第 58 卷 第 12 期/2021 年 6 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

融合空间注意力机制的图像语义描述算法

郭列[1*]，张团善[1]，孙威振[2]，郭杰龙[2]

[1]西安工程大学机电工程学院，西安市现代智能纺织装备重点实验室，陕西 西安 710600；

[2]中国科学院海西研究院泉州装备制造研究所，福建 泉州 362216

摘要 图像语义描述模型通常采用编码器-解码器方式实现图像语义描述，模型存在对图像特征利用不充分，图像目标的位置信息提取不足等问题。针对此问题，提出在编码器部分融合注意力机制的图像语义描述算法，通过解码器上下文信息对不同图像特征的注意力权重分配，从而提高图像语义描述的表达能力。并在 Flickr30k 和 MSCOCO 数据集上进行了验证，模型在 BLEU-4 评价指标上分别提升了 1.9%和 0.8%，实验证明了本文算法的有效性。

关键词 图像处理；注意力机制；深度卷积神经网络；长短时记忆网络

中图分类号 TP391 **文献标志码** A **doi**：10.3788/LOP202158.1210030

Image Semantic Description Algorithm with Integrated Spatial Attention Mechanism

Guo Lie[1*], Zhang Tuanshan[1], Sun Weizhen[2], Guo Jielong[2]

[1] *Xi'an Key Laboratory of Modern Intelligent Textile Equipment, College of Mechanical and Electrical Engineering, Xi'an Polytechnic University, Xi'an, Shaanxi* 710600, *China*;

[2] *Quanzhou Institute of Equipment Manufacturing, Haixi Institutes, Chinses Academy of Science, Quanzhou, Fujian* 362216, *China*

Abstract The image semantic description model usually adopts the encoder-decoder method to realize the image semantic description. The model has problems such as insufficient utilization of image features and insufficient location information extraction of image objects. In response to this problem, an image semantic description algorithm is proposed that integrates the attention mechanism in the encoder part, and the attention weight of different image features is allocated through the context information of the decoder, thereby improving the expressive ability of image semantic description. And verified on the Flickr30k and MSCOCO data sets, the model improves the BLEU-4 evaluation index by 1.9% and 0.8%, respectively. The experiment proves the effectiveness of the proposed algorithm.

Key words image processing; attention mechanism; deep convolutional neural network; long-short term memory

OCIS codes 100.4996; 100.3008; 100.2960

1 引 言

图像语义描述是对给定的图像自动生成自然语言描述的一类方法，近年来已经成为计算机视觉的热门研究领域之一。它相当于模仿人类将大量重要的视觉信息压缩成描述性语言的能力，这是人工智

收稿日期：2020-09-04；**修回日期**：2020-09-22；**录用日期**：2020-09-30

基金项目：国家自然科学基金青年基金(61806186)、机器人与系统国家重点实验室(SKLRS-2019-KF-15)、“福建省智能物流产业技术研究院建设”项目(2018H2001)、泉州市科技计划项目(2019C112，2019STS08)

* **E-mail**：lie_guo@163.com

能领域的一项重要挑战。图像语义模型不仅需要强大的图像识别技术来确定图像中的前景和背景，同时还需要建立精确的自然语言模型来生成准确的描述。由于模型复杂度较高，如何建立灵活的图像语义描述模型一直被视为计算机视觉领域一大挑战。

图像语义描述方法一般分为三大类：基于检索[1]的方法、基于模板匹配[2]的方法、深度神经网络[3]的方法。由于基于检索和模板匹配的方法依赖于人工提取特征和文本设计，效果往往并不理想。近年来，大多数图像语义描述的典型解决方案通常使用深度卷积神经网络(CNN)[4-7]对视觉特征进行编码，使用递归神经网络(RNN)生成语义描述。然而，人类视觉系统并不倾向于处理整张完整的图像，而是选择性地关注部分区域，因此我们期望图像语义模型也具备这个能力。近年来，融合注意力机制的图像语义描述取得了很大的进展，在图像语义描述中，视觉注意力[8-10]被证明是有效的。注意力机制让模型对局部区域投入更多的注意力资源，以获得更多目标的细节信息，并抑制其他不重要信息。融合注意力机制[11-13]的图像语义描述方法结合了上下文提取动态特征，从而产生更加丰富，更符合人类习惯的语义描述。

Kelvin 等首次提出了图像描述中的视觉注意力模型。通常，以上算法使用“硬”注意力机制来选择最有可能关注的区域[14]，或者使用“软”注意力机制的权重赋予空间特征。关于视觉问答(VQA)[15-16]，Zhu 等[17]采用“软”注意融合图像区域的特征。为了进一步细化空间注意，Yang 等应用了一种堆叠的空间注意模型，其中第二次注意力模型输入是基于第一次注意力模型调整后的特征图。上述模型的共同缺点是都采用特征图上的权重池化，导致空间信息的丢失和对图像特征利用不充分，进而影响模型语言描述的准确性。针对以上问题，本文分别在编码器和解码器部分融合注意力机制，解决了此问题，提高了模型的表达能力。

本文的主要贡献是同时将注意力机制引入编码器的特征提取模型中和解码器的语言模型中，其主要思想如下：

1) 在编码器部分，本文的目标是通过使用注意机制来增强表达能力，关注重要的特征并抑制不重要的特征。因此本文引入了空间注意力机制来增强通道维度上特征的提取。基于上述思想，我们可以让模型在通道维度上学习目标的类别和位置信息，从而通过学习需要强化或者抑制的位置信息，有效地帮助模型提取丰富的图像特征。

2) 在解码器部分，本文展示了引入注意力的一个优势，即通过可视化反应模型动态关注图像区域的效果，并且根据当前的解码状态自适应地确定描述图像的区域。我们研究的模型可以在生成描述的同时，对图像的重要部分分配更多的注意力资源，使得模型语义描述更加准确。

