基于显著性特征提取的图像描述算法

王 鑫 [1] 宋永红 [2] 张元林 [2]

摘 要 图像描述 (Image captioning) 是一个融合了计算机视觉和自然语言处理这两个领域的研究方向, 本文为图像描述设计了一种新颖的显著性特征提取机制 (Salient feature extraction mechanism, SFEM), 能够在语言模型预测每一个单词之前快速地向语言模型提供最有价值的视觉特征来指导单词预测, 有效解决了现有方法对视觉特征选择不准确以及时间性能不理想的问题. SFEM 包含全局显著性特征提取器和即时显著性特征提取器这两个部分: 全局显著性特征提取器能够从多个局部视觉向量中提取出显著性视觉特征, 并整合这些特征到全局显著性视觉向量中; 即时显著性特征提取器能够根据语言模型的需要, 从全局显著性视觉向量中提取出预测每一个单词所需的显著性视觉特征. 本文在 MS COCO (Microsoft common objects in context) 数据集上对 SFEM 进行了评估, 实验结果表明 SFEM 能够显著提升基准模型 (baseline) 生成图像描述的准确性, 并且 SFEM 在生成图像描述的准确性方面明显优于广泛使用的空间注意力模型, 在时间性能上也大幅领先空间注意力模型.

关键词 图像描述, 显著性特征提取, 语言模型, 编码器, 解码器

引用格式 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): 735−746

DOI 10.16383/j.aas.c190279

Salient Feature Extraction Mechanism for Image Captioning

WANG Xin[1] SONG Yong-Hong[2] ZHANG Yuan-Lin[2]

Abstract Image captioning is a research direction that combines computer vision and natural language processing. In this paper, a novel saliency feature extraction mechanism (SFEM) is designed to solve several key problems existing in current methods. It can quickly provide the most valuable visual features to the language model before which predict word. And it effectively solves the problems that the existing methods are inaccurate in selecting visual features and time-consuming. SFEM consists of global salient feature extractor and instant salient feature extractor: global salient Feature extractor extracts salient visual features from multiple local visual vectors and integrate these features into a global salient visual vector; the instant salient feature extractor can extract the saliency visual features required at each moment from the global saliency visual vector according to the needs of the language model. We evaluated SFEM on the MS COCO (Microsoft common objects in context) dataset. Experiments show that our SFEM can significantly improve the accuracy of baseline in caption generating. And SFEM is significantly better than the widely used spatial attention model in both the accuracy of generating caption and time performance.

Key words Image captioning, salient feature extract, language model, encoder, decoder

Citation Wang Xin, Song Yong-Hong, Zhang Yuan-Lin. Salient feature extraction mechanism for image captioning. *Acta Automatica Sinica*, 2022, **48**(3): 735−746

图像描述 (Image captioning) 是涉及到计算机视觉和自然语言处理这两个领域的一个重要的研究方向, 主要工作是实现图像到文本的多模态转换[1−3], 需要计算机能够识别图像上的对象, 理解对象的属性、对象之间的关系, 并用人类的语言表达出图像上的内容.

目前常用于图像描述的编码器−解码器 (Encoder-Decoder) 框架最早受启发于机器翻译[4−6], NIC (Neural image caption)[7] 模型作为第一个使用这个框架的图像描述模型, 以卷积神经网络 (Convolutional neural network, CNN) 作为编码器来提取图像上的视觉信息[8−9], 得到一个包含有整幅图像上视觉信息的全局视觉向量, 以单层的长短期记忆网络 (Long-short term memory, LSTM)[10] 作为解码器, 在生成图像描述的初始时刻将全局视觉向量输入 LSTM 网络中, 之后逐步生成图像描述中

收稿日期 2019-04-01 录用日期 2019-09-12
Manuscript received April 1, 2019; accepted September 12, 2019
陕西省自然科学基础研究计划 (2018JM6104), 国家重点研究开发项目 (2017YFB1301101) 资助
Supported by Natural Science Basic Research Program of Shaanxi (2018JM6104) and National Key Research and Development Program of China (2017YFB1301101)
本文责任编委 桑农
Recommended by Associate Editor SANG Nong
1. 西安交通大学软件学院 西安 710049 2. 西安交通大学人工智能学院 西安 710049
1. School of Software Engineering, Xi′an Jiaotong University, Xi′an 710049 2. College of Artificial Inteligence, Xi′an Jiaotong University, Xi′an 710049

