基于语境辅助转换器的图像标题生成算法

连 政[1, 2] 王 瑞[2] 李海昌[2] 姚 辉[2] 胡晓惠[2]

摘 要 在图像标题生成领域, 交叉注意力机制在建模语义查询与图像区域的关系方面, 已经取得了重要的进展. 然而, 其视觉连贯性仍有待探索. 为填补这项空白, 提出一种新颖的语境辅助的交叉注意力 (Context-assisted cross attention, CACA) 机制, 利用历史语境记忆 (Historical context memory, HCM), 来充分考虑先前关注过的视觉线索对当前注意力语境生成的潜在影响. 同时, 提出一种名为"自适应权重约束 (Adaptive weight constraint, AWC)" 的正则化方法, 来限制每个 CACA 模块分配给历史语境的权重总和. 本文将 CACA 模块与 AWC 方法同时应用于转换器 (Transformer) 模型, 构建一种语境辅助的转换器 (Context-assisted transformer, CAT) 模型, 用于解决图像标题生成问题. 基于 MS COCO (Microsoft common objects in context) 数据集的实验结果证明, 与当前先进的方法相比, 该方法均实现了稳定的提升.

关键词 图像标题生成, 注意力机制, 转换器, 视觉连贯性

引用格式 连政, 王瑞, 李海昌, 姚辉, 胡晓惠. 基于语境辅助转换器的图像标题生成算法. 自动化学报, 2023, **49**(9): 1889−1903

DOI 10.16383/j.aas.c220767

Context-assisted Transformer for Image Captioning

LIAN Zheng[1, 2] WANG Rui[2] LI Hai-Chang[2] YAO Hui[2] HU Xiao-Hui[2]

Abstract The cross attention mechanism has made significant progress in modeling the relationship between semantic queries and image regions in image captioning. However, its visual coherence remains to be explored. To fill this gap, we propose a novel context-assisted cross attention (CACA) mechanism. With the help of historical context memory (HCM), CACA fully considers the potential impact of previously attended visual cues on the generation of current attention context. Moreover, we present a regularization method, called adaptive weight constraint (AWC), to restrict the total weight assigned to the historical contexts of each CACA module. We apply CACA and AWC to the Transformer model and construct a context-assisted transformer (CAT) for image captioning. Experimental results on the MS COCO (microsoft common objects in context) dataset demonstrate that our method achieves consistent improvement over the current state-of-the-art methods.

Key words Image captioning, attention mechanism, transformer, visual coherence

Citation Lian Zheng, Wang Rui, Li Hai-Chang, Yao Hui, Hu Xiao-Hui. Context-assisted transformer for image captioning. *Acta Automatica Sinica*, 2023, **49**(9): 1889−1903

图像标题生成 (Image captioning) 是一项跨越计算机视觉与自然语言处理领域的多模态生成式任务[1−5], 其主要目标是自动为图像生成准确的描述性语句. 这要求计算机不仅要充分理解图像中的对象以及它们之间的关系, 还要通过流畅的自然语言表达出图像的内容. 图像标题生成技术具有广泛的应用价值. 在学术研究当中, 它可以推动图文检索、视觉问答等多模态领域技术的发展. 在实际生活当中, 这项技术在幼儿的早期教育和视障人群辅助设备的设计方面发挥着重要作用.

受神经机器翻译领域研究的启发, 早期的基于深度神经网络的图像标题生成算法[6] 采用了经典的编码器–解码器 (Encoder-decoder) 框架, 它将卷积神经网络 (Convolutional neural network, CNN) 作为编码器, 提取图像的全局特征, 再使用循环神经网络 (Recurrent neural network, RNN) 作为解码器对图像特征进行解码, 生成图像标题. 尽管经典的编码器–解码器框架在图像标题生成领域取得了巨大的成功, 但是两个固有的缺陷严重限制了该框架的序列解码能力: 1) 图像的全局信息在初始时刻被一次性地输入到解码器当中, 而解码器缺少特

收稿日期 2022-09-26 录用日期 2023-02-10
Manuscript received September 26, 2022; accepted February 10, 2023
国家重点研发计划 (2019YFB1405100), 国家自然科学基金 (61802380) 资助
Supported by National Key Research and Development Program of China (2019YFB1405100) and National Natural Science Foundation of China (61802380)
本文责任编委 白翔
Recommended by Associate Editor BAI Xiang
1. 中国科学院大学 北京 101408 2. 中国科学院软件研究所天基综合信息系统重点实验室 北京 100190
1. University of Chinese Academy of Sciences, Beijing 101408 2. Science & Technology on Integrated Information System Laboratory, Institute of Software, Chinese Academy of Sciences, Beijing 100190

征筛选的关键模块, 难以捕捉预测单词时所需的相关视觉特征; 2) 在整个序列预测的过程中, 作为解码器的循环神经网络会不断丢失一些重要的视觉信息, 从而导致语言模型在预测后期逐渐缺少了视觉信息的指导, 产生明显的误差累积, 降低图像标题的生成质量.

为了解决上述问题, 注意力机制 (Attention mechanism) 被引入到图像标题生成算法当中. 注意力机制拓展了经典的编码器–解码器框架, 它允许解码器在图像标题生成的不同时刻关注到与当前语义查询最为相关的图像信息. 具体来讲, 基于注意力机制的图像标题生成算法不再只是使用图像的全局特征, 而是首先通过 CNN 提取图像的局部区域特征, 再使用基于长短期记忆 (Long short-term memory, LSTM) 网络的解码器对图像特征进行解码. 在每一个解码时刻, 注意力模块会将 LSTM 提供的隐藏状态作为语义查询, 为图像的各个区域分配不同的注意力权重, 再通过对各部分图像特征进行加权求和, 得到当前时刻的注意力语境特征, 进而指导语言模型生成图像标题. 近年来, 转换器 (Transformer)[7] 在自然语言处理领域得到了广泛的应用, 它通过多头注意力 (Multi-head attention) 机制在多个语义空间中建模查询与键值对之间的关系. 在图像标题生成领域, Transformer 首先依靠自注意力 (Self-attention) 机制实现图像局部信息的融合, 然后通过解码器中的交叉注意力模块向语言模型中引入融合后的视觉特征, 实现不同模态的特征交互.

在当前主流的图像标题生成算法中, 交叉注意力机制在建模语义查询与图像区域之间的关系方面, 发挥着关键性的作用. 然而, 大多数现有的基于注意力机制的算法都忽视了视觉连贯性的潜在影响. 事实上, 我们人类往往会不由自主地回顾先前关注过的信息, 以便在当前时刻做出更加合理的注意力决策. 遗憾的是, 传统的交叉注意力机制无法实现这个意图. 为了弥补这项缺陷, 本文提出了一种新颖的语境辅助的交叉注意力 (Context-assisted cross attention, CACA) 机制. 具体来讲, 在每一个解码时刻, CACA 模块会首先根据当前输入的语义查询, 利用交叉注意力模块从图像特征中提取出与当前查询最为相关的临时语境特征, 并将其保存在历史语境记忆 (Historical context memory, HCM) 中, 然后将 HCM 中全部的历史语境特征与图像的局部特征相拼接, 作为键值对, 再次输入交叉注意力模块, 获取当前时刻最终的注意力语境特征. 同时, 为了限制每个 CACA 模块分配给历史语境的权重总和, 本文提出了一种名为“自适应权重约束 (Adaptive weight constraint, AWC)”的正则化方法, 从优化注意力权重分布的角度提升模型的泛化性能. 本文将 CACA 模块与 AWC 方法同时集成在转换器 (Transformer) 模型上, 构建了语境辅助的转换器 (Context-assisted transformer, CAT) 模型. 尽管 Transformer 模型可以通过自注意力层在一定程度上建模历史语义信息, 然而, 从信息论的角度来讲, 根据数据处理不等式[8] 可知, 输入模型的特征向量在神经网络逐层的特征处理与消息传递过程中, 势必会丢失一部分关键信息, 这将导致交叉注意力模块在某一时刻建模的语义信息无法完整地传递到后续解码过程中并得到充分利用. 为此, CAT 模型采用语境辅助的交叉注意力机制, 通过历史语境记忆保存了历史时刻中完整的交叉注意力语义特征, 充分利用序列预测过程中视觉信息的连贯性, 为解码过程提供更加丰富可靠的语境信息. 本文在流行的 MS COCO (Microsoft common objects in context) 数据集[9] 上, 以多个基于 Transformer 的图像标题生成算法作为基线模型, 通过向解码器中引入 CACA 模块与 AWC 方法, 对所提算法进行了评价. 实验结果表明, 与众多先进的基线模型相比, 本文提出的方法在它们的基础上均实现了稳定的提升.

本文的后续内容安排如下: 第 1 节主要介绍图像标题生成领域的相关工作; 第 2 节详细介绍本文提出的方法; 第 3 节通过大量的对比实验从众多角度对本文方法进行分析; 第 4 节总结本文的研究成果, 并提出下一步的工作设想.

1 图像标题生成算法综述

迄今绝大多数的图像标题生成模型都采用了经典的编码器–解码器框架. 该框架最早被提出并应用于神经机器翻译领域, 取得了显著的成就. 编码器–解码器框架的成功应用极大地促进了序列到序列 (Sequence-to-sequence) 任务的发展. 在早期的图像标题生成模型[6] 中, 该框架首先利用 CNN 提取图像的视觉表征, 再使用 RNN 解码图像特征生成图像标题. 在编码器–解码器框架下, 图像标题生成领域涌现出一大批出色的解决方案[10−12], 这些方法主要从编码器和解码器的组成结构上对图像标题生成模型进行了探索和改进, 然而, 由于在解码器中缺少特征选择的关键模块, 经典的编码器–解码器框架在序列预测能力上受到了很大的限制.

