北京大学学报(自然科学版) 第 57 卷 第 1 期 2021 年 1 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 57, No. 1 (Jan. 2021)
doi: 10.13209/j.0479-8023.2020.080

融合物体空间关系机制的图像摘要生成方法

万璋 张玉洁† 刘明童 徐金安 陈钰枫

北京交通大学计算机与信息技术学院, 北京 100044; † 通信作者, E-mail: yjzhang@bjtu.edu.cn

摘要 聚焦于图像中物体间位置关系这一特定信息, 提出一种融合空间关系机制的神经网络图像摘要生成模型, 以期为视觉问答和语音导航等下游任务提供物体方位或轨迹等关键信息。为了增强图像编码器的物体间位置关系学习能力, 通过改进 Transformer 结构来引入几何注意力机制, 显式地将物体间位置关系融合进物体外观信息中。为了辅助完成面向特定信息的抽取和摘要生成任务, 进一步提出相对位置关系的数据制作方法, 并基于 SpatialSense 数据集制作物体间位置关系的图像摘要数据集 Re-Position。与 5 个典型模型的对比测评实验结果表明, 所提模型的 5 个指标在公开测试集 COCO 上优于其他模型, 全部 6 个指标在本文制作的 Re-Position 数据集上优于其他模型。

关键词 图像摘要; 物体间位置关系; 注意力机制; Transformer 结构

Object Space Relation Mechanism Fused Image Caption Method

WAN Zhang, ZHANG Yujie†, LIU Mingtong, XU Jin'an, CHEN Yufeng

School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044;
† Corresponding author, E-mail: yjzhang@bjtu.edu.cn

Abstract Focusing on the specific information of the positional relationship between objects in the image, a neural network image summary generation model integrating spatial relationship mechanism is proposed, in order to provide key information (object position or trajectory) for downstream tasks such as visual question answering and voice navigation. In order to enhance the learning ability of the positional relationship between objects of the image encoder, the geometric attention mechanism is introduced by improving the Transformer structure, and the positional relationship between objects is explicitly integrated into the appearance information of the objects. In order to assist in the completion of specific information-oriented extraction and summary generation tasks, a data production method for relative position relations is further proposed, and the image abstract data set Re-Position of the position relations between objects is produced based on the SpatialSense data set. The experimental results of comparative evaluation with five typical models show that the five indicators of the proposed model are better than those of other models on the public test set COCO, and all six indicators are better than those of other models on Re-Position data set.

Key words image caption; positional relationship between objects; attention mechanism; Transformer structure

自然语言处理(natural language processing, NLP)和计算机视觉(computer vision, CV)是当前人工智能领域的研究热点。近期, 融合文本和图像信息的多模态信息处理问题引起研究者的极大兴趣。作为多模态信息处理的一项关键技术, 图像的摘要生成(image caption)最早由 Farhadi 等[1]提出, 给定二元组(I, S)(I 表示图像, S 表示摘要句子), 模型完成从图像到摘要句子(I→S)的多模态映射。最近, 研究人员注意到图像中一些特定的细粒度信息(如颜色和位置)能够为下游任务(如图片检索)提供重要的依据, 因此从图像中抽取特定信息生成摘要的需求日益增大。

国家自然科学基金(61876198, 61976015, 61976016)资助
收稿日期: 2020–06–09; 修回日期: 2020–08–14

在摘要生成中，图像信息通常用一句话表达，仅仅是对图像中某一部分信息的描述。现有数据集中，图像摘要的人工标注对具体对象和描述要素没有统一的规范，标注人员的关注点随意，未必包含特定信息。如此，面向特定信息抽取的摘要生成研究面临困境。

本文关注图像中物体间位置关系这一特定信息在文本摘要里的准确表达。物体间位置关系信息对理解图像内容至关重要，人类在对物理世界进行推理时也要使用这些信息。例如，相对位置信息的提取能够帮助生成“卧室内人坐在椅子上”，而不仅仅是“卧室内有人和椅子”。

为了增强图像编码器对物体间位置关系的学习能力，本文首次提出一种融合空间关系机制的神经网络图像摘要生成模型。我们对物体间的位置关系进行单独编码，获取位置关系的显式表示，并在Transformer结构中引入几何注意力机制，将位置关系融合进物体外观信息中。为了辅助完成面向特定信息的抽取和摘要生成任务，我们提出物体间位置关系数据制作方法，并基于SpatialSense数据集[2]制作位置关系数据集Re-Position。最后，在公开测试集COCO和本文制作的数据集Re-Position上进行验证，并与其他5个典型的模型进行对比。

