计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(5):245−251 [doi: 10.15888/j.cnki.csa.007396]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于双路细化注意力机制的图像描述模型①

丛璐文

(中国石油大学 (华东) 计算机科学与技术学院, 青岛 266580)
通讯作者: 丛璐文, E-mail: congve1@live.com

摘　要: 图像描述是连接计算机视觉与自然语言处理两大人工智能领域内的一项重要任务. 近几年来, 基于注意力机制的编码器-解码器架构在图像描述领域内取得了显著的进展. 然而, 许多基于注意力机制的图像描述模型仅使用了单一的注意力机制. 本文提出了一种基于双路细化注意力机制的图像描述模型, 该模型同时使用了空间注意力机制与通道注意力机制, 并且使用了细化图像特征的模块, 对图像特征进行进一步细化处理, 过滤掉图像中的冗余与不相关的特征. 我们在 MS COCO 数据集上进行实验来验证本文模型的有效性, 实验结果表明本文的基于双路细化注意力机制的图像描述模型与传统方法相比有显著的优越性.

关键词: 图像描述; 空间注意力; 通道注意力; 长短时记忆网络; 计算机视觉

引用格式: 丛璐文.基于双路细化注意力机制的图像描述模型.计算机系统应用,2020,29(5):245–251. http://www.c-s-a.org.cn/1003-3254/7396.html

Image Captioning Based on Dual Refined Attention

CONG Lu-Wen

(College of Computer Science and Technology, China University of Petroleum, Qingdao 266580, China)

Abstract: Image captioning is an important task, which connects computer vision and natural language processing, two major artificial intelligence fields. In recent years, encoder-decoder frameworks integrated with attention mechanism have made significant process in captioning. However, many attention-based methods only use spatial attention mechanism. In this study, we propose a novel dual refined attention model for image captioning. In the proposed model, we use not only spatial attention but also channel-wise attention and then use a refine module to refine the image features. By using the refine module, the proposed model can filter the redundant and irrelevant features in the attended image features. We validate the proposed model on MSCOCO dataset via various evaluation metrics, and the results show the effectiveness of the proposed model.

Key words: image captioning; spatial attention; channel-wise attention; Long Short Term Memory (LSTM); computer vision

1　引言

图像描述是计算机视觉领域与自然语言处理领域交叉的一项基本任务, 该任务是给定一张图像, 产生一句对应的自然语言描述, 并且具有广泛的应用, 例如为视觉有障碍的人提供帮助, 人机交互和视觉助手等. 然而, 用自然流畅的句子描述图像内容对机器来说是一项具有挑战性的任务. 它要求图像描述模型不仅识别图像中的显著对象, 而且识别这些对象之间的关系, 并使用自然语言来表达语义信息. 随着深度学习的兴起, 基于深度学习的图像描述模型逐渐发展起来. 但是目前的大部分图像描述方法都只采用了单一的注意力机制, 并且图像特征中存在冗余和不相关的信息, 这些信

① 收稿时间: 2019-10-07; 修改时间: 2019-11-07; 采用时间: 2019-11-18; csa 在线出版时间: 2020-05-07

息会误导注意力计算过程, 使解码器生成错误的句子. 本文针对上述问题, 提出了一种新的基于双路细化注意力机制的图像描述模型, 该模型首先使用 Faster R-CNN[1]目标检测算法提取图像区域特征, 然后使用空间注意力机制关注包含显著对象的区域, 同时利用通道注意力机制关注显著的隐藏单元, 该隐藏单元包含与预测单词更相关的语义信息. 在计算注意力权重时, 首先对解码器的隐藏状态应用卷积运算来过滤掉不相关的信息. 其次, 将经过注意力机制的特征输入到特征细化模块过滤掉其中的冗余信息, 并将这些细化的特征合并到模型中. 这样, 这些特征在语义上与图像内容更加相关.