的每个单词. 文献 [11]中提出了 g-LSTM (Guiding LSTM) 模型, 它与 NIC 模型最大的不同在于, 不仅将全局视觉向量作为 LSTM 网络的输入, 也将全局视觉向量用来构建 LSTM 网络的各个门, 作者尝试以这种方法来引导 LSTM 生成更加贴合于图像内容的描述. 文献 [12]中使用多标签分类的方法, 对图像进行多标签分类, 从而将图像上的多个高层属性编码进一个 0-1 向量中, 该向量的每一维都对应属性库中的一个属性, 如果图像上具有该属性, 向量对应维度的值取 1 否则取 0, 作者使用该向量代替编码器给出的全局视觉向量, 取得了比较好的效果.

虽然这几种编码器–解码器模型都取得了不错的效果, 但存在两个主要的问题:

1) 包含整幅图像视觉信息的全局视觉向量在初始时刻被输入解码器中, 解码器需要自己从中抽取预测单词所需的视觉信息, 造成解码器负担过重.

2) 作为解码器的 LSTM 网络在预测每个单词时都会接收新的输入并遗忘掉现有的部分信息, 这就造成了随着预测的进行一些重要的视觉信息会被遗忘掉, 从而导致语言模型[13−15] 预测出的单词逐渐缺乏图像上视觉信息的指导, 偏离了图像的真实内容.

在编码器–解码器框架的基础上, 相继提出了多种注意力模型. 文献 [16] 中最早将空间注意力机制引入到图像描述领域, 在预测单词时空间注意力模型能够根据 LSTM 的隐含层状态来为每个局部视觉向量分配不同的权重, 然后通过加权求和得到当前单词所需的视觉向量. 空间注意力模型与编码器–解码器模型的结合, 一定程度上解决了编码器–解码器模型的上述两个问题. 但是同时也产生了 3 个新的问题:

1) 在空间注意力模型中, 每个局部视觉向量只对应一个标量权重, 所以特征向量的每一维都需要乘以相同的权重, 空间注意力的这种操作相当于认为同一个图像区域中所有视觉特征具有同等重要性, 但实际情况并不是这样, 所以本文认为空间注意力模型对特征的选择是不准确的.

2) 空间注意力模型对局部视觉向量上视觉特征的选择是强制性的, 解码器在预测每个单词时, 空间注意模型都要求局部视觉向量权重之和为 1, 这就造成了局部视觉向量上没有解码器需要的视觉特征时, 空间注意力模型也会向解码器中输入视觉特征, 这些视觉特征就如同噪声一般, 会干扰解码器对单词的预测.

3) 空间注意力模型是一种自顶向下的注意力模型, 对于生成一个长度为 n 的句子, 空间注意力模块需要被执行 n 次, 并且每次执行空间注意力模块时所有的局部视觉向量都需要参与运算, 这无疑大大限制了模型的时间性能.

针对空间注意力模型存在的第 2 个问题, 文献 [17] 提出了自适应注意力机制 (Adaptive attention), 这种方法在局部视觉向量集合中添加一个编码有已生成单词序列语义信息的向量, 当局部视觉向量上没有解码器需要的视觉信息时, 该语义向量所对应的权重就会接近于 1, 从而可防止空间注意力模型强制向解码器中输入视觉特征. 但是自适应注意力机制没能解决第 1 个问题和第 3 个问题, 而且增加了空间注意力模型的参数量和计算复杂度. 文献 [18] 提出的 SCA-CNN (Spatial and channel-wise attention in CNN) 一定程度上对空间注意力模型的第 1 个问题做出了改进, 它的通道级注意力模型能够为编码器输出特征图的每一个通道赋予一个权重, 与空间注意力模型结合在一起既实现了对空间位置的选择也实现了对通道的选择. 但是通道级注意力模型本质上只能为特征图的每个通道计算一个权重, 这种对通道的筛选仍然不灵活、不充分, 并没有完全解决第 1 个问题. 另外 SCA-CNN 没有考虑解决第 2 个问题和第 3 个问题, 相反的通道级注意力模型同样作为一个自顶向下的注意力模型, 在空间注意力模型的基础上进一步增加了模型的参数量和计算复杂度.

