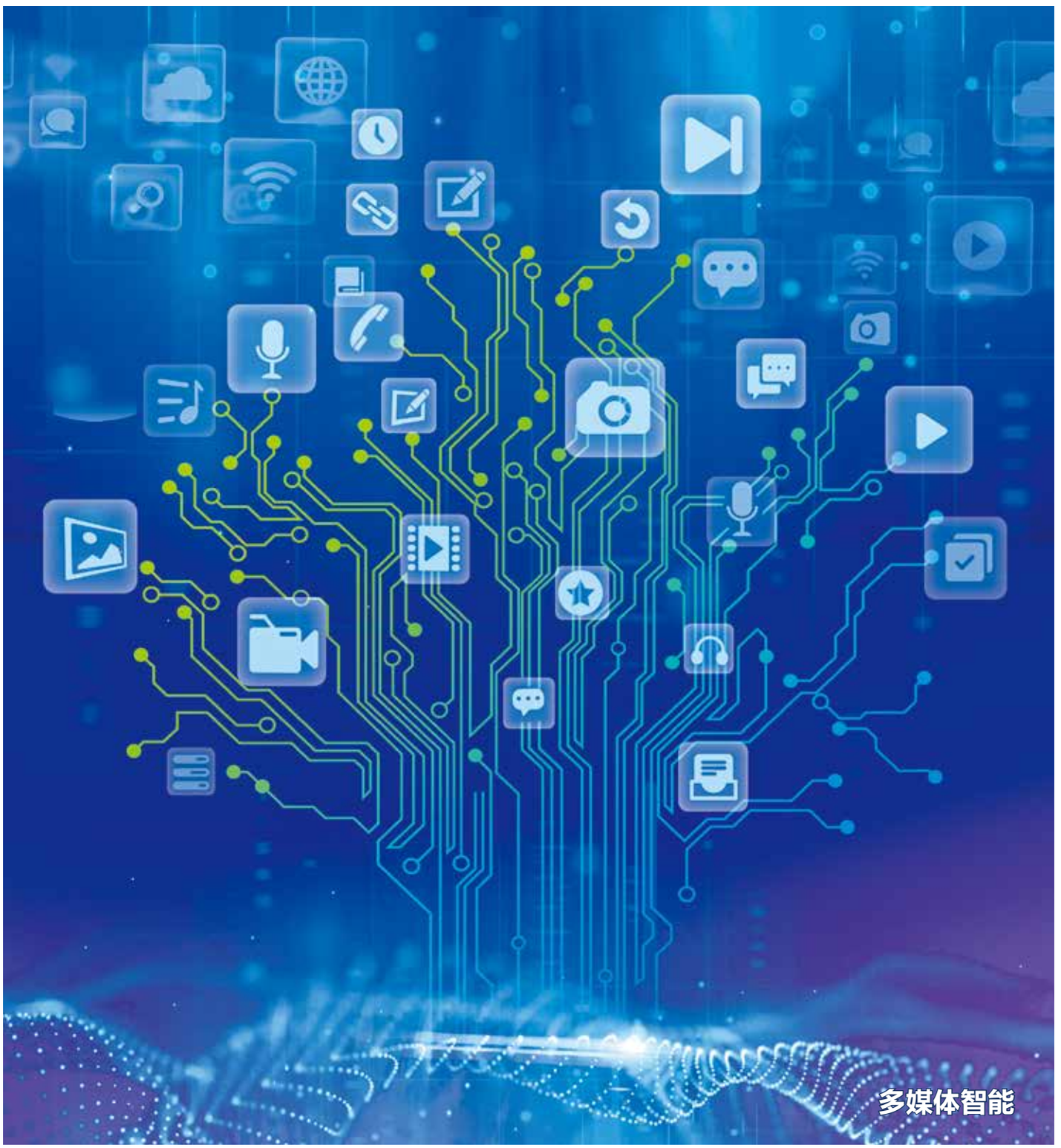
JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2022
09
VOL.27

ISSN1006-8961
CN11-3758/TB



中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第27卷第9期（总第317期）

2022年9月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 吴一戎
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@aircas.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@aircas.ac.cn
Telephone 010-58887035
Website www.cjig.cn

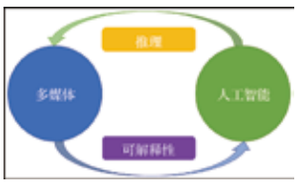
Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

2022年9月 第27卷 第9期（总第317期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

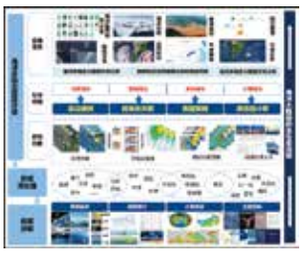
目次



多媒体智能：当多媒体遇到人工智能(第2551页)

学者观点

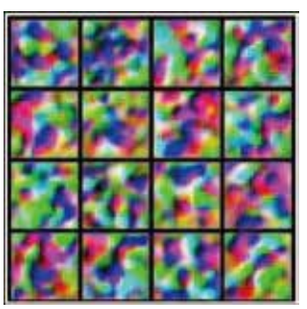
综述



面向海洋的多模态智能计算：挑战、进展和展望(第2589页)

多媒体智能安全

目标智能检测

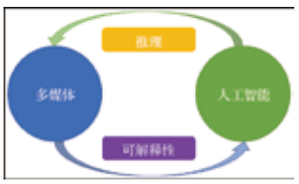


基于真实数据感知的模型功能窃取攻击(第2721页)

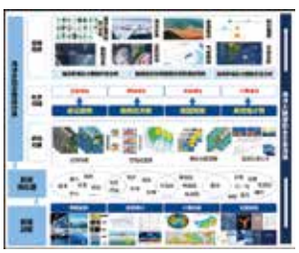
多媒体分析与理解

Volume 27, Number 9
Published September 16, 2022

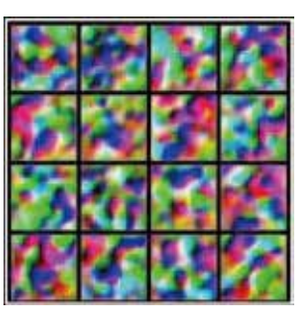
CONTENTS
JOURNAL OF IMAGE AND GRAPHICS



Multimedia intelligence: the convergence of multimedia and artificial intelligence(P2551)



Marine oriented multimodal intelligent computing: challenges, progress and prospects(P2589)



Model functionality stealing attacks based on real data awareness(P2721)

中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2022)09-2775-13

论文引用格式: Jiang W H, Zhan K, Cheng Y B, Xia X and Fang Y M. 2022. The integrated mechanism of hierarchical decoders and dynamic fusion for image captioning. Journal of Image and Graphics,27(09):2775-2787(姜文晖,占锟,程一波,夏雪,方玉明. 2022. 结合多层级解码器和动态融合机制的图像描述. 中国图象图形学报,27(09):2775-2787)[DOI:10.11834/jig.211252]

结合多层级解码器和动态融合机制的图像描述

姜文晖,占锟,程一波,夏雪,方玉明*

江西财经大学信息管理学院,南昌　330032

摘　要: **目的**　注意力机制是图像描述模型的常用方法,特点是自动关注图像的不同区域以动态生成描述图像的文本序列,但普遍存在不聚焦问题,即生成描述单词时,有时关注物体不重要区域,有时关注物体上下文,有时忽略图像重要目标,导致描述文本不够准确。针对上述问题,提出一种结合多层级解码器和动态融合机制的图像描述模型,以提高图像描述的准确性。**方法**　对 Transformer 的结构进行扩展,整体模型由图像特征编码、多层级文本解码和自适应融合等 3 个模块构成。通过设计多层级文本解码结构,不断精化预测的文本信息,为注意力机制的聚焦提供可靠反馈,从而不断修正注意力机制以生成更加准确的图像描述。同时,设计文本融合模块,自适应地融合由粗到精的图像描述,使低层级解码器的输出直接参与文本预测,不仅可以缓解训练过程产生的梯度消失现象,同时保证输出的文本描述细节信息丰富且语法多样。**结果**　在 MS COCO(Microsoft common objects in context)和 Flickr30K 两个数据集上使用不同评估方法对模型进行验证,并与具有代表性的 12 种方法进行对比实验。结果表明,本文模型性能优于其他对比方法。其中,在 MS COCO 数据集中,相比于对比方法中性能最好的模型,BLEU-1(bilingual evaluation understudy)值提高了 0.5,CIDEr(consensus-based image description evaluation)指标提高了 1.0;在 Flickr30K 数据集中,相比于对比方法中性能最好的模型,BLEU-1 值提高了 0.1,CIDEr 指标提高了 0.6;同时,消融实验分别验证了级联结构和自适应模型的有效性。定性分析也表明本文方法能够生成更加准确的图像描述。**结论**　本文方法在多种数据集的多项评价指标上取得最优性能,能够有效提高文本序列生成的准确性,最终形成对图像内容的准确描述。

