中图法分类号:(此号在中国图书馆分类法中查)　**文献标识码**:A　**文章编号**:1006-8961(年) -

论文引用格式：

引入语义匹配和语言评价的跨语言图像描述

张静[1,2,3]，郭丹[1,2,3]*，宋培培[1,2,3]*，李坤[1,2,3]，汪萌[1,2,3]

1. 合肥工业大学计算机与信息学院,合肥 230601； 2. 大数据知识工程教育部重点实验室（合肥工业大学）,合肥 230601；3. 智能互联系统安徽省实验室（合肥工业大学）,合肥 230601

摘　要：目的 目前基于深度学习的图像描述已经取得了显著效果，但大多数工作都是为图像生成英文描述，非英语母语者也应从现有的研究成果中受益。由于缺乏图像与目标语言域的成对数据，现有的跨语言描述方法都是基于轴（源）语言而转化为目标语言的，由于转化过程中的语义噪音干扰，生成的句子存在不够流畅以及与图像视觉内容关联弱等问题。因此，本文提出了一种引入语义匹配和语言评价的跨语言图像描述模型。**方法** 首先，本文选择了基于编码器-解码器的图像描述基准网络框架。其次，为了兼顾图像及其轴语言所包含的语义知识，本文构建了一个源域语义匹配模块；为了学习目标语言域的语言习惯，本文还构建了一个目标语言域评价模块。基于上述两个模块，对图像描述模型进行语义匹配约束和语言指导：1）图像&轴语言域语义匹配模块通过将图像、轴语言描述以及目标语言描述映射到公共嵌入空间来衡量各自模态特征表示的语义一致性。2）目标语言域评价模块则依据目标语言风格，对所生成的描述句子进行语言评分。**结果** 针对跨语言的英文图像描述任务，本文在MSCOCO（Microsoft common objects in context）数据集上进行了测试。与目前最先进的方法相比，本文方法在BLEU（bilingual evaluation understudy）-2、BLEU-3、BLEU-4和METEOR（metric for evaluation of translation with explicit ordering）等4个评价指标上的得分分别提升了1.4%，1.0%，0.7%和1.3%。针对跨语言的中文图像描述任务，本文在AIC-ICC（image Chinese captioning from AI challenge）数据集上进行了测试。与目前最先进的方法相比，本文方法在BLEU-1、BLEU-2、BLEU-3、BLEU-4、METEOR和CIDEr（consensus-based image description evaluation）等六个评价指标上的评分分别提升了5.7%，2.0%，1.6%，1.3%，1.2%和3.4%。**结论** 本文模型中图像&轴语言域语义匹配模块引导模型学习了更丰富的语义知识，目标语言域评价模块约束模型生成更加流畅的句子，因此本文模型适用于跨语言图像描述生成任务。

关键词： 跨语言；图像描述；强化学习；神经网络；轴语言

Cross-lingual image captioning based on semantic matching and language evaluation

Zhang Jing[1,2,3], Guo Dan[1,2,3]*, Song Peipei[1,2,3]*, Li Kun[1,2,3], Wang Meng[1,2,3]

1. School of Computer and Information Engineering, Hefei University of Technology, Hefei 230601, China;

2. Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology), Ministry of Education, Hefei 230601, China;

3. Intelligent Interconnected Systems Laboratory of Anhui Province (Hefei University of Technology), Hefei 230601, China.

收稿日期:　　；　**修回日期**:

* **通信作者**:郭丹　E-mail: guodan@hfut.edu.cn，　宋培培　E-mail: beta.songpp@gmail.com

基金项目:国家自然科学基金(62020106007，61876058)；中央高校基本科研业务费专项资金(JZ2020HGTB0020)

Supported by: National Natural Science Foundation of China (62020106007, 61876058)；Fundamental Research Funds for the Central Universities (JZ2020HGTB0020)

Abstract: Objective With the development of deep learning, image captioning has achieved great success. Image captioning can not only be applied to infant education, web search, and human-computer interaction but also can help visually impaired people to better obtain invisible information. Most of the image captioning works are proposed for captioning in the English language. However, the success of image captioning should not be restricted by language and non-native English speakers should benefit from existing English-based research results. Generating image captions in different languages is worth to be explored. The main challenge of cross-lingual image captioning is the lack of paired image-caption datasets in the target language. It is difficult and expensive to collect a large-scale image caption dataset for every target language. Benefiting from existing large-scale English captioning datasets and translation models, using the pivot language (*e.g.*, English) to bridge the image and the target language (*e.g.*, Chinese) is currently the main backbone framework for cross-lingual image captioning. However, such a language-pivoted approach suffers from disfluency and poor semantic relevance to images. To address these issues, a cross-lingual image captioning model based on semantic matching and language evaluation is proposed in this paper. **Method** First, the proposed model is constructed by a native encoder-decoder framework, in which the convolutional neural network extracts image features, and the recurrent neural network generates the description. The pivot language (source language) descriptions are transformed into the target language sentences by a translation API, which is regarded as pseudo captioning labels of the images. The proposed model is initialized with pseudo-labels. However, the captions generated by the initialized model are always in the repetitive combination of high-frequency vocabulary, or the language style of pseudo-labels, or poor relevancy with image content. It is worth noting that the pivot language written by humans is a correct description for the image content and contains the consistent semantics of the image. Therefore, considering the semantic guidance of the image content and pivot language, a semantic matching module based on the source corpus is proposed. Moreover, the language style of the generated captions differs greatly from the human-written target languages. To learn the language style of the target languages, a language evaluation module under the guidance of target language is proposed. The above two modules perform the constraints of semantic matching and language style on the optimization of the proposed captioning model. The methodological contributions are listed as follows. 1) The semantic matching module based on image and language labels in the source domain is a semantic embedding network. To tackle the semantic matching of image, pivot language, and generated sentence, we map these multimodal data into the embedding space and calculate semantic relevance. This model ensures the semantic enhancement of generated sentence to the visual content in the image. 2) The semantic evaluation module based on corpus in the target domain encourages the style of generated sentences to resemble the target language style. Under the joint rewards of semantic matching and language evaluation, the proposed model is optimized to generate sentences that more fluent and more semantically related to the image. As a fact, the semantic matching reward and language evaluation reward are performed in a reinforcement learning mode. **Result** In order to verify the effectiveness of our model, two sub-task experiments are carried out in this paper. 1) The cross-lingual English image captioning task is evaluated on the MSCOCO image-English dataset, which is trained under AIC-ICC image-Chinese dataset and MSCOCO English corpus. Compared with the state-of-the-art method, the metric values of BLEU-2, BLEU-3, BLEU-4, and METEOR of our method have increased by 1.4%, 1.0%, 0.7% and 1.3% respectively. 2) The cross-lingual Chinese image captioning task is evaluated on the AIC-ICC image-Chinese dataset, which is trained under MSCOCO image-English dataset and AIC-ICC Chinese corpus. Compared with the state-of-the-art method, the performances of BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, and CIDEr of our method have increased by 5.7%, 2.0%, 1.6% , 1.3%, 1.2%, and 3.4% respectively. **Conclusion** The semantic matching module guides the model to learn the relevant semantics in both image and pivot language description. The language evaluation module learns the data distribution and language style of the target corpus. The semantic reward and language reward obtained from these two modules have positive effects for cross-lingual image captioning, which not only improve the semantic relevance of the sentence but also further improve the fluency of the sentence.

