

# 基于枢轴语言的图像描述生成研究\*

张凯，李军辉，周国栋

(苏州大学 计算机科学与技术学院, 江苏省 苏州市 215006)

**摘要:** 当前图像描述生成的研究主要仅限于单语言(如英文), 这得益于大规模的已人工标注的图像及其英文描述语料。该文探索零标注资源情况下, 以英文作为枢轴语言的图像中文描述生成研究。具体地, 借助于神经机器翻译技术, 该文提出并比较了两种图像中文描述生成的方法: (1) 串行法, 该方法首先将图像生成英文描述, 然后由英文描述翻译成中文描述; (2) 构建伪训练语料法, 该方法首先将训练集中图像的英文描述翻译为中文描述, 得到图像-中文描述的伪标注语料, 然后训练一个图像中文描述生成模型。特别地, 对于第二种方法, 该文还比较了基于词和基于字的中文描述生成模型。实验结果表明, 采用构建伪训练语料法优于串行法, 同时基于字的中文描述生成模型也要优于基于词的模型, BLEU\_4 值达到 0.341。

**关键词:** 图像描述生成; 机器翻译; 神经网络; 枢轴语言

## Image Caption via Pivot Language

ZHANG Kai<sup>1</sup>, LI Junhui<sup>1</sup>, and ZHOU Guodong<sup>1</sup>

(School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu, 215006, China)

**Abstract:** Due to the publically available large-scale image dataset with manually labeled English captions, most studies on image caption aim at generating captions in a single language (e.g., English). In this paper, we explore zero-resource image caption to generate Chinese captions via English as the pivot language. Specifically, we propose and compare two approaches by taking advantage of recent advances in neural machine translation. The first approach, called pipeline approach, first generates English caption for a given image, then it translates the English caption into Chinese. The second approach, called building pseudo-training set approach, first translates all English captions in training sets and development set into Chinese to obtain image-Chinese caption datasets, therefore it then could directly train a model to generate Chinese caption for a given image. Moreover, we discuss both word-based and character-based Chinese caption generation models. Experimental results shows that the second approach, i.e., building pseudo-training set, is superior to the pipeline approach while the character-based Chinese caption generation model outperforms the word-based one by obtaining performance of 0.341 in BLEU\_4 score.

**Keywords :** image caption; machine translation; neural network; pivot language

---

\* 收稿日期: 定稿日期:

基金项目: 国家自然科学基金 (61401295)

作者简介: 张凯 (1993——), 男, 硕士研究生, 主要研究方向为自然语言处理, 机器翻译; 李军辉 (1983——), 男, 副教授, 主要研究方向为自然语言处理, 机器翻译; 周国栋 (1967——), 男, 教授, 主要研究方向为自然语言处理。

## 0 引言

自然语言处理 (Natural Language Processing, NLP) 和计算机视觉 (Computer Vision, CV) 是人工智能领域研究的两大热点。当前, 跨领域研究已经成为了未来研究的一种趋势, 引起了研究者极大的兴趣。图像的语言描述 (Image caption) 是结合计算机视觉和自然语言处理的一种跨领域研究, 该技术最早由 Farhadi<sup>[1]</sup>等人提出, 给定二元组  $(I, S)$ , 其中  $I$  表示图像,  $S$  表示对该图像的描述, 模型要完成  $I \rightarrow S$  的映射。学习到图片对应的描述, 然后用训练完成的模型实现随机给的一张图片, 描述图片的内容。

“看图说话”对正常人来说非常简单, 但这对于计算机来说是一项极大的挑战, 计算机不仅仅要识别图片的内容, 还有用人类的逻辑思维, 描述出人类可读的句子。

当前图像描述生成的主流方法是基于神经网络方法, 需要大量图像到目标语言的训练数据才能获得较好的性能。由于图像描述语料的目标端语言仅为英文, 目前图像描述生成的研究仅局限于英文图像描述的生成。但是在很多情况下, 图像到目标语言为非英文的标注语料资源为零, 因此研究图像到目标语言为非英文的描述生成是一项严峻的挑战。本文提出在只有图像到英文描述语料, 而没有图像到中文描述语料的情况下, 如何利用机器翻译技术进行图片的跨语言描述生成, 即生成图像的中文描述。基于这种假设下, 本文提出两种方法生成图像的中文描述。一种是串行法, 该方法先训练图像到英文描述模型, 然后利用训练好的模型, 将随机给定的一张图像, 生成其对应的英文描述, 再利用翻译模型将得到的英文描述翻译成中文描述; 另外一种构建伪语料法, 该方法先利用翻译模型将图片-英文描述语料中的英文描述翻译成中文描述, 这样得到同等规模的图像-中文描述伪标注语料; 这样可以训练图像到中文描述模型, 然后利用训练好的模型, 随机给定一张图像, 生成其对应的中文描述。

