智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于注意力融合的图像描述生成方法

莫宏伟， 田朋

引用本文:
莫宏伟, 田朋. 基于注意力融合的图像描述生成方法[J]. 智能系统学报, 2020, 15(4): 740–749.
MO Hongwei, TIAN Peng. An image caption generation method based on attention fusion[J]. *CAAI Transactions on Intelligent Systems*, 2020, 15(4): 740–749.

在线阅读 View online: https://dx.doi.org/10.11992/tis.201910039

DOI: 10.11992/tis.201910039

基于注意力融合的图像描述生成方法

莫宏伟，田朋
（哈尔滨工程大学 自动化学院，黑龙江 哈尔滨 150001）

摘　要：空间注意力机制和高层语义注意力机制都能够提升图像描述的效果，但是通过直接划分卷积神经网络提取图像空间注意力的方式不能准确地提取图像中目标对应的特征。为了提高基于注意力的图像描述效果，提出了一种基于注意力融合的图像描述模型，使用 Faster R-CNN（faster region with convolutional neural network）作为编码器在提取图像特征的同时可以检测出目标的准确位置和名称属性特征，再将这些特征分别作为高层语义注意力和空间注意力来指导单词序列的生成。在 COCO 数据集上的实验结果表明，基于注意力融合的图像描述模型的性能优于基于空间注意力的图像描述模型和多数主流的图像描述模型。在使用交叉熵训练方法的基础上，使用强化学习方法直接优化图像描述评价指标对模型进行训练，提升了基于注意力融合的图像描述模型的准确率。

关键词：图像描述；卷积神经网络；空间注意力；Faster R-CNN；注意力机制；名称属性；高层语义；强化学习
中图分类号：TP181　　**文献标志码：**A　　**文章编号：**1673−4785(2020)04−0740−10

中文引用格式：莫宏伟，田朋. 基于注意力融合的图像描述生成方法 [J]. 智能系统学报，2020，15(4)：740–749.
英文引用格式：MO Hongwei，TIAN Peng. An image caption generation method based on attention fusion[J]. CAAI transactions on intelligent systems，2020，15(4)：740–749.

An image caption generation method based on attention fusion

MO Hongwei，TIAN Peng
(College of Automation，Harbin Engineering University，Harbin 150001，China)

Abstract: The spatial attention mechanism and the high-level semantic attention mechanism can improve the effect of image captioning，but the method for extracting the spatial attention of image by directly dividing the convolutional neural network cannot accurately extract the features corresponding to target in the image. In order to improve the effect of image captioning based on attention，this paper proposes an image caption model based on attention fusion，using Faster R-CNN (faster region with convolutional neural network) as an encoder to exect image features and simultaneously detect the features of accurate position and noun attribute of the target object，then those features as high-level semantic attention and spatial attention respectively to guide the generation of word sequence. The experimental results on COCO dataset show that the performance of the image caption model based on attention fusion outperforms the image caption models based on spatial attention and most mainstream image caption models. Based on the cross entropy training method，we use reinforcement learning method to directly optimize the image caption evaluation index to train the model，which significantly improves the accuracy of the image caption model based on attention fusion.

Keywords: image caption; convolutional neural network; spatial attention; Faster R-CNN; attention mechanism; noun attribute; high-level semantic; reinforcement learning

图像描述是计算机视觉和自然语言处理的交叉学科，也是当前人工智能领域中研究的一个热点和难点问题[1]。图像描述是让计算机生成给定图像内容的文字性表述，相比于图像分类和目标检测识别等视觉任务，图像描述不仅需要检测识别出图像中的物体和关系，还需要使用自然语言将图像的主要语义信息进行准确地表述[2]。

收稿日期：2019−10−29.
基金项目：国家重点研发计划新一代人工智能重大专项（2018AAA0102702).
通信作者：莫宏伟. E-mail：honwei2004@126.com..

人类获取外部世界信息主要是通过视觉系统, 人类大脑的注意力机制能从环境中选择出感兴趣的目标区域, 并重点关注这些目标区域的相关信息。受此启发, 研究人员成功地将注意力机制应用于机器翻译和图像描述等诸多深度学习任务中[3-4]。

在基于编码解码结构的图像描述模型中, 卷积神经网络作为编码器将图像的主要特征提取为一个固定大小的特征向量, 长短期记忆网络作为解码器利用该特征向量生成描述文本[5-6]。长短期记忆网络沿着时间展开, 其网络中包含的图像特征信息会逐渐减少, 而且一次将全部特征信息送入到解码器中, 这不能使解码器充分地利用特征信息, 因此该模型无法取得非常好的描述效果。

在基于注意力机制的图像描述模型中, 卷积神经网络将提取的图像特征按照图像的空间位置划分为一定数量的局部特征向量, 注意力机制根据长短期记忆网络的隐藏状态从图像局部特征向量集合中动态选择与当前时刻生成单词有关的图像局部特征来指导当前时刻单词的生成。使用注意力机制改进的图像描述模型能够充分地利用图像的特征信息, 显著地提升了模型生成描述的效果[7]。

