计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(11):195−202 [doi: 10.15888/j.cnki.csa.008133]

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

基于 ViLBERT 与 BiLSTM 的图像描述算法①

许　昊[1], 张　凯[1], 田英杰[2], 种法广[1], 王子超[1]

[1](上海电力大学 计算机科学与技术学院, 上海 200090)
[2](国家电网公司 上海电器科学研究院, 上海 200437)
通讯作者: 张　凯, E-mail: zhangkaiecnu@163.com

摘　要: 传统图像描述算法存在提取图像特征利用不足、缺少上下文信息学习和训练参数过多的问题, 提出基于 ViLBERT 和双层长短期记忆网络 (BiLSTM) 结合的图像描述算法. 使用 ViLBERT 作为编码器, ViLBERT 模型能将图片特征和描述文本信息通过联合注意力的方式进行结合, 输出图像和文本的联合特征向量. 解码器使用结合注意力机制的 BiLSTM 来生成图像描述. 该算法在 MSCOCO2014 数据集进行训练和测试, 实验评价标准 BLEU-4 和 BLEU 得分分别达到 36.9 和 125.2, 优于基于传统图像特征提取结合注意力机制图像描述算法. 通过生成文本描述对比可看出, 该算法生成的图像描述能够更细致地表述图片信息.

关键词: 图像描述; ViLBERT; BiLSTM; 注意力机制

引用格式: 许昊,张凯,田英杰,种法广,王子超.基于 ViLBERT 与 BiLSTM 的图像描述算法.计算机系统应用,2021,30(11):195–202. http://www.c-s-a.org.cn/1003-3254/8133.html

Image Caption Algorithm Based on ViLBERT and BiLSTM

XU Hao[1], ZHANG Kai[1], TIAN Ying-Jie[2], CHONG Fa-Guang[1], WANG Zi-Chao[1]

[1](College of Computer Science and Technology, Shanghai University of Electric Power, Shanghai 200090, China)
[2](Shanghai Electrical Research Institute, State Grid Corporation of China, Shanghai 200437, China)

Abstract: Traditional image captioning has the problems of the under-utilization of extracted image features, the lack of context information learning and too many training parameters. This study proposes an image captioning algorithm based on Vision-and-Language BERT (ViLBERT) and Bidirectional Long Short-Term Memory network (BiLSTM). The ViLBERT model is used as an encoder, which can combine image features and descriptive text information through the co-attention mechanism and output the joint feature vector of image and text. The decoder uses a BiLSTM combined with attention mechanism to generate image caption. The algorithm is trained and tested on MSCOCO2014, and the scores of evaluation criteria BLEU-4 and BLEU are 36.9 and 125.2 respectively. This indicates that the proposed algorithm is better than the image captioning based on the traditional image feature extraction combined with the attention mechanism. The comparison of generated text descriptions demonstrates that the image caption generated by this algorithm can describe the image information in more detail.

Key words: image caption; Vision-and-Language BERT (ViLBERT); Bidirectional Long Short-Term Memory (BiLSTM); attention mechanism

① 基金项目: 国家自然科学基金 (61872230, 61802248, 61802249); 上海高校青年教师培养资助计划 (ZZsdl18006)
Foundation item: National Natural Science Foundation of China (61872230, 61802248, 61802249); Young Teacher Cultivating Program in Shanghai University (ZZsdl18006)
收稿时间: 2020-12-29; 修改时间: 2021-02-03; 采用时间: 2021-02-23; csa 在线出版时间: 2021-10-22

图像描述 (Image Caption) 是将计算机视觉和自然语言处理两个领域相结合的跨模态跨领域的任务. 一般的, 它将输入的图片通过卷积神经网络提取图像特征并利用循环神经网络等方法生成一段文字的描述, 这段描述要求和图片的内容高度相似. 这项技术在我们的生活中有着广泛的市场需求, 例如应用在对盲人的实时语音导航中. 基于这样的需求, 将实时采集的视频图像应用图像描述技术得到对应的文本描述, 再通过文本转语音技术实时传输到盲人的耳中, 让盲人能够实时地感受到周围的环境. 这在智能机器人领域也有着类似的需求, 这项技术相当于让它有了一双能够理解分析世界的“眼睛”. 除此之外, 该技术在图像检索系统、医学 CT 图像的报告生成和新闻标题生成方面都有着不错的应用前景.