对于第二种构建伪语料方法而言, 在训练的时候考虑到一个英文单词可能对应

于一个中文短语, 因此, 生成中文描述的时候, 必须结合中文的词组的形式, 可以先将中文句子分词, 进行基于词的图像中文描述生成; 又考虑到中文中每个字都可以和其他字组合成不同的词, 因此每个字都是一个独立的个体, 也可以不用对句子分词, 进行基于字的图像中文描述生成。

## 1 相关工作

本节先描述图片描述生成技术的相关工作, 再描述机器翻译领域基于枢轴语言的翻译技术。

与神经机器翻译类似, 图片描述 (image caption) 采用的神经网络模型通常由编码器和解码器两部分组成。编码器使用卷积神经网络 (CNN) 将每张图片转化为一个固定长度的向量, 也称图像的隐层表示; 解码器使用循环神经网络 (RNN) 将编码器输出的固定长度的向量解析为目标语言的句子。Mao<sup>[2]</sup>等人提出了在基于传统 CNN 编码器-RNN 解码器的神经网络模型的基础上, 提出并使用多模态空间为图像和文本建立联系。Vinyals<sup>[3]</sup>等人提出了神经图像描述 (Neural Image Caption, NIC) 模型, 该模型将图像和单词投影到多模态空间, 并使用长短时记忆网络生成英文描述。Jia<sup>[4]</sup>等人提出了 gLSTM 模型, 该模型使用语义信息指导长短时记忆网络生成描述。Xu<sup>[5]</sup>等人将注意力机制引入解码过程, 使得描述生成网络能够捕捉图像的局部信息。Li<sup>[6]</sup>等人构建了首个中文图像摘要数据集 Flickr8kCN, 并提出中文摘要生成模型 CS-NIC, 该方法使用 GoogleNet<sup>[7]</sup>对图像进行编码, 并使用长短时记忆网络 (LSTM) 对图像描述生成过程建模。

在零资源或低资源的情况下, 利用其它语言训练神经机器翻译模型近年来引起了广泛关注。Orhan Firat<sup>[8]</sup>等人提出多途径、多语言神经机器翻译, 所提出的方法使得单个神经翻译模型能够在多种语言之间进行翻译, 其中许多参数只与语言数量成线性增长。通过在所有语言对之间共享单个关注机制来实现。Cheng<sup>[9]</sup>等人提出基于轴的神

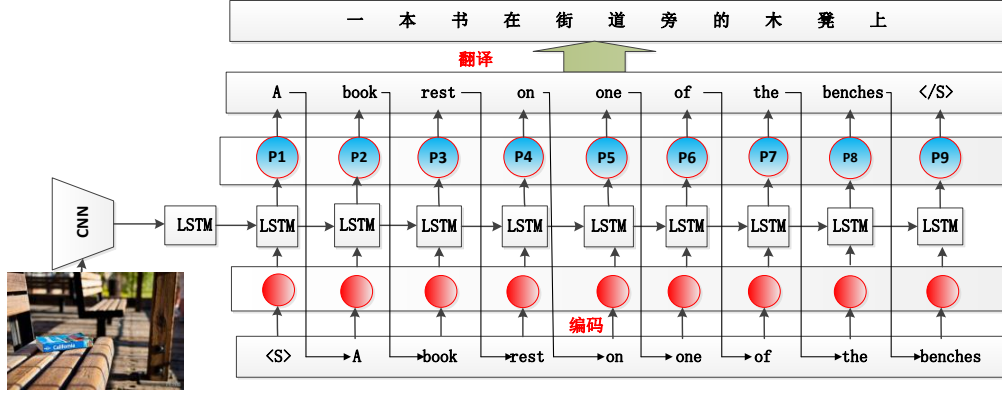


图 1 串行法图像中文描述生成的模型框架

机器翻译，为了提高源端到目标端翻译质量，同时提高源端到轴和轴到目标端的翻译质量。Nakayama and Nishida<sup>[10]</sup>等人实现零资源机器翻译，利用图像作为轴心，并训练多模态编码器以共享通用的语义表示。

