基于多重注意结构的图像密集描述生成方法研究

刘青茹 [1, 2] 李 刚 [1, 2] 赵 创 [1, 2] 顾广华 [1, 2] 赵 耀 [3]

摘 要 图像密集描述旨在为复杂场景图像提供细节描述语句. 现有研究方法虽已取得较好成绩, 但仍存在以下两个问题: 1) 大多数方法仅将注意力聚焦在网络所提取的深层语义信息上, 未能有效利用浅层视觉特征中的几何信息; 2) 现有方法致力于改进感兴趣区域间上下文信息的提取, 但图像内物体空间位置信息尚不能较好体现. 为解决上述问题, 提出一种基于多重注意结构的图像密集描述生成方法——MAS-ED (Multiple attention structure-encoder decoder). MAS-ED 通过多尺度特征环路融合 (Multi-scale feature loop fusion, MFLF) 机制将多种分辨率尺度的图像特征进行有效集成, 并在解码端设计多分支空间分步注意力 (Multi-branch spatial step attention, MSSA) 模块, 以捕捉图像内物体间的空间位置关系, 从而使模型生成更为精确的密集描述文本. 实验在 Visual Genome 数据集上对 MAS-ED 进行评估, 结果表明 MAS-ED 能够显著提升密集描述的准确性, 并可在文本中自适应加入几何信息和空间位置信息. 基于长短期记忆网络 (Long-short term memory, LSTM) 解码网络框架, MAS-ED 方法性能在主流评价指标上优于各基线方法.

关键词 图像密集描述, 多重注意结构, 多尺度特征环路融合, 多分支空间分步注意力

引用格式 刘青茹, 李刚, 赵创, 顾广华, 赵耀. 基于多重注意结构的图像密集描述生成方法研究. 自动化学报, 2022, **48**(10): 2537−2548

DOI 10.16383/j.aas.c220093

Dense Captioning Method Based on Multi-attention Structure

LIU Qing-Ru[1, 2] LI Gang[1, 2] ZHAO Chuang[1, 2] GU Guang-Hua[1, 2] ZHAO Yao[3]

Abstract Dense captioning aims to provide detailed description sentences for complex scenes. Although the existing research methods have achieved good results, there are still the following two problems: 1) Most methods only focus on the deep semantic information extracted by the network, and fail to effectively utilize the geometric information in the shallow visual features. 2) Existing methods are dedicated to improving the extraction of contextual information between regions of interest, but the spatial location information of objects in images cannot be well represented. To solve the above problems, this paper proposes a dense captioning generation method based on multiple attention structure-encoder decoder (MAS-ED). MAS-ED effectively integrates image features of multiple resolution scales through a multi-scale feature loop fusion (MFLF) mechanism, and designs a multi-branch spatial step attention (MSSA) at the decoding end to capture the spatial relationship between objects in the image, this enables the method model to generate more accurate dense description text. In this paper, MAS-ED is evaluated on the Visual Genome dataset. The experimental results show that MAS-ED can significantly improve the accuracy of dense captions, and can adaptively add geometric information and spatial location information to the text. Based on the long-short term memory (LSTM) decoding network framework, the performance of the MAS-ED method in this paper outperforms all baseline methods in mainstream evaluation indicators.

Key words Dense captioning, multi-attention structure, multi-scale feature loop fusion (MFLF), multi-branch spatial step attention (MSSA)

Citation Liu Qing-Ru, Li Gang, Zhao Chuang, Gu Guang-Hua, Zhao Yao. Dense captioning method based on multi-attention structure. *Acta Automatica Sinica*, 2022, **48**(10): 2537−2548

收稿日期 2022-02-10 录用日期 2022-05-17
Manuscript received February 10, 2022; accepted May 17, 2022
国家自然科学基金 (62072394), 河北省自然科学基金 (F2021203019), 河北省重点实验室项目 (202250701010046) 资助
Supported by National Natural Science Foundation of China (62072394), Natural Science Foundation of Hebei Province (F2021203019), and Hebei Key Laboratory Project (20225070 1010046)
本文责任编委 陈胜勇
Recommended by Associate Editor CHEN Sheng-Yong
1. 燕山大学信息科学与工程学院 秦皇岛 066004 2. 河北省信息传输与信号处理重点实验室 秦皇岛 066004 3. 北京交通大学信息科学研究所 北京 100044

图像密集描述是基于自然语言处理和计算机视觉两大研究领域的任务, 是一个由图像到语言的跨模态课题. 其主要工作是为图像生成多条细节描述语句, 描述对象从整幅图像扩展到图中局部物体细节. 近年来, 该任务颇受研究者关注. 一方面, 它具

1. School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004 2. Hebei Provincial Key Laboratory of Information Transmission and Signal Processing, Qinhuangdao 066004 3. Institute of Information Science, Beijing Jiaotong University, Beijing 100044

有实际的应用场景[1], 如人机交互[2]、导盲等; 另一方面, 它促进了众多研究任务的进一步发展, 如目标检测[3−4]、图像分割[5]、图像检索[6] 和视觉问答[7] 等.

作为图像描述的精细化任务, 图像密集描述实现了计算机对图像的细粒度解读. 同时, 该任务沿用了图像描述的一般网络架构. 受机器翻译[8] 启发, 目前的图像描述网络[9−11] 大多为编码器−解码器 (Encoder-decoder, ED) 框架, 因此图像密集描述任务也大多基于该传统结构. 该框架首先将卷积神经网络 (Convolutional neural network, CNN) 作为编码器来提取图像视觉信息[12], 得到一个全局视觉向量, 然后输入到基于长短期记忆网络 (Long-short term memory, LSTM)[13] 的解码器中, 最后逐步输出相应的描述文本单词.

基于上述编码−解码框架, 为实现图像区域密集描述, Karpathy 等[14] 试图在区域上运行图像描述模型, 但无法在同一模型中同时实现检测和描述. 在此基础上, Johnson 等[15] 实现了模型的端到端训练, 并首次提出了图像密集描述概念. 该工作为同时进行检测定位和语言描述提出了一种全卷积定位网络架构, 通过单一高效的前向传递机制处理图像, 不需要外部提供区域建议, 并且可实现端到端的优化. 虽然全卷积定位网络架构可实现端到端密集描述, 但仍存在两个问题:

1) 模型送入解码器的视觉信息仅为感兴趣区域的深层特征向量, 忽略了浅层网络视觉信息和感兴趣区域间的上下文信息, 从而导致语言模型预测出的单词缺乏场景信息的指导, 所生成的描述文本缺乏细节信息, 甚至可能偏离图像真实内容.

2) 对于单一图像的某个区域而言, 描述文本的生成过程即为一次图像描述. 图像描述中, 由于网络仅使用单一 LSTM 来预测每个单词, 故解码器未能较好地捕捉到物体间的空间位置关系[16], 从而造成描述文本的句式简单, 表述不够丰富.

为解决上下文场景信息缺失问题, Yang 等[17] 基于联合推理和上下文融合思想提出了一种多区域联合推理模型. 该模型将图像特征和区域特征进行集成, 实现了较为准确的密集描述. 但是提出的上下文信息过于粗糙, 且尚不完整. Yin 等[18] 通过相邻区域与目标区域间的多尺度信息传播, 提出一种上下文信息传递模块. 该模块引入了局部、邻居和全局信息, 从而获取较细粒度的上下文信息. 此外, Li 等[19] 通过目标检测技术揭示了描述区域与目标间的密切关系, 提出一种互补上下文学习架构, 也可实现上下文信息的细粒度获取. 在图像密集描述任务的最新进展中, Shao 等[20] 提出一种基于 Transformer 的图像密集描述网络, 打破了传统的编码−解码框架, 致力于改进 LSTM 网络和关注信息丰富区域. 上述工作在一定程度上解决了上下文场景信息的缺失问题, 但尚未有研究能解决浅层特征信息利用不完全和区域内空间位置信息获取不完备的问题.