经典图像描述算法可分为 3 类. 1) 通过模板填充的方法[1] 来生成图像描述, 它主要是通过局部二值模式、尺度不变特征转换或者方向梯度直方图等算法提取图像的视觉特征, 并根据这些特征检测对应目标、动作及属性对应的单词词汇, 最后将这些单词填入到模板中. 这样的方法虽然能够保证句型语法的正确性, 也有着很大的局限性, 由于使用的模板是固定的, 它也依赖于硬解码的视觉概念影响, 这样生成的语句格式相对固定且形式单一, 应用的场景也很局限. 2) 基于检索的方法[2], 它将大量的图片描述存于一个集合中, 然后通过比较有标签图片和训练生成图片描述两者间相似度来生成一个候选描述的集合, 再从中选择最符合该图片的描述. 这样的方法能保证语句的正确性, 但语义的正确性却难以保证, 因而对图像描述的正确率较低. 3) 基于生成的方法. 这类方法一般采用编码-解码器的结构, 编码器使用卷积神经网络 (CNN)[3] 提取图像特征, 解码器采用循环神经网络 (RNN)[4] 来生成文本描述. 这是在图像描述中普遍应用且效果最好的模型, 它在语句结构的完整性、语义的正确性以及泛化能力得到了一致的认可.

编码器-解码器的结构最初由 Vinyals 等[5] 提出. 该模型编码器使用基于 CNN 的 InceptionNet 网络提取图像特征信息, 解码器使用 RNN 处理输入的图像特征来生成描述. Fang 等[6] 对编码器进行了改进, 通过提取关键词作为输入来生成描述的方法为后续结合图像和语义的编码方法提供了借鉴. Wang 等[7] 提出了一种新型的解码结构. 由 Skel-LSTM 使用 CNN 提取的图像特征来生成骨架语句, 然后使用 Attr-LSTM 为骨架语句中的词语生成对应的属性词, 最后将这两部分结合生成完整的最终描述语句. Jyoti 等[8] 提出了一种不同于用 LSTM 或者 RNN 进行解码的方法, 该工作启发式地利用卷积来进行图像描述, 达到不比传统 LSTM 差的效果.

注意力机制的引入使得图像描述算法效果得到显著提高. Xu 等[9] 在 NIC 模型[5] 的基础上把注意力机制应用在图像描述的图像特征中. 其基本思想是将编码阶段获取的图像特征进行注意力处理, 解码阶段使用 LSTM. Lu 等[10] 提出注意力机制的改进工作. 这项工作对 LSTM 进行扩展, 加入“岗哨向量”, 存储着解码器中已有的知识信息. 同时提出新的自适应注意力机制, 通过空间注意力来决定关注图像的哪个区域. Anderson 等[11] 引入 Top-down, Bottom-up 机制, 使用 Faster R-CNN 提取图像区域特征.

对抗生成网络[12] 的应用使得描述生成更加多样化. 文献 [13,14] 将 Conditional GAN 运用在图像描述, 该方法生成的图片描述贴近人类的表达, 改善了句子的自然性和多样性. Zhang 等[15] 提出的模型由两个不同的 GAN 组成. Shekher 等[16] 拓展了 COCO 数据集, 并通过对抗样本验证了 Lavi 模型的鲁棒性, Dai 等[17] 则使用对抗样样本训练解决图像描述任务生成的描述缺少独特性的问题. 文献 [18–21] 通过强化学习的方法[22] 将不同评价标准作为奖励来训练模型, 能够显著提高生成描述的质量.

传统图像描述算法研究仍有这些不足: 1) 在编码器端一般使用传统 CNN 或者目标检测算法进行图像特征的提取, 存在图像特征与文本描述关联不紧密, 仅有部分注意力权重大的特征得到利用的问题. 2) 使用的解码器 LSTM 模型较为单一, 可使用更为复杂的模型提升性能.

针对上述问题, 本文提出基于 ViLBERT 和 BiLSTM 结合的图像描述算法. 主要有以下贡献:

1) 使用预训练的 ViLBERT 模型学习图像和文本间的内在联系, 并用联合特征向量统一表示, 能更好地学习到图像特征.

2) 使用 BiLSTM 处理联合特征向量, 加入注意力机制能显著改善生成描述的质量.

3) 在 MSCOCO2014 数据集上进行训练和测试, 通过实验结果表明该算法达到了优异的性能.

本文第 1 节对本文图像描述算法涉及的相关工作

进行介绍. 第 2 节介绍本文算法的具体框架及改进部分. 第 3 节介绍实验数据集、评价标准及实验的对比和分析. 第 4 节对本文算法进行总结并展望下一步研究工作.