Key words： cross-lingual; image captioning; reinforcement learning; neural network; pivot language

0 引　言

图像描述任务是指给定一张图片，计算机能够自动生成正确的语言描述（Farhadi 等，2010），涉及目标检测（Tang 等，2017；Li 等，2020）、关系推理（Hou 等，2020）、语言序列生成（Zhou 等，2019）等多项前沿技术。其成果不仅可应用于网页检索、人机交互等应用领域，还可以帮助视障人士更好地获取和理解信息。目前，得益于深度学习的快速发展和现有大规模成对的图像-句子数据集存在，图像描述任务已经取得了显著成果（Wang 等，2019；Ji 等，2020；Luo 等，2020）。然而，大多数现有工作关注于图片英文描述生成；非英语母语者也应该从现有的研究成果中受益。跨语言图像描述任务（例如，从英文描述迁移至中文描述）也逐渐成为研究的一种趋势（Lan 等，2017；Gu 等，2018；Song 等，2019）。

跨语言描述任务存在的一个客观原因在于缺少大规模目标语言的图像描述数据集。如图 1 所示，在训练数据集中，图像只有成对的轴语言描述（即源语言，例如英文）和无关的目标语料库（例如中文）。收集成对的图像-句子数据集是一项耗时费力的工作，为世界上任意一种语言都构建图像-句子成对数据集，代价更为昂贵。幸运地是，现已存在大规模的英文-图像对描述数据集。在具有丰富的目标语言语料库的前提下，将已有的轴语言描述（例如英文）视为连接图像和目标语言（例如中文）描述的桥梁是解决跨语言图像描述任务的一种常见做法。Lan 等人（2017）将轴语言数据通过翻译模型得到目标语言数据，视为句子伪标签，同时引入句子流畅性评估模型，根据流畅性得分奖励赋予伪标签相应的权重，减少不流畅的伪标签句子在模型训练中的作用。Gu 等人（2018）则先利用轴语言训练图像描述模型生成轴语言描述，再将其由翻译模型得到目标语言描述，通过正则化轴语言编码器和目标语言解码器的词嵌入优化模型来减少轴语言与目标语言的风格化差异。上述两种代表性工作各自存在明显的弊端：一是伪标签存在翻译误差，不如人工标注语言自然流畅，过度依赖伪标签会导致模型生成的句子质量受限；二是关注图像自身语义信息到轴语言的翻译，忽视了轴语言作为真实标准引入的语义知识。

此外，如图 1 所示，中文描述与英文描述存在语言风格差异，图像的源英文描述为“A photo of the back of a girl in a yellow dress”（一个穿黄色长裙的女孩的背影照片，短语句式），而目标域中文描述属于常规主谓宾的句式，翻译风格为“一个穿着黄色裙子的女生站在喷泉前”。而且，强调语义也不尽相同，虽然句子中都出现了“黄色长裙的女生”，但真实的中文描述以“女生”为描述中心，而英文描述却以“一张照片”为描述中心。

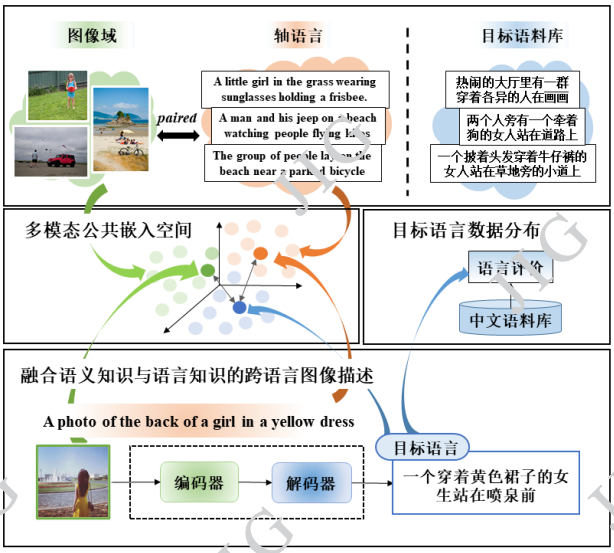

图 1 跨语言图像描述任务及本文解决方案

Fig.1 The task of cross-lingual image captioning and our solution

针对上述弊端及挑战，本文提出了一种引入语义匹配和语言评价奖励的跨语言模型图像描述方法。具体而言，为了兼顾图像自身语义及其轴语言所包含的语义知识，本文分别构建了一个源域语义匹配模块和一个目标语言域评价模块，从而对模型进行语义匹配约束和语言知识指导：1）图像&轴语言域语义匹配模块是一个多模态视觉语义嵌入网络，通过将图像、轴语言以及目标语言描述映射到公共嵌入空间来衡量各自模态特征表示的语义一致性。2）目标语言域评价模块在独立的目标语言语料库上学习目标语言的数据分布和表达方式，并依据目标语言风格，对所生成的描述句子进行语言知识评分。本文方法在语义匹配和语言知识的共同约束下，从而生成更加自然流畅、语义更相关的目标语言描述。

1 相关工作

1.1 图像描述生成

图像描述任务是涉及到计算机视觉和自然语言

处理两个研究领域的交叉任务。目前基于深度学习的图像描述工作已经取得了有效进展。Vinyals 等人（2015）首次提出端到端的 CNN（convolutional neural networks）编码器-RNN（recurrent neural network）解码器结构，以最大化输入图像的目标句子的似然概率为训练目标求解图像描述任务。此后，在编码器-解码器的框架基础上出现了各类融合注意力机制的方法。Xu 等人（2015）把图像分割为多个区域块，将区域块的各自空间注意力融合到图像卷积特征计算中，实现单词和局部视觉信息的对齐。Anderson 等人（2018）在图像区域级和对象级（object-level）特征上分别计算注意力。上述方法建立在视觉空间特征上，没有考虑字幕丰富性。Wang 等人（2019）提出了一个分层注意力网络，将文本特征与区域块、对象视觉特征一起输入特征金字塔层次结构同步计算，融合不同语义预测下一时刻的词。Ji 等人（2020）引入记忆机制，在序列生成过程中建立强记忆连接，关注不同时间步下注意力区域的变化以及关联性。

另一方面，Ranzato 等人（2015 年）早就指出图像编码器-句子解码器模型的改进并不能解决图像描述任务中训练-测试目标不匹配的问题。模型在训练时通常以真实单词最大似然概率为训练目标，在测试时却使用 BLEU，CIDEr 等评价指标。因此，强化学习的方法被引入到图像描述任务中。Rennie 等人（2017）提出自批判序列训练（self-critical sequence training），将当前模型在推理阶段生成的句子的特定指标（CIDEr）评分作为基准奖励以减少方差。比基准奖励得分高的句子得到鼓励，而比基准奖励得分低的句子被抑制，经过反复循环的强化训练，模型生成 CIDEr 奖励更好的句子。Liu 等人（2018）提出一个自检索模块以优化描述句子的多样性和独特性，该模块提供的奖励可以针对图像内容生成差异性描述句子。可见，语义指标的考量已被引入优化目标，成为传统图形描述任务的一个研究方向。本文延续采用编码器-解码器的基准框架求解跨语言图像描述生成任务，并将语义奖励优化引入本文方法。

1.2 跨语言图像描述生成

跨语言图像描述任务发展较慢，目前仍处于探索阶段。为了解决在不成对的图像-目标文本数据集上的图像描述问题，Lan 等人（2017）直接利用翻译模型得到图片在目标语言的伪标签，同时提出一个句子流畅性评估模块，根据流畅度评分对于流畅与不流畅的句子的目标损失赋予不同的权重，以抑制不流畅句子在训练中的负面作用。其实，即便生成不流畅的句子，也能包含正确的图片对象信息。目前，跨语言图像描述方法现大多采用基于轴语言转换的方法。Gu 等人（2018）提出了基于轴语言的跨语言描述模型，先使用图像描述模型为图像生成轴语言，然后利用翻译模型得到目标语言。为了克服不同语言的风格化差异，该模型进一步正则化轴语言的编码器和目标语言的解码器的词嵌入参数。当然，基于轴语言到目标语言的翻译误差也会被引入，翻译错误不会随着参数传递而缓解。