为提高图像区域描述的准确性, 本文提出一种基于多重注意结构的图像密集描述生成方法 —— MAS-ED (Multi-attention structure-encoder decoder). 该方法通过构建多尺度特征环路融合 (Multi-scale feature loop fusion, MFLF) 机制, 为解码器提供多尺度有效融合特征, 增加比较细节的几何信息; 并设计多分支空间分步注意力 (Multi-branch spatial step attention, MSSA) 解码器, 通过提取目标间的空间维度信息, 以加强文本中目标间的位置关系描述. 模型训练过程中, MFLF 机制和 MSSA 解码器之间交替优化、相互促进. 实验结果表明, 本文的 MAS-ED 方法在 Visual Genome 数据集上获得了具有竞争力的结果.

1 基于多重注意结构的密集描述

1.1 算法模型

本文提出的基于多重注意结构的密集描述生成方法网络框架如图 1 所示. 模型是一个端到端的网络模型. 据图 1 可知, MAS-ED 模型是基于残差网络和 LSTM 网络的编码−解码架构, 总体可分解为以下几个阶段.

1) 区域视觉特征获取. 选用在 ImageNet 数据集上预训练过的 ResNet-152 网络作为特征提取器, 用来获取含有整幅图像视觉信息的全局视觉向量, 然后将其送入区域建议网络 (Region proposal network, RPN), 得到高质量的区域建议候选框.

2) 上下文信息处理. 通过交并比 (Intersection over union, IoU) 计算两个区域图像块间的交并比分数, 并进行排序. 将分值最高的相邻图像块特征作为当前图像块的上下文特征. 全局特征的获取由全局池化层 (Global pooling layer, GAP) 来完成.

3) 多尺度环路融合特征提取. MFLF 机制会从残差网络的各 Block 层视觉特征中提取各向量上包含的几何信息和语义信息, 然后将其中显著性视觉信息编码进一个和 Block 层视觉特征维度相同的特征向量中. 最后将该向量送入 RPN 层, 以得到含有几何细节和语义信息丰富的多尺度环路融合特征.

4) 空间位置信息提取. 空间分步注意力 (Spatial step attention, SSA) 模块会根据上一解码器当前的隐含层状态, 动态决定从多尺度环路融合特征

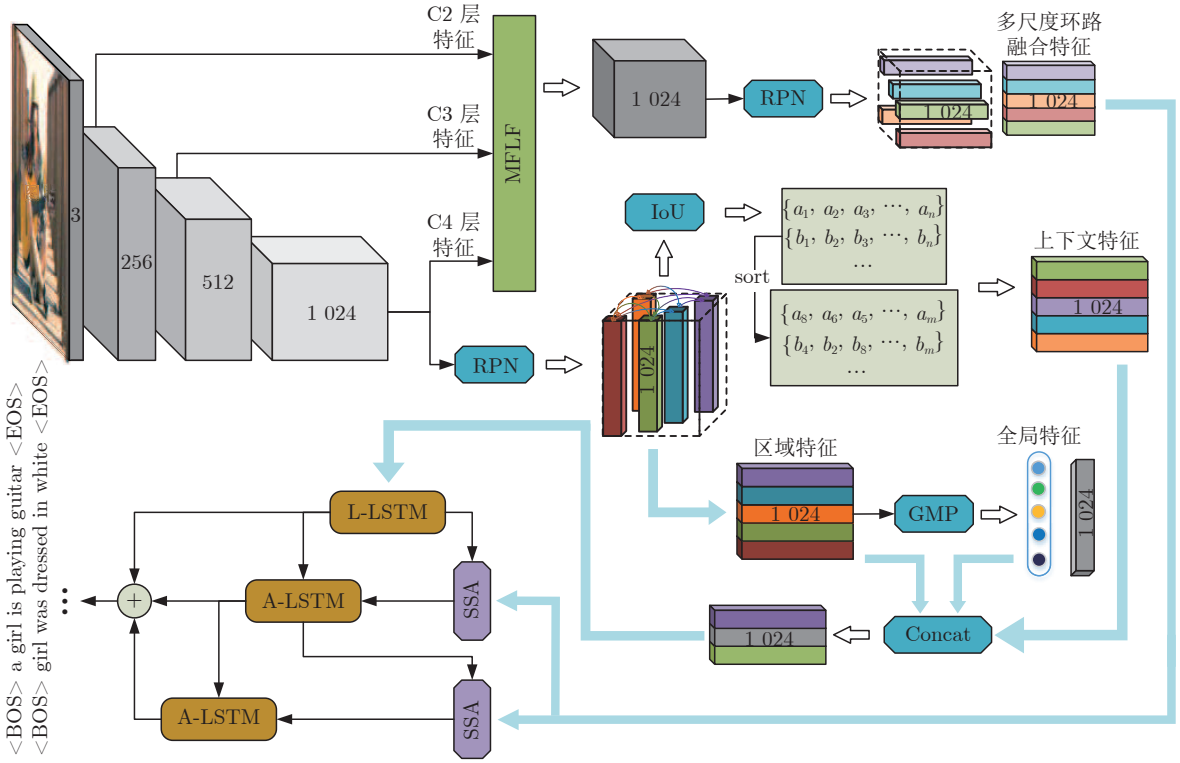


图 1　基于多重注意结构的图像密集描述生成方法

Fig. 1　Dense captioning method based on multi-attention structure

中获取哪些位置信息, 同时决定位置信息在当前单词预测时刻的参与比例, 从而向语言模型提供对预测本时刻单词最有用的位置关系特征.

5) 单词预测. 本文采用表示物体间空间位置关系的注意力特征来引导 LSTM 网络的单词序列建模过程. 图 1 中 L-LSTM 表示 Language-LSTM, 输入的视觉特征由区域特征、上下文特征和全局特征组成; A-LSTM 表示 Attention-LSTM, 输入的视觉特征是注意力引导的多尺度环路融合特征. 为使空间位置信息更好地融入到解码器的输出中, 本文将 SSA 模块和三个 LSTM 网络组成图 1 所示结构, 以形成选择和融合的反馈连接, 并称为多分支空间分步注意力 (MSSA) 解码器.

1.2　多尺度特征环路融合机制

图像密集描述兼具标签密度大和复杂性高两大难点, 其任务网络模型较为庞大. 现有研究方法仅将深层网络特征用于文本生成, 而浅层网络特征并未有效利用. 虽然深层网络特征语义信息表征能力强, 但其特征图分辨率低, 几何信息含量少. 而浅层网络特征的特征图分辨率高, 几何信息表征能力强. 故本文在增加少许网络参数量和计算量的情况下, 提出一种多尺度特征环路融合机制, 即 MFLF 机制, 将同一网络的深层和浅层特征进行多尺度融合, 使模型可更完备地提取出图中含有的几何信息和语义信息. 其结构如图 2 所示.

受到特征金字塔算法[21] 启发, MFLF 机制效仿其实现过程, 改进逐层流向结构, 以减少计算资源开支. MFLF 机制让高层网络输出特征流向低层网络输出特征, 以实现在低层特征图中加权有效的语义信息. 本文将此过程称为语义流, 其实现过程如图 2 中虚线子图框所示. 经几次语义流向过程后, 最底层特征图完成了全部有效语义信息的加权. 为使模型有效利用语义加权优化后低层特征图中的有效几何信息, MFLF 机制设计了从低层特征流向高层的网络结构, 以实现在高层特征图中加权有效几何信息的目的. 此过程称为几何流, 其实现过程如图 2 中实线子图框所示. 需要注意的是, 几何流的初始特征是经语义信息加权后的, 故可削弱冗杂信息的比重. 由图 2 可知, 语义流和几何流构成了闭合回路, 组成了多尺度特征环路融合 (MFLF) 机制.

