和语言特征、门控双边控制器控制视觉特征和语言特征的输入比例。Zhang 等[73]提出了可生成双语图像描述的模型，该模型利用了自注意力机制的长序列建模优势，并重点关注了两种语言描述之间可能存在的关联性，在 Up-Down 模型的基础上利用门控制网络等方法，使模型生成英文描述和日语描述。Yan 等[74]将重点放在优化注意力模块上，用任务自适应注意力模块替换 Transformer 解码器的注意力模块，使模型能动态关注图像特征与语言描述的关联，通过纠正视觉特征向量的注意力权重，来减少注意力模型对非视觉词生成过程的误导。



(a)基于 Transformer 解码器的结构　　(b)基于 Transformer 编码器-解码器的结构

图 7　两种常见的基于 Transformer 的语言输出子任务结构

Fig. 7　Two common language output subtask structures based on Transformer

3.2.2 基于 Transformer 编码器-解码器的优化方向

由于图像描述任务与机器翻译任务具有相似性，因此 Transformer 结构能轻易地融入图像描述任务。针对此特性，研究者使用 Transformer 编码器-解码器结构替换原有的 CNN-RNN 结构。总结相关研究可知，此类模型通常以目标检测技术作为特征提取器，再与 Transformer 编码器-解码器结构相连接，最终生成图像描述[50,75-76]。常见结构如图 7(b)所示。Yu 等[75]首次将 Transformer 整体结构引入图像描述模型，提出多模态 Transformer 模型，使模型能在统一的注意力机制中获取信息。Guo 等[50]改进了自注意力机制，提出规范化自注意力和几何感知自注意力，优化了模型对区域间相对几何关系的推理和利用能力，并使 Transformer 更适用于图像描述任务。

3.2.3 基于 Transformer 预训练策略的优化方向

随着可用于视觉-语言任务的数据集的规模不断扩大，预训练策略的优势也不断扩大，并被广泛应用于图像描述任务等多模态任务中[77]。预训练策略是自监督学习的一种，经过预训练的模型拥有在其他数据集中习得的知识，能有效提高模型在其他针对性训练中的性能与效率，同时语言输出模块也能够获得更丰富的参数知识。

经典预训练模型 BERT[78]是基于 Transformer 的双向编码器，BERT 模型的成功，使基于 BERT 范式的预训练策略获得广泛关注，并被应用于图像描述模型。这类模型通常将视觉与文本特征融合作为输入，利用 BERT 的结构与思想达到优化模型的目的。Li 等[77]提出多模态预训练方法 Oscar，通过随机替换图文三元组的标签，提升模型判别正确标签的能力。Zhou 等[79]提出了一种统一视觉-语言预训练模型，该模型在预训练阶段交替批处理双向自注意力掩码部分和序列到序列自注意力掩码部分，用于学习上下文的视觉语言表示。Zhang 等[80]的 VinVL 模型优化了 Oscar 方法，从

两种角度对图像描述模型进行预训练，一是针对图文三元组中的目标标签生成负样本，二是使用了掩码语言模型作为模型优化方向。Hu 等[81]提出的基于 VinVL 的模型极大地增大了模型规模与数据集规模。此外，Yi 等[82]提出的的分子图像描述模型也使用了大规模预训练模型作为图像特征提取器。

综合分析相关文献发现，Transformer 仍然面临复杂度高且需要更多数据资源与计算资源等问题。同时，当前大多数研究者更关注 Transformer 在 2.2 节中涉及的模态内注意力的优势，以及 3.2.3 节中涉及的预训练策略的优势。在此情况下，Transformer 其他优势的相关研究不如前两者丰富和深入，更不如它在计算机视觉和自然语言处理领域中那么自然与强势。但不可否认的是，Transformer 能有效解决长期依赖问题，且在一定程度上同步编码器与解码器的参数知识的能力不容忽视，是未来 IC 领域的研究重点。

3.3 从简单语句到新颖语句

前面提及的大多数模型生成的语句较为单一，仅围绕图像关键内容进行描述。需要强调的是，图像描述任务不仅需要提高准确性，它也可以从描述多样性、可控性和特定领域适用性等方面入手，使其在实际应用中更有意义。鉴于此，以生成新颖语句为目标的图像描述模型开始发展。本文将新颖语句的生成分为两个方向进行分析，一是可控图像描述，二是风格化图像描述。

