计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(7):32−40 [doi: 10.15888/j.cnki.csa.008010]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

融合双注意力与多标签的图像中文描述生成方法①

田　枫[1], 孙小强[1], 刘　芳[1], 李婷玉[2], 张　蕾[2], 刘志刚[1]

[1](东北石油大学 计算机与信息技术学院, 大庆 163318)
[2](中国石油天然气股份有限公司 冀东油田分公司 信息中心, 唐山 063004)
通讯作者: 刘　芳, E-mail: lfliufang1983@126.com

摘　要: 图像描述是目前图像理解领域的研究热点. 针对图像中文描述句子质量不高的问题, 本文提出融合双注意力与多标签的图像中文描述生成方法. 本文方法首先提取输入图像的视觉特征与多标签文本, 然后利用多标签文本增强解码器的隐藏状态与视觉特征的关联度, 根据解码器的隐藏状态对视觉特征分配注意力权重, 并将加权后的视觉特征解码为词语, 最后将词语按时序输出得到中文描述句子. 在图像中文描述数据集 Flickr8k-CN、COCO-CN 上的实验表明, 本文提出的模型有效地提升了描述句子质量.

关键词: 图像描述; 图像理解; 图像中文描述; 注意力机制; 图像多标签

引用格式: 田枫,孙小强,刘芳,李婷玉,张蕾,刘志刚.融合双注意力与多标签的图像中文描述生成方法.计算机系统应用,2021,30(7):32–40. http://www.c-s-a.org.cn/1003-3254/8010.html

Chinese Image Caption with Dual Attention and Multi-Label Image

TIAN Feng[1], SUN Xiao-Qiang[1], LIU Fang[1], LI Ting-Yu[2], ZHANG Lei[2], LIU Zhi-Gang[1]

[1](School of Computer & Information Technology, Northeast Petroleum University, Daqing 163318, China)
[2](Information Center, Jidong Oilfield Branch, PetroChina Co. Ltd., Tangshan 063004, China)

Abstract: Image caption represents a research hotspot in the field of image understanding. In view of the poor quality of sentences, we propose Chinese image caption combining dual attention and multi-label images. We extract visual features and multi-label text firstly, and then use multi-label text to enhance the correlation between the hidden state of the decoder and visual features. Next, we redistribute attention weights to the visual features according to the hidden state of the decoder and decode the weighted features into words. Finally, the words are output in a time sequence to obtain Chinese sentences. Experiments on Chinese image caption datasets, Flickr8k-CN and COCO-CN, reveal that the proposed method substantially improves the quality of sentences.

Key words: image caption; image understanding; Chinese image caption; attention mechanism; multi-label image

图像是目前信息传播的主流媒介之一, 随着成像设备的普及, 图像数据量增长迅速. 然而图像以像素的形式存储, 这与用户对图像的解读之间存在巨大的差异, 高效地对海量图像资源进行检索和管理极具挑战性. 如何使计算机依照人类理解的形式对图像进行描述, 已是目前图像理解领域的研究热点.

① 基金项目: 黑龙江省自然科学基金 (LH2020F003); 国家自然科学基金 (61502094); 黑龙江省省属本科高校基本科研业务费项目 (KYCXTD201903); 中央支持地方高校改革发展资金人才培养支持计划 (140119001); 东北石油大学研究生教育创新工程 (JYCX_11_2020); 东北石油大学引导性创新基金 (2020YDL-11)

Foundation item: Natural Science Foundation of Heilongjiang Province (LH2020F003); National Natural Science Foundation of China (61502094); Basic Research Foundation of Heilongjiang Provincial Higher Educations (KYCXTD201903); The Talent Training Support Program Funded By the Central Government to Support the Reform and Development of Local Colleges and Universities (140119001); Graduate Innovation Project of Northeast Petroleum University (JYCX_11_2020); Guiding Innovation Fund of Northeast Petroleum University (2020YDL-11)

收稿时间: 2020-10-22; 修改时间: 2020-11-28, 2020-12-12; 采用时间: 2020-12-25; csa 在线出版时间: 2021-06-30

图像描述 (Image Caption, IC)[1,2] 是指计算机针对给定的图像, 自动地以符合人类语法规则的句子将该图像的画面内容进行转换. 句子不仅可以作为图像检索的元数据, 从而提升对图像资源的检索和管理效率; 而且相比词汇标签能更形象、直观地传达图像内容, 因此图像描述任务吸引了众多研究者关注. 现有的图像描述研究可分为 3 类, 分别是基于模板的图像描述生成方法[3]、基于检索的图像描述生成方法[4] 以及基于翻译的图像描述生成方法[5–10], 其中基于翻译的图像描述生成方法借助深度学习端到端的训练特性, 通过在大规模图像句子对应数据集上进行学习, 模型生成的描述句子更为新颖. 现有的图像描述工作[3–10] 以研究如何为图像生成英文的描述句子为主, 显然此项研究不应该受限于语言, 将图像描述研究扩展到母语使用人口最多的中文环境, 具有更为重要的现实意义.