本文的创新点如下：1)引入了空间注意机制，有效地提高模型利用图像特征的能力，让模型关注主要目标，忽略次要信息；2)提出了融合注意力机制的编码器模型，该模型可以嵌入到不同的图像语义描述模型中，提升模型的性能。

2 融合注意力机制的图像语义描述模型

2.1 图像语义描述概述

图像语义描述是对给定图像进行简短描述的一项任务。近年来，针对深度神经网络(DNN)[18-20]在计算机视觉领域的巨大成功，计算机视觉领域的学者们对用于图像语义描述(NIC)[21-23]和视觉问题回答的编码器-解码器框架中的视觉注意力模型进行了进一步的研究，文献[24-26]提出了基于神经网络的图像语义描述生成方法。具体来说，这些方法使用编码器-解码器模型[27]，其中 CNN 将图像编码成一个静态的视觉特征向量，然后使用 RNN 或其变体[如门控递归单元(GRU)[28]和长短期记忆网络(LSTM)]生成图像语义描述。但是，静态向量无法使用图像特征关联当前句子的上下文。受机器翻译[29]中注意机制的启发，视觉注意模型在 NIC 和 VQA 中得到了广泛的应用，例如解码器动态地选择有用的源语言单词或子序列翻译成目标语言。本文将基于注意力模型的改进分为以下两个部分：1)编码器部分，通过通道方向的注意力提升了模型表达能力；2)解码器部分，动态选择感兴趣的区域。

本文采用主流的编码-解码器框架来生成图像语义描述，如图 1 所示，其中，CNN 首先将输入图像编码为矢量，然后 LSTM 将矢量解码为单词序列。CNN 通过融合空间注意力机制使原有的多层神经网络特征适应于句子的上下文。给定一张图像和相应的语义描述，通过以下目标直接对编码器-解码器模型进行优化，

$$\theta^{*}=\arg\max\sum_{(I,y)}\lg\ p(\boldsymbol{y}\mid\boldsymbol{I};\theta),\tag{1}$$

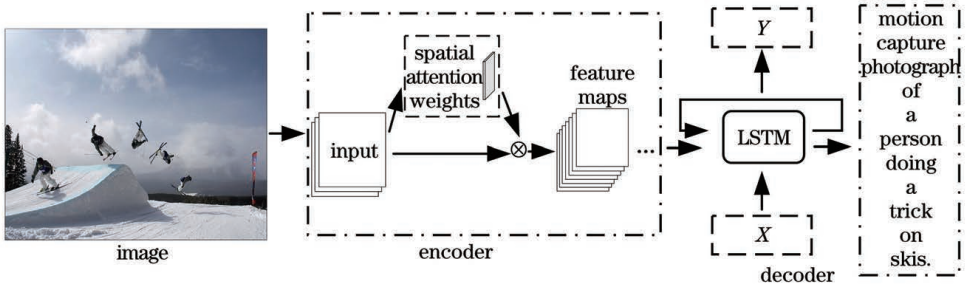

图 1　融合空间注意力机制的编码器-解码器模型

Fig. 1　Encoder-decoder model with integrated spatial attention mechanism

式中：θ 为模型参数；$\boldsymbol{I}$ 为给定图像；$\boldsymbol{y}=\{\boldsymbol{y}_0,\cdots,\boldsymbol{y}_{t-1}\}$ 是相应的语义描述。利用链式法则，可以将(1)式中联合概率分布的 lg 似然分解为有序条件，

$$\lg p(\boldsymbol{y})=\sum_{t=0}^{T}\lg p(\boldsymbol{y}_t\mid\boldsymbol{y}_0,\cdots,\boldsymbol{y}_{t-1},\boldsymbol{I})。\tag{2}$$

为了更方便地说明(1)式，(2)式中我们不考虑模型参数 θ。在编码器-解码器模型中，(2)式中每个单词条件概率表示为

$$\lg p(\boldsymbol{y}_t\mid\boldsymbol{y}_0,\cdots,\boldsymbol{y}_{t-1},\boldsymbol{I})=f(\boldsymbol{h}_t,\boldsymbol{c}_t),\tag{3}$$

式中：f 是非线性函数输出 $\boldsymbol{y}_t$ 的概率；$\boldsymbol{c}_t$ 是 t 时刻提取图像 $\boldsymbol{I}$ 的视觉上下文信息；$\boldsymbol{h}_t$ 是 LSTM 在 t 时刻的隐含层。(4)式中采用 LSTM 将图像特征解码成一个单词序列。LSTM 的隐藏单元的更新定义为

$$\boldsymbol{h}_t=\mathrm{LSTM}(\boldsymbol{x}_t,\boldsymbol{h}_{t-1},\boldsymbol{m}_{t-1}),\tag{4}$$

式中：$\boldsymbol{x}_t$ 为嵌入词表示；$\boldsymbol{h}_{t-1}$ 是 $t-1$ 时刻的隐藏状态；$\boldsymbol{m}_{t-1}$ 是 $t-1$ 时刻的记忆单元。通常，(3)式中上下文 $\boldsymbol{c}_t$ 是神经网络框架中的一个重要因素，它为生成语义描述提供了视觉信息。

2.2　空间注意力机制

利用特征的空间关系生成空间注意力。为了计算空间注意，首先沿着通道方向分别使用平均池化和最大池化，并将它们连接起来，以生成有效的图像特征。沿着通道方向采用池化操作可以有效地突出信息区域。在连接的图像特征上，使用卷积层来产生空间注意力。并使用两种池化操作整合通道信息，产生两个二维特征：$\boldsymbol{F}^s_{\mathrm{avg}}\in\mathbf{R}^{1\times H\times W}$ 和 $\boldsymbol{F}^s_{\max}\in\mathbf{R}^{1\times H\times W}$ 分别表示跨通道的平均池化和最大池化。然后将它们连接起来，并通过一个标准的卷积层进行卷积，生成二维空间注意力特征图。简而言之，空间注意力计算式为

$$\boldsymbol{M}_s(\boldsymbol{F})=\sigma\{f^{7\times7}\{[\mathrm{AvgPool}(\boldsymbol{F});\mathrm{MaxPool}(\boldsymbol{F})]\}\}=\sigma\{f^{7\times7}\{[\boldsymbol{F}^s_{\mathrm{avg}};\boldsymbol{F}^s_{\max}]\}\},\tag{5}$$

