结合全局和局部特征的 BiGRU-RA 图像中文描述模型

邓珍荣^{1,2)}, 张永林²⁾, 杨睿²⁾, 蓝如师^{1,2)*}, 黄文明^{1,2)}, 罗笑南^{1,2)}

¹⁾ (广西图像图形与智能处理重点实验室 桂林 541004)

²⁾ (桂林电子科技大学计算机与信息安全学院 桂林 541004)
(rslan@guet.edu.cn)

摘要: 针对目前基于全局特征的图像描述模型存在细节语义信息不足的问题, 提出结合全局和局部特征的图像中文描述模型. 该模型采用编码器-解码器框架, 在编码阶段, 分别使用残差网络(residual networks, ResNet)和 Faster R-CNN 提取图像的全局特征和局部特征, 提高模型对不同尺度图像特征的利用. 采用嵌入了残差连接结构和视觉注意力结构的双向门控循环单元(bi-directional gated recurrent unit, BiGRU)作为解码器(BiGRU with residual connection and attention, BiGRU-RA). 模型可以自适应分配图像特征和文本权重, 改善图像特征区域和上下文信息的映射关系. 此外, 加入基于强化学习的策略梯度对模型的损失函数进行改进, 直接对评价指标 CIDEr 进行优化. 在 AI Challenger 全球挑战赛图像中文描述数据集上进行训练和实验, 实验结果表明, 该模型获得更高的评分, 生成的描述语句更准确、更详细.

关键词: 图像描述; 双向门控循环单元; 视觉注意力; 强化学习; 残差连接
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2021.18262

BiGRU-RA Model for Image Chinese Captioning via Global and Local Features

Deng Zhenrong^{1,2)}, Zhang Yonglin²⁾, Yang Rui²⁾, Lan Rushi^{1,2)*}, Huang Wenming^{1,2)}, and Luo Xiaonan^{1,2)}

¹⁾ (Guangxi Key Laboratory of Image and Graphic Intelligent Processing, Guilin 541004)

²⁾ (School of Computer and Information Security, Guilin University of Electronic Technology, Guilin 541004)

Abstract: To address the problem of insufficient detailed semantic information in current global features-based image captioning models, an image Chinese captioning model combining global and local features is proposed. The proposed model adopts the encoder-decoder framework. In the coding stage, the residual networks (ResNet) and Faster R-CNN are used to extract the global and local features of images respectively, improving the model's utilization of image features at different scales. A bi-directional gated recurrent unit (BiGRU) with embedded visual attention structure and residual connection structure is applied as the decoder (BiGRU with residual connection and attention, BiGRU-RA). The model can adaptively allocate image features and text weights, and improve the mapping relationship between image feature regions and context information. Additionally, the reinforcement learning-based policy gradient is added to improve the loss function of the model and optimize the evaluation criteria CIDEr directly. The training and experiments are conducted on the Chinese captioning dataset of AI challenger. The comparative results show that the proposed model obtained better

收稿日期: 2020-03-04; 修回日期: 2020-09-30. 基金项目: 国家自然科学基金(61772149, 6202780103, 61762028); 国家重点研发计划(2018AAA0100300); 广西科技计划(AB20238013, ZY20198016, 2019GXNSFAA245014, AD18281079); 广西图像图形与智能处理重点实验室项目(GIIP2003); 桂林电子科技大学研究生教育创新计划(2020YCXS049). 邓珍荣(1977—), 女, 硕士, 研究员, 硕士生导师, 主要研究方向为图像处理; 张永林(1995—), 男, 硕士研究生, 主要研究方向为图像处理、深度学习; 杨睿(1996—), 女, 硕士研究生, 主要研究方向为计算机视觉; 蓝如师(1986—), 男, 博士, 副教授, 硕士生导师, 论文通讯作者, 主要研究方向为图像处理、机器学习; 黄文明(1963—), 男, 教授, 硕士生导师, 主要研究方向为大数据处理、图像处理; 罗笑南(1963—), 男, 博士, 教授, 博士生导师, 主要研究方向为图形学、数字图像处理.

scores and the generated caption are more accurate and detailed.

Key words: image captioning; bi-directional gated recurrent unit; visual attention; reinforcement learning; residual connection

随着智能拍照设备的普及和信息技术的高速发展,图像成为人们获取信息的重要途径.图像中包含众多信息,能够形象生动地表达出信息的内容.而面对大量的图像,如何让计算机代替人工理解图像信息,已成为计算机视觉领域的研究热点.图像的文本描述是融合计算机视觉和自然语言处理领域的交叉任务,能够完成从图像到文本的多模态转换^[1].通过对图像进行语义分析和理解,得到图像中各个实体对象信息和实体对象之间的联系,最终生成准确且符合人类语言习惯的描述语句.图像描述技术具有巨大的应用前景,在图像检索^[2]、人机交互系统、图像视频标注、辅助教育和自动驾驶等多方面具有重要的实用价值.