2 相关工作

近年来, 深度学习取得了重大进展, 研究者们提出了多种基于深度学习的图像描述模型. Vinyals 等[2]提出了基于编码器-解码器的图像描述模型, 该模型借鉴了机器翻译中常用的编码器-解码器架构, 与机器翻译不同的是, 该模型使用卷积神经网络 (Inception 网络模型[3]) 作为编码器提取图像特征, 使用长短时记忆网络 (LSTM)[4]作为解码器生成句子. 但是, 该模型仅在第一步使用图像特征, 而在随后的生成步骤中不使用图像特征. Wu 等[5]首先利用经过微调的多标签分类器来提取图像中的属性信息, 作为指导信息来指导模型生成描述, 提高了性能. Yao 等[6]首先利用经过多示例学习方法预训练的卷积神经网络提取图像中的属性信息, 同时使用卷积神经网络提取图像特征, 并且设计了 5 种架构来找出利用这两种表示的最佳方式以及探索这两种表示之间的内在联系.

强化学习的相关方法也被引入图像描述任务中. Ranzato 等[7]提出了一种直接优化模型评价标准的方法, 该方法利用了策略梯度方法来解决评价标准不可微且难以应用反向传播的问题. 通过使用蒙特卡罗采样方法来估计预期的未来回报, 该模型使得训练阶段更加高效和稳定. Rennie 等[8]提出了一种 SCST 训练方法, 该方法基于策略梯度强化学习算法, 并且使用模型自身解码生成的描述作为基准, 提高了训练过程的稳定性, SCST 训练方法显著地提高了图像描述模型的性能并且在一定程度上解决了图像描述模型训练阶段与测试阶段不匹配的问题.

受人类视觉系统中存在的注意力机制的启发, Xu 等[9]首次将注意力机制引入到图像描述模型中. 在解码阶段的每个时刻, 模型会根据解码器的隐藏状态来计算图像不同位置特征的权重. 这些权重衡量了图像区域和下一个生成的单词之间的相关性. You 等[10]提出了一种新的语义注意机制, 该方法首先会提取出图像的属性信息, 在模型生成描述的每个时刻, 选择最终要的属性信息为模型提供辅助信息. Lu 等[11]提出了一种自注意力机制, 该机制利用哨兵位置的概念, 当模型生成与图像内容无关的单词时, 会将注意力放在哨兵位置上, 以提高模型生成描述的准确性. Chen 等[12]提出了结合空间注意力与通道注意力的图像描述模型, 与之相比, 本文使用的是经过细化的空间注意力与通道注意力, 同时本文还使用 Faster R-CNN提取空间区域特征, 特征更加细化.

3 模型

如图 1 所示, 本文模型包含 5 个基本组件: 编码器、空间注意力机制、通道注意力机制、特征细化模块和解码器. 模型的整个流程如图 2 所示. 首先, 编码器使用 Faster R-CNN 目标检测算法提取图像区域特征. 然后, 在每个时刻, 空间注意力机制与通道注意力机制分别计算对应的特征权重, 特征细化模块通过过滤冗余和不相关的图像特征来细化经过权重修正的空间图像特征和通道图像特征. 在经过细化的图像特征的指导下, 解码器在每个时刻生成一个单词.

3.1 编码器

本文使用 Faster R-CNN 目标检测算法提取图像区域特征. Faster R-CNN 引入了区域建议网络 (Region Proposal Network, RPN), 提高了目标检测的准确率. 首先将图像输入到卷积神经网络中, 将高层卷积特征输入到 RPN 中得到建议区域, 然后再对建议区域与高层卷积特征共同使用感兴趣区域池化, 得到大小相同的特征图 (14×14), 然后将这些特征图输入到另一个卷积神经网络中, 将得到的特征经过平均区域池化即可得到对应的区域特征, 最后利用非极大值抑制过滤掉置信度不高的区域. 最终可以得到 L 个不同区域的特征, 将这些特征集合到一起, 记作 A, 如式 (1) 所示. 每个区域的特征包含 D 个通道.

$$A = \{a_1, \cdots, a_L\},\ a_i \in \mathbb{R}^D \tag{1}$$

全局图像特征可以用局部特征的平均来近似, 如式 (2) 所示.

