软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2022,33(9):3210−3222 [doi: 10.13328/j.cnki.jos.006623]

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

基于视觉关联与上下文双注意力的图像描述生成方法*

刘茂福[1], 施 琦[1], 聂礼强[2]

[1](武汉科技大学 计算机科学与技术学院, 湖北 武汉 430065)

[2](山东大学 计算机科学与技术学院, 山东 青岛 266237)

通信作者: 聂礼强, E-mail: nieliqiang@gmail.com

摘 要: 图像描述生成有着重要的理论意义与应用价值, 在计算机视觉与自然语言处理领域皆受到广泛关注. 基于注意力机制的图像描述生成方法, 在同一时刻融合当前词和视觉信息以生成目标词, 忽略了视觉连贯性及上下文信息, 导致生成描述与参考描述存在差异. 针对这一问题, 提出一种基于视觉关联与上下文双注意力机制的图像描述生成方法 (visual relevance and context dual attention, VRCDA). 视觉关联注意力在传统视觉注意力中增加前一时刻注意力向量以保证视觉连贯性, 上下文注意力从全局上下文中获取更完整的语义信息, 以充分利用上下文信息, 进而指导生成最终的图像描述文本. 在 MSCOCO 和 Flickr30k 两个标准数据集上进行了实验验证, 结果表明所提出的 VRCDA 方法能够有效地生成图像语义描述, 相比于主流的图像描述生成方法, 在各项评价指标上均取得了较高的提升.

关键词: 图像描述生成; 双注意力机制; 视觉关联注意力; 上下文注意力

中图法分类号: TP391

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

英文引用格式: Liu MF, Shi Q, Nie LQ. Image Captioning Based on Visual Relevance and Context Dual Attention. Ruan Jian Xue Bao/Journal of Software, 2022, 33(9): 3210–3222 (in Chinese). http://www.jos.org.cn/1000-9825/6623.htm

Image Captioning Based on Visual Relevance and Context Dual Attention

LIU Mao-Fu[1], SHI Qi[1], NIE Li-Qiang[2]

[1](School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China)

[2](School of Computer Science and Technology, Shandong University, Qingdao 266237, China)

Abstract: Image captioning is of great theoretical significance and application value, which has attracted wide attention in computer vision and natural language processing. The existing attention mechanism-based image captioning methods integrate the current word and visual cues at the same moment to generate the target word, but they neglect the visual relevance and contextual information, which results in a difference between the generated caption and the ground truth. To address this problem, this paper presents the visual relevance and context dual attention (VRCDA) method. The visual relevance attention incorporates the attention vector of the previous moment into the traditional visual attention to ensure visual relevance, and the context attention is used to obtain much complete semantic information from the global context for better use of the context. In this way, the final image caption is generated via visual relevance and context information. The experiments on the MSCOCO and Flickr30k benchmark datasets demonstrate that VRCDA can effectively describe the image semantics, and compared with several state-of-the-art methods of image captioning, VRCDA can yield superior performance in all evaluation metrics.

Key words: image captioning; dual attention mechanism; visual relevance attention; context attention

图像描述生成为一幅给定的图像自动生成一个合理通顺的自然语言描述, 是计算机视觉与自然语言处理交叉领域的一个研究热点和难点问题[1−3]. 随着社交网络平台的快速成长与移动摄影设备的低成本普及, 图像数量呈现

本文由“融合媒体环境下的媒体内容分析与信息服务技术”专题特约编辑汪萌教授、张勇东教授、俞俊教授以及张伟高级工程师推荐.
收稿时间: 2021-06-03; 修改时间: 2021-08-15, 2022-01-14; 采用时间: 2022-01-20; jos 在线出版时间: 2022-02-22

指数级增长. 根据新浪微博数据中心统计, 新浪微博日均图像发布量超过 1.2 亿. 然而, 无描述或错误描述的图像散落在网络中而无法被检索和利用, 造成大量资源浪费, 也为图像分类等语义理解带来巨大挑战[4]. 同时, 图像描述生成有着巨大实用价值, 例如, 社交媒体中图集或短视频自动配文、图像和文本跨媒体检索、视频摘要自动生成、辅助盲人图像理解等.

图像描述生成是一个非常艰难的任务, 面临着以下 3 个挑战: 1) 输入与输出是异质的媒体形式, 编码与解码的网络结构也是异构的, 存在跨模态语义鸿沟问题, 即单模态图像视觉特征在图像内容表达上存在多义性和不确定性; 2) 一图胜千言, 图像中的信息含量巨大, 拥有大量显式和隐式的视觉语义信息, 需要有针对性地找出最重要的对象和场景, 准确建立起视觉特征与生成文本之间的联系, 有侧重点地进行描述; 3) 生成的图像描述文本不但要符合自然语言的语法规则, 即形式上衔接良好, 而且要保证语义上的连贯性.

当前, 基于编解码框架的图像描述生成方法成为了主流, 该类方法通常使用卷积神经网络 (convolutional neural network, CNN) 为编码器提取图像视觉特征, 并采用循环神经网络 (recurrent neural network, RNN) 为解码器生成对应的描述文本. 但是, 这类方法在编码器和解码器之间仅通过一个固定大小的特征向量来连接, 并不能充分地表达图像数据与语义信息, 因此在解码阶段得不到足够有用的信息, 最终导致解码出的语句达不到理想效果. 为了解决这个问题, 借鉴注意力机制思想, 在传统编解码结构中添加注意力机制. 在基于注意力机制的图像描述生成方法中, CNN 将提取的图像特征按照图像的空间位置划分为一定数量的局部特征向量, 注意力机制根据 RNN 的隐藏状态, 动态选择与当前生成单词有关的图像局部特征向量, 来指导当前时刻单词的生成. 基于注意力机制的方法能够充分地利用图像的特征信息, 提升了生成描述的效果[5].