注意力机制是编码器–解码器框架的重要拓展, 它允许解码器在序列生成的每个时刻选择性地关注与当前查询最为相关的特征. 受到人类直觉与神经机器翻译领域研究的启发, Xu 等[13] 首次尝试将视觉注意力机制引入图像标题生成模型中, 以便在生

成描述时动态关注图像的显著区域. 随后, You 等[14] 通过一种语义注意力模型, 选择性地关注编码器提出的语义概念, 并将它们与循环神经网络的隐藏状态相结合. 该模型中的选择与融合形成了一个反馈, 连接了自顶而下和自底而上两种不同的计算方式. Lu 等[15] 提出了一种带有视觉哨兵的自适应注意力模型, 该模型可以决定是否关注视觉特征. Anderson 等[16] 介绍了一种组合的自底向上和自顶向下的注意力机制, 其中, 自底向上的注意力利用 Faster R-CNN 提取对象级别的图像特征, 而自顶向下的注意力负责预测视觉特征上的权重分布. Chen 等[17] 在文献中提出了一种增强的注意力机制, 它将基于刺激的注意力与自顶而下的注意力相结合, 为图像的显著区域提供可靠的先验知识. Huang 等[18] 设计了一种"注意力上的注意力"模块, 来确定注意力结果和查询之间的相关性. Pan 等[19] 提出了一种 X-线性注意力模块, 来模拟多模态输入的二阶相互作用. 最近, Yang 等[20] 提出了一种因果注意力机制, 来处理视觉−语言任务. 因果注意力从前门调整策略出发, 提出了样本内注意力机制和交叉样本注意力机制. 其中, 样本内注意力机制采用了经典的注意力网络, 来捕获语义查询与当前样本中图像特征的关系, 而交叉样本注意力机制负责在整个数据集的图像样本聚类后, 捕获语义查询与各个质心特征之间的关系. 王鑫等[21] 设计了一种显著性特征提取机制, 为语言模型提供最有价值的视觉特征, 指导单词的预测.

近年来, Transformer[7] 在图像标题生成领域得到了广泛的应用. Transformer 由堆叠的编码器层和解码器层组成, 每一个编码器层包括一个自注意力模块和一个前馈模块, 每一个解码器层包括一个掩码自注意力模块、一个交叉注意力模块和一个前馈模块. Herdade 等[22] 在标准 Transformer 模型的基础上, 对识别出的对象设计了一种几何注意力机制, 使得模型能够在编码图像的过程中考虑到对象在空间上的相对信息. Li 等[23] 沿用了 Transformer 架构, 在编码阶段使用了两个独立的 Transformer 编码器分别编码视觉信息和语义信息, 在解码器部分设计了一种纠缠注意力机制, 来弥补传统注意力在两类模态特征之间缺乏的互补性. 此外, Yu 等[24] 对 Transformer 进行了拓展, 提出了一种多模态 Transformer 模型, 该模型利用一种统一的注意力块同时捕获模态内与模态间的特征交互. 之后, Cornia 等[25] 提出了一种完全基于注意力机制的图像标题生成模型, 该模型首先通过记忆增强的编码器学习图像区域之间关系的多级表示, 整合从图像数据中学到的先验知识, 保存在记忆向量当中, 然后在解码阶段采用网状解码器同时利用底层和高层的视觉特征生成高质量的图像标题. Zhang 等[26] 提出了网格增强模块与适应性注意力模块, 并将二者嵌入到 Transformer 中构成 RSTNet. 其中, 网格增强模块通过融合图像网格间的相对几何特征增强模型的视觉表征能力, 适应性注意力模块在解码器做出单词预测的决策之前自适应地度量视觉和语言线索的贡献. Luo 等[27] 提出了一种双层协同 Transformer 网络, 充分利用了图像区域特征与网格特征之间的互补性. 最近, Zeng 等[28] 提出了空间与尺度感知的 Transformer, 它首先采用一个空间感知伪监督模块, 利用特征聚类帮助模型保存网格特征的空间信息, 然后通过一个简单的加权残差连接, 同时探索具有丰富语义的低级和高级编码特征. Wu 等[29] 在 Transformer 解码框架的基础上提出了一种双信息流网络, 它将全景分割特征作为网格特征之外的另一个视觉信息源, 来增强视觉信息对标题序列预测的贡献.

尽管交叉注意力机制在建模语义查询与图像区域之间的关系方面发挥了重要的作用, 极大地提升了编码器−解码器框架在图像标题生成任务上的性能, 但是, 其视觉连贯性对注意力语境生成的潜在影响尚未得到深入研究. 当前大多数基于注意力的图像标题生成算法都忽略了历史语境对产生当前注意力分布的影响. 截至目前, 只有少数研究在注意力机制的视觉连贯性方面进行了探索. Qin 等[10] 提出了回顾算法, 将上一时刻的注意力语境引入当前时刻的语义查询, 以适应人类的视觉连贯性. Lian 等[30] 使用注意力 LSTM 扩展了传统的时序注意力机制, 以捕获之前时间步中产生的注意力权重分布特征. 尽管上述两种解决方案充分考虑了注意力语境的历史信息, 有效地提升了图像标题生成模型的性能, 然而, 它们仅考虑了基于 LSTM 的解码框架, 尚未在流行的 Transformer 模型上实现进一步的探索. 本文在交叉注意力模块的设计上聚焦于 Transformer 解码框架, 充分考虑了 Transformer 在训练阶段的并行解码优势, 在不向注意力网络中添加额外的可训练参数的条件下, 引入视觉连贯性, 显著提升了基线模型的性能. 值得一提的是, 本文提出的 CACA 模块不仅可以扩展 Transformer 模型, 还同样适用于基于 LSTM 的解码框架.

2 基于语境辅助转换器的图像标题生成模型

为了更加清晰地阐述模型的细节, 本节首先回

顾了经典的多头注意力机制, 其次基于 Transformer 解码器结构介绍了语境辅助的交叉注意力机制, 以及其轻量级的网络结构设计, 然后介绍了基于语境辅助转换器的图像标题生成模型的整体框架, 最后提出了结合自适应权重约束的模型优化方法.

2.1 多头注意力机制

多头注意力机制 $f_{mhatt}(Q, K, V)$ 集成了多个并行的缩放点积注意力 (Scaled dot-product attention) 层, 以捕获不同特征子空间中与当前查询相关的语义信息. 具体而言, 它首先利用 h 组不同的线性转换层对输入的查询 Q, 键 K 和值 V 进行投影, 再利用缩放点积注意力网络 $f_{dpatt}(Q, K, V)$ 对每一组投影后的特征进行建模, 提取第 i 个子空间中的相关语义特征 $head_i$, 最后, 将这 h 组从特征子空间中提取到的语境向量拼接在一起, 通过另一个可学习的线性转换层进行投影, 得到最终的多头注意力语境特征. 在此, 本文假设 Q, K, V 的特征维度分别为 d_q, d_k, d_v. 如图 1 所示, 多头注意力机制可由如下公式表达:

$$\begin{cases} f_{mhatt}(Q, K, V) = Concat(head_1, \cdots, head_h)W^O \\ head_i = f_{dpatt}(Q_i, K_i, V_i) \\ f_{dpatt}(Q_i, K_i, V_i) = softmax\left(\dfrac{Q_i K_i^{\mathrm{T}}}{\sqrt{d_k}}\right)V_i \\ Q_i = QW_i^Q,\ K_i = KW_i^K,\ V_i = VW_i^V \end{cases} \tag{1}$$

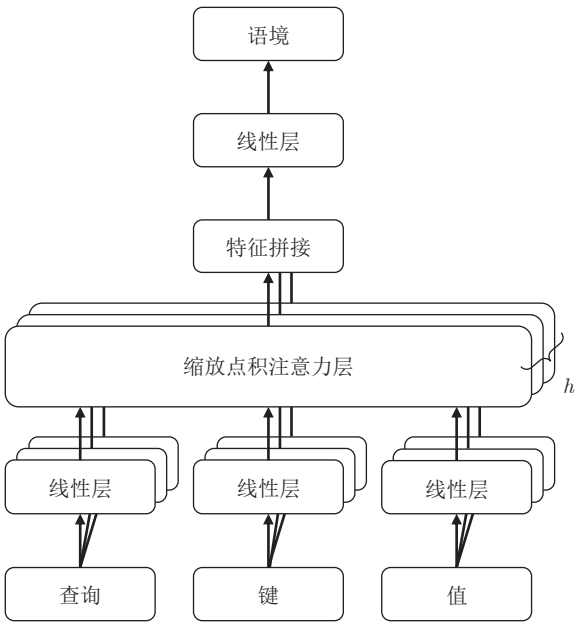


图 1 多头注意力机制的结构

Fig. 1 The structure of multi-head attention mechanism

其中, $W_i^Q \in \mathbf{R}^{d_q \times d_c}$, $W_i^K \in \mathbf{R}^{d_k \times d_c}$, $W_i^V \in \mathbf{R}^{d_v \times d_{c'}}$, $W^O \in \mathbf{R}^{hd_{c'} \times d_o}$ 是可学习的线性转换矩阵; d_c 是查询与键被投影后的特征维度; $d_{c'}$ 和 d_o 分别是子空间语境向量与最终语境向量的特征维度.