ResNet-152 网络可分为 4 个 Block, 第 1 个 Block 层的网络层数较少, 其特征图含有较多冗杂信息[22]. 因此在构建 MFLF 机制时, 仅考虑后 3 个 Block 的输出特征, 即图 2 中所示的 C2、C3 和 C4. 此外, 语义流和几何流的组合具有多种可能. 本文

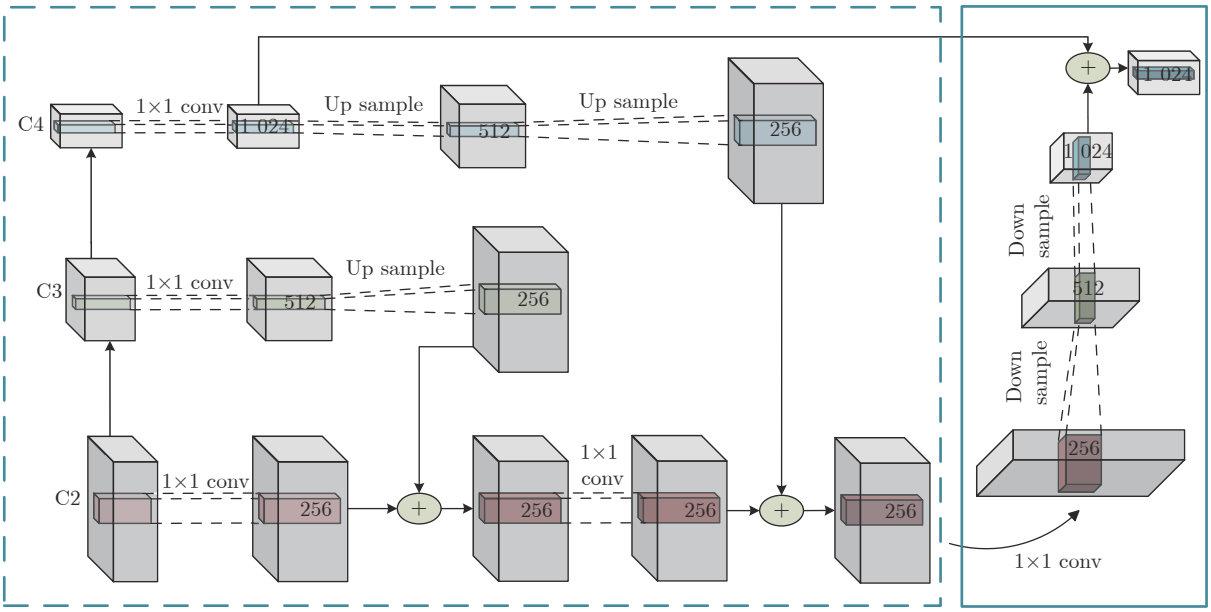


图 2 多尺度特征环路融合机制

Fig. 2 Multi-scale feature loop fusion mechanism

将在消融实验部分阐述如何选择语义流分支和几何流分支. 本文确定的最佳组合为语义流分支选择 C3-C2 和 C4-C2, 几何流分支选择 C2-C4, 其中 C3-C2 表示 C3 层特征信息流向 C2 层, 以此类推.

如图 2 所示, 单条语义流分支实现过程为: 1) 将两个不同尺度的特征图送入 1×1 卷积层, 以保留原有平面结构, 达到通道数统一; 2) 通过上采样将高层特征表示映射到低层特征表示空间; 3) 将上采样后的高层特征与低层特征进行元素级相加操作, 得到融合特征; 4) 将融合特征送入 1×1 卷积层完成通道数调整. 实际操作中, 若残差网络 Block 层输出特征通道数统一, 则不需要完成步骤 1) 和步骤 4). 本文为提高 MFLF 机制的健壮性和可迁移性, 特意增加这两个步骤. 单条几何流分支实现过程同单条语义流分支, 仅将其中的上采样操作更改为下采样操作即可. 最终, MFLF 机制将语义流分支和几何流分支融合形成一组多尺度视觉特征. 随着训练过程中网络参数的逐步优化, 各 Block 层的输出视觉特征也随之优化, 使 MFLF 机制动态调整几何信息和语义信息在输出特征中的比例, 为解码器提供了可动态优化的多尺度融合特征, 从而使模型能够准确生成含有丰富细节的文本描述.

1.3 多分支空间分步注意力解码器

1.3.1 空间分步注意力模块

注意力机制在各个研究领域中得到广泛应用[23–25]. 本文引入注意力机制获取目标位置信息, 并借鉴卷积块注意模块 (Convolutional block attention module, CBAM)[26] 模型方法, 同时考虑通道和空间两个维度, 以获得更好的注意效果. 如图 3 所示, 空间分步注意力模块 (SSA) 的类通道注意力模块 (Channel-like attention module, CLAM) 由维度变换操作和通道注意力模块[27] 共同组成, 且通道注意与空间注意交叉进行.

给定视觉特征 $F \in \mathbf{R}^{H\times W\times C}$ 和预测单词 $\boldsymbol{w} \in \mathbf{R}^{C}$, 其中 H, W, C 分别表示特征图的高、宽和通道. 首先扩充预测单词的空间维度 $S \in \mathbf{R}^{H\times W\times C}$, 并与视觉特征进行元素级加和及非线性 ReLu 函数激活, 得到携带预测单词信息的加和特征图 $F_S \in \mathbf{R}^{H\times W\times C}$:

$$F_S = \text{ReLu}(F + S) \tag{1}$$

由图 3 可知, SSA 模块包含上下两支路, 其作用过程类似. 以上支路为例, 先考虑预测单词在特征图 height 维度的加权, 后考虑 width 维度. SSA 模块将加和特征 F_S 输入 CLAM 中, 得到预测单词在特征图 height 维度的注意力权重图 A^H:

$$A^H = \text{CLAM}(F_S) = \sigma(\text{Maxpool}(f^{\text{T}}(F_S)) + \text{Avgpool}(f^{\text{T}}(F_S))) \tag{2}$$

其中, f^{T} 是维度变换函数, 目的是将特征图空间维度中的 height 维度信息映射到通道维度所在空间. 利用式 (3) 将注意力权重图 A^H 与视觉特征 F 相乘进行自适应特征优化, 得到经预测单词加权 height 维度后的特征矩阵向量 F^H:

$$F^H = \text{Matmul}(F,\ A^H) \tag{3}$$

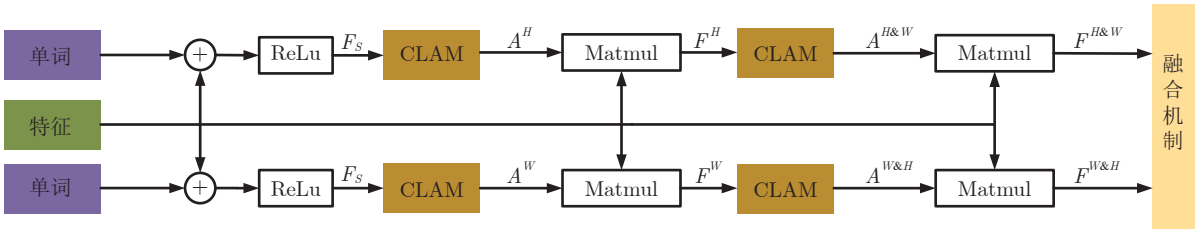


图 3　空间分步注意力模块

Fig. 3　Spatial step attention module

其中, Matmul 函数表示两个矩阵的乘积.

接下来介绍上支路中第二步空间注意, 即考虑预测单词在特征图 width 维度的加权. 将经预测单词加权 height 维度后的特征矩阵 F^H 送入式 (2), 得到预测单词在 width 维度各向量上的注意权重分布图 $A^{H\&W}$ ($H\&W$ 表示先考虑 height 维度, 后考虑 width 维度). 特别注意, 此次 f^{T} 函数是将特征图中的 width 维度信息映射到通道维度所在空间. 由此得到基于预测单词加权特征空间 height, width 两维度的特征图表示:

$$F^{H\&W} = \mathrm{Matmul}(F,\ A^{H\&W}) \tag{4}$$

图 3 中下支路的作用流程与上支路类似, 加和特征 F_S 经式 (2) ~ 式 (4) 操作后, 可得到基于预测单词加权特征空间 width, height 两维度的特征图 $F^{W\&H}$. 最后, SSA 模块将优化后的上、下两分支特征图进行元素级加和, 得到预测单词调整后的视觉特征:

$$F = F^{H\&W} + F^{W\&H} \tag{5}$$

综上, SSA 模块通过结合解码器上一时刻的预测单词, 实现了在空间维度和通道维度的交叉注意, 以加权视觉特征中的位置信息, 并将其用于指导解码器下一时刻的单词预测. 在解码器序列建模过程中, 模型可根据当前单词的预测结果, 完成有选择性地关注视觉特征中的空间位置关系.

