计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(9):191−197 [doi: 10.15888/j.cnki.csa.007513]

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

基于多模态神经网络生成图像中文描述①

陈　兴

(河海大学 计算机与信息学院, 南京 211100)
通讯作者: 陈　兴, E-mail: hsingchen@hhu.edu.cn

摘　要: 自动生成图片描述是自然语言处理和计算机视觉的热点研究话题, 要求计算机理解图像语义信息并用人类自然语言的形式进行文字表述. 针对当前生成中文图像描述整体质量不高的问题, 提出首先利用 FastText 生成词向量, 利用卷积神经网络提取图像全局特征; 然后将成对的语句和图像〈S, I〉进行编码, 并融合为两者的多模态特征矩阵; 最后模型采用多层的长短时记忆网络对多模态特征矩阵进行解码, 并通过计算余弦相似度得到解码的结果. 通过对比发现所提模型在双语评估研究 (BLEU) 指标上优于其他模型, 生成的中文描述可以准确概括图像的语义信息.

关键词: 图像中文描述; FastText 语言模型; 卷积神经网络; 长短时记忆网络

引用格式: 陈兴.基于多模态神经网络生成图像中文描述.计算机系统应用,2020,29(9):191–197. http://www.c-s-a.org.cn/1003-3254/7513.html

Generation of Chinese Image Description by Multimodal Neural Network

CHEN Xing

(College of Computer and Information, Hohai University, Nanjing 211100, China)

Abstract: Automatic image captioning is a hot topic which connects natural language processing and computer vision. It mainly completes the task of understanding image semantic information and expressing it in the form of human natural language. For the overall quality of Chinese image captioning is not very high, this study uses FastText to generate word vector, uses convolution neural network to extract the global features of the image, then encodes the pairs of sentences and images〈S, I〉, and finally merges them into a feature matrix containing both Chinese description and image information. Decoder uses LSTM model to decode the feature matrix, and obtains the decoding result by calculating cosine similarity. Through comparison, we find that the model proposed in this study is better than other models in BiLingual Evaluation Understudy (BLEU). The Chinese description generated by the model can accurately summarize the semantic information of the image.

Key words: Chinese image captioning; FastText; Convolutional Neural Network (CNN); Long and Short Term Memory network (LSTM)

生成图像描述是自然语言处理和计算机视觉的热点研究问题, 其任务主要为理解图像中物体和场景等语义信息并采用近似于人类语言的描述形式表达出来, 生成符合一定语法规则的文本信息. 图片描述作为解决图像信息到文本信息的跨模态转换的重要方法, 可广泛应用在人机交互、视觉辅助、图像标注等领域.

Vinyals 等[1] 提出了 GoogleNIC 模型, 借鉴机器翻译中常用的编码解码方式, 分别采用 InceptionV3[2] 预训练模型作为编码器提取图像特征信息, 长短时记忆网络 (LSTM)[3] 作为解码器生成图像描述. Xu 等[4] 在

① 收稿时间: 2020-01-04; 修改时间: 2020-01-22; 采用时间: 2020-02-11; csa 在线出版时间: 2020-09-04

解码过程中加入注意力机制, 更加关注图像的局部特征. 邓珍荣等[5] 也应用了注意力机制和卷积神经网络提取图像特征, 与 Xu 不同的是采用了循环门单元网络 (GRU)[6] 取代 LSTM 网络以此生成图像的描述.

随着生成对抗网络的发展流行, 生成对抗机制[7] 也逐渐被应用到图像描述生成任务中. 例如 Dai 等 [8] 通过控制条件对抗网络[9] 中噪音 Z 来生成图像的描述, 并通过实验证明加入对抗机制生成的描述在该任务上优于其他方法, 生成的图像描述在自然性和多样性两方面都有所提高. Shetty 等[10] 同样采用生成对抗网络结构, 通过对抗机制提高语句多样性. 其中生成器利用 CNN 提取图像特征, 在解码过程引入了目标检测模型, 额外的加入 RCNN[11] 检测图像中具体目标图像, 通过加入图像的先验知识来提高生成描述中包含图像目标的概率. 文献 [12] 也通过加入先验知识来改进模型生成质量, 提出了基于主题模型的图像描述生成模型, 不同于文献 [10] 加入目标检测, 而是通过预测图像的主题词的概率分布来实现先验知识的获取. 这种加入先验知识的方法虽然可以获得较好的生成效果, 但是增加了生成的计算和时间成本, 而且过度依赖于先验知识, 一旦接收了错误的先验知识, 生成的描述也会受到影响.