然而使用固定值均匀地划分图像获取空间注意力的方法存在着注意力不精确的问题, 分割出来的图像区域一般与目标的大小不符, 不利于单词序列的生成。为了解决基于空间注意力机制存在的不能准确选取目标对应特征的问题, 本文提出了一种基于空间注意力和高层语义注意力融合的图像描述模型, 使用具有卷积神经网络的快速区域目标检测模型(faster region with convolutional neural network, Faster R-CNN)作为图像描述模型的编码器, 使用高层语义注意力机制在提取图像特征的同时检测图像中的物体和显著视觉区域的位置和名称属性, 将位置特征和名称属性特征分别作为空间注意力机制和高层语义注意力机制的输入, 从而提高图像描述的准确性。

本文的主要贡献如下:

1) 针对空间注意力机制中存在的注意力不精确的问题, 提出使用 Faster R-CNN 作为编码器对空间注意力机制进行改进, 提高空间注意力机制的精度;

2) 提出一种融合空间注意力与高层语义注意力的注意力机制, 在提取图像特征的同时检测出图像中目标的准确位置和名称属性特征以指导单词的生成;

3) 使用强化学习方法训练基于注意力融合的图像描述模型, 提升模型在评价指标上的得分, 进一步提高模型的性能。

1　相关研究

图像描述涉及计算机视觉与自然语言处理两个研究领域, 近年来成为深度学习中的研究热点。图像描述的方法可以分为三大类: 基于模板的方法、基于检索的方法和基于编码解码结构的方法。

早期的图像描述研究主要是基于模板的方法, 该方法首先检测出图像中的物体及其属性等关键信息, 然后将这些信息通过特定的模板、语言模型或句法模型生成对应的描述。Farhadi 等[8]将图像中的物体、动作和场景检测出来, 形成对应的三元组, 根据模型信息生成描述; Girish 等[9]使用检测器识别图像中的物体、物体属性和相互关系, 然后使用条件随机场预测标签, 最后使用文本语料库生成图像的描述; Li 等[10]将图像中物体的相关信息表示为关系短语, 通过语言模型将短语组合生成描述语句。

基于模板的方法生成描述的质量依赖于特征提取部分的性能和模板的设定, 由于使用的模板是固定的, 其生成的描述虽然能够包含图像的主要语义信息, 但其描述格式单调, 表达生硬, 效果并不理想。

基于检索的方法将图像描述问题转化成图像检索问题, 在提取图像特征信息后通过相似度量算法来比较图像之间的相似度, 然后利用数据集中相似图像的描述经过合理地组织生成新的描述。Polina 等[11]在图像数据集中检索出与要描述图像相似的图像, 将这些图像对应的描述文本通过随机树形结构算法提取出词组以生成图像的描述语句。Yashaswi 等[12]使用图像的视觉特征作为衡量图像相似度的指标, 将与要描述图像相似的描述文本分解成短语, 通过图像的相似性等指标确定最优描述。Jacob 等[13]使用视觉几何组网络(visual geometry group network, VGG)模型[14]提取图像特征, 使用 KNN[15]找到与描述图像相似的图像以确定目标图像的描述。

基于检索的方法将图像描述看作是一种检索任务, 其性能依赖于标注的图像数据集的大小和检索算法的准确程度, 该方法过于依赖描述数据集, 所生成的图像描述局限于数据集中的描述。

基于编码解码结构的图像描述模型使用卷积神经网络作为编码器提取图像特征向量, 使用循环神经网络作为解码器, 根据提取到的图像特征向量来生成图像对应的描述文本。

Mao 等[16]首先提出编码解码结构的图像描述模型 m-RNN, 使用卷积神经网络将图像转化成特

征向量，利用循环神经网络根据之前生成的单词和图像的特征向量生成下一个单词，循环此过程直至生成完整的描述。Oriol 等[17]提出的图像描述模型使用长短期记忆网络（long short-term memory, LSTM）[18]替代循环神经网络作为解码器，LSTM 不仅能够记忆长期信息，而且能够解决梯度消失和梯度爆炸问题。

与基于模板和基于检索的方法相比，基于编码解码结构的方法利用卷积神经网络提取图像特征和循环神经网络生成序列数据上的优势，所生成的描述结构灵活、表达自然，而且模型泛化能力强、性能良好。

受到人的注意力会集中在感兴趣的物体上的启发，研究人员将视觉注意力机制引入到编码解码模型中，使得模型在生成不同的单词时关注图像中对应区域的特征。注意力机制改变了编码器与解码器之间的连接方式，使得模型生成的描述更符合图像的内容。目前使用的注意力机制主要分以下 3 类：

1) 空间注意力

Xu 等[19]最先提出将空间注意机制加入到编码解码模型中，将编码器提取到的图像特征先划分为不同的区域，利用当前时刻长短期记忆单元的隐藏状态和图像区域特征经过注意力模型决定各区域特征的权重，动态地选择与当前时刻生成单词相关的图像区域特征来指导单词的生成。

2) 高层语义注意力

You 等[20]提出使用目标检测算法检测图像中主要目标的名称属性，并将其作为高层语义信息，由此提出了基于高层语义注意力机制模型。该模型先将提取的物体名称属性向量送入注意力机制模型，然后通过动态地选择名称属性向量来指导单词序列的生成，从而提高生成描述的质量。

3) 层级注意力

Chen 等[21]提出了融合空间和高层语义注意力的基于层级注意力的图像描述模型，使用层级注意力机制动态地选择卷积神经网络的卷积特征图来指导单词的生成。该模型结合空间注意力机制、语义注意力机制和层级注意力机制，所生成的图像描述效果超过了同时期的其他图像描述模型。