关键词:图像描述;注意力机制;Transformer;多层级解码;动态融合;门机制

The integrated mechanism of hierarchical decoders and dynamic fusion for image captioning

Jiang Wenhui, Zhan Kun, Cheng Yibo, Xia Xue, Fang Yuming*

School of Information Management, Jiangxi University of Finance and Economics, Nanchang 330032, China

Abstract: **Objective** Image captioning aims at automatically generating lingual descriptions of images. It has a wide variety of applications scenarios like image indexing, medical imaging reports generation and human-machine interaction. To gener-

收稿日期:2022-01-20;**修回日期**:2022-05-05;**预印本日期**:2022-05-12

*通信作者:方玉明　fa0001ng@e.ntu.edu.sg

基金项目:科技创新 2030-"新一代人工智能"重大项目(2020AAA0109301);国家自然科学基金项目(62161013,62162029);江西省重点研发计划项目(20203BBE53033)

Supported by: National Key R&D Program of China (2020AAA0109301); National Natural Science Foundation of China (62161013, 62162029); Key R&D Program of Jiangxi Province, China (20203BBE53033)

ate fluent sentences of the gathered information all, an algorithm of image captioning is called to recognize the scenes, entities and their relationships of the image. A deep encoder-decoder framework has been developed to resolve the issue past decades. The convolutional neural networks based (CNNs-based) encoder extracts feature vectors of the image and the recurrent neural networks based (RNNs-based) decoder generates image descriptions. Recent image captioning is driven by the development of attention mechanism. It improves the performance of image captioners via attending to informative image regions. Most attention models are based on the previously generated words as inputs when the next attending phases are predicted. Due to the lack of relevant textual guidance, most existing works are challenged of "attention defocus", i. e., they fail to concentrate on correct image regions when generating the target words. As a result, contemporary models are prone to "hallucinating" objects, or missing informative visual clues, and make attention model be less interpretable. So, we facilitate an integrated hierarchical architecture and dynamic fusion strategy. **Method** The estimated word provides useful knowledge for predicting more grounded regions, although it is hard to localize the correct regions from the previously generated words at once. To refine the attention mechanism and improve the predicted words, we design a hierarchical architecture based on a series of captioning decoders. Our architecture is a hierarchical variant extended from the conventional encoder-decoder framework. Specifically, the first step is focused on the standard image captioning models, which generates a coarse description as a draft. To ground correct image regions with proper generated words, the latter one takes the outputs from the early decoder. Since the former decoder provides more predictable information to the target word, the attention accuracy is improved in latter decoders. To ground the final predicted words properly in this hierarchical architecture, attended regions from the early decoder can be well validated by the later decoders in a coarse-to-fine manner. Furthermore, we carry out a dynamic fusion strategy to aggregate the coarse-to-fine predictions from different decoders. Noteworthy, our manipulated gating mechanism is focused on the contributions from different decoders to the final word prediction. Differentiated from the previous gating mechanism managing the weight from each pathway, the contributions are jointed with a softmax schema from each decoder, which incorporates contextual information from all decoders to estimate the overall weight distribution. The dynamic fusion strategy provides rich fine-grained image descriptions and alleviates the problem of "vanishing gradients", which makes the learning of the hierarchical architecture easier. **Result** Our method is evaluated on Microsoft common objects in context (MS COCO) and Flickr30K, which are the common benchmark for image captioning. The MS COCO dataset is composed of 120 k images, and the Flickr30K includes 31 k examples. Each image of both datasets is provided with five descriptions. The model is trained and tested using the Karpathy splits. The quantitative evaluation metrics are related to bilingual evaluation understudy (BLEU), metric for evaluation of translation with explicit ordering (MEREOR), and consensus-based image description evaluation (CIDEr). We compare the performance of our model with 12 current methods. On MS COCO, our analysis is optimized by 0.5 and 1.0 of each beyond BLEU-1 and CIDEr. Our result achieves a CIDEr of 69.94 on Flickr30K. Compared to the baseline method (Transformer), our performance is optimized 4.6 of CIDEr on MS COCO and 3.8 on Flickr30K, which verifies that our method improves the accuracy of the predicted sentences effectively. In addition, our qualitative results demonstrate that the proposed method provides rich fine-grained image descriptions in comparison with other methods. Our method describes the number of appeared objects precisely when they belong to the same category. Our method could also describe small objects accurately. To further verify the effectiveness of the proposed hierarchical architecture, we visualize the attention mechanism and it shows that our method attends to discriminative parts of the target objects. In contrast, the baseline method may focus on irrelevant backgrounds, which leads to false predictions straightforward. **Conclusion** Our research is focused on a hierarchical architecture with dynamic fusion strategy for image captioning. The hierarchical architecture consists of a sequence of captioning decoders that refine the attention mechanism. To generate final sentence with rich fine-grained information, the dynamic fusion strategy aggregates different decoders. The ablation study demonstrates the effectiveness of each module in our proposed network. Our optimized results are demonstrated through the comparative experiments on MS COCO and Flickr30K datasets.

Key words: image captioning; attention mechanism; Transformer; hierarchical decoders; dynamic fusion; gating mechanism

0 引 言

图像描述任务(image captioning)旨在对一幅输入图像自动生成完整的自然语言描述。图像描述任务可以应用于人机对话、盲人导航以及医疗影像报告生成等场景,具有巨大的应用前景和研究价值。为生成完整的句子描述,该任务需要全面建模图像中物体的类别、属性以及与场景的交互关系等丰富信息,并将这些内容通过组织语言的方式流畅地进行描述。图像描述任务是计算机视觉和自然语言处理交叉领域的挑战性问题。

早期研究首先分析图像视觉内容,即检测图像中的物体及其属性,分析物体间的相对关系,并将这些内容映射为单词或短语等描述信息(Farhadi 等,2010)。然后通过自然语言技术,例如句子模板或语法规则,将这些基本描述单元转化为完整句子进行图像描述(Kuznetsova 等,2014)。然而模板和语法规则较大地限制了图像描述的多样性和独特性,且对数据集和人工设计的依赖性较强。

得益于深度学习(deep learning)的发展,大量研究工作将深度学习应用于自动图像描述领域(Wan 等,2022)。基于深度学习的主要框架是“编码器—解码器”模型。其中,编码器分析图像的语义内容,形成一组图像特征向量;解码器输入这些特征向量,通过语言生成模型输出完整的图像描述。相比于传统的方式,基于深度学习的模型脱离了具体的本文规则,能够生成变长、多样化的图像描述,并在描述准确性方面具有压倒性优势。因此,基于深度学习的方法是当前自动图像描述领域的主流模型。

注意力机制(attention mechanism)广泛融入编码器—解码器框架(Xu 等,2015),其主要优势在于生成描述语句的每个字符时,可以动态地改变输入特征的权重以指导文本生成,极大提高了图像描述模型的准确性。然而,通过可视化分析和量化分析,发现注意力机制普遍存在不聚焦问题(Liu 等,2017)。具体地,在生成描述单词时,注意力机制有时关注在物体不重要区域,例如人的身体,从而错误预测人的性别(Hendricks 等,2018);有时关注物体背景,导致幻想出与目标相关但未实际出现的物体(Rohrbach 等,2018);有时忽略图像中重要目标,导致描述中缺少重要信息。注意力机制的不聚焦问题严重影响了模型的可解释性。导致该问题的原因为:1)预测 t 时刻的单词时,注意力机制仅依赖 t 时刻之前生成的文本序列作为指导。因此,在未知待预测的目标单词条件下,显著性机制难以准确定位图像的正确区域。2)文本预测过程中,错误预测的单词将进一步误导注意力机制,从而对后续文本的生成产生误差累积。