2.2 语境辅助的交叉注意力机制

在图像标题生成领域, 交叉注意力模块的查询向量依赖于输入的文本特征, 而键值对往往采用固定不变的图像区域特征. 因此, 传统的交叉注意力机制无法捕获先前时刻被关注过的语境特征, 缺乏视觉信息的连贯性. 针对这一问题, 本文面向 Transformer 解码框架提出了一种语境辅助的交叉注意力 CACA 机制. 如图 2(a) 所示, CACA 拓展了传统的交叉注意力机制, 通过历史语境记忆 HCM 为每一个解码时刻提供丰富的历史语境特征. 具体而言, 在第 t 时刻, CACA 以当前的语义查询 $q_t \in \mathbf{R}^{d_v}$ 与键值对 $K, V \in \mathbf{R}^{n \times d_v}$ 作为输入, 利用交叉注意力模块与残差连接得到当前时刻的临时语境向量 c_t^{tmp}. 需要说明的是, Transformer解码器中的交叉注意力模块采用的是多头注意力机制.

$$\begin{cases} q_t^{ln} = LayerNorm(q_t) \\ c_t^{tmp} = q_t + f_{mhatt}(q_t^{ln}, K, V) \end{cases} \tag{2}$$

其中, $LayerNorm$ 是层归一化操作, $q_t^{ln}, c_t^{tmp} \in \mathbf{R}^{d_v}$. 随后, 临时语境向量 c_t^{tmp} 会被加入历史语境记忆 HCM 当中, 构建当前时刻完整的历史语境特征 c_t^{his}:

$$\begin{cases} c_{-1}^{his} = \emptyset \\ c_t^{his} = [c_{t-1}^{his}; c_t^{tmp}]^S \end{cases} \tag{3}$$

其中, $\emptyset$ 表示空集, $[\cdot;\cdot]^S$ 表示空间维度的特征拼接.

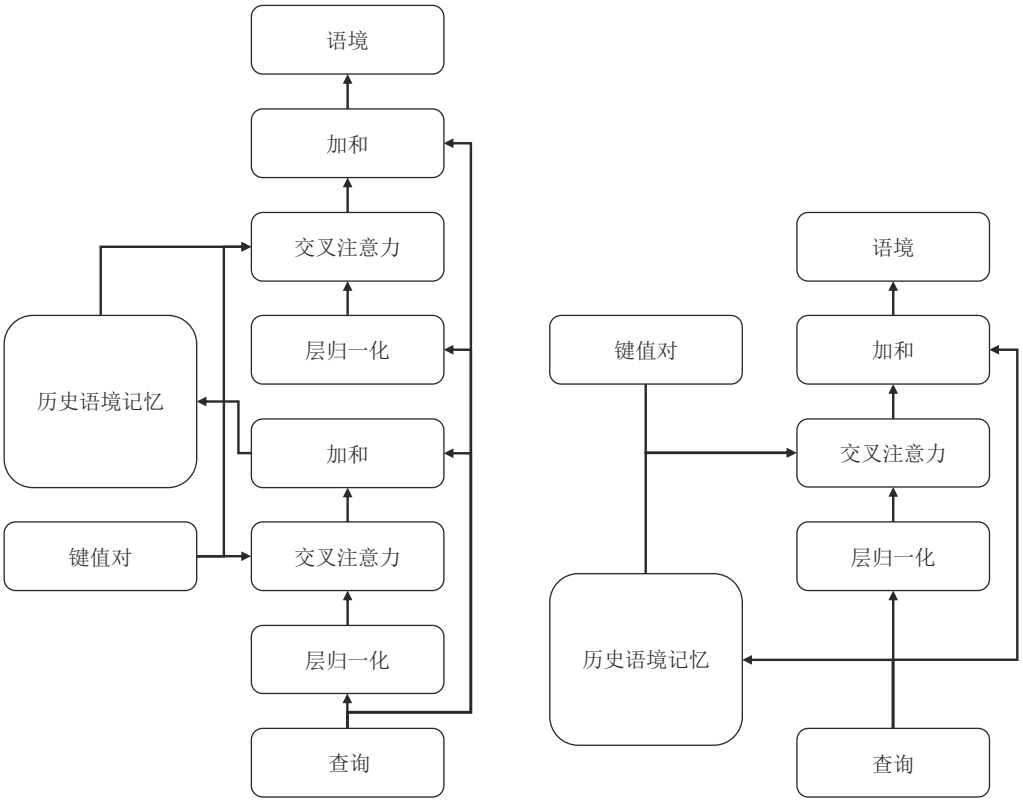
接下来, CACA 模块会将当前时刻的查询 q_t, 键值对 K, V, 以及完整历史语境 c_t^{his} 作为输入, 再一次应用交叉注意力模块, 获取当前时刻的最终语境特征 c_t:

$$\begin{cases} K' = [K; c_t^{his}]^S,\ V' = [V; c_t^{his}]^S \\ c_t = q_t + f_{mhatt}(q_t^{ln}, K', V') \end{cases} \tag{4}$$

值得一提的是, CACA 中两次使用的层归一化和多头交叉注意力机制分别共享相同的模型参数. 综上所述, 相较于 Transformer 解码器中传统的交叉注意力模块, CACA 在不添加任何参数的条件下, 引入了视觉信息的连贯性, 建模了每一时刻语义查询 q_t 与键值对 K, V 之间的关系, 得到了该时刻的最终语境特征 c_t:

$$c_t = f_{caca}(q_t, K, V) \tag{5}$$

其中, f_{caca} 表示语境辅助的交叉注意力机制.



(a) 语境辅助的交叉注意力机制
(a) Context-assisted cross attention mechanism

(b) 轻量级语境辅助的交叉注意力机制
(b) Light context-assisted cross attention mechanism

图 2　语境辅助的交叉注意力机制与其轻量级的模型结构

Fig. 2　Context-assisted cross attention mechanism and its light model structure

2.3　轻量级语境辅助的交叉注意力机制

语境辅助的交叉注意力机制通过历史语境记忆模块为每一个解码时刻提供了完整的历史语境特征, 向注意力模块中引入了视觉信息的连贯性. 然而, 两次使用交叉注意力机制大幅提高了模型推理的时间成本, 降低了模型的解码效率. 因此, 本文提出了一种轻量级的语境辅助的交叉注意力 (Light context-assisted cross attention, LightCACA) 模型, 在保证视觉连贯性的前提下, 以牺牲部分历史语境信息为代价, 换取与传统的交叉注意力机制接近的解码效率.

如图 2(b) 所示, LightCACA 首先将当前时刻的查询向量 q_t 加入到历史语境记忆当中, 构建当前完整的历史语境特征 c_t^{his}:

$$\begin{cases} c_{-1}^{his} = \emptyset \\ c_t^{his} = [c_{t-1}^{his}; q_t]^S \end{cases} \tag{6}$$

随后, 键值对 K, V 和完整历史语境特征 c_t^{his} 在空间维度上拼接, 供交叉注意力模块提取当前时刻的语境特征:

$$\begin{cases} K' = [K; c_t^{his}]^S,\ V' = [V; c_t^{his}]^S \\ q_t^{ln} = LayerNorm(q_t) \\ c_t = q_t + f_{mhatt}(q_t^{ln}, K', V') \end{cases} \tag{7}$$

轻量级语境辅助的交叉注意力机制与其标准模型的主要区别在于历史语境信息的不同. 在 Transformer 解码器的层级结构下, CACA 的历史语境信息由当前层的交叉注意力模块产生, HCM 存储的是当前层在每一时刻产生的临时语境特征, 而 LightCACA 的历史语境信息直接来源于当前层掩码自注意力模块的输出, 间接来源于上一层 LightCACA 产生的语境特征. 值得注意的是, 最底层 LightCACA 模块中 HCM 存储的历史语境信息来自解码器输入的文本序列特征.

2.4　语境辅助的转换器

图 3 展示了基于语境辅助转换器 (Context-assisted transformer, CAT) 的图像标题生成算法框架. 该框架主要包括三个部分: 提取图像对象级特征的 Faster R-CNN, 优化图像特征的 Transformer 编码器, 以及基于语境辅助的交叉注意力机制

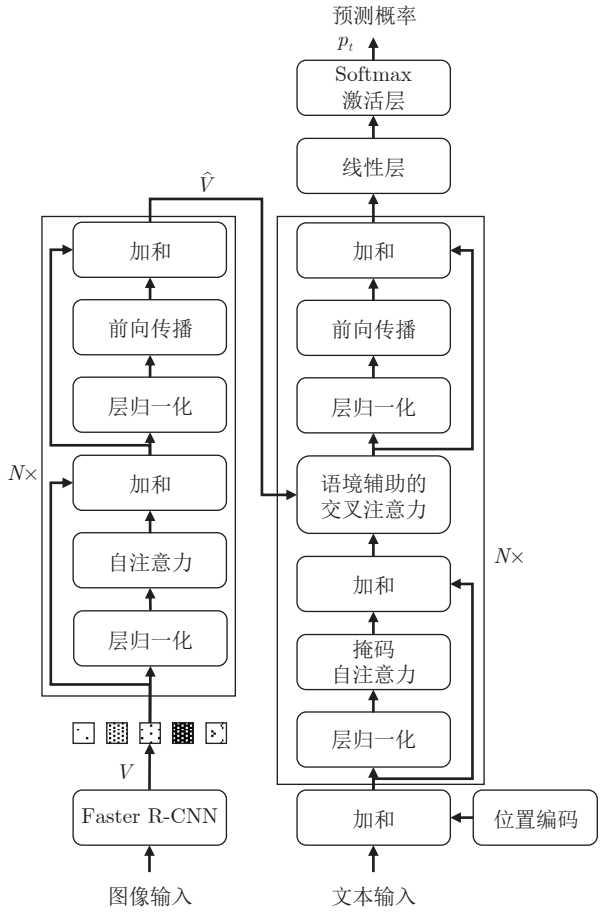


图 3 基于语境辅助转换器的图像标题生成模型

Fig. 3 Context-assisted transformer for image captioning

的 Transformer 解码器.

给定一幅图像 I, CAT 首先使用预训练好的 Faster R-CNN 从图像中提取出一组对象级别的视觉特征 $V=\{v_1, v_2, \cdots, v_m\}$, 其中, $v_i \in \mathbf{R}^{d_v}$, m 为从图像中提取到对象的数量. 需要说明的是, 在整个模型训练的过程中, Faster R-CNN 的参数固定.

$$V = f_{enc}(I) \tag{8}$$

其中, f_{enc} 表示 Faster R-CNN 特征提取模块.