在语义奖励方面，Song 等人（2019）则为了提升跨语言描述与图像的视觉相关性，提出了一种自监督的奖励模型（Self-supervised rewarding，SSR），利用句子级语义匹配和概念级语义匹配分别提供粗粒度和细粒度的视觉相关奖励。然而，由于不同标注者的主观关注点不同，同一张图片的不同描述可能包含不同的概念，得到的概念级语义奖励并不完全可靠。此外，得益于视觉概念检测 Faster R-CNN 模型（Ren 等，2015）的良好性能，Feng 等人（2019）提出了无监督图像描述模型，引入 Faster R-CNN（region based CNN）对生成句子进行概念约束，采取图像-句子双向语义重构的方法来进一步提升句子质量。Ben 等人（2021）提出一个语义约束自学习框架，迭代地进行伪标签生成和图像描述模型训练。这两个工作都由图像中检测出的对象（Object）作为引导，来加强输入图像和输出句子之间的语义对齐。然而，视觉概念检测器 Faster R-CNN 是在大规模的英文图像描述集上预训练好的，仅适用于英文概念检测；对于其他语言的概念尤其是在缺失训练数据集的情况下无法直接应用。

本文同样关注于语义奖励优化的正向反馈。不同于概念语义反馈，本文方法关注在特征映射空间中图像、轴语言句子、目标语言域句子三者之间的语义匹配（句子级语义反馈），还引入了目标语言域的文本语料对生成的翻译句子实现语言评分，以期待生成跟目标语言域风格一致的图像描述。

2 本文模型

如图 2 所示，本文所提出跨语言图像描述模型由 3 部分构成：1）朴素的图像编码器-句子解码器（图像描述生成）模块；2）图像&轴语言域语义匹配模块，用于提供语义匹配的奖励优化，兼顾了源域图像与轴语言的语义信息，映射图像、轴语言、

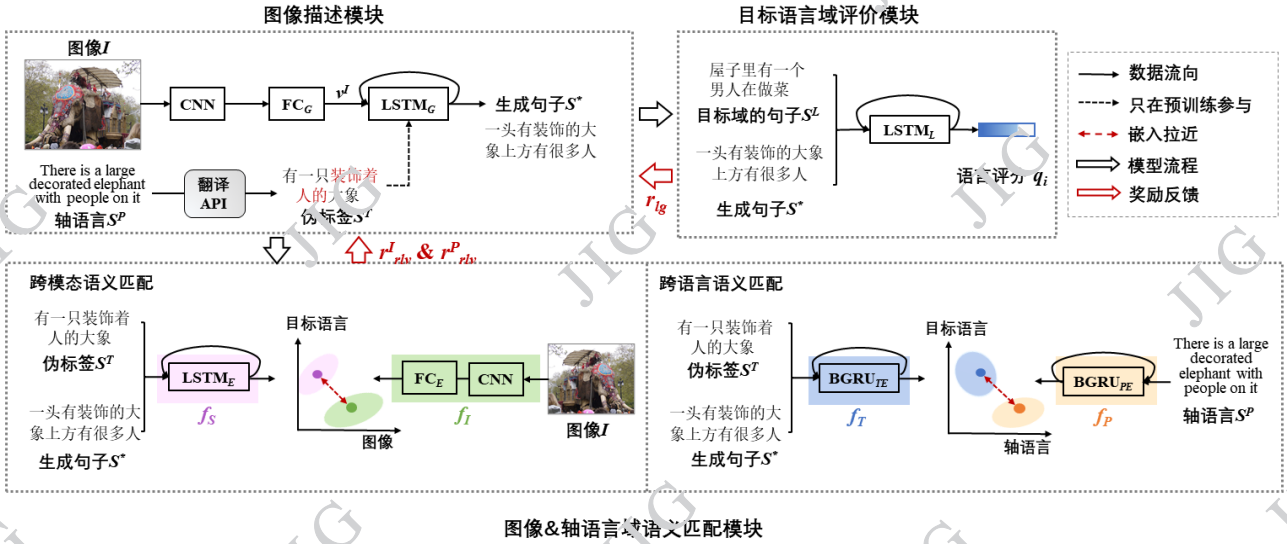

图 2 跨语言图像描述模型

Fig.2 Cross-lingual image captioning model

目标语言于公共嵌入空间进行语义匹配计算；3）目标语言域评价模块，用于提供语言评价奖励，引入目标域数据分布知识进行语言评价约束。第一个模块负责描述句子的生成，后者两个模块引导模型学习语义匹配约束和语言知识优化，使模型生成更加流畅和语义丰富的描述。

2.1 图像编码器-句子解码器模块

本文采用朴素的图像编码器-句子解码器框架生成描述句子。具体是指，本文使用预训练网络模型 ResNet（residual net）-101（He 等，2016）和一层全连接层（fully-connected layer，记为 FC_G）提取图像 $\boldsymbol{I}$ 的特征 $\boldsymbol{v}^{\boldsymbol{I}}$；使用单层 LSTM（long short-term memor，记为 $LSTM_G$）对 $\boldsymbol{v}^{\boldsymbol{I}}$ 进行解码生成当前时间步的单词。与前人工作（Lan 等，2017；Song 等，2019）类似，本文采用百度翻译 API[1]对图像 $\boldsymbol{I}$ 的源域描述语言 $\boldsymbol{S}^{\boldsymbol{P}}$ 获取目标域伪句子标签 $\boldsymbol{S}^{\boldsymbol{T}}$ 对此模块进行初始化。在模型初始化训练中，预训练模型 ResNet-101 不参与模型优化，全连接层 FC_G 和 $LSTM_G$ 参与。优化函数目标设为最小化句子中正确单词的负对数概率，如公式（1）所示：

$$L(\theta_G)=-\sum_{i=1}^{N}\log\left(p_{\theta_G}\left(\boldsymbol{w}_i^{(T)}\middle|\boldsymbol{v}^{\boldsymbol{I}},\boldsymbol{w}_{0:i-1}^{(T)}\right)\right) \tag{1}$$

式中，N 是句子 $\boldsymbol{S}^{\boldsymbol{T}}=\{\boldsymbol{w}_0^{(T)},\ \boldsymbol{w}_1^{(T)},\ \cdots,\ \boldsymbol{w}_N^{(T)}\}$的长度，单词 $\boldsymbol{w}_0^{(T)}$被设置为开始符<bos>，θ_G 是本模块 FC_G 和 $LSTM_G$ 的学习参数。

[1] http://api.fanyi.baidu.com

2.2 图像&轴语言域语义匹配模块

由 2.1 节初始化后的模型生成的描述具有如下特性：对伪标签的简单模仿或是高频词汇的重复组合，或缺少与图片内容的相关性。人工标注的轴语言具有丰富的语义，是对图片信息的切实描述。轴语言与图片应包含一致的语义信息。同时结合图片与轴语言两者语义信息，本文提出了一种多模态语义匹配模块进行语义相似度约束。