$$a^g = \frac{1}{L}\sum_{i=1}^{L} a_i \tag{2}$$

随后, 将局部图像特征与全局图像特征分别输入到单层感知机内, 并且使用 $ReLU$ 作为激活函数, 将这些特征投影到维度 d 的空间中.

$$q_i = ReLU(W_a a_i) \tag{3}$$

$$q^g = ReLU(W_b a^g) \tag{4}$$

式中, W_a与W_b是待学习参数, L 个区域图像特征组成局部图像特征$Q = \{q_i, \cdots, q_L\}$.

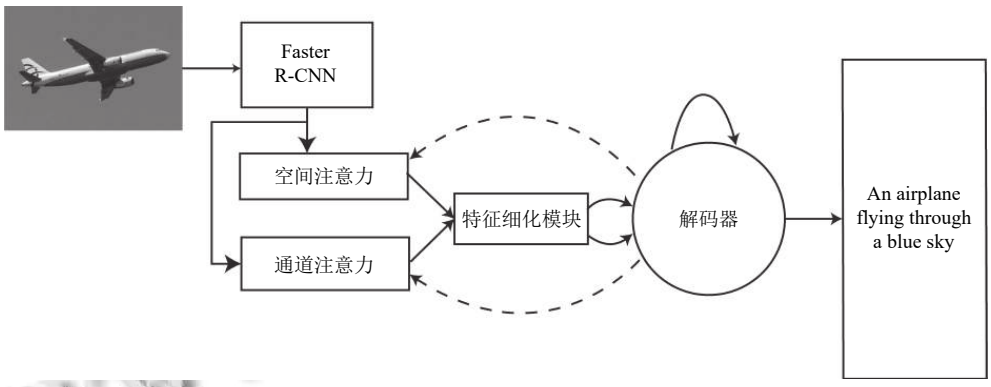

图 1 整体框架

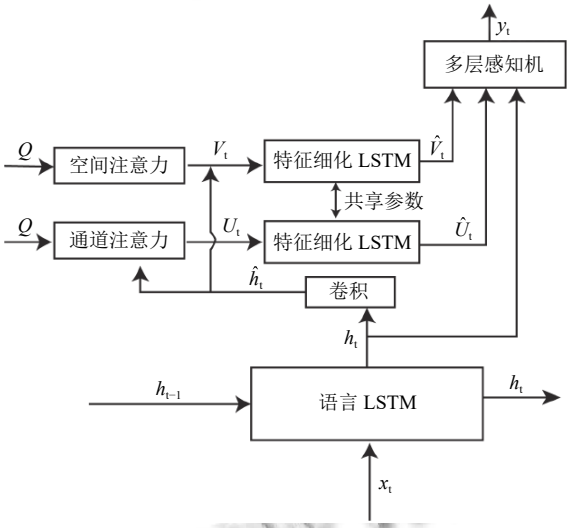

图 2 解码器结构

3.2 空间注意力模型

空间注意力机制广泛用于图像描述任务. 遵循编码器-解码器结构的传统模型仅使用全局图像特征. 基于空间注意力机制的模型更加关注图像中的显著区域, 并且能够捕捉显著区域的更多细节. 当生成与图像中物体相关的单词时, 空间注意力模型可以增加其对图像相应区域的权重. 本文模型也采用了空间注意力机制.

如图 2 所示, 给定局部区域特征$Q \in \mathbb{R}^{d\times L}$与解码器的当前时刻的隐藏状态$h_t \in \mathbb{R}^d$, 模型首先对隐藏状态进行卷积操作, 过滤掉其中的不相关的信息, 随后将这些信息输入到单层感知机中, 然后利用 Softmax 函数计算图像中 L 个区域的注意力分布. 计算过程如下列公式所示:

$$\widehat{h}_t = Conv(h_t) \tag{5}$$

$$z_t^s = w_{hs}^{\mathrm{T}} \tanh(W_{qs}Q + (W_{ss}\widehat{h}_t)1^{\mathrm{T}}) \tag{6}$$