尽管图像描述任务十分具有挑战性,但仍然引起了国内外学者的广泛关注.根据关键技术和文本描述方法的不同,图像文本描述方法一般分为3类:基于模板匹配的方法、基于检索的方法和基于深度学习的方法^[3-6].(1) 基于模板匹配的方法包括对象检测和描述生成2个过程.其在对象检测过程通过目标检测和属性特征检测算法识别出图像中包含的实体对象和对象之间的属性关系等;在描述生成过程将识别的单词填入场景模板的相应位置,形成该图像对应的描述语句.该方法的不足之处是局限于固定模板,生成的文本描述句式结构过于单一.(2) 基于检索的方法首先通过检索数据库中与输入图像相似的图像集,然后通过分析和重组该图像集的文本描述来形成输入图像的描述语句;但其描述结果依赖于句子池.(3) 基于深度学习的方法将机器翻译任务中的编解码思想应用到图像文本描述任务中,主要通过深度卷积神经网络对图像进行编码,利用循环神经网络建立语言模型,将图像信息和文本信息映射到循环神经网络中,进而利用图像信息指导描述语句的生成.该方法能够明确地学习图像与文本之间的对应关系,生成的描述语句具有多样性,可以广泛适用于任何场景.

虽然基于深度学习的图像描述方法已取得非常好的结果,但仍存在以下问题.(1) 图像特征提取不够充分,容易导致解码模型忽略图像的细节信息;传统的语言模型往往只考虑正向的上下文

信息,而序列的生成同样依赖反向的上下文信息,使得生成的描述文本不够详细.(2) 目前的图像描述任务主要是针对英文数据集,基于中文数据集的图像描述任务还没有取得很大的进展.

本文针对上述图像描述模型存在的细节语义信息利用不足、描述文本不够细致问题,提出结合全局和局部特征的 BiGRU-RA 图像中文描述模型,其采用编码器-解码器框架.在编码阶段,采用 ResNet-101 提取图像的全局特征,采用 Faster R-CNN 提取图像的局部特征.为了增强模型对上下文信息的利用,避免神经网络随着深度的增加引发的退化问题,采用基于残差连接的双向门控循环单元(bi-directional gated recurrent unit, BiGRU)作为解码器.在解码器中引入视觉注意力机制,通过自适应分配图像特征和文本权重改善图像特征区域和上下文信息的映射关系.此外,加入基于强化学习的策略梯度对模型的损失函数进行改进,直接对评价指标进行优化.

1 相关工作

随着深度学习的发展,基于深度学习的方法逐渐成为图像描述任务的主流方法,不断推动图像描述生成技术的发展. Mao 等^[6]提出首个基于神经网络的多模态图像描述生成模型,其使用卷积神经网络(convolutional neural networks, CNN)对图像建模,使用循环神经网络(recurrent neural networks, RNN)对描述文本建模,并使用多模态空间为图像和文本建立映射关系. Vinyals 等^[7]率先将编码器-解码器框架应用于图像描述任务,通过预训练好的 CNN 将图像编码成能够表征图像内容的特征,然后结合图像对应的训练文本提供的语义,输入到 RNN 将该特征解码成句子. Fang 等^[8]结合深度多模相似网络和最大熵语言模型生成图像的描述语句,在检测出图像中的属性信息后,利用属性信息替代原始图像信息输入到长短期记忆网络(long short-term memory, LSTM)语言模型生成描述语句. Xu 等^[9]在图像上引入注意力机制,将上下文信息引入编码器-解码器框架,在编码阶段,使用保留图像空间信息的较低层的卷积层作为图像特

