融合 word2vec 和注意力机制的图像描述模型

邓珍荣[1,2] 张宝军[1] 蒋周琴[1] 黄文明[1,2]
(桂林电子科技大学计算机与信息安全学院 广西 桂林 541004)[1]
(广西高校云计算与复杂系统重点实验室 广西 桂林 541004)[2]

摘 要 针对当前图像描述任务中,生成描述图像的语句整体质量不高的问题,提出一种融合 word2vec 和注意力机制的图像描述模型。在编码阶段,应用 word2vec 模型描述文本向量化操作,以增强词与词的相关性;应用 VGGNet19 网络提取图像特征,并在图像特征中融合注意力机制,使得模型在每一个时间节点上生成单词时能够突出相对应的图像特征。在解码阶段,应用 GRU 网络作为图像描述任务的语言生成模型,用以提高模型的训练效率和生成句子的质量。在 Flickr8k 和 Flickr30k 两个公共数据集上的实验结果表明,在同一训练环境下,GRU 模型的训练时长比 LSTM 模型节省了 1/3 的时间,在 BLEU 和 METEOR 评价标准上,所提模型的性能得到了显著提升。

关键词 图像描述,word2vec,注意力机制,GRU 模型
中图法分类号 TP391.41 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.04.042

Image Description Model Fusing Word2vec and Attention Mechanism

DENG Zhen-rong[1,2] ZHANG Bao-jun[1] JIANG Zhou-qin[1] HUANG Wen-ming[1,2]
(School of Computer and Information Security, Guilin University of Electronic Technology, Guilin, Guangxi 541004, China)[1]
(Guangxi Colleges and Universities Keys Laboratory of cloud Computing and Complex Systems, Guilin, Guangxi 541004, China)[2]

Abstract For the overall quality of the sentence describing the generated image is not high in the current image description task, and an image description model fusing word2vec and attention mechanism was proposed. In the encoding stage, the word2vec model is used to describe the text vectorization operations to enhance the relationship among words. The VGGNet19 network is utilized to extract image features, and the attention mechanism is integrated in the image features, so that the corresponding image features can be highlighted when the words are generated at each time node. In the decoding stage, the GRU network is used as a language generation model for image description tasks to improve the efficiency of model training and the quality of generated sentences. Experimental results on Flickr8k and Flickr30k data sets show that under the same training environment, the GRU model saves 1/3 training time compared to the LSTM model. In the BLEU and METEOR evaluation standards, the performance of the proposed model in this paper is significantly improved.

Keywords Image description, word2vec, Attention mechanism, GRU model

1 引言

图像描述任务是指使用自然语言对图像内容进行概括性描述,是一项非常具有难度的工作。图像描述模型不仅要对图像中的目标及场景进行描述,还要对目标与目标之间、目标与场景之间的关系进行表达,并能够生成符合一定语法和结构的自然语言句子[1]。图像描述有着广泛的应用领域,在视力缺陷人群的辅助、早期婴幼儿教育、智能人机交互及机器人开发等方面都有广阔的应用前景[2-5]。

图像描述的传统方法是基于模板[6-8]和语义迁移[9-11]两种。这两种方法分别存在生成句子结构单一的弊端;语义不足和理解偏差的弊端。针对上述问题,目前已有很多研究者利用深度学习的方法来完成图像描述任务。

Mao 等[2]提出了一种多模 RNN(m-RNN)模型,他们首先使用 CNN 模型(如 Alex-Net[12],VGG16/19[13]等)提取图像特征,同时生成单词的嵌入向量(Embedding Vector),然后将图像特征和嵌入向量结合在一起,最后将其输入到多模循环神经网络,以预测下一步产生的单词。Karpathy 等[3]提出为图像区域提供自然语言描述系统(NeuralTalk)。它使用 CNN 提取图像特征,利用多模式嵌入的方式将描述文本与图

到稿日期:2018-06-03 返修日期:2018-08-08 本文受广西高校云计算与复杂系统重点实验室项目(yf17106),广西自然科学基金(2018GXNSFAA138132),桂林电子科技大学研究生创新项目(2018YJCX55)资助。
邓珍荣(1977—),女,硕士,研究员,硕士生导师,主要研究方向为计算机软件架构及计算机视觉,E-mail:799349175@qq.com(通信作者);张宝军(1992—),男,硕士生,主要研究方向为计算机视觉、深度学习;蒋周琴(1994—),女,硕士生,主要研究方向为计算机视觉、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、图形图像处理。

