计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190900181

双语图像标题联合生成研究

张 凯 李军辉 周国栋
苏州大学计算机科学与技术学院 江苏 苏州 215006
(suda_zk@163.com)

摘 要 图像标题(Image Caption)的研究大多是对图像生成单一语言的标题,而在当今各国语言交汇融合的情况下,对一张图像生成两门甚至多门语言标题是必然趋势,以让不同母语的人理解其他人对同一张图片的评价。对此,提出一种双语图像标题,即图像同时生成两种语言标题的方法。该方法由一个编码器和两个不同的解码器组成,其中编码器基于卷积神经网络,用于提取图像特征;解码器基于长短时记忆网络,两个不同的解码器分别用于解码两种不同的语言特征。由于两种语言标题之间存在着互译的特性,因此提出了双语料图像标题的联合生成模型。具体地,在解码端采用交替的方式生成两种语言的标题,使得在预测某种语言的下一个单词时,不仅可以利用该语言标题的历史信息,还可以利用另一门语言标题的历史信息,同时促进两种语言标题生成的性能。基于MSCOCO2014数据集的实验结果表明,双语图像标题联合生成能够同时提高两门语言的性能,在英文上较英文单语言标题生成的性能提高了1.0个BLEU_4值和0.98个CIDEr值,在日文上较日文单语言标题生成的性能提高了1.0个BLEU_4值和0.31个CIDEr值。

关键词:图像双语标题;联合模型;交替生成
中图法分类号 TP391.1;TP391.41

Study on Joint Generation of Bilingual Image Captions

ZHANG Kai,LI Jun-hui and ZHOU Guo-dong
School of Computer Science and Technology,Soochow University,Suzhou,Jiangsu 215006,China

Abstract Most of the research on image caption is to generate a single language caption from an image,but in the context of the convergence of languages in various countries,it is necessary to generate two languages or even multiple languages from one image. Native speakers understand what other people say about the image,so this paper proposes an approach to generation of bilingual image captions,i.e.,generating two captions in two languages for an image. The architecture consists of an encoder and two decoders,in which the encoder uses convolutional neural network to extract image features while the decoders adopt Long Short-Term Memory networks. Motivated by the fact that the two captions of an image are semantically equivalent,we propose a joint model to generate bilingual image captions. Specifically,the two decoders generate image captions in alternative way,making the decoding history information of two languages are both available to predict the next word. The experimental results based on the MSCOCO2014 data set show that the joint generation of bilingual image caption can improve the performance of two languages at the same time. Compared with the single language image caption performance in English,the BLEU_4 increases by 1.0,CIDEr increases by 0.98 in Japanese. Compared with the Japanese single image caption generation performance,the BLEU_4 increases by 1.0,CIDEr increases by 0.31.

Keywords Bilingual image captions,Joint model,Alternative generation

1 引言

自然语言处理(Natural Language Processing,NLP)和计算机视觉(Computer Vision,CV)是人工智能领域的两大研究热点。当前,跨领域研究已经成为未来研究的一种趋势,引起了研究者的极大兴趣。图像标题(Image Caption)正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Ali等[1]提出,即给定二元组(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,在测试阶段,随机给定一张

到稿日期:2019-09-26 返修日期:2020-03-07 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61876120)
This work was supported by the National Natural Science Foundation of China(61876120).
通信作者:李军辉(jhli@suda.edu.cn)

图片，由模型自动描述图片的内容。“看图说话”对正常人来说非常简单，但对于计算机却是一项极大的挑战，计算机不仅要识别图片的内容，还要用人类的逻辑思维描述出人类可读的句子。

当前图像标题生成的主流方法是基于神经网络的方法。在此基础上，大部分研究都偏向于对图像生成单一语言（如英文）标题的研究，并且取得了很好的研究成果。但是在很多时候，需要用不同的语言来描述同一张图片，以便于不同母语的人理解其他人对同一张图片的评价。因此，对图像生成双语甚至多语言标题是一项必不可少的研究任务。本文提出了一种联合两种语言特征的方法，这种方法能同时提高图像双语标题的性能，其中用来联合的特征包括两种语言的词嵌入和标题生成模型的隐藏状态。

2 相关工作

图像标题（Image Caption）任务的传统做法是利用图像处理的一些算法提取图像特征，再经过支持向量机（Support Vector Machine，SVM）分类等得到图像中可能存在的目标对象[2]；然后根据提取到的对象以及它们的属性，利用条件随机场（Conditional Random Field，CRF）或者一些指定的规则来将图像特征恢复成对图像的描述。不难看出，这种做法非常依赖于图像特征的提取以及生成句子时所需要的规则，效果也并不理想。