1 相关研究

早期的基于神经网络模型[3−5]没有进行物体检测处理，图像编码器直接对整幅图像进行编码，因此无从获取物体间的位置关系信息。后来的研究中增加基于CNN的物体检测处理，检测出物体并提取相应的特征[6]，为每个物体生成单独的摘要，但图像编码器未对物体间的关系，尤其是相对位置关系进行建模。Anderson等[7]利用“自下而上”与“自上而下”(Up-Down模型)的注意力机制，对多个物体的特征向量进行编码，在图像摘要生成任务中取得最佳性能，但没有对物体间相对位置关系进行显示编码。Yao等[8]在图像编码器中对物体间位置设置11种关系，如“内部”、“覆盖”或“重叠”，采用图卷积网络构建物体间位置关系图，以边的类别表示位置关系类别，但其设置的关系类别数量有限，不能覆盖未知数据集中众多种类的物体间位置关系。之后，Yang等[9]利用知识图谱扩展物体间位置关系类别的数量，但仍无法处理知识图谱中不存在的关系类别。

我们的方法是根据数据集，动态地确定物体间位置关系类别的集合，即在图像编码器中使用Transformer结构来设计几何注意力机制，对物体检测框的大小和差异等特征进行物体间位置关系的显示编码，提高模型对数据集中出现的位置关系类别的覆盖程度，并针对物体间位置关系进行数据制作和评测。

2 融合空间关系机制的图像摘要生成模型

本文围绕位置关系抽取问题，提出融合空间关系机制的图像摘要模型。本文的任务如下：对图像中的n个(由数据集指定或由图像检测结果确定)物体，给出所有物体对之间的空间位置关系描述，最终生成所有物体对间的位置关系描述摘要。在摘要生成评测中使用BLEU等指标，计算生成摘要对参考摘要(包含所有物体对之间的位置关系描述)的覆盖度。

2.1 模型框架

本文提出的图像摘要生成模型由物体检测模块、图像编码器和文字解码器三部分构成，模型框架如图1所示。首先，利用物体检测模块(如Faster R-CNN)检测出图像中的n个物体，得到每个物体的特征向量；然后，利用图像编码器对n个物体的特征向量以及位置间关系信息进行编码，得到融合n个物体的图像表示；最后，文字解码器采用加入Attention机制的Bi-LSTM结构，对图像表示进行序列建模，生成摘要文本。另外，我们在图像编码器中引入几何注意力机制，对物体的空间位置进行单独编码，获得物体间位置关系的表示。

2.2 物体检测

本文使用Faster R-CNN[10]和ResNet-101[11]作为目标检测和特征提取的基础框架。为了得到物体的最佳候选检测框，我们利用非最大抑制算法，将重合程度超过阈值0.7的重叠检测框舍弃，并得到物体的几何特征；然后利用Faster R-CNN结构中的ROI层，将删选后的检测框转换至相同的维度(如14×14×2048)。为了预测每个物体检测框的类别标签，利用ResNet-101网络进行特征提取，得到物体的外观特征。进一步地，舍弃类别预测概率低于阈值0.2的物体检测框，以便得到物体的确定数量n($\leq$4)。最后，为每个物体生成包括几何特征(物体的