伴随着深度学习技术的快速发展，基于编码解码结构的图像描述方法逐渐成为主流，由于注意力机制的应用，图像描述的效果也在不断地提升。2015 年，Minh-thang 等[22]提出全局注意力和局部注意力模型，全局注意力考虑输入的隐状态来生成语境向量，局部注意力关注部分隐状态，该机制的难点是如何找到与预测词对应的隐状态。Xu 等[23]对图像描述的网络结构进行了改进，使用卷积层提取图像特征，每个时刻传入 LSTM 的是上一时刻的状态和经过加权处理后的卷积层特征，使用注意力机制对提取的特征进行加权。

2017 年，Marco 等[24]提出使用一种区域注意力模型，考虑了状态与预测单词以及图像区域与单词和状态之间的关系，使用卷积神经网络提取图像特征并生成描述单词。Li 等[25]提出一种使用注意力机制将局部特征与全局特征的权重进行组合的模型，首先分别提取全局特征与局部特征，然后使用注意力机制对全局特征和局部特征进行权重分配。Lu 等[26]提出一种带有视觉标记的自适应注意力模型，在解码过程中依据语义信息对不同的单词分配不同的注意力权重。

2018 年，Anderson 等[27]提出使用 Faster R-CNN[28]作为编码器在提取图像特征的同时检测目标及其所在的区域，将这些区域对应的特征向量送入到空间注意力模型中，经过注意力机制的动态分配来指导单词序列的生成。

当前大多数图像描述算法使用交叉熵作为损失函数训练模型存在着曝光偏差和衡量标准不一致的问题，曝光偏差会导致生成的单词与图像内容具有差异，影响下一个单词生成的准确性，衡量标准不一致导致模型在训练时无法充分地优化评价指标。为了解这个问题，研究人员提出了使用强化学习方法[29]来改进图像描述模型。Marc'aurelio 等[30]使用强化学习方法优化序列生成模型，Liu 等[31]使用全连接网络来估计基线，并使用更符合人类评价标准的图像描述指标 SPICE（semantic propositional image caption evaluation）。本文使用 REINFORCE 算法[32]对图像描述模型进行训练，将该算法应用到基于注意力融合的图像描述模型中，解决了交叉熵训练方法存在的曝光偏差和衡量标准不一致的问题，提高了基于注意力融合的图像描述模型生成描述的准确率。

2　模型框架

2.1　整体模型

基于注意力融合的图像描述模型主要由图像特征提取、特征编码、注意力模型和特征解码等部分组成。使用 ResNet-101[33]作为 Faster R-CNN 的特征提取网络以提高图像特征提取的能力和目标检测的精度，Faster R-CNN 作为编码器能够提取图像中物体和显著视觉区域的位置和名称属性信息，将目标对应的特征向量和名称属性

信息分别作为空间注意力机制和高层语义注意力机制的输入, 经过注意力模型整合处理后送入解码器, 最终由解码器生成单词序列, 基于注意力融合的图像描述模型结构如图 1 所示。