征, 然后结合注意力机制将其用于解码阶段. 该方法可有效地提取图像的视觉信息, 以生成更加准确的描述语句. Xu 等^[10]利用语义信息指导 LSTM 在各个时刻生成描述文本. Yao 等^[11]通过多实例学习的方法对图像属性信息进行提取, 研究图像属性特征对描述结果的影响. Wu 等^[12]在编码器-解码器框架中引入高级语义属性, 把原问题转换为多标签分类问题, 先利用 VggNet 进行多标签的预训练, 再通过 CNN 产生多标签的预测结果, 将预测结果经过最大池化处理后输入到 LSTM 产生描述. Jiang 等^[13]通过添加称为引导网络的组件来对编码器-解码器框架进行扩展, 引导网络对输入图像的属性进行建模, 并利用其在每个时间步的输出构成解码器的输入. 考虑原始注意力机制强制将每一个单词对应到图像的某一区域, 而某些单词没有对应图像的意义, Lu 等^[14]提出哨兵机制, 在生成每个单词时先计算这个单词是属于视觉词还是上下文词的概率, 即与图像的关联性, 再根据权重计算总体的特征. Chen 等^[15]提出一种融合空间和通道注意的模型, 在多层特征图中动态地调制句子生成上下文, 对视觉注意的位置和注意通道进行编码. Dai 等^[16]利用对比学习的方法构造损失函数, 利用负样本来参与训练, 以提高模型的

独特性. Anderson 等^[17]提出一种自上而下与自下而上相结合的注意力机制, 前者用于提取图像区域特征; 后者用于学习特征所对应的权重. 郭淑涛等^[18]提出一种面向中文数据集的多模态神经网络模型, 采用深度神经网络和注意力机制提取图像特征, 并在 RNN 中引入记忆力助手引导句子的生成.

2 本文模型

本文模型采用编码器-解码器框架, 由图像特征提取和语言生成 2 部分组成, 如图 1 所示. 编码器的主要任务是提取输入图像的特征并进行编码, 捕获图像内容中的对象主体以及属性关系, 其分别使用 ResNet^[19]和 Faster R-CNN^[20]提取图像的全局特征和局部特征. 解码器由带有残差连接的 BiGRU 网络和视觉注意力结构(图 1 中的 ATT)组成, 其主要任务是对图像特征进行解码, 并将图像特征和输入的词嵌入向量建立映射关系, 输出词典中词的概率分布. 本文将编码器提取到的全局图像特征和局部图像特征分别输入视觉注意力结构和 BiGRU 网络, 得到输入图像的上下文信息, 通过解码器计算每个词的概率, 最终生成描述语句.

图 1 本文模型总体框架

2.1 编码器

编码器由 ResNet 和 Faster R-CNN 组成. ResNet 通过引入残差网络结构, 可以有效地避免梯度消失和网络退化问题. 具体做法是在输入和输出之间采用如图 2a 所示的快捷连接(shortcut). 通过增加这个恒等映射连接, 使原本需要学习的函数 $F(x)$ 转换为 $F(x)+x$. 该做法不会增加网络的参数和计算量, 反而能够提升模型的训练速度和训练效果. 为了降低参数的数目, ResNet 内部采用如图 2b 所示的瓶颈层(bottleneck)设计. 第 1 个

图 2 快捷连接和瓶颈层设计

1×1 的卷积把 256 维通道降到 64 维, 然后通过下一个 1×1 卷积恢复为 256 维.

本文使用预训练好的 ResNet-101 模型来提取图像的全局特征. 网络共有 100 个卷积层和 1 个全连接层, 卷积层被分为 5 组, 卷积核的大小分别为 7×7 , 1×1 和 3×3 . 为了得到全局的图像特征图, 把最后一层全连接层去掉. 第 1 组卷积的输入大小为 224×224 , 输出大小为 7×7 , 共有 2048 维. 图像经过 ResNet-101 后得到的全局特征可表示为 V_g .

Faster R-CNN 常被用于目标检测任务中, 通过在深度神经网络中引入区域建议网络(region proposal networks, RPN), 以提高目标检测准确率. 采用预训练好的 Faster R-CNN 提取图像的局部特征, 通过 RPN 找到感兴趣的区域(region of interest, RoI), 从而得到包含图像视觉属性的局部特征. Faster R-CNN 的结构如图 3 所示, 图像输入卷积层中进行特征提取, 得到图像的特征图, 然后利用 RPN 生成候选区域, 这一层利用 Softmax 来确定当前锚点是前景还是背景, 再通过边界框回归修正锚点, 从而得到精确的候选区域. RoI 池化层通过收集之前的特征图和候选区域来得到候选特征图, 最后的分类层将 RoI 池化层形成固定大小的特征图进行全连接操作, 通过 Softmax 进行分类. 同时, 利用 L_1 损失完成边界框回归操作, 以获得物体的精确位置.

图 3 Faster R-CNN 结构

为得到图像的局部特征, 本文将 Faster R-CNN 的分类层去掉, 直接把候选特征图作为局部特征 V_l , 提取图像局部特征的过程如图 4 所示. 首先将图像输入卷积层中得到卷积特征图, 然后经过 RPN

得到特征图对应的推荐窗口, 再对推荐窗口进行 RoI 池化操作, 利用双线性插值算法对特征图进行裁剪, 最后得到固定尺寸的特征图作为局部特征.