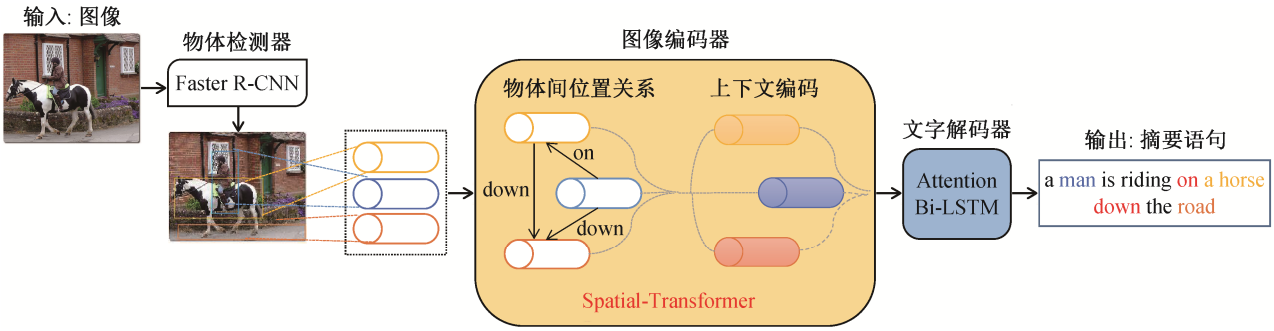


图 1　图像摘要生成模型基本框架

Fig. 1　Basic framework of image caption model

位置和大小)和外观特征(物体的类别)在内的特征向量(2048 维), 输入图像编码器中。

2.3 图像编码器

利用图像编码器, 对 n 个物体的特征向量进行编码, 得到图像表示。物体检测模块为每个物体生成一个特征向量, 向量信息之间没有联系。但是, 作为一幅图像中的物体, 相互之间存在一定的关系, 例如两个物体“房间”和“人”之间的关系为“房间里有人”, 因此图像编码器需要将物体之间的关系编码到图像表示中。为了表示物体之间的相互关系, 需要获取其他物体的信息, 可以通过计算物体间特征向量的相关性来实现, 并把这种相关性表示融合成物体的语义表示。

本文采用 Tansformer 结构[12]的编码部分作为图像编码器, 输入为 n 个特征向量, 对应 n 个物体。图像编码器的第一层有多个 Relation 模块, 每个模块输入一个物体的特征向量, 通过学习与其他物体之间的关系来更新物体的语义表示。图像编码器由多个编码层构成, 将前一个编码层的输出作为后一个编码层的输入, 将最后一个编码层的输出作为图像表示, 馈送到文字解码器生成摘要。

每个 Relation 模块负责获得相应物体与图像中其他所有物体之间的关系, 并更新该物体的语义表示, 由 Self-attention 机制来实现。对于物体 A_i ($1\leq i\leq n$), 首先根据式(1), 从其特征向量得到 quries (Q), keys (K)和 values (V):

$$\boldsymbol{Q}=\boldsymbol{X}\boldsymbol{W}_Q,\boldsymbol{K}=\boldsymbol{X}\boldsymbol{W}_K,\boldsymbol{V}=\boldsymbol{X}\boldsymbol{W}_V, \tag{1}$$

其中, $\boldsymbol{X}$表示图像中 n 个物体的特征向量矩阵; $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$ 和 $\boldsymbol{W}_V$ 是权重矩阵, 起到变化维度的作用, 可以通过模型训练得到。n 个物体的语义表示矩阵通过下式计算得到:

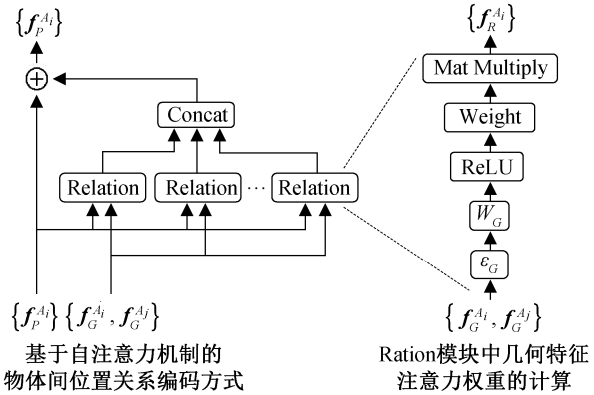
$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}。 \tag{2}$$

通过注意力机制得到的矩阵中, 每个向量对应一个物体, 代表融合了与其他物体关系的语义表示。Transformer 结构采用多头注意力机制, 我们通过拼接多头注意力机制计算得到的结果, 获得最终的语义表示:

$$\begin{aligned}&\text{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})\\&=\text{Concat}(\text{head}_1,\text{head}_2,...,\text{head}_h)\boldsymbol{W}_O。\end{aligned} \tag{3}$$

2.4 面向空间位置关系的图像编码器

由于图像中物体的类别、尺寸和位置不同, 导致难以对空间位置关系进行建模。鉴于物体的空间位置关系由各个物体的空间位置决定, 我们考虑充分利用物体特征向量中的几何特征来帮助获取空间位置关系, 为此提出基于几何注意力机制的图像编码方式, 如图 2 所示。