相比英文描述, 中文词语含义更加丰富, 中文句子结构也更为复杂, 因此图像中文描述任务更具有难度; 在模型构建方式上, 现有的图像英文描述研究利用编码器-解码器框架[5–7] 和融合注意力机制的编码器-解码器框架[8–10] 来构建模型, 而图像中文描述模型主要是基于编码器-解码器框架构建的. 例如, 2016 年 Li 等人[11] 将文献 [5] 中的模型在中文的环境下重新训练, 实现了首个图像中文描述模型 CS-NIC; 2019 年张凯等人[12] 通过利用机器翻译构建伪语料库, 从而将常规编码器-解码器框架在中文的环境下重新训练. 通过实验发现, 虽然现有的图像中文描述模型可以对图像进行描述, 但是描述句子的质量仍有待提升. 通过对现有的图像中文描述研究进行分析, 本文认为目前图像中文描述句子质量不高的原因可以归结为: 1) 现有研究利用编码器-解码器框架来构建模型, 该框架仅在解码器的仅接收一次图像特征, 由于解码器的“遗忘”特性, 导致模型生成的描述句子整体质量不高; 2) 中文词语的含义较为丰富, 现有研究在解码视觉特征时, 并未考虑视觉特征中的误差因素; 3) 现有方法的优化目标主要是基于输入视觉特征和已经生成的词语, 使预测的下一个词语是正确词语的概率最大化, 这在一定程度上忽略了最终生成句子整体语义与图像内容的关联度.

针对上述问题, 2016 年 Xu 等人[8] 在编码器-解码器框架中引入注意力机制, 使单词与图像视觉特征之间进行对齐, 提升了模型对视觉特征的鉴别能力. 注意力机制与人眼视觉特性相似, 其原理是使模型在生成文字序列时, 自主决定图像特征的权值, 从而实现模型动态地关注图像中重要的区域. 此外, 也有研究利用从图像中提取多标签信息对模型生成的描述句子质量进行改善. 例如, 2019 年蓝玮毓等人[13] 利用概率编码的图像多标签重排模型生成的描述句子候选集, 提升了模型生成的描述句子与图像内容的关联度. 因此为提升中文描述句子质量, 本文在现有研究的基础上提出融合双注意力与多标签的图像中文描述方法. 本文方法通过融合图像多标签文本信息, 增强解码器与图像内容的关联度; 通过利用注意力机制, 使模型能更好地利用视觉特征. 通过实验对比分析, 本文模型生成的图像描述句子更符合图像的内容, 对图像的背景等细节信息也能够进行描述.

1 相关工作

现有的图像描述生成方法可分为 3 个类别.

1) 基于模板的图像描述生成方法. 该类方法先利用计算机视觉技术识别出图像中视觉语义信息, 然后填充到模板句子中. 该类研究的代表性工作为 Fang 等人[3] 利用卷积神经网络 (Convolutional Neural Network, CNN)[14,15] 预测出一系列词语, 再利用最大熵语言模型生成描述句子. 该类方法往往能生成语法正确的描述句子, 但是由于模板句子的数量有限, 导致描述句子的多样性受限.

2) 基于检索的图像描述生成方法. 该类方法将相似图像描述句子作为输入图像的描述句子. 该类研究的代表性工作为 Ordonez 等人[4] 从 Flickr 网站收集大量图片, 通过使用数据清洗技术使最终检索库中的每幅图像均对应一个描述句子, 然后寻找与测试图像最相似图像, 将该图像的描述句子作为测试图像的描述句子. 该类方法往往能生成语义正确的描述句子, 但是严重依赖于检索算法与数据集质量, 当数据集中缺少与目标图像相似的图像, 将导致匹配失败.

3) 基于翻译的图像描述生成方法. 该类方法受机器翻译的启发, 将图像看作待翻译数据, 描述句子视为翻译结果, 利用编码器-解码器框架将图像内容进行翻译. 该类研究的代表性工作为 Vinyals 等人[5] 利用 CNN 作为编码器, 将图像编码为特定长度的语义向量, 然后利用长短时记忆网络 (Long Short Term Memory Neural Network, LSTM)[16] 作为解码器, 对语义向量进行解码, 模型使用最大似然概率函数进行训练. 汤鹏杰等人[7]

在编码器-解码器框架中融合场景信息, 使模型的性能得到提升. 该类方法生成的描述句子要更为新颖, 成为目前构建图像描述模型的主流方法.

Xu 等人[8] 将注意力机制引入到图像描述研究中, 先利用 CNN 提取图像的卷积层特征, 在生成每一个单词时, 根据解码器 LSTM 的隐藏状态计算出各个特征区域对应的权重, 通过权重乘上对应区域的特征对图像特征重新加权, 然后由解码器对加权后的特征进行解码. 该类方法通过使模型动态关注图像中的重要区域, 提升模型的性能, 吸引了众多研究者关注. 随后, 越来越多的研究者进一步地提出不同的注意力机制, 比如全局-局部注意力机制[9]、自适应注意力机制[10] 等.