1 本文工作

NIC 模型[7] 的应用揭示了单个全局视觉向量能够用来生成整幅图像对应的描述, 这就意味着全局视觉向量是对图像上的多种视觉信息的编码, 相应的每条局部视觉向量都是对局部图像上的多种视觉信息的编码. 换句话说, 视觉特征提取器输出的每条视觉向量都包含了多种视觉特征. 一般而言, 单个句子无法描述出图像中的所有内容, 所以语言模型在生成单条图像描述句子时, 也无法用到所有的视觉特征. 我们称视觉向量上对生成准确图像描述有用的特征为显著性视觉特征, 其余为非显著性视觉特征, 显然对于语言模型来说, 非显著性视觉特征就是噪声, 会影响其生成准确的图像描述. 由于神经网络模型的可解释性不强从而导致特征向量每一维的含义难以被人类所理解, 所以对显著性特征和非显著性特征的定义比较模糊, 但是我们仍然希望在这种思想的指导下, 设计出一种特征提取机制, 能够在训练过程中学会区分这两种特征, 提取显著性视觉特征, 过滤非显著性视觉特征, 本文称这种特征提取机制为显著性特征提取机制 (Salient feature extraction mechanism, SFEM). SFEM 由全局显著性特征提取器 (Global salient feature extractor, GE) 和即时显著性特征提取器 (Instant salient feature extractor, IE) 构成. 实验证明本文的 SFEM 能够有效解决编码器–解码器模型存在的

两个问题, 并且能够避免空间注意力模型所存在的三个问题.

本文在 MS COCO (Microsoft common objects in context) 数据集上对 SFEM 进行了评估, 使用编码器–解码器模型[7, 19] 作为基准模型 (baseline), 实验表明添加 SFEM 模块后, 模型在 BLEU (Bilingual evaluation understudy)/CIDER (Consensus-based image description evaluation) 值上比基准模型有 8.63%/11.24% 的提升. 并且 SFEM 可以完全取代空间注意力模型, 我们在与 SFEM 完全一致的基准模型上实现了空间注意力模型[16, 19], 实验表明 SFEM 在 BLEU4/CIDER 值上比空间注意力模型有 4.29%/5.79% 的提升. 另外本文还进行了两种模型在图形处理器 (Graphics processing unit, GPU)和中央处理器 (Central processing unit, CPU) 环境下的时间性能对比实验, 在单块 Nvidia TITAN X GPU 环境下本文模型的 FPS 值比空间注意力模型高 17.34%, 在 Intel Xeon CPU 环境下优势更加明显, 本文模型的 FPS (Frames per second) 值比空间注意力模型高 43.80%. 由于现有的大多数图像描述算法都是在空间注意力模型上添加新的模块而设计的[17−21], 时间复杂度在空间注意力模型的基础上都有不同程度的增加, 所以相比于其他目前先进的模型, 本文方法在时间性能上具有明显优势.

1.1　算法描述

本文的网络模型如图 1 所示, 整个模型分解为多个步骤, 主要是为了说明每个模块的作用, 实际中无论是前向传播还是反向传播, 本模型都是一个端到端的网络模型. 本文算法的主要步骤如下:

步骤 1. 视觉特征提取. 本文选用在 ImageNet 数据集上预训练过的 Inception-V4 模型作为特征提取器, 用来对输入图像提特征, 从而得到一个包含有多个特征向量的局部视觉向量集合, 以及一个全局视觉向量.

步骤 2. 全局显著性视觉特征提取. GE 会从局部视觉向量集合中提取出各个向量上包含的显著性视觉特征, 然后将整幅图像上的显著性视觉信息编码进一个和单个局部视觉向量维度相同的特征向量中, 本文将该特征向量称为全局显著性视觉向量.