像区域对齐，最后利用双向循环神经网络生成对图像区域描述的短句，当然也可以生成对整张图片的描述，只是没有对图像区域描述的效果好。Vinyals 等[4]提出了一种端到端(end to end)的图像描述模型(GoogleNIC)，利用深度卷积神经网络(GoogLeNet)作为编码器以提取图像特征，利用 LSTM 作为解码器来生成自然语言描述。Donahue 等[5]也提出了一种端到端的图像描述模型(LRCN)，直接使用深度卷积神经网络(Alex-Net)提取图像特征，然后将其与词嵌入向量一起输入到 LSTM 网络中。Jia 等[14]从生成描述文本的 LSTM 单元出发，改进了 LSTM 内部结构，提出了 g-lstm 模型，将从图像中提取的语义信息作为额外输入添加到 LSTM 块的每个单元，并以此为导向，引导 LSTM 网络训练模型。Xu 等[15]首次将注意力机制引入到图像描述模型中，注意力机制的本质作用是使模型了解在每个时间节点上应该关注图像的哪些重点区域。Xu 将图像分成 14 * 14 的区域，并利用 CNN 提取每个区域的特征；将提取到的整张图像的特征作为每个时间节点上 LSTM 单元的输入，在输入之前会对每个区域加入注意力机制，即为每个区域分配权值。分配完权值后，有两种输入方式：1)取其权值最大的区域作为当前时间节点的输入，即“hard-attention”；2)将每个区域特征与其对应的权值相乘、求和，然后将计算结果作为当前时间节点的输入，即“soft-attention”。

以上工作大部分采用 CNN-RNN 相结合的方式，利用 CNN 提取图像特征，然后将图像特征和文本向量放入到 RNN 中训练，以生成描述文本。但是他们大部分都把精力放到 RNN 网络的优化上，忽略了提取的图像特征和文本向量化的好坏对最后生成的描述句子的重要影响，因此会产生如下问题：1)利用 one-hot 方法处理连续型语句时会造成编码稀疏，导致文本词间关系被弱化，淹没潜在文本特征；2)在按时间序列生成单词时，每个时间节点所使用的图像特征信息都是“均等的”，但是在不同的时间节点下，图像不同区域的特征信息所起的作用是不一样的，因此要选择性地突出当前时间节点需要的图像特征信息。针对这些问题，文中采用了“端到端”的训练思想。在编码阶段：1)引入 word2vec[16-17] 词向量模型，通过 word2vec 词向量模型对图像描述的文本进行向量化处理，能够增强词与词之间相关性；2)对图像特征信息融合注意力机制，使得在训练模型时，在每个训练过程中都有图像的重点区域来“帮助”预测单词。在解码阶段：为了提高模型的训练效率和生成句子的质量，使用 GRU(Gated Recurrent Unit)[18] 网络作为语言生成模型。

2　融合 word2vec 和注意力机制的图像描述模型

针对描述文本编码稀疏、图像在每个时间节点的特征不突出和语言生成模型训练效率低的问题，文中提出一种融合 word2vec 和注意力机制的图像描述生成模型，以提高模型的训练效率和生成句子的质量。

2.1　word2vec 模型

传统的图像描述文本向量化是使用 one-hot 方法，对一句话，如“the cat on the chair”，进行编码，the:[1,0,0,0,0]，cat:[0,1,0,0,0]，on:[0,0,1,0,0]，以此类推，对所有的单词进行编码。其只是单地地对文本进行向量化操作，这种编码方式存在明显的问题，不能表示词与词之间的关系，这就导致许多潜在的对生成图像描述有用的文本特征被淹没，从而降低了模型生成文本的质量。

word2vec 提供了 CBOW 和 skip-gram 两种训练模型，CBOW 模型是根据词 $w(t)$ 前后各 c 个词来预测当前词；而 skip-gram 模型与之相反，它是根据词 $w(t)$ 来预测其前后各 c 个词。CBOW 模型图和 skip-gram 模型的示意图分别如图 1 和图 2 所示。