传统基于注意力机制的图像描述生成方法在预测目标词时, 仅融合当前时刻的单词和视觉信息, 这种处理方式既背离了人类视觉连贯性的习惯, 又忽略了上下文信息在生成目标词时的作用, 导致生成描述与参考描述存在差异. 显而易见, 当一个人描述一幅图像时, 视觉关注具有连贯性, 并且当前关注点会受到以前关注点的影响. 例如当使用“a black car”这 3 个单词表达同一个视觉对象时, 3 个单词在视觉上存在明显的连贯性; 又如图 1(a) 所示, 仅仅关注视觉上的“man”或“car”不足以得到“driving a car”这个文本短语, 只有同时关注到两者才能生成短语“driving a car”. 这说明在生成目标词时, 模型不仅要聚焦当前的视觉信息, 并且要关注以前的视觉信息, 这样才能保证视觉连贯性, 符合人类的习惯. 此外, 如图 1(b) 所示, 传统方法在预测目标词时仅依赖当前的视觉和文本信息, 故在生成“A man is”后会有极大的概率生成“standing”作为目标词, 但是如果考虑了视觉连贯性及上下文信息, 会有更大的概率生成“playing”这个更加准确且符合图像真实场景的目标词.

图 1 传统方法与 VRCDA 的对比

针对图像描述生成的 3 个挑战以及上述传统方法中存在的问题, 本文提出了一种基于视觉关联与上下文双注意力机制的图像描述生成方法 VRCDA. 该方法首先在传统视觉注意力中增加前一时刻的注意力向量, 提出了视觉关联注意力; 将视觉关联注意力同主流图像描述生成方法相结合得到对应的基础网络, 利用基础网络预先生成初始描述作为全局上下文信息; 使用上下文注意力处理全局上下文信息来获取更完整的语义信息, 指导生成图像语义描述文本.

本文的主要贡献包括以下 3 个方面.

(1) 本文提出了视觉关联注意力, 增加前一时刻的注意力向量有助于嵌入先前的视觉信息, 从而保证视觉连贯性并充分利用图像视觉信息.

(2) 本文使用上下文注意力从全局上下文信息中获取更完整的语义信息, 将上下文信息作为一种辅助信息去指导目标词的生成, 使模型具有更强的全局建模能力.

(3) 本文所提出的 VRCDA 方法有效地融合了视觉关联与上下文双注意力机制, 以生成更加准确且贴近图像真实场景的描述. 在标准数据集上进行了充分的实验验证, 并从定量和定性两个方面证明了 VRCDA 的有效性和优越性.

本文第 1 节介绍图像描述生成的相关工作和研究现状. 第 2 节详细介绍基于视觉关联与上下文双注意力机制的图像描述生成方法. 第 3 节将本文方法与当前主流方法进行对比实验, 以验证本文所提出方法的有效性. 第 4 节总结全文, 并对未来研究做出展望.

1 相关工作

图像描述生成具有信息量大、关系复杂、表达丰富等显著特性, 很多研究者展开了针对性探索. 目前, 主流的图像描述生成方法大多是基于编解码框架的, 该框架由图像编码器和语言解码器组成. Mao 等人[6]创造性地提出基于编解码框架的图像描述生成方法 m-RNN, 使用预训练 CNN 提取图像视觉特征, 利用 RNN 根据已生成的单词和图像的特征向量生成目标词, 不断循环该过程直至生成完整描述. 为了更好地提升模型的解码能力, Vinyals 等人[7]在 m-RNN 基础上, 利用长短期记忆网络 (long short-term memory, LSTM) 代替 RNN 作为解码器, 提出了经典的 NIC 模型, 以 LSTM 作为解码器的结构取得了突破性进展, LSTM 也成为了图像描述生成的通用解码器.

图像描述生成的编解码框架只是简单地将 CNN 最后全连接层的输出作为解码器初始输入的一部分, 仅利用了图像的全局信息, 这样会丢失图像的一些局部特征, 并且在解码过程中无法对部分图像区域进行精准解析. 受到人类的注意力会集中在感兴趣对象上这一现象的启发, 研究者将注意力机制引入到编解码框架中, 改变了编码器与解码器之间的连接方式[8,9], 得到了基于注意力机制的图像描述生成方法. Xu 等人[10]最先将视觉注意力引入到基于编解码框架的图像描述生成方法中, 利用编码器提取图像区域特征, 基于当前时刻 LSTM 隐藏状态和图像区域特征, 通过视觉注意力确定各区域特征的权重, 动态选择与当前时刻生成单词相关的区域特征来指导目标词的生成. Lu 等人[11]提出了一种基于自适应注意力的图像描述生成方法, 在解码过程中依据语义信息为不同的目标词分配不同的视觉注意力权重. Chen 等人[12]提出了一种基于层级注意力的图像描述生成方法, 使用层级注意力动态选择 CNN 的卷积特征图来指导单词生成. Pedersoli 等人[13]提出了基于区域注意力的图像描述生成方法, 不仅考虑了隐藏状态和预测单词之间的关系, 并且兼顾了图像区域和预测单词之间的关系. Anderson 等人[14]使用 Faster R-CNN[15]作为编码器, 提取图像特征的同时检测目标及其所在区域, 将这些区域对应的视觉特征向量送至视觉注意力, 经过自上而下的注意力机制动态分配权重来生成描述. You 等人[16]首先提出了基于文本注意力的图像描述生成方法, 使用目标检测算法获取图像中主要目标的名称属性, 将其作为高层语义信息用于生成描述. Zhou 等人[17]提出了一种基于文本条件注意力的图像描述生成方法, 利用文本信息指导目标词的生成.

上述研究都是基于视觉或文本单注意力机制的图像描述生成方法, 没有考虑视觉和文本在目标词生成时的相辅相成作用. 因此, 有些研究者将视觉与文本注意力结合并引导生成描述, 形成了基于双注意力机制的图像描述生成方法[18]. Liu 等人[19]提出了采用视觉注意力增强图像细节理解, 基于文本注意力完善图像语义, 利用视觉与文本双注意力机制生成描述. Wang 等人[20]采用视觉注意力处理图像视觉信息, 并在文本注意力的基础上提出了回忆词注意力, 将视觉与回忆词注意力结合来保证目标词的高效生成, 以提高生成描述的丰富性和准确性. 本文则在视觉关联与上下文双注意力机制基础上生成图像描述文本.