图 4 图像局部特征提取示意图

2.2 GRU 网络

门控循环单元(gated recurrent unit, GRU)^[21]网络是 LSTM^[22]的一种变体, 可以解决 RNN 中的长期依赖问题, 对长序列具有很好的学习能力. GRU 在保持了 LSTM 的效果同时又使结构更加简单, 因此在训练过程中, GRU 的参数更少, 收敛速度更快. GRU 的结构如图 5 所示, GRU 使用门控机制来跟踪序列的状态, 它有重置门和更新门 2 种类型. 重置门用来控制忽略上一时刻的状态信息的程度, 更新门用来控制上一时刻的状态信息被输入到当前状态中的程度. 在 t 时刻, GRU 的计算过程为

$$\begin{aligned} z_t &= \sigma(W_z \cdot [h_{t-1}, x_t]), \\ r_t &= \sigma(W_r \cdot [h_{t-1}, x_t]), \\ \tilde{h}_t &= \tanh(W \cdot [r_t * h_{t-1}, x_t]), \\ h_t &= (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t. \end{aligned}$$

其中, z_t 表示更新门; $\sigma(\cdot)$ 表示 Sigmoid 函数; r_t 表示重置门; $\tilde{h}_t$ 表示候选隐藏层; W_z, W_r, W 是需要学习的权重矩阵; $*$ 表示哈达玛积.

图 5 GRU 内部结构

在 GRU 对文本序列建模时, 每个 t 时刻的隐藏状态只能正向读取上下文信息, 无法反向读取; BiGRU 则可以利用并行通道, 同时读取前后方向的上下文信息. BiGRU 由 2 个方向相反的单向 GRU 组成, 如图 6 所示. 用 h_1 和 h_2 分别表示前向和后向 GRU 在 t 时刻隐藏状态的输出, 则双向 GRU 隐藏状态的输出可表示为 $h = [h_1, h_2]$.

略梯度对模型的损失函数进行改进, 直接对评价指标进行优化. 把描述文本生成当作强化学习的过程, 智能体为 GRU 网络, 环境为生成词和图像特征, 模型的参数 θ 定义了一个策略 p_θ , 根据策略会产生相应的动作, 即生成预测词. 在动作结束后, 模型就会更新 GRU 的细胞状态和隐藏状态, 智能体通过环境的变化来获得奖励, 该奖励设置为 CIDEr 的得分. 通过最小化奖励的负期望来选择动作, 此时, 损失函数定义为

$$\nabla_\theta L_\theta = -E_{w_s \sim p_\theta} [r(w_s) \nabla_\theta \log_{p_\theta}(w_s)].$$

其中, w_s 表示生成的句子; r 表示 CIDEr 得分. 为了加快模型的收敛, 加入一个基线值, 并在采样句子时采用集束搜索的方式, 即

$$\nabla_\theta L_\theta = -E_{w_s \sim p_\theta} [(r(w_s) - b) \nabla_\theta \log_{p_\theta}(w_s)].$$

其中, $b = r(\hat{w})$; $\hat{w}$ 表示通过集束搜索得到的句子.

3 实验结果与分析

3.1 数据集和实验环境

本文的实验数据集为 AI Challenger 全球挑战赛图像中文描述数据集. 该数据集的图像包含丰富的场景, 是目前规模最大的中文图像描述数据集. 数据集包含 30 万幅图像, 每幅图像对应有 5 句中文描述, 一共有 150 万句中文描述. 其中, 训练集有 21 万幅, 验证集有 3 万幅, 测试集有 6 万幅.

本文实验环境基于 Pytorch 深度学习框架, Ubuntu 14.04.1 操作系统, Intel(R) Xeon(R) E5-2698 v4 @ 2.20GHz CPU, 512GB 内存, NVIDIA Tesla P100-SXM2 16GB 显卡.