1.3.2　多分支解码器

一般来说, 若只将单一 LSTM 网络作为语言模型, 则在本时刻的单词预测仅可根据前几个时刻的信息来推断. 然而, 随着时间轴的不断延长, 解码器较大概率会出现错误累积现象[16]. 因此在当前时刻采用纠正手段来缓解错误累积, 可在一定程度上提高密集描述的准确率. 由第 1.3.1 节可知, SSA 模块可结合解码器上一时刻的预测单词, 来指导下一时刻的单词预测. 基于此, 本文设计如图 4 所示的多分支解码器结构以实现在当前时刻对预测单词的及时纠正. 多分支解码器结构由两个 SSA 模块、一个 L-LSTM 网络和两个 A-LSTM 网络组成. 三个 LSTM 网络的输入构成级联以实现同一时刻的错误纠正, 其输出构成并联以完成本时刻预测单词的反复验证.

三个 LSTM 网络的初始化向量均为局部特征、全局特征及上下文特征的串行连接向量 F_{concat}. 在密集描述文本生成前, 网络初始化过程为:

$$F_{\mathrm{concat}} = \mathrm{concat}(F_{\mathrm{local}},\ F_{\mathrm{global}},\ F_{\mathrm{context}}) \tag{6}$$

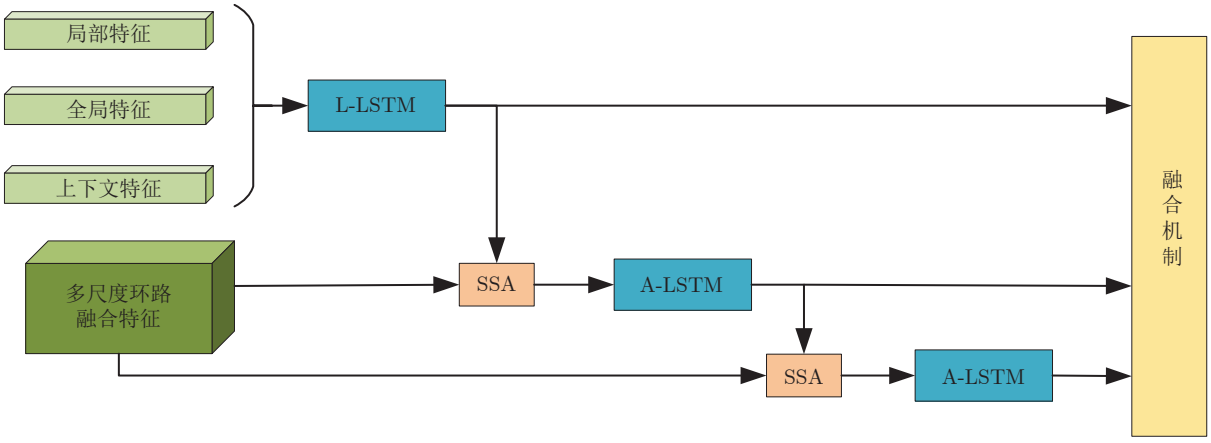


图 4　多分支空间分步注意力模块

Fig. 4　Multi-branch spatial step attention module

$$\begin{cases} h_1^L = \text{L-LSTM}(F_{\text{concat}},\ h_0^L) \\ h_1^{A1} = \text{A-LSTM}(F_{\text{concat}},\ h_0^{A1}) \\ h_1^{A2} = \text{A-LSTM}(F_{\text{concat}},\ h_0^{A2}) \end{cases} \tag{7}$$

其中, F_{local}, F_{global} 和 F_{context} 分别表示描述区域特征, 全局信息特征和上下文信息特征; F_{concat} 表示特征向量的拼接. 在 t 时刻下, 为生成预测单词 y_t, 解码器 L-LSTM 的向量转化如下:

$$h_t^L = \text{L-LSTM}(F_{\text{concat}},\ y_{t-1},\ h_{t-1}^L) \tag{8}$$

其中, h_t^L 代表 L-LSTM 网络在 t 时刻预测的单词向量. 为避免错误累积, 多分支解码器采用两个 A-LSTM 网络对单词向量进行纠正:

$$\begin{cases} F_1 = \text{SSA}(F,\ h_t^L) \\ h_t^{A1} = \text{A-LSTM}(F_1,\ y_{t-1},\ h_{t-1}^{A1}) \\ F_2 = \text{SSA}(F,\ h_t^{A1}) \\ h_t^{A2} = \text{A-LSTM}(F_2,\ y_{t-1},\ h_{t-1}^{A2}) \end{cases} \tag{9}$$

其中, h_t^{A1} 和 h_t^{A2} 表示经过 L-LSTM 解码器一次纠正和二次纠正后的预测单词向量, F_1 和 F_2 表示经 SSA 模块优化后的多尺度环路融合特征. 由此可知, 多分支解码器不仅可实现当前时刻预测单词的及时纠正, 还为单词预测过程引入了几何信息和空间位置信息, 从而使模型生成的描述文本更为精细. 最后, 多分支解码器更新当前隐藏状态 h_t:

$$h_t = \text{Add}(h_t^L + h_t^{A1} + h_t^{A2}) \tag{10}$$

1.4 算法复杂度分析

MAS-ED 方法主要包括多尺度特征环路融合、空间位置注意权重获取和多分支解码器建模几个步骤. 在多尺度特征环路融合中, 由于本文模型无需调整特征图通道数, 因此可去除 MFLF 机制的 1×1 卷积层, 故 MFLF 机制共有 3 次加法运算、3 次上采样和 2 次下采样. 实验中上采样和下采样由双线性插值函数来完成, 因此每个像素点坐标需完成 8 次乘法和 11 次加法运算. 因此 MFLF 机制的乘法运算次数为 $40 \times (w \times h)$, 加法运算次数为 $55 \times (w \times h) + 3$. 新增 8 个输出特征图, 故空间、时间复杂度分别为 $\text{O}(8 \times (w \times h \times C))$、$\text{O}(95 \times (w \times h) + 3)$. 而将同等 $w \times h$ 分辨率的高维特征图送入单个卷积层后, 其时间和空间复杂度可达到 $\text{O}(k^2 \times w \times h \times C_{\text{in}} \times C_{\text{out}})$ 和 $\text{O}(k^2 \times C_{\text{in}} \times C_{\text{out}})$. 由此可知, MFLF 机制增加的计算量和参数量尚不如一个卷积操作.

用 SSA 模块获取空间位置注意权重时, 模型需要完成 3 次加法运算、4 次矩阵乘法运算、2 次 ReLu 非线性变换和 4 次 CLAM 模块. 每个 CLAM 模块包含 2 次池化、2 次 ReLu 变换、4 次卷积和 1 次 Sigmoid 变换. 其中, 仅卷积操作和中间新增特征图涉及空间复杂度计算, 故 SSA 模块增加的参数量为 $\text{O}(k^2 \times C_{\text{in}} \times C_{\text{out}} + w \times h \times C)$, 增加的计算量为 $\text{O}(k^2 \times w \times h \times C_{\text{in}} \times C_{\text{out}} + C + C^2)$. 此外, 构建多分支解码器建模时, 模型仅增加了 1 次加法运算, 可以忽略.

基于编码器−解码器框架下, CAG-Net[18] 方法采用 VGG16 网络进行特征提取, 并将 3 个 LSTM 网络用于文本序列解码; 而 MAS-ED 则采用 ResNet-152 网络, 同样使用 3 个 LSTM 网络用于解码. VGG16 和 ResNet-152 的计算复杂度大致等同[23], 但前者参数量超出后者约 21 MB. 暂不考虑 CAG-Net 所提出的 CFE 和 CCI 这两个模块, 仅基础架构模型的参数量就已超 MAS-ED 所有参数量; 而且两者计算复杂度基本持平. TDC (Transformer-based dense captioner)[20] 模型同样采用参数量较少的 ResNet-152 网络, 但其后端解码网络使用了 Transformer[28]. 与 3 个 LSTM 网络相比, Transformer 网络增加的计算量和参数量相对较大. 综上可知, 相对于 CAG-Net 和 TDC, MAS-ED 虽然增加了 MFLF 机制和 MSSA 解码器两个模块, 但是增加的计算量和参数量均很小.