仅有少数的工作研究了面向非英语语种的图像描述. 李锡荣等人[11] 借助人工翻译、机器翻译得到首个中文的数据集 Flickr8k-cn, 并对文献 [5] 中的模型重新训练, 得到首个图像中文描述模型. 张凯等人[12] 通过利用机器翻译构建伪语料库, 从而完成了端到端的中文描述生成. 蓝玮毓等人[13] 提出利用概率向量编码的图像标签信息, 重排生成的图像描述文本集, 提升了模型生成的描述句子与图像内容的关联度. 这些工作都是利用编码器-解码器框架构建模型, 解码器仅接收一次视觉特征, 而且在解码过程中对视觉特征的利用方式简单, 虽然注意力机制可以根据解码器的隐藏状态增强视觉特征的利用方式, 但是融合注意力机制的编码器-解码器框架仅在图像英文描述生成中被证实是可行的, 由于中文与英文之间的差异, 注意力机制能否应用到图像中文描述研究中仍有待验证; 此外, 现有利用图像多标签改善描述句子质量的研究是使用概率向量编码的图像多标签进行的, 且并没有利用图像多标签信息生成新的描述句子, 对于一幅图像, 其中的对象、场景、行为等信息往往是确定的, 如何使用非概率编码的多标签文本辅助模型生成更高质量的描述句子仍需要实验进行验证; 最后, 本文根据图像中包含的目标类型和数量, 对模型的描述能力进行分析.

2 融合双注意力与多标签的图像中文描述生成模型

1) 优化目标

图像多标签不仅能反映图像内容, 而且能作为描述句子中的词语, 可为模型生成更高质量的中文描述文本提供帮助. 因此对于输入图像 I_i, 令中文词表为 D, 本文为该图像预测一个中文标签集合 $\{L_i\}$, 标签 L_i 与 D 中的单词对应. 即本文利用 L_i 辅助模型生成更高质量的中文描述句子 S. 模型的训练目标为式 (1) 所示:

$$\theta* = \arg\max \sum_{i=1}^{N}\sum_{t=1}^{M} \log p(S|L_i, I_i; \theta) \tag{1}$$

其中, θ 是模型需要学习的参数; N 是训练集图像的数量; i 是指数据集中第 i 幅图像, M 是指第 i 幅图像对应的描述句子 S 的长度, $S=\{s_1, s_2, \cdots, s_n\}$. 因为 CNN 提取的视觉特征相比图像 I_i 本身, 能更好地反映其高层语义, 因此令 $V(I_i)=I_i$, $V(I_i)$ 表示图像对应的视觉特征, $W(I_i)=L_i$, $W(I_i)$ 表示图像对应的多标签文本信息. 利用链式求导法则, 式 (1) 可转化式 (2):

$$J(\theta) = \sum_{i=1}^{N}\sum_{t=1}^{M} \log(s_t|s_1, s_2, \cdots, s_{t-1}, W(I_i), V(I_i)) \tag{2}$$

本文利用图像语义编码网络和双注意力解码网络对式 (2) 进行求解. 图像语义编码网络用于提取视觉特征 $V(I_i)$ 和多标签文本 $W(I_i)$; 双注意力解码网络根据图像多标签文本 $W(I_i)$, 更好地对视觉特征 $V(I_i)$ 进行解码, 从而生成更高质量的描述句子 S. 融合双注意力与多标签的图像中文描述生成模型框架如图 1 所示.

2) 图像语义编码网络

如图 1 所示, 图像语义编码网络由两部分组成, 分别是视觉特征编码网络和多标签文本生成网络. 对于输入图像 I_i, 视觉特征编码网络输出该图像的视觉特征 $V(I_i)$, 多标签文本生成网络输出多标签文本 $W(I_i)$. 为防止模型在训练阶段的损失相互干扰, 图像多标签文本预测网络与图像视觉特征编码网络分离开进行训练. 接下来本文分别介绍这两个子网络.

① 视觉特征提取网络

本文以 ResNet101 作为视觉特征提取网络. ResNet-101 是 ResNet[14] 衍生出的一种网络, 通过在大规模图像分类数据集 ImageNet[15] 上进行训练, ResNet101 在目标识别、目标检测等领域仍能有效刻画图像视觉信息. 对于输入图像 I_i, 将其缩放到 256×256 个像素, 利用视觉特征提取网络输出其视觉特征 $V(I_i)$.

② 多标签文本生成网络

本文微调 AlexNet[15] 网络结构, 将微调后的网络作为本文的多标签文本生成网络. AlexNet 是深度学习

的一个代表性网络, 不仅在图像分类等任务上表现优异, 而且比 ResNet 等网络的计算量少. 但是 AlexNet 网络本身并不适用于多标签分类, 因此本文将 AlexNet 网络输出层的神经元结点的数量修改为中文词表 D 的长度, 并将最后一层的激活函数改为适合多分类的 Sigmoid 函数. 训练过程中以 BCEloss[17] 作为模型的损失函数, 其数学表达式如式 (3) 所示:

$$BCEloss(O,T)=-\frac{1}{N}\sum_{i}^{N}\sum_{j}^{m}(T_{ij}*\log(O_{ij})+(1-T_{ij})*\log(1-O_{ij})) \tag{3}$$

其中, N 是图像数量, m 是标签数量, 其中 $O\in\{0,1\}^{n*o}$, 表示样本的真实标签, $T\in R^{n*o}$, 是模型对不同标签的预测概率输出. 多标签文本生成网络的训练过程如图 2 所示.