为了解决交叉熵损失在训练中存在曝光偏差以及训练目标和评估指标不一致的问题, 研究者提出使用强化学习方法来优化模型的生成结果. Ranzato 等人[21]首先提出了一种基于 RNN 的策略梯度强化学习方法, 直接在评价指标上优化模型的生成结果. Rennie 等人[22]提出了一种基于强化学习的自评估序列训练方式 (self-critical sequence

training, SCST), 将模型在测试时的生成描述作为训练的基线, 优化模型以生成更优的描述. 此外, Transformer 也逐渐应用到了图像描述生成中, 用于提升模型的性能. Li 等人[23]提出了纠缠注意力, 使 Transformer 能够同时利用语义和视觉信息. Huang 等人[24]在 Transformer 基础上通过确定注意力结果间的相关度, 拓展了传统注意力机制. Yu 等人[25]受 Transformer 机制的启发, 提出了模块化共同关注网络, 通过共同注意力机制更新视觉特征和文本特征. 本文遵循 SCST 方式对 VRCDA 进行优化处理, 以此解决交叉熵损失在训练中存在的问题, 从而提升 VRCDA 生成描述的效果.

2 方 法

本文提出了一种基于视觉关联与上下文双注意力机制的图像描述生成方法 VRCDA, 该方法采用预训练 Faster R-CNN[15]提取输入图像 I 的视觉特征, 表示为 $V=\{v_1,v_2,...,v_k\}$, 其中 k 是图像区域数量, $v_i\in R^{d_I}$ 表示第 i 个区域的视觉特征, d_I 为特征向量维度; 为了更好地解释其工作机制, 我们选择了 Up-Down 方法[14]与视觉关联注意力相结合得到对应的基础网络, 在视觉特征 V 的基础上, 利用基础网络预先生成初始描述作为全局上下文, 表示为 $C=\{c_1,c_2,...,c_T\}$, $c_i\in R^{d_E}$ 表示 t 时刻的上下文信息, d_E 是词向量维度; 上下文注意力从全局上下文中获取更完整的语义信息, 将其作为一种辅助信息去指导生成最终的图像语义描述, 表示为 $Y=\{y_1,y_2,...,y_T\}$, 其中 T 是生成描述的最大长度. 图 2 展示了 VRCDA 方法的整体架构, 本节将对 VRCDA 的细节和实现过程进行详细介绍.

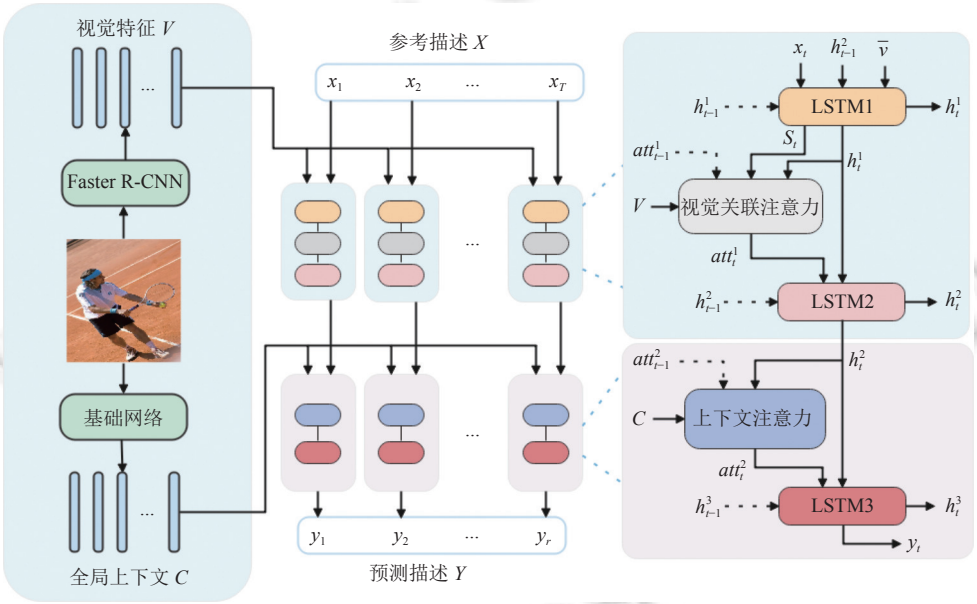

图 2 VRCDA 方法概览

2.1 视觉关联注意力

传统视觉注意力已经被证明对图像描述生成是非常有效的; 实际上, 在生成目标词时, 不仅需要当前时刻的视觉信息, 并且还要依赖先前时刻的视觉信息. 因此, 本文改进了传统视觉注意力, 提出了视觉关联注意力, 主要是通过增加前一时刻的注意力向量来嵌入先前的视觉信息以保证视觉连贯性, 充分利用图像视觉信息.

视觉关联注意力同样对图像区域特征 $V=\{v_1,v_2,...,v_k\}$ 进行加权整合, 即 $att_t^1=\sum_{i=1}^{k}\alpha_{i,t}v_i$, 在每个时刻将加权后的注意力向量作为视觉信息输入到解码器, 让解码器在不同时刻关注到图像的不同区域, 从而充分利用图像中的区域特征信息. 第 i 个区域特征在 t 时刻的注意力权重 $\alpha_{i,t}$ 由解码器的隐藏状态 h_t、前一时刻的注意力向量 att_{t-1}^1 和图像的视觉特征向量 V 共同来计算.

$$\begin{cases} u_{i,t} = W_u^{\mathrm{T}} \tanh(W_v v_i + W_h([h_t; att_{t-1}^1])) \\ \alpha_{i,t} = \dfrac{\exp(u_{i,t})}{\sum_{i=1}^{k} \exp(u_{i,t})} \end{cases} \tag{1}$$

其中, $[\cdot;\cdot]$ 表示向量拼接操作, $u_{i,t}$ 表示当前隐藏状态 h_t 与图像区域特征 v_i 之间的相关性变量, $W_u \in R^{d_A}$ 、 $W_v \in R^{d_A \times d_I}$ 和 $W_h \in R^{d_A \times (d_H + d_I)}$ 是权重矩阵, d_I 、 d_H 和 d_A 分别为视觉特征向量、隐藏状态和注意力向量的维度.