目前并未发现零资源情况下，图像描述生成方面的研究。

## 2 图像中文描述生成方法

在本文提出的两种图像中文描述生成方法中，都包含两部分模块，分别对应翻译模型和图像描述生成模型。翻译模型将英文描述翻译成中文描述，图像生成描述模型将图像生成相应语言的描述。目前具有代表性的翻译模型包括基于 RNN 和注意力机制的序列到序列模型<sup>[11]</sup>、基于 CNN 的序列到序列模型<sup>[12]</sup>以及基于自注意力机制的序列到序列模型<sup>[13]</sup>。由于翻译模型本身并不是本文的研究重点，本文调用谷歌翻译 API (<https://translate.google.cn>) 获取描述的中文翻译。

### 2.1 图像描述生成模型

图像描述生成模型利用编码-解码的架构。首先编码器对图像进行编码，提取图像视觉特征；再使用解码器对视觉特征进行解码，生成句子。给一张图片和对应的描述，编码器-解码器模型直接用公式 (1) 最大化目标：

$$\theta^* = \arg \max_{\theta} \sum_{(I,S)} \log p(S|I;\theta) \quad (1)$$

其中  $\theta$  是模型要训练的参数， $I$  表示图片，

$S = \{S_0, \Lambda S_N\}$  表示对应图像的描述。因为

$S$  表示任意长度的句子，它的长度不固定。利用链式法则，可以将联合概率分布的对数可能性分解为有序条件：

$$\log p(S|I) = \sum_{t=0}^N \log p(S_t|I, S_0, \Lambda, S_{t-1}) \quad (2)$$

本文使用卷积神经网络(CNN)和长短期记忆网络(LSTM)对图像中文描述生成过程建模。在  $t=-1$  时刻， $X_{-1} = \text{CNN}(I)$  表示 LSTM 在这一时刻接收图像内容， $t \geq -1$  时，网络的计算过程为  $h_t = \text{LSTM}(x_t, h_{t-1})$ ，其中

$x_t \in \mathbf{R}^m$  为  $t$  时刻的输入，即  $x_t = W_e S_t$ ，

$S_t \in \mathbf{R}^n$  为图像对应描述的每个词编码成 one-hot<sup>[14]</sup> 向量，其维度大小等于字典的大小  $n$ 。在每个句子序列中我们用  $S_0$  表示特殊的开始符，用  $S_N$  表示特殊的结束符。

$W_e \in \mathbf{R}^{n \times m}$  表示词向量字典， $m$  表示词向量

大小， $h_t \in \mathbf{R}^d$  为隐藏单元状态， $\text{LSTM}(\cdot)$

函数表示为下列形式：

$$i_t = \sigma(W_i x_t + U_i h_{t-1})$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1})$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1})$$

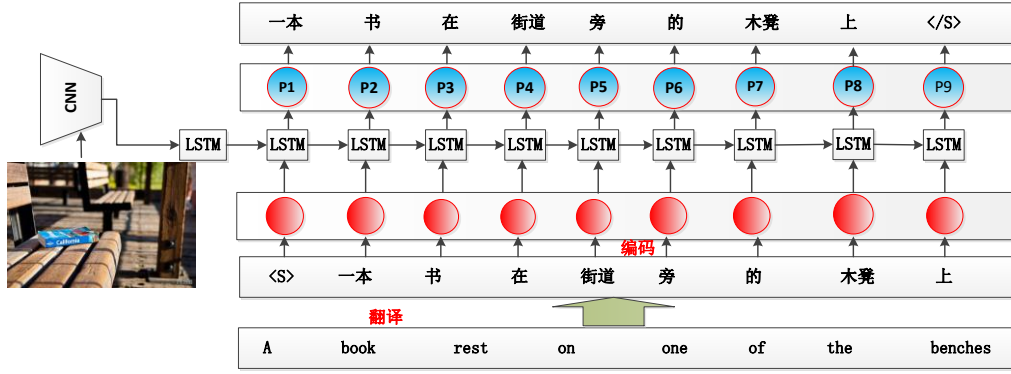


图 2 基于词的模型框架

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c x_t + U_c h_{t-1})$$

$$h_t = o_t \odot \tanh(c_t) \quad (3)$$

其中  $i_t \in \mathbf{R}^d$  为输入门,  $f_t \in \mathbf{R}^d$  为遗忘门,

$o_t \in \mathbf{R}^d$  为输出门,  $c_t \in \mathbf{R}^d$  为单元状态门,

$W_{i, f, o, c} \in \mathbf{R}^{c \times m}$  和  $U_{i, f, o, c} \in \mathbf{R}^{c \times d}$  矩阵为要训练的参数,  $c$  为词向量长度,  $d$  为隐藏状态长度, ( $c=d$ )。公式(3)得到的隐藏单元状态和全连接层网络的权重矩阵  $W_{net} \in \mathbf{R}^{d \times n}$