随后, 视觉特征 V 将被输入 Transformer 编码器进行优化, 建立不同对象特征之间的语义关系. 值得一提的是, 本文认为不同对象之间不存在明显的位置顺序, 所以并未给视觉特征添加位置编码信息. 除此之外, 本文方法与传统 Transformer 编码器的算法流程一致. 第 n 层 Transformer 编码器 $f^n_{trm\text{-}enc}$ 的操作可总结如下:

$$V_{n+1} = f^n_{trm\text{-}enc}(V_n) \tag{9}$$

其中, 第一层 Transformer 编码器的输入向量 $V_1 = V$. 在此, 假设 Transformer 编码器共 N 层, 则其优化后的视觉特征 $\hat{V}$ 可由如下操作得到:

$$\hat{V} = f^N_{trm\text{-}enc}(V_N) \tag{10}$$

CAT 解码器使用 CACA 模块替换了传统的交叉注意力机制, 其余模型结构与 Transformer 解码器[7] 完全一致. 第 n 层 CAT 解码器 $f^n_{cat\text{-}dec}$ 在第 t 时刻的算法流程可总结如下:

$$y_{\leq t,\, n+1} = f^n_{cat\text{-}dec}(y_{\leq t,\, n}, \hat{V}) \tag{11}$$

本文假设 CAT 解码器的层数与编码器层数相同, 在第 t 时刻, 解码器生成单词 w_t 的过程可由如下公式表示:

$$\begin{cases} \hat{y_{\leq t}} = f^N_{cat\text{-}dec}(y_{\leq t,\, N}, \hat{V}) \\ w_t \sim p_t = Softmax(\hat{y_{\leq t}} W^p + b^p) \end{cases} \tag{12}$$

其中, $W^p \in \mathbf{R}^{d_v \times |\Sigma|}$, $b^p \in \mathbf{R}^{|\Sigma|}$, $|\Sigma|$ 是字典 Σ 中单词的数量.

轻量级语境辅助的转换器 (Light context-assisted transformer, LightCAT) 在模型设计的思路上与 CAT 完全相同, 区别仅在于使用 LightCACA 替换了 CAT 中的 CACA 模块.

2.5 模型优化

在图像标题的生成过程中, 传统的交叉注意力机制只能够关注到图像区域的特征, 而本文提出的语境辅助的注意力机制还可以回顾先前时刻已经关注过的历史语境. 为了约束每个 (Light)CACA 模块分配给图像特征与历史语境特征的权重总和, 产生一个更加有效的权重分布, 本文提出了一种名为 "自适应权重约束 (Adaptive weight constraint, AWC)"的正则化方法. 对于第 n 层 (Light)CAT 解码器中的 (Light)CACA 模块而言, 假设第 h 个头部在整个序列预测的过程中为历史语境特征分配了权重向量 $\alpha^{n,\,h} = \{\alpha^{n,\,h}_1, \cdots, \alpha^{n,\,h}_l\}$, 其中, $\alpha^{n,\,h}_t = \{\alpha^{n,\,h}_{t,\,1}, \cdots, \alpha^{n,\,h}_{t,\,t}\}$, $t \in [1, l]$, l 为生成标题的长度, $\alpha^{n,\,h}_{t,\,i} \in \mathbf{R}^1$, $i \in [1, t]$, 作为辅助损失函数, 则第 n 层第 h 个头部的 AWC 损失 $\mathcal{L}^{n,\,h}_{awc}(\theta)$ 可由如下公式计算得到:

$$\mathcal{L}^{n,\,h}_{awc}(\theta) = \frac{1}{l}\sum_{t=1}^{l}\left(\sum_{i=1}^{t}\alpha^{n,\,h}_{t,\,i} - \beta^{n,\,h} + \epsilon\right)^2 \tag{13}$$

其中, $\beta^{n,\,h}$ 是一个可学习的参数, ϵ 用于防止训练过程中的梯度爆炸. 本文设置 ϵ 为 1×10^{-8}.

本文遵循了以往研究工作中的训练策略. 给定真实标题序列 $y^{\star}_{1:T}$, 本文首先通过联合优化交叉熵 (Cross entropy, CE) 损失 $\mathcal{L}_{ce}(\theta)$ 与 AWC 损失 $\mathcal{L}_{awc}(\theta)$ 来训练 (Light)CAT 模型:

$$\mathcal{L}_{dl}(\theta)=\mathcal{L}_{ce}(\theta)+\gamma\mathcal{L}_{awc}(\theta)=$$
$$-\sum_{t=1}^{T}\ln(p_\theta(y_t^\star|y_{1:(t-1)}^\star,I))+$$
$$\frac{\gamma}{NH}\sum_{j=1}^{N}\sum_{k=1}^{H}\mathcal{L}_{awc}^{j,k}(\theta)\tag{14}$$

其中, γ 是两项损失的平衡因子, H 是多头注意力模块的头部数量, 本文依据经验将其设置为 0.5, N 为 (Light)CAT 解码器的层数.

随后, 本文在强化学习阶段采用自我批判序列训练 (Self-critical sequence training, SCST) 算法[31] 直接优化了不可微分的评价指标:

$$\mathcal{L}_{rl}(\theta)=-\mathrm{E}_{w_{1:l}\sim p_\theta}[r(w_{1:l})]\tag{15}$$

其中, $w_{1:l}$ 是生成的图像标题, 本文中的奖励 $r(\cdot)$ 采用了流行的 CIDEr-D[32] 分数.

3　实验与分析

3.1　数据集与评价标准

本文在 MS COCO (Microsoft common objects in context) 数据集[9] 上评估了 (Light)CAT 的性能. 该数据集共包含 123 287 幅图像, 每幅图像由不同的 AMT (Amazon mechanical turk) 工作人员用至少 5 条标题进行标注. 为了与其他先进的基线方法进行公平的比较, 本文采用了“Karpathy”分割[33] 进行离线评估, 其中, 113 287 幅图像用于训练, 5 000 幅用于验证, 另外 5 000 幅用于测试. 本文使用的评价方法包括 BLEU[34], METEOR[35], ROUGE-L[36], CIDEr-D[32], 以及 SPICE[37].

3.2　实现细节

本文采用在视觉基因组 (Visual genome) 数据集[38] 上预训练好的 Faster R-CNN 作为图像特征提取器, 该编码器为每一幅图像检测出 10 ~ 100 个不同区域, 每个区域特征向量的维数为 2 048, 随后将它们投影到 512 维后输入到 Transformer 编码器当中进行特征优化. 对于 Transformer 编码器与 (Light)CAT 解码器而言, 本文参照了之前的研究工作[25], 将二者的层数设定为 3, 多头注意力机制的头数为 8, 每个模块输出的向量维度为 512, 每一个注意力网络和前向网络都采用了 Dropout 方法, 丢失率为 0.1. 在训练过程中, 本文首先采用联合优化交叉熵损失和自适应权重约束损失的方式训练模型, 其中包括了 10 000 次热身 (Warm-up) 训练. 之后, 在优化 CIDEr-D 分数时, 本文采用了固定的学习率 5×10^{-6}, 当 CIDEr-D 分数在连续五轮训练中均未出现提升时, 终止训练过程. 在两个训练阶段, 本文都将批量大小设置为 50, 集束搜索的大小设置为 5.

3.3　语境辅助交叉注意力机制的性能分析

为了验证语境辅助的交叉注意力机制在 Transformer 解码框架中的有效性和通用性, 本文采用 Transformer, $\mathcal{M}^2$ Transformer[25], DLCT[27], $\mathcal{S}^2$ Transformer[28], DIFNet[29] 作为基线模型, 在 MS COCO 数据集上设计了 5 组对比实验. 每一组实验均使用 CACA 模块与 LightCACA 模块替换了基线模型中的传统的交叉注意力机制, 除 (Light)CACA 模块外, 改进模型与原模型在结构上完全一致. 同时, 改进模型在训练过程中加入了自适应权重约束, 来寻求一个更具泛化性的交叉注意力权重分布. 如表 1 所示, 采用 (Light)CACA 模块改进后的模型在绝大多数评价指标中都超越了基线模型的性能. 值得一提的是, 在与当前最先进的 $\mathcal{S}^2$ Transformer 和 DIFNet 模型的比较中, 采用标准 CACA 模块的改进模型实现了对基线方法的全面超越, 在 BLEU 与 CIDEr-D 分数上均取得了明显的提升. 同时, 标准 CACA 模块给模型带来的性能提升比 LightCACA 模块更加明显. 举例而言, 以 Transformer 为基线模型, LightCAT 模型在 BLEU-4 和 CIDEr-D 分数上较 Transformer 分别提升了 1.1% 和 1.0%, 而 CAT 模型带来的提升为 2.4% 和 2.5%. 该结果从定量分析的角度有力地证明了当前层交叉注意力语境特征对解码过程的实用价值.

正如上文所提到的, 本文设计的 CACA 模块与自适应权重约束同样适用于基于 LSTM 的解码框架. 在此, 本文以 Att2in[31], BUTD[16], LB[10] 作为基线模型, 在 MS COCO 数据集上设计了 3 组对比实验. 由于这些基线模型的解码器中只存在一个交叉注意力模块, 所以自适应权重约束中的参数 $N=1$. 表 2 是上述三种基于 LSTM 的图像标题生成模型结合 CACA 模块后在 MS COCO 数据集上的性能表现. 实验结果表明, 本文提出的 CACA 模块不仅适用于 Transformer 解码框架, 还可以大幅提升 LSTM 解码模型的性能.