步骤 3. 即时显著性视觉特征提取. IE 根据解码器当前的隐含层状态, 动态决定从全局显著性视觉向量中获取哪些视觉特征, 同时决定视觉特征在当前步预测单词时的参与比例, 从而向语言模型提供对预测本时刻单词最有用的显著性视觉特征.

步骤 4. 单词预测. 本文选用单层 LSTM 网络作为语言模型, 其需要凭借显著性视觉特征和上文的语义特征, 预测本时刻的输出单词. 如果输出单词不是句子终止符号, 则转到步骤 3, 否则完成预测.

1.2　本文的主要贡献

1) 提出了全局显著性特征提取器. 本文的全局显著性特征提取器有三方面的作用: 首先, 全局显

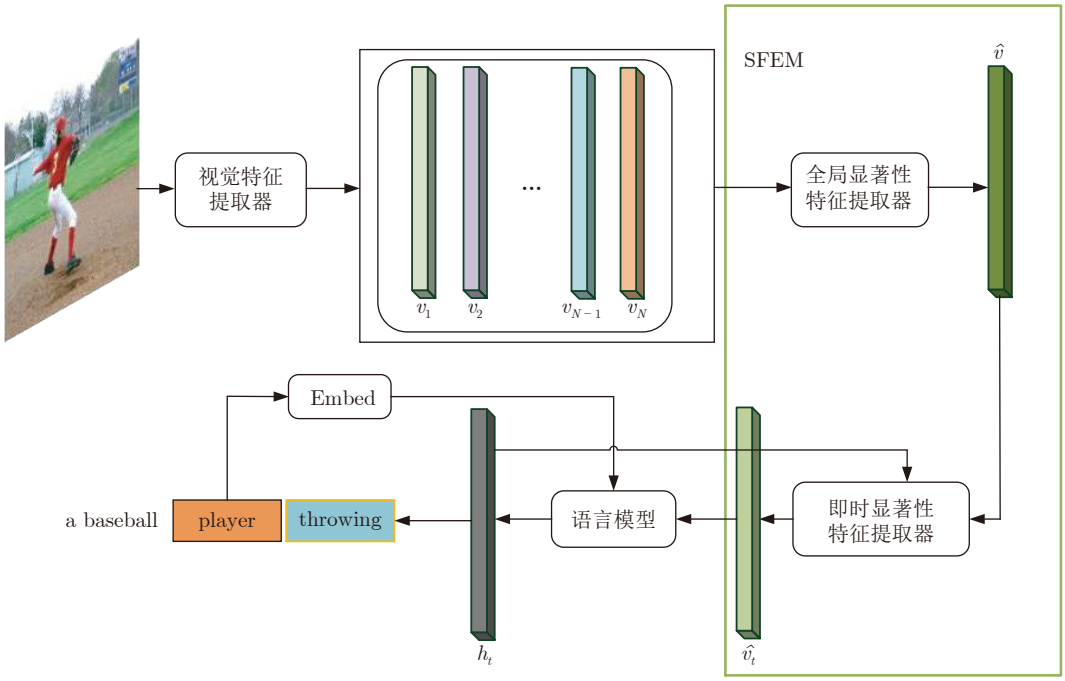


图 1　本文网络模型

Fig. 1　Structure of our network

著性特征提取器会从各个局部视觉向量中提取并整合显著性视觉特征, 这个操作会为局部视觉向量的每一维都生成一个权重, 能够有效克服空间注意力模型对特征选择不准确的问题; 其次, 全局显著性特征提取器不需要使用自上而下的语义信息, 所以对于单幅图像其只需要提取一次显著性视觉特征就可以用来生成任意长度和任意数量的句子; 最后, 全局显著性特征提取器只输出一条全局显著性视觉向量, 能够显著减少解码器端提取视觉信息时的计算量.

2) 提出了即时显著性特征提取器. 本文的即时显著性特征提取器有两方面的作用: 首先即时显著性特征提取器能够根据解码器当前的隐含层状态, 动态决定从全局显著性视觉向量中获取哪些视觉特征, 并有效控制视觉特征在语言模型预测单词时的参与比例, 该比例可以为 0, 避免了空间注意力模型强制向语言模型输入视觉特征的问题; 其次即时显著性特征提取器的计算量明显小于空间注意力模型, 执行速度要优于空间注意力模型.