$\boldsymbol{f}_G^{A_i}$ 表示物体 A_i 的几何特征, $\boldsymbol{f}_P^{A_i}$ 表示物体 A_i 的外观特征, $\boldsymbol{f}_G^{A_j}$ 表示物体 A_j 的几何特征, $\boldsymbol{f}_P^{A_j}$ 表表示物体 A_j 的外观特征, $\boldsymbol{f}_R^{A_i}$ 表示物体 A_i 与其他所有物体间的位置关系信息

图 2　基于几何注意力机制的物体间位置关系编码方法

Fig. 2　Coding method of position relationship between objects based on geometric attention mechanism

我们利用Self-attention机制设计空间位置关系编码方法。具体地，在2.3节描述的图像编码器基础上增加基于几何注意力机制的编码部分。式(2)只考虑物体间的关系，利用物体的特征向量获得物体间的注意力权重。为了增强对物体空间位置关系的学习，我们对物体 A_i 的特征向量中的几何特征 $\boldsymbol{f}_G^{A_i}$ 专门设计注意力机制，获取物体间(如 $\boldsymbol{f}_G^{A_j}$ 与 $\boldsymbol{f}_G^{A_i}$)空间位置的注意力权重，然后与式(2)中物体间关系的注意力权重组合作为物体间总的注意力权重，最后以这些权重融合其他物体的信息($\boldsymbol{f}_P^{A_i}$ 和 $\boldsymbol{f}_P^{A_j}$)为每个物体的语义表示 $\boldsymbol{f}_R^{A_i}$。

对于物体 A_i，计算其几何特征与其他物体(如 A_j)几何特征的注意力权重，获取与其他物体的空间位置关系，并融合物体 A_i 的外观特征表示，作为其最终语义表示进行输出。其中，两个物体 A_i 和 A_j 特征向量中的几何特征可表示为 $(x_{A_i}, y_{A_i}, w_{A_i}, h_{A_i})$ 和 $(x_{A_j}, y_{A_j}, w_{A_j}, h_{A_j})$ (x 和 y 表示物体的中心坐标，w 和 h 表示物体的宽度和高度)。图像中不同物体间距离的变化范围很大，容易导致训练结果发散，因此对物体的几何特征按照式(4)进行变换操作:

$$\lambda(A_i, A_j) = \left\{ \log\left(\frac{|x_{A_i} - x_{A_j}|}{w_{A_j}} \right), \log\left(\frac{|y_{A_i} - y_{A_j}|}{h_{A_i}} \right), \log\left(\frac{w_{A_i}}{w_{A_j}} \right), \log\left(\frac{h_{A_j}}{h_{A_i}} \right) \right\}。 \tag{4}$$

为了计算给定物体 A_i 与 A_j 间的位置关系，我们设计式(5)来计算几何特征注意力权重:

$$w_G^{A_iA_j} = \max\{0, \boldsymbol{W}_G \cdot \varepsilon_G(\boldsymbol{f}_G^{A_i}, \boldsymbol{f}_G^{A_j})\}, \tag{5}$$

其中，ε_G 是由余弦函数和正弦函数构成的升维函数，给 $\boldsymbol{f}_G^{A_i}$ 和 $\boldsymbol{f}_G^{A_j}$ 两个向量提升维度; W_G 是模型可以学习的参数，可将升维后的高维特征映射到表示两个物体之间位置关系密切程度的得分，分值越大表示关系越密切。

图像编码器有多个Relation模块，每个Relation模块都将物体的几何特征作为输入，采用下式计算当前物体 A_i 与另一物体 A_j 间的位置关系:

$$\boldsymbol{f}_R^{A_i} = w_G^{A_iA_j} \cdot \boldsymbol{V}, \tag{6}$$

这里的 V 仅表示物体的外观特征，含义与2.3节不同。最后，我们融合多个Relation模块得到 $\boldsymbol{f}_R^{A_i}$，并与当前模块的外观特征融合，作为当前物体的语义表示，计算公式如下:

$$\boldsymbol{f}_P^{A_i} = \boldsymbol{f}_P^{A_i} + \text{Concat}[\boldsymbol{f}_R^1(n), ..., \boldsymbol{f}_R^{(N_r)}(n)], \tag{7}$$