$$\alpha_t = Softmax(z_t^s) \tag{7}$$

其中, $Conv$ 是包含一个卷积层的块, 卷积层后面跟随 $ReLU$ 激活函数. 1^{T}是所有元素都为 1 的向量. W_{qs}, $W_{ss} \in \mathbb{R}^{L\times d}$、$w_{hs} \in \mathbb{R}^L$是待学习的权重参数. $\alpha_t \in \mathbb{R}^L$是图像中 L 个区域的注意力分布. 所关注的局部图像特征V_t可以通过以下方式计算:

$$V_t = \sum_{i=1}^{L} \alpha_{ti} q_i \tag{8}$$

与文献[11]相同, 本文也使用解码器的当前时刻隐藏状态而不是上一时刻的隐藏状态来计算对局部图像特征的空间注意力.

3.3 通道注意力模型

Zhou 等[13]发现每个隐藏单元可以与不同的语义概念对齐. 然而, 在基于空间注意力的模型中, 通道特

征是相同的, 忽略了语义差异. 如图 2 所示, 本文同时也采用了通道注意力机制. 将局部区域特征$Q \in \mathbb{R}^{d\times L}$与解码器的当前时刻的经过卷积的隐藏状态$\widehat{h}_t$输入单层感知机中, 随后用 *Softmax* 函数计算局部图像特征在通道上的注意力分布:

$$z_t^c = w_{hc}^{\mathrm{T}}(W_{qc}Q^{\mathrm{T}} + (W_{sc}\widehat{h}_t)1^{\mathrm{T}}) \tag{9}$$

$$\beta_t = Softmax(z_t^c) \tag{10}$$

其中, $w_{hc} \in \mathbb{R}^d$, $W_{qc} \in \mathbb{R}^{d\times L}$, $W_{sc} \in \mathbb{R}^{d\times d}$为待学习的权重参数. 1^{T}是所有元素都为 1 的向量. $\beta_t \in \mathbb{R}^d$是局部图像特征中隐藏单元上的注意力分布. 基于通道注意力的通道局部图像特征U_t可以由式 (11) 计算获得.

$$U_t = \sum_{i=1}^{d}\beta_{ti}Q_i^{\mathrm{T}} \tag{11}$$

其中, Q_i表示每个区域特征中第 i 个通道组成的向量.

在解码生成描述的每个时刻, β_{ti}确定了第 i 个通道特征与生成的下一个单词之间的相关性.

3.4 特征细化模块

通常提取到的图像特征中会包含一些冗余或与生成描述不相关的特征. 为了减少这些特征的影响, 本文设计了一个特征细化模块来细化图像特征, 过滤掉冗余的和不相关的特征. 如图 2 所示, 该模块使用单层 LSTM 作为细化模块. LSTM 被命名为特征细化 LSTM. 在计算关注的局部图像特征V_t和关注的通道图像特征U_t之后, 首先通过单层感知器将这些图像特征投影到相同的维度 d. 然后, 将这些图像特征输入到细化 LSTM, 并通过 n 个时间步长来细化图像特征. 最后, 得到细化的关注空间图像特征和细化的关注通道图像特征:

$$V_t' = W_{vd}V_t \tag{12}$$

$$U_t' = W_{ud}U_t \tag{13}$$

$$h_n^v = f_{\mathrm{LSTM}}(V_t', h_{n-1}^v) \tag{14}$$

$$h_n^u = f_{\mathrm{LSTM}}(U_t', h_{n-1}^u) \tag{15}$$

$$\widehat{V}_t = h_n^v \tag{16}$$

$$\widehat{U}_t = h_n^u \tag{17}$$

其中, $W_{vd} \in \mathbb{R}^{d\times d}$和$W_{ud} \in \mathbb{R}^{d\times L}$是待学习的权重参数. 本文使用共享参数的特征细化 LSTM, 以降低训练过程中的存储成本.