1 相关工作

1.1 BERT

BERT (Bidirectional Encoder Representations from Transformers)[23] 即由 Transformer 进行的双向编码表示形式, 是广泛使用的自然语言处理 (NLP) 模型. BERT 的特征之一是可以读取上下文. BERT 具有一个称为 Transformer[24] 的内置体系结构, 并且通过从两个方向 (句子的开头和结尾) 学习句子来实现“读取上下文”. 由于 BERT 是双向的, 它们可以通过从左右两个方向读取数据来使语句联系上下文. 这样可以建立单词之间的关系, 并帮助模型对相关单词做出更明智的预测. 使用转换器架构对大量未标记数据集进行预训练, 因此不需要按顺序处理数据序列, 从而可以进行并行计算.

BERT 的主要模型结构是多层 Transformer 编码器. 原始形式的 Transformer 包含两种独立的机制: 读取文本输入的编码器和产生任务预测的解码器. 由于 BERT 的目标是生成语言模型, 因此仅需要编码器机制. BERT 有简单和复杂两种结构, 结构参数见表 1. 使用 *GELU* 作为非线性激活函数, 如式 (1) 所示, 其中 tanh 为双曲正切函数.

$$GELU \approx 0.5x\left(1+\tanh\left[\sqrt{\frac{2}{\pi}}\left(x+0.0447\,15x^3\right)\right]\right) \quad (1)$$

表 1 BERT 模型参数

模型	网络层数	隐层维度	注意力头	参数量 (M)
$\text{BRET}_{\text{base}}$	12	768	12	110
$\text{BRET}_{\text{large}}$	24	1024	16	340

图 1 是对 BERT 模型的描述. 输入是一系列标识 (tokens), 这些标识首先嵌入向量中, 然后在神经网络中进行处理. 输出是大小为H的向量序列, 其中每个向量对应于具有相同索引的输入标识.

1.2 BiLSTM

BiLSTM[25], 即双向 LSTM, 是一个由两个 LSTM 网络组成的序列处理模型: 一个 LSTM 将向量正向依次输入, 另一个 LSTM 则将该向量反向依次输入. BiLSTM 有效地增加了网络可用的信息量, 改善了算法可用的上下文信息.

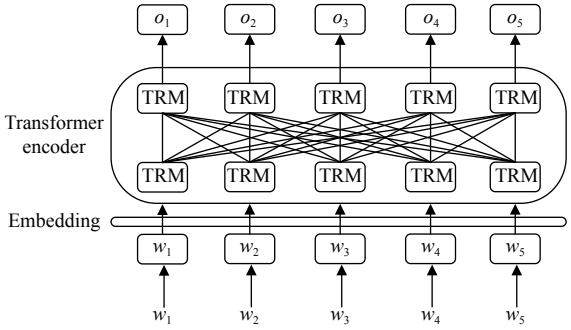

图 1 BERT 模型

BiLSTM 结构图如图 2 所示. 前向 LSTM 层的输出向量序列$\overrightarrow{h}$由前 T–n 到 T–1 时刻的正序输入迭代计算得到, 而后向 LSTM 层的输出向量序列$\overleftarrow{h}$由前 T–n 到 T–1 时刻的逆序输入迭代计算得到. 正向和反向层的输出都由标准 LSTM 更新公式计算, 见式 (2)–式 (7). BiLSTM 层最终生成一个输出向量$Y_T = [h_{T-n}, \cdots, h_{T-1}]$, 其中每个元素通过式 (8) 计算.

$$h_t = \sigma_h(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \quad (2)$$

$$y_t = \sigma_y(W_{hy}h_t + b_y) \quad (3)$$

$$f_t = \sigma_g\left(W_f x_t + U_f h_{t-1} + b_f\right) \quad (4)$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \quad (5)$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \quad (6)$$

$$\tilde{C}_t = \tanh(W_C x_t + U_C h_{t-1} + b_C) \quad (7)$$

$$y_t = \sigma\left(\overrightarrow{h}_t, \overleftarrow{h}_t\right) \quad (8)$$

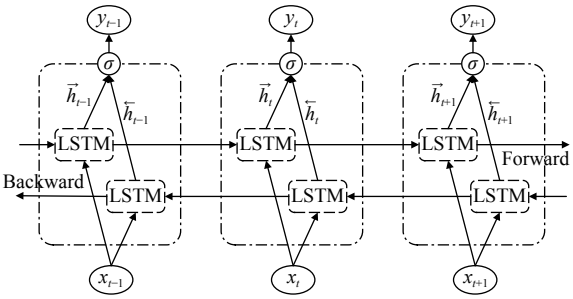

图 2 BiLSTM 模型

其中, h_t是 t 时刻隐藏层向量, x_t是输入向量, W_{xh}是从输入层到隐藏层的加权矩阵, W_{hy}是从隐藏层到输出层的加权矩阵, W_{hh}是连续隐藏层间的加权矩阵. b_h, b_y分