为了自适应地处理视觉和文本信息, 引入了一种自适应门控机制. 该机制的作用是自动调节视觉和文本信息在生成目标词过程中所占的权重, 具体计算方式如下:

$$\begin{cases} g_t = \sigma(W_X X_t + W_h h_{t-1}) \\ s_t = g_t \odot \tanh(m_t) \\ \beta_t = W_s^{\mathrm{T}} \tanh(W_s s_t + W_h h_t) \end{cases} \tag{2}$$

其中, $X_t \in R^{d_E + d_H + d_I}$ 是 t 时刻 LSTM 的全部输入, s_t 是该门控机制中对文本语义信息的一种表示, $m_t \in R^{d_H}$ 是 t 时刻 LSTM 的记忆单元, $W_X \in R^{d_E \times (d_E + d_H + d_I)}$ 和 $W_s \in R^{d_A \times d_I}$ 是权重矩阵, $\odot$ 表示对应项相乘; 标记权重 $\beta_t \in [0,1]$ 表示当前预测单词所关注的文本信息与视觉信息的比例.

视觉关联注意力的结构如图 3 所示, 其完整计算公式如下:

$$\begin{cases} u_{i,t} = W_u^{\mathrm{T}} \tanh(W_v v_i + W_h([h_t^1; att_{t-1}^1])) \\ \alpha_{i,t} = \dfrac{\exp(u_{i,t})}{\sum_{i=1}^{k} \exp(u_{i,t})} \\ \widetilde{att_t^1} = \sum_{i=1}^{k} \alpha_{i,t} v_i \\ att_t^1 = \beta_t s_t + (1 - \beta_t) \widetilde{att_t^1} \end{cases} \tag{3}$$

视觉关联注意力不仅会考虑前一时刻的注意力向量, 以保证视觉连贯性, 并且可以自适应地处理视觉和文本信息, 自动确定两种信息在生成目标词时所占权重.

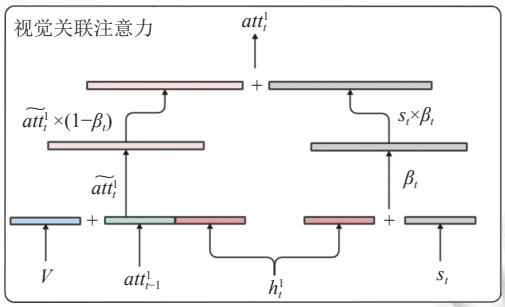

图 3 视觉关联注意力

2.2 上下文注意力

本文采用上下文注意力对全局上下文进行处理, 全局上下文是基础网络预先生成的完整描述, 表示为 $C = \{c_1, c_2, ..., c_T\}$, $c_i \in R^{d_E}$ 表示 t 时刻的上下文信息. 上下文注意力计算方式和传统视觉注意力机制类似:

$$\begin{cases} u_{i,t} = W_u^{\mathrm{T}} \tanh(W_c c_i + W_h h_t^2) \\ \alpha_{i,t} = \dfrac{\exp(u_{i,t})}{\sum_{i=1}^{T} \exp(u_{i,t})} \\ att_t^2 = \sum_{i=1}^{T} \alpha_{i,t} c_i \end{cases} \tag{4}$$

其中, h_t^2 是图 2 中语言解码器 LSTM2 的隐藏状态, $W_u \in R^{d_A}$ 、 $W_c \in R^{d_A \times d_E}$ 和 $W_h \in R^{d_A \times d_H}$ 是权重矩阵, att_t^2 是 t 时刻的上下文注意力向量.

上下文注意力可以保证每一时刻目标词的生成都有选择地回顾上文和展望下文, 以充分利用上下文信息, 将

其作为一种辅助信息去指导生成最终的图像语义描述.

2.3 视觉关联与上下文双注意力机制

为了在生成描述时可以更好地利用图像的视觉信息和上下文信息, 本文使用视觉关联与上下文双注意力机制, 利用视觉关联注意力保证视觉连贯性, 采用上下文注意力获取更完整的语义信息, 以充分利用上下文信息, 进而指导生成图像语义描述.

图 4 展示了 VRCDA 中视觉关联与上下文双注意力机制的结构, 该结构首先利用 LSTM1 处理原始的视觉特征和单词向量, 得到视觉关联注意力所需要的输入 s_t 和 h_t^1, s_t 的计算见公式 (2), h_t^1 的计算公式如下:

$$h_t^1 = LSTM1([x_t; h_{t-1}^2; \overline{v}], h_{t-1}^1) \tag{5}$$

其中, $\overline{v} = \frac{1}{k}\sum v_i$ 表示图像所有区域的平均视觉特征向量. 通过视觉关联注意力计算得到该时刻的视觉注意力向量 att_t^1, 由 att_t^1 引导语言解码器 LSTM2 计算目标词概率 p_t^1.

$$\begin{cases} att_t^1 = f_{\text{VR-att}}(V, [h_t^1; att_{t-1}^1], s_t) \\ h_t^2 = LSTM2([h_t^1; att_t^1], h_{t-1}^2) \\ y_t^1 \sim p_t^1 = softmax(W_p^1 h_t^2) \end{cases} \tag{6}$$

其中, $f_{\text{VR-att}}(\cdot)$ 表示视觉关联注意力, $W_p^1 \in R^{d_E \times d_H}$ 是权重矩阵.

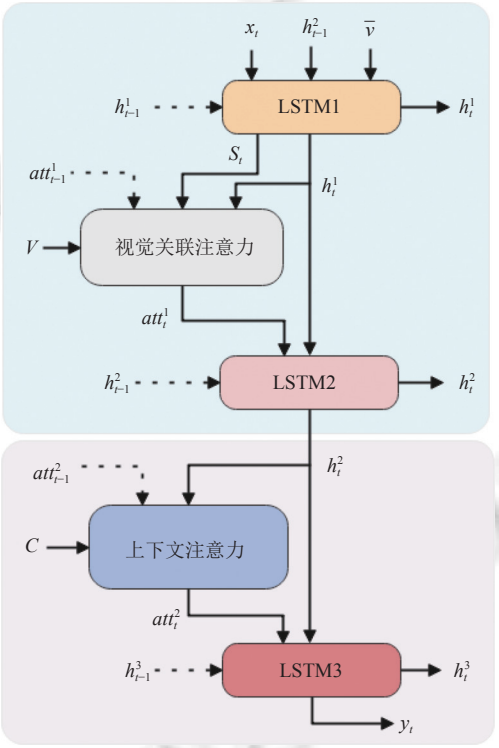

图 4 VRCDA 方法中视觉关联与上下文双注意力

利用上下文注意力去处理全局上下文信息得到上下文注意力向量 att_t^2, 将其作为辅助信息指导语言解码器 LSTM3 计算目标词概率 p_t^2.