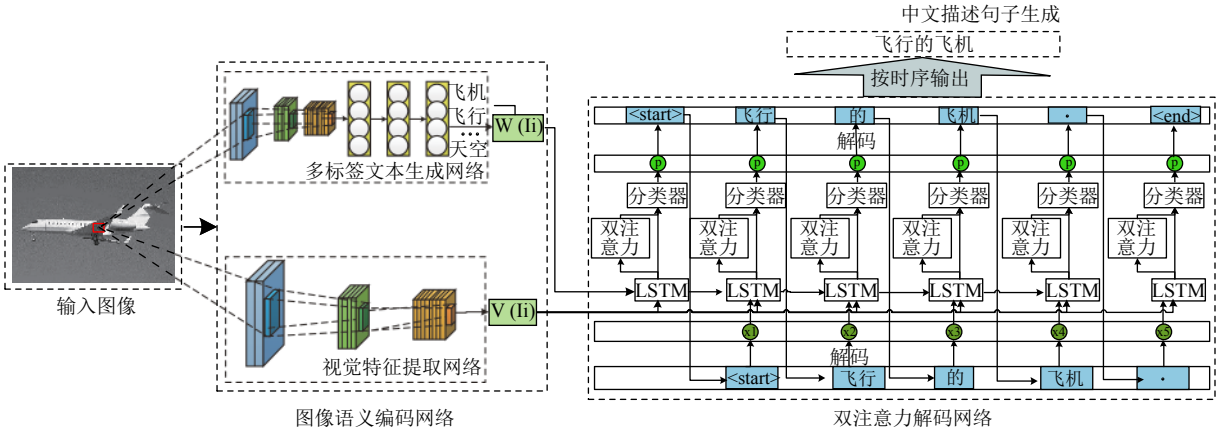

图 1 模型框架图

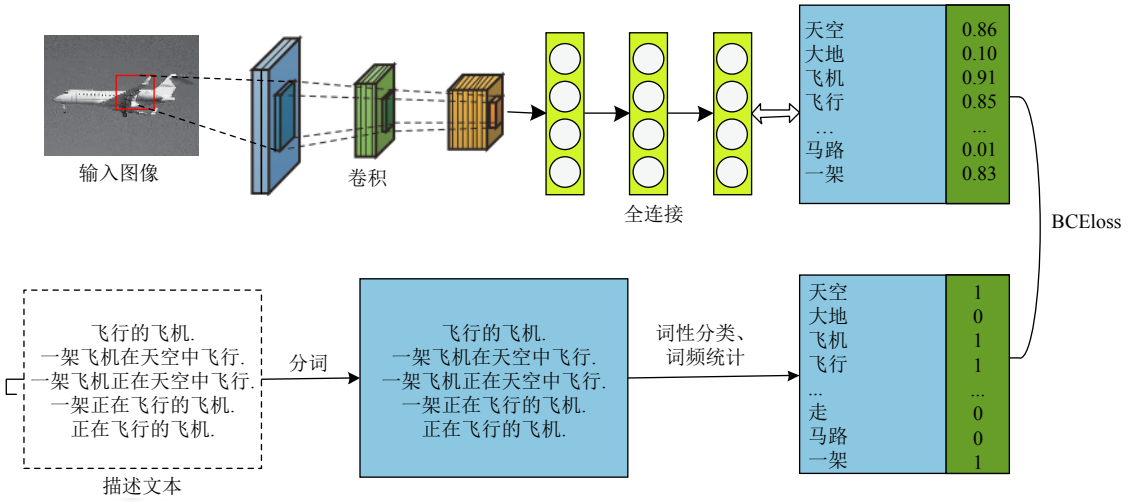

图 2 多标签文本生成网络训练过程

对图像描述数据集进行预处理, 得到图像多标签数据集, 将微调后的 AlexNet 网络在图像多标签数据集上进行训练, 将训练后模型的参数迁移到本文的多标签文本生成网络中. 对于输入图像 I_i, 将其缩放到 256×256 个像素后, 再利用多标签文本生成网络输出概率编码的图像多标签, 最后通过设置阈值输出图像对应的多标签文本 $W(I_i)$.

由图 1 可知, 一幅图像 I_i, 模型提取其视觉特征 $V(I_i)$ 和多标签文本 $W(I_i)$ 后, 将其输入到双注意力的解码网络中, 由双注意力解码网络对视觉特征进行解码生成描述词语.

3) 双注意力解码网络

由于中文词语的含义较为丰富, 因此合理地利用视觉特征对于图像描述生成尤为重要. 本文模型在解

码器中引入注意力机制, 使解码器可以根据 LSTM 内部的隐藏状态 h_t, 加权出与当前输出词语关联度高的视觉特征, 进而对加权后的视觉特征进行解码生成描述词语. 本文的双注意力解码网络工作流程解码流程如图 3 所示.

双注意力解码网络首先利用视觉特征 $V(I_i)$ 和解码器上一次输出词语更新 LSTM 内部的隐藏状态 h_t, LSTM 内部更新公式如式 (4) 所示:

$$\begin{cases} x_t = E_w W_{t-1} \text{ for } t > 0 \\ i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \\ f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \\ o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \\ c_t = i_t \odot \phi(W_c x_t + W_I I_i + U_c h_{t-1} + b_c) + f_t \odot c_{t-1} \\ h_t = o_i \odot \tanh(c_t) \end{cases} \tag{4}$$

其中, E_w 是词嵌入矩阵, W_{t-1} 是 LSTM 的上一次的输出词语, h_{t-1} 是 LSTM 上一次的隐藏状态, x_t 是 LSTM 当前的输入, σ 是指 Sigmoid 激活函数, f、i、o 分别表示 LSTM 内部是否忘记此前信息、是否接受新的输入以及是否输出当前信息的“闸门”, W、U 和 b 是 LSTM 结构中需要训练的模型参数, ⊙表示对应向量与闸门取值的乘积, c_t 是 LSTM 当前的记忆单元状态, h_t 是 LSTM 当前的隐藏状态.