为解决以上问题,本文提出一种结合多层级解码器和动态融合机制的图像描述模型。该模型是对标准的编码器—解码器结构的扩展,出发点是虽然通过 t 时刻之前预测的单词不足以指导 t 时刻文本生成,但是该预测结果能够提供更加有效的反馈信息,并进一步指导注意力机制定位到准确的图像区域。首先,设计解码器级联的结构实现注意力机制的渐进式精化。其中,第 1 级解码器采用标准的文本预测结构,以前时刻预测的单词为输入,输出粗略的图像描述。其次,后级解码器以前级解码器的预测单词为输入。由于该输入与预测的目标单词更相关,注意力机制能够更有效地聚焦到图像的关键区域,从而生成更准确的文本序列,并缓解误差累积。同时,本文提出一种解码器动态融合策略,根据每级解码器的输出,动态调整其对应权重,自适应地融合由粗到精的文本信息,最终生成细节信息丰富且准确多样的图像描述。动态融合结构使低层级解码器的输出直接参与最终的文本预测,为不同层级的解码器直接引入了监督信息,解决了传统级联结构容易产生的梯度消失现象,使模型训练更加稳定。

为验证模型的有效性,在 MS COCO(Microsoft common objects in context)(Lin 等,2014)和 Flickr30K(Plummer 等,2015)数据集上进行实验。结果表明,本文设计的模型效果显著,在 BLEU(bilingual evaluation understudy)、METEOR(metric for evaluation of translation with explicit ordering)和 CIDEr(consensus-based image description evaluation)等多项评价指标上优于其他对比方法。定性分析结果也验证了本文模型能够生成更加准确的图像描述。

1 相关工作

自动图像描述任务主要以编码器—解码器为主要架构。编码器提取输入图像的语义特征,解码器

对编码器的输出结果进行处理,形成完整的文本描述。鉴于深度神经网络的灵活性和较强的建模能力,当前的主要工作是基于深度神经网络分别对编码器和解码器的结构进行建模(谭云兰 等,2021)。编码器广泛采用卷积神经网络(convolutional neural network,CNN),例如使用 ResNet(residual network)和 VGG(Visual Geometry Group network)等深层网络进行图像的特征表示(汤鹏杰 等,2017)。解码器广泛采用循环神经网络(recurrent neural network,RNN)和长短时记忆网络(long short-term memory,LSTM)对较长的文本序列进行关联建模(罗会兰和岳亮亮,2020)。基于深度神经网络的结构不依赖文本规则,生成的图像描述语法灵活。

1.1 注意力机制

随着注意力机制在机器翻译领域的广泛应用,越来越多的研究将其引入编码器—解码器结构。Xu 等人(2015)将注意力机制引入自动图像描述任务,提出软注意力机制(soft attention),通过隐状态估算图像中不同空间特征的权重,使每一时刻的文本预测都能自适应地关注图像中的不同区域,从而提高下一时刻文本预测的准确性。然而,注意力机制学习的权重在模型中是隐变量,缺少显式的监督信息指导,导致注意力机制普遍存在离焦问题(Liu 等,2017)。为解决该问题,Lu 等人(2017)提出并不是每个本文都对应具体的图像区域,对于部分虚词和注意力机制不置信的情况,引入视觉信息将误导文本预测的结果。因此提出一种“哨兵”模型,当注意力机制的输出结果不足以对预测的单词提供有效的指导信息时,依赖语言模型进行文本预测。Huang 等人(2019a)通过分析注意力机制预测的结果与输入单词的相关性,提取可靠信息对图像编码特征和输入词向量进行加权,以修正注意力机制的输出结果。除此之外,Zhou 等人(2019)额外引入名词在图像中的位置信息,显式地监督注意力机制的学习。然而,收集描述中的名词在图像对应位置的标注信息引入了额外的标注成本。Zhou 等人(2020)提出基于图像和文本的匹配模型进行知识蒸馏,以提高注意力机制的定位能力,降低了监督信息的标注成本。Ma 等人(2020)提出对预测的单词重建作为对注意力机制的规则化,以避免注意力机制关注不相关的图像区域。Zhang 等人(2021)通过视觉图模型和语言图模型的对齐操作提高注意力机制的准确性。这些方法都一定程度地改善了注意力机制,但准确性远低于预期效果。

1.2 语言生成模型

语言生成模型旨在预测句子中文本生成的概率。当前,图像描述任务中的语言模型可以分为两类,一类是基于 LSTM 的模型(Vinyals 等,2015),主要结构基于单层 LSTM 或多层 LSTM 进行序列预测;另一类是基于 Transformer 的模型(Vaswani 等,2017)。

LSTM 可以对时间序列进行关联建模,为生成复杂的文本序列奠定了基础。在该方案中,图像的特征编码作为 LSTM 的第 1 个词向量输入,其后每一时刻以前一时刻预测的文本作为词向量的输入,预测下一时刻的输出单词(Vinyals 等,2015)。然而,该过程较大程度地依赖语言模型,忽视了图像的视觉信息。Gu 等人(2018)设计了一种双层 LSTM 序列生成器,第 1 层 LSTM 生成粗略的图像描述,第 2 层 LSTM 以第 1 层 LSTM 的输出作为输入,生成更加准确的图像描述。Huang 等人(2019b)进一步改进多层 LSTM 结构,针对 LSTM 预测不够准确的问题,提出基于每层输出结果的置信度,动态决定是否需要引入更深的 LSTM 修正预测结果。Guo 等人(2020)提出先通过标准的 LSTM 模型输出完整的图像描述,随后结合完整描述的上下文对每个单词进行修正。然而,LSTM 对较长的序列建模能力不足。同时,LSTM 的训练过程是串行的,导致模型训练较为耗时。

Transformer 的模型结构广泛用于自然语言处理领域(Vaswani 等,2017),并逐渐应用于自动图像描述任务。标准的 Transformer 编码器采用多层的自注意力操作(self-attention)实现图像的上下文关联。解码器对生成的单词采用掩膜化的自注意力操作(masked self-attention),建模文本序列的上下文信息,同时采用跨模态注意力模块(cross attention)动态地更新图像的特征编码,以输出正确文本。同时,解码器通过自堆叠形成更加鲁棒的词汇预测。然而,堆叠增加了模型的深度,伴随而来的梯度消失使模型训练困难。

本文对 Transformer 的结构进行扩展,提出一种新颖的多层级解码器动态融合的图像描述模型。该模型通过解码器级联实现注意力机制的渐进式精化,并设计动态融合策略,自适应地融合由粗到精的

文本信息，提高文本描述的准确性。同时，缓解了梯度消失现象，使模型训练更加稳定。

2 模型设计

本文模型的整体结构如图 1 所示。模型采取编码器—解码器架构。对于输入图像 $\boldsymbol{I}$，其对应的语言描述为 $y_{1:T}$，其中 T 为文本描述的最大长度。$\boldsymbol{I}$ 经过卷积神经网络抽取图像的网格特征（grid features）。对于 $w \times h$ 的网格特征，每个特征向量都对应于原始图像特定区域的高层语义表示。将网格特征扁平化排列后（flatten），通过自注意力机制进一步增强，最终得到图像的视觉特征编码 $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$，其中，$\boldsymbol{x}_i$ 是 d_x 维的特征向量，$N = w \times h$ 是网格特征的总数。解码器则基于图像的编码特征预测描述图像内容的语句。不同于标准的解码器，本文提出的解码器采取级联结构，下一级解码器以上一级解码器预测的文本为指导，由粗到精地逐渐提高预测精度，从而生成更加准确的图像描述。同时，设计了一种自适应融合模型，结合多层次文本的输出实现对序列的综合预测，使图像描述更加准确。