3.3.1 可控图像描述

如 2.3.2 节所述，针对视觉输入子任务，研究者们通过添加场景图以达到控制描述生成的目的。而在语言输出子任务中，将图像语义分析与标注技术应用于图像描述能使描述内容更加具有可解释性与可控性。

Cornia 等[58]提出利用名词块序列来控制描述生成的 SCT 模型。该模型基于依存关系树的概念，将名词与对应修饰词组合成名词块，再利用名词块之间不同的排列组合为模型提供描述多样性。Chen 等[59]的特定动词语义角色 CIC 模型则是从更具体的语义分析技术入手，加入特定动词语义角色作为控制信号。特定动词语义角色由一个动词和几个对应的语义角色组成，表示目标活动和该活动涉及的目标及其对应语义角色。该控制信号兼顾事件兼容性和样本适合性，在可控性和多样性方面取得了较好的效果。但该模型依赖动词，且用于对齐的语义角色定位标注模块面临目标检测与语义角色标注的双重误差[83]。

将上述 CIC 模型与基线模型 Up-Down 进行性能对比，如表 4 所列，分数数据来源于 Chen 等[59]的实验测试结果，所有结果均来自 MSCOCO 数据集。令 CIC 模型与 IC 模型各自生成 6 句相关图像描述，分别从描述准确性与描述多样性进行对比。其中，D-1 和 D-2 代表 Div-n 分数[84-85]，分数越高则多样性程度越高；s-C 代表 self-CIDEr 分数[86]。由表 4 可知，CIC 模型在描述多样性上有显著优势，与此同时，标准评估指标也有略微提升。由此可见，CIC 模型的宗旨是在兼顾准确性与语法正确性的基础上，提高描述多样性。此类优化策略主要通过添加额外的 NLP 技术辅助描述生成，从而获得新颖语句。

表 4 CIC 模型与经典 IC 模型的评估指标

Table 4 Evaluation metrics of CIC models and classical IC models

模型名称	模型种类	标准评估指标					多样性评估指标		
		B4	M	R	C	S	D-1	D-2	s-C
Up-Down[26]	IC	20.9	25.4	52.1	209.5	47.9	22.7	35.6	53.9
SCT[58]	CIC	22.0	26.5	55.4	222.5	54.9	27.7	45.7	69.1
VSR[59]	CIC	26.6	30.2	59.8	267.3	56.6	25.1	43.8	67.0

3.3.2 风格化图像描述

人类描述图像时可以从不同的角度出发，得到侧重点不同的多样性描述。为了使图像描述模型的生成语句更加多样化，大多数研究者将多样性图像描述模型设计的重点放在调整描述的角度与描述的复杂程度上。风格化图像描述(Stylized Image Captioning)旨在在准确描述图像内容的前提下，将语言风格融入描述中，如积极、消极、幽默和浪漫等语言风格。在 3.1.3 节中提到的 SentiCap 模型[70]就是典型的风格化图像描述模型。Chen 等[87]在模型中加入风格-事实 LSTM 模块，分别获取事实知识与风格化知识。Gan 等[88]的模型则通过训练因式分解 LSTM 模块，分解传统 LSTM 的参数矩阵，融合特定风格的矩阵。Chen 等[89]的解耦-检索-生成模型使用基于检索的方法，添加了风格化分类器来获取风格词汇，并利用预训练图像描述模型生成的普通描述来检索相似风格词汇，最终生成风格化图像描述。Lin 等[90]采用解纠缠表示学习辅助模型控制风格描述和事实描述。

上述模型依赖于成对的风格文本语料库，数据获取代价较高。因此，擅长无监督学习的生成式对抗网络(Generative Adversarial Networks，GAN)开始受到广泛关注，使用 GAN 来训练风格化描述的模型逐渐增多[91-93]。Guo 等[94]提出的多风格图像描述模型在语言输出部分的设计思路来源于 GAN 的生成器和鉴别器交替训练，最后由反向翻译网络[95]关联图像与描述。

然而基于 GAN 的图像描述始终面临准确性不高的问题。为改善此问题，Deshpande 等[85]提出了一种利用词性标注的模型，通过词性对图像不同区域进行整合，并以词性标注序列作为生成语句的基础。在词性的影响下，描述语句会出现更多拥有形容词词性的单词，增加了描述多样性。Mathews 等[96]提出的模型同样利用词性标注，结合 FrameNet 框架代替动词，该模型同样无需配对的语料库。Cheng 等[97]的模型则利用对比学习来处理成对数据和非成对数据，解决风格文本语料库的问题。