以上工作都是基于英文领域的图像描述生成方法.因为中文词汇和语句表达的特殊性, 在理解中文语义信息往往需要断句和分词, 且中文语法表达极为灵活, 句式更加多变, 因此相比较英文的图像描述难度更大. Li 等[13] 在 Flickr8K 数据集基础上通过机器翻译和人工标注两种方式提出构建首个中文图片描述的 Flickr8K-CN 数据集. Wu 等[14] 构建首个大规模的图像中文描述数据集 AIC-ICC, 包含 30 万张图片和 150 万句的中文描述. 刘泽宇等[15] 基于编码解码结构构建了针对图像中文描述的生成模型, 并在 Flickr8K-CN 数据集上验证了该模型的有效性.

比较以上, 本文提出了多模态神经网络生成图像中文描述模型, 不同于现有的图像中文描述方法, (1) 本文采用 FastText[16] 预训练模型对词汇进行词嵌入表示. 相较于常用的 one-hot 编码, 一个词汇的表示往往需要词库大小的维度, 这个问题在中文词汇上尤其严重. 词嵌入表示形式维度相较于 one-hot 编码维度更小, 可以更准确的保留词汇之间的语义和用法的相关性; (2) 相较于利用注意力机制放大图像的局部特征, 本文更加关注图像的整体特征, 通过融合编码器设计, 将卷积神经网络中提取到的图片全局特征与图片对应的中文描述的词向量表示融合编码; (3) 损失函数由两部分组成, 一部分衡量生成的词向量矩阵与真实词向量特征矩阵的余弦距离, 另一部分衡量生成描述单一词汇与目标之间的复现比例.

本文在数据集 AIC-ICC、Flickr8K-CN 都进行了实验, 并通过双语评估研究 (BLEU)[17] 值的客观指标进行评估, 与现有的模型进行对比发现 BLEU 值有不同程度上的提高.

1 图像中文描述网络结构

本文实现图像的中文描述同样采用 encoder-decoder 结构, 其中 encoder 模型包含 FastText 词嵌入模型和卷积网络, 前者生成语句的词向量矩阵, 后者提取图像的全局特征, encoder 将成对的语句、图像〈S,I〉进行编码, 最终融合为既包含中文描述又包含图像信息的多模态特征矩阵. Decoder 采用多层 LSTM 模型对多模态特征矩阵进行解码, 通过计算余弦相似度得到解码结果.

1.1 FastText

词汇如果通过常用的 one-hot 编码方式进行表示, 例如在表示词汇“运动员”时, 需要在“运动员”对应的维度上设置为 1, 其他维度上设置为 0, 这就意味着每一个词都需要独占一维空间. 那么, 一个词汇的维度等于词表长度, 该词向量中绝大多数的维度都没有被利用到. 同时, 词汇 one-hot 编码形式无法反映出词汇之间的相关性. 本文采用的 FastText 预训练模型对词汇进行词嵌入表示可以大大减少 one-hot 编码带来的冗余和稀疏问题. 一个包含 10^4 量级大小的词表, 在 FastText 中只需要 10^2 量级就可以表征, 通过余弦距离的计算也可以反映词汇之间的相关程度.

FastText[16] 一种高效快速的文本分类模型. 该模型首先将文本词汇通过 n-gram[18] 格式分解, 然后和原单词相加, 得到的文本序列$(x_1,x_2,\cdots,x_{n-1},x_n)$作为网络的输入, 通过单层隐藏层学习, 最终输出该文本的分类类别. FastText 采用分层 Softmax[19] 根据类别频率构造霍夫曼树, 相较于标准的 Softmax 层计算的时间复杂度从$O(kh)$下降到$O(h(\log_2 k))$, 其中 k 为类别数量, h 为文本特征维数, 训练和分类效率都得到了较大的提升. FastText 模型如图 1 所示.