相乘, 馈送给 Softmax 函数, 产生所有单词的概率分布, 如公式(4)所示。

$$P_{t+1} = \text{Softmax}(W_{net} h_t) \quad (4)$$

其中  $W_{net}$  是要训练的参数。

模型的损失值是每一个正确单词的预测概率的对数负值总和, 如公式(5)所示。

$$L(I, S) = -\sum_{t=1}^N \log p_t(s_t) \quad (5)$$

## 2.2 串行法图像中文描述生成

串行法首先训练图像到英文描述生成的模型, 再利用训练好的模型, 将随机给定的图片, 预测图片生成的英文描述。然后将预测的英文描述使用翻译模型翻译成中文描述。

图 1 是串行法图像中文描述生成框架, 首先将给定的一张图像经过 CNN 提取图像特征, 图像对应的英文描述句子加上特殊的开始符  $\langle S \rangle$  和特殊的结束符  $\langle /S \rangle$ , 编码成 one-hot<sup>[14]</sup> 向量, 图中用圆球表示。接着将图像特征和编码向量用 LSTM 解码, 图像特征作为 LSTM 的初始值, 只在零时刻使用, 编码向量作为 LSTM 的输入。生成英文描述的概率用 PN,  $N=\{0,1,\dots\}$  表示。最后得到英文描述, 利用翻译模型翻译成中文描述。

## 2.3 构建伪语料法图像中文描述生成

构建伪语料的方法是先将图像对应的英文描述用翻译模型翻译成中文描述, 形成图像和中文描述的伪标注语料, 接着用该伪标注语料训练图像到中文描述生成模型。这种方法可以直接由图像得到它对应的中文描述。考虑到生成中文描述的方式, 本文提出了两种训练生成模型的方法, 分别是基于词的方法和基于字的方法。

### 2.3.1 基于词的中文描述生成

基于词的模型框架, 是将经过中文描述以词的形式进行编码。比如给定一句中文描述[一本书在街道旁的木凳上], 其分词结果为[“一本” “书” “在” “街道” “旁” “的” “木凳” “上”]。

图 2 表示基于词的训练模型框架, 首先是将英文描述使用翻译模型, 得到中文句子, 然后将中文句子分词并且加上特殊的开始符  $\langle S \rangle$  和特殊的结束符  $\langle /S \rangle$ , 将处理后的

表 1 串行法图像中文描述生成实验结果

CNN	CIDEr	BLEU_4	BLEU_3	BLEU_2	BLEU_1	ROUGE_L	METEOR
inceptionV3	0.989	0.298	0.395	0.523	0.692	0.498	0.251
Inception-Resnet-v2	1.038	0.315	0.412	0.539	0.703	0.507	0.256

表 2 图像英文描述生成实验结果(Image-EN)以及正确英文描述情况下的中文描述生成实验结果(Gold EN-ZH)

	CNN	CIDEr	BLEU_4	BLEU_3	BLEU_2	BLEU_1	ROUGE_L	METEOR
Image - EN	InceptionV3	0.864	0.264	0.368	0.511	0.688	0.492	0.240
	Inception-Resnet-v2	0.876	0.274	0.379	0.522	0.698	0.516	0.245
Gold EN -ZH	-	2.105	0.708	0.768	0.837	0.919	0.789	0.4692

句子编码成 one-hot 向量；图像特征使用 CNN 提取。类似地，训练过程中，图像特征作为 LSTM 的初始值，只在零时刻使用，编码向量作为 LSTM 的输入。生成中文描述的概率用 PN,  $N=\{0,1,\dots\}$  表示。最后得到中文描述。

### 2.3.2 基于字的中文描述生成

基于字的方法是将中文描述以字为单位进行编码。比如给定一句中文句子[一本书在街道旁的木凳上]，将整个句子切分成字的形式[“一” “本” “书” “在” “街” “道” “旁” “的” “木” “凳” “上”]。基于字的模型除了中文句子形成 one-hot 向量时是以每个字为单位之外，其余和基于词的模型是一致的。