2 实验与分析

2.1 数据集和评估指标

本文使用标准数据集 Visual Genome 对 MAS-ED 方法进行测试. 该数据集有 V1.0 和 V1.2 两个版本, V1.2 比 V1.0 标注重复率更低, 标注语句也更符合人类习惯. 对数据集的处理同文献 [15], 将出现次数少于 15 的单词换为 <UNK> 标记, 得到一个包含 10 497 个单词的词汇表; 将超过 10 个单词的注释语句去除, 来提高运行效率. 本文的数据划分方式同基线方法, 77 398 张图片用于训练, 5 000 张图片用于验证和测试. 本文基于 V1.0 和 V1.2 两个版本的数据集来验证方法的有效性.

与目标检测任务的平均准确均值 (Mean average precision, mAP) 指标不同, 本文所用的 mAP 指标专门用来评估图像密集描述任务, 由文献 [15] 首次提出. 该指标的计算过程为: 首先, 利用交并比函数 (IoU), 将区域间重叠分值处于 {0.2, 0.3, 0.4, 0.5, 0.6} 的几种精度均值 (Average precision, AP) 作为预测区域性定位的准确性度量; 之后, 使用 METEOR 指标[29] 将语义相似度处于 {0, 0.05, 0.10, 0.15, 0.20, 0.25} 的几种精度均值 (AP), 作为预测文

本和真值标注间的语义相似度度量; 最后, 计算这几组 AP 的平均值作为最终的 mAP 分值.

2.2　实验设置

本文采用文献 [17] 的近似联合训练方法来实现模型的端到端训练, 并使用随机梯度下降来优化模型, 其学习率和迭代数的设置均与基线方法相同. 训练过程中, 图像批大小设为 1, 且每次前向训练中为单个图像生成 256 个感兴趣区域. 实验使用具有 512 个隐藏节点的 LSTM单元, 并将单词序列长度设为 12. 对于测试评估, 将高度重叠的框合并为具有多个参考标题的单个框, 来预处理验证/测试集中的真值标注区域. 具体地, 对于每个图像, 迭代选择具有最多重叠框的框(基于阈值为 0.7 的 IoU), 将它们合并在具有多个标注的单个框中. 之后排除该组, 并重复以上过程.

2.3　MAS-ED 评估

为验证 MAS-ED 方法的有效性和可靠性, 本文选取几种典型的基线方法来完成对比实验. 基线方法根据网络框架分为两组: 基于 LSTM 解码网络框架和基于 Transformer 解码网络框架. 其中, 仅 TDC[20] 模型为基于 Transformer 解码网络框架. 密集描述模型性能由 mAP 分值来评估.

基于 LSTM 解码网络框架下的各模型性能如表 1 所示. 针对 V1.0 数据集, 与 FCLN 相比, MAS-ED 的 mAP 分值提高了 98.01%, 性能提升明显; 与 T-LSTM 和 COCG 相比, MAS-ED 的 mAP 分别提升了 14.64% 和 8.76%. 由于 T-LSTM 和 COCG 模型仅致力于上下文信息的改进, 而 MAS-ED 不仅考虑到上下文关系, 还有效利用浅层特征和空间位置关系, 所以本文 mAP 性能得到有效提升. 与最先进的 CAG-Net 方法相比, 为公平起见, MAS-ED 未使用 ResNet-152 网络而使用 VGG16 网络, 其 mAP 性能仍提升 1.55%. 这表明, MAS-ED 优于 CAG-Net. 针对 V1.2 数据集, MAS-ED 性能同样优于基线方法, 与最先进的 COCG 相比, MAS-ED 获得了 6.26% 的性能优势.

表 2 所示为基于 Transformer 解码网络框架下的模型性能. 由表 2 可见, MAS-ED 方法的 mAP 分值优于 TDC 方法, 在 V1.2 数据集上 mAP 分值达到了 11.04; 而与 TDC + ROCSU 模型相比, MAS-ED 性能稍差. 但 TDC + ROCSU 模型算法复杂度远高于 MAS-ED. 具体来说, TDC + ROCSU 模型选用 Transformer 作为序列解码器, 而本文选用 LSTM 网络, 前者所增加的计算量和参数量远远大于后者; 其次, TDC + ROCSU 模型在使用 ROCSU 模块获取上下文时, 部分网络不能进行 on-line 训练, 无法实现整个网络的端到端训练, 而 MAS-ED 却可实现端到端的网络优化; 最后, TDC + ROCSU 致力于获取准确的文本描述, 而 MAS-ED 不仅考虑文本描述的准确性, 还试图为文本增加几何细节和空间位置关系, 在一定程度上增加了文本的丰富度. 所以相比于 TDC + ROCSU 模型, 本文方法 MAS-ED 算法复杂度低, 可端到端优化且能提高文本丰富性.

表 1　基于 LSTM 解码网络密集描述算法 mAP 性能
Table 1　mAP performance of dense caption algorithms based on LSTM decoding network

模型	V1.0	V1.2
FCLN[15]	5.39	5.16
T-LSTM[17]	9.31	9.96
ImgG[19]	9.25	9.68
COCD[19]	9.36	9.75
COCG[19]	9.82	10.39
CAG-Net[18]	10.51	–
MAS-ED	10.68	11.04

表 2　基于非 LSTM 解码网络密集描述算法 mAP 性能
Table 2　mAP performance of dense caption algorithms based on non-LSTM decoding network

模型	V1.0	V1.2
TDC	10.64	10.33
TDC + ROCSU	11.49	11.90
MAS-ED	10.68	11.04

2.4　消融实验

本文共实现了三种基于注意结构的密集描述模型: 1) 多尺度特征环路融合模型 (MFLF-ED), 使用深、浅层网络的融合特征作为视觉信息, 由标准三层 LSTM 解码; 2) 多分支空间分步注意力模型 (MSSA-ED), 仅使用深层网络特征作为视觉信息, 由多分支空间分步注意力解码器解码; 3) 多重注意结构模型 (MAS-ED), 使用深、浅层网络的融合特征作为视觉信息, 由多分支空间分步注意力解码器解码. 为验证两个模块的有效性, 在相同实验条件下, 本文设置了如表 3 所示的对比实验.

由表 3 可知, 在两种不同网络框架下, MSSA-ED 模型和 MFLF-ED 模型的性能表现均优于基线模型, 这表明浅层细节信息和空间位置信息都利于图像的密集描述. 此外, MSSA-ED 模型要比 MFLF-ED 模型表现更优. 这是因为在 MSSA 解码器中,

表 3 VG 数据集上密集描述模型 mAP 性能
Table 3 mAP performance of dense caption models on VG dataset

模型	VGG16	ResNet-152
Baseline[17]	9.31	9.96
MFLF-ED	10.29	10.65
MSSA-ED	10.42	11.87
MAS-ED	10.68	11.04

SSA 模块通过上一解码器的预测单词指导下一解码器的单词生成时, 模块有额外视觉特征输入, 所以 MSSA-ED 模型除了可获取物体的空间位置信息, 还在一定程度上利用了视觉特征中区域目标的相关信息. 而 MFLF-ED 模型仅使用 MFLF 机制来融合多尺度特征, 增加几何信息, 以此提升小目标的检测精度和增加大目标的描述细节. 因此相对而言, MSSA-ED 模型的改进方法较为多元, 实验效果较好.