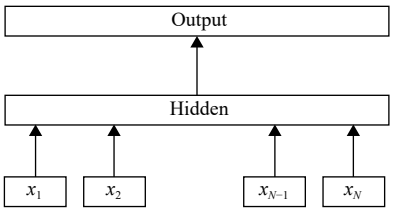

图 1 FastText 模型结构

本文中利用 FastText 模型对图像描述中的每个词汇转换为词向量形式, 计算两个词向量的余弦相似度, 其计算公式为式 (1). 计算实例见表 1.

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \tag{1}$$

若余弦值越大, 则两个词汇在高维空间的表征形式越相近, 说明两个词汇有着相近的语义或强相关的用法, 所以 FastText 模型在词向量表征学习具有良好的性能.

表 1 余弦相似度实例

Word(a)	References(b)	*Similarity*(cos(a,b))
运动员	運動員	0.727463603019714 4
	篮球	0.704215347766876 2
	司职	0.701143264770507 8
	选手	0.661421418190002 4
	球员	0.659771442413330 1
甜点	甜點	0.639626264572143 6
	點心	0.637476801872253 4
	蛋糕	0.615930676460266 1
	料理	0.613221883773803 7
	菜式	0.597212612628936 8

图 2 为将若干个词向量通过 PCA 降成两维后的可视化表示. 可以看出, “教师”与“授课”、“学校”、“老师”等词汇相关度较高, 在图中距离也较近. 以“教师”、“运动员”、“甜点”、“商品”为关键词的词汇在图中形成 4 个词汇区域, 可见 FastText 在表征词汇上可以很好地保留其语义信息, 词汇的相关性也得以体现.

1.2 卷积网络提取图像特征

卷积神经网络是一种有效的图像特征提取方法, 广泛应用在图像识别、图像检测等相关任务上, 并且表现出相当好的性能. 卷积网络通过局部区域感受野和权值共享的设计, 大大减少了模型的复杂度, 使得模型更加易于训练. 卷积网络的网络结构也随着问题的逐渐复杂而逐渐加深[20–22].

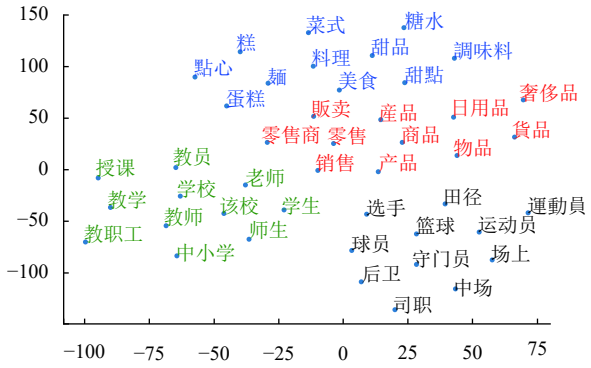

图 2 部分词向量的可视化表示

本文中使用了类似于 VGG-16 的卷积神经网络提取图像特征, 如表 2 所示, 输入图像首先经过裁剪和随机翻转增加图像的多样性, 最终变形为 224×224 的三通道 RGB 图像, 经过第一个卷积模块, 包含两次 64 个 3×3 的卷积层, 每层卷积层使用 ReLU 激活函数, 最终采用最大池化层向下采样, 输出 128 个 112×112 的特征图; 第二个卷积模块与第一个卷积模块一致, 输出 256 个 56×56 的特征图; 第三、第四、第五卷积模块包含 3 个卷积层, 经 ReLU 激活函数后采用平均池化层降采样, 进一步压缩和编码图像, 输出 512 个 7×7 的图像特征图; 最后经过两个全连接层, 大小分别为 4096、300, 最终输出 1×300 的图像全局特征向量, 该特征向量的特征维数与 FastText 所编码的文本词向量特征维度一致.