## 3 实验与分析

### 3.1 数据集

本文使用的数据集为 MSCOCO2014<sup>[15]</sup>，其中训练集包含 82783 张图片，对应的英文句子描述共有 414114 句，平均每张图片对应 5 句不同的英文描述；开发集包含 40504 张图片，共 202655 句英文描述，类似地每张图片对应 5 句不同的英文描述；测试集包含 40775 张图片。特别地，为了评估中文描述生成的性能，我们从测试集随机选取 500 张图片，参照英文描

述标注的方法，从不同角度由人工标注，为每张图片标注 5 句不同的中文描述。因此，本文使用的测试集中仅限于这人工标注中文描述的 500 张图片。

### 3.2 实验设置

#### 3.2.1 视觉特征提取网络设置

视觉特征提取网络 CNN(I)完成  $I \rightarrow V(I)$  的特征映射，其中 I 为输入的图像，输出为视觉特征向量  $V(I)$ 。视觉特征的提取使用 Inception-Resnet-v2<sup>[16]</sup>和 InceptionV3<sup>[17]</sup>两种结构，两种结构均在大规模单标签分类任务 ImageNet<sup>[18]</sup>上进行训练。本文使用两种已经训练好的结构提取视觉特征，用两种结构的网络隐藏层的输出来表示提取到的视觉特征。首先对图像进行缩放、裁剪、调整对比度清晰度得到大小为 299X299 的三通道 RGB 图像，然后使用这两种结构进行图像特征提取。

- CNN 模型使用 Inception-Resnet-v2<sup>[14]</sup>时，隐藏层的输出(32X8X8X1536)作为提取到的视觉特征。将视觉特征使用卷积核大小为(8X8)的平均池化层、平铺之后得到(32X1536)的视觉特征。其中 32 是批处理的个数。在使用输出为 512 的全连接层将得到的视觉特征映射为 (32X512)的特征矩阵作为最终的视觉特征。
- CNN 模型使用 InceptionV3<sup>[15]</sup>时，隐藏

表 3 构建伪语料法图像中文描述生成实验结果

	CNN	CIDEr	BLEU_4	BLEU_3	BLEU_2	BLEU_1	ROUGE_L	METEOR
基于词	InceptionV3	1.036	0.321	0.415	0.536	0.695	0.511	0.255
	Inception-Resnet-v2	1.071	0.328	0.425	0.543	0.697	0.529	0.268
基于字	InceptionV3	1.064	0.332	0.427	0.546	0.701	0.523	0.256
	Inception-Resnet-v2	1.101	0.341	0.434	0.553	0.710	0.522	0.271



0: A man surfing a small wave with a paddle

1: 一名男子用桨冲浪

2: a man riding a paddle board on top of a body of water .

3: 一名男子在水体顶部骑着桨板。

4: 一个人在水中冲浪板上。

5: 一名男子在水中冲浪板上。



0: A giraffe bending over while standing on green grass.

1: 一头长颈鹿低着头站在绿色的草地上。

2: a giraffe is standing in a grassy field .

3: 长颈鹿站在草地上。

4: 长颈鹿站在草地上。

5: 一头长颈鹿站在草地上。

图 3 inception-Resnet-v2 模型下预测的结果。0 和 1 分别表示人工标注的英文描述和中文描述。2 表示串行法中得到的自动英文描述，3 表示串行法中自动英文描述经过翻译模型得到的中文描述。4 和 5 是在构建伪语料方法下基于词和基于字的方法分别得到的结果。

层的输出(32X2048)作为提取到的视觉特征，在使用输出为 512 的全连接层将得到的视觉特征映射为(32X512)的特征矩阵作为最终的视觉特征。

### 3.2.2 两种图像描述生成方法设置

在串行法图像中文描述生成方法中，需要建立图像到英文描述的生成模型。为此，本文对图像英文描述句子进行断词处理，对每个句子加入开始符“<S>”和结束符“</S>”，词汇大小  $n$  设置为 12000，未登录词用<UNK>表示。词向量长度  $m$  为 512，解码器 LSTM 的隐蔽状态长度  $d$  也设置为 512；词向量和模型参数的初始值在区间 [-0.08,0.08]按均匀分布中得到，实验采用初始学习率为 2.0 的随机梯度下降算法，学习衰减率为 0.5。在测试的时候，本文使用大小为 4 的柱状搜索算法<sup>[19]</sup>。处理大小为 32。在训练过程中我们使用 Dropout 正则化和归一化处理来提高模型的泛化能力<sup>[20]</sup>。在构建伪语料法图像中文描述生成方法中，需要建立图像到中文描述的生成模型。在基于词的中文描述生成时，使用结巴分词工具(<https://pypi.python.org/pypi/jieba>)获取句子