受神经机器翻译的启发，将机器翻译中编码源文字的循环神经网络（Recursive Neural Network，RNN）换成卷积神经网络（Convolutional Neural Network，CNN）来编码图像，图像标题生成便转化为机器翻译问题。从翻译的角度来看，此处的源文字就是图像，目标文字就是生成的标题。因此，图像标题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 CNN 将图片转化为一个固定长度的向量，也称作图像的隐层表示；解码器使用 RNN 将编码器输出的固定长度的向量解析为目标语言句子。Vinyals 等[3]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（Long Short Term Memory Network，LSTM）生成英文描述。Karpathy 等[4]提出利用片段图像生成局部区域的描述。Mao 等[5]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。Xu 等[6]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标题，解决了图像仅在开始时传入 LSTM 的问题。在此基础上，Wu 等[7]提出了 att-LSTM 模型，该模型通过图像多标签分类来提取图像中可能存在的属性，这种方法解决了图像特征总是使用全局特征的问题。Xu 等[8]将注意力机制引入解码过程，使得标题生成网络能够捕捉图像的局部信息。然而，这种加入注意力的方法也存在一些缺点，即每个词都会对应一个图像区域，但是有些介词和动词并不能对应实体；除此之外，注意力机制是基于卷积层的加权，对应到图像中略显模糊，而且不能准确定位图中的对应区域。为了解决这些问题，Lu 等[9]提出了一种自适应性的注意力机制，使模型可以自己决定是根据先验知识（模板）还是图像中的区域生成单词。前人所做的工作都基于解码器 RNN，然而 CNN 也是不可忽略的一个重点。Chen 等[10]提出利用卷积层的不同通道做注意力，同时还利用了空间注意力机制。Li 等[11]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[12]对图像进行编码，并使用 LSTM 对图像描述生成过程建模。Rennie 等[13]提出 SCST 模型，利用强化学习的方法生成有区别度的标题。Anderson 等[14]提出了 Bottom-Up and Top-Down 模型，该模型结合 Bottom-Up 和 Top-Down 视觉注意力机制，Bottom-Up 机制用来提取视觉特征，Top-Down 机制用于关注词向量特征。Dognin 等[15]指出图像标题任务的评价指标不够全面；且为了“跨模型评估”，受 inception score 指标的启发，使用了 semantic score，也就是利用有监督的标签构建了一个分类器。Biten 等[16]提出了基于新闻场景的图像描述任务，他们指出当前的 caption 任务都是生成一个描述性的句子，但是人类都是带着先验知识来理解和描述图片的，因此将新闻作为先验条件可以生成包含实际地名的句子。Kim 等[17]提出基于关系的图像生成标题方法，首先检测一个关系，然后生成句子。

以上工作均是对图像生成单一语言（如英文）标题的研究。本文主要研究对图像同时生成两门语言的标题，具体地，在预测某种语言的下一个单词时，不仅可以利用该语言标题的历史信息，还可以利用另一门语言标题的历史信息，采用这种交叉训练的方式，可以同时提高两门语言标题生成的性能。

3 单语图像标题生成模型

本文的单语图像标题生成模型采用 Anderson 等[14]提出的 Top-Down 框架。如图 1 所示，该模型由两层 LSTM 和一个视觉注意力单元组成。第一个 LSTM 层称为 Top-Down Attention LSTM，第二层 LSTM 称为 Language LSTM。视觉注意力单元称为 Attend。为方便起见，本文将使用下面的公式表示 t 时刻 LSTM 的操作：

$$\boldsymbol{h}_t = LSTM(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}) \tag{1}$$

其中，$\boldsymbol{x}_t$ 是 $LSTM$ 的输入向量，$\boldsymbol{h}_{t-1}$ 是前一时刻 LSTM 的输出向量，$\boldsymbol{h}_t$ 是 LSTM 的输出向量。

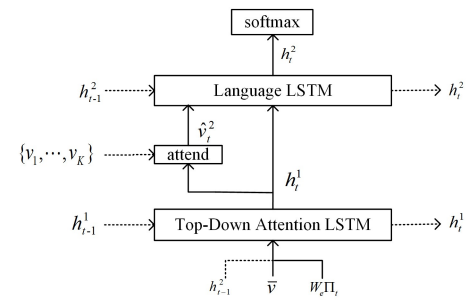

图 1 单语图像标题生成模型

Fig. 1 Monolingual image caption generation model

3.1　基于自上而下的注意力 LSTM

在单语图像标题生成模型中，本文将第一层 LSTM 称为基于自上而下的注意力 LSTM（图 1 中的 Top-Down Attention LSTM），在后面的公式中用上标数字表示层数。在 t 时刻，第一层 LSTM 的输入包括第二层 LSTM 前一时刻的隐藏状态输出 h_{t-1}^2、经过平均池化后的图像特征 $\bar{v}=\frac{1}{k}\sum_i v_i$，以及 t 时刻的单词嵌入 $\boldsymbol{W}_e\boldsymbol{\Pi}_t$，表示如下：