表 2 卷积神经网络结构

网络层	输入	输出	卷积核	步幅	激活函数
Input	224×224×3	224×224×3	—	—	—
Conv1	224×224×3	224×224×64	3×3,64	1	ReLU
Conv2	224×224×64	224×224×64	3×3,64	1	ReLU
MaxPool1	224×224×64	112×112×64	—	2	—
Conv3	112×112×64	112×112×128	3×3,128	1	ReLU
Conv4	112×112×128	112×112×128	3×3,128	1	ReLU
MaxPool2	112×112×128	56×56×128	—	2	—
Conv5	56×56×128	56×56×256	3×3,256	1	ReLU
Conv6	56×56×256	56×56×256	3×3,256	1	ReLU
Conv7	56×56×256	56×56×256	3×3,256	1	ReLU
AvgPool1	56×56×256	28×28×256	—	2	—
Conv8	28×28×256	28×28×512	3×3,512	1	ReLU
Conv9	28×28×512	28×28×512	3×3,512	1	ReLU
Conv10	28×28×512	28×28×512	3×3,512	1	ReLU
AvgPool2	28×28×512	14×14×512	—	2	—
Conv11	14×14×512	14×14×512	3×3,512	1	ReLU
Conv12	14×14×512	14×14×512	3×3,512	1	ReLU
Conv13	14×14×512	14×14×512	3×3,512	1	ReLU
AvgPool3	14×14×512	7×7×512	—	2	—
FC1	7×7×512	4096	—	—	—
FC2	4096	300	—	—	—

1.3 图像中文描述生成模型

如图 3 所示, 图像中文描述模型由两部分构成, 融合编码器和 LSTM 解码器.

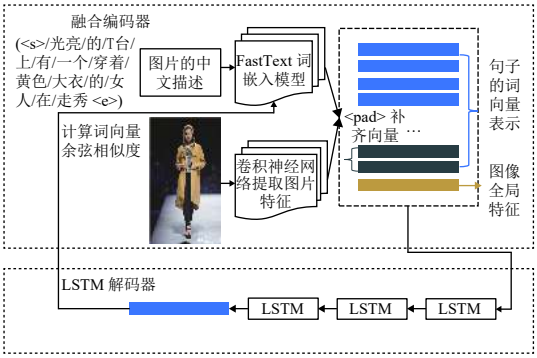

图 3 图像中文描述生成模型

融合编码器对成对的语句和图像〈S,I〉进行编码. 图像的中文描述在编码之前首先需要分词, 将原始的中文描述 S 转换为多个词汇的序列结构$\{w_1, w_2, \cdots, w_{n-1}, w_n\}$, 其中$w_i, i \in (1, n)$表示一个中文词汇, 词汇结构通过 FastText 词嵌入模型进一步转换为 (n×300) 的词向量矩阵. 与描述文本对应的图像送入表 2 所示的卷积神经网络提取图像的全局特征, 生成的图像特征, 最终对两个编码模型的生成结果以拼接的形式整合为 ((n+1)×300) 的多模态融合特征矩阵.

融合编码器所编码后形成的多模态特征矩阵依赖于 LSTM 网络作为解码器进行解码, LSTM 网络为三层, 最终解码生成的 (1×300) 词向量在 FastText 词嵌入模型中计算余弦相似度, 寻找与该词向量最相似的词汇, 作为本次解码的结果.