其中，Concat表示对所有向量进行拼接操作。

3 实验设计与结果分析

3.1 面向位置关系摘要的数据集制作方法

目前，没有专门面向物体间位置关系的数据集。如图3所示，现有的数据集中，或者只有一个物体，或者摘要没有关注物体间的位置关系。为此，我们设计利用现有数据制作物体间位置关系数据集的方法，分为如下4个步骤。

1) 设计物体间位置关系的标签集合，包含in, on和left等共21个词语，如表1所示。

2) 人工判断并选取现有数据集中只包含两个物体的检测框，且两个物体之间有明确位置关系的图片。

3) 利用数据集中物体的位置坐标(x, y, w, h)呈现的检测框区域(图4)，人工判断物体间位置关系，并使用步骤1的标签进行标注。

4) 利用数据集给定的物体名称以及步骤3得到的位置关系标注，人工制作摘要，并按照COCO数据集的摘要格式存储。

我们利用上述数据制作方法，在SpatialSense数据集[2]上得到物体间空间位置关系的图像摘要数据集Re-Position。图4为本文制作的物体间位置关系的图像摘要示例，每张图片包含两个物体的检测框以及它们之间的位置关系描述。模型直接将图像和物体的位置坐标共同作为输入，可以避免因图像检测中物体识别错误导致的摘要生成错误，使模型评测实验重点关注图像编码和文字解码部分。Re-Position数据集共有1000张图片，每张图片对应1条摘要。本文将该数据集分割为训练集、开发集和测试集，分别为600张、200张和200张图片。

3.2 公开评测数据集

我们同时利用广泛使用的公开数据集Microsoft COCO (MS-COCO) Captions进行评测，共有123287张图片，每张图片有5条摘要。本文设置与文献[7,13]相同的训练集、开发集和测试集，分别为113287张、5000张和5000张图片，并将数据集中的摘要部分转换为小写。

3.3 实验参数设置

模型训练中采用softmax交叉熵作为损失函数，

an airplane is flying　　a person threw out the frisbee

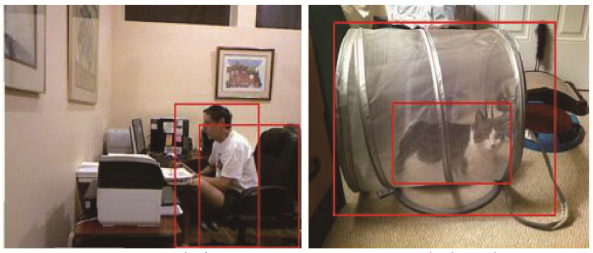
图 3　现有数据集示例

Fig. 3　Examples of existing data sets

表 1　Re-Position 数据集中表示空间位置关系词语的分布概率

Table 1　Distribution probability statistics of words representing spatial position relations on Re-Position datasets

词语	分布概率/%
on	16
left	11
right	12
beside	12
in front of	7
next to	10
close to	4
其他	2



a man on chair　　a cat in laundry

图 4　Re-Position 数据集示例

Fig. 4　Examples of Re-Position data sets

将学习率设为 0.003，warmup 设为 20000，文字解码 Bi-LSTM 设为 500 维，词向量维度设为 500，在层之间采用 dropout 正则化技术，drop 率设为 0.3。采用 Adam 优化算法[14]进行优化，训练轮数为 30，Batch 大小设为 64。将出现次数少于 8 的单词丢弃，最终得到 10118 个词汇。

3.4　评测结果

本研究使用的摘要评测指标包括 CIDEr-D[15]，BLEU-N[16]，METEOR[17]，SPICE[18]和 ROUGE-L。在 Re-Position 数据集上，对比评测基于编码解码框架的 5 种代表性图像摘要生成模型：1) Show and Tell 模型[3]，采用标准 CNN 结构的图像编码和 RNN 结构的文字解码；2) SCST 模型[19]，在编码器与解码器之间引入改进的视觉注意机制生成摘要，还设计一种自临界序列训练策略来训练采用句子级奖励损失函数的 LSTM 结构；3) ADP-ATT 模型[20]，采用标准 CNN 结构的图像编码和 LSTM 结构的文字解码，在编码器与解码器之间使用注意力机制；4) LSTM-A 模型[21]，在解码部分结合外部知识(如语义属性信息)生成摘要；5) Up-Down 模型[7]，使用自下而上和自上而下的注意力机制。评测结果(表 2)显示，与其他 5 个模型相比，本文模型 6 个评测指标的得分均显著提升。其中，Up-Down 模型同样将物体检测坐标作为图像编码器的输入，与其相比，本文模型的 BLEU-1，BLEU-4，CIDEr-D，METEO，SPICE 以及 ROUGE-L 分别提高 3.6%，1.9%，2.3%，0.5%，0.9%和 1.6%。