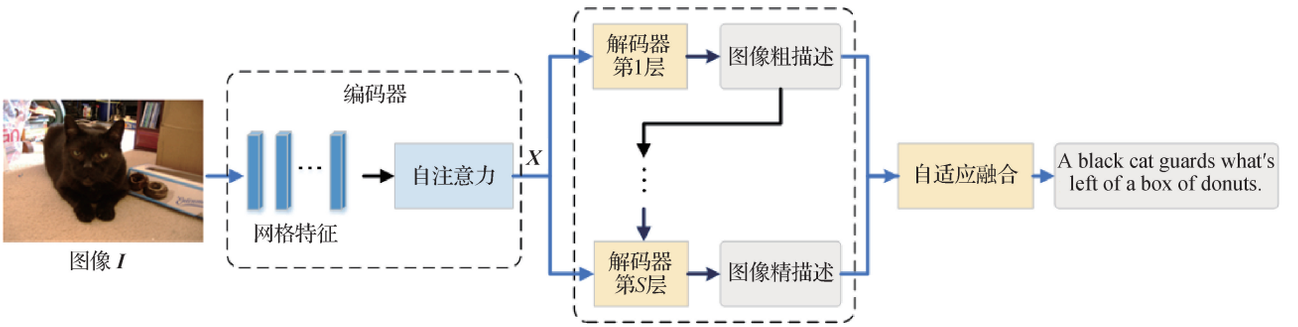


图 1　基于多层级解码器和自适应融合的图像描述模型的整体框架

Fig. 1　Overall framework of the proposed method

2.1 标准解码器结构

本文提出的解码器基本结构是标准 Transformer 解码器（Vaswani 等，2017），包含 1 个跨模态注意力模块和 1 个文本生成模块。跨模态注意力模块通过基于点乘的注意力机制（dot-product attention）建模文本与图像之间的跨模态关联。具体地，该机制以查询矩阵 $\boldsymbol{Q} \in \mathbf{R}^{M\times d}$、键矩阵 $\boldsymbol{K} \in \mathbf{R}^{N\times d}$ 和值矩阵 $\boldsymbol{V} \in \mathbf{R}^{N\times d}$ 为输入。查询矩阵由 M 个 d 维向量构成，键矩阵和值矩阵由 N 个 d 维向量构成。首先，通过计算查询矩阵与键矩阵之间的相似性预测在 N 个不同的值向量上的权重矩阵，计算为

$$\boldsymbol{\alpha} = A(\boldsymbol{Q},\boldsymbol{K}) = f_{\text{softmax}}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right) \tag{1}$$

式中，$\boldsymbol{\alpha} = [a_1, a_2, \cdots, a_N]$ 描述了不同的值向量对应的注意力权重，$A(\,)$ 为权重计算函数。较大的权重表示对应的值向量与查询的相关性更大。随后，结合权重矩阵和值矩阵，对不同的值向量加权融合，得到经过注意力机制聚合后的向量表示，具体为

$$\boldsymbol{Z} = f_{\text{Attention}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = A(\boldsymbol{Q},\boldsymbol{K})\boldsymbol{V} \tag{2}$$

式中，$f_{\text{Attention}}(\,)$ 为注意力机制的计算函数。在图像描述任务中，以文本序列编码矩阵 $\boldsymbol{Y}$ 和视觉特征编码 $\boldsymbol{X}$ 为输入。跨模态注意力模块首先将 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 通过线性映射形成查询矩阵、键矩阵和值矩阵，并通过多头注意力机制预测对下一时刻输出单词具有区分性的视觉特征，并通过前馈神经网络（feed forward network）输出最终的特征向量 $\boldsymbol{F}$。即

$$\begin{aligned}\boldsymbol{Z} &= f_{\text{Attention}}(\boldsymbol{W}_q\boldsymbol{Y},\boldsymbol{W}_k\boldsymbol{X},\boldsymbol{W}_v\boldsymbol{X})\\ \boldsymbol{F} &= FFN(\boldsymbol{Z})\end{aligned} \tag{3}$$

式中，$\boldsymbol{W}_q$、$\boldsymbol{W}_k$ 和 $\boldsymbol{W}_v$ 是可学习的参数，$FFN(\,)$ 为前馈神经网络的计算函数。由于文本序列生成过程中，第 t 个单词的预测仅由 t 时刻之前的文本进行推测。因此，文本序列的编码矩阵 $\boldsymbol{Y}$ 由前时刻所有的预测单词 $\tilde{y}_{1:t-1}$ 经过掩膜化的自注意力操作（masked self-attention）形成。即

$$\boldsymbol{Y} = SA_{\text{mask}}(\tilde{y}_{1:t-1}) \tag{4}$$

式中，$SA_{\text{mask}}(\,)$ 为经过掩膜化的自注意力函数。

最后，基于生成的加权图像特征编码，预测输出单词的概率分布，以预测该时刻的目标单词。具体为

$$\tilde{y}_t \sim f_{\text{softmax}}(\boldsymbol{W}_e\boldsymbol{F}) \tag{5}$$

式中，$\boldsymbol{W}_e$ 是可学习的投影矩阵，将 $\boldsymbol{F}$ 映射为输出单词的概率分布。

由于 $\tilde{y}_{1:t-1}$ 与 t 时刻的真实文本 y_t 可能差别较大，且 $\tilde{y}_{1:t-1}$ 可能包含错误，通过有限的文本序列难以提供充分信息使注意力机制关注在正确的图像区域，导致文本预测不够准确。为解决这一问题，本文提出解码器级联结构对预测的文本进行修正，以提高文本预测的精度。

2.2 级联解码器结构

本文提出的解码器级联结构如图 2 所示。首先，通过标准的解码器，基于 t 时刻之前预测的单词 $\{\tilde{y}_1,\tilde{y}_2,\cdots,\tilde{y}_{t-1}\}$，预测 t 时刻的输出 $\tilde{y}_t^1$。在之后的第 s 层解码器（$s\in\{2,3,\cdots,S\}$），结合前一层解码器预测的单词 $\tilde{y}_t^{s-1}$ 和前时刻的所有预测单词 $\tilde{y}_{1:t-1}$，经过掩膜化的自注意力操作形成文本序列编码。具体为

$$\boldsymbol{Y}^s = SA_{\text{mask}}([\tilde{y}_{1:t-1},\tilde{y}_t^{s-1}]) \tag{6}$$

式中，$[\cdot,\cdot]$ 是拼接操作。对于第 s 级解码器，跨模态注意力模块以图像的视觉特征编码 $\boldsymbol{X}$ 和文本序列的编码矩阵 $\boldsymbol{Y}^s$ 为输入，对 t 时刻的预测单词进行更新。具体为

$$\begin{aligned}\boldsymbol{Z}^s &= f_{\text{Attention}}(W_q\boldsymbol{Y}^s,W_k\boldsymbol{X},W_v\boldsymbol{X})\\ \boldsymbol{F}^s &= FFN(\boldsymbol{Z}^s)\\ \tilde{y}_t^s &\sim f_{\text{softmax}}(\boldsymbol{W}_e\boldsymbol{F}^s)\end{aligned} \tag{7}$$

高层解码器相比于标准的单层解码器，引入了额外的文本信息 $\tilde{y}_t^{s-1}$ 为指导。$\tilde{y}_t^{s-1}$ 相比 $\tilde{y}_{t-1}$ 更接近真实预测的单词 y_t，因此，$A(\boldsymbol{Y}^s,\boldsymbol{X})$ 生成的注意力权重更真实地反映了单词对视觉特征 $\boldsymbol{X}$ 的重要性，从而提高预测单词的准确性。随着 $\tilde{y}_t^s$ 的精确度逐渐提升，后续解码器的注意力机制将更加准确，从而渐进式地提高文本预测的准确性。