$$x_t^1=[h_{t-1}^2,\bar{v},\boldsymbol{W}_e\boldsymbol{\Pi}_t] \tag{2}$$

其中，$\boldsymbol{W}_e\in R^{E\times|V|}$ 是词嵌入矩阵，$\boldsymbol{E}$ 表示词向量大小，$|V|$ 是字典的大小；$\boldsymbol{\Pi}_t$ 为 t 时刻输入单词编码生成的 one-hot[18] 向量，其维度等于字典的大小。

3.2　图像注意力机制

在 t 时刻，自上而下的注意力 LSTM 得到的输出 h_t^1，为 K 个图片特征 $\{v_1,\cdots,v_K\}$ 的每一个特征 $v_i\in R^D$ 生成标准化的注意力权重 α_{it}（图 1 中的 attend），公式如下：

$$a_{it}=\omega_a^{\mathrm{T}}\tanh(W_{va}v_i+W_{ha}h_t^1) \tag{3}$$

$$\alpha_t=softmax(a_t) \tag{4}$$

其中，$W_{va}\in R^{H\times D}$，$W_{ha}\in R^{H\times H}$，$\omega_a\in R^H$ 是模型参数，D 表示图像特征向量大小，H 是隐藏状态维度大小。我们的目标是得到所有图像特征的权重和 $\hat{v}t$，即：

$$\hat{v}_t=\sum_{i=1}^{K}\alpha_{it}v_i \tag{5}$$

3.3　基于语言模型的 LSTM

单语图像标题生成模型的第二层 LSTM 被称为基于语言模型的 LSTM（图 1 中的 Language LSTM）。Language LSTM 的输入 x_t^2 为式(5)得到的 $\hat{v}_t$ 和第一层 LSTM 的输出 h_t^1 的拼接，即 $x_t^2=[\hat{v}_t,h_t^1]$。本文使用 $y_{1:T}$ 表示一个单词序列 $(y_1,\cdots,y_T)$，在 t 时刻，使用式(6)表示 Language LSTM 可能输出单词的条件概率：

$$p(y_t|y_{1:t-1})=softmax(W_ph_t^2+b_p) \tag{6}$$

其中，$W_p\in R^{|V|\times H}$，$b_p\in R^{|V|}$ 分别表示要学习的权重和偏执项。

给定输入图片 I，整个输出序列上的条件分布为所有单词的条件概率的乘积：

$$p(y_{1:T})=\prod_{t=1}^{T}p(y_t|y_{1:t-1},I) \tag{7}$$

图像标题生成模型的训练目标是最小化式(8)中的交叉熵损失：

$$L(\theta)=-\sum_{t=1}^{T}\log(p_\theta(y_t^*|y_{1:t-1}^*,I)) \tag{8}$$

其中，$y_{1:T}^*$ 表示真实标签，θ 表示单语图像标题生成模型要训练的参数。

4　双语图像标题联合生成模型

为了得到图像的双语标题，一种可行的做法是分别为每种语言生成标题时采用单语标题生成模型。但该做法往往忽略了两种语言标题之间存在着互译的特性。本节首先提出基于自注意力机制的单语图像标题生成模型；然后在此基础之上提出双语图像标题联合生成模型。

图 2 给出本文提出的双语图像标题联合生成模型的框架。两门语言共享 Encoder 端，即图中的 CNN 提取图像特征，图中虚线部分表示交叉训练的联合生成方式，这部分会在 4.2 小节详细说明，Top-Down 表示图 1 的单语标题生成模型。