的分词结果。在建立图像到基于词和基于字的中文描述生成方法中，目标端词汇的大小都设置为 16000，未登录词用<UNK>表示。其他参数设置与串行法建立的图像到英文描述的生成模型一致。

### 3.2.3 评测指标

本文的评测指标使用 BLEU-1, 2, 3, 4<sup>[21]</sup>、METEOR<sup>[22]</sup>、RougeL<sup>[23]</sup>和 CIDEr<sup>[24]</sup>六种指标衡量图像描述生成结果的质量。其中 BLEU 一般是用在机器翻译评测翻译质量的，反映了生成结果与参考答案之间的  $N$  元文法准确率。METEOR 测量基于单精度加权调和平均数和单字召回率。RougeL 与 BLEU 类似，它是基于召回率的相似度衡量方法。CIDEr 是基于共识的评价方法，优于上述其他指标。此外，为了减少分词错误对评测的影响，本文在评测图像到中文描述的性能，中文端都是以字为单位进行评测。

## 3.3 结果与分析

以下我们分别报告并比较本文提出的两种方法生成的中文描述的质量。



表 4 构建伪语料法中，基于词和基于字的中文描述概率最高的四种预测结果及其概率，其中第一行表示人工标注的中文描述，称为金标（Gold），第二行是基于词的结果及其概率，第三行是基于字的结果及其概率。



	Gold: 一个婴儿手里拿着一把牙刷。	
	0) 一个 婴儿 拿 着 牙刷 和 牙膏 。 (p=0.000524) 1) 一个 婴儿 拿 着 牙刷 在 他 的 嘴里 。 (p=0.000213) 2) 一个 婴儿 拿 着 牙刷 和 牙膏 (p=0.000195) 3) 一个 婴儿 拿 着 牙刷 在 他 的 嘴里 (p=0.000100)	
	0) 一个 年 轻 的 孩 子 拿 着 一 把 牙 刷 。 (p=0.000111) 1) 一个 年 轻 的 男 孩 拿 着 一 把 牙 刷 。 (p=0.000084) 2) 一个 年 轻 的 孩 子 拿 着 一 把 牙 刷 (p=0.000030) 3) 一个 年 轻 的 男 孩 拿 着 一 把 牙 刷 ， 用 牙 刷 刷 牙 。 (p=0.000001)	

表 5 基于词和基于字的情况下，模型预测的中文描述及其描述的概率，以及中文描述中每个字或词预测下个字或词的概率。第一行表示人工标注的中文描述，称为金标（Gold），第一列表示要预测描述的图片，第二列表示基于词的情况下模型预测的结果和概率，第三列表示基于字的情况下，模型预测的结果和概率。

	Gold: 背着背包的一群人站在滑雪板上准备滑雪	
	一群 人 在 雪 地 里 滑 雪 。 (p=0.008602)	一 群 人 在 雪 地 上 滑 雪 。 (p=0.007592)
	一群 (0.3749)，人 (0.7207)，在 (0.3917)，雪地 (0.7285)，里 (0.5089)，滑雪 (0.7419)，。(0.9996)	一 (0.7063)，群 (0.4664)， 人，(0.5900)，在 (0.4661)，雪 (0.8427)，地 (0.5554)，上 (0.4998) 滑 (0.9916)，雪 (0.7861)，。(0.9998)

### 3.3.1 串行法图像中文描述生成实验结果

表1给出了串行法图像中文描述生成的实验结果。从中可以看出，Inception-Resnet-v2 视觉特征提取算法取得的性能要优于 InceptionV3,这与大规模单标签分类任务的观察结果一致<sup>[16]</sup>。

本文的串行法图像中文描述生成共包含两个环节，首先是图像的英文描述生成，以及英文描述的中文翻译。为了更清晰地理解这两个环节的性能，表2分别给出了英文描述生成的实验结果，以及假定英文描述没有错误情况下，中文描述生成的结果，即翻译模型在测试集上的性能。如表2所示，Inception-Resnet-v2 视觉特征抽取方法在第一个环节得到的英文描述性能 BLEU\_4 值为 0.274，表明自动生成的英文描述与人工