图 4 进一步说明了多模态特征矩阵的解码过程, 首先图像 I 经过如表 2 所述的卷积神经网络中得到图像的全局特征向量$\{s\}$, $\{s\}$与<start>开始标记对应的词向量$\{v_0\}$融合为多模态特征矩阵在t_1时刻输入到 LSTM 网络中 (事实上, 为了训练方便固定了多模态特征的大小, 其中长度不足的用<pad>标记进行补齐), 经过三层 LSTM 网络对输入多模态特征矩阵$\{v_0, s\}$进行解码得到词向量$\{v_1\}$, $\{v_1\}$通过 FastText 词嵌入模型计算余弦相似度可以得到该词向量所代表的具体词汇$\{y_1\}$. 在t_2时刻将t_1时刻的输入$\{v_0, s\}$与词向量$\{v_1\}$融合为新的多模态特征矩阵$\{v_0, v_1, s\}$作为此时 LSTM 网络的输入, 直到生成的词向量$\{v_m\}$代表的词汇$\{y_m\}$为结束标记<end>为止. 其中序列$\{v_0, v_1, \cdots, v_m, s\}$为最终解码得到的多模态特征矩阵, $\{y_0, y_1, \cdots, y_m\}$表示经 FastText 模型后生成的词汇序列. 其中$\{y_0\}$代表<start>开始标记.

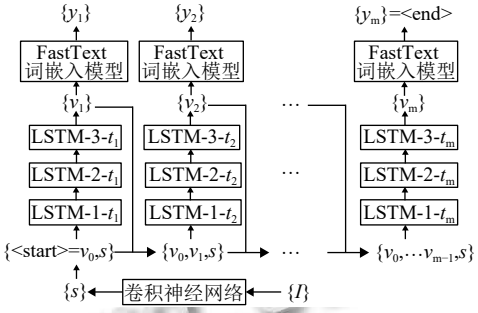

图 4 多模态特征矩阵的解码过程

1.4 损失函数

在 one-hot 编码中, 词向量每一个维度的值都为 0 或 1 的离散型数据, 可以选择交叉熵函数作为损失函数, 最小化损失函数减少生成词向量和目标词向量的距离. 本文为了降低 one-hot 编码形式带了冗余, 采用 FastText 模型得到每个词汇的词向量, 也就是说, 在产生最终词汇序列$\{y_0, y_1, \cdots, y_m\}$之前, 必然会有与之相对应的多模态特征矩阵$\{v_0, v_1, \cdots, v_m, s\}$.

设$\{w_0, w_1, \cdots, w_m\}$为图像 I 的生成语句目标. 词汇序列$\{w_0, w_1, \cdots, w_m\}$经 FastText 模型可进一步得到该序列在词空间上的表示$\{v'_0, v'_1, \cdots, v'_m\}$. 那么衡量生成的词向量矩阵与真实词向量特征矩阵平均余弦相似度为:

$$l_1 = \frac{1}{m+1}\sum_{i=0}^{m}\cos\left(v_i, v'_i\right) \tag{2}$$

$$l_1 = \frac{1}{m+1}\sum_{i=0}^{m}\frac{\sum_{j=0}^{n-1} v_{ij} \times v'_{ij}}{\sqrt{\sum_{j=0}^{n-1}\left(v_{ij}\right)^2} \times \sqrt{\sum_{j=0}^{n-1}\left(v'_{ij}\right)^2}} \tag{3}$$

其中, $m+1$表示词向量的个数, n为词向量的维度大小.

l_1值越大表明生成的词向量矩阵和目标矩阵越相似. l_1值严格限制了词向量的序列, 即词向量v_i仅和v'_i计算相似度, 如果v_i与$v'_j (i \neq j)$相似但在l_1中无法得到奖励, 于是需要进一步量化生成描述中单一词汇与目标之间的复现比例, 即:

$$l_2 = \frac{1}{m+1}\sum_{i=0}^{m} g\left(\mathrm{y}_i, W\right) \tag{4}$$

$$g(y_i, W) = \begin{cases} 1, & \text{if}(y_i \in W) \\ 0, & \text{if}(y_i \notin W) \end{cases} \tag{5}$$

最终的损失函数可以表示为:

$$l = \frac{1}{l_1} + \lambda \frac{1}{l_2} \tag{6}$$

其中, λ为平衡两种量化指标的因子, 在实验中设置为 0.1. 最小化损失函数l即意味着最大化词向量矩阵$\{v_0, v_1, \cdots, v_m\}$与$\{v'_0, v'_1, \cdots, v'_m\}$余弦相似度, 和最大化词序列$\{y_0, y_1, \cdots, y_m\}$在目标词序列$\{w_0, w_1, \cdots, w_m\}$中的复现比例.