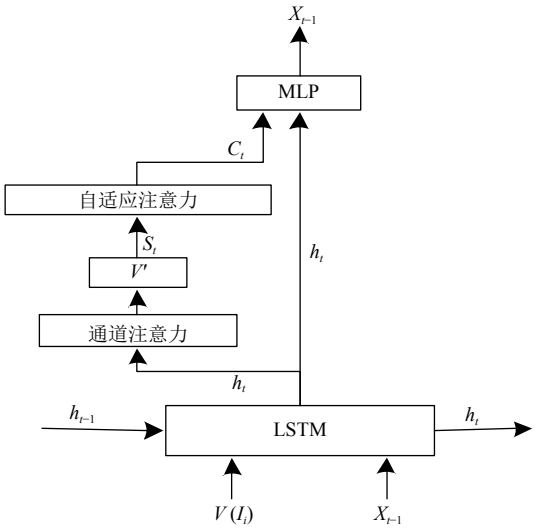

图 3 双注意力的解码网络工作流程

由图 3 可知, 在 LSTM 内部的隐藏状态 h_t 更新后, 双注意力解码网络根据 LSTM 内部的隐藏状态 h_t 利用通道注意力机制加权出与当前输出关联度较高的视觉特征 $V'(I_i)$. 通道注意力机制从特征通道的角度分析与不同通道的视觉特征与当前输出词语的关联度, 从而降低特征通道层的误差干扰, 其内部的数学计算为式 (5) 所示:

$$\begin{cases} V'(I_i) = \sum_{i=1}^{B} \beta_t^i V(I_i) \\ b = \tanh((W_a \otimes V + b_a) \oplus W_{h_t} h_t) \\ \beta = Softmax(W_b b + b_b) \end{cases} \tag{5}$$

其中, B 是视觉特征的通道数, tanh 和 *Softmax* 为激活函数, b_a、b_b、W_a、W_{ht}、W_b 是网络要学习的参数, V 表示视觉特征的每一个通道平均池化后的通道特征, h_t 表示 LSTM 在 t 时刻的隐藏状态, β 的每个值表示每个通道特征的权重, ⊙表示逐元素相乘, ⊕表示逐元素相加.

如图 3 所示, 双注意力解码网络加权出与当前输出词语关联度高的视觉特征 $V'(I_i)$ 后, 利用自适应注意力机制计算视觉特征 $V'(I_i)$ 与当前输出词语的视觉关联度. 自适应注意力机制利用视觉监督向量 s_t 对 LSTM 进行扩展, 视觉监督向量 s_t 通过对已经生成的文本信息和当前输入的视觉特征进行建模, 分析解码器输出的当前词语是否需要关注视觉特征. 当模型生成非语义词语时, 可以通过视觉监督向量 s_t 直接生成, 而不需要再关注图像特征信息. 自适应注意力解码网络内部的计算为:

$$\begin{cases} g_t = \sigma(E_w x_t + U_t h_t) \\ s_t = g_t \odot \tanh(o_t) \end{cases} \tag{6}$$

其中, g_t 表示 LSTM 内部记忆单元 o_t 中的候选状态, σ 是指 Sigmoid 激活函数, E_w 是词嵌入矩阵, x_t 表示在 t 时刻 LSTM 网络的输入单词.

最后双注意力机制解码网络将原有的上下文向量 c_t 与视觉监督向量 s_t 进行加权生成一个新的上下文向量c_t'.

$$c_t' = \alpha_t s_t + (1 - \alpha_t) c_t \tag{7}$$

其中, 参数 α_t 的取值范围为 0 到 1 之间. 从式 (7) 中可以看出, 当 α_t=1 时, 新的上下文向量c_t'为视觉监督向量 s_t, 此时双注意力解码网络只需利用已生成的文本信息可以预测下一个词语; 反之, 当 α_t=0 时, 模型更关注视觉特征信息生成下一个单词.

为使双注意力解码网络更好地解码视觉特征, 本文利用多标签文本 $W(I_i)$ 初始化 LSTM, 增强 LSTM 内部的隐藏状态, 初始化方式如式 (8) 所示:

$$\begin{cases} x_{-1} = W_v V(I_i) + E_w W(I_i) \\ h_{-1} = W_w W(I_i) \\ c_{-1} = W_w W(I_i) \\ h_0, c_0 = LSTM(x_{-1}.h_{-1}, c_{-1}) \end{cases} \tag{8}$$

其中, W_v, W_w 是模型需要学习的参数, $V(I_i)$ 是视觉特征, $W(I_i)$ 是多标签文本.

4) 中文描述生成

在 MLP 层利用 *Softmax* 函数将c'_t与词表 D 建立映射连接:

$$y_t \sim p_t = Softmax(W_t c'_t) \tag{9}$$

其中, y_t 是指在 t 时刻 LSTM 的输出, p_t 是 MLP 对词表 D 中不同单词的预测概率, $Softmax$ 是激活函数, W_t 是网络的学习参数.

3 实验

本节对本文方法的实验环境与具体参数设置进行介绍, 并结合实验对本文模型进行分析.