2.2.1 跨模态语义匹配

针对异构的图像与句子，本文首先将图像和句子映射到公共嵌入空间，衡量语义的关联度。如图 2 所示，图像语义嵌入网络 f_I 由 CNN 编码器（本文使用预训练网络模型 ResNet-101）和一层全连接层（记为 FC_E）构成。文本语义嵌入网络 f_S 由单层 LSTM（记为 $LSTM_E$）构成。$LSTM_E$ 最后时刻的隐向量被定义为输入句子在公共嵌入空间的语义向量。本文将图像-伪标签数据对（$\boldsymbol{I}$，$\boldsymbol{S}^{\boldsymbol{T}}$）输入，即可得到图像 $\boldsymbol{I}$ 在公共语义空间的嵌入特征 $f_I(\boldsymbol{I})$，句子 $\boldsymbol{S}^{\boldsymbol{T}}$ 在公共语义空间的嵌入特征 $f_S(\boldsymbol{S}^{\boldsymbol{T}})$。对于匹配对（$\boldsymbol{I}$，$\boldsymbol{S}^{\boldsymbol{T}}$），寻找同组 batch 的句子集中与 $\boldsymbol{I}$ 不匹配的负例 $\boldsymbol{S}^{\boldsymbol{T}'}$，同组 batch 的图像集中与 $\boldsymbol{S}^{\boldsymbol{T}}$ 不匹配的负例 $\boldsymbol{I}'$。以最小化双向 ranking 损失对公共语义空间进行预训练，如公式（2）所示：

$$L(\theta_\mu)=\sum_{\boldsymbol{I}}\sum_{\boldsymbol{S}^{T'}}\max\left(0,\Delta-f_I(\boldsymbol{I})f_S(\boldsymbol{S}^{\boldsymbol{T}})+f_I(\boldsymbol{I})f_S(\boldsymbol{S}^{\boldsymbol{T}'})\right)+\sum_{\boldsymbol{I}'}\sum_{\boldsymbol{S}^{\boldsymbol{T}}}\max\left(0,\Delta-f_I(\boldsymbol{I})f_S(\boldsymbol{S}^{\boldsymbol{T}})+f_I(\boldsymbol{I}')f_S(\boldsymbol{S}^{\boldsymbol{T}})\right) \tag{2}$$

式中，Δ表示界限超参数；θ_μ 是本模块 FC_E 和 $LSTM_E$ 的学习参数。

2.2.2 跨语言语义匹配

同时，本文还有轴语言句子-伪标签句子对（$\boldsymbol{S^P}$，$\boldsymbol{S^T}$）。本节引入跨语言语义匹配计算增强句子的语义相关性，采用类似 2.2.1 节的语义嵌入网络机制对齐目标语言与轴语言嵌入向量。本节中目标语言和轴语言的编码器都采用单层 BGRU（bidirectional gated recurrent unit）结构，以 BGRU 最后时刻的隐向量作为句子特征向量。记为 f_P 是轴语言特征映射器（$BGRU_{PE}$），f_T 是目标语言特征映射器（$BGRU_{TE}$）。同样地，以最小化双向 ranking 损失对公共语义空间进行预训练，如公式（3）所示：

$$L(\theta_\rho)=\sum_{S^P}\sum_{S^{T'}}\max\left(0,\Delta-f_P(\boldsymbol{S^P})f_T(\boldsymbol{S^T})+f_P(\boldsymbol{S^P})f_T(\boldsymbol{S^{T'}})\right)+\sum_{S^{P'}}\sum_{S^T}\max\left(0,\Delta-f_P(\boldsymbol{S^P})f_T(\boldsymbol{S^T})+f_P(\boldsymbol{S^{P'}})f_T(\boldsymbol{S^T})\right) \tag{3}$$

式中，对于匹配对（$\boldsymbol{S^P}$，$\boldsymbol{S^T}$），$\boldsymbol{S^{T'}}$ 是同组 batch 的伪标签句子集中与 $\boldsymbol{S^P}$ 不匹配的负例，$\boldsymbol{S^{P'}}$ 是同组 batch 的轴语言句子集中与 $\boldsymbol{S^T}$ 不匹配的负例。θ_ρ 是本模块 $BGRU_{PE}$ 和 $BGRU_{TE}$ 的学习参数。

2.3 目标语言域评价模块

由于生成的描述目前与目标语料缺乏关联，因此生成的描述句子与真实目标句子常常语言风格差异明显，本节利用目标域语料集提供目标域语言表达进而优化描述语言质量。具体来说，本模块独立在目标语言数据集上预训练一个可以提供语言评价奖励的模块，对于输入的单词要做正确的分类检测。该模块采用 LSTM（记为 $LSTM_L$），将句子逐词输入 $LSTM_L$，再利用 $LSTM_L$ 预测当前输入词的概率。将目标语料库中长度为 N 的句子 $\boldsymbol{S^L}=\{\boldsymbol{w_0^{(L)}},\boldsymbol{w_1^{(L)}},\cdots,\boldsymbol{w_N^{(L)}}\}$ 作为 $LSTM_L$ 的输入，预训练的目标函数为最小化句子中正确单词的负对数概率，如公式（4）所示：

$$L(\theta_\omega)=-\sum_{i=1}^{N}\log\left(p_{\theta_\omega}\left(\boldsymbol{w_i^{(L)}}\middle|\boldsymbol{w_{0:i-1}^{(L)}}\right)\right) \tag{4}$$

式中，θ_ω 是本模块 $LSTM_L$ 的学习参数。

2.4 基于语义匹配和语言奖励的模型优化

在进行了上述三个模块初始化的预训练自学习后，本节联合三个模块一起实现 2.1 节中图像编码器-句子解码器模块的奖励优化学习。具体而言，利用 2.2 节的语义匹配奖励和 2.3 节的语言评价奖励对 2.1 节模块进行优化。其中，语义匹配奖励衡量目标语言与图像、轴语言在视觉对象（object）、对象关系（relation）上的一致性。首先，输入图像 $\boldsymbol{I}$，由 2.1 节自动生成目标语言域的句子 $\boldsymbol{S^*}$。其次，计算如下语义匹配奖励和语言评价奖励：

1）图像-句子匹配奖励。图像 $\boldsymbol{I}$ 经由视觉语义嵌入网络 f_I 映射，句子 $\boldsymbol{S^*}$ 经由文本语义嵌入网络 f_S 映射到公共嵌入空间，其跨模态语义匹配奖励可以定义为：

$$r_{rlv}^{I}(\boldsymbol{S^*})=\frac{f_I(\boldsymbol{I})f_S(\boldsymbol{S^*})}{\|f_I(\boldsymbol{I})\|\|f_S(\boldsymbol{S^*})\|} \tag{5}$$

2）轴语言-句子匹配奖励。同样地，源域句子 $\boldsymbol{S^P}$ 经由轴语言特征映射器 f_P 映射，句子 $\boldsymbol{S^*}$ 经由目标语言特征映射器 f_T 映射，其跨语言语义匹配奖励可以定义为：

$$r_{rlv}^{P}(\boldsymbol{S^*})=\frac{f_P(\boldsymbol{S^P})f_T(\boldsymbol{S^*})}{\|f_P(\boldsymbol{S^P})\|\|f_T(\boldsymbol{S^*})\|} \tag{6}$$

式中，$\boldsymbol{S^P}$ 是与图像 $\boldsymbol{I}$ 匹配的轴语言描述。

3）目标域句子语言评价奖励。将句子 $\boldsymbol{S^*}$ 的每个单词迭代输入 2.3 节在目标语言域训练好的模块 $LSTM_L$，语言评价的过程如下：

$$[q_i,h_i^L]=\mathrm{LSTM}_L\left(\boldsymbol{w_i^{(*)}},h_{i-1}^L;\theta_\omega\right),i\in\{1\ldots N\} \tag{7}$$