2 实验与结果分析

2.1 数据集

目前公开的图像中文描述数据集有 AIC-ICC 以及 Flickr8K-CN. 其详细信息如表 3 所示.

表 3 图片中文描述数据集 (单位: K)

Dataset	Train	Valid	Test-1	Test-2	Captions
AIC-ICC	210	30	30	30	1500
Flicr8K-CN	6	1	1	—	40

由表 3 可见, AIC-ICC 数据集共包含 30 万张图像和 150 万句图片描述, 本文主要选择该数据集进行实验, 其中 AIC-ICC-Train 的 21 万张图像作为训练集, AIC-ICC-Valid 的 3 万张图像作为验证集. 数据样本如图 5 所示.

中文描述
1. 绵延的山顶上站着一个举着双手的人
2. 陡峭的山顶站着一个背着包双手举起的人
3. 蓝天白云下有一个高举双手的男人站在大地上
4. 阳光明媚的山顶上站着一个举着双臂的人
5. 高高的山顶上站着一个双手举起的男人

图 5 AIC-ICC 数据实例

2.2 实验设置

本文采用结巴分词对中文描述进行分词, 统计语句经过分词后的句子长度. 如图 6 所示, 训练集分词后的最大句子长度为 32; 除此之外, 加入<start>和<end>标记向量作为语句起始和结束标志, 通过 FastText 产生的语句描述词向量矩阵固定为 (34×300), 当语句长度不足 34 时, 使用<pad>标记进行补齐, FastText 中不存在的词向量的词汇用标记<unk>替换. 加上 CNN 网络提取到的图像特征向量 (1×300), 所以最终送入 LSTM 解码器的融合矩阵大小固定为 (35×300).

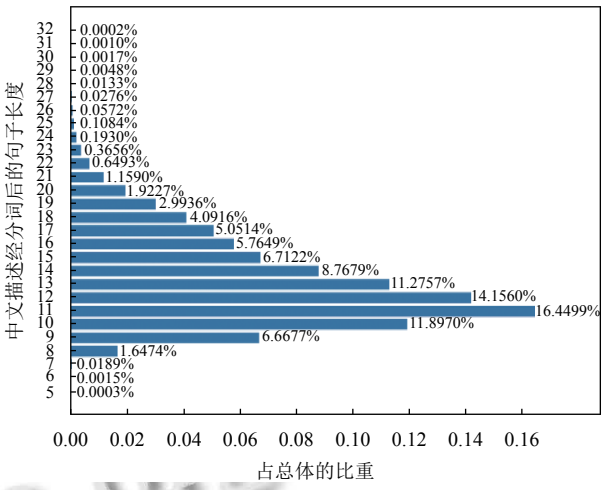

图 6 中文分词后的句子长度统计

输入图像大小统一设置为 3×224×224, batch_size 设置为 256, 进行 5 轮共计约 150 000 次迭代训练, 使用 Adam 优化算法, 其中学习速率为 0.0001, beta1 为 0.9, beta2 为 0.999. 每 10 000 次迭代完成后保存一次模型, 最终在验证集上随机选取测试图片进行测试.

2.3 评价指标

BLEU[17] 是 2002 年提出的衡量生成语句的质量的评价指标. 最先应用在对比机器翻译语句和人工翻译的参考语句之间的相似度, 衡量机器翻译所生成的质量好坏, 现已广泛应用在生成式的自然语言评价上. BLEU 针对一元词汇、二元词汇、三元词汇和元词汇分别又有 BLEU-1、BLEU-2、BLEU-3、BLEU-4. 其公式为:

$$f_{\text{BLEU}} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{7}$$

其中, BP 为机器翻译长度小于参考语句时的惩罚因子, 即:

$$BP = \begin{cases} 1, & \text{if} \quad c > r \\ \exp^{1-r/c}, & \text{if} \quad c \le r \end{cases} \tag{8}$$

其中, c 为机器语句长度, r 为参考语句的长度. p_n为:

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\_gram \in C} Count_{clip}(n\_gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\_gram' \in C'} Count(n\_gram')} \tag{9}$$

p_n直接反映了机器翻译语句在参考译文的在 n 元组上准确率. BLEU-1 只考虑 1 元组的准确率, BLEU-2 同时对 1 元组和 2 元组加权求和, 以此类推. BLEU 值越高, 说明模型生成的语句与参考语句越相似, 生成

语句的质量越高.