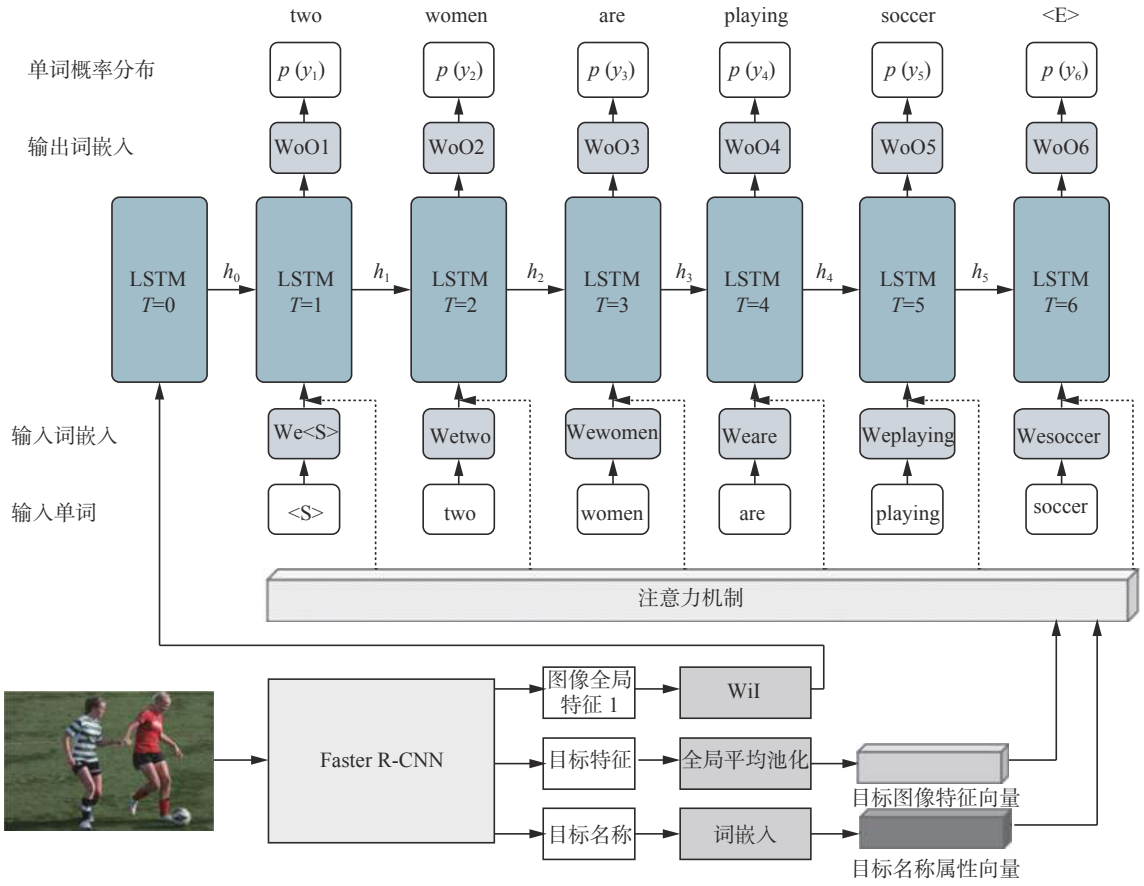

图 1　基于注意力融合的图像描述模型结构

Fig. 1　Image caption model structure based on attention fusion

使用 Faster R-CNN 模型对输入图像进行检测, 并提供空间注意力和高层语义注意力, 目标对应的图像特征隐式地包含了名称属性信息, 对该图像特征进行推断得到目标的名称属性, 在其中筛选置信度大于 0.3 的目标作为注意力机制的输入。被检测到目标的空间位置信息对应在 ResNet-101 最后一层卷积层的特征图, 将其进行平均池化处理得到 2 048 维的图像特征向量作为空间注意机制的输入, 将目标的名称属性经过词嵌入表示为 512 维的名称属性向量作为高层语义注意力的输入。将 ResNet-101 最后一层卷积层的特征图经过平均池化处理得到 2 048 维的图像全局特征向量作为编码器初始时刻的输入, 目标对应的图像特征向量和名称属性向量经过注意力机制的分配在解码器生成单词的过程中动态地指导单词序列的生成。使用在 ImageNet 数据集上预先训练好的 Faster R-CNN 作为编码器, 在训练图像描述模型时固定 Faster R-CNN 的参数, 仅对注意力机制和解码器的参数进行训练。本文直接设置目标对应的特征向量的权重与名称属性向量的权重相等, 使用注意力模型同时决定两者的权重。

2.2　注意力机制

本文使用的注意力模型选取目标对应的特征向量 $\{v_1, v_2, \cdots, v_n\}$和名称属性向量$\{\alpha_1, \alpha_2, \cdots, \alpha_n\}$ 以及解码器中长短期记忆网络上一时刻的隐藏状态 h_{i-1} 来决定当前时刻选取的特征向量和名称属性向量的权重 α_{ij}, 计算公式为

$$e_{ij} = f_{\text{att}}(h_{i-1}, v_j, a_j) \tag{1}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \tag{2}$$

其中注意力机制 f_{att} 是一个多层感知机模型, 对其输出结果使用 Softmax 进行归一化可以得到目标对应的特征在时刻 i 的权重分布, 这些权重就表示描述模型对图像各目标的重视程度。

当前时刻输入的视觉上下文信息 z_i 为

$$z_i = \left(\sum_{j=1}^{L} \alpha_{ij} v_j, \sum_{j=1}^{L} \alpha_{ij} a_j \right) \tag{3}$$

将图像全局特征 V 分别通过两个独立的多层

感知机计算得到长短期记忆网络的细胞单元状态 c_0 和隐藏状态的初始值：

$$c_0 = f_{\text{init},c}(V) \tag{4}$$

$$h_0 = f_{\text{init},h}(V) \tag{5}$$

然后，根据前一时刻的输出 y_{i-1}、前一时刻的隐藏状态 h_{i-1} 和视觉上下文 z_i 可以计算得到当前时刻的隐藏状态 h_i，由长短期记忆网络的公式计算可得：

$$h_i = \text{LSTM}(y_{i-1}, h_{i-1}, z_i) \tag{6}$$

由当前时刻的隐藏状态、视觉上下文信息以及前一时刻的输出通过 Softmax 可以得到当前输出单词的概率分布：

$$p(y_i|z_i, y_{i-1}) = \text{softmax}(Ey_{i-1} + L_h h_i + L_z z_i) \tag{7}$$

模型使用交叉熵损失函数进行训练，给定人工标注描述 y^*，使用 θ 表示模型中的参数，交叉熵损失函数 $L(\theta)$ 的表达式为

$$L(\theta) = -\sum_{i=1}^{N} \log(p(y_i^* | z_i, y_{i-1}^*)) + \lambda_\theta \| \theta \|_2^2 \tag{8}$$

其中 $\lambda_\theta \| \theta \|_2^2$ 表示 L2 正则化项，可以防止模型过拟合，加快模型收敛的速度。

2.3 模型优化

REINFORCE 算法是强化学习中一种常用的策略梯度算法，首先将策略参数化用以估计目标函数对策略参数的梯度，然后根据梯度下降算法不断对策略参数进行更新，直至找到最优策略。

由于 REINFORCE 算法在模型的训练过程中能够对离散和不可微的指标进行优化，适用于处理文本序列生成问题，在训练时可以直接优化评价指标对模型进行训练以提高模型生成描述的效果。所以将 REINFORCE 算法应用到基于注意力融合的图像描述模型中，将图像描述评价指标作为奖励函数，把模型的参数看成状态，把生成的描述在图像描述评价指标上的得分作为奖励，并利用得到的奖励来更新模型参数。