式中，$\boldsymbol{S^*}=\{\boldsymbol{w^{(*)}_0},\boldsymbol{w^{(*)}_1},\cdots,\boldsymbol{w^{(*)}_N}\}$，$\boldsymbol{w^{(*)}_0}$ 是句子的开始标志<bos>，N 是句子 $\boldsymbol{S^*}$ 的长度；h_i^L 是第 i 时间步的隐向量，q_i 是第 i 时间步在词汇字典上的概率向量，维度等于词汇量大小。在概率向量 q_i 中，单词 $\boldsymbol{w^{(*)}_i}$ 对应的预测概率表示为 $q_i(\boldsymbol{w^{(*)}_i})$。本文将句子 $\boldsymbol{S^*}$ 的语言评价奖励定义为所有时间步下单词 $\boldsymbol{w^{(*)}_i}$ 的对数预测概率的期望：

$$r_{lg}(\boldsymbol{S^*})=\frac{1}{N}\sum_{i=0}^{N}\log\left(q_i\left(\boldsymbol{w_i^{(*)}}\middle|\boldsymbol{w_{0:i-1}^{(*)}}\right)\right) \tag{8}$$

整个跨语言描述模型的总奖励设置为：

$$r_{total}=\alpha r_{lg}+\beta r_{rlv}^{I}+\gamma r_{rlv}^{P} \tag{9}$$

式中 α、β 和 γ 是超参数，取值范围[0,1]。α、β 和 γ 为经验参数，最佳值设置见 3.2 节。

为减少模型训练时的期望梯度方差，本文遵循自批判序列训练方式。当前模型利用多项式分布采样方式得到句子 $\boldsymbol{S^*}$，另外默认按照最大概率贪婪采样方式得到句子 $\boldsymbol{S}$，以 $r_{total}(\boldsymbol{S})$ 作为基准奖励。对句子 $\boldsymbol{S^*}$ 的总体奖励可表示为 $r_{total}(\boldsymbol{S^*})$- r_{total}

（$\boldsymbol{S}$）：比基准奖励得分高的句子得到鼓励，而比基准奖励得分低的句子被抑制，经过反复循环的强化训练，模型生成语义匹配奖励更好和语言评价奖励更好的句子。因此，跨语言描述模型的最终目标损失可定义为：

$$L_{total}=-\sum_{i=1}^{N}\left(\begin{array}{l}\left(r_{total}\left(\boldsymbol{S}^{s}\right)-r_{total}\left(\boldsymbol{S}\right)\right)\times\\ \log \mathrm{P}_{\theta_G}\left(\boldsymbol{w}_{i}^{(*)}\left|\boldsymbol{v}^{\boldsymbol{I}},\boldsymbol{w}_{0:i-1}^{(*)}\right.\right)\end{array}\right) \tag{10}$$

式中，θ_G 是图像描述模块的参数。

3 实验及结果分析

为了验证模型在跨语言图像描述任务上的有效性，本文分别进行了两个子任务实验：以中文为轴语言实现图像英文描述和以英文为轴语言实现图像中文描述。

3.1 数据集及评价指标

本文采用两个基准数据集进行评测，如表 1 所示。1）英文数据集 MSCOCO（Lin 等，2014）包含 123,287 张图片，每张图片至少有 5 个人工标注的英文描述。实验遵循 Lin 等人（2014）提出的划分方式：113,287 张图片用作训练集，5,000 张图片用在验证集，5,000 张图片用作测试集。中文单词划分使用“结巴”工具[2]，保留出现频率不少于 5 的中文单词，同时将所有长度大于 16 的中文句子进行截断。英文单词的划分使用“斯坦福解析”工具[3]，保留出现频率不少于 5 的英文单词，同时将所有句子长度大于 20 的英文句子进行截断。2）中文数据集 AIC-ICC（Wu 等，2017）训练集有 208,354 张图片，验证集有 30000 张图片，每张图片包含 5 个人工标注的中文描述。AIC-ICC 没有官方公布的测试集，实验遵循 Song 等人（2019）提出的划分方式：在 30,000 张的验证集中随机采样 5000 张图片作为测试集，5000 张图片作为验证集，剩余 20,000 张图片归到训练集中。注意 AIC-ICC 和 MSCOCO 两者数据集中图像和句子各不相同。

在从中文跨到英文的图片描述任务中，以 AIC-ICC 中文数据集联合 MSCOCO 英文语料训练，使用 MSCOCO 测试集进行评测。在从英文跨到中文的图片描述任务中，以 MSCOCO 数据集联合 AIC-ICC 中文语料训练，使用 AIC-ICC 测试集进行评测。实验中，采用语义评估指标 BLEU，METEOR 和 CIDEr 对生成的图像描述进行评测。

[2] https://github.com/fxsjy/jieba
[3] http://nlp.stanford.edu:8080/parser/index.jsp

表 1 实验使用的数据集信息

Table 1 Statistics of the datasets used in our experiments

数据集	语言	训练集	验证集	测试集	描述
MSCOCO	英文	113287	5000	5000	5
AIC-ICC	中文	228354	5000	5000	5

注：“描述”表示每张图片对应的描述句子数量

3.2 训练设置

在图像编码器-句子解码器模块（图像描述生成模块）中，图像特征 $\boldsymbol{v}^{\boldsymbol{I}}$ 由预训练模型 ResNet-101 和一层全连接层提取，维度 d=512；并将其作为解码器 LSTM_G 第 0 时刻的隐向量输入。在跨模态语义匹配模块中，图像语义嵌入网络由预训练模型 ResNet-101 和一层全连接层组成，目标语言编码器采用单层的 LSTM_E 结构。在跨语言语义匹配模块中，轴语言和目标语言的编码器分别采用单层的 BGRU_{PE} 和 BGRU_{TE} 框架，隐藏层维度均是 512 维，BGRU 输出的维度是 1024 维。在目标语言域评估模块中，语言序列模型使用单层的 LSTM_L。本文所有的 LSTM 结构的隐藏层维度和单词嵌入维度均为 d=512 维。两个子任务实验在整个模型训练过程中，dropout 设置为 0.3，预训练时的 batchsize 设为 128，强化训练时的 batchsize 设为 256。

在语义匹配模块（2.2 节）和语言优化模块（2.3 节）预训练结束后，学习参数 θ_μ，θ_ρ 和 θ_ω 都保持固定。二者提供奖励共同引导图像描述生成模块（2.1 节）学习更多的源域语义知识和目标域语言知识。1）以中文为轴语言实现图像英文描述：图像描述生成模块预训练的学习率是 1e-3，源域语义匹配模块和目标语言域评价模块的预训练的学习率设为 2e-4。在使用语言评价奖励和多模态语义奖励训练时，图像描述生成模块的学习率是 4e-5，α、β 和 γ 分别取值 1，1，0.15。2）以英文为轴语言实现图像中文描述：图像描述生成模块预训练的学习率是 1e-3，源域语义匹配模块和目标语言域评价模块预训练的学习率设为 4e-4。在使用语言评价奖励和多模态语义匹配奖励训练时，图像描述生成模块的学习率是 1e-5，α、β 和 γ 的值分取值 1，1，1。

3.3 实验结果分析

表 2 不同奖励对于跨语言英文图像描述任务在 MSCOCO 测试集上的贡献和不同奖励对于跨语言中文图像描述任务在 AIC-ICC 测试集上的贡献

Table 2 The contribution of different rewards for cross-lingual English image captioning on MSCOCO test dataset and cross-lingual Chinese image captioning on AIC-ICC test dataset

任务	$L(\theta_G)$	r_{lg}	r^{I}_{rlv}	r^{P}_{rlv}	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
跨语言英文图像描述	√				41.9	21.4	10.3	5.3	14.3	14.1
	√		√	√	42.0	22.6	10.9	5.5	14.6	15.3
	√	√			49.0	26.3	13.1	6.9	14.0	16.4
	√	√	√		48.0	28.9	17.4	11.0	14.6	24.4
	√	√	√	√	**51.7**	**31.4**	**18.9**	**11.8**	**15.5**	**27.9**
跨语言中文图像描述	√				45.5	28.6	17.3	10.7	22.0	16.8
	√		√	√	46.4	29.2	17.9	11.4	22.6	18.2
	√	√			48.0	30.1	18.8	12.0	22.5	19.0
	√	√	√		51.0	31.5	19.5	12.2	22.9	20.4
	√	√	√	√	**51.7**	**32.9**	**20.9**	**13.6**	**24.0**	**21.7**