2.4 结果分析

在 AIC-ICC 和 Flickr8K-CN 的验证集上测试本文所提的模型, 测试结果如图 7、图 8 所示. 每幅图片中共有 6 句中文描述, 其中第 1 句为模型的生成结果, 后 5 句为数据集中的中文描述, 由此可见自动生成中的中文图像描述与图像相关, 图像中的人物、物体和场景在描述中都可以准确地表达出来, 生成结果与图像真实的人工描述相近.

生成描述:
平整的 T 台上有一位穿着时尚的女士在走秀
人工描述:
1. 光亮的 T 台上有一个穿着黄色大衣的女人在走秀
2. 秀场上有一个双手抓着衣襟的女人在走秀
3. 平坦的 T 台上有一位穿着黄色上衣的女士在走秀
4. 一个双手抓着衣服边的女人在 T 台上走秀
5. 一个身穿长款外套的女人在 T 台上走秀

(a) AIC-ICC 验证集测试结果示例一

生成描述:
洒满阳光的道路上有一个短头发的男人在骑电动车
人工描述:
1. 一个骑助力车的男人行驶在洒着阳光的道路上
2. 一个戴着墨镜的男人在道路上骑电动车
3. 道路上有一个穿着白色短袖的男人在骑电动车
4. 平整的道路上有一个戴墨镜的男人在骑摩托车
5. 一个戴着墨镜的男人在宽敞的马路上骑电动车

(b) AIC-ICC 验证集测试结果示例二

图 7 AIC-ICC 验证集上的测试结果

生成描述:
一个小女孩正在公园里玩旋转滑梯
人工描述:
1. 滑梯上的一个孩子
2. 一个孩子正在操场上从螺旋滑梯上滑下
3. 一个女孩从公园里蓝色和黄色的滑梯上滑下来
4. 一个小女孩在公园滑下滑梯
5. 在游乐场一个小女孩从一个螺旋滑梯上滑下来

(a) Flickr8K-CN 验证集测试结果示例一

生成描述:
一只狗在山顶的草地上玩游戏
人工描述:
1. 一只黑白相间的狗嘴里叼着一根棍子站在一座小山上
2. 一只狗在草地上玩接东西游戏
3. 一只狗站在山顶上眺望
4. 一只狗站在一座小山上, 望着山谷, 嘴里叼着一个木棍
5. 嘴叼着棒子的棕色狗

(b) Flickr8K-CN 验证集测试结果示例二

图 8 Flickr8K-CN 验证集上的测试结果

本文选取 GoogleNIC、Hard-Attention、Soft-Attention、gLSTM、Multimodal-RNN、CNIC-E 等 6 种图像描述生成模型进行对比, 其中 GoogleNIC、Hard-Attention、Soft-Attention、gLSTM、Multimodal-RNN 模型为 Flickr8K 上的测试数据, CNIC-E、本文是在 Flickr8K-CN 数据集上进行 BLEU-1、BLEU-2、BLEU-4、BLEU-4 指标测试. 测试结果见表 4.

表 4 各个模型的 BLEU 值对比

模型	BLEU-1	BLEU-2	BLEU-3	BLEU-4
GoogleNIC[1](flickr8k)	63.0	41.0	27.0	--
HardAttention[2] (flickr8k)	67.0	45.7	31.4	21.3
SoftAttention[2](flickr8k)	67.0	44.8	29.9	19.5
MultiRNN[6](flickr8k)	57.9	38.3	24.5	16.0
gLSTM[23](flickr8k)	64.7	45.9	31.8	21.6
CNIC-E[15] (flickr8kCN)	69.1	45.9	30.8	20.5
本文 (flickr8kCN)	**69.8**	**47.8**	**34.8**	**21.9**

由表 4 所知, 本文所提模型的 BLEU 值与现有的模型对比, 在 BLEU 指数上有不同程度的提高. 本文所述模型在 Flickr8K-CN 数据集上取得了最好的结果, 生成的语句更加贴近人工图片描述的参考语句, 符合人类自然语言表达.