3.2 模型分析及参数设置

本文模型采用编码器-解码器框架, 在编码阶段, 通过 ResNet-101 的最后一层卷积层提取图像的全局特征, 图像通过 100 个卷积层后, 输出图像特征大小为 7×7 , 共有 2048 维. 通过 Faster R-CNN 提取图像的局部特征. 具体做法如下. 先将图像输入到卷积层中得到卷积特征图, 然后经过 RPN 得到特征图对应的推荐窗口; 再对推荐窗口进行 RoI 池化操作, 得到固定尺寸的特征图作为图像的局部特征. 对于中文的描述语句, 本文采用 jieba 分词对文本进行分词, 通过逐个对比文本和词库中出现的字词实现分词, 分词效果比较稳定, 避免了中文语句的语义歧义带来的不确定性. 然后统计出高频词, 设置句子最大长度为 40, 用“start”表示句子开始, “end”表示句子结束, “un-known”表示未知

词; 最后利用 word2vec 对分词进行编码. 解码器由带有残差连接的 BiGRU 网络和视觉注意力结构组成. 图像的全局特征输入到视觉注意力结构中, 由注意力自动分配特征权重. 图像的局部特征和文本向量输入到第 1 层 GRU 网络中, 使模型关注更多的视觉属性信息. 通过残差连接, 减少模型的参数计算量, 避免信息丢失. 在模型训练过程中, GRU 隐藏神经元个数设置为 512, 学习率设置为 10^{-4} . 同时采用 dropout 方法防止网络过拟合, 其值设置为 0.5. 批大小设置为 64, epoch 设置为 10k. 解码过程中采用 Adam 算法进行优化; 模型训练完成后, 加入策略梯度优化损失函数对模型进行微调, 从而得到最终的图像中文描述模型.

3.3 评价方法

图像描述的评价指标主要有 BLEU^[24], METEOR^[25], ROUGE^[26] 和 CIDEr. BLEU 通过计算预测语句和参考语句之间 N 元词共现的程度, 衡量这 2 个句子的相似程度. 加权平均后的相似度得分计算公式为

$$B_N(C, S) = b(C, S) \exp \left(\sum_{n=1}^N \omega_n \ln C_n(C, S) \right).$$

其中,

$$C_n(C, S) = \frac{\sum_i \sum_k \min(h_k(c_i), \max_{j \in m} h_k(s_{ij}))}{\sum_i \sum_k h_k(c_i)}$$

表示准确度指标, 与句子的长度有关; $h_k(c_i)$ 表示 N 元词在预测语句中出现的次数; $h_k(s_{ij})$ 表示 N 元词在参考语句中出现的次数; $b(C, S) = \begin{cases} 1, & l_c > l_s \\ e^{1-l_s/l_c}, & l_c \leq l_s \end{cases}$ 表示惩罚项; l_c 表示预测语句 c_i 的长度; l_s 表示参考语句 s_{ij} 的长度.

METEOR 在计算 2 个句子相似程度的同时, 考虑了同义词等相关信息, 使用同义词库作为辅助信息, 计算生成描述句在参考描述句上的准确率和召回率的调和平均. METEOR 评价得分计算公式为 $M_{\text{score}} = (1 - b_{\text{penalty}}) F_{\text{mean}}$. 其中,

$b_{\text{penalty}} = \gamma \left(\frac{c_h}{m} \right)^\theta$ 表示惩罚系数; $F_{\text{mean}} = \frac{PR}{\alpha P + (1 - \alpha)R}$ 表示准确率和召回率的调和平均; α , γ 和 θ 均为用于评价的默认参数; m 表示预先给定的一组校准; c_h 表示语句中连续有序的块;

$P = \frac{|m|}{\sum_k h_k(c_i)}$ 和 $R = \frac{|m|}{\sum_k h_k(s_{ij})}$ 分别表示准确率和召回率.

ROUGE 源自文本摘要, 是基于最长公共子序列(longest common subsequence, LCS)的一种测量方法. ROUGE 评价得分计算公式为

$$R_{\text{score}}(c_i, s_{ij}) = \frac{(1 + \beta^2)RP}{R + \beta^2 P}.$$

其中, $R = \max_j \frac{l(c_i, s_{ij})}{|s_{ij}|}$ 和 $P = \max_j \frac{l(c_i, s_{ij})}{|c_i|}$ 分别表示召回率和准确率; $l(c_i, s_{ij})$ 表示预测语句和参考语句的最长公共子序列长度; $\beta = R_l/P_l$ 表示权重系数, 当 β 很大时, 评价得分更加关注召回率; c_i 为预测语句; s_{ij} 为参考语句.