从输出简单语句到输出新颖语句的优化策略，是图像描述任务从以准确性为导向到以实用性为导向的转变。与追求高准确性、输出精确描述语句的优化策略相比，输出新颖语句的优化策略在保持一定准确性的基础上，更

注重模型的实用价值与灵活性。

4 总结与展望

图像描述是视觉-语言理解任务的重要研究领域，对于人工智能理解图像与文本有特殊意义。图像描述融合了计算机视觉和自然语言处理两个领域，是一项有特殊意义且充满挑战的多模态任务。图像描述任务可以参考其他两个领域的思路，但也将同时面临其他两个领域的挑战。总体来说，近年来基于深度学习的图像描述模型的准确性在不断提高，但其灵活性与多样性还有待改善，同时由于其结构的限制，模型准确性和稳定性始终无法到达令人满意的水平，这影响了它在实际中的应用。

总结视觉输入子任务和语言输出子任务的优化策略，可以将未来图像描述领域的优化趋势归纳为以下两种，分别是由简入繁的策略与化繁为简的策略。

由简入繁的策略共分为 3 点：1)从卷积全局特征到区域特征的优化策略，这是检测精细程度的由简入繁。不断发展的检测技术与注意力机制，将视觉输入原有的简单编码转换为更具指向性的精细编码。在此基础上，精细程度更高的语义分割技术也被提及。此类细粒度识别类的优化方向或许是未来 IC 研究的方向，但与此同时，这类由简入繁的策略也带来了问题，即如何平衡越发精细的检测技术与逐渐增加的检测代价之间的矛盾。2)从单 LSTM 到多 LSTM 的优化策略，这是语言输出部分结构的由简入繁。近年来基于 LSTM 的语言输出子任务结构逐渐趋于一致，大部分语言输出仍选择使用双层 LSTM 以及与其思路相近的 LSTM 变体。3)从简单语句到新颖语句的优化策略，这是模型输出内容的由简入繁。这里的“繁”与前两点不同，不仅仅代表繁多与复杂，更代表了繁盛与多样。随着语义分析与标注、GAN 等技术的加入，IC 模型输出语句变得更加丰富且多样，可控且更类似于人类的判断逻辑。这样生成的新颖语句在一定程度上更具实际应用价值，这也是图像描述领域需要继续探索的方向。同样重要的还有在特定领域数据集加持下的新颖描述，如多语言描述生成[73]、针对视力受损人群的应用开发[98]、医学影像描述[99]、新闻描述生成[100-101]等。

化繁为简的策略可分为 4 点：1)从注意力机制到自注意力机制的优化策略，这是模型结构与资源代价的化繁为简。利用网格特征与 Transformer 代替目标检测技术使得模型消耗减少。如何使无检测器模型在准确性上超越有检测器的模型，获得更轻量、更高效的 IC 模型，是未来 IC 模型研究方向。2)从卷积表征学习到图表征学习的优化策略，这是视觉输入内容的化繁为简。“图”的加入，看似增加了模型结构的复杂程度，实际上是“图”辅助整合了图像内容，使模型能够获得更清晰、更简洁明了的图像信息，从而进行更高效的推理。这也在一定程度上减少了模型对数据规模的依赖。3)从异质架构到同质架构的优化策略，这是模型空间结构的化繁为简。此类优化策略将异质架构中不同的网络组件转换为相同的结构，实现编解码器参数同步等效果，简化了视觉信息和语言信息的表示空间。利用同质架构的优势，使 Transformer 更有效地服务于 IC 模型，是 IC 模型未来的一项研究重点。4)基于 Transformer 预训练策略的优化方向，这是模型训练方式的化繁为简。预训练策略减轻了原本繁琐复杂的训练负担，使模型能够额外获得更多的数据集知识，是一个发展趋势良好的研究方向。预训练策略和 BERT-like 多模态模型异军突起，也昭示着模型从使用单视觉特征到使用视觉文本融合特征的改变，它将是一项新的、有效的优化策略。

参 考 文 献

[1] KULKARNI G, PREMRAJ V, DHAR S, et al. BabyTalk: Understanding and Generating Simple Image Descriptions[C]// Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2011: 1601-1608.