注：加粗字体为每项指标的最优值

3.3.1 消融实验

为了验证图像&轴语言域语义匹配模块和目标语言域评价模块的有效性，本文进行了消融实验。表 2 展示了从中文跨到英文的图片描述任务上的消融实验结果和从英文跨到中文的图片描述任务上的消融实验结果。表 2 中以公式（1）中 $L(\theta_G)$为目标函数的模型为 Baseline 基准模型；“√”表示相应奖励/损失参与了模型训练。具体地，使用 r^{I}_{rlv} 奖励表示跨模态语义匹配模块（2.2.1 节）参与训练，使用 r^{P}_{rlv} 奖励表示跨语言语义匹配模块（2.2.2 节）参与训练，使用 r_{lg} 奖励表示目标语言域评价模块（2.3 节）参与训练。联合使用奖励 r^{I}_{rlv}、r^{P}_{rlv} 和 r_{lg} 表示本文提出的模型。

据表 2 统计，在多模态语义相关性奖励 r^{I}_{rlv} &r^{P}_{rlv} 的作用下，各项指标性能得到提升，与 Baseline 相比，CIDEr 得分分别提升 1.2%和 1.4%。结果表明，图像&轴语言域语义匹配模块提升了句子语义相关性。目标域语言奖励 r_{lg} 在跨语言英文图片描述任务和跨语言中文图片描述任务上都起到积极的作用，与 Baseline 相比，CIDEr 得分分别上升了 2.3%和 2.2%。另外，在目标域语言奖励 r_{lg} 和图像-句子语义匹配奖励 r^{I}_{rlv} 的共同作用下，在两个跨语言图片描述任务上性能继续提升。与 Baseline 相比，CIDEr 得分分别上升了 10.3%和 3.6%。这表明对图像描述模型同时叠加奖励 r^{I}_{rlv} 和 r_{lg} 也使生成的描述与图片在语义上更加一致。随后，在奖励 r_{lg}、r^{I}_{rlv} 和 r^{P}_{rlv} 的共同作用下，各项指标上又出现了明显提升，与 Baseline 相比，CIDEr 得分分别上升了 13.8%和 4.9%。由此表明，在图像&轴语言域和目标语言域的共同指导下，跨语言图像描述模型学习到最佳丰富语义知识，提升了句子的流畅性和语义相关性。

此外，与仅有 r_{lg} 参与的实验相比，在 r_{lg} & r^{I}_{rlv} 的奖励下，CIDEr 得分分别上升了 8.0%和 1.4%，奖励 r^{I}_{rlv} 对两个子任务的作用差异明显。这是因为在前者跨语言英文图像描述子任务中，测试集 MSCOCO 的图片包含各种场景，例如人、动物、静物等，视觉语义丰富且更加多样性。此时，来自图像-句子语义匹配奖励 r^{I}_{rlv} 展示了优异的语义补充能力（上升 8.0%）。而在后者跨语言中文图像描述子任务中，测试集 AIC-ICC 视觉场景单一（图片多以人物为主），可以补充的视觉语义有限，本文方法依然提升了 1.4%。

3.3.2 跨语言英文图像描述主性能分析

表 3 展现了不同方法关于跨语言英文图像描述任务在 MSCOCO 测试集上的实验结果。本文工作与现有跨语言图像描述实验进行了对比，这些实验具体包括：1）Baseline：仅利用伪标签和公式（1）中损失函数初始化的模型（见 2.1 节）；2）2-Stage pivot - Google API（Gu 等，2018）使用图像描述生成模块生成轴语言，再将轴语言通过 Goggle 翻译器得到英文描述（目标语言）；3）2-Stage pivot（Gu

表 3 不同方法关于跨语言英文图像描述任务在 MSCOCO 测试集上的性能比较

Table 3 Performance comparison with different methods for cross-lingual English image captioning evaluated on the MSCOCO test dataset

任务	方法	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
跨语言英文图像描述	Baseline	41.0	21.4	10.3	5.3	14.3	14.1
	2-Stage pivot（Gu 等，2018）- Baidu API	41.1	20.9	9.8	4.9	14.0	14.0
	2-Stage pivot- Google API（Gu 等，2018）	42.2	21.8	10.7	5.3	14.5	17.0
	2-Stage pivot -joint model（Gu 等，2018）	46.2	24.0	11.2	5.4	13.2	17.7
	SSR-Baseline & CIDEr Reward（Song 等，2019）	44.0	22.0	10.5	5.3	13.0	14.6
	SSR（Song 等，2019）	**52.0**	30.0	17.9	11.1	14.2	**28.2**
	本文	51.7	**31.4**	**18.9**	**11.8**	**15.5**	27.9

注：加粗字体为每项指标的最优值

表 4 不同方法关于跨语言中文图像描述任务在 AIC-ICC 测试集上的性能比较

Table 4 Performance comparison with different methods for cross-lingual Chinese image captioning evaluated on the AIC-ICC test dataset

任务	方法	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
跨语言中文图像描述	Baseline	45.5	28.6	17.3	10.7	22.0	16.8
	2-Stage pivot（Gu 等，2018）- Baidu API	35.5	20.9	10.9	6.3	18.4	11.1
	SSR-Baseline & CIDEr Reward（Song 等，2019）	41.6	24.4	13.3	7.3	21.1	11.6
	SSR（Song 等，2019）	46.0	30.9	19.3	12.3	22.8	18.3
	本文	**51.7**	**32.9**	**20.9**	**13.6**	**24.0**	**21.7**

注：加粗字体为每项指标的最优值

等，2018） - Baidu API 采用上述（2）中图像描述模型框架生成轴语言描述后，采用翻译 API 得到目标语言描述。不同之处在于将 Goggle API 替换为 Baidu API。本文增加了 Baidu API 的测试；4）2-Stagepivot - joint model 是 Gu 等人（2018）提出的一种跨语言图像描述方法，同样先将图片通过图像描述模块生成轴语言描述，再将轴语言描述通过翻译模块得到目标语言，与前面两者不同之处在于共享了两个模块的编码器和解码器嵌入参数来减少风格化差异；5）SSR 是 Song 等人（2019）针对不成对的图像-句子数据集提出的跨语言描述模型，利用句子级相关性奖励和概念级相关性奖励来提高描述的视觉语义相关性。为了实验对比，Song 等人（2019）还设置了 SSR-Baseline & CIDEr Reward 模型，引入朴素的 CIDEr 评分奖励替换所提出的句子兼概念语义奖励，进行强化学习训练。

如表 3 所示，与 2-Stage pivot - Google API 相比，2-Stage pivot - Baidu API 性能表现不佳，尤其在所有方法中为最低性能。同样也表明了 Google API 在英语翻译上比 Baidu API 有优势。尽管 2-Stage pivot Google API 的指标得分有所提升，相比之下，本文模型的 BLEU-4 得分和 CIDEr 得分高出了 6.5%和 10.9%。由此表明，与直接使用轴语言描述作监督信息参与模型训练、再进行翻译的两阶段方法相比，本文模型表现更为优越。与 2-Stage pivot - joint model 相比，本文方法在 BLEU-4 评分上提升了 6.4%，在 CIDEr 评分上提上了 10.2%。结果表明，目标语言域评价模块引导模型学习了丰富的目标域语言表达方式，降低了翻译模型中不流畅句子对模型的负面影响。与 SSR-Baseline & CIDEr Reward 实验结果相比，本文模型在所有评价指标上都有明显提升，其中 CIDEr 评分提升 13.3%。这一结果表明，仅使用 CIDEr Reward 强化学习策略，对求解复杂的跨语言图像描述任务还远远不够。与目前最好性能 SSR 方法相比，本文方法在 BLEU-2、BLEU-3、BLEU-4 和 METEOR 等 4 个评价指标上的得分分别提升了 1.4%，1.0%，0.7%和 1.3%。结果表明，相比 SSR 对生成句子使用句子级和概念级语义奖励机