CIDEr 通过对每个 N 元组进行词频-逆文本频率 (term frequency-inverse document frequency, TF-IDF) 权重计算, 以此来衡量图像描述的一致性. CIDEr 是为图像描述任务而定制的, 其计算公式为

$$C(c_i, s_i) = \sum_{n=1}^N \omega_n C_n(c_i, s_i).$$

其中, $C_n = \frac{1}{m} \sum_j \frac{\mathbf{g}^n(c_i) \cdot \mathbf{g}^n(s_{ij})}{\|\mathbf{g}^n(c_i)\|_2 \cdot \|\mathbf{g}^n(s_{ij})\|_2}$ 表示相似度得分; s_i 表示参考语句集合; c_i 为预测语句; s_{ij} 为参考语句; $\mathbf{g}^n(c_i)$ 和 $\mathbf{g}^n(s_{ij})$ 为 TF-IDF 向量.

3.4 实验对比

为了验证模型的有效性, 本文在中文数据集上进行对比实验. 实验分为 2 个部分: 一个是在基准模型上加入不同策略的得分对比; 另一个是所提模型与当前主流模型的得分对比.

表 1 所示为在基准模型上加入不同策略的实验对比. 其中, Base 是基准模型, 即结合全局特征和视觉注意力的 BiGRU 模型; Base_L 表示基准模型加入局部特征; Base_R 表示基准模型加入残差连接; Base_P 表示基准模型加入策略梯度进行优化; Base_LR 表示基准模型加入局部特征和残差连接; Base_LRP 表示基准模型加入局部特征、残差连接和策略梯度. 通过实验对比结果可知, 在基准模型中分别加入 3 种策略后, 在 BLEU 和 CIDEr 指标上的得分均有显著的提高. 其中, Base_LRP 模型表现最好, 将其作为本文最终的模型.

表 1 不同策略的评分比较

策略	局部特征	残差连接	策略梯度	BLEU-1	BLEU-4	CIDEr
Base	×	×	×	0.733	0.463	1.837
Base_L	√	×	×	0.745	0.468	1.840
Base_R	×	√	×	0.751	0.476	1.857
Base_P	×	×	√	0.742	0.471	1.864
Base_LR	√	√	×	0.787	0.482	1.891
Base_LRP	√	√	√	0.791	0.487	1.920

注. 粗体表示最高评分.

表 2 所示为本文模型与当前主流模型的实验对比. 其中, NIC^[6]使用 GoogLeNet 提取图像特征, 使用 LSTM 作为语言生成模型; Adaptive^[13]在 LSTM 解码器中使用自适应的注意力机制; LSTM-A^[10]使用多实例学习提取图像属性特征, 并把属性特征作为 LSTM 的原始输入; Up-Down^[17]使用 Faster R-CNN 提取图像特征, 使用 2 层 LSTM 作为语言模型, 并引入自上而下的注意力; ILAM^[18]采用 Incepresv2 作为编码器, 采用 LSTM 作为解码器, 并加入注意力机制和记忆助手, 在图像中文数据集中进行实验. BiGRU-RA 表示本文模型. 其中, NIC, Adaptive, LSTM-A, Up-Down 模型在原始文献中均使用英文数据集. 英文和中文描述的主要区别在于语言的语法不同. 英文一句话中每个单词用空格隔开, 英文常用的分词方法为 word2vec, 而中文在分词的时候常常以一个词组为单位分开. 在对比实验中, 以上模型均采用同一个中文数据集, 采用 jieba 分词工具对文本进行切分. 在相同的环境下进行训练, 得到实验结果.

表 2 不同模型的评分比较

模型	METEOR ^[25]	ROUGE_L	CIDEr
NIC ^[6]	0.380	0.612	1.603
Adaptive ^[13]	0.392	0.631	1.735
LSTM-A ^[10]	0.396	0.673	1.810
Up-Down ^[17]	0.407	0.701	1.852
ILAM ^[18]	0.412	0.688	1.848
BiGRU-RA	0.413	0.709	1.920

注. 粗体表示最高评分.

3.5 实验分析

由表 1 可知, 模型中加入局部特征、残差连接和策略梯度后, 得分均有不同程度提升. 由表 2 可知, 本文模型的各项得分得到了提升, 其中 METEOR 得分为 0.413, ROUGE_L 得分为 0.709, CIDEr 得分为 1.920. 与 Adaptive 模型相比, 本文所提模型的

METEOR, ROUGE_L 和 CIDEr 得分相对提高了 5.4%, 12.3% 和 10.7%, 原因是在自适应注意力模型中加入残差连接起到了一定作用, 可以减少由网络层数增加带来的额外的信息丢失问题. 与 Up-Down 模型相比, 本文模型的 CIDEr 得分提升最大, 原因是加入策略梯度优化损失函数可以让模型寻找最优的参数优化指标, 使得模型的图像描述能力提升, 从而获得较高的评价分数. 与 ILAM 模型相比, 本文模型的 ROUGE_L 和 CIDEr 得分相对提高 3.1% 和 3.8%. 与使用单层 LSTM 作为解码器的 NIC 和 LSTM-A 模型相比, 本文所提模型的得分表现最好, 说明使用 BiGRU 网络作为解码器可以缓解 RNN 中的长期依赖问题.