[2] FARHADI A, HEJRATI M, SADEGHI M A, et al. Every Picture Tells a Story: Generating Sentences from Images[C]// European Conference on Computer Vision. Berlin: Springer, 2010: 15-29.

[3] KIROS R, SALAKHUTDINOV R, ZEMEL R. Multimodal Neural Language Models[C]// Proceedings of the 31st International Conference on International Conference on Machine Learning. 2014: II-595-II-603.

[4] BAI S, AN S. A Survey on Automatic Image Caption Generation[J]. Neurocomputing, 2018, 311: 291-304.

[5] MIAO Y, ZHAO Z S, YANG Y L, et al. Survey of Image Captioning Methods[J]. Computer Science, 2020, 47(12): 149-160.

[6] LI Z X, WEI H Y, ZHANG C L, et al. Research Progress on Image Captioning[J]. Journal of Computer Research and Development, 2021, 58(9): 1951-1974.

[7] MING Y, HU N N, FAN C X, et al. Visuals to Text: A Comprehensive Review on Automatic Image Captioning[J]. IEEE/CAA Journal of Automatica Sinica, 2022, 9(8): 1339-1365.

[8] HOSSAIN Z M, SOHEL F, SHIRATUDDIN M F, et al. A Comprehensive Surveys of Deep Learning for Image Captioning[J]. ACM Computing Surveys, 2019, 51(6): 1-36.

[9] SHI Y L, YANG W Z, DU H X, et al. Overview of Image Captions Based on Deep Learning[J]. Acta Electronica Sinica, 2021, 49(10): 2048-2060.

[10] STEFANINI M, CORNIA M, BARALDI L, et al. From Show to Tell: A Survey on Image Captioning[C]// IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021.

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

[12] SZEGEDY C, LIU W, JIA Y, et al. Going Deeper with Convolutions[C]// Conference on Computer Vision and Pattern Recognition(CVPR). 2015: 1-9.

[13] MAO J H, XU W, YANG Y, et al. Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN)[C]// International Conference on Learning Representations. 2015.

[14] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet Classification with Deep Convolutional Neural Networks[C]// Annual Conference on Neural Information Processing Systems. 2012.