标注存在着一定的差距。另一方面，基于正确的英文描述标注，Gold EN-ZH 的 BLEU\_4 性能达到 0.708，这说明机器翻译在短句翻译上能够取得较好的性能，这也与机器翻译领域的研究一致<sup>[25]</sup>。综合表1和表2，可以发现，当英文描述由正确描述变为自动描述时，中文描述的性能 BLEU\_4 值由 0.708 急剧下降为 0.315，这也说明图像描述生成任务本身较中英机器翻译任务更具挑战性。

### 3.3.2 构建伪语料法图像中文描述生成实验结果

以下分析，视觉特征抽取方法均使用 Inception-Resnet-v2。

表3给出了构建伪语料法图像中文描述生成的实验结果，分为基于词和基于字的中

文描述生成。从表中可以看出，一方面，基于字的性能要优于基于词的性能。例如，从基于词到基于字 BLEU\_4 从 0.328 提高到 0.341，METEOR 从 0.268 提高到 0.271。另一方面，无论是基于词还是基于字，Inception-Resnet-v2 视觉特征提取算法的性能均优于 InceptionV3，这和串行法得到的结论一致。

从表 1 和表 3 可以看出，构建伪语料的方法性能明显优于串行法。例如，从串行法到构建伪语料法，BLEU\_4 从 0.315 提高到 0.341，METEOR 的值从 0.256 提高到 0.271。

### 3.3.3 实例分析

图3给出了两张图片生成的图像描述结果。从图中可以看出，两种方法得到的结果和人工标注的结果相比，都存在一定的预测错误。例如，左边的图片，在串行法中，由于自动生成的英文描述存在着不常用的表达句式“on top of a body of water”，导致翻译得到的中文描述存在生硬及不自然的表达“在水体顶部”。相对而言，构建伪语料法得到的中文描述在表达上更加自然，特别地，构建伪语料法中基于字的模型得到的中文描述最好。

表 4 给出在构建伪语料的方法中，基于词和基于字的中文描述概率最高的四种预测结果及其概率，其中 CNN 模型采用 Inception-Resnet-v2 特征提取方法。基于词的预测结果为“一个 婴儿 拿着 牙刷 和 牙膏。”，错误地在识别了图像中并没有的物体牙膏”，同时遗漏“牙刷”的数量修饰词“一把”；另一方面，基于字的预测结果为“一个 年 轻 的 孩 子 拿 着 一 把 牙 刷。”，正确地预测了“牙刷”的数量修饰词“一把”，但将“婴儿”不是很准确地预测为“一个 年 轻 的 孩 子”。

表 5 给出基于词和基于字的情况下，随机给定的一张图片，模型预测的图像的中文描述及其概率，以及每个字或词预测下一个字或词的概率，即模型经过 softmax 得到一个词或字的概率后。在本例子中，基于词和基于字的预测结果差异性很小，仅在介词选择上不同。基于词预测的中文描述为“雪地

里”，而基于字预测的中文描述使用更加准确的介词“雪 地 上”。另外，基于词预测的中文描述概率( $p=0.008602$ )高于基于字预测的中文描述概率( $p=0.007592$ )，这说明虽然基于字的预测结果长度一般要长于基于词的预测结果，但预测结果概率相差不多。

## 4 结论与未来工作

本文提出了以英文为枢轴语言的图像中文描述生成的两种方法：串行法和构建伪语料法。实验结果表明构建伪语料法得到的系统性能高于串行法。另外本文在构建伪语料法的前提下，比较了基于词和基于字的中文描述生成方法，实验结果表明使用基于字的方法优于基于词的方法。

不论本文的串行法，还是构建伪语料法，图像描述生成都仅为一种语言（英文或中文）。在未来工作中，我们将考虑同时输出两门语言的描述，通过共享图像的表示向量，目标端的两门语言之间存在语义相近的特点，能够帮助同时取得两门语言上更好的描述。

## 参考文献

- [1] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, et al. Every Picture Tells a Story: Generating Sentences from Images [C]// Proceedings Part IV of the 11th European Conference on Computer Vision. Heraklion, Crete, Greece: Springer, 2010:15-29.
- [2] JunHua Mao, Wei Xu, Yi Yang, et al. Deep captioning with multimodal recurrent neural networks (m-rnn)[J]. arXiv preprint arXiv:1412.6632, 2014.
- [3] Oriol Vinyals, Alexander Toshev, Samy Bengio, et al. Show and tell: A neural image caption generator [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015: 3156-3164.
- [4] Xu Jia, Efstratios Gavves, Basura Fernando, et al. Guiding the long-short term memory for image