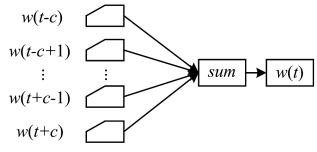

图 1　CBOW 模型示意图

Fig. 1　Schematic diagram of CBOW model

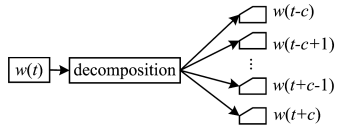

图 2　skip-gram 模型示意图

Fig. 2　Schematic diagram of skip-gram model

本文使用 skip-gram 模型，并用 Flickr8k[19] 和 Flickr30k[20] 数据集的文本部分作为语料库来训练词向量模型。假设一个句子 S 由 T 个词 $w_1,w_2,w_3,\cdots,w_T$ 组成，则 c-skip-n-gram 为：

$$C=\{w_{t_1},w_{t_2},w_{t_3},\cdots,w_{t_n}\mid\sum_{i=1}^{n-1}t_{i+1}-t_i\leqslant c+1\} \tag{1}$$

其中，c 代表选取当前词上下文的长短，在 skip-gram 模型中就是指 c-skip-n-gram 中 n 的大小，即训练文本的大小。当 $n=0$ 时，skip-n-gram 完全退化为 skip-gram 模型；gram($n=2$)则适合用小规模语料集。$n=2$ 时，其目标函数是最大化平均对数概率：

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leqslant j\leqslant c}\log p(w_{t+j}\mid w_t) \tag{2}$$

通过对 skip-gram 模型的训练得到隐层的参数，也即词的向量表示。

2.2　GRU 模型

图像描述领域中，生成描述文本时，使用 LSTM 模型作为语言生成模型，LSTM 表现出了较好的性能，但是因为 LSTM 内部具有繁琐的门控单元，使得在训练图像描述模型时效率较低。在图像描述任务中，GRU 模型表现出良好的性能，且拥有简单的内部结构，因此生成文本描述时，使用 GRU 网络作为语言生成模型。

GRU 单元用“更新门(Update Gate)”来代替 LSTM 单元中的“遗忘门(Forget Gate)”和“输入门(Input Gate)”，并把“细胞状态(Cell State)”和“隐藏状态(h_t)”合并在一起，依然具有在每个时间节点上调节内部信息流的门控单元，GRU 能够在不同时间下，适应性地捕捉每个循环单元的依赖关系，但是它没有单独的存储单元。

1)首先,介绍 GRU 的第一个门,即“重置门(Reset Gates)”。⊗代表元素乘法,σ 是 sigmoid 函数,当 r_t 接近 0 时,之前的隐藏层的信息就会被丢弃,允许模型丢弃一些与未来无关的信息,使其忘记以前计算的状态。其计算公式如下:

$$r_t=\sigma(W_r\cdot[h_{t-1},x_t]) \tag{3}$$

2)GRU 的第二个门,即“更新门(Update Gates)”。更新门控制当前时间节点的隐藏层输出 h_t 需要保留之前的隐藏层信息量。z_t 接近 1,相当于我们把之前的隐藏层的信息全部拷贝到当前的时刻,这样可以学习到长距离依赖。z_t 的计算公式如下:

$$z_t=\sigma(W_z\cdot[h_{t-1},x_t]) \tag{4}$$

3)计算候选隐藏层(Candidate Hidden Layer) $\tilde{h}_t$。候选隐藏层与 LSTM 类似,可以看成是当前时间节点的新信息,其中 r_t 用来控制需要保留多少之前的记忆,若 r_t 为 0,则 h_t 只包含当前词 $h=g(vA_t+v'A_t')$ 的信息,其计算公式如下:

$$\tilde{h}_t=\tanh(W_{\tilde{h}}\cdot[r_t*h_{t-1},x_t]) \tag{5}$$

4)z_t 控制需要从前一时间节点的隐藏层 h_{t-1} 中遗忘的信息量,以及需要加入多少当前时间节点的隐藏层信息 $\tilde{h}_t$,以得到隐藏层输出信息 h_t。与 LSTM 的不同之处在于,GRU 中没有“输出门(output gates)”。h_t 的计算公式如下:

$$h_t=(1-z_t)\cdot h_{t-1}+z_t\cdot\tilde{h}_t \tag{6}$$

图 3 是 GRU 单元的更新过程。

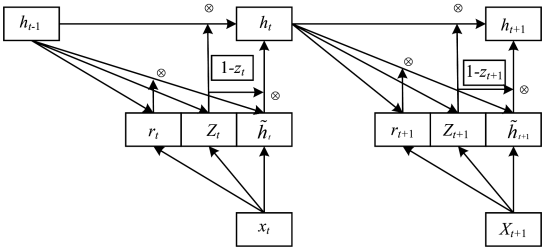

图 3 GRU 模型结构

Fig. 3 Structure of GRU model

2.3 注意力机制

传统的图像描述模型在对句子中的每一个词进行预测时,没有考虑突出与这个词相对应的图片区域,接受的输入仅仅是上一步的隐藏层输出。注意力机制的本质作用是将图像的文本描述与图像的不同区域做一个映射,使得图像的文本描述可以更好地对应到图像的相应区域。传统的图像描述模型利用 $h_t=f(h_{t-1},x_{t-1})$ 预测 t 时刻的单词,h_t 是当前时间节点隐藏层的输出(输出当前时间状态下需要预测的词),h_{t-1} 是前一个时间节点隐藏层的输出,图像相关的特征向量 $ImageVec=CNN(I)$ 只在最初时被输入到 RNN 模型,以后每个时间节点预测单词时,均是利用最初输入的图像特征。然而,注意力机制对每一步预测时的输入均进行改变,从原来的 $h_t=f(h_{t-1},x_{t-1})$ 形式改为 $h_t=f(h_{t-1},c_{t-1})$,让模型在每个时间节点预测单词时能关注图像中重点的区域信息,而不仅仅是在训练模型的最开始就把整个图像的特征向量输进 GRU 网络。下文将描述 c_t 是如何计算的。其注意力机制结构如图 4 所示。

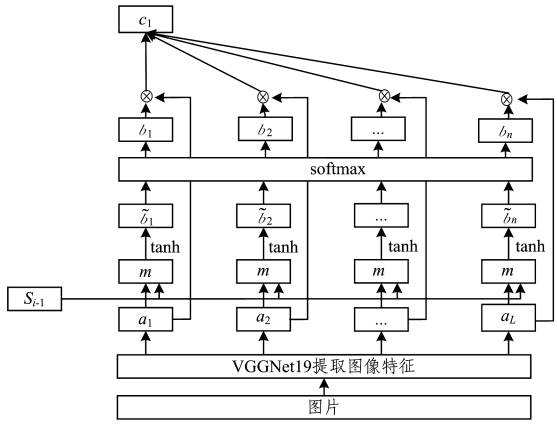

图 4 注意力机制模型的结构

Fig. 4 Structure of attention mechanism model

在说明注意力机制之前,先对一些数学符号做一些约定。首先训练 attention 机制获取 c_t 时,模型接受的输入仍然是一张图像和其对应的描述,同样地,每一个描述句子都使用 word2vec 模型向量化,这里记为 S。在使用 Vgg19/Vgg16 提取图像特征时,我们并没有使用完整的 Vgg19/Vgg16 结构,而是直接从较低的卷积层提取图像特征,并把提取到的特征划分成各特征图(Feature Map)子块。假设从一张图像中得到的特征图的数量为 L,维度为 D,则每张特征图代表着图片中的相应位置。用 a 表示特征图,得到整张图片的待关注区域的集合:

$$a=\{a_1,a_2,\cdots,a_L\},a_i\in R^D \tag{7}$$

该集合中的每一个向量都对应着图像中某个区域里的特征,我们的目的就是将描述的句子 S 中的每个单词对应到图片的特定区域 a_i 上。首先,分别计算出 S 中的每个单词与 a_i 的相似度或者相关性,也即图 4 中 m 的操作,这里 m 是一种相似度匹配计算,且是相互独立的;然后进行激活函数,本文使用 tanh 激活函数,得到 $\tilde{b}_i$。为了更加突出当前时间节点上重要特征的权值以及使得权值归一化,对 $\tilde{b}_i$ 进行归一化处理,使得 $\sum_{i=1}^{L}b_i=1$。最后将得到的权值 b_i 与特征 a_i 相乘后累加,得到 c_t。计算公式如下:

$$\tilde{b}_i=\tanh(w_{sm}S+w_{am}a_i) \tag{8}$$

$$b_{ti}=\text{soft max}(\tilde{b}_i)=\frac{\exp(\tilde{b}_{ti})}{\sum_{k=1}^{L}\exp(\tilde{b}_{tk})} \tag{9}$$

$$c_t=\emptyset(\{a_i\},\{b_i\})=\sum_{i=1}^{L}b_i\cdot a_i \tag{10}$$

需要说明的是对某个 $\tilde{b}_i$ 来说,其输入是 a_i 本身和前一步 GRU 的输出 S,其输出是 $\tilde{b}_i$,每个时间节点的 $\tilde{b}_i$ 值都不一样,且与上一步的输出有关。例如:当模型的上一步输出单词“drink”时,下一步就给图像中类似“water”“liquor”这样的位置赋予更高的权重,以利于模型在下一步预测出“water”或者“liquor”。