此外, MAS-ED 模型性能优于两个单独模型. 这是因为在 MAS-ED 模型训练过程中, MSSA 解码器通过反向传播机制, 促使 MFLF 机制不断调整视觉融合特征中语义信息和几何信息的参与比例; 同时, MFLF 机制通过提供优质融合特征, 来辅助 MSSA 解码器尽最大可能地获取区域实体间的空间位置关系. 最后, 由表 3 可知, 基于 ResNet-152 的三个消融模型性能比基于 VGG16 更优越. 说明密集描述模型不仅需要具有几何细节的浅层特征, 也需要包含丰富语义的深层特征, 从而也证明本文将深层残差网络 ResNet-152 作为特征提取网络的正确性.

2.4.1 MFLF-ED

为探索 MFLF 机制的最佳实现方式, 本文设计了不同语义流和几何流支路组合的性能对比实验, 实验结果如表 4 所示. 由 MFLF 机理可知, 语义流的源特征层应为最高的 C4 层, 以保证最优的语义信息可流向低层特征图; 其目的特征层应为最低的 C2 层, 以确保较完整的几何细节可流向高层特征图. 而几何流的源特征层和目的特征层应与语义流相反, 从而几何流和语义流构成环路融合. 语义流有 4 种情况: C4-C2, C4-C3 & C3-C2, C4-C2+(C3-C2), C4-C2+(C4-C3 & C3-C2), 同样几何流有 C2-C4, C2-C3 & C3-C4, C2-C4+(C3-C4) 和 C2-C4+(C2-C3 & C3-C4). 本文将从源特征层直接流向目的特征层的分支 (如 C4-C2) 称为直接流向分支, 而将途经其他特征层的分支 (如 C4-C3 & C3-C2) 称为逐层流向分支.

由表 4 可知, 当语义流和几何流均采用单条直接流向分支 [C4-C2]+[C2-C4] 时, 其性能 (10.530) 优于两者均采用单条逐层流向分支 [C4-C3 & C3-C2]+ [C2-C3 & C3-C4](10.349), 更优于两者均采用逐层流向分支和直接流向分支 [C4-C2+(C4-C3 & C3-C2)]+[C2-C4+(C2-C3 & C3-C4)](7.704). 这是由于直接流向结构可确保源特征图信息完整地融入目的特征图, 而逐层流向结构会造成信息丢失. 此外, 若同时使用两种结构进行信息传播, 由于信息含量过多且较为冗杂, 会造成显著性信息缺失, 从而性能表现最差.

当语义流和几何流均选用单条直接流向分支和部分逐层流向分支 [C4-C2+(C3-C2)]+[C2-C4+(C3-C4)] 时, 其模型性能 (10.504) 虽优于逐层流向结构模型 (10.349), 但劣于直接流向结构模型 (10.530). 为进一步提高模型性能, 本文选择分开考虑语义流和几何流配置. 当语义流选用直接流向分支, 而几何流选用直接流向分支和部分逐层流向分支 [C4-C2]+[C2-C4+(C3-C4)] 时, 其模型性能较差 (9.727). 而当语义流选用直接流向分支和部分逐层流向分支, 几何流选用直接流向分支 [C4-C2+(C3-C2)]+[C2-C4] 时, 其模型性能 (10.654) 要优于直接流向结构模型 (10.530).

除此之外, 由表 4 中前 2 行数据可知, C4 层中的优质语义信息多于 C3 层, C2 层中的几何细节信息也比 C3 层多, 从而进一步证明了 MFLF 机制将 C4 层和 C2 层作为源特征层和目的特征层的正确性.

综上, [C4-C2+(C3-C2)]+[C2-C4] 是 MFLF 机

表 4 不同分支组合模型的 mAP 性能比较
Table 4 Comparison of mAP performance of different branch combination models

语义流	几何流			
	C2-C4	C2-C3 & C3-C4	C2-C4 + (C3-C4)	C2-C4 + (C2-C3 & C3-C4)
C3-C2	9.924	10.245	10.268	7.122
C4-C2	10.530	10.371	9.727	8.305
C4-C3 & C3-C2	10.125	10.349	10.474	10.299
C4-C2+(C3-C2)	10.654	10.420	10.504	10.230
C4-C2+(C4-C3&C3-C2)	10.159	10.242	10.094	7.704

制的最优组合方式. 为了更加直观, 本文将各模型的描述结果可视化如图 5 所示. 当语义流和几何流均采用直接流向和逐层流向的双通路实现时, 由于信息冗杂, 语句中含有的信息量少, 甚至出现错误信息, 如 “A shelf of a shelf”. 当单独采用直接流向或逐层流向时, 语句中含有的语义和几何信息有所提升, 如 “wood”和 “yellow”. 随着网络结构不断优化, 生成语句中的语义信息更抽象, 如 “kitchen room”, 几何信息也更加具体, 如 “many items”.

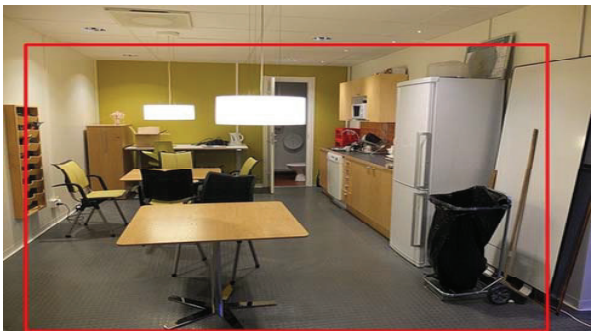
2.4.2 MSSA-ED

1) SSA 模块. 基于相同实验条件下, 本文在模型 MSSA-ED 上对 SSA 模块中上下两分支进行冗余性分析, 实验结果如表 5 所示. 表中 Up-ED 表示仅使用 SSA 模块上支路, 即先考虑预测单词在特征图 height 维度的加权, 后考虑 width 维度; Down-ED 则仅使用 SSA 模块下支路, 维度加权顺序与上支路相反. 由表 5 可知, 两个单支路模型的性能相差不大, 而采用双支路的 MSSA-ED 性能优于两个单支路模型. 这是因为每个支路对两个空间维度 (height 维度和 width 维度) 都进行加权考虑, 加权先后顺序对模型性能影响并不大, 若将上下两支路所得到的加权信息融合, 模型便可获得更加准确的空间位置信息.

各模型的可视化效果如图 6 所示. Up-ED 能检测出 “sign”与 “wall”的左右关系, Down-ED 则捕捉到目标物体与 “refrigerator”的高低关系, 而 MSSA-ED 则通过融合两个位置信息得出最符合真值标注的预测语句.

2) 多分支解码器. 本文通过设计对比实验来确定多分支解码器的支路数, 实验结果见表 6. 其中单支路表示仅添加一条 A-LSTM 通路, 依此类推两支路与三支路表示. 由表 6 可知, 基于三种不同 SSA 模块, 两支路模型的性能都优于单支路模型和三支路模型. 这是因为采用 A-LSTM 对预测单词进行实时纠正时, 过少支路的模型不能在复杂特征信息中准确定位描述目标; 而过多支路的模型, 虽对单目标区域十分友好, 但在多目标区域描述时, 会过度关注每个目标, 导致模型忽略目标间的语义关系.

为了更加直观, 图 7 将基于 MSSA-ED 的三种不同支路模型的注意权重可视化. 图中从左到右依次为原图、单支路注意图、两支路注意图和三支路注意图, 图下方为各模型的预测语句. 其中单支路模型的注意权重分布较分散, 无法准确捕捉到目标; 三支路对单目标注意相对集中, 但对多目标注意权重图成点簇状; 而两支路不仅能突出描述区域内的目标, 并且可关注到区域内目标间的空间位置关系.



[C4-C2]+[C4-C3]+[C3-C2]
[C2-C4]+[C2-C3]+[C3-C4]: A large room.

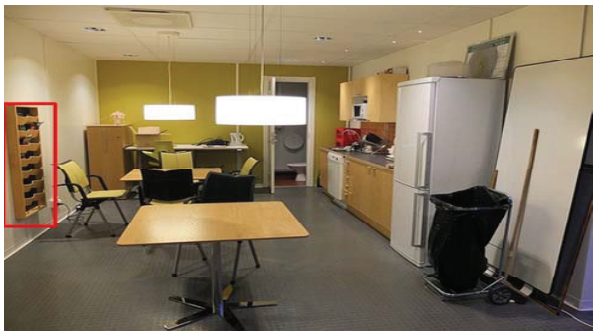
[C4-C2]
[C2-C4]+[C3-C4]: A large yellow room.