式中：σ 表示 sigmoid 激活函数；$\boldsymbol{F}$ 表示输入的图像特征；AvgPool 和 MaxPool 分别表示平均池化和最大池化；$f^{7\times7}$ 表示卷积操作，卷积核的大小为 7×7。

空间注意力机制用来确定图像目标的位置信息。虽然基于通道的注意力已被证明是有效的，但它没有考虑到图像区域的空间结构。事实上，缺乏空间结构会导致位置不准确，从而影响生成图像语义描述的准确性。为了保留图像区域的空间结构，本文引入了卷积空间注意力机制。如图 2 所示：一方面，本文的卷积空间注意力机制保留了图像的空间结构；另一方面，本文使用卷积核为 7×7 的卷积运算来提供更大的感受野并精确地确定每一步应该关注的区域，使得模型关注主要信息，忽略次要信息。

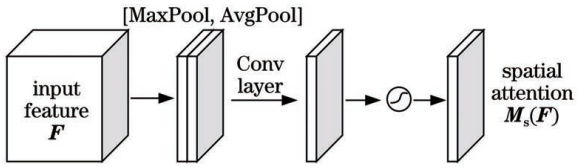

图 2　空间注意力模块图

Fig. 2　Diagram of spatial attention module

2.3　融合注意力机制

传统的编码器-解码器框架中，(3)式中 $\boldsymbol{c}_t$ 仅仅依赖于 CNN 编码器，CNN 提取图像特征，最后一层全连接层作为全局的图像特征。在生成单词的过程中，上下文向量 $\boldsymbol{c}_t$ 保持不变，不依赖于解码器隐含层。在基于注意力的框架中，$\boldsymbol{c}_t$ 依赖于编码器和解码器。在 t 时刻，解码器根据隐藏状态，关注图像的特定区域，利用 CNN 的一个卷积层的空间图像特征来计算 $\boldsymbol{c}_t$，研究表明，注意力模型可以显著提高图像语义描述的性能。

注意机制[30]是产生 $\boldsymbol{c}_t$ 的关键

$$\boldsymbol{V}=\boldsymbol{M}_s(\boldsymbol{F})\otimes\boldsymbol{F},\boldsymbol{c}_t=\mathrm{att}(\boldsymbol{V},\boldsymbol{h}_{t-1}),\tag{6}$$

式中：$\boldsymbol{M}_s(\boldsymbol{F})$ 为空间注意力机制权重系数；$\boldsymbol{F}$ 为图像特征；att(•)为注意力机制函数；$\boldsymbol{V}\in\mathbf{R}^{C\times W\times H}$($C$、$W$ 和 H 分别代表通道、宽度和高度)为空间注意力

图像特征。对于空间注意模型，(6)式中上下文向量 $\boldsymbol{c}_t$ 公式为

$$\begin{cases} z_t = \mu^T \tanh(W_s \boldsymbol{V} + W_h \boldsymbol{h}_{t-1}) \\ \boldsymbol{\alpha}_{ti} = \mathrm{softmax}(z_t) \\ \boldsymbol{c}_t = \sum_{i=1}^{H \times W} \boldsymbol{\alpha}_{ti} \boldsymbol{V} \end{cases}, \tag{7}$$

式中：μ^T、W_s、W_h 为待学习的参数；$\boldsymbol{\alpha}_{ti}$ 是 $\boldsymbol{V}$ 的注意力权重。如图 3 所示，融合注意力机制的模型结构，编码器引入空间注意力，改善了空间结构位置不准确的问题。输入图像通过 CNN 提取特征，该特征作为空间注意力机制的输入，然后生成具有空间注意力的编码器特征。解码器部分的注意力机制 $\boldsymbol{c}_t$，在生成每一个词的时候，根据隐藏层动态选择适当的区域。