2.4 语言生成模型

融合 word2vec 和注意力机制的图像描述模型包含三部

分内容：第一部分是对描述语句中的每个单词使用 word2vec 方法生成词向量，以增强词与词之间的联系；第二部分是利用 VGGNet19 模型提取图像特征，并将特征区域划分为 L，将所得特征和向量化的描述文本一起放入注意力机制进行训练，使得描述词在特定特征上有更高的权重，并将得到的含有权重的特征放入 GRU 网络中；第三部分是使用 GRU 网络生成描述句子。由于在前文中已经对第一部分和第二部分进行了介绍，本节重点介绍第三部分，其模型结构如图 5 所示。

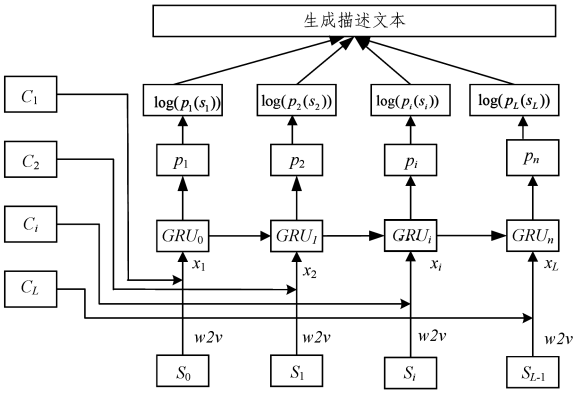

图 5　图像描述模型结构

Fig. 5　Structure of image description model

在图像描述中，首先用“EOF”设定 S_0 和 S_L 作为句子特殊的开始词和结束词。将第一部分生成的词向量 S 与第二部分含有权重的图像特征 C 相融合，即 $x=g(c,s)$，然后将 x 输入到 GRU 单元，即 $p_t=GRU(S_{t-1},x_t)$。其输出经过 softmax 函数生成下一个时间节点上的单词，然后使用交叉熵函数计算该单词与下一个时间节点所对应的真实单词的损失。将整个描述句子上的所有损失的总和即为这次迭代的总的代价函数，其后使用 BPTT 算法对 GRU 网络中的参数进行更新。其模型优化的目标函数为：

$$\theta^* = \arg\max_{\theta} \sum_{C,S} \log p(S|C;\theta) \tag{11}$$

其中，θ 为模型要学习的参数，S 为正确描述图片的句子，C 为含有权重的特征图：$C\in\{c_1,c_2,\cdots,c_t\}$。因为 S 是不固定长度的，而且 S_t 是依赖 S_{t-1} 生成的，所以利用链式法则，式(11)的后半部分可以写成：

$$\log p(S|C;\theta) = \sum_{t=0}^{N} \log p(S_t|C,\theta,S_0,S_1,\cdots,S_{t-1}) \tag{12}$$

其中，N 为句子长度，S_i 为句子包含的每一个词，其损失函数为：

$$L(C,S) = -\sum_{t=1}^{N} \log p_t(S_t) \tag{13}$$

其目标是更新 GRU 的参数，使得每一个正确的词出现的概率最大，也即让此损失函数越小。

在 Ovinyals 等[12]的训练过程中，图像特征只是在训练的最开始输入到 LSTM 模型中，其目的是为了防止过拟合和图像中的噪声。本文首先对提取到的图像特征采取注意力机制，因此在每个时间节点输入图像特征时，避免了图像特征中的噪音部分，以及在生成单词时，权重大的特征区域被重点突出，避免了过拟合，使得生成的单词更加准确。

在训练效率上，本文采用了 GRU 模型，这是因为该模型的内部结构较为简单。在相同的环境下，该模型的训练时间要少于 LSTM 模型，并且模型性能超过了 LSTM 模型，3.3.2 节将给出两者训练时间的对比数据。