$$\begin{cases} att_t^2 = f_{\text{C-att}}(C, [h_t^2; att_{t-1}^2]) \\ h_t^3 = LSTM3([h_t^2; att_t^2], h_{t-1}^3) \\ y_t^2 \sim p_t^2 = softmax(W_p^2 h_t^3) \end{cases} \tag{7}$$

其中, $f_{\text{C-att}}(\cdot)$ 表示上下文注意力, $W_p^2 \in R^{d_E \times d_H}$ 是权重矩阵. 最后, 联合 p_t^1 和 p_t^2 计算目标词 y_t 的生成概率:

$$p_t = p_t^1 + \mu p_t^2 \tag{8}$$

其中, $\mu \in [0,1]$ 为权重系数, $p_t \in R^D$ 是预测的目标词概率向量, D 为词典大小.

在模型的训练中, 首先使用交叉熵损失进行训练, 给定参考描述 $y^* = \{y_1^*, y_2^*, ..., y_T^*\}$, 用 θ 表示模型中的参数, 最小化模型的交叉熵损失函数 $L(\theta)$, 即在每个时刻, 最大化正确参考单词的概率,

$$L(\theta) = -\sum_{t=1}^{T} \log_{p_t}(y_t^*) + \lambda_\theta \|\theta\|_2^2 \tag{9}$$

其中, $\lambda_\theta \|\theta\|_2^2$ 表示 L2 正则化项, 可以在一定程度上防止模型过拟合, 加快模型收敛.

交叉熵损失在训练过程中会存在曝光偏差以及训练目标和评估指标不一致的问题, 为了能够得到更好的生成结果, 本文遵循 SCST 方式[22]对评价指标 CIDEr[26]进行了优化. 基于强化学习方法的训练目标是最大限度地减小负奖励期望:

$$L(\theta) = -E_{Y\text{-}p_\theta}[r(Y)] \approx -r(Y) \tag{10}$$

其中, $r(Y)$ 表示模型生成描述 Y 的 CIDEr 得分. 梯度 $\nabla_\theta L(\theta)$ 可由蒙特卡罗方法近似估计:

$$\nabla_\theta L(\theta) = -E_{Y\text{-}p_\theta}[r(Y)\nabla_\theta \log_{p_\theta}(Y)] \approx -r(Y)\nabla_\theta \log_{p_\theta}(Y) \tag{11}$$

本文遵循 SCST 方式, 使用模型在测试时生成的描述 $\widehat{Y}$ 作为基线, 来鼓励模型生成相对于基线更好的描述:

$$\nabla_\theta L(\theta) \approx -(r(Y) - r(\widehat{Y}))\nabla_\theta \log_{p_\theta}(Y) \tag{12}$$

基于强化学习的训练方法直接在评价指标上优化了描述的生成, 使得模型在训练和测试过程中保持一致, 解决了交叉熵损失训练中存在的问题, 极大地提升了模型性能.

3 实 验

为了验证本文 VRCDA 方法的有效性, 选用 MSCOCO[27]和 Flickr30k[28]两个标准数据集, 将 VRCDA 与其他主流图像描述生成方法进行了对比实验, 并从定量和定性两个角度进行了实验结果分析. 本节将对 VRCDA 方法的实验细节及其结果进行详细介绍与分析.

3.1 数据集与评价指标

MSCOCO 数据集是当前图像描述生成公开通用的大型英文数据集, 共有 164 062 幅图像. MSCOCO 的原测试集并没有提供参考描述, 因此本文采用 Karpathy[29]的划分方式, 从原验证集分别选取 5 000 幅图像及其参考描述作为验证集和测试集, 再将验证集剩余图像与原训练集重新组合, 得到包含 113 287 幅图像及其参考描述的训练集, 其中的每幅图像对应至少 5 条人工标注的参考描述. Flickr30k 数据集是公开通用的小型英文数据集, 共有 31 783 幅图像, 其中 29 000 幅图像构成训练集, 分别选择 1 000 幅图像构成验证集和测试集, 每幅图像同样对应 5 条人工标注的参考描述.

在测试阶段, 本文采用 5 种常见的评价指标, 分别为 BLEU[30]、METEOR[31]、ROUGE-L[32]、CIDEr[26]、SPICE[33]. 其中, BLEU 衡量机器生成描述的准确性, 使用了 n-gram 统计生成描述和参考描述之间的覆盖率; METEOR 是 BLEU 的改进版, 其更注重语句中单词的召回率和准确率; ROUGE-L 是一种评估文本摘要质量的指标, 基于最长公共子串来计算准确率和召回率; CIDEr 用于评测图像描述一致性和丰富度, 基于 TF-IDF 计算生成描述与参考描述的余弦相似度来衡量文本的一致性; SPICE 是一种基于场景图和语义概念的评估指标, 用于衡量生成描述是否有效地描述了图像中对象、属性及它们之间的关系.

3.2 实验设置

Faster R-CNN 图像编码器从每幅图像提取 36 个显著区域, 每个区域使用 2 048 维的特征向量表示. 语言解码器的词向量、注意力层和 LSTM 隐藏层维度皆设置为 1 024, 其他网络参数采用随机初始化.

在训练阶段, 设置批量大小为 100, 最大迭代周期为 50. 使用 Adam 优化器, 设置初始学习率为 5×10^{-4}, 动量参数设置为 0.9. 为加快模型收敛, 在第 10 轮后, 学习率每 3 轮衰减一次, 衰减率为 0.8. 在每轮训练结束后, 在验证

集上评估模型的性能, 最后选择在验证集上具有最高 CIDEr 的模型用于测试. 在测试阶段, 模型使用 Beam Search 方法[34]进行解码, Beam Size 设置为 3.