在 Re-Position 数据集上的实验结果(表 2)表明，本文模型引入的几何注意力机制可以增强对物体间位置信息的表示能力，对物体间的位置关系进行有效的编码，最终在解码时能够准确地生成关于物体

表 2　Re-Position 数据集上不同模型的对比评测结果

Table 2　Experimental results of different models on Re-Position data sets

模型	评测指标得分					
	BLEU-1/%	BLEU-4/%	CIDEr-D	METEOR/%	SPICE/%	ROUGE-L/%
Show and Tell	65.5	30.7	105.8	28.3	–	53.8
SCST	66.2	31.8	106.5	28.7	–	55.2
ADP-ATT	72.5	30.0	108.4	29.2	–	–
LSTM-A	73.9	32.5	109.3	29.5	20.6	56.7
Up-Down	75.3	33.2	111.2	30.1	21.4	56.8
本文模型	78.9 (3.6↑)	35.1 (1.9↑)	113.5 (2.3↑)	30.6 (0.5↑)	22.3 (0.9↑)	58.4 (1.6↑)

说明：括号内数字表示本文模型与 Up-Down 模型评测得分相差的百分点，↑代表提升，↓代表下降，下同。

间位置关系的摘要。

在COCO数据集上，与同样5个代表性模型进行对比评测。由于COCO数据集没有物体的位置坐标，Up-Down模型和本文模型需要进行图像检测，因此均采用Faster-RCNN作为图像检测器的基本框架。评测结果(表3)显示，与前4个模型相比，本文模型的6个评测指标均提升。与Up-Down模型相比，BLEU-1, BLEU-4, CIDEr-D, SPICE和ROUGE-L分别提高0.3%, 0.5%, 2.9%, 0.7%和1.7%；本文模型的METEOR得分略低于Up-Down模型(降低0.2%)。在COCO数据集上的实验结果表明，本文模型在公开数据集上同样超过现有代表性模型的性能，从而验证了本文模型的有效性。

3.5 消融实验

本文模型性能的提升是完全来自Transformer的优势，还是与几何注意力机制的引入相关，需要通过消融实验来验证。我们为此构建两个模型：一个是在Up-Down模型(同样使用目标检测)中加入几何注意力机制，记为Up-Down+Geom_Attn；另一个是在本文模型中去掉几何注意力机制，记为Transf +Bi-LSTM。然后，分别进行评测，并与本文模型进行对比。对比评测结果(表4)显示，本文模型去除几何注意力机制后性能下降，METEOR，CIDEr-D，BLEU-1，BLEU-4，SPICE以及ROUGR-L分别下降0.6%，5.7%，0.4%，0.4%和0.8%，说明3.4节的评测结果中，本文模型性能的提升的确有来自几何注意力机制的贡献，并非完全来自Tranformer结构的使用。另一方面，将几何注意力机制引入Up-Down模型后，也会带来性能的提升(METEOR，CIDEr-D，BLEU-1，BLEU-4，SPICE以及ROUGR-L分别提高0.2%，2.2%，0.1%，0.3%和0.4%)，进一步说明本文提出的几何注意力机制可以提升模型性能。消融实验结果表明，本文提出的几何注意力机制可以显著地提升物体间位置关系的表示能力，从而提升摘要生成的质量。

3.6 实例分析

为了进一步分析本文模型的性能，我们选择CIDEr-D得分有明显提升的摘要实例与Up-Down模型进行对比，结果如图5所示。图5(a1)中，Up-Down模型错误地生成“人在椅子的前面”，本文模型正确地生成“人在椅子上”；图5(a2)中，Up-Down模型错误地生成“孩子站在水里”，本文模型正确地生成“孩子在水面上”；图5(b1)中，本文模型正确地生成杯子、电脑和桌子的三者关系；图 5(b2)中，本文模型正确地给出孩子的数量，说明本文模型中的物体检测器能够正确地识别出两个孩子。这一实