为了分析语境辅助的交叉注意力机制对模型推理效率的影响, 本文从 MS COCO 测试集中随机选出了 1 000 幅图像, 分别使用 Transformer, CAT 和 LightCAT 模型生成图像标题. 具体而言, 每一轮解码过程的输入为 50 幅图像, 集束搜索算法的束大小为 5. 本组实验在单块 NVIDIA TITAN XP GPU 环境下进行, CUDA 版本为 10.1. 表 3 记录

表 1 基于 Transformer 的图像标题生成模型结合 (轻量级) 语境辅助的交叉注意力机制在 MS COCO 数据集上的性能表现 (%)

Table 1 Performance of Transformer-based image captioning models combined with (Light)CACA on MS COCO dataset (%)

模型名称	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr-D	SPICE
Transformer	80.0	38.0	28.5	57.9	126.5	22.4
Transformer + CACA (CAT)	**80.8**	**38.9**	**28.9**	**58.6**	**129.6**	**22.6**
Transformer + LightCACA (LightCAT)	80.6	38.4	28.6	58.2	127.8	22.5
$\mathcal{M}^2$ Transformer[25]	80.8	39.1	29.2	58.6	131.2	22.6
$\mathcal{M}^2$ Transformer + CACA	**81.2**	**39.4**	**29.5**	**59.0**	**132.4**	**22.8**
$\mathcal{M}^2$ Transformer + LightCACA	**81.2**	39.3	29.4	58.8	131.9	**22.8**
DLCT[27]	81.4	39.8	29.5	59.1	133.8	23.0
DLCT + CACA	**81.6**	**40.2**	**29.6**	**59.2**	**134.3**	**23.2**
DLCT + LightCACA	81.4	40.0	29.5	**59.2**	134.1	23.0
$\mathcal{S}^2$ Transformer[28]	81.1	39.6	29.6	59.1	133.5	23.2
$\mathcal{S}^2$ Transformer + CACA	**81.5**	**40.0**	**29.7**	**59.3**	**134.2**	**23.3**
$\mathcal{S}^2$ Transformer + LightCACA	81.3	39.7	29.6	**59.3**	133.8	**23.3**
DIFNet[29]	81.7	40.0	29.7	59.4	136.2	23.2
DIFNet + CACA	**82.0**	**40.5**	**29.9**	**59.7**	**136.8**	**23.4**
DIFNet + LightCACA	81.9	40.1	29.7	59.5	136.4	23.2

表 2 基于 LSTM 的图像标题生成模型结合语境辅助的交叉注意力机制在 MS COCO 数据集上的性能表现 (%)

Table 2 Performance of LSTM-based image captioning models combined with CACA on MS COCO dataset (%)

模型名称	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr-D	SPICE
Att2in[31]	—	33.3	26.3	55.3	111.4	—
Att2in + CACA	**77.8**	**36.7**	**27.5**	**57.1**	**119.7**	**21.0**
BUTD[16]	79.8	36.3	27.7	56.9	120.1	21.4
BUTD + CACA	**80.4**	**38.1**	**28.3**	**58.2**	**126.4**	**22.1**
LB[10]	79.6	37.7	28.4	58.1	124.4	21.8
LB + CACA	**80.8**	**38.6**	**28.6**	**58.6**	**128.1**	**22.3**

表 3 语境辅助的交叉注意力机制对 Transformer 推理效率的影响 (ms)

Table 3 The effect of context-assisted cross attention mechanism on Transformer′s reasoning efficiency (ms)

模型名称	单轮贪心解码时间	单轮集束搜索解码时间
Transformer	4.7	63.9
CAT	6.1	86.6
LightCAT	4.9	68.1

了 3 种模型对每一轮输入图像的平均解码时间. 尽管语境辅助的交叉注意力机制大幅提高了图像标题的质量, 但由于两次使用交叉注意力模块, 不可避免地导致了解码效率的下降, 在贪心和集束搜索算法下, 使模型的解码时间分别上升 29.8% 和 35.5%. 对于轻量级的语境辅助的交叉注意力机制而言, 其模型结构与传统的交叉注意力模块相似, 仅通过扩充数据信息的方式引入视觉连贯性, 所以, Light-CACA 可以在保证解码效率的同时提升模型的性能. 虽然 CACA 模块的结构较为复杂, 需要更长的解码时间, 但总体来讲, 它为模型带来的性能提升更加明显, 且解码效率仍在可接受的范围之内, 所以, 本文中的大部分实验均以 CACA 模块为代表, 体现本文算法的优势.

3.4 语境辅助转换器与先进基线方法的比较

本文将基于不同基线模型的语境辅助转换器与当前先进的基线方法在 MS COCO 数据集上进行了比较. 这些基线方法包括: 1) Att2in 与 Att2all[31], 使用视觉注意力机制, 并采用不可微分的评价指标对模型进行优化; 2) BUTD[16], 使用 Faster R-CNN 提取图像特征, 再采用自顶向下的解码器对视觉特征进行解码; 3) AoANet[18], 使用注意力门从被关注的语境特征中筛选与语义查询切实相关的知识; 4) $\mathcal{M}^2$ Transformer[25], 通过网状连接的编解码

框架充分利用低层与高层的视觉特征; 5) X-LAN 与 X-Transformer[19], 使用空间与管道双线性注意力机制来建模不同模态间的二阶相互作用; 6) DLCT[27], 通过图像区域特征与网格特征的协作互补, 增强视觉信息的表达能力; 7) RSTNet[26], 建立了一个基于 BERT 的语言模型来捕获文本上下文信息, 并通过自适应注意力模块来衡量视觉与文本线索的贡献; 8) CATT[20], 使用前门调整策略来消除视觉−语言模型中难以捕捉的混淆效应; 9) $\mathcal{S}^2$ Transformer[28], 采用空间和尺度感知的 Transformer 将图像网格特征高效地融入图像标题生成模型; 10) DIFNet[29], 将图像的全景分割特征作为网格特征之外的另一个视觉信息源, 以增强视觉信息对图像标题生成的贡献; 11) CIIC[39], 通过后门调整策略缓解由无法观测的混淆因素引起的虚假相关性. 与当前先进方法的对比结果如表 4 所示. 本文的 DIFNet + CACA 模型在全部评价指标上都取得了当前最优的效果, 其中, 在 BLEU-4 和 CIDEr-D 上分别达到了 40.5 与 136.8.

3.5　语境辅助交叉注意力机制的消融实验

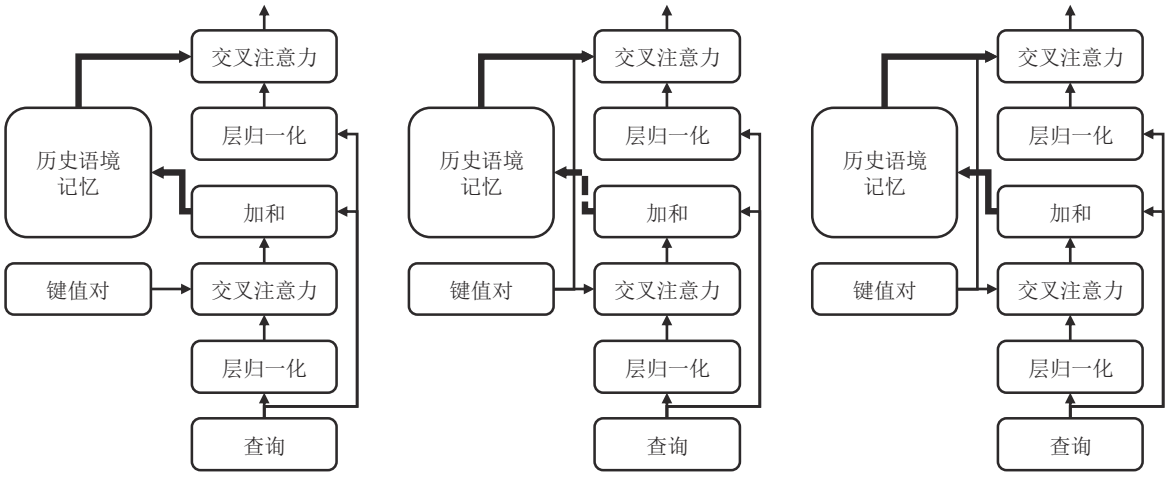
为了更加清晰地说明语境辅助的交叉注意力机制的设计思路, 分析它为基线模型带来的性能提升, 本文以经典的 Transformer 解码框架为基础, 使用三种不同的语境辅助策略增强解码器中传统的交叉注意力 (Traditional cross attention, TCA) 模块, 在 MS COCO 数据集上进行了对比实验. 具体而言, 不同语境辅助策略的主要区别在于历史语境特征的引入形式不同. 如图 4 所示, 左侧的 CACA 模块在引入历史语境特征时, 并未与视觉特征相结合, 而是仅将历史语境记忆中的特征向量作为键值对 (Only historical contexts, OHC), 通过二次使用交叉注意力模块, 提取当前时刻的语境特征; 中间的 CACA 模块将之前时刻的历史语境特征与视觉特征相拼接, 构建交叉注意力模块的键值对输入, 此处的历史语境特征不包括当前时刻首次使用交叉注意力模块时产生的临时语境特征 (Incomplete historical contexts, IHC); 右侧的 CACA 模块则是本文在第 2.2 节中提到的方法, 它为交叉注意力模块同时提供了完整的历史语境特征 (Complete historical contexts, CHC) 与视觉信息. 为公平起见, 本组对比实验中均未加入自适应权重约束.