使用图像描述评价指标 CIDEr 作为奖励函数，模型生成的图像描述为 y，数据集中图像对应的人工标注描述为 s，则奖励 $r(y)$ 表示为

$$r(y) = \text{CIDEr}(y, s) \tag{9}$$

图像描述模型训练的目标是最大化期望奖励，其表达式为

$$L(\theta) = -E_{y^s \sim p_\theta}[r(y^s)] \tag{10}$$

其中 $y^s = \{y_1^s, y_2^s, \cdots, y_T^s\}$ 表示模型采样得到的单词所构成的单词序列，图像描述生成的单词是从单词的概率分布 p_θ 中采样得到的，单个采样样本对应的损失函数可以近似为

$$L(\theta) \approx -r(y^s), y^s \sim p_\theta \tag{11}$$

损失函数关于参数 θ 的梯度为

$$\nabla_\theta L(\theta) = -E_{y^s \sim p_\theta}[r(y^s) \nabla_\theta \log p_\theta(y^s)] \tag{12}$$

训练时使用蒙特卡罗方法从概率分布 p_θ 中采样得到的句子序列 y^s可以近似得到期望梯度，对于每个训练样本有：

$$\nabla_\theta L(\theta) \approx -r(y^s) \nabla_\theta \log p_\theta(y^s) \tag{13}$$

使用链式法则，将损失函数关于参数 θ 的梯度表示为

$$\nabla_\theta L(\theta) = \sum_{t=1}^{T} \frac{\partial L(\theta)}{\partial z_t} \frac{\partial z_t}{\partial \theta} \tag{14}$$

其中 z_t 表示 softmax 层的输入。

将模型通过贪婪解码方法得到的句子 $y = \{y_1, y_2, \cdots, y_T\}$ 在图像描述指标上的得分 $r(y)$ 作为基线，可以得到：

$$\frac{\partial L(\theta)}{\partial z_t} \approx (r(y^s) - r(y))(p_\theta(y_t|h_t) - 1_{y_t^s}) \tag{15}$$

由式（1）可得，当采样所得句子的得分高于基线时，模型会向增大生成该类型句子概率的方向调整参数，若低于基线，模型会向降低生成该类型句子概率的方向调整参数。这种训练方法可使模型在训练时通过采样得到的结果优于测试测时通过贪婪解码方式得到的结果，不仅能减少梯度的方差，而且使模型的训练更加稳定。

3 实验设置

3.1 数据集预处理

本文使用 COCO[34] 数据集，采用 Karpathy 分割方法[35] 将数据集分为训练集、验证集和测试集，选取 113 287 张图像和对应的人工标注描述作为训练集，分别选取 5 000 张图像和对应的人工标注描述作为验证集和测试集。

训练模型之前需要先对数据集中的图像和人工标注描述进行预处理。对于人工标注描述，先将其中所有单词转换成小写的形式，使用空格替代标点符号，然后统计所有单词出现的次数，将出现频率超过 5 次的单词构成单词表，并使用<UNK>替换出现频率小于 5 次的单词，以避免罕见单词不利于描述文本的生成，最终得到的单词表包含 9 487 个单词。

3.2 模型训练

实验中使用的图像全局特征和目标对应的局部特征的维度均设置为 2 048，词向量的维度设置为 512，长短期记忆网络隐藏层的维度设置为 512，生成的描述文本最大长度设置为 16。

模型训练使用的批量大小设置为 64，最大迭代周期设置为 30。使用 Adam 优化算法[36] 作为训

练的优化器, 参数设置为 $a = 0.9$, $\beta = 0.999$, $\varepsilon = 10^{-8}$。学习率的初始值设置为 0.004, 为了加快模型收敛使得模型更接近最优解, 在第 6 轮之后, 学习率的初始值每 3 轮减小为原来的 0.8 倍。模型在测试时解码器使用 Beam Search[37] 方法进行解码, 集束大小设置为 5, 然后从生成的描述结果中选取在 CIDEr 指标上得分最大的句子作为最终的描述语句。

4　实验结果分析

4.1　不同注意力机制对比实验结果分析

将基于注意力融合的图像描述模型 SA-Attention 与当前主流的几种图像描述模型在 COCO 数据集上的生成效果进行对比, 对比的图像描述模型包括 Xu 等[19] 提出的基于空间注意力机制 Soft-Attention 模型和 Hard-Attention 模型、You 等[20] 提出的基于高层语义注意机制 Semantic-Attention 模型和 Lu 等[26] 提出的自适应注意力机制 Adaptive-Attention 模型。使用预先训练好的 ResNet-101 模型对注意力机制和解码器的参数进行训练, 将训练好的模型作为基线模型(Baseline)。

为验证基于 Faster R-CNN 改进的空间注意力机制描述模型的生成效果, 将仅使用空间注意力机制的模型记为 Spatial-Attention, 将基于空间注意力和高层语义注意力融合的图像描述模型记为 SA-Attention。表 1 给出了 Baseline 模型、Spatial-Attention 模型和 SA-Attention 模型以及上述几种对比模型在 COCO 上的测试结果, 各指标数值没有单位, 模型在指标上的得分数值越大表示生成的图像描述与数据集中的人工标注描述越相近, 所生成的描述效果越好。

表 1　基于注意力融合的图像描述模型与其他主流模型实验结果对比

Table 1　Comparison of experimental results between the image caption models based on attention fusion and other mainstream models