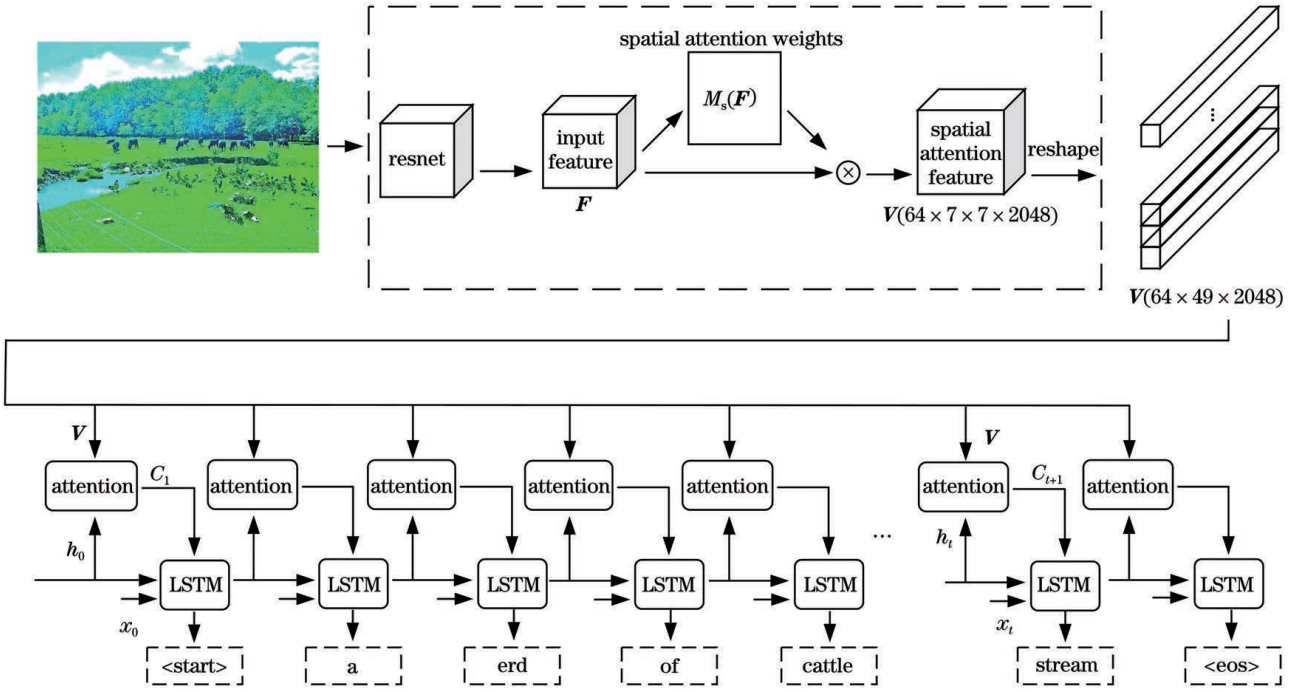

图 3　融合空间注意力机制的编码器-解码器网络结构图

Fig. 3　Diagram of encoder-decoder network with integrated spatial attention mechanism

3　实验结果及分析

3.1　实验环境介绍

本文实验所使用计算机是北京灵思创奇科技有限公司深度学习专用服务器，软件开发 IDE 为 Pycharm2018，编程语言为 Python3.6，深度学习框架为 PyTorch1.1.0，具体的硬件设备如表 1 所示。

表 1　实验所用服务器配置

Table 1　Server configuration used for the experiment

Instruction	Parameter
Memory	16.0 GB
CPU	Inter(R)Core(TM)i7-6700 CPU @3.40 GHz
GPU	NVIDIA GeForce GTX 1080 Ti

3.2　数据集

本文在两个常用的数据集上进行实验，表 2 为实验服务器配置。

1) MSCOCO[31] 数据集由微软公司提供，广泛应用于图像识别检测、图像分割，以及图像语义描述。该数据集包含两个版本，2014 版和 2017 版。本文使用 2014 版，其中训练集 82783 张图像，测试集 40504 张图像，验证集 40775 张图像。数据集中大多数图像来源于实际生活，具有复杂的背景信息，图像中包含多个目标，平均每张图像拥有 7.7 个对象，小目标众多；同时，数据集中每张图片有 5 个对应的人工语义描述。

2) Flickr30k[32] 数据集由雅虎公司 Flickr 网站提供，主要应用于图像语义描述。该数据集的图像场景大部分是一项活动，每张图片有 5 个对应的人工语义描述。数据集包含 31783 张图像。由于缺少正式的切分模式参考，我们使用 Karpathy 中给出的拆分标准。它使用 29000 张图像用于训练，1000 张图像用于验证，1000 张图像用于测试。

为了评估本文方法，采用了机器翻译和图像语

义描述任务中最普遍的客观量化评分方法 BLEU(Bilingual Evaluation Understudy，包含 BLEU1、BLEU2、BLEU3、BLEU4[33])。BLEU 广泛用于评估模型生成的句子和实际句子差异的指标，它的取值范围为 0.0～1.0，如果两个句子完全匹配，那么 BLEU 是 1.0。该指标计算代价小，容易理解，与人类评价结果高度相关，广泛应用于学术界和工业界。

表 2 实验服务器配置

Table 2 Experimental server configuration

Dataset name	Train	Valid	Test
Flickr30k	29783	1000	1000
MSCOCO	82783	40504	40775

3.3 实验结果及分析

本文的图像语义描述模型中，在编码部分，考虑到 VGG(Visual Geometry Group)和 ResNet 强大的特征提取能力，以及控制变量与其他方法相比的方便性，本文最终采用了广泛使用的 CNN 架构：VGG 模型和 ResNet-50 模型提取图像特征作为输入。通过 ImageNet 数据集做预训练，经过预训练的模型参数作为模型的初始化参数。在提取特征时，原始图像不进行裁剪或缩放，利用自适应空间平均池化，最终生成固定大小为 2048×7×7 的特征矩阵。在解码部分，单词嵌入和注意力维度设置为 512。使用 Adam[34]梯度下降法来优化网络，并将学习速率设置为 0.0001。图像的批大小为 64。在测试阶段，采用了 beam 搜索策略，从候选图像中选择最佳的语义描述，beam size 为 3。每个时刻都会保留 3 个概率最大的选择作为当前的最优选择，然后当解码下一时刻时，继续选择与之前保留的 3 个最优选择组合起来以后的概率最大的 3 个选择，依次循环迭代下去，直到编码结束。

本文融合注意力机制的图像语义描述模型方法与其他先进的模型比较。Deep VS、m-RNN[35]和 Google NIC 都是端到端多模态网络，将 CNN 用于图像编码，RNN 用于序列建模解码。Deep VS 提取句子和图像特征，将其嵌入共同的语义空间，视作检索任务，这种传统的方法不能自动生成丰富的描述，Deep VS 模型的编码器采用 VGG 模型，通过 ImageNet 数据集做预训练，经过预训练的模型参数作为模型的初始化参数。使用随机梯度下降(SGD)来优化网络，在测试阶段，采用了 beam 搜索策略，从候选图像中选择最佳的语义描述，beam size 为 7。m-RNN 和 Google NIC 在编码器和解码器都没有引入注意力机制，因此，在生成语义描述的过程中，不能准确捕捉图像的重要区域，进而导致语义描述的准确率偏低。m-RNN 的编码器同样采用 VGG 模型以及 ImageNet 预训练模型，解码器部分使用两层词嵌入系统的 m-RNN 网络。Google NIC 的编码器模型使用 Google NIC 模型以及 ImageNet 预训练模型，在测试阶段，采用了 beam 搜索策略，从候选图像中选择最佳的语义描述，beam size 为 3。SAT(Show Attend and Tell)模型中的软-注意和硬-注意都是解码器部分单一的注意模型。“软”注意权重从空间维度将视觉特征归纳为注意力特征，“硬”注意权重将区域特征随机抽取为注意力特征，SAT 模型的编码器部分没有引入注意力机制，因此图像语义描述的准确率偏低。