制，本文提出的强调多模态的语义匹配和语言指导模型，更重视图像、轴语言和目标语言的语义一致性约束，从不同模态数据出发向一致性语义表达优化，能学习到更丰富准确的语义知识。

图 3 是本文模型关于跨语言英文图像描述任务在 MSCOCO 测试集的可视化效果，红色字体表示来自 Baseline 模型翻译的错误语义，绿色字体表示来自本文模型翻译的正确语义。图 3 表明，一方面，本文模型生成的描述更贴近图像视觉内容。例如，本文模型可以识别出物体属性：将错误的人物对象“woman”替换为“boy”；可以推理对象关系：一个男人“sitting on the green grass”纠正为“sitting on a horse in the grass”。另一方面，本文模型生成的句子与目标语言风格差异更小。例如，本文模型生成的句子更偏向目标语言风格的“某人在某地做某事”句式：“a man is skiing in the snow in the mountains.”（一个男人在山里的雪地上滑雪），而 Baseline 模型倾向给对象添加定语修饰：“a man with a ski pole in both hands was skiing in the snow”（一个双手拿着滑雪杖的人在滑雪）。

3.3.3 跨语言中文图像描述主性能分析

表 4 展现了不同方法关于跨语言中文图像描述任务在 AIC-ICC 测试集上的评分效果。本文与 4 项跨语言中文图像描述实验进行了对比：1）Baseline 方法；2）SSR-Baseline & CIDEr Reward（Song 等，2019）方法；3）2-Stage pivot（Gu 等，2018） - Baidu API 方法；4）SSR（Song 等，2019）方法。这些方法与 3.3.2 节中所提相同。

如表 4 所示，2-Stage pivot - Baidu model 在所有的方法中取得了最低性能。相比于 2-Stage pivot - Baidu API 方法，本文模型的 BLEU-4 和 CIDEr 得分分别高出 7.3%和 10.6%。这表明，针对跨语言中文图像描述任务，与两阶段的图像-轴语言-目标语言的方法相比，本文模型更具优越性。与 Baseline 方法相比，本文模型在所有指标上都取得了明显提升，其中 BLEU-4 和 CIDEr 分别提升了 2.9%和 4.9%。与 SSR-Baseline & CIDEr Reward 方法相比，本文模型在 BLEU-4 和 CIDEr 得分分别提升了 6.3%和 10.1%。与性能最好的 SSR 方法相比，本文方法在 BLEU-1、BLEU-2、BLEU-3、BLEU-4、METEOR 和 CIDEr 等六个评价指标上的评分上分别提升了 5.7%，2.0%，1.6%，1.3%，1.2%和 3.4%。以上结果表明，跨语言中文图像描述任务中，在语义匹配模块和语言评价模块的共同作用下，也同样生成更加语义完整和流畅的句子。

图 4 是本文模型关于跨语言中文图像描述任务在 AIC-ICC 测试集的可视化效果，红色字体表示来自 Baseline 模型翻译的错误语义，绿色字体表示来自本文模型翻译的正确语义。从图 4 可见，一方面，本文模型生成的描述与真实描述语义更相关。例如，

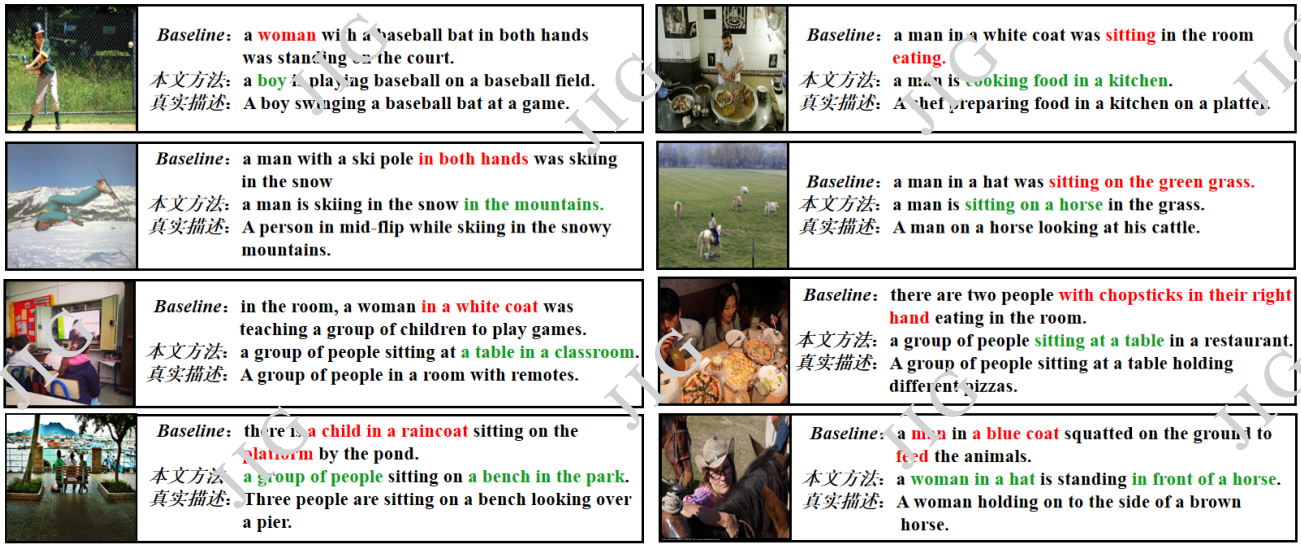

图 3 跨语言英文图像描述在 MSCOCO 测试集的样例

Fig.3 Examples of the cross-lingual English image captioning from the MSCOCO testing set

图 4 跨语言中文图像描述在 AIC-ICC 测试集的样例

Fig.4 Examples of the cross-lingual Chinese image captioning from the AIC-ICC testing set

本文模型可以对缺少的、有误的视觉信息进行补充和替换：Baseline 模型生成的句子“一个穿着西装打着领带的男人站在一起”只检测出一个人物对象且句子不流畅，本文模型的句子“一个穿着西装的男人站在一个穿着裙子的女人旁边”，关注了更丰富的语义信息且句子更加流畅；将错误的视觉信息“手里拿着冲浪板”修正为“站在海滩上的岩石上”。另一方面，本文模型生成的句子与真实描述语言风格更相近。例如本文模型生成的句子更偏向真实描述的“连续且简短的”描述风格，符合目标语料的风格“一个女人在外面的桌子旁吃东西”，而 Baseline 模型更倾向于生成“逗号分隔的”复杂句式“一个女人坐在一张桌子旁，手里拿着一部手机。”

4 结　论

针对现有的跨语言图像描述方法在缺乏成对图像-句子数据集下生成的目标语言描述与图像语义关联弱、与真实目标语言风格差异明显等问题，本文提出了一种引入语义匹配和语言评价的跨语言图像描述模型。在以编码器-解码器为基准架构的模型上，本文设计了图像&轴语言语义匹配模块，通过对目标语言、源域图像、轴语言句子进行语义匹配计算来约束描述的语义相关性。同时本文设计了目标语言评价模块，通过学习目标语料集中的语言表达来优化描述的语言质量。在语义匹配奖励和语言评价奖励的指导下，模型生成语义更准确和语言更流畅的描述。