3　实验结果及分析

3.1　数据集

在图像描述领域，一般使用 Flickr8k 和 Flickr30k 数据集，本文也遵循此规则。这两种数据集均由雅虎发布，并且可以免费使用。两种数据集的结构一样，只是数据量不一样，Flickr30k 含有 30 多万张图像，Flickr30k 含有 8 千张图像。两种数据集中的每一张图像都配有 5 句亚马逊的人工注解，也即概括图像大意的语言。在实验过程中，训练集、测试集和验证集的数据量分配方式如表 1 所列。

表 1　训练集、测试集和验证集的分配方法

Table 1　Distribution method of training data, test data and verification data

（单位：万张）

数据集	Flickr8k	Flickr30k
训练集	0.6	2.8
验证集	0.1	0.1
测试集	0.1	0.1

3.2　评价方法

在本文的实验评价标准中，采用 BLEU[21] 和 METEOR[22] 两种评价方法对生成的描述文本进行整体评价。BLEU 方法是在翻译中使用的评价方法，因为图像描述属于特殊的翻译，所以利用 BLEU 评价方法亦可。该方法的思想是：首先计算正确的描述文本和生成的描述文本之间 n-gram 的匹配数量，然后将匹配数量与生成描述文本中 n-gram 数量的比值作为评价指标。METEOR 方法是基于召回率提出的，基于召回率的标准与基于精度的标准（如 BLEU）相比，其判断的结果与人工判断的结果具有较高的相关性，因此利用 METEOR 方法来评价生成文本的质量。

3.3　实验结果及分析

3.3.1　实验生成句子及分析

本文利用所提模型生成描述文本，如图 6 和图 7 所示。图 6 是利用 Flickr8k 数据集生成的描述，图 7 是利用 Flickr30k 数据集生成的描述，其中 K 代表训练样本集中人工标注的表述，Q 代表模型生成的描述。

由图 6 和图 7 中生成的描述性语句可以看出，本文模型能够较准确地描述图像中的主要场景以及场景与目标之间的关系，有的描述甚至超过了人工标注。如图 6(b)所示，K_1 只描述了目标和场景，没有结合目标和场景，缺少“玩耍”这个动作；K_2 属于抽象的描述，这对于模型来说很难学习到这个抽象的行为；K_3 较好地描述了整张图像；K_4 和 K_5 只关注了目标的动作和场景，并没有对目标的属性进行描述。本文所提模型与 K_3 相似，较好地表达了目标的属性、场景以及目标与场景之间的关系，这种描述属于较为准确的描述；然而也有一些较为不好的描述，图 7(b)只是描述了目标、动作和场景，并没有对目标的属性进行描述。

K_1: A backpacker in the mountains using his hiking stick to point at a glacier.
K_2: A backpacker points to the snow-capped mountains as he stands on a rocky plain.
K_3: A hiker is pointing towards the mountains.
K_4: A hiker poses for a picture in front of stunning mountains and clouds .
K_5: A man with a green pack using his pole to point to snowcapped mountains.
Q: A backpacker uses a pole to point to a mountain covered by snow in the distance.

(a)

K_1: A boy in his blue swim shorts at the beach.
K_2: A boy smiles for the camera at a beach.
K_3: A young boy in swimming trunks is walking with his arms outstretched on the beach.
K_4: Children playing on the beach.
K_5: The boy is playing on the shore of an ocean.
Q: Smiling boy in swimming trunks is playing on the shore of an ocean.

(b)

图 6 Flickr8k 数据集上生成的部分句子示例

Fig. 6 Examples of candidate sentence generated with Flickr80k data set

K_1: A woman figure skater in a blue costume holds her leg in the air by the blade of her skate.
K_2: A lady in a blue outfit is practicing figure skating.
K_3: A man and a woman ice skating on a rink.
K_4: A very graceful ice skater.
K_5: Woman dancing on the ice.
Q: A girl is dancing on the rink.

(a)

K_1: A little boy with a pacifier in his mouth stands in a park with a goose.
K_2: A little child with a red cap and a duck are in the grass.
K_3: A child looking at camera and bird in background.
K_4: The little boy is in the park near a duck.
K_5: A young boy was watching a goose.
Q: A little boy with a pacifier stands with a duck in the park.