1) 实验环境

本文实验在深度学习服务器上运行, 显卡其型号是 NVIDIA 1070Ti, 内存大小为 8 GB. 数据的预处理过程与模型的训练和测试过程均在 Python3、PyTorch 0.4 上进行.

2) 数据集

本文在 Flickr8k-CN[11]、COCO-CN[18] 两个图像中文描述数据集上进行实验.

Flickr8k-CN[11] 是首个图像中文描述数据集, 数据集中的图像大多来源于人类真实生活场景, 且图像中的描述目标较为显著. 该数据集中共有 8000 张图像, 其中每幅图像对应 5 个描述文本, 每个描述文本从不同的角度描述图像的内容, 其中训练集 6000 张图像, 验证集 1000 张图像, 测试集 1000 张图像.

COCO-CN[18] 数据集中图像的场景更为多样化, 图像中的干扰元素更多, 每幅图像对应的描述文本由 1 个到 5 个不等, 该数据集共有 20341 张图像, 其中训练集 18341 张图像, 验证集 1000 张图像, 测试集 1000 张图像. 数据集构成与示例如表 1 所示.

3) 实验设置

由于中文句子缺乏自然分隔符, 本文利用 THULAC[19] 中文分词工具对数据集中的描述句子进行分词. 为避免罕见单词不利于描述文本生成, 本文统计词频大于 5 的词语, 并且增加“< start> ”表示句子开始、“< end > ”表示句子结尾、“< UNK> ”表示未知单词、“< PAD> ”表示填充单词 4 个具有特殊意义的单词, 建立中文词表 D. 利用词表 D 对数据集中的文本进行向量化. 表 2 给出了不同的数据集对应的词表 D 大小.

① 多标签文本生成网络参数设置

利用词表 D 对图像中文描述数据集中的出现频率大于 5 的名词、动词进行映射, 得到图像中文多标签数据集. 使用在 ImageNet 数据集上预训练的 AlexNet 网络参数对多标签文本生成网络进行初始化. 在网络的训练过程中, 输入的图像分辨率设置为 256×256 个像素, 学习率大小设置为 0.001. 为避免过拟合, 训练过程中采用 Dropout 对网络的隐藏输出采样. 本文将多标签文本生成网络输出概率较大的作为该图像的多标签文本, 通过在验证集上进行搜索, 选取在验证集上取得最好效果为 0.9.

表 1 数据集构成与示例

数据集	构成		数据集中图像	数据集中图像对应的描述文本
Flickr8k-CN	描述	数量		棕色和白色的狗在雪地上运行./狗在雪地上运行./狗通过雪地上运行./白色和棕色的狗通过一个白雪覆盖的田野上奔跑./白色和棕色的狗正在运行雪的表面上.
	总数	8000		
	训练集	6000		
	验证集	1000		
	测试集	1000		
COCO-CN	描述	数量		许多公交车排成队在广场上停着.
	总数	20341		
	训练集	18341		
	验证集	1000		
	测试集	1000		

② 融合双注意力与多标签模型的参数设置

将双注意力解码网络的 LSTM 的隐藏层维度设置为 512, 利用 Adam 优化器优化模型的误差, 批训练样本的大小设置为 32. 为了避免过拟合, 采用 Dropout 对

网络的隐藏输出采样. 在测试阶段采用了集束搜索策略, beam_size 大小为 1.

表 2 不同数据集合对应的词表 D 大小

数据集	标签数量
Flickr8k-CN	1447
COCO-CN	2069

③ 模型的评价指标设置

本文使用的评价指标为: BLEU[20]: 机器翻译的评价指标, 能够分析机器生成语句和参考语句间的 N 元文法准确率, 根据 N 元文法的选择该指标有 BLEU-1、BLEU-2、BLEU-3、BLEU-4 被广泛使用. METEOR[21]: 利用单精度的加权调和平均数和单字召回率的方法改善 BLEU 指标存在的问题. ROUGE[22]: 通过比较召回率的相似度来度量指标.

4) 实验结果与分析

① 实验 1. 数据集上模型效果对比

实验中选择 CS-NIC[8]、软注意力机制 (Soft-ATT)[11]、自适应注意力机制 (Adaptive)[23]、通道注意力机制 (SCA-CNN)[24] 作为对比. 其中 CS-NIC 是作为首个图像中文描述模型有重要的参考价值; 软注意力机制、通道注意力机制与自适应注意力机制在图像英文描述研究中是有效的, 为了验证注意力机制是否能应用于中文环境, 本文将在中文的环境下对注意力模型重新训练. 表 3 与表 4 是以上模型在 Flickr8k-CN 和 COCO-CN 数据集上的表现.