表 3 COCO 数据集上不同模型的对比评测结果

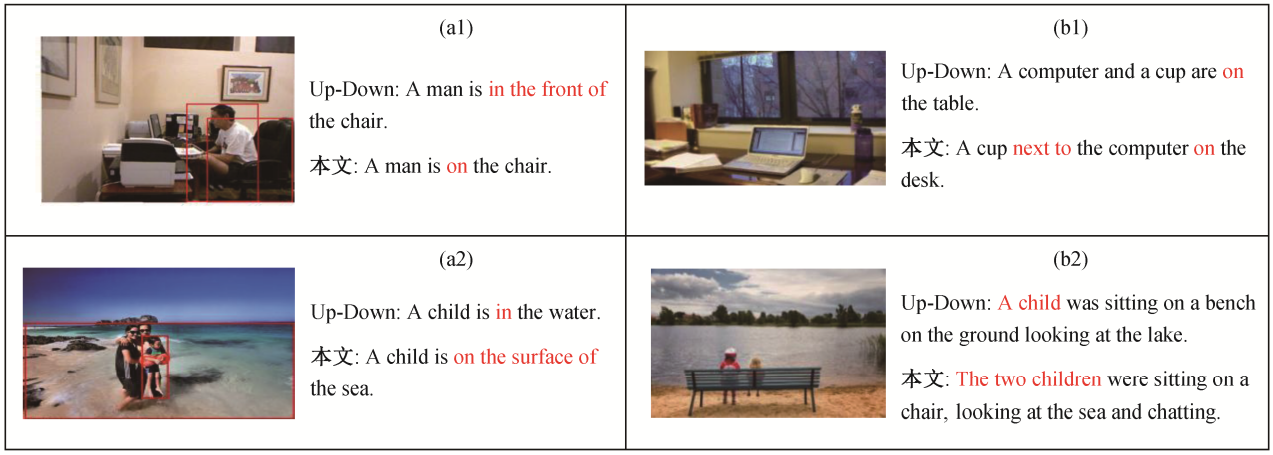
Table 3 Experiment results of different models on COCO dataset

模型	评测指标得分					
	BLEU-1/%	BLEU-4/%	CIDEr-D	METEOR/%	SPICE/%	ROUGE-L/%
Show and Tell	–	31.5	106.3	24.6	–	52.3
SCST	–	34.2	108.9	24.7	–	53.6
ADP-ATT	–	–	–	–	–	–
LSTM-A	76.5	35.6	112.7	26.1	20.6	55.8
Up-Down	79.8	37.2	113.5	27.5	20.8	56.3
本文模型	80.1 (0.3↑)	37.7 (0.5↑)	117.4 (2.9↑)	27.3 (0.2↓)	21.5 (0.7↑)	58.0 (1.7↑)

表 4 消融实验结果

Table 4 Results of ablation experiments

模型	评测指标得分					
	BLEU-1/%	BLEU-4/%	CIDEr-D	METEOR/%	SPICE/%	ROUGR-L/%
Up-Down	79.8	37.2	113.5	27.5	20.8	56.3
+Gemo_Attn	79.9	37.5	115.7	27.7	21.1	56.7
Transf+Bi-LSTM	79.7	37.3	111.7	26.7	19.4	57.6
+Gemo_Attn (本文)	80.1	37.7	117.4	27.3	21.5	58..4



(a1)和(a2)在 Re-Position 数据集上生成的实例；(b1)和(b2)在 COCO 数据集上生成的实例。
红字为两个模型生成的摘要中表示物体间位置关系的词语

图 5　在 Re-Position 和 COCO 数据集上本文模型与 Up-Down 模型的生成实例对比

Fig. 5　Comparison of generation examples of this model and Up-Down model on the Re-Position and COCO datasets

例分析结果表明，本文引入几何注意机制对物体检测精度的提升也有帮助，这一发现与 Hu 等[22]的结论一致。COCO 数据集上的对比实例显示，本文模型在包含两个以上物体的图片摘要生成中获得质量更好的结果。