[15] SIMONYAN K, ZISSERMAN A. Very Deep Convolutional Net-

works for Large-Scale Image Recognition[C]// International Conference on Learning Representations. 2015.
[16] HE K M, ZHANG X Y, REN S Q, et al. Deep Residual Learning for Image Recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2016.
[17] CHEN L, ZHANG H W, XIAO J, et al. SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning[C]// IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2017: 6298-6306.
[18] WU Q, SHEN C, LIU L, et al. What Value Do Explicit High Level Concepts Have in Vision to Language Problems[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2016: 203-212.
[19] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common Objects in Context[C]// Computer Vision — ECCV 2014. 2014: 740-755.
[20] YOU Q Z, JIN H L, WANG Z W, et al. Image Captioning with Semantic Attention[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2016: 4651-4659.
[21] GAN Z, GAN C, HE X D, et al. Semantic Compositional Networks for Visual Captioning[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1141-1150.
[22] CHEN F H, JI R R, SU J S, et al. StructCap: Structured Semantic Embedding for Image Captioning[C]// Proceedings of the 25th ACM International Conference on Multimedia. 2017: 46-54.
[23] XU K, BA J, KIROS R, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention[C]// Proceedings of the 32nd International Conference on International Conference on Machine Learning. 2015: 2048-2057.
[24] YAO T, PAN Y W, LI Y H, et al. Boosting Image Captioning with Attributes[C]// IEEE International Conference on Computer Vision(ICCV). 2017: 4904-4912.
[25] LU J S, XIONG C M, PARIKH D, et al. Knowing When to LookAdaptive Attention via A Visual Sentinel for Image Captioning[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2017: 3242-3250.
[26] ANDERSON P, HE X D, BUEHLER C, et al. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018.
[27] EGLY R, DRIVER J, RAFAL R. Shifting Visual Attention Between Objects and Locations: Evidence From Normal and Parietal Lesion Subjects[J]. Journal of Experimental Psychology: General, 1994, 123(2): 161.
[28] SCHOLL B J. Objects and attention: the state of the art[J]. Cognition, 2001, 80(1/2): 1-46.
[29] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich Feature Hierarchies for Accurate Object Detection andSemantic Segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2014: 580-587.
[30] KARPATHY A, LI F F. Deep Visual-Semantic Alignments for Generating Image Descriptions[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2015: 3128-3137.
[31] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[C]// Conference and Workshop on Neural Information Processing Systems. 2016: 1-10.
[32] KRISHNA R, ZHU YK, 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.
[33] QIN Y, DU J J, ZHANG Y H, et al. Look Back and Predict Forward in Image Captioning[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2019: 8359-8367.
[34] YAO T, PAN Y W, LI Y H, et al. Hierarchy Parsing for Image Captioning[C]// IEEE/CVF International Conference on Computer Vision(ICCV). 2019: 2621-2629.
[35] DATTA S, SIKKA K, ROY A, et al. Align2Ground: Weakly Supervised Phrase Grounding Guided by Image-Caption Alignment [C]// IEEE/CVF International Conference on Computer Vision (ICCV). 2019: 2601-2610.
[36] LU J S, YANG J W, BATRA D, et al. Neural Baby Talk[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018.
[37] YAO T, PAN Y W, LI Y H, et al. Exploring Visual Relationship for Image Captioning[C]// European Conference on Computer Vision. 2018.
[38] ZHENG Y, LI Y L, WANG S J. Intention Oriented Image Captions with Guiding Objects[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2019: 8387-8396.
[39] LI Y S, YAN B Y, ZHOU J L. Fully Convolutional Image Description Model Based on Semantic Segmentation[J]. Computer Engineering and Design, 2023, 44(1): 210-217.
[40] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: A Method for Automatic Evaluation of Machine Translation[C]// Proceedings of the 40thAnnual Meeting of the Association for Computational Linguistics. 2002: 311-318.
[41] BANERJEE S, LAVIE A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments[C]// Proceedings of the Association for Computational Linguistics. 2005: 65-72.
[42] LAVIE A, AGARWAL A. Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments[C]// Proceedings of the Second Workshop on Statistical Machine Translation. 2007: 228-231.
[43] LIN C Y, OCHF J. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-bigram Statistics[C]// Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. 2004: 21-26.
[44] VEDANTAM R, ZITNICK C, PARIKH D. CIDEr: Consensus-based Image Description Evaluation[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2015: 4566-4575.
[45] ANDERSON P, FERNANDO B, JOHNSON M, et al. SPICE: Semantic Propositional Image Caption Evaluation[C]// Computer Vision — ECCV 2016. Springer International Publishing,