模型	B_1	B_2	B_3	B_4	M	R	C	S
Baseline	0.741	0.573	0.431	0.323	0.257	0.541	1.019	0.191
Spatial-Attention	0.752	0.590	0.450	0.342	0.263	0.552	1.060	0.196
Soft-Attention	**0.707**	0.592	**0.344**	**0.243**	0.239	—	—	—
Hard-Attention	0.718	**0.504**	0.357	0.250	**0.230**	0.516	0.865	—
Semantic-Attention	0.709	0.537	0.402	0.304	0.243	0.543	1.042	—
Adaptive-Attention	0.742	0.580	0.439	0.332	0.266	0.550	1.037	—
SA-Attention	**0.771**	**0.612**	**0.473**	**0.364**	**0.275**	**0.568**	**1.132**	**0.208**

注: B_1表示BLEU-1[38]、B_2表示BLEU-2、B_3表示BLEU-3、B_4表示BLEU-4、M表示METEOR[39]、R表示ROUGE_L[40]、C表示CIDEr[41]、S表示SPICE[42]。

由表 1 可得, Spatial-Attention 模型在各个指标上的得分都高于 Baseline 模型, 在 CIDEr 指标下, Spatial-Attention 模型得分最高, 这说明基于 Faster R-CNN 改进的空间注意力机制优于当前多个主流的图像描述模型所使用的空间注意力机制, 证明了基于 Faster R-CNN 改进的空间注意力机制的有效性。

从表 1 可以得出, 基于空间注意力和高层语义注意力融合的模型 SA-Attention 在上述指标上的得分均高于空间注意力机制模型 Spatial-Attention, 在 CIDEr 指标下, SA-Attention 模型得分比 Spatial-Attention 模型的得分高出 0.072, 这表明高层语义注意力可以进一步提升模型的效果。与基线模型相比, 基于注意力融合的图像描述模型 SA-Attention 在上述评价指标上的得分均具有显著地提升, 这说明本文所提出的基于空间注意力和高层语义注意力融合的图像描述模型能够有效提升图像描述的准确性。与其他几种主流图像描述模型相比, 基于空间注意力和高层语义注意力融合的模型 SA-Attention 在以上指标上的得分都高于对比的主流图像描述模型, 在 B_1 指标下, SA-Attention 模型得分比 Soft-Attention 模型得分高出 0.064, 在 B_2 指标下, SA-Attention 模型得分比 Hard-Attention 模型得分高出 0.108, 在 B_3 指标下, SA-Attention 模型得分比 Soft-Attention 模型得分高出 0.129, 在 B_4 指标下, SA-Attention 模型得分比 Soft-Attention 模型得分高出 0.121, 在 M 指标下, SA-Attention 模型得分比 Hard-Attention 模型得分高出 0.045, 以上对比结

果表明，使用 Faster R-CNN 改进的将空间注意力和高层语义注意力进行融合的方法能够有效提升图像描述的性能。图 2 给出了使用交叉熵损失函数训练得到的模型 SA-Attention（XE）和使用强化学习优化 CIDEr 评价指标训练得到的模型 SA-Attention（CIDEr）的部分图像描述效果。

SA-Attention (XE):
a dog and a cat are sitting on a table
SA-Attention (CIDEr):
a dog and a cat playing with each other

(a) 示例 1

SA-Attention (XE):
a group of children playing soccer on afield
SA-Attention (CIDEr):
a group of young children playing soccer on a field

(b) 示例 2

SA-Attention (XE):
panda bear sitting on a tree branch
SA-Attention (CIDEr):
two panda bears sitting on the top of a tree branch

(c) 示例 3

SA-Attention (XE):
a man skiing down a snowy hill on skis
SA-Attention (CIDEr):
a woman is skiing down a snowy hill

(d) 示例 4

SA-Attention (XE):
a little girl holding a toothbrush in her mouth
SA-Attention (CIDEr):
a little girl brushing her teeth with a toothbrush

(e) 示例 5

SA-Attention (XE):
a horse drawn carriage with a man standing in front of it
SA-Attention (CIDEr):
a black and white photo of horses pulling drawn carriage

(f) 示例 6

图 2　SA-Attention（XE）模型和 SA-Attention（CIDEr）模型生成图像描述

Fig. 2　Image caption generated by SA-Attention (XE) model and SA-Attention (CIDEr) model

4.2　不同训练方法对比实验结果分析

使用交叉熵损失和强化学习两种训练方法对基于注意力融合的图像描述模型进行训练，将使用交叉熵损失训练得到的模型表示为 SA-Attention（XE）。在强化学习训练过程中首先使用交叉熵损失训练模型，再使用 REINFORCE算法直接优化 CIDEr 评价指标对模型进一步训练，模型记为 SA-Attention（CIDEr）。将使用强化学习方法训练得到的模型与使用交叉熵损失训练得到的模型在 COCO 数据集上进行测试，测试结果如表 2 所示。

表 2　强化学习使用不同指标优化模型的实验结果

Table 2　Experimental results of reinforcement learning uses different indicators to optimization model

模型	B_1	B_2	B_3	B_4	M	R	C	S
SA-Attention（XE）	0.771	0.612	0.473	**0.364**	0.275	0.568	1.132	0.208
SA-Attention（CIDEr）	**0.794**	**0.633**	**0.484**	**0.364**	**0.277**	**0.574**	**1.231**	**0.212**