3) 提出由全局显著性特征提取器和即时显著性特征提取器组成的 SFEM, 使用 SFEM 能够大幅提高编码器−解码器模型生成图像描述的准确性, 并且相比于广泛使用的空间注意力模型, SFEM 在生成图像描述的准确性和时间性能两方面都具有明显的优势.

4) 将全局显著性特征提取器和即时显著性特征提取器分别与空间注意力模型组合使用, 实验结果表明本文的全局显著性特征提取器和即时显著性特征提取器单独使用时也能提升空间注意力模型生成图像描述的准确性.

2 基于显著性特征提取的图像描述模型

2.1 视觉特征提取器

视觉特征提取器通常也称为编码器, 主要作用是从输入图像中提取整张图像上的视觉特征. 本文选用在 ImageNet 数据集上预训练过的 Inception-V4 作为编码器. 首先将任意尺寸的图像预处理为 229×229 像素, 然后将图像送入编码器中提取其视觉特征. 在 Inception-V4 中, 第 3 个 Inception-C 模块输出 1 536 个通道的特征图, 每个特征图的尺寸为 8×8, 将这些特征图由 $C \times W \times H$ 形变为 $(W \times H) \times C$, 从而得到局部视觉向量集合 $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_{64}\}$, $\boldsymbol{v}_i \in \mathbf{R}^{1\,536}$, 如图 2 所示, 本文将图像划分为规则的网格, $\boldsymbol{v}_i$ 的感受野对应于图像上第 i 个格子, 另外取平均层的输出为全局视觉向量 $\boldsymbol{g} \in \mathbf{R}^{1\,536}$, 对应的感受野是整幅图像.

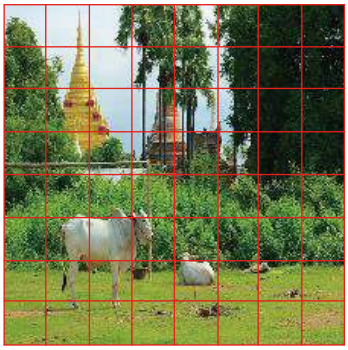


图 2 局部视觉向量与图像的对应关系

Fig. 2 Correspondence between local visual vectors and image

2.2 语言模型

语言模型通常也称为解码器, 对于给定的一幅图像 $\mathcal{I}$, 我们的目标是生成描述这幅图像内容的一条句子 $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_N\}$, 其中 $\mathcal{S}_i$ 表示句子中第 i 个单词. 遵循图像描述中有监督学习的优化方式, 建立语言模型时的优化目标是最大化产生图像正确描述的概率, 所以理想情况下, 模型的参数 θ 应该满足

$$\theta^* = \arg\max_{\theta} \sum_{(\mathcal{I}, \hat{\mathcal{S}})} \log p\left(\hat{\mathcal{S}}|\mathcal{I}; \theta\right) \tag{1}$$

其中, θ 是模型的参数, $\mathcal{I}$ 是一幅图像, $\hat{\mathcal{S}}$ 是这幅图像对应的正确描述. 使用链式法则展开 $p(\hat{\mathcal{S}}|\mathcal{I}; \theta)$

$$\log p\left(\hat{\mathcal{S}}|\mathcal{I}\right) = \sum_{t=1}^{N} \log p\left(\hat{\mathcal{S}}_t|\mathcal{I}, \hat{\mathcal{S}}_1, \cdots, \hat{\mathcal{S}}_{t-1}\right) \tag{2}$$

为了在表达上简洁, 我们去掉了 θ. 本文使用单层的 LSTM 网络对 $p(\hat{\mathcal{S}}_t|\mathcal{I}, \hat{\mathcal{S}}_1, \cdots, \hat{\mathcal{S}}_{t-1})$ 进行建模, 即