2016:382-398.
[46] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is All You Need[C]// Proceedings of the 31st International Conference on Neural InformationProcessing Systems. 2017:6000-6010.
[47] LI G,ZHU L C,LIU P,et al. Entangled Transformer for Image Captioning[C]// IEEE/CVF International Conference on Computer Vision(ICCV). 2019.
[48] CORNIA M,STEFANINI M,BARALDI L,et al. Meshed-Memory Transformer for Image Captioning[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020:10575-10584.
[49] JI J Y,LUO Y P,SUN X S,et al. Improving Image Captioning by Leveraging Intra-and Inter-layer Global Representation in Transformer Network[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021:1655-1663.
[50] GUO L T,LIU J,ZHU X X,et al. Normalized and Geometry-Aware Self-Attention Network for Image Captioning[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020.
[51] HUANG L,WANG W M,CHEN J,et al. Attention on Attention for Image Captioning[C]// IEEE International Conference on Computer Vision. 2019.
[52] ZHUO Y Q,WEI J H,LI Z X. Research on Image Captioning Based on Double Attention Model[J]. ACTA ELECTRONICA SINICA,2022,50(5):1123-1130.
[53] FANG Z J,ZHANG J,LI D D. Spatial Encoding and Multi-layer Joint Encoding Enhanced Transformer for Image Captioning[J]. Computer Science,2022,49(10):151-158.
[54] WANG M Z,JI J Z,JIA A Z,et al. Cross-scale Feature Fusion Self-attention for Image Captioning[J]. Computer Science,2022,49(10):191-197.
[55] FANG Z Y,WANG J F,HU X W,et al. Injecting Semantic Concepts into End-to-End Image Captioning[C]// Conference on Computer Vision and Pattern Recognition(CVPR). 2022.
[56] ZHANG X Y,SUN X S,LUO Y P,et al. RSTNet:Captioning with Adaptive Attention on Visual and Non-Visual Words[C]// Conference on Computer Vision and Pattern Recognition (CVPR). 2021:15460-15469.
[57] YANG X,TANG K H,ZHANG H W,et al. Auto-Encoding Scene Graphsfor Image Captioning[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2019:10677-10686.
[58] CORNIA M,BARALDI L,CUCCHIARA R. Show Control and Tell A Framework for Generating Controllable and Grounded Captions[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2019.
[59] CHEN L,JIANG Z H,XIAO J,et al. Human-like Controllable Image Captioning with Verb-specific Semantic Roles[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2021:16841-16851.
[60] CHEN S Z,JIN Q,WANG P,et al. Say As You Wish Fine-grained Control of Image Caption Generation with Scene Graphs [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2020:9959-9968.
[61] HOCHREITER S,SCHMIDHUBER J. Long Short-Term Memory[J]. Neural Computation,1997,9(8):1735-1780.
[62] YANG Z L,YUAN Y,WU Y X,et al. Review Networks for Caption Generation[C]// Proceedings of the 30th International Conference on Neural Information Processing Systems. 2016:2369-2377.
[63] KE L,PEI W J,LI R Y,et al. Reflective Decoding Network for Image Captioning[C]// IEEE/CVF International Conference on Computer Vision(ICCV). 2019:8887-8896.
[64] SHI Z,ZHOU X,QIU X P,et al. Improving Image Captioning with Better Use of Captions[C]// The Association for Computational Linguistics. 2020.
[65] LI Z X,WEI H Y,HUANG F C,et al. Combine Visual Features and Scene Semantics for Image Captioning[J]. Chinese Journal of Computers,2020,43(9):1624-1640.
[66] WANG C,YANG H J,BARTZ C,et al. Image Captioning with Deep Bidirectional LSTMs[C]// Proceedings of the 24th ACM International Conference on Multimedia. 2016:988-997.
[67] FENG Y,MA L,LIU W,et al. Unsupervised Image Captioning [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). 2019:4120-4129.
[68] TAI K S,SOCHER R,MANNING C. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015:1556-1566.
[69] DAI B,YE D M,LIN D H. Rethinking the Form of Latent States in Image Captioning[C]// Computer Vision — ECCV 2018. Springer International Publishing,2018:294-310.
[70] MATHEWS A,XIE L,HE X. SentiCap:generating image descriptions with sentiments[C]// Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. AAAI Press,2016:3574-3580.
[71] XU Y,LI L,XU H Y,et al. Image Captioning In the Transformer Age[J]. arXiv:2204.07374v1,2022,
[72] ZHU X X,LI L X,LIU J,et al. Captioning Transformer with Stacked Attention Modules[J]. Applied Sciences,2018,8(5):739-749.
[73] ZHANG K,LI J H,ZHOU G D. Study on Joint Generation of Bilingual Image Captions[J]. Computer Science,2020,47(12):183-189.
[74] YAN C G,HAO Y M,LI L,et al. Task-Adaptive Attention for Image Captioning[J]. IEEE Transactions on Circuits and Systems for Video Technology,2022,32(1):43-51.
[75] YU J,LI J,YU Z,et al. Multimodal Transformer With Multi-View Visual Representation for Image Captioning[J]. IEEE Transactions on Circuits and Systems for Video Technology,2019,30(12):4467-4480.
[76] HERDADE S,KAPPELER A,BOAKYE K,et al. Image Captioning:Transforming Objects into Words[C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019:11137-11147.
[77] LI X J,YIN X,LI C Y,et al. Oscar:Object-Semantics Aligned Pre-training for Vision-Language Tasks[C]// Computer Vision —