3.5 解码器

LSTM 通常用于现有的图像描述模型中, 因为 LSTM 在对长期依赖关系建模方面具有强大的力量. 本文遵循常用的 LSTM 结构, 基本 LSTM 块中的门控单元和存储单元定义如下:

$$\begin{cases} x_t = [W_e y_{t-1}; q^g], \ \text{for } t \geq 1 \\ f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f) \\ i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \\ o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \\ c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{cx}x_t + W_{ch}h_{t-1} + b_c) \\ h_t = o_t \odot \tanh(c_t) \end{cases} \tag{18}$$

其中, x_t、f_t、i_t、o_t、c_t、h_t分别是时刻 t 的输入向量、遗忘门、输入门、输出门、存储单元和隐藏状态. y_{t-1}是前一个单词的单热向量, 具有字典大小的维度. W_e 是单词嵌入矩阵. $[\cdot;\cdot]$是两个向量的拼接. x_t是词向量和全局图像特征的组合. $\sigma(\cdot)$是 Sigmoid 非线性激活函数, $\tanh(\cdot)$是双曲正切激活函数. $\odot$ 表示元素乘法.

通过使用隐藏状态h_t、细化的关注局部图像特征$\widehat{V}_t$、细化的通道关注局部图像特征$\widehat{U}_t$, 当前时刻生成的单词条件概率分布可由式 (19) 计算.

$$p(y_t|y_1,\cdots,y_{t-1}, I) = Softmax(W_p(h_t + \widehat{U}_t + \widehat{V}_t)) \tag{19}$$

本文训练过程的第一个阶段使用交叉熵损失函数作为目标函数进行训练, 如式 (20) 所示, 第二个阶段使用 SCST 训练方法, 目标函数如式 (21) 所示.

$$L_{XE}(\theta) = -\sum_{t=1}^{T}\log(p_\theta(y_t^*|y_1^*,\cdots,y_{t-1}^*)) \tag{20}$$

$$L_R = -E_{y_{1:T}\sim p_\theta}[r(y_{1:T})] \tag{21}$$

式中,$y_t^*, y_1^*, \cdots, y_{t-1}^*$表示参考描述中的单词序列, $y_{1:T}$表示单词序列$(y_1, y_2, \cdots, y_T)$的缩写

在训练过程中, 将参考描述的单词序列输入到模型中, 可以得到每个时刻预测的单词概率分布, 随后计算目标函数, 进行优化.

在推理过程中, 选择每个时刻概率最大的单词作为生成的单词或者使用集束搜索 (beam search), 每次选择概率最大的前 k 个单词作为候选, 最终输出联合概率最大的描述作为最终的描述结果.

4 实验分析

4.1 实验数据集与评价标准

本文模型在用于图像描述的 MS COCO 数据集[14]

上进行实验. COCO 数据集包含 82 783 张用于训练的图像、40 504 张用于验证的图像和 40 775 张用于测试的图像. 它还为在线测试提供了一个评估服务器. 本文使用文献[15]中的数据划分, 该数据划分中包含 5000 张用于验证的图像, 5000 张用于测试的图像, 其余图像用于训练.

为了验证本文模型生成描述的质量, 并与其他方法进行比较, 本文使用了广泛使用的评价指标, 包括 BLEU[16]、METEOR[17]、ROUGE-L[18]和 CIDEr[19]. 本文使用文献[20]提供的评估工具来计算分数. BLEU 分数衡量生成的句子和参考句子之间的 n-gram 精度. ROUGE-L 分数测量生成的句子和参考句子之间最长公共子序列 (LCS) 的 F-Score. METEOR 评分通过添加生成的句子和参考句子之间的对应关系, 与人类的评价标准更加相关. 与上述指标不同, CIDEr 评分是为图像描述设计的. 它通过计算每个 n-gram 的 TF-IDF 权重来测量生成描述与参考描述之间的一致性.