$$\begin{aligned}
&\boldsymbol{i}_t = \sigma\left(W_{ix}\boldsymbol{x}_t + W_{ih}\boldsymbol{h}_{t-1} + W_{i\hat{v}}\hat{\boldsymbol{v}}_t\right)\\
&\boldsymbol{f}_t = \sigma\left(W_{fx}\boldsymbol{x}_t + W_{fh}\boldsymbol{h}_{t-1} + W_{f\hat{v}}\hat{\boldsymbol{v}}_t\right)\\
&\boldsymbol{o}_t = \sigma\left(W_{ox}\boldsymbol{x}_t + W_{oh}\boldsymbol{h}_{t-1} + W_{o\hat{v}}\hat{\boldsymbol{v}}_t\right)\\
&\boldsymbol{g}_t = \tanh\left(W_{gx}\boldsymbol{x}_t + W_{gh}\boldsymbol{h}_{t-1} + W_{g\hat{v}}\hat{\boldsymbol{v}}_t\right)\\
&\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \boldsymbol{g}_t\\
&\boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh\left(\boldsymbol{c}_t\right)\\
&p\left(\hat{\mathcal{S}}_t|\mathcal{I}, \hat{\mathcal{S}}_1, \cdots, \hat{\mathcal{S}}_{t-1}\right) = \delta\left(\mathrm{softmax}\left(\boldsymbol{h}_t\right), k\right)\\
&\boldsymbol{x}_0 = \tanh\left(W_{x\hat{v}}\hat{\boldsymbol{v}}\right)
\end{aligned} \tag{3}$$

其中, $\delta(\boldsymbol{x}, k)$ 表示取向量 $\boldsymbol{x}$ 第 k 维上的值, $\hat{\boldsymbol{v}}$ 表示全局显著性视觉向量, $\hat{\boldsymbol{v}}_t$ 表示解码器在 t 时刻所需的显著性视觉向量, W 表示网络权重.

2.3 SFEM

在图像描述领域, 解码器之所以可以生成描述

图像内容的句子, 核心之处在于向解码器中输入了视觉特征, 这些视觉特征能够指导编码器生成与图像内容相关的图像描述. 而如何在合适的时间向解码器中输入合适的视觉特征则是让解码器生成最符合图像内容的描述的关键之处. 本文提出了显著性特征的概念, 并在提取显著性视觉特征, 过滤非显著性视觉特征的思想指导下设计出 SFEM, 如图 3 所示, SFEM 包含 GE 和 IE 两个部分. GE 能够自适应地提取视觉向量 $\boldsymbol{v}_i$ 上的显著性视觉特征, 过滤掉非显著性视觉特征. 然后 GE 会将所有局部视觉向量 $\boldsymbol{v}_i$ 上的显著性视觉特征整合到唯一的一条特征向量中, 称其为全局显著性视觉特征向量 $\hat{\boldsymbol{v}}$, 之后解码器所需的一切视觉信息只需要从 $\hat{\boldsymbol{v}}$ 上获取. GE 为

$$\hat{\boldsymbol{v}} = \phi(\boldsymbol{g}, V) \tag{4}$$

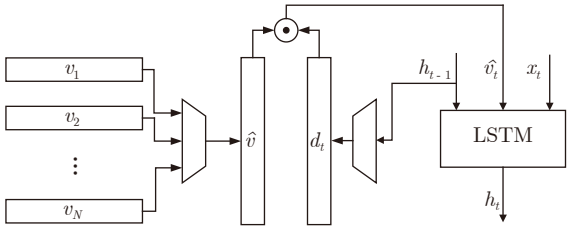


图 3　SFEM 网络结构

Fig. 3　Structure of SFEM

IE 能够根据 LSTM 的隐含层状态 $\boldsymbol{h}_{t-1}$ 从 $\hat{\boldsymbol{v}}$ 中自适应地提取 t 时刻所需的显著性视觉特征, 并能够灵活地控制视觉信息在解码器中的参与比例, 避免无关的视觉信息干扰解码器预测单词, 这一点十分有益于解码器生成语法和语义上正确的句子. IE 为

$$\hat{\boldsymbol{v}}_t = \gamma\left(\hat{\boldsymbol{v}}, \boldsymbol{h}_{t-1}\right) \tag{5}$$