ECCV 2020. Springer International Publishing, 2020: 121-137.
[78] DEVLIN J, CHANG M W, LEE K, etal. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [C]// Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2018.
[79] ZHOU L W, PALANGI H, ZHANG L, et al. Unified Vision-Language Pre-Training for Image Captioning and VQA[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 13041-13049.
[80] ZHANG P C, LI X J, HU X W, et al. VinVL: Revisiting Visual Representations in Vision-Language Models[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 5575-5584.
[81] HU X W, GAN Z, WANG J F, et al. Scaling Up Vision-Language Pre-training for Image Captioning[C]// 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021.
[82] YI J C, WU C K, ZHANG X C, et al. MICER: A Pre-trained Encoder-Decoder Architecture for Molecular Image Captioning[J]. Bioinformatics, 2022, 38(19): 4562-4572.
[83] WEI M, CHEN L, JI W, et al. Rethinking the Two-Stage Framework for Grounded Situation Recognition[C]// Proceedings of the AAAI Conference on Artificial Intelligence 2022: 2651-2658.
[84] ANEJA J, AGRAWAL H, BATRA D, et al. Sequential Latent Spaces for Modeling the Intention During Diverse Image Captioning[C]// IEEE/CVF International Conference on Computer Vision (ICCV). 2019.
[85] DESHPANDE A, ANEJA J, WANG L, et al. Fast, Diverse and Accurate Image Captioning Guided By Part-of-Speech[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019.
[86] WANG Q Z, CHAN A B. Describing Like Humans: On Diversity in Image Captioning[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019.
[87] CHEN T L, ZHANG Z P, YOU Q Z, et al. "Factual" or "Emotional": Stylized Image Captioning with Adaptive Learning and Attention[C]// Computer Vision— ECCV 2018. Springer International Publishing, 2018: 527-543.
[88] GAN C, GAN Z, HE X, et al. StyleNet: Generating Attractive Visual Captions with Styles[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 955-964.
[89] CHEN Z H, XIONG Y. Stylized Image Captioning Model Based on Disentangle-Retrieve-Generate[J]. Computer Science, 2022, 49(6): 180-186.
[90] LIN Z H, LI G D, ZENG X J, et al. A Stylized Image Caption Approach Based on Cross-Media Disentangled Representation Learning. Computer Science, 2022, 45(12): 2510-2527.
[91] SHETTY R, ROHRBACH M, HENDRICKS L, et al. Speaking the Same Language: Matching Machine to Human Captions by Adversarial Training[C]// IEEE International Conference on Computer Vision (ICCV). 2017.
[92] DAI B, LIN D, URTASUN R, et al. Towards Diverse and Natural Image Descriptions via a Conditional GAN[C]// 2017 IEEE International Conference on Computer Vision (ICCV). 2017.
[93] LI D Q, HE X D, HUANG Q Y, et al. Generating Diverse and Accurate Visual Captions by Comparative Adversarial Learning [C]// Annual Conference and Workshop on Neural Information Processing Systems. 2018.
[94] GUO L T, LIU J, YAO P, et al. MSCap Multi-Style Image Captioning with Unpaired Stylized Text[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 4199-4208.
[95] JOHNSON M, SCHUSTER M, LE Q, et al. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation[J]. Transactions of the Association for Computational Linguistics, 2016, 5(2): 339-351.
[96] MATHEWS A, XIE L X, HE X M. SemStyle: Learning to Generate Stylised Image Captions Using Unaligned Text[C]// IEEE/CVF Conference onComputer Vision and Pattern Recognition. 2018: 8591-8600.
[97] CHENG K Z, MA Z, ZONG S, et al. ADS-Cap: A Framework for Accurate and Diverse Stylized Captioning with Unpaired Stylistic Corpora[C]// Natural Language Processing and Chinese Computing. Springer International Publishing, 2022: 736-748.
[98] GURARI D, ZHAO Y, ZHANG M, et al. Captioning Images Taken by People Who Are Blind[C]// European Conference on Computer Vision. 2020: 417-434.
[99] BEDDIAR DR, OUSSALAH M, SEPPäNEN T. Automatic Captioning for Medical Imaging (MIC): a Rapid Review of Literature [J]. Artificial Intelligence Review, 2022, 56(5): 4019-4076.
[100] BITEN A, GOMEZ L, RUSIÑOL M, et al. Good News, Everyone! Context Driven Entity-aware Captioning for News Images [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 12466-12475.
[101] TAN R, PLUMMER B A, SAENKO K, et al. NewsStories: Illustrating Articles with Visual Summaries[C]// European Conference on Computer Vision. Springer Nature Switzerland, 2022: 644-661.

ZHOU Ziyi, born in 1998, postgraduate. Her main research interests include image captioning and computer vision.

XIONG Hailing, born in 1971, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include database and intelligent information processing, cellular automata theory and application, digital agriculture, computer simulation of agricultural ecology and environmental process.

(责任编辑:何杨)