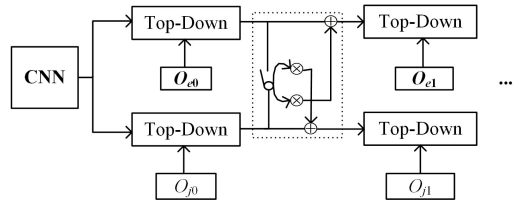

图 2　双语图像标题联合生成模型

Fig. 2　Models on joint generation of bilingual image captions

4.1　带自注意力机制的单语图像标题生成模型

从式(2)可以看出，在当前时刻，生成单词时，只使用了前一时刻的单词输入，虽然这种方法在预测当前单词的时候能够使用到前面时刻的单词信息，但是随着预测序列长度的增加，更早之前预测的单词信息在当前时刻的作用就会减小。Werlen 等[19] 认为不同的单词根据它们与下一个单词的预测关系有着不同的重要性，因此采用一种自注意力机制的方式来预测当前单词时，获取当前单词对前面所有单词的依赖程度。这种机制的目的是在一定程度上对单词之间重要的非顺序依赖关系进行建模，并将其作为循环层的补充内容。借鉴该思想，我们用 $y_t=W_e\Pi_t$ 表示第一层 LSTM 在 t 时刻的单词输入，那么 $x_t^1=[h_{t-1}^2,\bar{v},y_t]$ 就可以表示为 $x_t^1=[h_{t-1}^2,\bar{v},d_t]$。$d_t$ 的表示如下：

$$d_t=\sum_{i=1}^{t-1}\alpha_i^t y_i \tag{9}$$

$$\alpha_i^t=softmax(e_i^t) \tag{10}$$

$$e_i^t=v^{\mathrm{T}}\tanh(\boldsymbol{W}_y y_i) \tag{11}$$

其中，$v\in R^H$，$\boldsymbol{W}_y\in R^{H\times H}$，$\boldsymbol{W}_s\in R^{H\times H}$ 表示权重矩阵，是模型要学习的参数。

4.2　双语图像标题联合生成模型

本文提出了两种双语图像标题联合生成模型的方法，第一种方法是基于词嵌入的联合模型，另一种方法是基于隐藏状态的联合模型。两种方法均按交替的方式生成双语图像标题。为方便起见，根据第 5 节的实验设置，以下分别以英文和日文为例说明双语图像标题的联合生成模型，同时以 en 和 ja 为右下标区别英文生成模型和日文生成模型的符号表示。例如，在 t 时刻，先生成英文单词 $y_{t_{en}}$，接着生成日文单词 $y_{t_{ja}}$。这样，生成英文单词的历史标题信息包含 $(y_{1_{en}}\cdots y_{t-1_{en}})$ 和 $(y_{1_{ja}}\cdots y_{t-1_{ja}})$，历史状态信息包括 $(h_{1_{en}}^2\cdots h_{t-1_{en}}^2)$ 和 $(h_{1_{ja}}^2\cdots h_{t-1_{ja}}^2)$，生成日文单词的历史标题信息包含 $(y_{1_{en}}\cdots y_{t_{en}})$ 和 $(y_{1_{ja}}\cdots y_{t-1_{ja}})$，历史状态信息包括 $(h_{1_{en}}^2\cdots h_{t_{en}}^2)$ 和 $(h_{1_{ja}}^2\cdots h_{t-1_{ja}}^2)$。

4.2.1　基于词嵌入的联合模型

t 时刻基于词嵌入的联合模型单元如图 3 所示。英文和日文的词嵌入经过式(9)分别得到 $d_{t_{en}}$ 和 $d_{t_{ja}}$，通过门控机制 $sigmoid(\cdot)$ 来决定在生成当前语言句子时，使用到另一种语言已经得到的部分单词信息的多少。我们使用下面的公式来

表示词嵌入联合模型单元：

$$g=sigmoid(W_{en}d_{t_{en}}+W_{ja}d_{t_{ja}}) \tag{12}$$

$$\hat{d}_{t_{en}}=d'_{t_{en}}+g\circ d_{t_{ja}} \tag{13}$$

$$\hat{d}_{t_{ja}}=d'_{t_{ja}}+g\circ d_{t_{en}} \tag{14}$$

其中，$W_{en}\in R^H$，$W_{ja}\in R^H$ 是模型参数，$\circ$表示矩阵对应位置的元素相乘，$d'_{t_{en}}$ 和 $d'_{t_{ja}}$ 分别表示 t 时刻双语标题联合模型的英文和日文的实际单词输入。那么，联合训练模型的英文输入就可以表示为 $x^1_{t_{en}}=[h^2_{t-1_{en}},v\_,\hat{d}_{t_{en}}]$，日文输入表示为 $x^1_{t_{ja}}=[h^2_{t-1_{ja}},\bar{v},\hat{d}_{t_{ja}}]$。通过这种门控制方式，可以在联合模型生成英文句子时使用部分已经得到的日文单词信息，生成日文句子时使用部分已经得到的英文单词信息。