别为隐藏层和输出层的偏差向量. W_f, W_i, W_o, W_C 分别为隐藏层映射到遗忘门、输入门、输出门和细胞状态的加权矩阵, U_f, U_i, U_o, U_C 为前一 LSTM 模块细胞输出连接当前对应门和细胞状态的加权矩阵, b_f, b_i, b_o, b_C 为对应的偏差向量. σ 和 tanh 分别为 Sigmoid 函数和双曲正切函数. 相比单一 LSTM, BiLSTM 能更有效地从数据集中学习时空特征, 并且在大数据集中有很好的性能.

2 基于 ViLBERT 和 BiLSTM 图像描述算法

基于 ViLBERT[26] 和 BiLSTM 图像描述算法主要由 VilBERT 模型和结合注意力机制的 BiLSTM 组成. 整体模型框架见图 3.

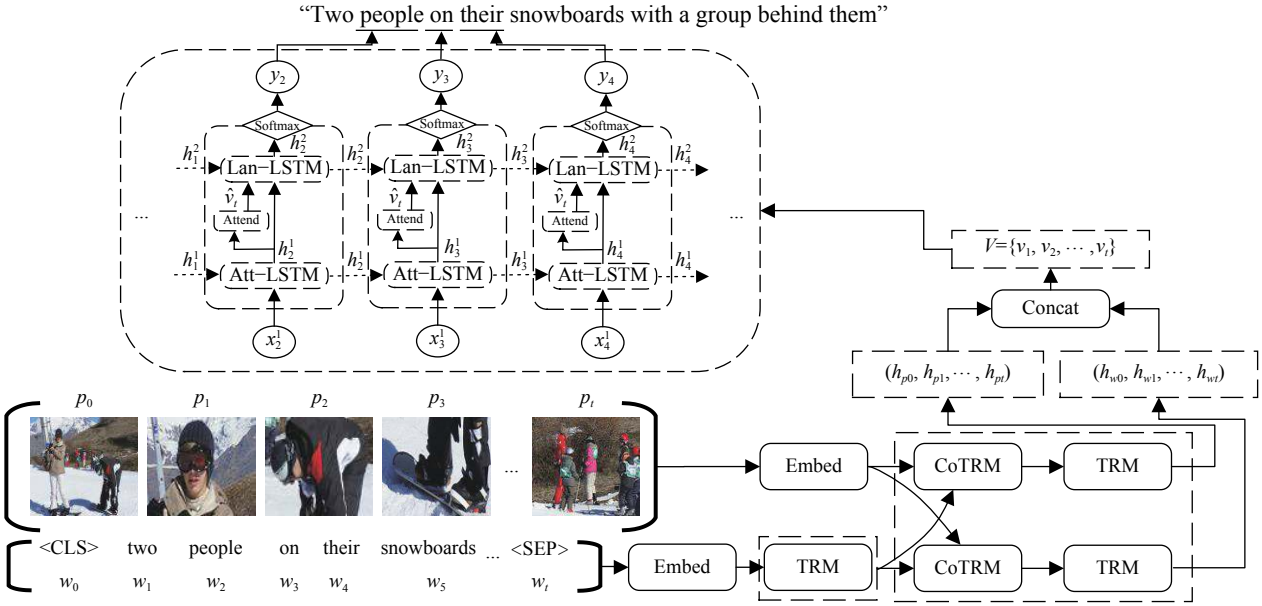

图 3 基于 ViLBERT 和 BiLSTM 图像描述框架

2.1 改进 ViLBERT 模型

ViLBERT 是 BERT 技术的扩展. 为了将 BERT 模型应用于视觉和语言, 其研究小组使用一个概念性标题的数据集, 该数据集由大约 300 万张图片和对应文本组成. 他们的方法是屏蔽掉图像的随机部分, 然后要求模型根据相关的文本重建图像的其余部分.