3 结论与展望

本文提出的图像中文描述生成方法, 首先将描述语句和图像<S,I>进行编码, 既包含对中文描述语句的分词表示又包含对图像的全局特征提取, 最终语句图像对<S,I>融合为一个多模态特征矩阵. 使用三层 LSTM 模型对多模态特征矩阵进行解码, 通过计算余弦相似度得到解码的结果. 所提模型生成的中文描述可以准确的概括图像的语义信息. 在 *BLEU* 指标上优于其他模型.

但是, 本文模型仍存在一些问题. 例如, 对较为复杂的场景和多人物的识别不够准确, 生成的语句不够细腻等, 仍有待提升的空间. 该模型较多的考虑图像的全局特征, 而忽略了图像的局部特征的提取, 从而导致生成语句在图像的细节上体现不足. 未来可以加入注意力机制, 加强模型对局部细节的把握.

参考文献

1 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.
2 Szegedy C, Liu W, Jia YQ, *et al*. Going deeper with convolutions. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 1–9.
3 Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 1998, 6(2): 107–116. [doi: 10.1142/S0218488598000094]
4 Xu K, Ba J, 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.
5 邓珍荣, 张宝军, 蒋周琴, 等. 融合 Word2Vec 和注意力机制的图像描述模型. 计算机科学, 2019, 46(4): 268–273. [doi: 10.11896/j.issn.1002-137X.2019.04.042]
6 Cho K, Van Merrienboer B, Gulcehre C, *et al*. Learning phrase representations using RNN encoder-decoder for statistical machine translation. Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar. 2014. 1724–1734.
7 Goodfellow IJ, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial nets. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, QC, Canada. 2014. 2672–2680.
8 Dai B, Fidler S, Urtasun R, *et al*. Towards diverse and natural image descriptions via a conditional GAN. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy. 2017. 2989–2998.
9 Mirza M, Osindero S. Conditional generative adversarial nets. arXiv: 1411.1784, 2014.
10 Shetty R, Rohrbach M, Hendricks LA, *et al*. Speaking the same language: Matching machine to human captions by adversarial training. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy. 2017. 4155–4164.
11 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: 10.1109/TPAMI.2016.2577031]
12 余燕. 基于视觉注意力与主题模型的图像中文描述生成方法研究 [硕士学位论文]. 武汉: 武汉科技大学, 2019.
13 Li XR, Lan WY, Dong JF, *et al*. Adding chinese captions to images. Proceedings of 2016 ACM on International Conference on Multimedia Retrieval. New York, NY, USA. 2016. 271–275.
14 Wu JH, Zheng H, Zhao B, *et al*. AI challenger: A large-scale dataset for going deeper in image understanding. arXiv: 1711.06475, 2017.
15 刘泽宇, 马龙龙, 吴健, 等. 基于多模态神经网络的图像中文摘要生成方法. 中文信息学报, 2017, 31(6): 162–171. [doi: 10.3969/j.issn.1003-0077.2017.06.022]
16 Joulin A, Grave E, Bojanowski P, *et al*. Bag of tricks for efficient text classification. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia, Spain. 2017. 427–431.
17 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, FL, USA. 2002. 311–318.
18 Cavnar WB, Trenkle JM. N-gram-based text categorization. Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval. Las Vegas, NV, USA. 1994. 161−175.
19 Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. Proceedings of the 1st International Conference on Learning Representations. Scottsdale, AZ, USA. 2013.
20 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Proceedings of the 26th Annual Conference on Neural Information Processing Systems. Lake Tahoe, NV, USA. 2012. 1106–1114.
21 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Proceedings of the 3rd International Conference on Learning Representations. San Diego, CA, USA. 2015.
22 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 770–778.
23 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.