4　结语

本文围绕物体间位置关系特定信息抽取这一问题，提出利用几何注意力机制对物体间位置关系进行编码，获取物体间位置关系的显式表示，从而增强模型对物体间位置关系的学习能力。实验结果显示，本文模型在位置关系显示编码上的有效性可以帮助提升摘要中相关描述生成的准确性。为了辅助完成面向特定信息的抽取和摘要生成任务，我们提出物体间位置关系数据制作方法，并基于 Spatial-Sense 数据集[2]，制作物体间位置关系的图像摘要数据集 Re-Position。在 MS-COCO 数据集上的测评结果表明，本文模型的摘要生成能从物体间位置关系信息中受益，提高摘要生成的质量。在 Re-Position 数据集上的测评结果表明，本文模型对物体间位置关系信息的表示能力显著增强。定性的实例分析结构说明，引入几何注意机制能产生更好的表示物体位置关系的图像摘要。

目前，本文模型仅在编码阶段考虑了物体间位置关系的信息。今后的工作中，我们拟在解码器的交叉注意层中也融入几何注意力机制，进一步提升模型的性能。

参考文献

[1] Farhadi A, Hejrati M, Sadeghi A, et al. Every picture tells a story: generating sentences from images // Proceeding of Part IV of the 11th European Conference on Computer Vision. Heraklion, 2010: 15–29

[2] Yang K, Russakovsky O, Deng J, et al. Spatial sense: an adversarially crowdsourced benchmark for spatial relation recognition // 2019 IEEE International Conference on Computer Vision. Seoul, 2019: 2051–2060

[3] Vinyals O, Toshev A, Bengio S, et al. Show and tell: a neural image caption generator // 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, 2015: 3156–3164

[4] Kiros R, Salakhutdinov R, Zemell R. Multimodal neural language models // International Conference on Machine Learning. Beijing, 2014: 595–603

[5] Mao J, Xu W, Yang J, et al. Deep captioning with multimodal recurrent neural networks (m-RNN). arXiv preprint. 2014, arXiv: 1412.6632

[6] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation // 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, 2014: 580–587

[7] Anderson P, He, X, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering // 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake, 2018: 6077–6086

[8] Yao T, Pan Y, Li Y, et al. Exploring visual relationship for image captioning // 2018 European Conference on Computer Vision. Munich, 2018: 711–727

[9] Yang X, Tang K, Zhang H, et al. Auto-encoding scene graphs for image captioning // 2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, 2019: 10685–10694

[10] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell, 2017, 39(6): 1137–1149

[11] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition // 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 770–778

[12] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need // 2017 Conference and Workshop on Neural Information Processing Systems. Long Beach, 2017: 5998–6008

[13] Rennie S J, Marcheret E, Mroueh Y, et al. Self-critical sequence training for image captioning // 2017 IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, 2017:1179–1195

[14] Kingma D P, Ba J. Adam: a method for stochastic optimization // International Conference on Learning Representations. San Diego, 2015: 1–15

[15] Vedantam R, Lawrence Zitnick, C, Parikh D. Cider: consensus-based image description evaluation // 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, 2015: 4566–4575

[16] Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation // 2002 Annual Meeting of the Association for Computational Linguistic. Philadelphia, 2002: 311–318

[17] Banerjee S, Lavie A. Meteor: an automatic metric for MT evaluation with improved correlation // 2005 Annual Meeting of the Association for Computational Linguistic. Michigan, 2005: 65–72

[18] Anderson P, Fernando B, Johnson M, et al. Spice: semantic propositional image caption evaluation // 2016 European Conference on Computer Vision. Amsterdam, 2016: 382–398

[19] Rennie S J, Marcheret E, Mroueh Y, et al. Self-critical sequence training for image captioning // 2017 IEEE Conference on Computer Vision and Pattern Recognition. Hawai, 2017: 1179–1195

[20] Lu J, Xiong C, Parikh D. Knowing when to look: adaptive attention via avisual sentinel for image captioning // 2017 IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, 2017: 3242–3250

[21] Yao T, Pan Y, Li Y, et al. Boosting image captioning with attributes // 2017 IEEE International Conference on Computer Vision. Venice, 2017: 4904–4912

[22] Hu H, Gu J, Zhang Z, et al. Relation networks for object detection // 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake, 2018: 3588–3597