本文模型中, ViLBERT 由两个并行的 BERT 模型组成, 分别处理图像和文本部分, 见图 3. 每一部分都是由一系列 Transformer 块 (TRM) 和联合注意力 Transformer 层 (CoTRM) 组成, 联合注意力 Transformer 层能够实现图像和文本不同 Transformer 块的信息交互. 给定一张图片, 用 $p_1, p_2, \cdots, p_t$ 表示图像不同区域的特征, 这里使用 Faster R-CNN 使用 Resnet152 模块在 Visual Genome 数据集上进行了预训练, 来提取图像的区域特征. 图片对应的文本用 $w_0, w_2, \cdots, w_t$ 表示. 图像文本两个流之间的交换被限制在特定的层之间. 文本流在与视觉特征交互之前需要经过一个 Transformer 块的处理, 这是因为视觉特征经过神经网络深层次的提取已经有高层语义的信息, 而句子中的单词提取只需要有限的上下文聚合. 联合注意力 Transformer 层使用将图像和文本特征进行残差相加后能使图像和文本流都具有多模态的特征. 图像流部分最终输出为 $h_{p0}, h_{p1}, \cdots, h_{pt}$, 文本流部分最终输出为 $h_{w0}, h_{w1}, \cdots, h_{wt}$. 然后将两部分进行加权相加, 在本文中, 加权因子取 0.5, 即 $v_i = 0.5h_{pi} + 0.5h_{wi}$. 得到最终的图像特征 $V = \{v_1, v_2, \cdots, v_t\}$.

2.2 融合注意力 BiLSTM

采用加入注意力机制的双层长短期记忆网络, 不同于双向长短期记忆网络两个 LSTM 分别正向和逆向处理同一数据. 这里使用两个正向的 LSTM 处理不同的数据, 并在第 2 步 LSTM 中加入注意力权重向量. 每个 LSTM 采用式 (2)–式 (7) 计算. 具体结构见图 3 上部分.

在每一个时间步长上 LSTM 使用式 (9) 计算:

$$h_t = LSTM(x_t, h_{t-1}) \tag{9}$$

其中, x_t 为 LSTM 的输入向量, h_{t-1} 为上一时间步长的 LSTM 的输出向量. 每一时间步长的 LSTM 模块由一

个 Att-LSTM 和一个 Lan-LSTM 组成.

BiLSTM 中 t 时刻的 Att-LSTM 的输入为 $x_t^1=[h_{t-1}^2,\bar{v},W_e\prod_t]$, 其中 h_{t-1}^2 为上一时间步长 Lan-LSTM 的输出, $\bar{v}$ 为 ViLBERT 模型最终输出的图像向量的均值池化值, 即 $\bar{v}=\frac{1}{k}\sum_i v_i$, $W_e\prod_t$ 为之前生成单词的编码向量, W_e 为加权矩阵, $\prod_t$ 为 t 时刻驶入单词的 one-hot encoding (独热编码), 最后将这三者拼接得到 x_t^1. 将其输入到 Att-LSTM 中, 输出得到 h_t^1. 随后在每一个时间步长结合图像特征 V 生成一个标准化注意力权重 $\alpha_{i,t}$, 并得到加权后的图像特征 $\widehat{v}_t$, 计算公式如下:

$$\alpha_{i,t}=\omega_a^{\mathrm{T}}\tanh\left(W_{va}v_i+W_{ha}h_t^1\right) \tag{10}$$

$$\widehat{v}_t=\sum_{i=1}^{K}\alpha_{i,t}v_i \tag{11}$$

其中, $W_{va}\in\mathbb{R}^{H\times V}$, $W_{ha}\in\mathbb{R}^{H\times M}$, $\omega_a\in\mathbb{R}^H$ 是已学习的参数, ω_a^{T} 为 ω_a 的转置矩阵, v_i 是图形特征 V 的第 i 个区域特征, h_t^1 为 Att-LSTM 的 t 时刻的输出隐藏向量.

Lan-LSTM 的输入为 $x_t^2=[\widehat{v}_t,h_t^1]$, 由 Att-LSTM 的输出 h_t^1 和上文得到的有注意力权重的图像特征组成, 输出得到 h_t^2.