2.3 多层级解码器自适应融合

解码器级联结构包含了文本由粗到精的预测结果，蕴含了描述图像内容的丰富细节。为进一步提高模型预测的准确性，本文提出一种自适应融合结构，以最大化利用不同层级解码器的输出结果。具体地，基于门机制（gating mechanism），动态地预测权重 β，以控制不同解码器的输出信息流。如图 3 所示，第 s 层解码器的权重 β^s 由注意力机制的输出 $\boldsymbol{F}^s$ 和输入文本序列的编码 $\boldsymbol{Y}^s$ 共同决定。即

$$c^s = \boldsymbol{W}_s[\boldsymbol{Y}^s,\boldsymbol{F}^s]$$

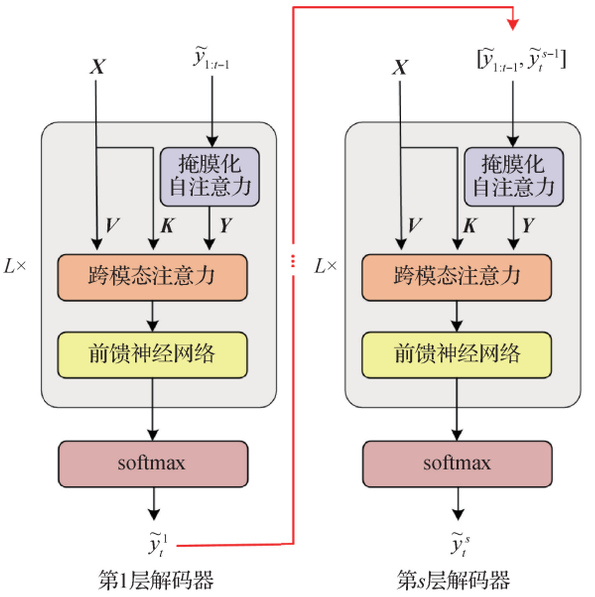


图 2 级联解码器结构示意图

Fig. 2 Architecture of the hierarchical decoders

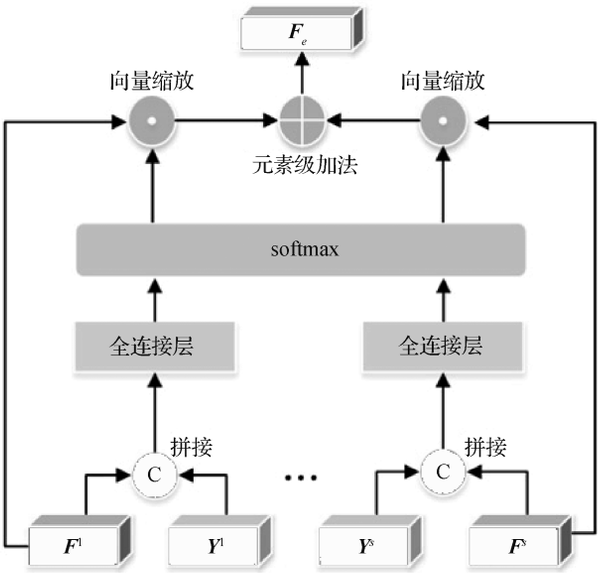


图 3 自适应融合结构示意图

Fig. 3 Architecture of the dynamic fusion module

$$\boldsymbol{\beta} = f_{\text{softmax}}([c^1,c^2,\cdots,c^S]) \tag{8}$$

式中，$[\cdot,\cdot]$ 是拼接操作，$\boldsymbol{W}_s\in\mathbf{R}^{1\times 2d}$ 是可学习权重矩阵，$\boldsymbol{\beta}=[\beta^1,\beta^2,\cdots,\beta^S]$ 代表不同解码器的权重。不同于传统的门机制仅依赖单路信息流预测其对应的权重（Cornia 等，2020），本文提出的门机制同时输入多路信息流，引入具有互斥功能的 softmax 函数感知不同层解码器的上下文信息，融合全局信息流，以指导权重的自适应调整。

随后，自适应融合模块基于已学习的权重对不同层的注意力特征进行集成。即

$$\boldsymbol{F}_e = \sum_{s=1}^{S} \beta^s \boldsymbol{F}^s \tag{9}$$

最后，基于集成后的特征预测最终的输出单词。具体为

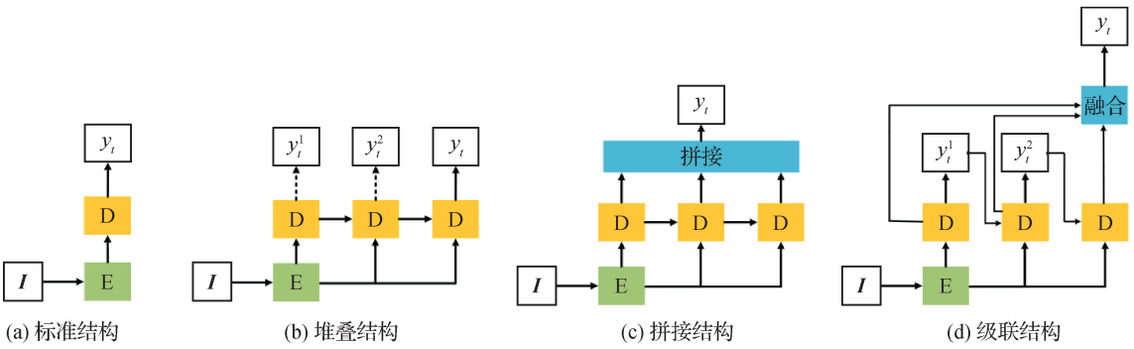
$$\tilde{y}_t \sim \boldsymbol{p} = f_{\text{softmax}}(\boldsymbol{W}_e \boldsymbol{F}_e) \tag{10}$$

动态融合结构能为多层级解码器更好地引入监督信息并缓解梯度消失。以最容易形成梯度消失现象的第 1 级解码器为例，设模型学习的损失函数为 L，第 1 级解码器的参数为 θ^1。由式(8)—式(10)可知，θ^1 的梯度计算为

$$\frac{\partial L}{\partial \theta^1} = \frac{\partial L}{\partial \boldsymbol{F}_e}\left[\sum_{s=1}^{S}\left(\frac{\partial \beta^s}{\partial \theta_1}\boldsymbol{F}^s + \frac{\partial \boldsymbol{F}^s}{\partial \theta_1}\beta^s\right)\right] \tag{11}$$

由式(11)可见，梯度包含 $\left(\beta^1 \frac{\partial L}{\partial \boldsymbol{F}_e}\right)\frac{\partial \boldsymbol{F}^1}{\partial \theta_1}$，该部分不受其他各级解码器影响。由此可见，由于第 1 层解码器的输出 $\boldsymbol{F}^1$ 直接参与最终的文本预测，相应地，模型学习过程中的梯度也直接反馈给 θ^1，从而解决了传统级联结构容易产生的梯度消失现象，使模型训练更加稳定。

以 3 层解码器为例，本文提出的级联解码器与其他多层解码器结构相比，主要区别如图 4 所示。图中，E 代表标准编码器，D 代表标准解码器，y_t 表示最终预测文本，y_t^1 和 y_t^2 为中间预测结果，虚线连接的结构仅在模型训练时有效。首先，在多层级解码器的结构上，本文提出的级联结构以前一级解码器预测的单词作为指导信息，提高了下一级解码器的精度。其次，本文提出一种自适应融合策略集成不同解码器的预测结果，提高了模型预测的准确性。



(a) 标准结构　(b) 堆叠结构　(c) 拼接结构　(d) 级联结构

图 4　不同解码结构的对比示意图

Fig. 4　The architectures of different decoders ((a) vanilla decoder architecture; (b) stacked multi-layer decoder; (c) concatenated multi-layer decoder; (d) hierarchical decoders)

2.4　学习策略

本文采用图像自动描述的标准训练方法(Rennie 等，2017)，将训练过程分为两个阶段。第 1 阶段对每个时刻生成的单词采用交叉熵损失函数(cross-entropy loss)进行训练，第 2 阶段采用强化学习对描述生成的模型进行调优。

在以交叉熵损失函数为目标的训练阶段，通过输入真实文本 $y_{1:t-1}$，预测与之对应的下一单词。记模型的参数为 θ，损失值为 L_{XE}。采用最大似然估计，以最大化真实单词 y_t 的预测概率，具体为

$$L_{\text{XE}}(\theta) = -\sum_{t=1}^{T} \log(p_\theta(y_t \mid y_{1:t-1})) \tag{12}$$