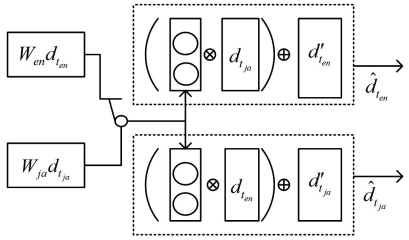

图 3 词嵌入联合单元

Fig. 3 Word embedding joint unit

4.2.2 基于隐藏状态的联合模型

t 时刻基于隐藏状态的联合单元如图 4 所示。基于隐藏状态的联合模型和基于词的联合模型方法类似，在交替生成英文和日文句子时，不使用词嵌入联合，而是联合双语标题生成模型第二层 LSTM 的隐藏状态的输出 $h^2_{t_{en}}$ 和 $h^2_{t_{ja}}$：

$$g=sigmoid(W^h_{en}h^2_{t_{en}}+W^h_{ja}h^2_{t_{ja}}) \tag{15}$$

$$h'^2_{t_{en}}=h^2_{t_{en}}+g\circ h^2_{t_{ja}} \tag{16}$$

$$h'^2_{t_{ja}}=h^2_{t_{ja}}+g\circ h^2_{t_{en}} \tag{17}$$

其中，$W^h_{en}\in R^d$ 和 $W^h_{ja}\in R^d$ 是模型参数，$h'^2_{t_{en}}$ 和 $h'^2_{t_{ja}}$ 分别表示双语标题生成模型第二层 LSTM 的隐藏状态经过联合后的英文和日文的最终隐藏状态的输出。这样，式(6)就可以写成：

$$p_{en}(y_{t_{en}}|y_{1:t-1_{en}})=softmax(W_{p_{en}}h'^2_{t_{en}}+b_{p_{en}})$$

$$p_{ja}(y_{t_{ja}}|y_{1:t-1_{ja}})=softmax(W_{p_{ja}}h'^2_{t_{ja}}+b_{p_{ja}})$$

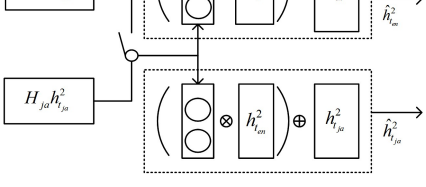

图 4 隐藏状态联合单元

Fig. 4 Hidden state joint unit

生成英文和日文标题的训练目标分别是最小化下面公式中的交叉熵损失：

$$L_{en}(\theta_{en})=-\sum_{t=1}^{T}\log(p_{\theta_{en}}(y^*_{en}|y^*_{1:t-1_{en}},I))$$

$$L_{ja}(\theta_{ja})=-\sum_{t=1}^{T}\log(p_{\theta_{ja}}(y^*_{t_{ja}}|y^*_{1:t-1_{ja}},I))$$

其中，$y^*_{1:T_{en}}$ 和 $y^*_{1:T_{ja}}$ 分别表示英文和日文的真实标签，θ_{en} 和 θ_{ja} 表示双语标题生成模型的参数。因此双语图像标题联合生成模型的最终目标损失为 $L_{loss}(\theta)=\beta L_{en}(\theta)+(1-\beta)L_{ja}(\theta)$。其中，$\beta$ 表示超参数，在实验中设置为 $\beta=0.5$。

5 实验与分析

5.1 数据集

本文使用的数据集为 MSCOCO2014[20]，日文数据集使用 Yoshikawa 等[21]的 STAIR Captions 数据集。我们使用 Karpathy 等[4]的方法将 MSCOCO 数据集分成训练集、验证集和测试集，其中训练集共有 113 287 张图片，验证集和测试集各有 5 000 张图片，每张图片都提供 5 句不同的英文标题以及对应的日文标题。

本文的评测使用 BLEU-1，2，3，4[22]，METEOR[23]，ROUGE_L[24]，CIDEr[25]和 SPICE[26] 5 种指标来衡量图像生成标题的质量。其中，BLEU 指标一般是用于机器翻译领域评测翻译质量，反映了生成结果与参考答案之间的 N 元文法的准确率；METEOR 用于测量基于单精度加权调和平均数和单字召回率；ROUGE_L 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，这个指标是将每个句子都看成“文档”，将其表示为向量的形式，然后计算参考标题与模型生成标题的余弦相似度作为打分；SPICE 是一种语义命题图像标题评估方法，用于评测某些句子虽然在 n-gram 上的重叠度很低，但是表达的意思相近的情况，尽可能多地考虑到每句话的语义命题。该评测指标通过将候选标题和参考标题转换为一种被称为场景图的基于图的语义表示来评估标题质量。场景图显式地对图像标题中的对象、属性和关系进行编码，并在此过程中抽象出自然语言的大部分词汇和句法特性。