(b)

图 7 Flickr30k 数据集上生成的部分句子示例

Fig. 7 Examples of candidate sentence generated with Flickr30k data set

3.3.2 实验统计结果及分析

本文实验环境的操作系统为 Ubuntu 14.04.1，处理器为 Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20 GHz，运行内存(RAM)为 512 GB，显卡为 NVIDIA Tesla P100-SXM2 16 GB，程序运行框架为 Tensorflow 1.3.0 平台。

使用本文所提模型生成句子时，将其分别与 multimodal-RNN，m-RNN，gLSTM，LRCN，Google NIC，Hard-Attention，Soft-Attention 模型在 Flickr8k 和 Flickr30k 数据集上进行指标对比。比对实验数据显示，本文所提模型在多个评价方法上的评价指标大部分超过了对比模型。

从表 2 和表 3 可以看出，在 Flickr8k 和 Flickr30k 数据集中，本文模型在 BLEU_1，BLEU_2，BLEU_3，BLEU_4，METEOR 评价标准上都得到了较高的分数(评价标准获取的分数越高，说明模型越好)。由于 multimodal-RNN 模型提出时间较早，在生成图像描述时采用双向循环神经网络，因此本文模型整体上要优于 multimodal-RNN 模型。gLSTM 模型只是改进了 LSTM 的内部结构，在训练时通过提取图像语义信息来引导模型训练，但是并没有优化描述文本的编码方式，因此该模型的整体效果没有本文模型好。Google NIC 模型在图像描述领域具有非常重要的意义，之后的模型基本上都是基于它的框架结构设计的，因此其具有很好的代表性，但是其效果不如后者。Hard-Attention 和 Soft-Attention 模型与本文模型较为接近，但本文模型加入了 word2vec 向量化方法，这种编码方法要优于其他 word embedding 方法，因此所提模型优于 Hard-Attention 和 Soft-Attention 模型，但是在 BLEU_1 分数上并没有 Hard-Attention 和 Soft-Attention 模型高。

表 2 不同模型在 Flickr8k 图像集上的性能评价

Table 2 Performance evaluation of different models on Flickr8k dataset

模型	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR
multimodal-RNN[18]	57.9	38.3	24.5	16.0	16.7
m-RNN-VGGNet16[15]	—	—	—	—	—
gLSTM[23]	64.7	45.9	31.8	21.6	20.2
LRCN[20]	—	—	—	—	—
Google NIC[19]	63.0	41.0	27.0	—	—
Hard-Attention[15]	**67.0**	45.7	31.4	21.3	**20.3**
Soft-Attention[15]	**67.0**	44.8	29.9	19.5	18.9
ours	66.8	**46.0**	**32.1**	**22.0**	20.2

表 3 不同模型在 Flickr30k 数据集上的性能评价

Table 3 Performance evaluation of different models on Flickr30k dataset

模型	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR
multimodal-RNN	57.3	36.9	24.0	15.7	15.3
m-RNN-VGGNet16	60	41	28	19	—
gLSTM	64.6	44.6	30.5	20.6	17.9
LRCN	58.7	39.1	25.1	16.5	—
Google NIC	66.3	42.3	27.7	18.3	—
Hard-Attention	**66.9**	43.9	29.6	19.9	18.5
Soft-Attention	66.7	43.4	28.8	19.1	18.5
ours	66.6	**45.0**	**31.0**	**21.0**	**18.9**

另外，在生成描述句子时，本文分别使用了 LSTM 网络和 GRU 网络，并进行了对比。利用前文提到的配置训练模型，从表 4 可以看出两个生成模型在性能上基本相似，GRU 模型的性能略好，但是差别不大。但是在训练时间上，GRU 模型所用时间明显比 LSTM 少，大约只用了 LSTM 2/3 的时间就达到了与 LSTM 相近的水平。因此同等条件下，GRU 模型在效率上优于 LSTM 模型。

表 4　LSTM 和 GRU 的性能对比

Table 4　Performance comparison between LSTM and GRU

（单位：h）

模型	时间	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR
Att-LSTM	168	66.5	45.1	30.8	20.7	18.8
Att-GRU	113	66.6	45.0	31.0	21.0	18.9