式中，T 为句子的长度。交叉熵损失函数预测过程简单，但是每个单词独立优化，导致生成的句子完整性和流畅性不足。

为解决该问题，本文以交叉熵损失函数训练得到的 θ 作为初始值，以 SCST(self-critical sequence training)强化学习(Rennie 等，2017)为模型训练的第 2 阶段，进一步优化文本描述的评价指标。具体地，解码器的输出作为“实体”与外部环境进行交互。“行为”则是对下一个单词预测。在预测完整的文本序列后，“实体”收到一个奖励(reward)。本文定义奖励为预测的图像描述与真实描述之间的相似性，用语言评价指标 CIDEr 描述。强化学习的目标是最小化负的期望奖励，具体为

$$L_{\text{RL}}(\theta) = -E_{\tilde{y}_{1:T}\sim p_\theta}[r(\tilde{y}_{1:T})] \tag{13}$$

式中，L_{RL} 为第 2 阶段的损失值，$r(\cdot)$ 是奖励函数，即生成文本的 CIDEr 得分。$\tilde{y}_{1:T}$ 是模型依据 p_θ 使用

蒙特卡洛采样生成的句子描述，E 为期望。使用 SCST 方法，梯度损失可以近似为

$$\nabla_\theta L_{\mathrm{RL}}(\theta) = -\frac{1}{N}\sum_{t=1}^{T}(r(\tilde{y}_{1:T}) - b)\nabla_\theta \log(p(y_t)) \tag{14}$$

式中，b 代表基础模型生成的图像描述对应的奖励分数。本文采用贪婪算法（greedy decoding）作为基础模型。

在序列的预测过程中，本文采用集束搜索策略（beam search），即每个时刻从解码器的概率分布中采样概率最大的前 k 个单词，并在解码过程中始终保留置信度最高的前 k 个文本序列。最后，将置信度最高的序列作为预测的文本描述。

3 实验结果与分析

3.1 数据集和评估指标

实验在 MS COCO（Lin 等，2014）和 Flickr30K（Plummer 等，2015）公开数据集上进行，对图像描述模型进行评价。MS COCO 数据集包含 123 287 幅图像，Flickr30K 数据集包含 31 783 幅图像。两组数据集均涵盖广泛的自然场景，每幅图像均提供 5 条参考描述。实验采用 Karpathy 和 Li（2015）提出的训练集和测试集划分方法对模型进行训练和评估。对 MS COCO 数据集，分别取 82 783、5 000 和 5 000 幅图像及其描述作为训练集、验证集和测试集。对 Flickr30K 数据集，分别取 29 000、1 000 和 1 000 幅图像及其描述作为训练集、验证集和测试集。

为评估模型生成图像描述的质量，采用 BLEU-1、BLEU-4（Papineni 等，2002）、METEOR（Banerjee 和 Lavie，2005）和 CIDEr（Vedantam 等，2015）等标准的图像描述评估标准验证模型的效果。以上指标分别记为 B-1、B-4、M 和 C。B-1 和 B-4 评价预测语句与参考语句之间 1 元组和 4 元组共同出现的程度，衡量预测语句的准确性；METEOR 描述句子中连续且顺序相同的文本数量，反映语句的流畅度；CIDEr 使用语法匹配测量生成句子与参考语句之间的语义相似性，与人类的主观评价一致。

3.2 实施细节

本文基于深度学习框架 Pytorch 实现所述模型，模型的训练和测试均使用 2080TI GPU。在图像的编码器部分，采用 Jiang 等人（2020）的方法抽取图像的网格特征，其中网格大小为 7 × 7，每个特征表示为 2 048 维的向量。文本的编码采用标准的词嵌入模型（Cornia 等，2020）。模型的实现细节中，本文参照 Transformer 的一般设置，将维度 d 设为 512，FFN 的隐藏层特征维度设为 2 048，dropout 的概率为 0.1。对于每层解码器，L 设为 1。采用 ADAM（adaptive momentum estimation）优化器进行模型训练，批处理大小（batch size）设为 50。在交叉熵学习阶段，初始学习率设为 0.000 5，学习率变化过程参照模型训练的一般设置（Cornia 等，2020）。如果训练过程中，验证集的 CIDEr 连续下降 5 个训练周期（epoch），则进入强化学习阶段。在强化学习阶段，学习率固定为 5×10^{-6}。当验证集的 CIDEr 连续下降 5 个训练周期后，模型训练结束。在测试过程中，集束搜索中 k 值设为 5。

3.3 消融实验与分析

为验证多层级解码器动态融合的有效性，设计 4 种不同结构与本文提出的模型进行对比。第 1 种结构（图 4（b））为级联结构中每层解码器独立地设计损失函数，预测过程依靠最终解码器输出的结果，该结构记为堆叠；第 2 种结构（图 4（c））对不同解码器的输出拼接后预测文本序列，该结构记为拼接；第 3 种结构将式（8）采用的 softmax 门函数替换为 sigmoid 门函数，以独立计算不同解码器的权重；第 4 种结构将式（8）中的 $\boldsymbol{W}^s$ 设为 $d\times2d$ 的权重矩阵，$\boldsymbol{\beta}^s$ 此时为与解码器输出特征维度相同的矢量，对不同维度的特征赋予不同的融合权重。不同的解码器结构性能对比结果如表 1 所示。

从表 1 可以看出，相比于堆叠和拼接，自适应加权融合方法在 MS COCO 和 Flickr30K 数据集都具有明显优势。具体地，堆叠结构的 CIDEr 在 MS COCO 数据集上下降了 1.4，在 Flickr30K 数据集上下降显著，比本文方法低 4.3。拼接结构结果相似。在门函数设计方面，采用 sigmoid 门函数预测不同层解码器的权重使 CIDEr 在 MS COCO 数据集上下降了 1.3，在 Flickr30K 数据集上下降了 0.06。这意味着通过 softmax 操作引入不同层解码器的上下文关联对于解码器的权重控制十分重要。最后，对比基于矢量权重的融合方法，标量权重能够显著提高图像描述的准确性。特别地，基于矢量权重的融合方法在 Flickr30K 数据集上的 CIDEr 仅为 62.0，显著低于基于标量权重的融合方法。原因是矢量权重增

表 1　不同的解码器结构对图像描述性能的影响

Table 1　Ablation study on different decoder architectures

结构	融合模式	门函数	标量权重	MS COCO				Flickr30K			
				B-1	B-4	M	C	B-1	B-4	M	C
基准	基准	–	–	80.1	38.8	28.7	127.2	71	28	21.6	66.1
对比 1	堆叠	–	–	80.9	39.1	29.2	130.4	71.4	29.4	22	65.6
对比 2	拼接	–	–	81	39.7	29.1	130.2	72.1	29.5	21.9	65.9
对比 3	级联	sigmoid	√	80.1	38.9	**29.3**	130.1	73.1	**31.1**	22.4	69.8
对比 4	级联	softmax	–	**81.1**	**39.7**	**29.3**	130.8	70.4	28.5	21.7	62
本文	级联	softmax	√	81	38.9	**29.3**	**131.8**	**73.5**	31	**22.7**	**69.9**

注：加粗字体表示各列最优结果。“–”表示不添加函数，“√”表示添加对应函数。

加了模型参数量，使预测结果对噪声干扰更加敏感，因此在较小的 Flickr30K 数据集上性能下降更加明显。

为进一步分析级联结构的有效性，实验对 S 的变化对图像描述性能的影响进行分析，结果如图 5 和图 6 所示。可以看出，当 S 取 3 时，模型在 MS COCO 和 Flickr30K 测试集上性能均达到最佳，这与标准的 Transformer 堆叠的数量一致。因此，在后续实验中，本文将 S 设置为 3。