表 5 列出了在 Transformer 解码框架下, 传统交叉注意力机制在结合三种不同语境辅助策略时的性能表现. 从实验结果中可以看出, TCA + OHC 与传统方法相比, 在多数评价指标中分数均有所下

表 4　本文模型与先进方法在 MS COCO 数据集上的性能对比 (%)

Table 4　Performance comparison between our models and the state-of-the-art (%)

模型名称	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr-D	SPICE
Att2in[31]	—	33.3	26.3	55.3	111.4	—
Att2all[31]	—	34.2	26.7	55.7	114.0	—
BUTD[16]	79.8	36.3	27.7	56.9	120.1	21.4
AoANet[18]	80.2	38.9	29.2	58.8	129.8	22.4
$\mathcal{M}^2$ Transformer[25]	80.8	39.1	29.2	58.6	131.2	22.6
X-LAN[19]	80.8	39.5	29.5	59.2	132.0	**23.4**
X-Transformer[19]	80.9	39.7	29.5	59.1	132.8	**23.4**
DLCT[27]	81.4	39.8	29.5	59.1	133.8	23.0
RSTNet (ResNext101)[26]	81.1	39.3	29.4	58.8	133.3	23.0
BUTD + CATT[20]	—	38.6	28.5	58.6	128.3	21.9
Transformer + CATT[20]	—	39.4	29.3	58.9	131.7	22.8
$\mathcal{S}^2$ Transformer[28]	81.1	39.6	29.6	59.1	133.5	23.2
DIFNet[29]	81.7	40.0	29.7	59.4	136.2	23.2
$\text{CIIC}_{\mathcal{O}}$ [39]	81.4	40.2	29.3	59.2	132.6	23.2
$\text{CIIC}_{\mathcal{G}}$ [39]	81.7	40.2	29.5	59.4	133.1	23.2
Transformer + CACA (CAT)	80.8	38.9	28.9	58.6	129.6	22.6
$\mathcal{M}^2$ Transformer + CACA	81.2	39.4	29.5	59.0	132.4	22.8
DLCT + CACA	81.6	40.2	29.6	59.2	134.3	23.2
$\mathcal{S}^2$ Transformer + CACA	81.5	40.0	29.7	59.3	134.2	23.3
DIFNet + CACA	**82.0**	**40.5**	**29.9**	**59.7**	**136.8**	**23.4**



(a) TCA + OHC (b) TCA + IHC (c) TCA + CHC

图 4 传统交叉注意力机制的三种语境辅助策略

Fig. 4 Three context-assisted strategies of traditional cross attention

表 5 传统交叉注意力机制结合不同语境辅助策略在 MS COCO 数据集上的表现 (%)

Table 5 Performance of the traditional cross attention mechanism combined with different context-assisted strategies on MS COCO dataset (%)

模型名称	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr-D	SPICE
TCA (base)	80.0	38.0	28.5	57.9	126.5	22.4
TCA + OHC	80.4	37.8	28.2	57.4	126.8	21.8
TCA + IHC	80.8	38.2	28.5	58.1	128.2	22.2
TCA + CHC (CACA)	**81.2**	**38.6**	**28.6**	**58.2**	**128.9**	**22.6**

降, 导致此结果的原因是, 交叉注意力模块在生成最终语境特征时缺少了原始视觉特征的参与和指导, 同时, 每一时刻历史语境记忆能够为注意力模块提供的特征向量十分有限, 严重限制了注意力模块的选择能力. TCA + IHC 相较于传统方法, 在大多数评价指标上均有所提升, 说明历史语境特征的加入丰富了交叉注意力模块的选择空间, 为当前语境特征的生成提供了更加丰富且有效的信息, 也从侧面反映出视觉连贯性在序列预测任务当中的重要性. TCA + CHC 是本文提出的 CACA 模型, 与传统交叉注意力机制相比, 该方法在所有的评价指标上均取得了明显的提升. 同时, 从 TCA + CHC 与 TCA + IHC 的性能对比中可以得出结论, 临时语境特征的加入有助于每一个 CACA 模块产生更高质量的最终语境特征, 进而指导语言模型生成更加合理的图像标题.

本文在 CACA 模块上的设计理念是, 在不添加任何额外的可训练模型参数的条件下, 通过 CACA 模块引入视觉信息的连贯性, 提升基线模型的性能. 具体来讲, 在 CACA 中两次使用的交叉注意力模块共享 (Shared) 相同的参数. 为了分析在不共享 (Not shared) 模型参数的条件下 CACA 模块的性能表现, 本文在 MS COCO 数据集上以不同解码器层数的 CAT 模型为基础进行了对比实验. 在本组实验中, 不同 CAT 模型的编码器层数固定为 3 层, 且在训练过程中同样未加入自适应权重约束.

表 6 展示了三组不同解码器层数的 CAT 模型在共享与不共享交叉注意力模块参数时的性能表现. 当解码器层数为 2 层时, 从实验结果中可以看出, 无论是否共享交叉注意力模块的参数, 使用 CACA 模块的 CAT 模型的性能在所有评价指标上都超越了使用 TCA 模块的模型的性能. 进一步对 CACA 模块进行分析, 与共享参数的 CACA 模型相比, 不共享参数的模型拥有更多的可训练参数, 且模型性能明显优于共享参数的模型. 3 层解码器的模型实验反映出了相似的实验结论, 不同的是, 相较于共享参数的 CACA 模型, 不共享参数的模型性能提升较小. 同时, 在 4 层解码器的模型实验中, TCA 模型的性能较 3 层解码器的 TCA 模型有所降低, 且 CACA 模块对基线模型的性能产生了负面影响. 综合表 6 中的实验结果及上述分析, 本文得出了以下两点结论: 1) 当基于 TCA 的模型尚未出现过拟

合现象时, 共享参数的 CACA 模块能够有效提升基线模型的性能, 而不共享参数的 CACA 模块在提升模型性能的同时, 由于加入了更多的参数, 模型可能出现过拟合问题; 2) 当基于 TCA 的模型已经出现过拟合现象时, CACA 模块将扩大过拟合产生的负面影响, 尤其是不共享参数的 CACA 模块, 将大幅降低图像标题的质量.

3.6　自适应权重约束的消融实验

本文在 MS COCO 数据集上设计了一组消融实验来解释自适应权重约束给 CAT 模型带来的性能提升. 通过观察 AWC 损失与 CE 损失的数量级, 本文依据经验将损失权衡系数 γ 设置为 0.5. 在本组实验中, CAT 解码器的层数为 3 层. 从表 7 列出的实验结果中可以看出, 当 CAT 模型采用固定值作为 CACA 模块的权重约束时, 其性能表现随 β 值的增大, 先缓慢提升, 在 $\beta = 0.5$ 附近达到最优, 随后迅速下降. 结合表 5 中的信息, 本文发现, 当固定权重约束 $\beta = 0.1$ 时, 即在少量引入历史语境特征的条件下, CAT 模型的性能就可在仅使用 TCA 的基础上实现大幅提升, 模型的 CIDEr-D 分数由 126.5 提升至 127.8. 同时, 当固定权重约束 $\beta = 0.9$ 时, 即几乎将全部的权重都分配给历史语境特征时, CAT 模型的性能将偏向表 5 中 TCA + OHC 的实验结果, 过度关注历史语境信息而忽略原始的视觉信息, 导致图像标题的质量严重下降. 当固定权重约束 $\beta = 0.5$ 时, 模型在视觉特征与历史语境特征上的权重分配相对平衡, 一定程度上提升了 CAT 模型的性能. 与固定权重约束相比, 自适应权重约束更加灵活, 它能够依据数据和模型的需要, 学习到一组更具泛化性的参数. 从实验结果上看, 自适应权重约束为 CAT 模型带来的提升要明显优于固定权重约束, 同时, 与无权重约束的模型相比, 采用 AWC 的 CAT 模型在所有评价指标中均超越了基线模型.

3.7　注意力图的可视化分析

为了深入阐释历史语境记忆的重要作用以及自适应权重约束的有效性, 本文基于一组完整的图像标题生成示例, 对视觉特征和历史语境记忆上的注意力分布进行了可视化分析. 考虑到顶层解码器的输出特征与图像标题的生成结果直接相关, 本文以 Transformer 模型顶层解码器中的 CACA 模块为例展开讨论.

如图 5 所示, 中间部分展示了原始图像, 以及

表 6　不同解码器层数的 CAT 模型在共享与不共享交叉注意力模块参数时的性能表现 (%)
Table 6　Performance of CAT models with different decoder layers when sharing or not sharing parameters of the cross attention module (%)

解码器层数	交叉注意力模块设置	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr-D	SPICE
$N = 2$	TCA	78.8	37.4	28.0	57.4	125.4	21.8
$N = 2$	CACA (Shared)	80.4	38.0	28.2	57.8	128.0	22.3
$N = 2$	CACA (Not shared)	**80.8**	**38.4**	**28.5**	**58.2**	**128.8**	**22.5**
$N = 3$	TCA	80.0	38.0	28.5	57.9	126.5	22.4
$N = 3$	CACA (Shared)	**81.2**	38.6	28.6	58.2	128.9	22.6
$N = 3$	CACA (Not shared)	81.0	**38.8**	**28.8**	**58.3**	**129.3**	**22.7**
$N = 4$	TCA	79.6	**37.8**	**28.5**	**57.8**	**126.2**	**22.2**
$N = 4$	CACA (Shared)	**79.8**	37.5	28.4	57.6	125.8	21.9
$N = 4$	CACA (Not shared)	79.0	36.8	28.1	57.1	124.3	21.5

表 7　采用自适应权重约束的 CAT 模型在 MS COCO 数据集上的表现 (%)
Table 7　Performance of the CAT model with adaptive weight constraint on MS COCO dataset (%)

权重约束方式	BLEU-4	METEOR	ROUGE-L	CIDEr-D
无权重约束	38.6	28.6	58.2	128.9
固定权重约束 $\beta = 0.1$	38.4	28.4	58.1	127.8
固定权重约束 $\beta = 0.3$	38.7	28.6	58.3	128.7
固定权重约束 $\beta = 0.5$	**38.9**	28.7	58.4	129.3
固定权重约束 $\beta = 0.7$	38.5	28.4	58.1	128.4
固定权重约束 $\beta = 0.9$	38.1	28.2	57.6	127.2
自适应权重约束	**38.9**	**28.9**	**58.6**	**129.6**

采用 AWC 优化的 CACA 模块在每个解码时刻分配给图像特征的注意力权重分布图. 图 5 顶部的折线图展示了 CACA 模块在对应时刻为历史语境记忆分配的注意力权重总和. 其中, 橙黄色实线与金黄色虚线分别代表了“采用”与“未采用”AWC 优化的 CACA 模块给历史语境记忆的权重分配结果. 在此, 本文首先通过橙黄色的实验数据深入分析历史语境记忆存在的重要意义. 在第一个解码时刻, 采用 AWC 优化的 CACA 模块将大部分注意力给予了图像特征, 仅为历史语境记忆分配了 0.0732 的注意力权重. 直观分析, 在序列生成的初始时刻, 解码器亟待充分理解图像中的显著特征, 同时, 历史语境记忆能够为解码过程提供的语义信息十分有限, 因此, CACA 模块主要依靠图像特征完成第一个时间步的单词预测. 在后续的时刻中, 随着历史语境记忆中的特征向量逐渐丰富, CACA 模块为其分配的注意力权重也迅速增加, 并最终稳定在 0.2 左右. 在图像标题的生成过程中, 解码器不断寻求历史语境记忆的指导, 说明历史语境记忆蕴含了大量有价值的信息, 进一步证实了该模块存在的必要性.