实验结果如表 3 所示，本文在 VGG 网络结构上融合了空间注意力机制，在 MSCOCO 上 BLEU-1 提升了 1.2%，BLEU-2 提升了 2.7%，BLEU-3 提升了 2.8%，BLEU-4 提升了 1.9%；在 Flickr30k 数据集上，BLEU-1 提升了 1.2%，BLEU-2 提升了 1.0%，BLEU-3 提升了 0.9%，BLEU-4 提升了 0.8%。实验数据损失曲线如图 4 所示，图 4(a)表示在 MSCOCO 数据集上的损失曲线，图 4(b)表示在 Flickr30k 数据集上的损失曲线。

表 3 VGG 网络结构融合空间注意力机制的实验数据对比

Table 3 Experimental comparison of spatial attention mechanism in VGG network

Model(VGG)	MSCOCO				Flickr30k			
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Deep VS	62.5	45.0	32.1	23.0	57.3	36.9	24.0	16.0
Log bilinear	70.8	48.9	34.4	24.3	60.0	38.0	25.4	17.1
SAT	70.7	49.2	34.4	24.3	61.0	40.5	27.3	18.2
Proposed	**71.9**	**51.9**	**37.2**	**26.2**	**62.2**	**41.5**	**28.2**	**19.0**

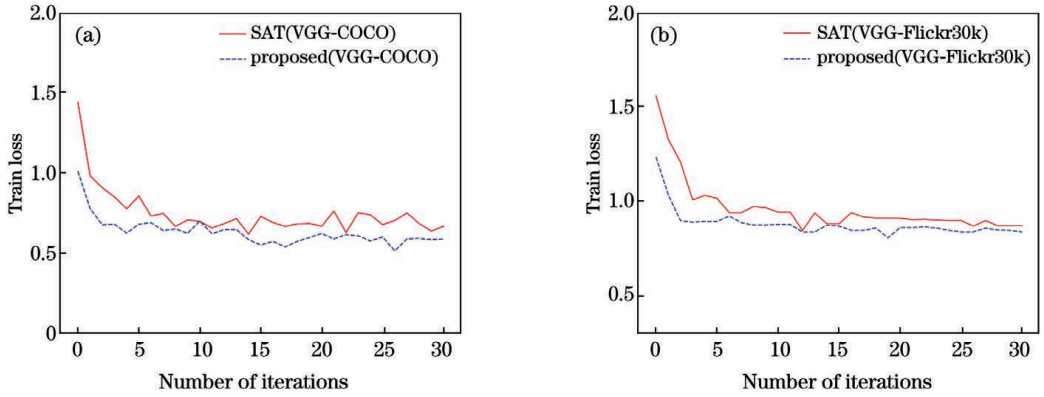

图 4　VGG 网络结构融合空间注意力机制的实验数据损失曲线。(a) VGG(MSCOCO);(b) VGG(Flickr30k)

Fig. 4　Experimental data loss curves of spatial attention mechanism in VGG network. (a) VGG(MSCOCO); (b) VGG(Flickr30k)

实验结果如表 4 所示，本文在 ResNet-50 网络结构上融合了空间注意力机制，在 MSCOCO 上 BLEU-1 提升了 0.3%，BLEU-2 提升了 0.7%，BLEU-3 提升了 1.1%，BLEU-4 提升了 1.2%；在 Flickr30k 上，BLEU-1 提升了 1.2%，BLEU-2 提升了 1.3%，BLEU-3 提升了 1.1%，BLEU-4 提升了 0.9%。实验数据损失曲线如图 5 所示，图 5(a)表示在 MSCOCO 数据集上的损失曲线，图 5(b)表示在 Flickr30k 数据集上的损失曲线。

表 4　ResNet 网络融合空间注意力机制的实验数据对比

Table 4　Experimental comparison of spatial attention mechanism in ResNet network

Model (ResNet-50)	MSCOCO				Flickr30k			
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Deep VS	62.5	45.0	32.1	23.0	57.3	36.9	24.0	16.0
Google NIC	66.6	46.1	32.9	24.6	66.3	42.3	27.7	18.3
m-RNN	67.0	49.0	35.0	25.0	60.0	41.0	28.0	19.0
SAT	72.7	52.8	37.9	26.7	63.4	42.6	29.2	19.7
Proposed	**73.0**	**53.5**	**39.0**	**27.9**	**64.6**	**43.9**	**30.3**	**20.6**

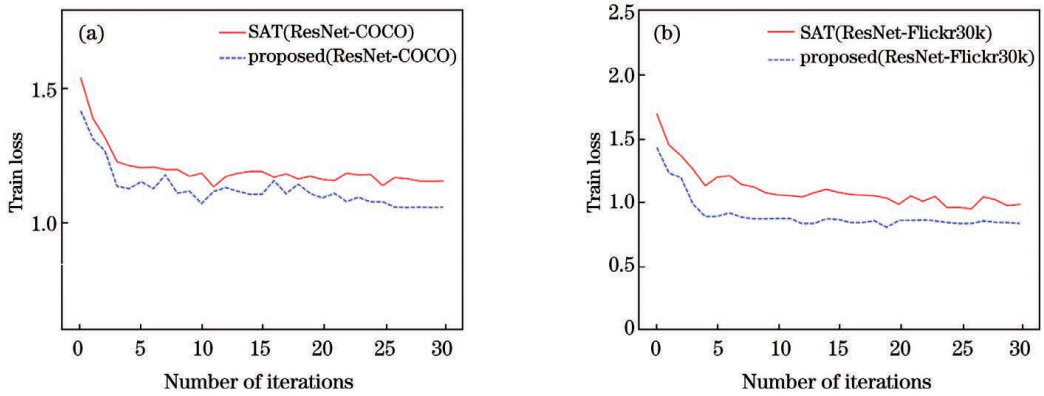

图 5　ResNet 网络结构融合空间注意力机制的实验数据损失曲线。(a) ResNet-50(MSCOCO);(b) ResNet-50(Flickr30k)