5.2 基准模型

本文使用的单语图像标题生成的基准模型与 Up-Down：ResNet-101[14]方法相同，即编码器使用 ResNet-101 结构提取图像特征，解码器使用第 2 节介绍的 Top-Down 模型。

双语图像标题生成模型使用的是一个共享的编码器和两个不同的解码器。其中，共享的编码器用来提取视觉特征，使用 ResNet-101[27]结构；两个不同的解码器分别用来解码英文句子和日文句子，解码器使用 Anderson 等[14]提出的 Top-Down 模型。

5.3 实验设置

5.3.1 视觉特征提取网络的设置

视觉特征提取网络 CNN(I)完成 $I\rightarrow V(I)$ 的特征映射，其中 I 为输入图像，输出 $V(I)$ 为视觉特征，一般而言，视觉特征提取方法使用 ResNet-101 结构，这种结构在大规模单标签分类任务 ImageNet[28]上进行训练。本文使用已经训练好的结构提取视觉特征，用这种结构的最后一个卷积层的输出作为提取到的视觉特征。实验中使用 ResNet-101 网络结构层的最后一个卷积层的输出经过平均池化层，将图像特征映射为(50，14×14，2 048)的特征矩阵；再经过隐藏单元数为 512

的全连接层,将视觉特征映射为(50,14×14,512)的特征矩阵后作为最终的视觉特征。在整个模型训练的过程中,ResNet-101 的模型参数不更新。

5.3.2　双语图像标题生成模型的设置

双语图像标题生成模型使用 Top-Down 机制,两层 LSTM 的隐藏状态长度 H 均为 512,词向量的长度 E 为 512。英文词汇表大小为 9487,日文词汇表大小为 11874,未登录词用〈UNK〉表示。词向量和模型参数的初始值在区间[-0.1,0.1]按均匀分布得到,使用自适应矩估计(Adam)算法[29]对损失函数进行优化,学习率为 5×10^{-4}。训练时批处理大小为 50,测试时批处理大小为 10,并且在测试时使用大小为 4 的柱状搜索算法[30]。在训练过程中,每一层使用 Dropout 正则化和归一化处理来提高模型泛化能力[31]。

5.4　结果与分析

本节对本文提出的基于词嵌入和基于隐藏状态的联合模型方法生成英文和日文标题的质量进行对比。

5.4.1　英文标题的实验结果

表 1 比较了本文提出的同时使用基于词和基于隐藏状态的联合模型与已有工作的实验结果,本文的基准方法和表中 Up-Down:ResNet-101 的方法相同。从表中实验结果可以得知,基准方法的性能与 Up-Down:ResNet-101 相当,虽然 CIDEr 值低于 Up-Down:ResNet-101,但是 BLEU_4 值更高。由数据得出,本文提出的 Both_joint 方法较基准模型的性能提高了 1.0 个 BLEU_4 值和 0.98 个 CIDEr 值。

表 1　在 MSCOCO Karpathy 测试集上本文提出的英文图像标题方法与已有方法的比较

Table 1　Comparisons of our English image captioning approach and existing methods on MSCOCO Karpathy test split

Method	*CIDEr*	*BLEU_4*	*BLEU_3*	*BLEU_2*	*BLEU_1*	*ROUGE_L*	*METEOR*	*SPICE*
DeepVS[5]	0.660	0.230	0.321	0.450	0.625	—	0.195	—
Hard-attention[9]	—	0.250	0.357	0.504	0.718	—	0.230	—
Adaptive[10]	1.085	0.332	0.439	0.580	0.742	—	0.257	—
SCST:Att2all[14]	1.140	0.342	—	—	—	0.557	0.267	—
Up-Down:ResNet-101[15]	1.054	0.334	—	—	0.754	—	0.261	0.192
Our:baseline	1.048	0.338	0.447	0.558	0.752	0.550	0.263	0.190
Our:Both_joint	**1.146**	**0.348**	**0.465**	**0.613**	**0.777**	**0.561**	**0.267**	**0.203**

表 2 给出了英文标题生成的实验结果。从表中可以看出,在基准模型(Baseline)的基础上加入自注意力机制(Self_attentive),CIDEr 值从 1.048 提高到 1.062,其他评测指标也有提升。我们在自注意力机制的基础上分别进行词嵌入联合(Word_joint)、第二层 LSTM 隐藏状态输出联合(Hidden_joint)以及两者同时联合(Both_joint)。从实验结果中可以看到,无论是词嵌入联合,还是第二层 LSTM 隐藏状态联合,CIDEr 值都有提高,从 1.062 分别提高到 1.115 和 1.133;第二层 LSTM 隐藏状态联合的效果好于词嵌入联合;最好的效果是两者都联合,CIDEr 值达到 1.146。英文实验结果表明,本文所提方法的性能均好于基准模型,最好的结果 CIDEr 值从 1.048 提高到 1.146,BLEU_4 值从 0.338 提高到 0.348。