在表 2 中，对比使用强化学习方法训练模型与使用交叉熵损失训练模型的实验结果，可以看出使用强化学习方法优化 CIDEr 评价指标训练的模型在多数指标上的得分均高于使用交叉熵损失训练的模型，这说明使用强化学习的训练方法可以在评价指标上进一步优化使用交叉熵损失训练方法得到的模型，能够显著地提高图像描述模型的性能，证明了使用强化学习训练方法的有效性。

当使用交叉熵方法将模型训练稳定时，再使用强化学习优化评价指标对模型进一步训练。使用强化学习方法优化 CIDEr 指标训练时，模型在 CIDEr 指标上的得分随训练步数的变化情况如图 3 所示，从 5.3 万次迭代开始使用强化学习方法直

接优化 CIDEr 评价指标对模型进行训练, 模型在 CIDEr 评价指标上的得分有了大幅度的提升。使用强化学习方法在 CIDEr 评价指标上对基于注意力融合的图像描述模型进行优化能够进一步提高模型的性能, 使得模型在各评价指标上的得分均有一定幅度的增加, 这表明了使用强化学习训练方法的有效性。

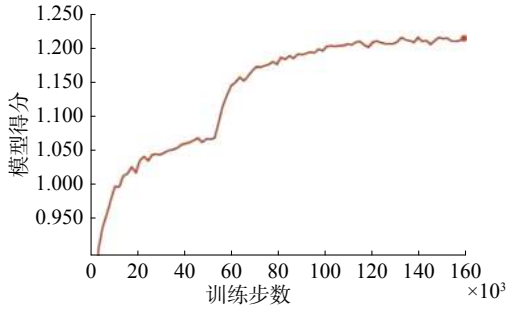

图 3　在 CIDEr 指标上模型得分随训练步数变化曲线

Fig. 3　Score change curve with the number of training steps on the CIDEr indicator

4.3　改进的空间注意力机制运行机制分析

对基于 Faster R-CNN 改进的空间注意力机制进行进一步分析, 在测试时记录生成每个单词的注意力机制权重, 通过可视化的方式还原生成单词时注意力机制的工作过程。图 4 给出了空间注意力机制在生成单词时的可视化过程。

图 4　基于 Faster R-CNN 改进的空间注意力机制的工作过程可视化

Fig. 4　Visualization of the working process of spatial attention mechanism improved based on Faster R-CNN

从图 4 中可以得出, 当生成与图像中目标相关的单词时, 模型会关注到与当前时刻生成单词相关的图像区域, 注意力机制会将目标对应区域的图像特征送入到解码器, 例如, 当生成单词“cake”时, 图像描述模型会选取图像中被检测到的“cake”目标区域的特征。基于 Faster R-CNN 改进的注意力机制能够自适应地选择与当前时刻生成单词相关的目标区域的图像特征来指导单词的生成, 基于 Faster R-CNN 改进的空间注意力机制能够精确地选取与生成单词相关的目标特征, 使生成的单词更切合图像内容, 准确性更高。

5　结束语

本文提出一种基于注意力融合的图像描述模型, 使用 Faster R-CNN 作为编码器来检测提取图像中目标的准确位置和名称属性信息, 将目标的名称属性和位置信息分别作为高层语义注意力与空间注意力来指导单词序列的生成, 以提高图像描述的准确性。使用强化学习方法训练模型, 先使用交叉熵损失训练模型至稳定状态, 再使用 REINFORCE 算法直接优化评价指标对模型进一步训练, 该方法显著地提高了基于注意力融合的图像描述模型的生成效果。

参考文献:

[1] 李亚栋, 莫红, 王世豪. 基于图像描述的人物检索方法 [J]. 系统仿真学报, 2018, 30(7): 377–383.
LI Yadong, MO Hong, WANG Shihao. Person retrieval method based on image caption[J]. Journal of system simulation, 2018, 30(7): 377–383.

[2] WU Jie, XIE Siya, SHI Xinbao, et al. Global-local feature attention network with reranking strategy for image caption generation[J]. Optoelectronics letters, 2017, 13(6): 448–451.

[3] 邓珍荣, 张宝军, 蒋周琴. 融合 word2vec 和注意力机制的图像描述模型 [J]. 计算机科学, 2019, 46(4): 274–279.
DENG Zhenrong, ZHANG Baojun, JIANG Zhouqin. Image description model fusing Word2vec and attention mechanism[J]. Journal of computer science, 2019, 46(4): 274–279.

[4] 陶云松, 张丽红. 基于双向注意力机制图像描述方法研究 [J]. 测试技术学报, 2019, 33(4): 346–351.
TAO Yunsong, ZHANG Lihong. Research on image description method based on bidirectional attentional mechanism[J]. Journal of test and measurement technology, 2019, 33(4): 346–351.

[5] QU Shiru, XI Yuling. Visual attention based on long-short term memory model for image caption generation[C]// 2017 29th Chinese Control and Decision Conference. Chongqing, China, 2017: 4789-4794

[6] XU Jia, EFSTRATIOS G. Guiding the Long-Short Term Memory Model for Image Caption Generation[C]// 2015 IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 2407-2415

[7] JIN Junqi, FU Kun, CUI Runpeng, et al. Aligning where to

see and what to tell: image caption with region-based attention and scene factorization[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 39(12): 2321–2334.