- caption generation [C]// Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015:, 2015: 2407-2415.
- [5] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention [C]// Proceedings of the 32nd International Conference on Machine Learning. Lille, France: JMLR.org, 2015:2048-20 57.
- [6] Xirong Li, Weiyu Lan, Jianfeng Dong, et al. Adding Chinese Captions to Images [C]// Proceedings of the 2016 Association for Computing Machinery(ACM) on International Conference on Multimedia Retrieval. New York, USA: ACM, 2016: 271-275.
- [7] Christian Szegedy, Wei Liu, Yang Qing Jia. Going deeper with convolutions [C]// Proceedings of the 32nd International Conference on Machine Learning. Lille, France: JMLR. org, 2015: 1-9.
- [8] Orhan Firat, Kyunghyun Cho, Yoshua Bengio. 2016a. Multi-way, multilingual neural machine translation with a shared attention mechanism. In HLT-NAACL.
- [9] Yong Cheng, Yang Liu, Qian Yang et al. 2016a. Neural machine translation with pivot languages. CoRR abs/ 1611.04928.
- [10] Hideki Nakayama and Noriki Nishida. 2016. Zero resource machine translation by multimodal encoder-decoder network with multimedia pivot. CoRR abs/1611.04503
- [11] Dzmitry Bahdanau, KyungHyun Cho, Bengio Yoshua. Neural machine translation by jointly learning to align and translate [C]// Proceedings of the International Conference on Learning Representations, 2015.
- [12] Jonas Gehring, Michael Auli, David Grangier, et al. Convolutional Sequence to Sequence Learning [C]// International Conference on Machine Learning. 2017: 1243-1252.
- [13] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need [C]// Advances in Neural Information Processing Systems. 2017: 6000-6010.
- [14] Subhasish Mitra, LaNae J Avra , Edward J McCluskey. Scan synthesis for one-hot signals [C]// Test Conference, 1997. Proceedings International. IEEE, 1997: 714-722.
- [15] Tsung-Yi Lin, Michael Maire, Serge Belongie, et al. Microsoft coco: Common objects in context [C]// European conference on computer vision. Springer, Cham, 2014: 740-755.
- [16] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, et al. Inception-v4, inception-resnet and the impact of residual connections on learning [C]// American Association for Applied Linguistics (AAAI). 2017, 4: 12.
- [17] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, et al. Rethinking the Inception Architecture for Computer Vision[C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016: 2818-2826.
- [18] Olga Russakovsky, Jia Deng, Hao Su, et al. ImageNet Large Scale Visual Recognition Challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.
- [19] Sam Wiseman, AlexanderM. Rush. Sequence-to-sequence learning as beam-search optimization[J]. arXiv preprint arXiv: 1606. 02960, 2016.
- [20] Sergey Ioffe, Christian Szegedy. Batch Normalization. Accelerating Deep Network Training by Reducing Internal Covariate Shift [C]// Proceedings of the 32nd International Conference on Machine Learning. Lille, France: JMLR.org, 2015: 448-456
- [21] Kishore Papineni, Salim Roukos, Todd Ward, et al. Bleu: a Method for Automatic Evaluation of Machine Translation [C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, PA, USA: ACL, 2002: 311-318.
- [22] Michael Denkowski, Alon Lavi. Meteor universal: Language specific translation evaluation for any target language[C]// Proceedings of the ninth workshop on statistical machine translation. 2014: 376-380.
- [23] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries [C]// Proceedings of the 42nd Annual Meeting of the

Association for Computational Linguistics.  
Barcelona, Spain: ACL, 2004:10-18.

- [24] Ramakrishna Vedantam, C. Lawrence Zitnick, Devi Parikh. CIDEr: Consensus-based image description evaluation [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 4566-4575.
- [25] Junhui Li, Deyi Xiong, Zhaopeng Tu, et al. Modeling source syntax for neural machine translation [C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 688-697.



张凯（1993—），硕士研究生，主要研究领域为自然语言处理，机器翻译。

E-mail: 1421390187@qq.com



李军辉（1983—），副教授，主要研究领域为自然语言处理，机器翻译。

E-mail: lijunhui26@gmail.com



周国栋（1967—），教授，主要研究领域为自然语言处理。

E-mail: gdzhou@suda.edu.cn