表 2　英文图像标题的实验结果

Table 2　Experimental results of English image caption

Method	*CIDEr*	*BLEU_4*	*BLEU_3*	*BLEU_2*	*BLEU_1*	*ROUGE_L*	*METEOR*	*SPICE*
Baseline	1.048	0.338	0.447	0.558	0.752	0.550	0.263	0.190
Self_attentive	1.062	0.341	0.448	0.585	0.748	0.551	0.260	0.192
Word_joint	1.115	0.343	0.450	0.588	0.751	0.550	0.259	0.195
Hidden_joint	1.133	0.347	**0.467**	**0.615**	**0.779**	0.561	**0.269**	0.201
Both_joint	**1.146**	**0.348**	0.465	0.613	0.777	**0.561**	0.267	**0.203**

5.4.2　日文标题的实验结果

表 3 给出了日文标题生成的实验结果。从中可以得到与英文标题实验结果类似的结论:无论是词嵌入联合模型,还是第二层 LSTM 隐藏状态联合模型,性能都较基准模型有提升;性能最好的是两种模型同时联合,其中 BLEU_4 值从基准的 0.388 提升到 0.398,CIDEr 值从 0.970 提升到 1.001。

表 3　日文图像标题的实验结果

Table 3　Experimental results of Japanese image caption

Method	*CIDEr*	*BLEU_4*	*BLEU_3*	*BLEU_2*	*BLEU_1*	*ROUGE_L*	*METEOR*	*SPICE*
Baseline	0.970	0.388	0.491	0.615	0.759	0.579	0.309	0.274
Self_attentive	0.976	0.391	0.490	0.617	0.761	0.580	0.308	0.276
Word_joint	0.991	0.398	0.496	0.623	0.764	0.583	0.310	0.280
Hidden_joint	0.997	0.397	0.496	0.621	0.765	0.582	0.298	0.279
Both_joint	**1.001**	**0.398**	**0.497**	**0.623**	**0.764**	**0.587**	**0.306**	**0.280**

5.5　实例分析

图 5 给出了两张图片同时生成英文和日文标题的结果。其中,Baseline 是基准模型生成的标题,Self_attentive 是在基准模型基础上加入自注意力机制生成的标题,Word_joint 和

Hidden_joint 分别表示在加入自注意力基础上词嵌入联合模型和隐藏状态联合模型生成的标题，Both_joint 表示词嵌入模型和隐藏状态模型同时联合生成的标题。

英文标题生成结果：
Baseline：a group of people standing on top of a beach
Self_attentive：a group of people on a beach with surfboards
Word_joint：a group of people walking across a beach
Hidden_joint：a group of people standing on a beach holding surfboards
Both_joint：a group of people walking across a beach holding surfboards

日文标题生成结果：
Baseline：サーフ ボード を 持っ た 人 が 海 に 入っ て いる
Self_attentive：サーフ ボード を 持っ た 人 が 2 人 いる
Word_joint：サーフ ボード を 持っ た 人 が 歩い て いる
Hidden_joint：サーフ ボード を 持っ た 人 が 二 人 いる
Both_joint：サーフ ボード を 持っ た 人 が 波打ち際 を 歩い て いる

英文标题生成结果：
Baseline：a person holding a dog in a car
Self_attentive：a person holding a dog in a car
Word_joint：a dog sitting in the back of a car
Hidden_joint：a person holding a dog in a car
Both_joint：a dog is looking out the windows of a car

日文标题生成结果：
Baseline：车 の 窓 から 顔 を 出し て いる 犬
Self_attentive：车 の 窓 から 顔 を 出し て いる 犬
Word_joint：车 の 窓 から 顔 を 出し て いる 犬
Hidden_joint：车 の 运転 席 に 犬 が 座っ て いる
Both_joint：车 の 窓 から 犬 が 顔 を 出し て いる

图 5 模型预测结果

Fig. 5 Prediction resuls of models

第一张图片中，从英文标题生成结果来看，基准模型预测的结果是“a group of people standing on top of a beach”，但是从给出的图片中我们人类大体上将这张图片描述为“a group of people walking across a beach holding surfboards”，因此与人类描述的结果相比，基准模型预测的结果遗漏了“holding surfboards”，并且把动词“walking”预测成了“standing”。相对而言，基于隐藏状态联合模型生成的标题则描述得更加准确一些，其中最准确的是同时联合词嵌入和隐藏状态的方法，其预测的结果不但全面，而且表达更加自然。从第二张图片可以更明显地看到，同时联合词嵌入和隐藏状态的方法预测的结果更加全面，其成功预测到了“windows of a car”；而其他几种模型都没能预测到这一表述。由此可见，本文提出的同时联合生成模型的方法可以捕获到更多的图像信息。