Fig. 5　Experimental data loss curves of spatial attention mechanism in ResNet network. (a) ResNet-50(MSCOCO); (b) ResNet-50(Flickr30k)

3.4　实验可视化及分析

为了更好地理解本文模型，图 6～9 中分别提供四个代表性的例子。其中(a)表示 MSCOCO 数据集中的测试集图片，(b)表示 SAT 模型可视化结果，(c)表示本文模型可视化结果。同时，(b)和(c)中每一张图像分别表示模型在每一个单词预测步骤

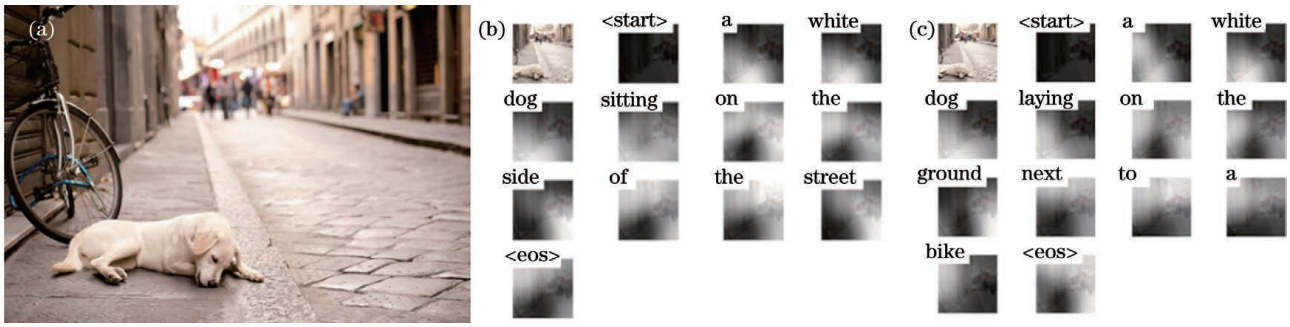

图 6 可视化结果对比。(a)测试集;(b)SAT 模型可视化结果;(c)本文模型可视化结果

Fig. 6 Comparison of visualization results. (a) Test set; (b) SAT model visualization results; (c) proposed model visualization results

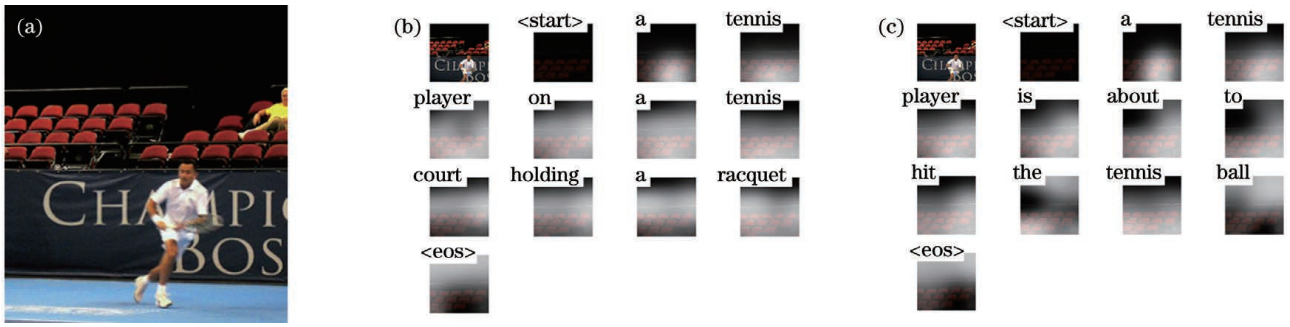

图 7 可视化结果对比。(a)测试集;(b)SAT 模型可视化结果;(c)本文模型可视化结果

Fig. 7 Comparison of visualization results. (a) Test set; (b) SAT model visualization results; (c) proposed model visualization results

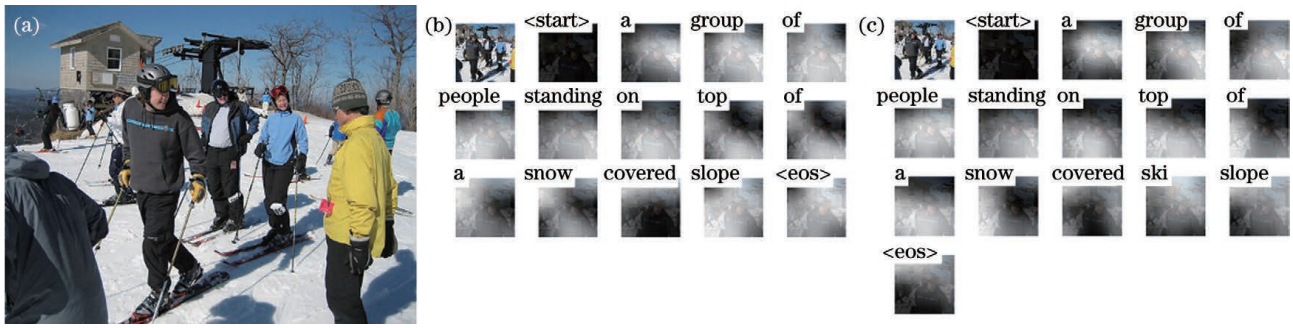

图 8 可视化结果对比。(a)测试集;(b)SAT 模型可视化结果;(c)本文模型可视化结果

Fig. 8 Comparison of visualization results. (a) Test set; (b) SAT model visualization results; (c) proposed model visualization results