此外, 实验设置了如公式 (8) 所示的权重参数μ, 以控制全局上下文信息对生成描述的影响程度, 从而针对不同的目标词能够更好地利用其对应的上下文信息. 首先实验验证了μ值对生成描述的影响, 结果如图 5 所示, 其中横坐标表示μ的取值, 纵坐标对应指标 BLEU-4 和 CIDEr 的值.

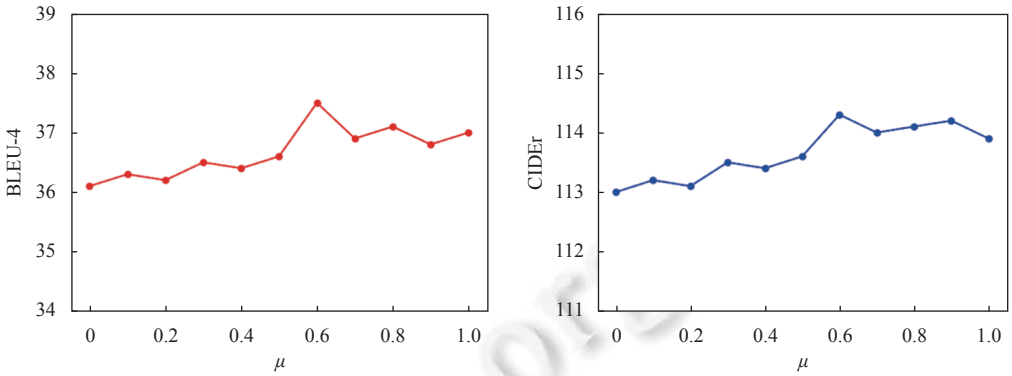

图 5 不同μ值对 VRCDA 性能的影响

从图 5 可以看出, 当$\mu=0$时, VRCDA 整体性能最差, 因为没有利用全局上下文信息; 当$\mu>0$时, 无论是 BLEU-4 还是 CIDEr 都有明显提升, 说明了本文 VRCDA 方法的有效性; 在$\mu=0.6$时, VRCDA 整体性能最好; 因此, 后续实验将μ设置为 0.6.

3.3 实验结果定量分析

为了验证本文方法 VRCDA 的有效性, 选择 Up-Down 方法[14]与视觉关联注意力相结合得到对应的基础网络. 同时, 为了展示更加公平的对比, 本文首先将 Up-Down 作为基线方法, 使用完全相同的数据和参数来进行后续对比实验. 表 1 展示了 VRCDA 在 MSCOCO 数据集上与基线方法的对比结果, 其中 B1、B4、M、R、C 和 S 分别表示 BLEU-1、BLEU-4、METEOR、ROUGE-L、CIDEr 和 SPICE.

表 1 VRCDA 及其消融模型与基线方法在 MSCOCO 数据集上的结果对比 (%)

Models	Cross-Entropy Loss						CIDEr Optimization					
	B1	B4	M	R	C	S	B1	B4	M	R	C	S
基线方法	76.4	36.0	26.9	56.4	112.7	20.3	79.3	36.2	27.5	56.9	119.8	21.2
VRCDA-PA	76.6	36.3	27.3	56.4	113.0	20.4	79.8	36.6	27.7	56.9	122.6	21.4
VRCDA-GC	77.2	36.7	27.6	56.7	113.4	20.5	80.3	37.0	27.8	57.0	123.1	21.5
VRCDA	**77.5**	**37.2**	**28.0**	**57.3**	**114.3**	**20.9**	**80.6**	**37.9**	**28.4**	**58.2**	**123.7**	**21.8**

由表 1 可以看出, 本文方法 VRCDA 在交叉熵损失训练下, 各项指标上仅略高于基线方法. 由于在交叉熵损失训练下, 存在曝光偏差以及训练目标和评估指标不一致的问题, 模型很难取得最大的性能提升, 经过强化学习优化后, 整体性能获得了最大程度的提升, 在各项指标上明显优于基线方法, 其中反映描述准确性的 BLEU-4 提高了 1.7 个百分点, 反映语义丰富程度的 CIDEr 提高了 3.9 个百分点. VRCDA 的 BLEU-4 与 CIDEr 分别达到了 37.9% 与 123.7%, 这表明, VRCDA 采用视觉关联和上下文双注意力机制可以一定程度上保证视觉和语言连贯性, 从而得到更加丰富、更加准确且符合图像真实场景的文本描述.

同时, 为了验证在 VRCDA 中引入前一时刻的注意力向量和全局上下文信息的有效性, 对 VRCDA 进行了消融实验, 实验结果如表 1 所示, 其中 VRCDA-PA 和 VRCDA-GC 分别表示在基线方法基础上使用视觉关联注意力和采用上下文注意力处理全局上下文信息. 由表 1 可以看出, VRCDA-PA、VRCDA-GC 相比于基线方

法在各项指标上均有所提高, 表明使用视觉关联注意力以及采用上下文注意力处理全局上下文信息的有效性.

将 VRCDA 与当前主流图像描述生成方法在 MSCOCO 数据集上进行实验对比, 包括 Xu 等人[10]基于空间注意力的 Soft-Attention 方法和 Hard-Attention 方法、Lu 等人[11]基于自适应注意力的 Adaptive-Attention 方法、Chen 等人[12]基于层级注意力的 SCA-CNN 方法、Pedersoli 等人[13]基于区域注意力的 Area-Attention 方法、You 等人[16]基于语义注意力的 Semantic-Attention 方法、Rennie 等人[22]基于 SCST 方式的 Att2in 方法和 Att2all 方法、Anderson 等人[14]基于自上而下注意力的 Up-Down 方法、Wang 等人[35]基于视觉关系注意力的 A_R_L 方法及 Li 等人[36]基于改进的视觉注意力的 IVAIC 方法. 表 2 给出了 VRCDA 和上述方法在 MSCOCO 上的实验结果.