与此同时, 通过比较两条折线中数据点的大小, 本文发现, 未采用 AWC 优化的 CACA 模型对历史语境记忆的利用率远不及采用 AWC 优化的 CACA 模型. 结合前文的结论, 若不采用 AWC 对模型进行优化, CACA 模块则难以充分利用历史语境记忆中的有效信息为解码过程提供丰富的语义特征. 综上所述, 自适应权重能够提升 CACA 模块对历史语境记忆的利用率, 为解码器提供更多有价值的信息, 从而提高图像标题的生成质量.

在图 5 的底部, 本文对注意力权重在历史语境记忆中的具体分配情况进行了可视化分析. 为了清晰起见, 本文挑选了三个具有代表性的时间步进行讨论. 具体而言, 当历史语境记忆中的一条特征向量获得大于 0.05 的注意力权重时, 则通过一条连线指向当前时刻生成的单词. 此处展示的图像标题为采用 AWC 优化的模型生成的结果. 值得一提的是, 连线的颜色越深, 表示特征被分配的权重越大. 如图 5 所示, 当模型预测单词“man (男人)”和“holding (拿着)”时, CACA 对当前时刻新加入历史语境记忆的特征向量格外关注, 表明视觉特征在此刻发挥着重要作用; 而当模型预测单词“on (在…之上)”时, 由于图像中缺少明显的视觉线索表达这一概念, 因此, CACA 重点关注了历史语境记忆中可以辅助推断当前词的语义特征. 上述事实说明, 历史语境记忆可以发挥视觉哨兵[16] 的作用, 为 CACA 模块提供一个回退选项, 在必要时舍弃部分低价值的视觉特征, 利用之前时刻的历史语境特征, 协助解码器完成单词的预测.

3.8 图像标题生成示例

为了进一步证明本文方法在传统的交叉注意力机制上的改进, 本文在图 6 中展示了八组图像标题生成的案例. 其中, 每组案例包括了一幅图像, Transformer 基线模型生成的标题, CAT 模型生成的标题, 以及图像对应的真实 (Ground truth, GT) 标题. 举例来讲, 在第一个案例中, Transformer 与 CAT 模型都关注到了图像中的主要目标“dog”与“frisbee”, 这得益于它们拥有相同的编码器结构 Faster R-CNN 与 Transformer 编码器, Faster R-CNN 能够提取到图像中的显著目标, Transformer 编码器则可以隐性地建模不同目标之间的关系. 然而, 由于缺少动作信息捕捉的相关模块, 这便要求解码器承担相应的职责. 从模型结构来看, Transformer 解码器通过传统的交叉注意力机制与图像

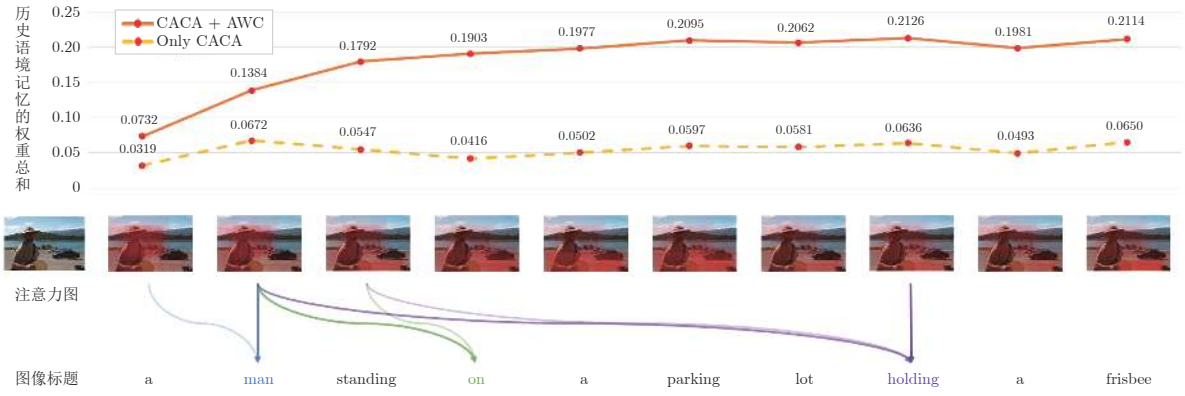


图 5 由语境辅助的交叉注意力模块分配给图像特征与历史语境记忆的注意力分布可视化

Fig. 5 Visualization of attention distribution assigned to both image features and historical context memory by our CACA module

特征进行交互, 认为图像中狗是叼着飞盘在沙滩上“奔跑 (running)”, 然而实际上, 图像中的狗是通过“跳跃 (jumping)”来接住空中的飞盘. 本文提出的 CAT 模型利用语境辅助的交叉注意力机制, 在解码过程中, 不仅能够关注到与当前语义查询最为相关的图像信息, 还能够从历史语境特征中受到启发. 在这一案例中, CAT 模型通过 CACA 模块, 进一步捕获到历史时刻与狗相关的语境特征, 从而生成了更加符合图像事实的描述“狗跳起接住 (jumping to catch) 飞盘”. 另外, 本文在图 6 中展示了一个失败的案例. 如案例八所示, 图中有一块砧板, 上面放着一块被刀切开的奶酪. 从两个模型生成的标题来看, 它们都错误地将奶酪 (cheese) 描述成“橘子 (orange)”. 导致这一结果的原因主要有两点: 1) 形如图中的奶酪在整个数据集中出现的次数较少, 深度模型难以捕捉其内在的判别特征; 2) 奶酪的颜色与生活中常见的橘子相似, 外加明亮的白光环境, 使得编码器提取到的特征难以将二者进行区分. 本文提出的 CACA 模块主要作用于模型的解码器部分, 对编码器的特征提取能力影响较小, 难以解决上述问题. 针对此类现象, 可以通过平衡数据分布、增强编码器、采用小样本学习[37] 等方式提升模型性能.

3.9　人工评价

在人工评价环节, 本文从 MS COCO 的测试集中随机选择了 500 幅图像, 使用 Transformer 模型与 CAT 模型为其生成图像标题. 为了提高评价的可信度, 本文将每组标题随机打乱, 并提供给 5 名评测人员, 由他们对标题的“相关性”和“一致性”分别进行比较和评价. 其中, 相关性的评价标准是图像与标题之间的相关程度, 而一致性代表了标题的流畅程度与语义一致性. 对于每一幅图像, 评测人员必须在上述两种评价指标上选出质量更高的一条标题, 当 2 名以上评测人员对某一条标题的相关性或一致性表示更加认可时, 本文则认定该条标题在对应指标上表现更好. 从表 8 中可以看出, 在图像与标题的相关性方面, Transformer 与 CAT 具备相近的生成能力. 然而, 本文提出的 CAT 模型生成的标题具有更强的一致性, 评价结果明显优于 Trans-

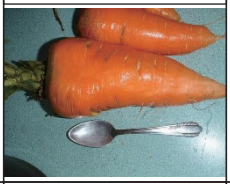
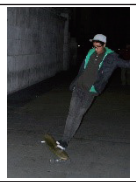
案例一 	Transformer: a dog running on the beach with a frisbee in its mouth CAT: a dog jumping to catch a frisbee on a beach GT: a dog jumping up into the air to catch a frisbee	案例二 	Transformer: a carrot and a carrot on a table CAT: a couple of carrots and a spoon on a table GT: a couple of carrots sit next to a spoon
案例三 	Transformer: a dog sticking its head out of a car window CAT: a dog sitting in the passenger seat of a car GT: an image of a dog sitting in the passenger seat of a car	案例四 	Transformer: a man holding a skateboard on a sidewalk CAT: a man doing a trick on a skateboard in the street GT: a man is doing a trick on a skateboard
案例五 	Transformer: a group of zebras grazing in a field CAT: a herd of zebras grazing in a field with a rainbow in the background GT: a herd of zebras grazing with a rainbow behind	案例六 	Transformer: a white toilet in a bathroom with a floor CAT: a white toilet in a bathroom with a black and white checkered floor GT: a white toilet in a bathroom with a checkers floor
案例七 	Transformer: a group of airplanes parked at an airport window CAT: a group of airplanes parked on the runway at an airport GT: a truck driving towards some planes parked on the runway	案例八 	Transformer: an orange and a knife on a cutting board CAT: an orange cut in half on a cutting board with a knife GT: a block of cheese on a cutting board with a knife in it

图 6　Transformer 与 CAT 生成的图像标题展示

Fig. 6　Image captions generated by the Transformer and the CAT

表 8 Transformer 与 CAT 模型的人工评价 (%)
Table 8 Human evaluation of Transformer and CAT (%)

模型名称	更强的相关性	更强的一致性
Transformer	8.8	7.4
CAT	10.2	12.4

former 模型, 这得益于 CACA 模块可以回顾历史语境特征的能力, 使语言模型在标题生成的过程中, 不断参考过去关注过的信息, 体现了视觉连贯性的优势.