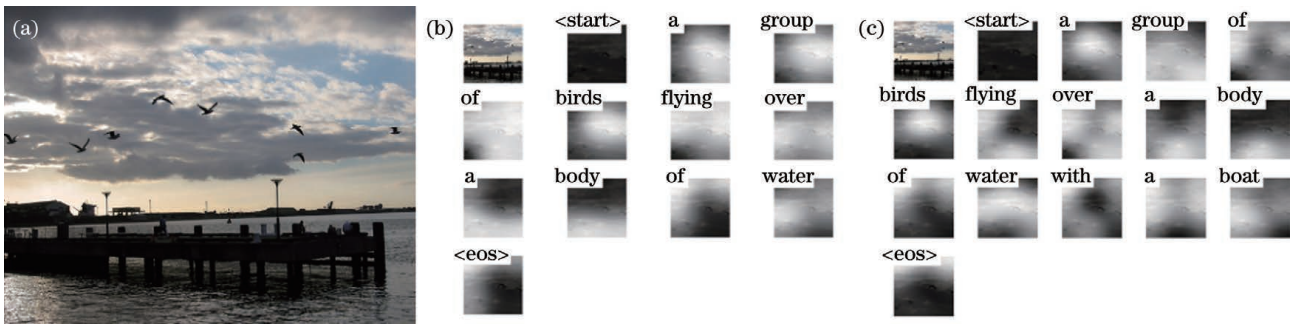

图 9 可视化结果对比。(a)测试集;(b)SAT 模型可视化结果;(c)本文模型可视化结果

Fig. 9 Comparison of visualization results. (a) Test set; (b) SAT model visualization results; (c) proposed model visualization results

中注意力可视化结果。(b)和(c)图体现了空间注意力机制让模型关注主要目标，忽略次要信息，白色区域表示空间注意力机制投入更多的注意力资源，黑色区域代表空间注意力机制投入了更少的注意力资源。每个示例包含三个语义描述的句子，分别对应于 SAT 模型、本文模型和真值(GT)。图 6 代表静态的图像，图 7 代表动态的图像，图 8 代表细节信息，图 9 代表模型效果不佳的例子。由于本文模型融合了注意力机制，在以下前三种场景下的语义描述均优于 SAT 模型，体现了融合注意力机制的有效性。

图 6 的可视化实验结果如下："GT：The white dog lays next to the bicycle on the sidewalk. SAT：A white dog sitting on the side of the street. Proposed：A white dog laying on the ground next to a bike. "其中，SAT 模型没有自行车(bike)的描述，"坐着"(sitting)没有准确描述图像信息。本文模型中，"躺着"(laying)相对于"坐着"(sitting)更加准确地描述了图像中狗的状态，同时由于本文引入注意力机制，模型能够更好地捕捉到自行车(bike)的信息。

图 7 的可视化实验结果如下："GT：A tennis player runs to hit the ball. SAT：A tennis player on a tennis court holding a racquet. Proposed：A tennis player is about to hit the tennis ball. "其中，SAT 模型仅体现了拿着球拍(holding a racquet)的静态描述，没有生动的动作描述。本文模型生成了主语将要(is about to)击打网球(hit the tennis ball)的动作状态，使得图像语义描述更加准确。

图 8 的可视化实验结果如下："GT：A group of people riding skis on top of a ski slope. SAT：A group of people standing on top of a snow covered slope. Proposed：A group of people standing on top of a snow covered ski slope." 其中，SAT 模型没有滑雪板(ski)的描述，属于重要信息遗漏。本文模型由于引入了空间注意力机制有效地提取图像目标信息滑雪板(ski)，使其描述更加接近标准描述。

图 9 的可视化实验结果如下："GT：Some birds are flying over a beach pier. SAT：A group of birds flying over a body of water. Proposed：A group of birds flying over a body of water with a boat. "其中，SAT 模型没有海滩码头(beach pier)的描述，属于信息遗漏。本文模型虽然引入了空间注意力机制，但是目标信息海滩码头(beach pier)与船只(boat)相似，模型没有准确地捕捉到图像目标信息海滩码头(beach pier)，使得图像语义描述与标准描述有偏差。

4 结 论

本文提出了融合空间注意力的图像语义描述算法，通过引入空间注意力机制，提高了编码器的表达能力。融合了注意力机制的编码器，使得模型对图像特征使用更加充分，以及模型的图像语义描述更加丰富、准确。实验表明，该算法在 MSCOCO 数据集的 VGG 和 Resnet-50 网络结构上，BLEU-4 分别提高了 0.9%和 1.2%；在 Flickr30k 数据集的 VGG 和 Resnet-50 网络结构上，BLEU-4 分别提高了 0.8%和 0.9%，证明了通过注意力机制模型能够精确地提取图像目标的位置信息，从而更加细致、生动地描述图像的内容。

参 考 文 献

[1] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), June 20-25, 2005, San Diego, CA, USA. New York: IEEE Press, 2005: 886-893.

[2] Fang H, Gupta S, Iandola F, et al. From captions to visual concepts and back[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 1473-1482.

[3] Xu K, Ba J, Kiros R, et al. Show, attend and tell: neural image caption generation with visual attention [EB/OL]. (2016-04-19) [2020-09-04]. https://arxiv.org/abs/1502.03044.

[4] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[5] Tao Z Y, Li J, Tang X L. Texture images classification algorithm combining wavelet transform and capsule network[J]. Laser & Optoelectronics Progress, 2020, 57(24): 241002.
陶志勇, 李杰, 唐晓亮. 融合小波变换与胶囊网络的纹理图像分类算法[J]. 激光与光电子学进展, 2020, 57(24): 241002.