为了进一步验证模型生成描述语句的性能, 选取测试集的图像进行对比实验. 选取的图像和实验生成的描述语句如图 9 所示, 展示的 6 幅图像包含了不同场景和不同的目标数量. 通过对比不同模型生成的描述文本可知, 本文所提模型的描

述效果更好, 可以更准确、更细致地描述出图像包含的场景、目标和目标之间的属性信息. 在图 9a 中, 本文模型不仅能识别出图像里有 1 个人在滑雪, 还能给物体加上一些形容词, 如“白雪皑皑”, 使生成的描述文本更加形象生动. 在图 9b 中, 虽然有 3 个对比模型都能描述出有 2 个人在踢球, 但本文模型生成描述的是“足球场上有 2 个穿着不同球衣的男人在抢球”, “穿着不同球衣”以及“抢球”更能体现出图像的细节信息. 在图 9d 中, NIC 模型生成的描述为“有 2 个人在打球”, 但实际上是 2 个人在进行拳击比赛, 而本文模型能够描述出擂台上有 2 个戴着拳套的男人在打拳击, 表述出了更多的视觉属性信息, 说明局部特征中包含的视觉属性信息对图像的整体描述生成有着很好的提升作用. 在图 9e 中, 本文模型不仅能够准确地识别出图像中的主体, 即小孩和狗, 还能描述出主体之间的位置关系, 即小孩在狗的旁边. 同样地, 本文模型生成的其他描述语句中包含有更多的修饰性的

NIC: 1 个人站在道路上
Adaptive: 室外有 1 个穿着黄色衣服的人
LSTM-A: 1 个穿着黄色衣服的人站在雪地上
Up-Down: 山上有 1 个人在滑雪
本文: 白雪皑皑的雪地上有 1 个戴着帽子的人在滑雪

a. 场景1

NIC: 有 2 个人站在草地上
Adaptive: 草地上有 2 个人在踢球
LSTM-A: 运动场上有 2 个人在踢球
Up-Down: 运动场上有 2 个男人正在踢球
本文: 足球场上有 2 个穿着不同球衣的男人在抢球

b. 场景2

NIC: 有 1 个人在水里
Adaptive: 水面下有 1 个人在游泳
LSTM-A: 有 1 个人在水里游泳
Up-Down: 泳池里有 1 个人在游泳
本文: 1 个穿着潜水服的人在海里潜水

c. 场景3

NIC: 有 2 个人在打球
Adaptive: 室内有 2 个人在打比赛
LSTM-A: 室内有 2 个男人在搏斗
Up-Down: 舞台上有 2 个男人在打拳
本文: 擂台上 2 个戴着拳套的男人在打拳击

d. 场景4

NIC: 室内有 1 只猫和 1 个人
Adaptive: 有 1 只猫和 1 个人在睡觉
LSTM-A: 床上有 1 个人和 1 只狗
Up-Down: 屋子里有 1 个人和 1 只狗躺在床上
本文: 房间里 1 只狗的旁边有 1 个穿着白色上衣的小孩躺在沙发上

e. 场景5

NIC: 室内有 3 个人坐着
Adaptive: 有 3 个人在电脑桌前坐着
LSTM-A: 有 1 群人围在桌子旁边
Up-Down: 房间里有 1 群人在看电脑
本文: 室内的桌子旁边有 1 群穿着西装的人在交谈

f. 场景6

图 9 不同模型生成的描述语句对比

词,使生成的描述语句更加准确和详细.由此可见,本文模型的图像描述效果更好,生成的描述语句能够表述出图像的场景信息和图像主体的属性以及关系.

4 结 语

本文提出结合全局和局部特征的 BiGRU-RA 图像中文描述模型,其通过利用图像全局特征和局部特征提高模型的表征能力.同时采用嵌入视觉注意力结构和残差连接结构的 BiGRU 解码器.此外,加入基于强化学习的策略梯度对损失函数进行优化.本文模型能够生成较为准确和详细的语句,但语句的新颖性有待提高.下一步,将考虑结合目标检测与卡槽填充的图像描述方法,提高模型在复杂场景下的鲁棒性,生成更加新颖的描述文本.