使用符号 $y_{1:T}$ 表示单词序列 $(y_1,\cdots,y_T)$, 在 t 时间步长输出单词的概率分布由下式给出:

$$p(y_t|y_{1:t-1})=softmax(W_ph_t^2+b_p) \tag{12}$$

其中, $W_p\in\mathbb{R}^{|\Sigma|\times M}$, $b_p\in\mathbb{R}^{|\Sigma|}$ 是已学习权重和偏差参数. 完整的序列输出分布为:

$$p(y_{1:T})=\prod_{t=1}^{T}p(y_t|y_{1:t-1}) \tag{13}$$

2.3 损失函数

基于给定的真实人工描述序列 $y_{1:T}^*$ 和本文图像描述算法训练得到的参数 θ, 采用交叉熵函数来最小化熵损失, 公式如下:

$$L(\theta)=-\sum_{t=1}^{T}\lg\left(p_\theta\left(y_t^*|y_{1:t-1}^*\right)\right) \tag{14}$$

其中, y_t^* 表示第 t 个真实人工描述.

3 实验及结果分析

3.1 数据集

实验采用 MSCOCO2014 数据集[27], 该数据集旨在通过将对象识别问题置于更广泛的场景理解问题的上下文中, 从而提高对象识别的最新水平, 并通过收集包含自然环境中常见对象的图像来实现. 该数据集使用专业机构人为地对图片进行描述, 每张图片收录 5 句或者 15 句参考描述, 标注集一般以 JSON 格式保存. 该数据集有超过 33 万张图片, 其中 20 万有标注描述, 包含 91 类目标, 32.8 万张图像中总共有 250 万个带有标签的实例, 这也是目前最大的语义分割数据集.

3.2 评价指标

实验采用 BLEU (BiLingual Evaluation Understudy)[28] 和 CIDEr[29] 两种评价指标对本文算法进行评估.

BLEU 算法对生成的待评价语句和人工标注语句间的差异进行评分, 得分输出在 0–1 之间. 该标准现已成为图像描述算法应用最广泛的计算标准之一.

对于图像 I_i, 图像描述模型对于这个图像生成的描述语句 c_i, 人工标注的 5 个描述语句集合 $S_i=\{s_{i1},\cdots,s_{i5}\}\in S$, 我们要对 c_i 进行评价. BLEU 的计算公式如下所示:

$$CP_n(C,S)=\frac{\Sigma_i\Sigma_k\min\left(h_k(c_i),\max_{j\in m}h_k\left(s_{ij}\right)\right)}{\Sigma_i\Sigma_k h_k(c_i)} \tag{15}$$

$$b(C,S)=\begin{cases}1, & \text{if } l_C>l_S\\ \mathrm{e}^{1-l_S/l_C}, & \text{if } l_C<l_S\end{cases} \tag{16}$$

$$BLEU_N(C,S)=b(C,S)\exp\left(\sum_{n=1}^{N}\omega_n\lg CP_n(C,S)\right) \tag{17}$$

其中, 每一个语句用 n 元组 ω_k 来表示的, n 元组 ω_k 在人工标注语句 s_{ij} 中出现的次数记作 $h_k\left(s_{ij}\right)$, n 元组 ω_k 在待评价语句 $c_i\in C$ 中出现的次数记作 $h_k(c_i)$, l_C 是待评价语句 c_i 的总长, l_S 是人工标注语句的总长度. BLEU 得分越高, 性能也就越好.

CIDEr 是专门设计用于评价图像描述模型的, 它通过计算每个 n 元组的 TF-IDF 权重得到待评价语句和参考语句之间的相似度, 以此评价图像描述的效果.

一个 n 元组 ω_k 在人工标注语句 s_{ij} 中出现的次数记作 $h_k\left(s_{ij}\right)$, 在待评价语句中出现的次数记作 $h_k(c_i)$, n 元组 ω_k 的 TF-IDF 权重 $g_k\left(s_{ij}\right)$ 如下所示:

$$g_k\left(s_{ij}\right)=\frac{h_k\left(s_{ij}\right)}{\Sigma_{\omega_l\in\Omega}}\lg\left(\frac{|I|}{\Sigma_{I_p\in I}\min\left(1,\sum_q h_k\left(s_{pq}\right)\right)}\right) \tag{18}$$

其中, Ω 是所有 n 元组的语料库, I 是数据集中所有图像的集合. CIDEr 的得分越高, 生成的语句的质量也就越好.

3.3 实验环境

实验环境基于 Ubuntu 18.04 系统, CPU 为 Inteli9-9900k, GPU 为 NVIDIA GeForce RTX 2080Ti, 16 GB

内存, Python 3.7+CUDA 10.1 的 PyTorch 深度学习环境.