[8] FARHADI A, HEJRATI M. Every picture tells a story: Generating sentences from images[C]//European Conference on Computer Vision. Berlin, Heidelberg, 2010: 15-29.

[9] GIRISH K, VISRUTH P, SAGNIK D, et al. Babytalk: Understanding and generating simple image descriptions[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(12): 2891–2903.

[10] LI Siming, GIRISH K, TAMARA L B, et al. Composing simple image descriptions using web-scale n-grams[C]// Proceedings of the Fifteenth Conference on Computational Natural Language Learning. Portland, Oregon, USA, 2011: 220-228.

[11] POLINA K, VICENTE O, ALEXANDER C, et al. Collective generation of natural image descriptions[C]// Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju: Republic of Korea, 2012: 359-368.

[12] YASHASWI V, ANKUSH G. Generating image descriptions using semantic similarities in the output space[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Washington, USA, 2013: 288-293.

[13] JACOB D, CHENG Hao, FANG Hao, et al. Language models for image captioning: The quirks and what works[J]. arXiv preprint arXiv:1505.01809, 2015.

[14] KAREN S, ANDREW Z. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[15] NAOMI.S ALTMAN. An introduction to kernel and nearest-neighbor nonparametric regression[J]. The american statistician, 1992, 46(3): 175–185.

[16] MAO Junhua, XU Wei, YANG Yi, et al. Explain images with multimodal recurrent neural networks[J]. arXiv preprint arXiv:1410.1090, 2014.

[17] ORIOL V, ALEXANDER T, SAMY B, et al. Show and tell: a neural image caption generator [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 3156-3164.

[18] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735–1780.

[19] XU KELVIN, BA JIMMY, KIROS RYAN, et al. Show, attend and tell: neural image caption generation with visual attention[C]//International Conference on Machine Learning. Lille, France, 2015: 2048-2057.

[20] YOU Quanzeng, JIN Hailin, WANG Zhaowen, et al. Image captioning with semantic attention[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 4651-4659.

[21] CHEN Long, ZHANG Hanwang, XIAO Jun, 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. Honolulu, USA, 2017: 5659-5667.

[22] MINH-THANG L, HIEU P, CHRISTOPHER D. Manning. Effective approaches to attention-based neural machine translation[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015: 1412-1421.

[23] XU K, JIMMY L B. Show, attend and tell: neural image caption generation with visual attention [C]// Proceedings of the 32th International Conference on Machine Learning. Lille, France, 2015: 2048-2057.

[24] MARCO P, THOMAS L, CORDELIA S, et al. Areas of attention for image captioning[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 1242-1250.

[25] LI Linghui, TANG Sheng, DENG Lixi, et al. Image caption with global-local attention[C]//Thirty-First AAAI Conference on Artificial Intelligence. San Francisco, USA, 2017: 4133-4138.

[26] LU Jiasen, XIONG Caiming, DEVI P, 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 Pattern Recognition. Honolulu, USA, 2017: 375-383.

[27] ANDERSON P, HE Xiaodong, 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. Salt Lake, USA, 2018: 6077-6086.

[28] REN Shaoqing, HE Kaiming, GIRSHICK R, et al. Faster r-cnn: towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis & machine intelligence, 2017, 39(6): 1137–1149.

[29] RICHARD S, SUTTON A, BARTO G. Reinforcement learning: an introduction[M]. Cambridge: MIT Press, 1998.

[30] MARC'AURELIO R, SUMIT C, MICHAEL A, et al. Se-

quence level training with recurrent neural networks[J]. arXiv preprint arXiv:1511.06732, 2015.

[31] LIU Siqi, ZHU Zhenhai, YE Ning, et al. Improved image captioning via policy gradient optimization of spider[C]// Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 873-881.

[32] RONALD J. Simple statistical gradient-following algorithms for connectionist reinforcement learning[J]. Machine learning, 1992, 8: 229–256.

[33] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016.

[34] LIN Tsungyi, MICHAEL M, SERGE B, et al. Microsoft coco: common objects in context[C]//European Conference on Computer Vision. Zürich, Switzerland, 2014: 740-755.

[35] ANDREJ K, LI Feifei. Deep visual-semantic alignments for generating image descriptions[J]. IEEE transactions on pattern analysis and machine intelligence, 2016: 664–676.

[36] DIEDERIK P. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv: 1412.6980, 2014.

[37] WANG Pidong, HWEEe T N. A beam-search decoder for normalization of social media text with application to machine translation[C]//Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, Georgia, 2013: 471-481.

[38] KISHORE P, SALIM R, TODD W, et al. BLEU: a method for automatic evaluation of machine translation[C]// Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, USA, 2002: 311-318.

[39] SATANJEEV B, ALON L. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments[C]// Proceedings of the acl workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Michigan, USA, 2005: 65-72.

[40] LIN C, EDUARD H. Automatic evaluation of summaries using n-gram co-occurrence statistics [C]//Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. Edmonton, Canada, 2003: 71-78.

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

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

作者简介:

莫宏伟, 教授, 博士生导师, 主要研究方向为人工智能、类脑计算、智能机器人。承担完成国家自然科学基金、国防预研等项目 17 项, 授权发明专利 7 项。发表学术论文 70 余篇, 出版专著 6 部。

田朋, 博士研究生, 主要研究方向为图像描述、视觉关系检测和场景理解。