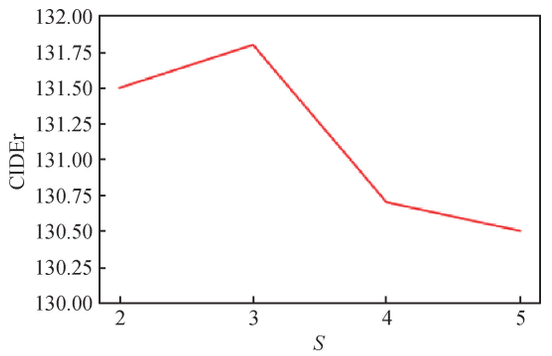


图 5　参数 S 对 MS COCO 测试集性能的影响

Fig. 5　The impact of S on MS COCO test set

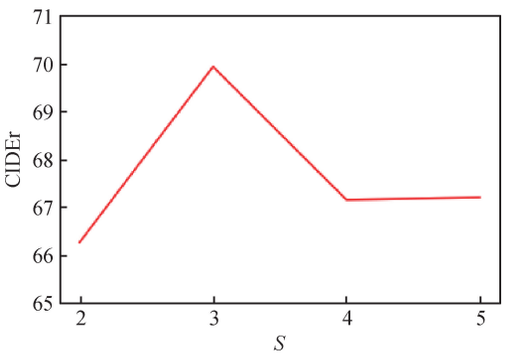


图 6　参数 S 对 Flickr30K 测试集性能的影响

Fig. 6　The impact of S on Flickr30K test set

3.4　对比实验与分析

实验挑选 12 种代表性方法与本文提出的模型开展定量比较。包括 Up-Down（Anderson 等，2018）、Transformer（Vaswani 等，2017）、M2（meshed-memory Transformer）（Cornia 等，2020）、POS-SCAN（part-of-speech enhanced stacked cross attention）（Zhou 等，2020）、GVD（grounded video description）（Zhou 等，2019）、Stack-Cap（Gu 等，2018）、AAT（adaptive attention time）（Huang 等，2019b）、RD（ruminant decoding）（Guo 等，2020）、CGRL（consensus graph representation learning）（Zhang 等，2021）、Cyclical（Ma 等，2020）、SOCPK（scene and object category prior knowledge）（汤鹏杰 等，2017）和 CMFF/CD（cross-layer multi-model feature fusion and causal convolutional decoding）（罗会兰和岳亮亮，2020）。其中，Up-Down 和 Transformer 是基准模型；M2 是目前性能最好的图像描述模型；SCAN、CGRL 和 GVD 通过修正注意力机制提高图像描述的准确性；Stack-Cap、RD 和 Cyclical 通过引入解码器级联结构提高图像描述的性能；SOCPK 和 CMFF/CD 通过改善图像的特征表示提高图像描述的准确性。

表 2 展示了不同方法在 MS COCO 和 Flickr30K 数据集上的对比结果。

在 MS COCO 数据集的实验结果表明，本文方法显著改善了基于 Transformer 的基准模型，同时高于其他对比方法。具体地，对于描述短语重叠率的评估指标，B-1 指标比 M2 提高了 0.5，说明本文提出的模型能精确地输出描述图像的单词；对于描述

表 2　不同方法在 MS COCO 和 Flickr30K 测试集的性能比较

Table 2　Comparison of performance among different methods on the MS COCO and Flickr30K test set

模型	MS COCO				Flickr30K			
	B-1	B-4	M	C	B-1	B-4	M	C
Up-Down	79.8	36.3	27.7	120.1	–	26.4	21.5	57.0
Transformer	80.1	38.8	28.7	127.2	71.0	28	21.6	66.1
M2	80.5	38.9	29.2	130.8	73.0	30.9	22.4	67.7
SOCPK	71.0	28.1	23.9	88.2	62.7	21.7	19.7	43.9
CMFF/CD	72.1	31.0	24.6	94.6	64.6	19.7	19.1	39.5
Cyclical	–	–	–	–	69.9	27.4	22.3	61.4
GVD	–	–	–	–	69.9	27.3	22.5	62.3
POS-SCAN	80.2	38	28.5	126.1	73.4	30.1	22.6	69.3
Stack-Cap	78.6	36.1	27.4	120.4	–	–	–	–
AAT	–	38.7	28.6	128.6	–	–	–	–
RD	–	38.6	28.7	128.3	–	26.8	20.5	57.0
CGRL	–	–	–	–	72.5	27.8	22.4	65.2
本文	**81.0**	**38.9**	**29.3**	**131.8**	**73.5**	**31.0**	**22.7**	**69.9**

注：加粗字体表示各列最优结果，“–”表示该方法原文未提供数据。

句子流畅程度的指标，M 指标相比对比方法中的最好结果也略有改善。对于描述语义相似性的指标，CIDEr 提升更显著，相比当前最好的模型 M2 提高 1.0，说明模型能更好地输出与人类主观描述一致的文本序列。对比 Transformer、M2、Stack-Cap、AAT 和 RD 在各项指标上的性能，本文方法性能均高于对比方法。值得注意的是，在 Transformer 和 M2 结构中，堆叠的参数 L 设置为 3，与本文的层级 S 取值一致，表明本文模型复杂度与 Transformer 和 M2 等对比方法接近，也间接证明了本文提出的级联结构设计的有效性。

在 Flickr30K 数据集上的实验结果表明，本文模型在较小数据集上能够保持良好描述效果。具体地，相比 M2 模型，本文方法在 CIDEr 上提高了 2.2。B-1、B-4 和 M 指标也均高于 M2。相比引入额外监督信息的 SCAN 和 GVD 方法，本文提出的模型在 CIDEr 指标上分别高出 0.6 和 7.6。以上结果表明，本文提出模型同时关注了图像描述的准确性、流畅性和语义的正确性。

3.5　可视化分析

图 7 展示了本文模型与 Transformer 基准模型在 MS COCO 测试集上对部分图像的描述对比。整体来看，本文方法能够输出更加准确和丰富的图像描述。例如，图 7 第 1 行，本文模型能够准确预测出猫旁边小物体是 a box of donuts，而不是 toy；图 7 第 2 行，本文模型能够在同类物体密集出现条件下正确预测量词。为了进一步验证多层次解码器的有效性，本文对跨模态注意力机制进行可视化分析。由图 7(b)可见，Transformer 基准模型关注的视觉区域更分散，受背景干扰较大。例如，图 7 第 1 行，注意力机制部分关注于“猫”后方的背景区域，从而对描述“猫”周围环境时造成干扰。对比图 7(c)可见，本文提出的级联解码结构能够准确定位至图像的相关区域，从而生成更加准确的文字描述。以上可视化分析结果从另一角度验证了本文方法的有效性。

4　结　论

本文提出了一种结合多层级解码器和动态融合机制的图像描述模型。通过设计解码器级联结构实现注意力机制的渐进式精化。其中，高层级的解码器以低层级解码器的预测结果为输入。由于该输入与预测的目标单词更相关，注意力机制能够更有效

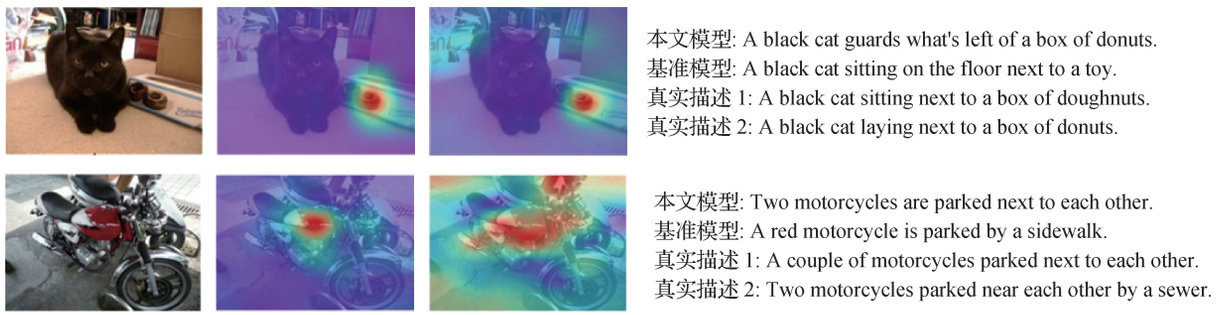


图 7　本文模型与 Transformer 基准模型在 MS COCO 测试集上的注意力机制可视化结果和生成的描述对比

Fig. 7　Examples of image captioning results by Transformer and the proposed model ((a) original images; (b) ours; (c) baseline)