2.3.1　全局显著性特征提取器

使用编码器对给定图像提特征得到全局视觉向量 $\boldsymbol{g} \in \mathbf{R}^D$ 和局部视觉向量集合 $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_N\}$, $\boldsymbol{v}_i \in \mathbf{R}^K$. $\boldsymbol{g}$ 是对整幅图像上视觉信息的编码, $\boldsymbol{v}_i$ 是对图像上局部区域上视觉信息的编码. 正如之前所提到的, 我们认为在每条视觉向量上都存在显著性视觉特征和非显著性视觉特征, 其中显著性视觉特征对于解码器生成图像描述有用, 需要保留下来, 而非显著性视觉特征则会作为噪声干扰解码器生成图像描述, 需要过滤掉. 对于 $\boldsymbol{v}_i = [\alpha_1, \alpha_2, \cdots, \alpha_K]^{\mathrm{T}}$, 在 GE 中直观地将 $\boldsymbol{v}_i$ 的每个维度 α_i 视为一种特征, 并定义权重系数 d_j^{ϕ} 对该特征进行过滤

$$\begin{aligned} \alpha_j' &= \alpha_j d_j^{\phi} \\ d_j^{\phi} &= \sigma\left(\sum_h w_{jh}^{v} \alpha_h + \sum_h w_{jh}^{g} \beta_h\right) \end{aligned} \tag{6}$$

其中, 参数 w_{jh}^{v} 和 w_{jh}^{g} 需要网络在训练中学习, σ 表示 sigmoid 函数, β_h 是全局视觉向量 $\boldsymbol{g}$ 第 h 维的值. 因此相比于空间注意力模型为 $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_N\}$ 构建 N 个权重系数, 本文的 GE 能够会为其构建 $N \times K$ 个权重系数, 这意味着每个局部视觉向量的每个特征值都能被关注到, 从而最大程度地保证了模型对非显著性特征的过滤, 为单个视觉向量构建权重系数如下:

$$\boldsymbol{d}_i^{\phi} = \sigma\left(W_{vd}\boldsymbol{v}_i + W_{gd}\boldsymbol{g}\right) \tag{7}$$

其中, $W_{vd} \in \mathbf{R}^{K \times K}$, $W_{gd} \in \mathbf{R}^{K \times D}$. 值得注意的是, 对于任意的 $\boldsymbol{v}_i$, GE 为其构建 $\boldsymbol{d}_i^{\phi}$ 时共用同一套参数 W_{vd} 和 W_{gd}. 所以就网络的参数量而言, GE 其实和空间注意力模型基本一致. 另外为了减少解码器一端的计算量, 本文将 GE 从各个局部视觉向量中提取到的显著性视觉特征融合到 $\hat{\boldsymbol{v}}$ 中, $\hat{\boldsymbol{v}}$ 的计算方式为

$$\hat{\boldsymbol{v}} = \frac{\sum_{i=1}^{N} \boldsymbol{d}_i^{\phi} \odot \boldsymbol{v}_i}{N} \tag{8}$$

$\hat{\boldsymbol{v}}$ 实际上包含了整幅图像上所有重要的视觉信息, 所以解码器只需要从 $\hat{\boldsymbol{v}}$ 获取视觉信息就能够生成正确的图像描述, 从而能够减少解码器提取视觉特征时的计算量. 最后由于本文的 GE 位于解码器一端, 所以对于单幅图像 GE 只需要执行一次就可以用来生成任意数量、任意长度的图像描述. 而生成一个长度为 n 的句子, 空间注意力模型需要执行 n 次.

2.3.2　即时显著性特征提取器

解码器在预测图像描述时需要两种信息的支持, 首先是前文的语义信息, 其次是图像上的视觉信息. 在本文方法中, 通过 GE 对局部视觉向量集合 $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_N\}$ 中的显著性视觉特征进行提取, 大量的非显著性视觉特征已经被过滤, 但是解码器是按时间顺序逐个预测单词来生成图像描述的, 对于不同的单词, 解码器所需的显著性视觉特征不同, 而于同一个单词, 在图像描述中出现第 i 次和第 $i+1$ 次时, 其所需的显著性视觉特征也不相同. 为此本文提出 IE 用来从 $\hat{\boldsymbol{v}}$ 中提取解码器在每一时刻所需显著性视觉特征. 对于 $\hat{\boldsymbol{v}} = [\alpha_1, \alpha_2, \cdots, \alpha_K]^{\mathrm{T}}$, 在 IE 中采取与 GE 类似的方法, 为每一维的特征值 α_i 赋予一个权重, 来衡量 α_i 的显著性程度. 所以对于 $\hat{\boldsymbol{v}}$, IE 需要为其生成 K 维的权重向量 $\boldsymbol{d}_t^{\gamma}$. $\boldsymbol{d}_t^{\gamma}$ 的构建使用解码器的 $t-1$ 时刻的隐含层状态 $\boldsymbol{h}_{t-1} \in \mathbf{R}^L$ 作为指导, 因为 $\boldsymbol{h}_{t-1}$ 包含了已生成单词序列的语义信息, 所以训练好的模型可以通过 $\boldsymbol{h}_{t-1}$ 来选择 t 时刻预测单词可能会用到的显著性视觉特征, 即