[C4-C3]+[C3-C2]
[C2-C3]+[C3-C4]: The scene is in a room.

[C4-C2]+[C3-C2]
[C2-C4]+[C3-C4]: A large kitchen table in the room.

[C4-C2]
[C2-C4]: The scene is in a large room.

[C4-C2]+[C3-C2]
[C2-C4]: The picture of large kitchen room.



[C4-C2]+[C4-C3]+[C3-C2]
[C2-C4]+[C2-C3]+[C3-C4]: A shelf of a shelf.

[C4-C2]
[C2-C4]+[C3-C4]: A shelf in the wall.

[C4-C3]+[C3-C2]
[C2-C3]+[C3-C4]: The shelf is made of wood.

[C4-C2]+[C3-C2]
[C2-C4]+[C3-C4]: A wooden shelf.

[C4-C2]
[C2-C4]: A shelf of a brown frame.

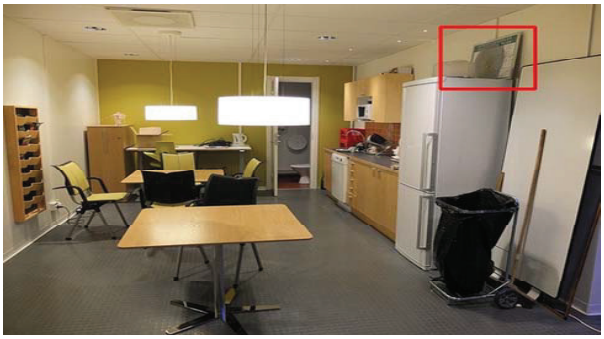
[C4-C2]+[C3-C2]
[C2-C4]: A wooden shelf with many items.

图 5　不同分支组合模型结果可视化 (图中每行上面 “[·]”表示语义流, 下面 “[·]”表示几何流)

Fig. 5　Visualization of results of different semantic flow branching models (The upper “[·]” of each line in the figure represents the semantic flow, and the lower “[·]” represents the geometric flow)

表 5 SSA 模块支路模型的 mAP 性能
Table 5 mAP performance of SSA module branch model

模型	Up-ED	Down-ED	MSSA-ED
mAP	10.751	10.779	10.867



Up-ED: A white sign on the wall.
Down-ED: The top of refrigerator.
MSSA-ED: White paper on top of refrigerator.

图 6 SSA 模块支路模型的结果可视化
Fig. 6 Visualization of results from the SSA module branch model

2.5 可视化分析

为进一步直观表明各个模块实验效果, 图 8 给出了多个密集描述模型的定性表现. 由图中的描述语句可得, MFLF-ED 模型可以描述出灌木丛 “bush” 的 “small” 和 “green”, 建筑物 “building” 和公交车

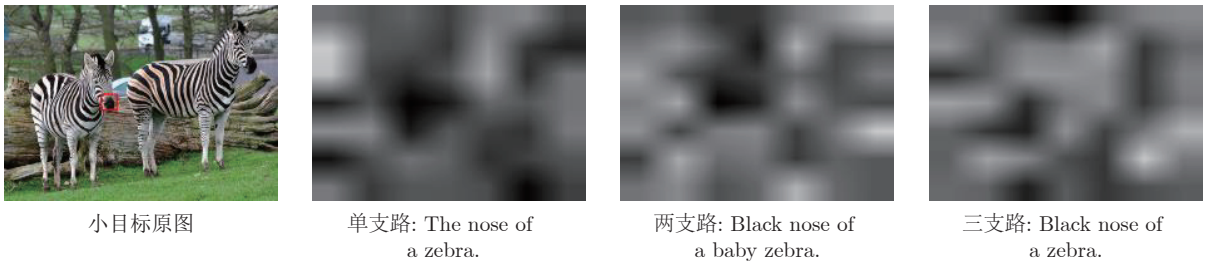
表 6 不同支路数对多分支解码器性能的影响
Table 6 Effects of different branch numbers on the performance of multi-branch decoders

模型	单支路	两支路	三支路
Up-ED	10.043	10.751	10.571
Down-ED	10.168	10.779	10.686
MSSA-ED	10.347	10.867	10.638

“bus” 的颜色 “red” 等细节信息, 说明 MFLF 机制能为密集描述增加有效几何信息, 但描述语句均为简单句, 较少体现物体间的逻辑关系; MSSA-ED 模型能够捕捉到建筑物 “building” 与植物 “plants”、树 “trees” 与大象 “elephant” 间的空间位置关系, 证明 MSSA 解码器能为密集描述获取有效位置关系, 但因缺乏几何细节, 左子图中 “bush” 的信息表述模糊, 采用了广泛的 “plant” 来表述; 而 MAS-ED 模型不仅可检测出灌木丛 “bush”、建筑物 “building” 以及公交车 “bus” 的颜色、大小细节, 而且还在一定程度上能够表达出各物体间的空间位置关系, 如 “side”, “behind” 等.

值得注意的是, MAS-ED 模型的预测语句沿用了 MSSA-ED 中的 “growing on” 词组, 这表明 “bush” 的一种生长状态, 是基准描述语句中未体现的. 类似地, 右子图中的 “beard man” 也没有存在于基准语句中, 这些都体现了 MAS-ED 方法可为密集描述增加丰富度, 能够生成灵活多样的描述语句.

特殊地, 对于大目标物体的细节信息, 如 “build-



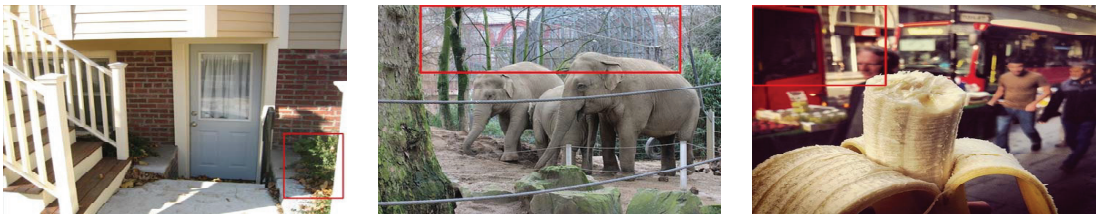
小目标原图 | 单支路: The nose of a zebra. | 两支路: Black nose of a baby zebra. | 三支路: Black nose of a zebra.

(a) 小目标区域的模型性能
(a) Model performance on small target regions

多目标原图 | 单支路: The tree in the background. | 两支路: The trees are behind the zebra. | 三支路: A tree in the background.

(b) 多目标区域的模型性能
(b) Model performance on multi-target regions

图 7 注意图可视化
Fig. 7 Attentional map visualization



MFLF-ED:	A small green bush.	A red building in the background.	The bus is red.
MSSA-ED:	Green plants growing on the side of the building.	The bare trees behind the elephant.	A red bus in the background.
MAS-ED:	A green bush growing on the side of the building.	A red brick building behind the elephant.	The window of red bus behind a beard man.
GT:	Small green bush on the side of the building.	A red and white building behind the elephant.	The window of the red bus on the street.

图 8　图像密集描述模型的定性分析

Fig. 8　Qualitative analysis of image dense captioning model

ing", MAS-ED 模型指出了该物体的颜色 "red" 和组成 "brick". 但 GT 和 MFLF-ED 模型的语句中仅体现了颜色这一细节, 因此 "brick" 是 MAS-ED 模型自适应添加的几何细节, 且该几何细节完全符合图中物体. 此外, MAS-ED 还一定程度上增加了小目标物体的精确检测, 如 GT 语句中未体现 "beard man". 该目标是 MAS-ED 模型在描述语句中自适应增加的, 并且由图 8 可知当前描述区域中的确含有这一目标. 此外, 图 8 中间子图的密集描述语句体现了 MAS-ED 模型可自适应加入位置信息. 在该子图中, MSSA-ED 模型捕捉到了 "tress" 与 "elephant" 间的位置关系, 但 MAS-ED 模型中却未体现, 而是指出了 "building" 与 "elephant" 间的关系. 这是由于 MAS-ED 模型经训练后, 有选择地筛选出了最为突出的目标间位置信息.