4.2 实现细节

首先将 COCO 数据集中所有的描述转换成小写并且将描述的最大长度设置为 15. 如果描述的长度超过 15, 则会截断之后单词. 本文过滤掉训练集中出现不到 5 次的所有单词, 并且增加了四个特殊的单词. “< BOS >”表示句子的开头, “< EOS >”表示句子的结尾, “< UNK >”表示未知单词, 而“< PAD >”是填充单词. 经过这样的处理以后, 得到的字典长度为 10 372.

本文将 LSTM 的隐藏单元的数量设置为 512, 随机初始化词嵌入向量, 而不是使用预训练的词嵌入向量. 我们使用 Adam 优化器[21]来训练本文的模型. 在使用交叉熵训练的阶段, 基础学习率设置为5×10^{-4}, 并且使用1×10^{-6}的重量衰减, 批大小设置为 256, 每三轮学习率衰减 0.8 倍. 训练轮次的最大数量被设置为 30. 在 SCST 训练阶段, 选择交叉熵训练阶段 CIDEr 得分最高的模型作为初始模型, 学习率固定为5×10^{-5}, 训练轮次设置成 40. 整个训练过程在一个 NVIDIA TITAN X 图形处理器上需要大约 50 小时. 本文的模型使用 Pytorch 深度学习框架实现.

4.3 实验对比方法介绍

Goole NIC[2]使用编码器-解码器框架, 使用卷积神经网络作为编码器, 使用 LSTM 作为解码器.

Hard-Attention[9]将空间注意力机制引入图像描述模型, 根据解码器的状态动态地为图像不同区域的特征分配权重.

MSM[6]共同利用了图像属性信息与图像全局特征.

AdaAtt[11]使用了自适应注意力机制, 如果要生成的单词与图像内容无关, 则注意力放在一个虚拟的“哨兵”位置上.

文献[22]中的模型使用了视觉属性注意力并且引入了残差连接.

Att2all[8]首次提出并使用了 SCST 训练方法.

SCA-CNN[12]同时使用了空间与通道注意力.

4.4 实验分析

如表 1 所示, 与 SCA-CNN 模型相比, 本文模型使用的双路细化注意力以及空间区域特征对生成图像描述有着更强的指导作用. 相较于只是用单一空间注意力机制的 Hard-Attention 模型、AdaAtt 模型、文献[21]中的模型、Att2all 模型相比, 本文模型使用的双路细化注意力机制, 可以生成更加紧凑, 冗余信息更少的特征, 并且除了在空间位置上施加注意力, 也在通道上施加注意力, 使得模型可以更好地利用与生成描述相关地特征.

表 1 本文模型与经典算法比较

方法	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Goole NIC[2]	66.6	45.1	30.4	20.3	-	-	-
Hard-Attention[9]	71.8	50.4	35.7	25.0	23.0	-	-
MSM[6]	73.0	56.5	42.9	32.5	25.1	53.8	98.6
AdaAtt[11]	74.2	58.0	43.9	33.2	22.6	-	108.5
文献[21]	74.3	57.6	43.1	32.5	26.2	-	102.9
Att2all[8]	-	-	-	34.2	26.7	55.7	114.0
SCA-CNN[12]	71.9	54.8	41.1	31.1	25.0	53.1	95.2
本文算法	78.0	61.7	46.7	34.9	26.9	56.5	117.1

为研究本文中不同模块的有效性, 设计了不同的模型进行比较, 实验结果见表 2. 基准模型为只使用 Faster R-CNN 目标检测算法提取图像区域特征, 不使用注意力机制与特征细化模块, 表中的“X”表示该模型

在基准模型的基础上使用该模块. 从表 2 中可见, 空间注意力机制、通道注意力机制、特征细化模块都可提高模型性能. 同时使用两种注意力机制的模型 3 相较于只使用一种注意力机制的模型 2 与模型 1, 性能有进一步的提高, 证明本文提出的双路注意力机制的有效性. 模型 5、模型 6、本文算法在模型 1、模型 2、模型 3 的基础上增加了特征细化模块, 最终模型性能也有提高, 证明了特征细化模块的有效性.