表 2 在 MSCOCO 数据集上与其他主流方法对比实验结果 (%)

Models	Cross-Entropy Loss						CIDEr Optimization					
	B1	B4	M	R	C	S	B1	B4	M	R	C	S
Soft-Attention [10]	70.7	24.3	23.9	—	—	—	—	—	—	—	—	—
Hard-Attention [10]	71.8	25.0	23.0	51.6	86.5	—	—	—	—	—	—	—
Adaptive-Attention [11]	74.2	33.2	26.6	—	108.5	—	—	—	—	—	—	—
SCA-CNN [12]	71.9	31.1	25.0	53.1	95.2	—	—	—	—	—	—	—
Area-Attention [13]	—	31.9	25.2	—	98.1	—	—	—	—	—	—	—
Semantic-Attention [16]	70.9	30.4	24.3	54.3	104.2	—	—	—	—	—	—	—
SCST: Att2in [22]	—	31.3	26.0	54.3	101.3	—	—	33.3	26.3	55.3	111.4	—
SCST: Att2all [22]	—	30.0	25.9	53.4	99.4	—	—	34.2	26.7	55.7	114.0	—
Up-Down [14]	77.2	36.2	27.0	56.4	113.5	20.3	79.8	36.3	27.7	56.9	120.1	21.4
A_R_L [35]	75.9	35.8	27.8	56.4	111.3	—	—	—	—	—	—	—
IVAIC [36]	76.5	36.3	27.6	56.4	113.7	20.5	79.9	37.9	27.8	58.1	122.3	21.5
Ours: VRCDA	**77.5**	**37.2**	**28.0**	**57.3**	**114.3**	**20.9**	**80.6**	**37.9**	**28.4**	**58.2**	**123.7**	**21.8**

由表 2 可以看出, VRCDA 依然具有很强的竞争优势, 在各项指标上全面优于其他主流图像描述生成方法, 进一步证明了 VRCDA 的有效性.

最后, 为了证明 VRCDA 在其他数据集上的泛化能力, 将 VRCDA 与当前主流图像描述生成方法在 Flickr30k 数据集上进行实验对比, 实验结果如表 3 所示.

表 3 在 Flickr30k 数据集上与其他主流方法对比实验结果 (%)

Models	B1	B4	M	R	C	S
Soft-Attention [10]	66.7	19.1	18.5	—	—	—
Hard-Attention [10]	66.9	19.9	18.5	—	—	—
Adaptive-Attention [11]	67.7	25.1	20.4	—	53.1	—
SCA-CNN [12]	66.2	22.3	19.5	—	—	—
Semantic-Attention [16]	64.7	23.0	18.9	—	—	—
A_R_L [35]	69.8	27.7	21.5	48.5	57.4	—
IVAIC [36]	70.8	30.6	22.5	49.8	63.0	16.8
Ours: VRCDA	**73.2**	**30.6**	**22.7**	**50.6**	**66.0**	**16.8**

由表 3 可以看出, VRCDA 与其他主流图像描述生成方法相比在各项指标上都有更好的表现, 这充分说明, 无论数据集的大小, 本文提出的视觉关联注意力和利用上下文注意力处理全局上下文的思想都是非常有效的. 最终的实验结果也证明了 VRCDA 在 Flickr30k 这种小型数据集上依然能取得良好的性能表现和泛化能力, 可以显著地提升生成描述的效果.

3.4 实验结果定性分析

图 6 展示了 VRCDA 与 Up-Down 生成的一些描述示例对比, 其中 GT 表示参考描述, 可以很直观地发现, VRCDA 的生成描述在对象动作、细节以及上下文连贯性上明显优于 Up-Down.

				
Up-Down	A young man holding a tennis ball on a court.	A lttle girl holding a toothbrush in her mouth.	A traffic light on a city street with cars.	A man holding a frisbee in the field.
VRCDA	A young man hitting a tennis ball with a tennis racket.	A lttle girl brushing her teeth with a toothbrush.	A group of cars on city street with a traffic light.	A man holding a yellow frisbee in front of a building.
GT	A boy attempts to hit the tennis ball with the racquet.	A little girl brushing her teeth with an electric toothbrush.	A traffic light and some cars on a city street.	A man holding a frisbee in the field close to some buildings.
	(a)	(b)	(c)	(d)

图 6 VRCDA 与 Up-Down 生成描述实例分析

(1) VRCDA 可以更加精确地描述对象的动作, 使得生成的描述更贴近图像内容. 如图 6(a), VRCDA 可以精确地生成“A young man hitting a tennis ball with a tennis racket”, 而不是像 Up-Down 那样粗略地生成“A young man holding a tennis ball on a court”.

(2) VRCDA 的生成描述在语言上更加平滑与连贯, 符合人类表达习惯. 如图 6(b), VRCDA 生成的“A little girl brushing her teeth with a toothbrush”显然比 Up-Down 生成的“A little girl holding a toothbrush in her mouth”更加衔接良好且语义连贯.

(3) VRCDA 可以获得更丰富的词汇, 检测出更准确的细节, 具有更强的语言表现力. 如图 6(c), VRCDA 用“a group of”去修饰“cars”而不是单调地预测出“cars”; 又如图 6(d), VRCDA 可以检测出飞盘的颜色, 并准确预测出了“man”和“building”之间的位置关系, 这是 Up-Down 无法实现的.

图 7 的图像描述可以进一步说明 VRCDA 的生成描述同图像真实场景的关系. 图 7(a) 中, VRCDA 可以精准地描述出“bears”之间的位置关系是“next to each other”, 而参考描述简单地叙述为“bears standing”; 图 7(b) 中, VRCDA 可以非常精细地给出“chairs”与“umbrella”的位置关系是“under”, 不是简单地像参考描述那样使用“and”来连接两个对象. 在方法中引入全局上下文信息以及前一时刻的视觉注意力向量也可能会带来一些干扰信息, 从而导致生成描述质量不高, 如图 7(c) 和 (d) 所示. 图 7(c) 中, VRCDA 为“man”错误预测了“jumping”, 这原本应该是“dog”的动作, 图 7(d) 中, VRCDA 将最后一个单词错误预测成“table”, 原本应该是“fork”.