4 结束语

本文面向图像标题生成任务, 针对传统的交叉注意力机制缺乏视觉连贯性的问题, 提出了一种语境辅助的交叉注意力 (CACA) 机制, 通过历史语境记忆为注意力模块提供先前关注过的语义信息, 为语言模型提供更加丰富的语境特征, 从而提升图像标题的生成质量. 为了限制每一个 CACA 模块分配给历史语境特征的权重总和, 本文设计了一种自适应权重约束 (AWC), 来提升模型的泛化能力. 本文将 CACA 模块与 AWC 方法集成到 Transformer 解码框架中, 构建了一种语境辅助的转换器 (CAT) 模型. 基于 MS COCO 数据集的实验结果表明, 与现有的多个基线模型相比, 本文提出的方法均取得了稳定的提升. 本文未来的研究工作将围绕历史语境特征在 Transformer 中的跨层交互展开探索.

References

1 Ji J, Luo Y, Sun X, Chen F, Luo G, Wu Y, et al. Improving image captioning by leveraging intra- and inter-layer global representation in Transformer network. In: Proceedings of the AAAI Conference on Artificial Intelligence. Virtual Conference: 2021. 1655−1663

2 Fang Z, Wang J, Hu X, Liang L, Gan Z, Wang L, et al. Injecting semantic concepts into end-to-end image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, Louisiana, USA: IEEE, 2022. 18009−18019

3 Tan J H, Tan Y H, Chan C S, Chuah J H. Acort: A compact object relation transformer for parameter efficient image captioning. *Neurocomputing*, 2022, **482**: 60−72

4 Fei Z. Attention-aligned Transformer for image captioning. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, British Columbia, Canada: 2022. 607−615

5 Stefanini M, Cornia M, Baraldi L, Cascianelli S, Fiameni G, Cucchiara R. From show to tell: A survey on deep learning-based image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, **45**(1): 539−559

6 Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE Transactions on Multimedia*, 2016, **39**(4): 652−663

7 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems. Long Beach, USA: 2017. 5998−6008

8 Cover T M, Thomas J A. *Elements of Information Theory*. New York: John Wiley & Sons, 2012.

9 Lin T Y, Maire M, Belongie S J, Hays J, Perona P, Ramanan D, et al. Microsoft coco: Common objects in context. In: Proceedings of European Conference on Computer Vision. Zurich, Switzerland: 2014. 740−755

10 Qin Y, Du J, Zhang Y, Lu H. Look back and predict forward in image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019. 8367−8375

11 Aneja J, Deshpande A, Schwing A G. Convolutional image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah, USA: IEEE, 2018. 5561−5570

12 Tang Peng-Jie, Wang Han-Li, Xu Kai-Sheng. Multi-objective layer-wise optimization and multi-level probability fusion for image description generation using LSTM. *Acta Automatica Sinica*, 2018, **44**(7): 1237−1249
(汤鹏杰, 王瀚漓, 许恺晟. LSTM 逐层多目标优化及多层概率融合的图像描述. 自动化学报, 2018, **44**(7): 1237−1249)

13 Xu K, Ba J, Kiros R, Cho K, Courville A C, Salakhutdinov R, et al. Show, attend and tell: Neural image caption generation with visual attention. In: Proceedings of the 32nd International Conference on Machine Learning. Lille, France: 2015. 2048−2057

14 You Q, Jin H, Wang Z, Fang C, Luo J. Image captioning with semantic attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Las Vegas, Nevada, USA: 2016. 4651−4659

15 Lu J, Xiong C, Parikh D, Socher R. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE, 2017. 3242−3250

16 Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, et al. Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah, USA: IEEE, 2018. 6077−6086

17 Chen S, Zhao Q. Boosted attention: Leveraging human attention for image captioning. In: Proceedings of European Conference on Computer Vision. Munich, Germany: 2018. 68−84

18 Huang L, Wang W, Chen J, Wei X. Attention on attention for image captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019. 4633−4642

19 Pan Y, Yao T, Li Y, Mei T. X-linear attention networks for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020. 10968−10977

20 Yang X, Zhang H, Qi G, Cai J. Causal attention for vision-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual Conference: 2021. 9847−9857

21 Wang Xin, Song Yong-Hong, Zhang Yuan-Lin. Salient feature extraction mechanism for image captioning. *Acta Automatica Sinica*, 2022, **48**(3): 735−746
(王鑫, 宋永红, 张元林. 基于显著性特征提取的图像描述算法. 自动化学报, 2022, **48**(3): 735−746)

22 Herdade S, Kappeler A, Boakye K, Soares J. Image captioning: transforming objects into words. In: Proceedings of Advances in Neural Information Processing Systems. Vancouver, Canada: 2019. 11135−11145

23 Li G, Zhu L, Liu P, Yang Y. Entangled transformer for image captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019. 8927−8936

24 Yu J, Li J, Yu Z, Huang Q. Multimodal transformer with multi-view visual representation for image captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019, **30**(12): 4467−4480

25 Cornia M, Stefanini M, Baraldi L, Cucchiara R. Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog-

nition. Seattle, USA: IEEE, 2020. 10575−10584

26 Zhang X, Sun X, Luo Y, Ji J, Zhou Y, Wu Y, et al. Rstnet: Captioning with adaptive attention on visual and non-visual words. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual Conference: 2021. 15465−15474

27 Luo Y, Ji J, Sun X, Cao L, Wu Y, Huang F, et al. Dual-level collaborative transformer for image captioning. In: Proceedings of the AAAI Conference on Artificial Intelligence. Virtual Conference: 2021. 2286−2293

28 Zeng P, Zhang H, Song J, Gao L. S^2 transformer for image captioning. In: Proceedings of the International Joint Conferences on Artificial Intelligence. Vienna, Austria: 2022.

29 Wu M, Zhang X, Sun X, Zhou Y, Chen C, Gu J, et al. Difnet: Boosting visual information flow for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, Louisiana, USA: 2022. 18020−18029

30 Lian Z, Li H, Wang R, Hu X. Enhanced soft attention mechanism with an inception-like module for image captioning. In: Proceedings of the 32nd International Conference on Tools With Artificial Intelligence. Virtual Conference: 2020. 748−752

31 Rennie S J, Marcheret E, Mroueh Y, Ross J, Goel V. Self-critical sequence training for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Hawaii, USA: 2017. 1179−1195

32 Vedantam R, Zitnick C L, Parikh D. Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Boston, USA: 2015. 4566−4575

33 Karpathy A, Li F F. Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Boston, USA: 2015. 3128−3137

34 Papineni K, Roukos S, Ward T, Zhu W J. Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA: 2002. 311−318

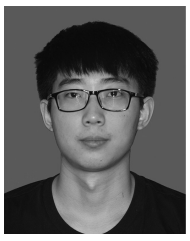
35 Denkowski M J, Lavie A. Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the 9th Workshop on Statistical Machine Translation. Baltimore, Maryland, USA: 2014. 376−380

36 Lin C Y. Rouge: A package for automatic evaluation of summaries. In: Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004. Barcelona, Spain: 2004. 74−81

37 Anderson P, Fernando B, Johnson M, Gould S. Spice: Semantic propositional image caption evaluation. In: Proceedings of European Conference on Computer Vision. Amsterdam, Netherlands: 2016. 382−398

38 Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 2017, **123**(1): 32−73

39 Liu B, Wang D, Yang X, Zhou Y, Yao R, Shao Z, et al. Show, deconfound and tell: Image captioning with causal inference. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, Louisiana, USA: 2022. 18041−18050



连 政 中国科学院软件研究所博士研究生. 2017 年获得西安电子科技大学学士学位. 主要研究方向为图像标题生成和自然语言处理.
E-mail: lianzheng2017@iscas.ac.cn
(**LIAN Zheng** Ph.D. candidate at the Institute of Software, Chinese Academy of Sciences. He received his bachelor degree from Xidian University in 2017. His research interest covers image captioning and natural language processing.)

王 瑞 中国科学院软件研究所高级工程师. 2012 年获得山东大学硕士学位. 主要研究方向为深度强化学习和多媒体技术.
E-mail: wangrui@iscas.ac.cn
(**WANG Rui** Senior engineer at the Institute of Software, Chinese Academy of Sciences. She received her master degree from Shandong University in 2012. Her research interest covers deep reinforcement learning and multimedia technology.)

李海昌 中国科学院软件研究所副教授. 2016 年获得中国科学院自动化研究所博士学位. 主要研究方向为计算机视觉和遥感技术.
E-mail: haichang@iscas.ac.cn
(**LI Hai-Chang** Associate professor at the Institute of Software, Chinese Academy of Sciences. He received his Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences in 2016. His research interest covers computer vision and remote sensing.)

姚 辉 中国科学院软件研究所网络工程师. 1997 年获得中国人民解放军装备指挥技术学院学士学位. 主要研究方向为智能信息处理和网络工程.
E-mail: iscasyh@sina.com
(**YAO Hui** Network engineer at the Institute of Software, Chinese Academy of Sciences. He received his bachelor degree from Equipment Command and Technology College of the Chinese People′s Liberation Army in 1997. His research interest covers intelligent information processing and network engineering.)

胡晓惠 中国科学院软件研究所教授. 2003 年获得北京航空航天大学博士学位. 主要研究方向为大数据分析和协同多智能体系统. 本文通信作者.
E-mail: hxh@iscas.ac.cn
(**HU Xiao-Hui** Professor at the Institute of Software, Chinese Academy of Sciences. He received his Ph.D. degree from Beihang University in 2003. His research interest covers big data analysis and cooperative multi-agent systems. Corresponding author of this paper.)