参考文献(References):

- [1] Farhadi A, Hejrati M, Sadeghi M A, *et al.* Every picture tells a story: generating sentences from images[C] //Proceedings of the 11th European Conference on Computer Vision. Heidelberg: Springer, 2010: 15-29
- [2] Wang Zhiming, Zhang Hang. A fast image retrieval method based on multi-layer CNN features [J]. Journal of Computer-Aided Design & Computer Graphics, 2019, 31(8): 1410-1416(in Chinese)
(王志明, 张航. 融合多层卷积神经网络特征的快速图像检索方法[J]. 计算机辅助设计与图形学学报, 2019, 31(8): 1410-1416)
- [3] Lyu Fan, Hu Fuyuan, Zhang Yanning, *et al.* Feedback attention model for image captioning[J]. Journal of Computer-Aided Design & Computer Graphics, 2019, 31(7): 1122-1129(in Chinese)
(吕凡, 胡伏原, 张艳宁, 等. 面向图像自动语句标注的注意力反馈模型[J]. 计算机辅助设计与图形学学报, 2019, 31(7): 1122-1129)
- [4] Aker A, Gaizauskas R. Generating image descriptions using dependency relational patterns[C] //Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2010: 1250-1258
- [5] Karpathy A, Li F F. Deep visual-semantic alignments for generating image descriptions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 664-676
- [6] Mao J H, Xu W, Yang Y, *et al.* Deep captioning with multimodal recurrent neural networks (m-RNN)[OL]. [2020-03-04]. <https://arxiv.org/abs/1412.6632>
- [7] Vinyals O, Toshev A, Bengio S, *et al.* Show and tell: a neural image caption generator[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 3156-3164
- [8] Fang H, Gupta S, Iandola F N, *et al.* From captions to visual concepts and back[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 1473-1482
- [9] Xu K, Ba J L, Kiros R, *et al.* Show, attend and tell: neural image caption generation with visual attention[C] //Proceedings of the 32nd International Conference on International Conference on Machine Learning. New York: ACM Press, 2015: 2048-2057
- [10] Xu J, Gavves E, Fernando B, *et al.* Guiding the long-short term memory model for image caption generation[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2015: 2407-2415
- [11] Yao T, Pan Y W, Li Y H, *et al.* Boosting image captioning with attributes[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 4894-4902
- [12] Wu Q, Shen C H, Liu L Q, *et al.* What value do explicit high level concepts have in vision to language problems?[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 203-212
- [13] Jiang W H, Ma L, Chen X P, *et al.* Learning to guide decoding for image captioning[OL]. [2020-03-04]. <http://arxiv.org/abs/1804.00887>
- [14] Lu J S, Xiong C M, Parikh D, *et al.* Knowing when to look: adaptive attention via a visual sentinel for image captioning[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 3242-3250
- [15] Chen L, Zhang H W, Xiao J, *et al.* SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 5659-5667
- [16] Dai B, Lin D H. Contrastive learning for image captioning [OL]. [2020-03-04]. <https://arxiv.org/abs/1710.02534>
- [17] Anderson P, He X D, Buehler C, *et al.* Bottom-up and top-down attention for image captioning and visual question answering[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 6077-6086
- [18] Guo Shutao, Zhao Dexin. A Chinese image caption model based on deep learning[J]. Journal of Tianjin University of Technology, 2020, 36(3): 30-35(in Chinese)
(郭淑涛, 赵德新. 一种基于深度学习的中文图像描述模型[J]. 天津理工大学学报, 2020, 36(3): 30-35)
- [19] He K M, Zhang X Y, Ren S Q, *et al.* Deep residual learning for image recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 770-778
- [20] Ren S Q, He K M, Girshick R, *et al.* Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelli-

- gence, 2017, 39(6): 1137-1149
- [21] Cho K, van Merriënboer B, Gulcehre C, *et al.* Learning phrase representations using RNN encoder-decoder for statistical machine translation[C] //Proceedings of the Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2014: 1724-1734
- [22] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780
- [23] Vedantam R, Lawrence Zitnick C, Parikh D. CIDEr: consensus-based image description evaluation[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 4566-4575
- [24] Papineni K, Roukos S, Ward T, *et al.* BLEU: a method for automatic evaluation of machine translation[C] //Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2002: 311-318
- [25] Lavie A, Agarwal A. METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments[C] //Proceedings of the 2nd Workshop on Statistical Machine Translation. Stroudsburg: Association for Computational Linguistics, 2007: 228-231
- [26] Lin C Y. ROUGE: a package for automatic evaluation of summaries[C] //Proceedings of the Workshop on Text Summarization Branches Out. Stroudsburg: Association for Computational Linguistics, 2004: 74-81