结束语　图像描述任务是一项非常有难度的任务，该任务要求既要了解计算机视觉，又要了解自然语言处理。本文按照“编码-解码”的整体框架思想：在编码阶段，提取图像特征部分加入了注意力机制，文本向量化采用 word2vec 方法；在解码阶段，使用了 GRU 网络，使得模型不仅提高了描述文本的质量，也减少了训练时间，提高了训练效率。但是本文在提取图像特征时是直接利用已经训练好的 VGGNet19 模型，虽然可以基本满足任务需求，但是 VGGNet19 模型的深度较浅，且没有针对性地对模型参数进行调优。

由于 Flickr8k 和 Flickr30k 数据集的数据量还是较小，下一步将采用更大规模的数据集 COCO 进行模型训练和验证。在提取图像特征方面计划采用更深层的 ResNet 网络来提取图像特征，优化 ResNet 网络底层参数，使得提取的特征更加符合图像描述任务，并继续改进注意力机制。

参 考 文 献

[1] OLIVA A，TORRALBA A. The role of context in object recognition[J]. Trends in Cognitive Sciences，2007，11(12)：520-527.

[2] MAO J，XU W，YANG Y，et al. Deep captioning with multimodal recurrent neural networks (m-rnn)[J]. Preprint arXiv：1412.6632v5.

[3] KARPATHY A，LI F F. Deep visual-semantic alignments for generating image descriptions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston，Massachusetts，2015：3128-3137.

[4] 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. Boston，Massachusetts，2015：3156-3164.

[5] DONAHUE J，HENDRICKS L A，GUADARRAMA S，et al. Long-term recurrent convolutional networks for visual recognition and description[C]// IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2015：2625-2634.

[6] 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. Heraklion，Crete，reece：Springer，2010：15-29.

[7] MITCHELL M，HAN X F，DODGE J，et al. Midge：generating image descriptions from computer vision detections[C]// Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon，France：ACL，2012：747-756.

[8] KULKARNI G，PREMRAJ V，ORDONEZ V，et al. BabyTalk：understanding and generating simple image descriptions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2013，35 (12)：2891-2903.

[9] KUZNETSOVA P，ORDONEZ V，BERG A C，et al. Generalizing image captions for image-text parallel corpus[C]// Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia，Bulgaria：ACL，2013：790-796.

[10] MASON R，CHARNIAK E. Nonparametric method for data driven image captioning[C]// Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore，Maryland，USA：ACL，2014：592-598.

[11] SOCHER R，KARPATHY A，LE Q V，et al. Grounded compositional semantics for finding and describing images with sentences[J]. Transactions of the Association for Computational Linguistics，2014，2：207-218.

[12] OVINYALS A. TOSHEV S . BENGIO D. Erhan，Show and tell：a neural image caption generator[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston，Massachusetts，2015：3156-3164.

[13] SIMONYAN K，ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]// International Conference on Learning Representations (ICLR). 2014.

[14] JIA X，GAVVES E，FERNANDO B，et al. Guiding the Long-Short Term Memory model for Image Caption Generation[C]// IEEE International Conference on Computer Vision (ICCV). 2015：2407-2415.

[15] XU K，BA J，KIROS R，et al. Show，attend and tell：Neural image caption generation with visual attention[C]// International Conference on Machine Learning(ICML). 2015.

[16] MIKOLOV T，KOPECK J，BURGET L，et al. Neural network based language models for highly inflective languages [C]// IEEE International Conference on Acoustics. IEEE Computer Society，2009：126-129.

[17] HINTON G E，MCCLELLAND J L，RUMELHART D E. Distributed Representations[M]// Parallel Distributed Processing：Explorations in the Microstructure of Cognition. Cambridge：MIT Press，1986.

[18] CHO K，MERRIENBOER B V，GULCEHRE C，et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha：Association for Computational Linguistics，2014：1724-1734.

[19] LIN T Y，MAIRE M，BELONGIE S et al. Microsoft coco：common objects in context[C]// Proceedings of the 13th European Conference on Computer Vision. Zurich，Switzerland：Springer，2014：740-755.

[20] YOUNG P，LAI A，HODOSH M，et al. From image descriptions to visual denotations：new similarity metrics for semantic inference over event descriptions[J]. Transactions of the Association for Computational Linguistics，2014，2：67-78.

[21] 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. Philadelphia，Pennsylvania：Association for Computational Linguistics，2002：311-318.

[22] BANERJEE S，LAVIE A. METEO R：an automatic metric for MT evaluation with improved correlation with human judgments [C]// Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and /or SUMMARIZATION. ANN ARBO：ACL，2005：65-72.