本文在 MSCOCO 和 AIC-ICC 两个数据集上与其他现有方法分别进行了跨语言英文图像描述和跨语言中文图像描述测试和比较。定量对比结果表明，本文模型在多个测评指标上达到最好，生成的描述与真实的目标语言描述更加接近，具有较好的鲁棒性和有效性。定性对比结果表明，本文模型提升了描述与图像的语义一致性。同时消融实验结果表明，本文提出的语义匹配奖励、语言评价奖励对模型都产生了积极作用。

由于本文模型对图像细节的关注较弱，生成的描述在精度上仍有不足。因此，在后续工作中将考虑引入注意力机制，探索更加细粒度的跨语言图像描述。

参考文献(References)

Anderson P, He X D, Buehler C, Teney D, Johnson M, Gould S and Zhang L. 2018. Bottom-up and top-down attention for image captioning and visual question answering//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City, UT, USA: IEEE: 6077-6086 [DOI: 10.1109/CVPR.2018.00636]

Ben H X, Pan Y W, Li Y H, Yao T, Hong R C, Wang M and Mei T. 2021. Unpaired Image Captioning with Semantic-Constrained Self-Learning. IEEE Transactions on Multimedia, [DOI: 10.1109/TMM.2021.3060948]

Denkowski M and Lavie A. 2014. Meteor universal: Language

specific translation evaluation for any target language//Proceedings of the ninth workshop on statistical machine translation. Baltimore, Maryland, USA: Association for Computational Linguistics: 376-380 [DOI: 10.3115/v1/W14-3348]

Farhadi A, Hejrati M, Sadeghi M A, Young P, Rashtchian C, Hockenmaier J and Forsyth D. 2010. Every picture tells a story: Generating sentences from images///Proceedings of the 11th European Conference on Computer Vision. Heraklion, Crete, Greece: Springer: 15-29 [DOI: 10.1007/978-3-642-15561-1_2]

Feng Y, Ma L, Liu W and Luo J B. 2019. Unsupervised image captioning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach，CA，USA: IEEE: 4125-4134 [DOI: 10.1109/CVPR.2019.00425]

Gu J X, Joty S, Cai J F and Wang G. 2018. Unpaired image captioning by language pivoting//Proceedings of the European Conference on Computer Vision. Munich, Germany: Springer: 503-519 [DOI: 10.1007/978-3-030-01246-5_31]

He K M, Zhang X Y, Ren S Q, and Sun J. 2016. Deep residual learning for image recognition//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas, NV, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Hou J Y, Wu X X, Zhang X X, Qi Y Y, Jia Y D and Luo J B. 2020. Joint Commonsense and Relation Reasoning for Image and Video Captioning//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA: AAAI: 10973-10980 [DOI: 10.1609/aaai.v34i07.6731]

Ji J Z, Xu C, Zhang X D, Wang B Y and Song X H. 2020. Spatio-temporal memory attention for image captioning. IEEE Transactions on Image Processing, 29: 7615-7628 [DOI: 10.1109/TIP.2020.3004729]

Lan W Y, Li X R and Dong J F. 2017. Fluency-guided cross-lingual image captioning//Proceedings of the 25th ACM international conference on Multimedia. Mountain View, CA USA: Association for Computing Machinery: 1549-1557 [DOI: 10.1145/3123266.3123366]

Li Z X, Wei H Y, Huang F C, Zhang C L, Ma H F and Shi Z Z. 2020. Combine visual features and scene semantics for image captioning. Chinese Journal of Computers, 43(09):1624-1640. (李志欣，魏海洋，黄飞成，张灿龙，马慧芳，史忠植. 2020. 结合视觉特征和场景语义的图像描述生成. 计算机学报，43(09)：1624-1640.) [DOI: 10.11897/SP.J.1016.2020.01624]

Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick C L. 2014. Microsoft coco: Common objects in context//Proceedings of the European conference on computer vision. Zurich, Switzerland: Springer: 740-755 [DOI: 10.1007/978-3-319-10602-1_48]

Liu X H, Li H S, Shao J, Chen D P and Wang X G. 2018. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data//Proceedings of the European Conference on Computer Vision. Munich, Germany: Springer: 338-354 [DOI: 10.1007/978-3-030-01267-0_21]

Luo H L and Yue L L. 2020. Image caption based on causal convolutional decoding with cross-layer multi-model feature fusion. Journal of Image and Graphics，25(08) : 1604-1617(罗会兰，岳亮亮. 2020. 跨层多模型特征融合与因果卷积解码的图像描述. 中国图象图形学报，25(08): 1604-1617) [DOI: 10. 11834 /jig.190543]

Ranzato M A, Chopra S, Auli M, andZaremba W. 2015. Sequence level training with recurrent neural networks[EB/OL]. [2021-06-11].https://arxiv.org/pdf/1511.06732.pdf.

Ren S Q, He K M, Girshick R and Sun J. 2016. Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence , 39(6): 1137-1149 [DOI: 10.1109/TPAMI.2016.2577031]

Rennie S J, Marcheret E, Mroueh Y and Goel V. 2017. Self-critical sequence training for image captioning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE: 7008-7024 [DOI: 10.1109/CVPR.2017.131]

Song Y Q, Chen S Z, Zhao Y D and Jin Q. 2019. Unpaired cross-lingual image caption generation with self-supervised rewards//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France: Association for Computing Machinery: 784-792 [DOI: 10.1145/3343031.3350996]

Tang P J, Tan Y L and Li J Z. 2017. Image description based on the fusion of scene and object category prior knowledge. Journal of Image and Graphics，22(9) : 1251-1260. (汤鹏杰, 谭云兰, 李金忠. 2017. 融合图像场景及物体先验知识的图像描述生成模型. 中国图象图形学报, 22(9) : 1251-1260.) [DOI: 10. 11834 /jig. 170052]

Vinyals O, Toshev A, Bengio S and Erhan D. 2015. Show and tell: A neural image caption generator//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE: 3156-3164 [DOI: 10.1109/CVPR.2015.7298935]

Wang W X, Chen Z H and Hu H F. 2019. Hierarchical attention network for image captioning//Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii, USA: AAAI: 8957-8964 [DOI: 10.1609/AAAI.V33I01.33018957]

Wu J H, Zheng H, Zhao B, Li Y X, Yan B M, Liang R, Wang W J,

Zhou S P, Lin G, Fu Y W, Wang Y Z and Wang Y G. 2017. Ai challenger: A large-scale dataset for going deeper in image understanding [EB/OL]. [2021-05-18].https://arxiv.org/pdf/1711.06475.pdf.

Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R and Bengio Y. 2015. Show, attend and tell: Neural image caption generation with visual attention//Proceedings of the 32nd International Conference on Machine Learning. Lille, France: PMLR: 2048-2057

Zhou L W, Palangi H, Zhang L, Hu H D, Corso J and Gao J F. 2020, Unified vision-language pre-training for image captioning and vqa//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA: AAAI: 13041-13049 [DOI: 10.1609/aaai.v34i07.7005]

作者简介

张静，1996 年生，女，硕士研究生，主要研究方向为图像描述生成、深度学习。
E-mail: hfutzhangjing@gmail.com

郭丹，通信作者，女，教授，计算机学会(CCF)会员，主要研究方向为视频分析、模式识别、深度学习。
E-mail: guodan@hfut.edu.cn

宋培培，通信作者，女，博士研究生，主要研究方向为计算机视觉、自然语言处理、深度学习。
E-mail: beta.songpp@gmail.com

李坤，男，博士研究生，主要研究方向为视频理解、时序动作定位。
E-mail: kunli.hfut@gmail.com

汪萌，男，博士，教授，计算机学会(CCF)会员，IEEE Fellow，主要研究方向为模式识别、数据挖掘、多媒体信息处理。
E-mail: eric.mengwang@gmail.com