3.4 预处理及参数设置

在处理注释文件时, 删除了非字母字符, 将剩余的字符转化为小写字母, 并将所有出现小于 5 次的单词替换为特殊的单词 UNK. 最终在 MSCOCO 数据集中得到 9517 个单词, 也就是最终使用的语料库.

将生成语句的最大长度设为 16, 解码器生成采用 beamsearch (集束搜索), N 值设为 3 时算法的评价指标得分最佳, 采用 Dropout 方法防止过拟合, 参数设为 0.5. 使用 Adam 优化器训练本文模型, 在训练损失函数阶段, 学习率设为3×10^{-4}, 权重衰减为1×10^{-6}, 批处理大小为 128, 训练轮数设为 30.

3.5 实验结果对比

表 2 对比本文算法 (VB-BL) 和文献 [5,10,11,13,19–21,30] 这些经典算法在 B@1、B@4 和 CIDEr 评价指标下的评分.

表 2 不同图像描述算法评价指标评分对比

模型	B@1	B@4	CIDEr
文献[5]	66.6	24.6	85.5
文献[10]	74.8	33.6	104.2
文献[11]	79.8	36.3	120.1
文献[13]	—	20.7	79.5
文献[19]	75.4	33.2	101.3
文献[20]	71.3	30.4	93.7
文献[21]	—	35.4	117.5
文献[30]	58.9	18.6	54.9
本文算法	**81.1**	**36.9**	**125.2**

注: B@1和B@4分别表示BLEU-1和BLEU-4评价标准.

可以看出, 本文的算法在 B@1, B@4 得分比该表中最优得分分别高出 1.3 和 0.6, 有一定程度的提高, 在 CIDEr 得分比该表中最优得分高出 5.1, 有 4.2% 的提升. 在编码阶段结合图像和文本特征之间的联系, 能够有效增强提取图像中的重要部分, 抑制与文本不相关、不重要的部分, 使得图像中各目标间的关联更加紧密, 起到注意力机制的效果. 在此基础上再结合双向注意力机制的 LSTM, 能够有效地提高本文算法的效果.

接着验证本文算法不同模块的效果, 进行了两个消融实验: 1) 将 ViLBERT 换成 CNN; 2) 将结合注意力机制的 BiLSTM 换成 BiLSTM. 实验对比见表 3.

从表 3 中可以看出, 将 ViLBERT 提取后的图像和文本特征进行加权处理后能够高效利用图像中的关键信息, 为后续的高质量的文本生成提供保证. 在 BiLSTM 中加入注意力机制也可以有效提升算法的效果. 综合对比可以看出, 在编码阶段使用文本信息参与对图像特征进行交融, 能使算法更加关注图像中的关键信息, 这也是本文算法取得优化的关键原因.

表 3 本文算法消融实验

模型	B@1	B@4	CIDEr
CNN+AttBiLSTM	71.8	28.6	101.1
ViLBERT+ BiLSTM	75.1	33.4	115.4
VB-BL	81.1	36.9	125.2

注: B@1和B@4分别表示BLEU-1和BLEU-4评价标准.

图 4 中选取 6 张图片进行效果的展示, 本文算法生成描述与 NIC 模型的对比见表 4.

对比可以看出, 本文算法生成的描述已经能够详细地表述出图中的主要内容, 并能够描述出各主要目标的细节特征. 在缺少人类主观因素和知识储备的情况下, 已经表现出优异的性能.

(a) (b) (c) (d) (e) (f)

图 4 对比效果图片

表 4 NIC 模型和本文算法描述对比

图像标号	描述
图4(a)	A man holding a pair of scissors in a room. A woman soldier cutting a man's necktie.
图4(b)	A group of people playing frisbee in a field. Three young women are trying to catch a frisbee.
图4(c)	Two plates of food on a table. Two plates and a bowl full of food sitting on a table.
图4(d)	A dog sitting on top of a toilet. A brown and white dog sitting on top of a toilet.
图4(e)	A group of people flying kites on a beach. A crowd of people standing on a beach flying two kites.
图4(f)	A group of people standing on top of a snow covered slope. Two men standing on a hill in snow skis.

注: 每张图第一行为NIC模型[5], 第二行为本文算法描述.