表 3 不同模型在 Flickr8k-CN 数据集上的表现

模型	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGLE
CS-NIC[8]	0.611	0.411	0.221	—	—	—
Soft-ATT[11]	0.701	0.515	0.404	0.313	0.321	0.604
SCA-CNN[23]	0.479	0.316	0.175	0.086	0.199	0.480
Adaptive[24]	0.697	0.506	0.392	0.303	0.321	0.603
本文(Ours)	**0.715**	**0.532**	**0.410**	**0.314**	**0.314**	**0.605**

表 4 不同模型在 COCO-CN 数据集上的表现

模型	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGLE
CS-NIC[8]	—	—	—	—	—	—
Soft-ATT[11]	0.387	0.251	0.174	0.124	0.209	0.387
SCA-CNN[23]	0.177	0.01	—	—	0.014	0.167
Adaptive[24]	0.375	0.234	0.155	0.107	0.200	0.368
本文(Ours)	**0.392**	**0.239**	**0.177**	**0.127**	**0.198**	**0.384**

图 4 是对表 3 的可视化. 从图 4 中可知, 相比目前的主流的图像中文描述模型 CS-NIC, 本文的模型通过融合双注意力机制与图像多标签文本, 在 BLEU-1、BLEU-2、BLEU-3 上均有提升, 这证明本文提出的模型是有效的. 具体地本文模型相比 CS-NIC 模型, 在 BLEU-1 上的提升 10. 3%, 在 BLEU-2 上的提升 12. 1%, 在 BLEU-3 上的提升 17. 8%; 此外, 将注意力机制在中文的环境下重新训练后, 相比 CS-NIC 模型来说在不同的评价指标上, 也均有一定的提升, 这说明注意力机制可以应用到中文环境; 另外, 相比自适应注意力机制和通道注意力机制, 本文模型在 BLEU 评价指标上也有提升, 这一点在表 3 和表 4 中均有所体现, 这说明通过利用多标签文本初始化双注意力解码网络, 可以生成更高质量的图像描述句子.

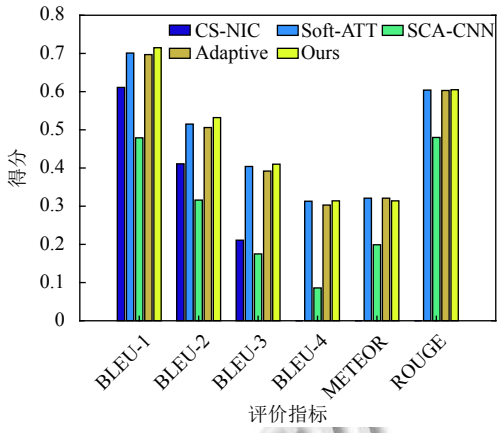

图 4 Flickr8k-CN 数据集上不同模型在评价指标上的得分

② 实验 2. Flickr8k-CN 数据集上消融实验

表 5 给出了本文模型不同组成部分对模型提升贡献度. 通过表 5 可知, 相比自适应注意力机制, 本文通过融合通道注意力机制, 在 BLEU-1 上提升 0. 1%, 在 BLEU-2 上提升 0. 9%, 在 BLEU-3 上提升 0. 7%, BLEU-4 上提升 0. 9%, 这说明降低在视觉通道特征中误差因素的干扰, 模型可以生成更高质量的描述句子; 本文模型利用多标签文本初始化 LSTM 内部的隐藏状态, 在 BLEU-1 上提升 0. 4%, 在 BLEU-2 上提升 0. 4%, 在 BLEU 上提升 0. 3%, 在 BLEU-4 上提升 0. 9%, 这验证了通过利用多标签文本初始化 LSTM 内部的隐藏状态, 可以提升图像中文描述模型的效果.

5) 可视化实例分析

根据图像中描述对象的类型和数量进行分类, 本文将数据集中的描述场景分为 3 种类型, 即单类单目

标场景、单类多目标场景和多类多目标场景. 表 6 是软注意力机制 (Soft-ATT)、自适应注意力机制 (Adaptive)、通道注意力机制 (SCA-CNN) 以及本文模型对不同场景的图像的描述效果. 从表 6 可看出, 本文提出的模型对图像中物体的识别和语义的理解还是比较准确的.

表 5 不同模型在 Flickr8k-CN 数据集上的消融实验

模型	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGLE
SCA-CNN	0.479	0.316	0.175	0.086	0.199	0.480
Adaptive	0.697	0.506	0.392	0.303	0.321	0.603
融合双注意力	0.698	0.515	0.399	0.312	0.313	0.598
融合多标签文本	0.701	0.510	0.395	0.312	0.323	0.603
本文	**0.715**	**0.532**	**0.410**	**0.314**	**0.314**	**0.605**

表 6 不同场景下对比模型生成的图像中文描述文本

场景	输入图像	数据集中的描述	不同模型生成的描述	
单类单目标场景		一个小孩穿着泳衣躺在水里./女孩躺在海边./女人穿比基尼躺在海水中./一个女人在水里躺着./一个穿着泳装的女人躺在海水里.	Soft-ATT	穿着泳衣的女孩.
			SCA-CNN	一个小女孩在玩水
			Adaptive	躺在地上的小女孩.
			Ours	在水中的小女孩.
		徒步旅游的男人们./趴着的狗./一条狗趴在沙子上./狗狗在雪地上./一只在雪地中的金毛犬.	Soft-ATT	两只狗在雪地里玩耍.
			SCA-CNN	一只狗在<UNK>.
			Adaptive	一只狗在雪地里.
			Ours	一只狗在雪地里.
单类多目标场景		坐着的人和站着的人./坐在路边休息的情侣./坐在路边的人们./坐在台阶上的情侣./两个坐在地上的男女和一个站着的男人.	Soft-ATT	坐着的男人.
			SCA-CNN	坐在台阶上的人们.
			Adaptive	坐在台阶上的两个人.
			Ours	坐在台阶上的男人.
		两只狗在玩耍./两只小狗在打架./两只狗在草坪玩耍./两只小狗在玩耍.两只狗在草地上玩耍.	Soft-ATT	两只狗在玩耍.
			SCA-CNN	两只狗在玩耍.
			Adaptive	两只狗在玩耍.
			Ours	两只狗在草地上玩耍.
多目标场景		一个小孩在训狗./1个男孩和1只狗在玩耍./一只小狗咬着一个小男孩拿着的木棍./一只狗正在咬着一个小男孩拿着的棍子./一只狗正在咬一根木头.	Soft-ATT	一个人在玩球.
			SCA-CNN	两只狗在玩耍
			Adaptive	一个女人在训狗.
			Ours	一个人和一只狗在玩耍.