地聚焦到图像的关键区域,从而生成更准确的文本序列。此外,设计了一种解码器动态融合策略,根据每级解码器的输出动态地调整输出权重,自适应地融合由粗到精的文本信息,提高图像描述的鲁棒性。同时,动态融合为不同层次解码器引入监督信息,进一步解决了级联结构容易产生的梯度消失现象,使模型训练更加稳定。但是自动图像描述的准确率还有进一步提升空间。下一步工作将尝试改进图像的特征表达以提高图像描述的丰富性,优化图像的视觉特征和语言模型的关联以提高自动图像描述模型的鲁棒性。

参考文献(References)

Anderson P, He X D, Buehler C, Teney D, Johnson M, Gould S and Zhang L. 2018. Bottom-up and top-down attention for image captioning and visual question answering//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 6077-6086 [DOI: 10.1109/CVPR.2018.00636]

Banerjee S and Lavie A. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments//Proceedings of 2005 ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Ann Arbor, USA: Association for Computational Linguistics: 65-72

Cornia M, Stefanini M, Baraldi L and Cucchiara R. 2020. Meshed-memory transformer for image captioning//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 10575-10584 [DOI: 10.1109/CVPR42600.2020.01059]

Farhadi A, Hejrati M, Sadeghi M A, Young P, Rashtchian C, Hockenmaier J and Forsyth D. 2010. Every picture tells a story: generating sentences from images//Proceedings of the 11th European Conference on Computer Vision. Heraklion, Greece: Springer: 15-29 [DOI: 10.1007/978-3-642-15561-1_2]

Gu J X, Cai J F, Wang G and Chen T. 2018. Stack-captioning: coarse-to-fine learning for image captioning//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI: 6837-6844

Guo L T, Liu J, Lu S C and Lu H Q. 2020. Show, tell, and polish: ruminant decoding for image captioning. IEEE Transactions on Multimedia, 22(8): 2149-2162 [DOI: 10.1109/TMM.2019.2951226]

Hendricks L A, Burns K, Saenko K, Darrell T and Rohrbach A. 2018. Women also snowboard: overcoming bias in captioning models//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 793-811 [DOI: 10.1007/978-3-030-01219-9_47]

Huang L, Wang W M, Chen J and Wei X Y. 2019a. Attention on attention for image captioning//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 4633-4642 [DOI: 10.1109/ICCV.2019.00473]

Huang L, Wang W M, Xia Y X and Chen J. 2019b. Adaptively aligned image captioning via adaptive attention time//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates, Inc.: 8942-8951

Jiang H Z, Misra I, Rohrbach M, Learned-Miller E and Chen X L. 2020. In defense of grid features for visual question answering//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 10264-10273 [DOI: 10.1109/CVPR42600.2020.01028]

Karpathy A and Li F F. 2015. Deep visual-semantic alignments for generating image descriptions//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE:

3128-3137 [DOI: 10.1109/CVPR.2015.7298932]

Kuznetsova P, Ordonez V, Berg T L and Choi Y. 2014. TreeTalk: composition and compression of trees for image descriptions. Transactions of the Association for Computational Linguistics, 2: 351-362 [DOI: 10.1162/tacl_a_00188]

Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick C L. 2014. Microsoft COCO: common objects in context//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer: 740-755 [DOI: 10.1007/978-3-319-10602-1_48]

Liu C X, Mao J H, Sha F and Yuille A. 2017. Attention correctness in neural image captioning//Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, USA: AAAI: 4176-4182

Lu J S, Xiong C M, Parikh D and Socher R. 2017. Knowing when to look: adaptive attention via a visual sentinel for image captioning//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 3242-3250 [DOI: 10.1109/CVPR.2017.345]

Luo H L and Yue L L. 2020. Image caption based on causal convolutional decoding with cross-layer multi-model feature fusion. Journal of Image and Graphics, 25(8): 1604-1617 (罗会兰, 岳亮亮. 2020. 跨层多模型特征融合与因果卷积解码的图像描述. 中国图象图形学报, 25(8): 1604-1617) [DOI: 10.11834/jig.190543]

Ma C Y, Kalantidis Y, AlRegib G, Vajda P, Rohrbach M and Kira Z. 2020. Learning to generate grounded visual captions without localization supervision//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 353-370 [DOI: 10.1007/978-3-030-58523-5_21]

Papineni K, Roukos S, Ward T and Zhu W J. 2002. BLEU: a method for automatic evaluation of machine translation//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, USA: ACL: 311-318 [DOI: 10.3115/1073083.1073135]

Plummer B A, Wang L W, Cervantes C M, Caicedo J C, Hockenmaier J and Lazebnik S. 2015. Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 2641-2649 [DOI: 10.1109/ICCV.2015.303]

Rennie S J, Marcheret E, Mroueh Y, Ross J and Goel V. 2017. Self-critical sequence training for image captioning//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1179-1195 [DOI: 10.1109/CVPR.2017.131]

Rohrbach A, Hendricks L A, Burns K, Darrell T and Saenko K. 2018. Object hallucination in image captioning//Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: ACL: 4035-4045 [DOI: 10.18653/v1/d18-1437]

Tan Y L, Tang P J, Zhang L and Luo Y P. 2021. From image to language: image captioning and description. Journal of Image and Graphics, 26(4): 727-750 (谭云兰, 汤鹏杰, 张丽, 罗玉盘. 2021. 从图像到语言: 图像标题生成与描述. 中国图象图形学报, 26(4): 727-750) [DOI: 10.11834/jig.200177]

Tang P J, Tan Y L and Li J Z. 2017. Image description based on the fusion of scene and object category prior knowledge. Journal of Image and Graphics, 22(9): 1251-1260 (汤鹏杰, 谭云兰, 李金忠. 2017. 融合图像场景及物体先验知识的图像描述生成模型. 中国图象图形学报, 22(9): 1251-1260) [DOI: 10.11834/jig.170052]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Vedantam R, Zitnick C L and Parikh D. 2015. CIDEr: consensus-based image description evaluation//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 4566-4575 [DOI: 10.1109/CVPR.2015.7299087]

Vinyals O, Toshev A, Bengio S and Erhan D. 2015. Show and tell: a neural image caption generator//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 3156-3164 [DOI: 10.1109/CVPR.2015.7298935]

Wan B Y, Jiang W H, Fang Y M, Zhu M W, Li Q and Liu Y. 2022. Revisiting image captioning via maximum discrepancy competition. Pattern Recognition, 122: #108358 [DOI: 10.1016/j.patcog.2021.108358]

Xu K, Ba J L, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel R S and Bengio Y. 2015. Show, attend and tell: neural image caption generation with visual attention//Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille, France: PMLR: 2048-2057

Zhang W Q, Shi H C, Tang S L, Xiao J, Yu Q and Zhuang Y T. 2021. Consensus graph representation learning for better grounded image captioning//Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtual: AAAI: 3394-3402

Zhou L W, Kalantidis Y, Chen X L, Corso J J and Rohrbach M. 2019. Grounded video description//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 6571-6580 [DOI: 10.1109/CVPR.2019.00674]

Zhou Y E, Wang M, Liu D Q, Hu Z Z and Zhang H W. 2020. More grounded image captioning by distilling image-text matching model//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 4776-4785 [DOI: 10.1109/CVPR42600.2020.00483]

作者简介

姜文晖，1989 年生，男，讲师，主要研究方向为计算机视觉、跨媒体分析、深度学习。
E-mail：jiang1st@ bupt. cn

方玉明，通信作者，男，教授，主要研究方向为计算机视觉、多媒体信号处理、视觉质量评估。
E-mail：fa0001ng@ e. ntu. edu. sg

占锟，男，硕士研究生，主要研究方向为计算机视觉与深度学习。E-mail：zhankun1008@ gmail. com
程一波，男，硕士研究生，主要研究方向为跨媒体分析。
E-mail：592891032@ qq. com
夏雪，女，讲师，主要研究方向为图像处理和语义分割。
E-mail：yeziandkuma@ qq. com