4 结束语

本文提出基于 ViLBERT 和 BiLSTM 结合的图像描述算法. 使用 ViLBERT 模型创新地融合了图像和文

本间的特征, 使得提取到的图像特征具有类似视觉注意力的特性, 该模型精简的参数量能有效缩减训练的时间. 结合采用了融合注意力机制的双层长短期记忆网络能改善注意力权重的可解释性. 该算法进一步统一视觉图像和语言理解间的跨模块特征. 实验表明, 该算法在各评价指标上都有着优异的表现. 在未来的工作中, 将结合图卷积神经网络来加强图像内各目标间的联系来展开进一步的研究.

参考文献

1 Kulkarni G, Premraj V, Ordonez V, *et al*. Babytalk: Understanding and generating simple image descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2891–2903. [doi: 10.1109/TPAMI.2012.162]

2 Kuznetsova P, Ordonez V, Berg TL, *et al*. $T_{REE}T_{ALK}$: Composition and compression of trees for image descriptions. Transactions of the Association for Computational Linguistics, 2014, 2: 351–362. [doi: 10.1162/tacl_a_00188]

3 LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436–444. [doi: 10.1038/nature14539]

4 Zaremba W, Sutskever I, Vinyals O. Recurrent neural network regularization. International Conference on Learning Representations 2015. San Diego: ICLR, 2015.

5 Vinyals O, Toshev A, Bengio S, *et al*. Show and tell: A neural image caption generator. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015. 3156–3164.

6 Fang H, Gupta S, Iandola F, *et al*. From captions to visual concepts and back. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015. 1473–1482.

7 Wang YF, Lin Z, Shen XH, *et al*. Skeleton key: Image captioning by skeleton-attribute decomposition. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 7378–7387.

8 Aneja J, Deshpande A, Schwing AG. Convolutional image captioning. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 5561–5570.

9 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 International Conference on Machine Learning-Volume 37. Lille: ACM, 2015. 2048–2057.

10 Lu JS, Xiong CM, Parikh D, *et al*. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 3242–3250.

11 Anderson P, He XD, Buehler C, *et al*. Bottom-up and top-down attention for image captioning and visual question answering. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6077–6086.

12 Goodfellow IJ, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial nets. Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2. Montreal: ACM, 2014. 2672–2680.

13 Dai B, Fidler S, Urtasun R, *et al*. Towards diverse and natural image descriptions via a conditional GAN. 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017. 2989–2998.

14 Shetty R, Rohrbach M, Hendricks L A, *et al*. Speaking the same language: Matching machine to human captions by adversarial training. 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017. 4155–4164.

15 Zhang H, Xu T, Li HS, *et al*. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017. 5908–5916.

16 Shekhar R, Pezzelle S, Klimovich Y, *et al*. FOIL it! find one mismatch between image and language caption. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver: Association for Computational Linguistics, 2017. 255–265.

17 Dai B, Lin DH. Contrastive learning for image captioning. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: ACM, 2017. 898–907.

18 Ranzato MA, Chopra S, Auli M, *et al*. Sequence level training with recurrent neural networks. 4th International Conference on Learning Representations. San Juan: ICLR, 2016.

19 Liu SQ, Zhu ZH, Ye N, *et al*. Improved image captioning via policy gradient optimization of spider. 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017. 873–881.

20 Ren Z, Wang XY, Zhang N, *et al*. Deep reinforcement learning-based image captioning with embedding reward.

2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 1151–1159.

21 Rennie SJ, Marcheret E, Mroueh Y, *et al*. Self-critical sequence training for image captioning. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 1179–1195.

22 Sutton RS, Barto AG. Reinforcement Learning: An Introduction. 2nd ed. Cambridge: MIT Press, 2018.

23 Devlin J, Chang MW, Lee K, *et al*. Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2019. 4171–4186.

24 Correia GM, Niculae V, Martins AFT. Adaptively sparse transformers. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong: Association for Computational Linguistics, 2019. 2174–2184.

25 Cornegruta S, Bakewell R, Withey S, *et al*. Modelling radiological language with bidirectional long short-term memory networks. Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis. Austin: Association for Computational Linguistics, 2016. 17–27.

26 Lu JS, Batra D, Parikh D, *et al*. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in Neural Information Processing Systems. Vancouver: NeurIPS, 2019. 13–23.

27 Chen XL, Fang H, Lin TY, *et al*. Microsoft COCO captions: Data collection and evaluation server. arXiv: 1504.00325v2, 2015.

28 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: ACM, 2002. 311–318.

29 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: ACL, 2005. 65–72.

30 Feng Y, Ma L, Liu W, *et al*. Unsupervised image captioning. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 4120–4129.