				
VRCDA	A couple of bears standing next to each other on a rock.	Two beach chairs under an umbrella on the beach.	A man jumping in the air to catch a frisbee.	A plate of food on a table with a table.
GT	A couple of bears standing on top of a rock.	Two beach chairs and an umbrella on a beach.	A man on pier with dog jumping for frisbee into water	A plate of food on a table with a fork.
	(a)	(b)	(c)	(d)

图 7 VRCDA 生成描述与参考描述实例分析

4 总结与展望

本文提出了一种基于视觉关联与上下文双注意力机制的图像描述生成方法 VRCDA, 该方法在生成描述的过程中考虑了视觉连贯性以及上下文信息, 显著地提升了图像描述的生成效果. 在 MSCOCO 和 Flickr30k 数据集上进行了验证, 实验结果表明本文所提出的 VRCDA 方法相比于其他主流的图像描述生成方法具有明显的性能优势. 此外, 本文提出的视觉关联注意力和利用上下文信息去指导生成描述的思想可以应用于大多数基于注意力机制的图像描述生成方法中, 表现出了较强的通用性和泛化能力.

未来计划优化视觉关联与上下文双注意力机制, 尤其是其有效的组合方式, 以充分利用图像视觉和上下文信息, 生成更加准确且丰富的图像描述文本. 此外, 在解码器和文本生成阶段引入语言句法和语义信息, 以增强基于深度学习的图像描述生成方法的可解释性.

References:

[1] Zhou YE, Wang M, Liu DQ, Hu ZZ, Zhang HW. More grounded image captioning by distilling image-text matching model. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 4776–4785. [doi: 10.1109/CVPR42600.2020.00483]

[2] Wang JC, Zhou YE, Hu ZZ, Zhang X, Wang M. Sequential image encoding for vision-to-language problems. Multimedia Tools and Applications, 2021, 80(11): 16141–16152. [doi: 10.1007/s11042-019-08439-7]

[3] Liu XL, Wang M, Zha ZJ, Hong RC. Cross-modality feature learning via convolutional autoencoder. ACM Trans. on Multimedia Computing, Communications, and Applications, 2019, 15(S1): 7. [doi: 10.1145/3231740]

[4] Xue ZY, Guo PY, Zhu XB, Zhang NG. Image description method based on generative adversarial networks. Ruan Jian Xue Bao/Journal of Software, 2018, 29(S2): 30–43 (in Chinese with English abstract).

[5] Fu K, Jin JQ, Cui RP, Sha F, Zhang CS. Aligning where to see and what to tell: Image caption with region-based attention and scene-specific contexts. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2321–2334. [doi: 10.1109/TPAMI.2016.2642953]

[6] Mao JH, Xu W, Yang Y, Wang J, Huang ZH, Yuille AL. Deep captioning with multimodal recurrent neural networks (m-RNN). In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego, 2015. 1–17.

[7] Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: A neural image caption generator. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 3156–3164. [doi: 10.1109/CVPR.2015.7298935]

[8] Yu J, Li J, Yu Z, Huang QM. Multimodal transformer with multi-view visual representation for image captioning. IEEE Trans. on Circuits and Systems for Video Technology, 2020, 30(12): 4467–4480. [doi: 10.1109/TCSVT.2019.2947482]

[9] Zha ZJ, Liu DQ, Zhang HW, Zhang YD, Wu F. Context-aware visual policy network for fine-grained image captioning. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2022, 44(2): 710–722. [doi: 10.1109/TPAMI.2019.2909864]

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

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

[12] Chen L, Zhang HW, Xiao J, Nie LQ, Shao J, Liu W, Chua TS. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6298–6306. [doi: 10.1109/CVPR.2017.667]

[13] Pedersoli M, Lucas T, Schmid C, Verbeek J. Areas of attention for image captioning. In: Proc. of the 2017 Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 1251–1259. [doi: 10.1109/ICCV.2017.140]

[14] Anderson P, He XD, Buehler C, Teney D, Johnson M, Gould S, Zhang L. Bottom-Up and top-down attention for image captioning and visual question answering. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6077–6086. [doi: 10.1109/CVPR.2018.00636]

[15] Ren SQ, He KM, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: 10.1109/TPAMI.2016.2577031]

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

[17] Zhou LW, Xu CL, Koch P, Corso JJ. Watch what you just said: Image captioning with text-conditional attention. In: Proc. of the Thematic Workshops of ACM Multimedia. Mountain: ACM, 2017. 305–313. [doi: 10.1145/3126686.3126717]

[18] Yu LT, Zhang J, Wu Q. Dual attention on pyramid feature maps for image captioning. IEEE Trans. on Multimedia, 2021, 24: 1775–1786. [doi: 10.1109/TMM.2021.3072479]

[19] Liu MF, Li LJ, Hu HJ, Guan WL, Tian J. Image caption generation with dual attention mechanism. Information Processing & Management, 2020, 57(2): 102178. [doi: 10.1016/j.ipm.2019.102178]

[20] Wang L, Bai ZC, Zhang YH, Lu HT. Show, recall, and tell: Image captioning with recall mechanism. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI, 2020. 12176–12183. [doi: 10.1609/aaai.v34i07.6898]

[21] Ranzato MA, Chopra S, Auli M, Zaremba W. Sequence level training with recurrent neural networks. In: Proc. of the 4th Int'l Conf. on Learning Representations. 2016. 1–16.

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

[23] Li G, Zhu LC, Liu P, Yang Y. Entangled transformer for image captioning. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 8927–8936. [doi: 10.1109/ICCV.2019.00902]

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

[25] Yu Z, Yu J, Cui YH, Tao DC, Tian Q. Deep modular co-attention networks for visual question answering. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 6274−6283. [doi: 10.1109/CVPR.2019.00644]

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

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

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

[29] Karpathy A, Li FF. Deep visual-semantic alignments for generating image descriptions. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(4): 664–676. [doi: 10.1109/tpami.2016.2598339]

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

[31] Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proc. of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Ann Arbor: ACL, 2005. 65–72.

[32] Lin CY. ROUGE: A package for automatic evaluation of summaries. In: Proc. of Text Summarization Branches Out. Barcelona: ACL, 2004. 74–81.

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

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

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

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

附中文参考文献:

[4] 薛子育, 郭沛宇, 祝晓斌, 张乃光. 一种基于生成式对抗网络的图像描述方法. 软件学报, 2018, 29(S2): 30–43.

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

刘茂福(1977－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为自然语言处理.

聂礼强(1985－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为多媒体内容分析与搜索.

施琦(1997－), 男, 硕士生, 主要研究领域为自然语言处理.