结束语 本文提出的图像双语标题联合生成的方法紧密结合了计算机视觉和自然语言理解任务，通过门控制网络来联合两门语言特征。使用门控制的方法，解码时我们可以进行词嵌入联合，也可以进行隐藏状态之间的联合。实验结果表明，无论哪种联合方法，性能都优于基准模型的性能，尤其是当两种联合方法同时进行时，模型性能最好。这种方法虽然有效地利用了两门语言特征，但是由于不同语种的语序有所不同，因此两种语言的联合方法可能存在融合噪声的问题，未来会致力于减小这种噪声对方法性能的影响；同时，考虑模型的泛化能力，利用 Transformer[32] 模型，在解码端使用多头注意力机制（Multi-Head Attention）代替 LSTM 作为图像标题生成模型，使模型能应用在多种语言中。

参 考 文 献

[1] ALI F, HEJRATI M, AMIN M S, 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] 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.

[3] 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, MA, USA: IEEE, 2015: 3156-3164.

[4] 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. 2015: 3128-3137.

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

[6] 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. 2015: 2407-2415.

[7] 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. 2016: 203-212.

[8] XU K, BA J, KIROS R, 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: 2048-2057.

[9] 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 Pat-tern Recognition. 2017:375-383.

[10] 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. 2017:5659-5667.

[11] LIXR, LANWY, DONGJF, 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.

[12] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C]// Proceedings of the 32nd International Conference on Machine Learning. Lille, France: JMLR. org, 2015:1-9.

[13] RENNIE S J, MARCHERET E, MROUEH Y, et al. Self-critical sequence training for image captioning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:7008-7024.

[14] 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 Conference on Computer Vision and Pattern Recognition. 2018:6077-6086.

[15] DOGNIN P L, MELNYK I, MROUEH Y, et al. Adversarial Semantic Alignment for Improved Image Captions[J]. arXiv: 1805.00063v3.

[16] BITEN A F, GOMEZ L, RUSIÑOL, MARÇAL, et al. Good News, Everyone! Context driven entity-aware captioning for news images[J]. arXiv:1904.01475.

[17] KIM D J, CHOI J, OH T H, et al. Dense Relational Captioning: Triple-Stream Networks for Relationship-Based Captioning[J]. arXiv:1903.05942v3.

[18] MITRA S, AVRA L J, MCCLUSKEY E J, et al. Scan synthesis for one-hot signals[C]// Proceedings International Test Conference. IEEE, 1997:714-722.

[19] WERLEN L M, PAPPAS N, RAM D, et al. Self-attentive residual decoder for neural machine translation[J]. arXiv: 1709.04849.

[20] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[C]// European conference on computer vision. Cham: Springer, 2014:740-755.

[21] YOSHIKAWA Y, SHIGETO Y, Takeuchi A. Stair captions: Constructing a large-scale japanese image caption dataset[J]. arXiv:1705.00823, 2017.

[22] PAPINENI K, ROUKOS S, WARD T, 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.

[23] DENKOWSKI M, LAVI A. Meteor universal: Language specific translation evaluation for any target language[C]// Proceedings of the Ninth Workshop on Statistical Machine Translation. 2014:376-380.

[24] LIN C Y. Rouge: A package for automatic evaluation of summaries[C]// Post-Conference Workshop of ACL 2004. 2004.

[25] VEDANTAM R, ZITNICK C L, PARIKH D, et al. CIDEr: Consensus-based image description evaluation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:4566-4575.

[26] ANDERSON P, FERNANDO B, JOHNSON M, et al. Spice: Semantic propositional image caption evaluation[C]// European Conference on Computer Vision. Cham: Springer, 2016: 382-398.

[27] 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. 2016:770-778.

[28] RUSSAKOVSKY O, DENG J, SU H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3):211-252.

[29] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. arXiv:1412.6980.

[30] WISEMAN S, RUSH A M. Sequence-to- sequence learning as beam-search optimization[J]. arXiv:1606.02960.

[31] IOFFE S, SZEGEDY C. 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.

[32] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017:5998-6008.

ZHANG Kai, born in 1992, graduate student, is a member of China Computer Federation. His main research interests include natural language processing, machine translation and image caption.

LI Jun-hui, born in 1983, associate professor. His main research interests include natural language processing and machine translation.