表 2 本文模型不同模块效果比较

方法	空间注意力	通道注意力	特征细化模块	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
基准模型				76.0	60.0	45.2	33.8	25.8	55.8	111.8
模型 1	X			76.3	60.4	45.7	34.1	26.4	56.0	113.8
模型 2		X		76.2	60.5	45.5	34.2	26.2	55.9	113.9
模型 3	X	X		77.3	61.2	46.3	34.7	26.6	56.3	115.2
模型 4			X	76.2	60.1	45.3	34.0	25.7	55.8	112.2
模型 5	X		X	77.4	61.2	46.3	34.7	26.6	56.3	114.9
模型 6		X	X	77.2	60.9	46.0	34.4	26.5	56.2	114.7
本文算法	X	X	X	78.0	61.7	46.7	34.9	26.9	56.5	117.1

5 结论与展望

本文提出了一种新的基于双路细化注意力机制的图像描述模型. 本文模型整合了空间注意力机制和通道注意力机制. 首先使用卷积运算来过滤隐藏状态的不相关信息, 然后计算注意力. 为了对减少关注图像特征中的冗余和不相关特征的影响, 本文设计了一个特征细化模块来细化关注图像特征, 使关注图像特征更加紧凑和有区分度. 为了验证本文模型的有效性, 我们在 MS COCO 数据集上进行了实验, 实验结果表明, 本文提出模型性能优越.

参考文献

1 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems. 2015. 91–99.

2 Vinyals O, Toshev A, Bengio S, *et al*. Show and tell: A neural image caption generator. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 3156–3164.

3 Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv: 1502.03167, 2015.

4 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]

5 Wu Q, Shen CH, Liu LQ, *et al*. What value do explicit high level concepts have in vision to language problems? Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 203–212.

6 Yao T, Pan YW, Li YH, *et al*. Boosting image captioning with attributes. Proceedings of the 20017 IEEE International Conference on Computer Vision. Venice, Italy. 2017. 4894–4902.

7 Ranzato MA, Chopra S, Auli M, *et al*. Sequence level training with recurrent neural networks. arXiv preprint arXiv: 1511.06732, 2015.

8 Rennie SJ, Marcheret E, Mroueh Y, *et al*. Self-critical sequence training for image captioning. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 7008–7024.

9 Xu K, Ba J, Kiros R, *et al*. Show, attend and tell: Neural image caption generation with visual attention. arXiv preprint arXiv: 1502.03044, 2015. 2048–2057.

10 You QZ, Jin HL, Wang ZW, *et al*. Image captioning with semantic attention. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 4651–4659.

11 Lu JS, Xiong CM, Parikh D, *et al*. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 375–383.

12 Chen L, Zhang HW, Xiao J, *et al*. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 5659–5667.

13 Zhou BL, Bau D, Oliva A, *et al*. Interpreting deep visual representations via network dissection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(9):

2131–2145. [doi: 10.1109/TPAMI.2018.2858759]

14 Lin TY, Maire M, Belongie S, *et al*. Microsoft coco: Common objects in context. Proceedings of European Conference on Computer Vision (ECCV 2014). Zurich, Switzerland. 2014. 740–755.

15 Karpathy A, Li FF. Deep visual-semantic alignments for generating image descriptions. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 3128–3137.

16 Papineni K, Roukos S, Ward T, *et al*. BLEU: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, PA, USA. 2002. 311–318.

17 Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Ann Arbor, MI, USA. 2005. 65–72.

18 Lin CY. Rouge: A package for automatic evaluation of summaries. Proceedings of the Workshop on Text Summarization Branches Out. Barcelona, Spain. 2004. 74–81.

19 Vedantam R, Zitnick CL, Parikh D. Cider: Consensus-based image description evaluation. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 4566–4575.

20 Chen XL, Fang H, Lin TY, *et al*. Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv: 1504.00325, 2015.

21 Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv: 1412.6980, 2015.

22 周治平, 张威. 结合视觉属性注意力和残差连接的图像描述生成模型. 计算机辅助设计与图形学学报, 2018, 30(8): 1536–1542, 1553.