3　结论

本文提出了一种基于多重注意结构的图像密集描述生成方法, 该方法通过构建一个多尺度特征环路融合机制, 为文本描述增加了较为细节的几何信息; 并设计了多分支空间分步注意力解码器, 以加强描述目标间的空间位置关系. 实验结果表明, 基于 LSTM 解码网络框架, 本文 MAS-ED 方法的性能优于其他图像密集描述方法.

References

1　Miao Y Q, Lin Z J, Ma X, Ding G G, Han J G. Learning transformation-invariant local descriptors with low-coupling binary codes. *IEEE Transactions on Image Processing*, 2021, **30**: 7554−7566

2　Khavas Z R, Ahmadzadeh S R, Robinette P. Modeling trust in human-robot interaction: A survey. In: Proceedings of the 2020 International Conference on Social Robotics. Berlin, Germany: Springer, 2020. 529−541

3　Cao J L, Pang Y W, Han J G, Li X L. Hierarchical regression and classification for accurate object detection. *IEEE Transactions on Neural Networks and Learning Systems*, DOI: 10.1109/TNNLS.2021.3106641

4　Jiang Hong-Yi, Wang Yong-Juan, Kang Jin-Yu. A survey of object detection models and its optimization methods. *Acta Automatica Sinica*, 2021, **47**(6): 1232−1255
(蒋弘毅, 王永娟, 康锦煜. 目标检测模型及其优化方法综述. 自动化学报, 2021, **47**(6): 1232−1255)

5　Chu Jun, Shu Wen, Zhou Zi-Bo, Miao Jun, Leng Lu. Combining semantics with multi-level feature fusion for pedestrian detection. *Acta Automatica Sinica*, 2022, **48**(1): 282−291
(储珺, 束雯, 周子博, 缪君, 冷璐. 结合语义和多层特征融合的行人检测. 自动化学报, 2022, **48**(1): 282−291)

6　Xu X, Wang T, Yang Y, Zuo L, Shen F M, Shen H T. Cross-modal attention with semantic consistence for image-text matching. *IEEE Transactions on Neural Networks and Learning Systems*, 2020, **31**(12): 5412−5425

7　Bao Xi-Gang, Zhou Chun-Lai, Xiao Ke-Jing, Qin Biao. Survey on visual question answering. *Journal of Software*, 2021, **32**(8): 2522−2544
(包希港, 周春来, 肖克晶, 覃飙. 视觉问答研究综述. 软件学报, 2021, **32**(8): 2522−2544)

8　Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv: 1409.0473, 2016.

9　Kulkarni G, Premraj V, Ordonez V, Dhar S, Li S M, Choi Y, et al. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, **35**(12): 2891−2903

10　You Q Z, Jin H L, Wang Z W, Fang C, Luo J B. Image captioning with semantic attention. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. NewYork, USA: IEEE, 2016. 4651−4659

11　Wang Xin, Song Yong-Hong, Zhang Yuan-Lin. Salient feature extraction mechanism for image captioning. *Acta Automatica Sinica*, 2022, **48**(3): 745−756
(王鑫, 宋永红, 张元林. 基于显著性特征提取的图像描述算法. 自动化学报, 2022, **48**(3): 745−756)

12　Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. NewYork, USA: IEEE, 2014. 580−587

13　Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Computation*, 1997, **9**(8): 1735−1780

14　Karpathy A, Li F F. Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. NewYork, USA: IEEE, 2015. 3128−3137

15　Johnson J, Karpathy A, Li F F. Densecap: Fully convolutional

localization networks for dense captioning. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2016. 4565−4574

16 Jia X, Gavves E, Fernando B, Tuytelaars T. Guiding the long-short term memory model for image caption generation. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. New York, USA: IEEE, 2015. 2407−2415

17 Yang L J, Tang K, Yang J C, Li L J. Dense captioning with joint inference and visual context. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2017. 2193−2202

18 Yin G J, Sheng L, Liu B, Yu N H, Wang X G, Shao J. Context and attribute grounded dense captioning. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2019. 6241−6250

19 Li X Y, Jiang S Q, Han J G. Learning object context for dense captioning. In: Proceedings of the 2019 AAAI Conference on Artificial Intelligence. Menlo Park, California: AAAI, 2019. 8650−8657

20 Shao Z, Han J G, Marnerides D, Debattista K. Region-object relation-aware dense captioning via transformer. *IEEE Transactions on Neural Networks and Learning Systems*, DOI: 10.1109/TNNLS.2022.3152990

21 Lin T Y, Dollár P, Girshick R, He K M, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2017. 2117−2125

22 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2016. 770−778

23 Lu J S, Xiong C M, Parikh D, Socher R. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2017. 375−383

24 Anderson P, He X D, Buehler C, Teney D, Johnson M, Gould S, et al. Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the 2018 IEEE Conference on Computer Vsion and Pattern Recognition. New York, USA: IEEE, 2018. 6077−6086

25 Zhang Z Z, Lan W J, Zeng W J, Jin X, Chen Z B. Relation-aware global attention for person re-identification. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2020. 3183−3192

26 Woo S, Park J, Lee J Y, Kweon I S. Cbam: Convolutional block attention module. In: Proceedings of the 2018 European Conference on Computer Vision. Berlin, Germany: Springer, 2018. 3−19

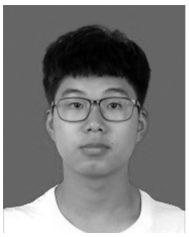
27 Hu J, Shen L, Albanie S, Sun G, Wu E H. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, **42**(8): 2011−2023

28 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. Attention is all you need. In: Proceedings of the 2017 Advances in Neural Information Processing Systems. California, USA: Curran Associates Inc, 2017. 6000−6010

29 Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the 2005 ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Stroudsburg, PA, USA: ACL, 2005. 65−72

刘青茹 燕山大学信息科学与工程学院硕士研究生. 2019 年获得中北大学学士学位. 主要研究方向为图像语义描述.
E-mail: ysu_lqr@163.com
(**LIU Qing-Ru** Master student at the School of Information Science and Engineering, Yanshan University. She received her bachelor degree from North China University in 2019. Her main research interest is image semantic description.)

李 刚 燕山大学信息科学与工程学院副教授. 2009 年获得燕山大学电路与系统专业博士学位. 主要研究方向为图像语义分类, 模式识别.
E-mail: lg@ysu.edu.cn
(**LI Gang** Associate professor at the School of Information Science and Engineering, Yanshan University. He received his Ph.D. degree in circuits and systems from Yanshan University in 2009. His research interest covers image semantic classification and pattern recognition.)



赵 创 燕山大学信息科学与工程学院硕士研究生. 2020 年获得燕山大学学士学位. 主要研究方向为跨模态检索. E-mail: zhaocccchuang@163.com
(**ZHAO Chuang** Master student at the School of Information Science and Engineering, Yanshan University. He received his bachelor degree from Yanshan University in 2020. His main research interest is cross-modal retrieval.)

顾广华 燕山大学信息科学与工程学院教授. 2013 年获得北京交通大学信号与信息处理专业博士学位. 主要研究方向为图像理解, 图像检索. 本文通信作者.
E-mail: guguanghua@ysu.edu.cn
(**GU Guang-Hua** Professor at the School of Information Science and Engineering, Yanshan University. He received his Ph.D. degree in signal and information processing from Beijing Jiaotong University in 2013. His research interest covers image understanding and image retrieval. Corresponding author of this paper.)

赵 耀 北京交通大学信息科学研究所教授. 1996 年获得北京交通大学信号与信息处理专业博士学位. 主要研究方向为多媒体技术.
E-mail: yzhao@bjtu.edu.cn
(**ZHAO Yao** Professor at the Institute of Information Science, Beijing Jiaotong University. He received his Ph.D. degree in signal and information processing from Beijing Jiaotong University in 1996. His main research interest is multimedia technology.)