4 总结与展望

为提升图像中文描述句子质量, 本文在验证注意力机制可用于图像中文描述生成的基础上, 提出融合双注意力与图像多标签的图像中文描述生成方法. 通过在图像中文描述数据集上进行评测, 在多个图像描述评价指标上优于目前主流的图像中文描述生成模型. 然而本文模型所使用的注意力机制是英文环境下的注意力机制迁移而来的, 由于中文与英文语法的差异, 因此结合中文语法规则设计出符合中文环境的注意力机制是该领域的目标. 此外, 在对不同场景的图像分析过程中, 本文模型在单类单目标场景和多类多目标场景下, 生成的描述句子更符合图像本身的内容, 语义也更为饱满, 但是对于多类单目标的场景, 本文模型生成的图像中文描述句子容易只描述出图像中的部分区域, 因此在未来的工作中会专注于提升模型对图像全局语义的理解能力.

参考文献

1 王晓光, 徐雷, 李纲. 敦煌壁画数字图像语义描述方法研究. 中国图书馆学报, 2014, 40(1): 50–59. [doi: 10.3969/j.issn.1001-8867.2014.01.005]

2 Farhadi A, Hejrati M, Sadeghi MA, *et al*. Every picture tells a story: Generating sentences from images. Proceedings of the 11th European Conference on Computer Vision. Heraklion, Greece. 2010. 15–29.

3 Fang H, Gupta S, Iandola F, *et al*. From captions to visual concepts and back. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 1473–1482.

4 Ordonez V, Kulkarni G, Berg TL, *et al*. Im2Text: Describing images using 1 million captioned photographs. Proceedings

of Advances in Neural Information Processing Systems. Granada, Spain. 2011. 1143–1151.
5 Vinyals O, Toshev A, Bengio S, *et al*. Show and tell: A neural image caption generator. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 3156–3164.
6 Jia X, Gavves E, Fernando B, *et al*. Guiding the long-short term memory model for image caption generation. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile. 2015. 2407–2415.
7 汤鹏杰, 谭云兰, 李金忠. 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, 2017, 22(9): 1251–1260. [doi: 10.11834/jig.170052]
8 Xu K, Ba JL, Kiros R, *et al*. Show, attend and tell: Neural image caption generation with visual attention. Proceedings of the 32nd International Conference on Machine Learning. Lille, France. 2015. 2048–2057.
9 Li LH, Tang S, Deng LX, *et al*. Image caption with global-local attention. Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, CA, USA. 2017. 4133–4139.
10 Huang L, Wang WM, Xia YX, *et al*. Adaptively aligned image captioning via adaptive attention time. Proceedings of Advances in Neural Information Processing Systems. Vancouver, BC, Canada. 2019. 8942–8951.
11 Li XR, Lan WY, Dong JF, *et al*. Adding Chinese captions to images. Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval. New York, NY, USA. 2016. 271–275.
12 张凯, 李军辉, 周国栋. 基于枢轴语言的图像描述生成研究. 中文信息学报, 2019, 33(3): 110–117.
13 蓝玮毓, 王晓旭, 杨刚, 等. 标签增强的中文看图造句. 计算机学报, 2019, 42(1): 136–148.
14 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 770–778.
15 Krizhevsky A, Sutskever I, Hinton GE, *et al*. ImageNet classification with deep convolutional neural networks. Proceedings of Advances in Neural Information Processing Systems. Lake Tahoe, CA, USA. 2012. 1097–1105.
16 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]
17 Liu Y, Dai YT, Doan AD, *et al*. In defense of OSVOS. arXiv preprint arXiv: 1908.06692, 2019.
18 Li XR, Xu CX, Wang XX, *et al*. COCO-CN for cross-lingual image tagging, captioning, and retrieval. IEEE Transactions on Multimedia, 2019, 21(9): 2347–2360. [doi: 10.1109/TMM.2019.2896494]
19 Li ZG, Sun MS. Punctuation as implicit annotations for Chinese word segmentation. Computational Linguistics, 2009, 35(4): 505–512. [doi: 10.1162/coli.2009.35.4.35403]
20 Papineni K, Roukos S, Ward T, *et al*. BLEU: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, PA, USA. 2002. 311–318.
21 Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Ann Arbor, MI, USA. 2005. 228–231.
22 Lin CY. ROUGE: A package for automatic evaluation of summaries. Proceedings of the Workshop on Text Summarization Branches Out. Barcelona, Spain. 2004. 74–81.
23 Lu JS, Xiong CM, Parikh D, *et al*. 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, HI, USA. 2017. 3242–3250.
24 Chen L, Zhang HW, Xiao J, *et al*. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 6298–6306.