[6] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[C]//Thirty-first AAAI

conference on artificial intelligence, February 4-9, 2017, San Francisco, California, USA. Virigina: AIAA, 2017: 4278-4284.

[7] Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 2261-2269.

[8] You Q Z, Jin H L, Wang Z W, et al. Image captioning with semantic attention[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 4651-4659.

[9] Huang L, Wang W M, Xia Y X, et al. Adaptively aligned image captioning via adaptive attention time [EB/OL]. (2020-01-06)[2020-09-04]. https://arxiv.org/abs/1909.09060v3.

[10] Huang L, Wang W M, Chen J, et al. Attention on attention for image captioning[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 4633-4642.

[11] Liu F L, Liu Y X, Ren X C, et al. Aligning visual regions and textual concepts for semantic-grounded image representations[EB/OL]. (2019-11-04)[2020-09-04]. https://arxiv.org/abs/1905.06139v3.

[12] Yang X, Tang K H, Zhang H W, et al. Auto-encoding scene graphs for image captioning[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 10677-10686.

[13] Zhao X H, Yin L F, Zhao C L. Image captioning based on global-local feature and adaptive-attention [J]. Journal of Zhejiang University (Engineering Science), 2020, 54(1): 126-134.
赵小虎, 尹良飞, 赵成龙. 基于全局-局部特征和自适应注意力机制的图像语义描述算法[J]. 浙江大学学报(工学版), 2020, 54(1): 126-134.

[14] Lu J S, Xiong C M, Parikh D, et al. Knowing when to look: adaptive attention via a visual sentinel for image captioning[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 3242-3250.

[15] Yang Z C, He X D, Gao J F, et al. Stacked attention networks for image question answering[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 21-29.

[16] Xu H J, Saenko K. Ask, attend and answer: exploring question-guided spatial attention for visual question answering[EB/OL](2015-11-17)[2020-09-04]. https://arxiv.org/abs/1511.05234v1.

[17] Zhu Y K, Groth O, Bernstein M, et al. Visual7W: grounded question answering in images[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 4995-5004.

[18] Dong Y F, Yang Y X, Wang L Q, Image semantic segmentation based on multi-scale feature extraction and fully connected conditional random fields[J]. Laser & Optoelectronics Progress, 2019, 56(13): 131007.
董永峰, 杨雨沂, 王利琴. 基于多尺度特征提取和全连接条件随机场的图像语义分割方法[J]. 激光与光电子学进展, 2019, 56(13): 131007.

[19] Yue S Y. Image semantic segmentation based on hierarchical context information [J]. Laser & Optoelectronics Progress, 2019, 56(24): 241005.
岳师怡. 基于多层级上下文信息的图像语义分割[J]. 激光与光电子学进展, 2019, 56(24): 241005.

[20] Wang Y H. Image caption based on multi-fusion model[J]. Henan Science and Technology, 2019 (14): 34-36.
王媛华. 基于多融合模型的图像语义描述研究[J]. 河南科技, 2019(14): 34-36.

[21] Vinyals O, Toshev A, Bengio S, et al. Show and tell: a neural image caption generator[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 3156-3164.

[22] Vinyals O, Toshev A, Bengio S, et al. Show and tell: lessons learned from the 2015 MSCOCO image captioning challenge [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39 (4): 652-663.

[23] Li R F, Liang H Y, Feng F X, et al. Paragraph image captioning with deep fully convolutional neural networks[J]. Journal of Beijing University of Posts and Telecommunications, 2019, 42(6): 155-161.
李睿凡, 梁昊雨, 冯方向, 等. 全卷积神经结构的段落式图像描述算法[J]. 北京邮电大学学报, 2019, 42(6): 155-161.

[24] Karpathy A, Li F F. Deep visual-semantic alignments for generating image descriptions[C]//2015 IEEE Conference on Computer Vision and

Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 3128-3137.

[25] Chen L, Zhang H W, Xiao J, et al. SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6298-6306.

[26] Ren Z, Wang X Y, Zhang N, et al. Deep reinforcement learning-based image captioning with embedding reward[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI. New York: IEEE Press, 2017: 1151-1159.

[27] Bengio S, Vinyals O, Jaitly N, et al. Scheduled sampling for sequence prediction with recurrent neural networks[EB/OL]. (2015-09-23)[2020-09-04]. https://arxiv.org/abs/1506.03099.

[28] Chung J Y, Gulcehre C, Cho K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling [EB/OL]. (2014-12-11)[2020-09-04]. https://arxiv.org/abs/1412.3555v1.

[29] Luong M T, Pham H, Manning C D. Effective approaches to attention-based neural machine translation [EB/OL]. (2015-09-20)[2020-09-04]. https://arxiv.org/abs/1508.04025.

[30] Corbetta M, Shulman G L. Control of goal-directed and stimulus-driven attention in the brain[J]. Nature Reviews Neuroscience, 2002, 3(3): 201-215.

[31] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context[M]//Fleet D, Pajdla T, Schiele B, et al. Computer vision-ECCV 2014. Cham: Springer, 2014, 8693: 740-755.

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

[33] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation [C] // Proceedings of the 40th Annual Meeting on Association for Computational Linguistics-ACL'02, July 7-12, 2002, Philadelphia, Pennsylvania. Morristown: Association for Computational Linguistics, 2002.

[34] Kingma D P, Ba J. Adam: a method for stochastic optimization[EB/OL]. (2017-01-30)[2020-09-04]. https://arxiv.org/abs/1412.6980v9.

[35] Mao J H, Xu W, Yang Y, et al. Deep captioning with multimodal recurrent neural networks (m-RNN) [EB/OL]. (2015-06-11)[2020-09-04]. https://arxiv.org/abs/1412.6632v2.