$$\boldsymbol{d}_t^{\gamma} = W_{hv}\boldsymbol{h}_{t-1} \tag{9}$$

其中, $W_{hv} \in \mathbf{R}^{K\times L}$ 是网络需要在训练中学习的参数. 将 $\boldsymbol{d}_t^{\gamma}$ 与 $\hat{\boldsymbol{v}}$ 对应元素相乘就可以获得 t 时刻输入解码器的显著性视觉向量 $\hat{\boldsymbol{v}}_t$

$$\hat{\boldsymbol{v}}_t = \boldsymbol{d}_t^{\gamma} \odot \hat{\boldsymbol{v}} \tag{10}$$

从网络的参数量上来看, 本文的 IE 是非常少的, 并且 IE 在计算上也非常精简, 因为虽然本文的 IE 需要在预测每个单词时都执行一次, 但 IE 只涉及到 $\boldsymbol{h}_{t-1}$ 和 $\hat{\boldsymbol{v}}$ 这两个输入, 并且计算过程仅仅是对 $\boldsymbol{h}_{t-1}$ 进行一个线性变换, 再加一个向量间的对应元素相乘的操作, 所以本文的 IE 每次执行所需的时间要远远小于包括空间注意力模型在内的自定向下注意力模型. 值得注意的是, 本文的 IE 不会强制向解码器输入视觉信息, 因为 $\|\boldsymbol{d}_t^{\gamma}\| \geq 0$, 所以当 $\hat{\boldsymbol{v}}$ 中没有解码器在 t 时刻需要的视觉特征时, $\boldsymbol{d}_t^{\gamma}$ 每一维的值都为 0, 从而将 $\hat{\boldsymbol{v}}$ 上所有特征都作为非显著性特征进行过滤.

2.4 模型优化

本文网络采用端到端的训练方式, 训练过程中固定视觉特征提取器的参数, 只对 SFEM 和语言模型进行训练. 语言模型的损失函数也是整个网络的损失函数, 即

$$\mathcal{L} = -\log p\left(\hat{\mathcal{S}}|\mathcal{I}\right) = -\sum_{t=2}^{N}\log p\left(\hat{\mathcal{S}}_t|\mathcal{I}, \hat{\mathcal{S}}_1, \cdots, \hat{\mathcal{S}}_{t-1}\right) \tag{11}$$

3 实验与分析

3.1 数据集和评价标准

我们使用 MS COCO 数据集[22] 来评价本文提出的方法. MS COCO 的训练集有 82 783 幅图像, 验证集中有 40 504 幅图像, 并且每幅图像对应 5 个标注句子, 这 5 个句子的表达方式以及描述内容不尽相同, 但这 5 个句子都是对图像中内容的描述. 本文使用 Karpathy[23] 中的数据划分方式进行模型的训练和评估, 训练集不变, 依旧是 82 783 幅图像, 从原来的验证集中选择 5 000 幅图像来做验证集, 选择 5 000 幅图像来做测试集. 对数据集的处理参照文献 [19] 的一系列处理方式, 包括将句子中的字母都转换为小写、删除非字母数字的符号、使用空格将单词分割等. 本文只保留在所有句子组成的集合中, 至少出现 5 次的单词, 这样一来, 本文最终的词库大小为 10 516. 对于句子长度, 本文限制在 30 个单词以内, 并且这 30 个单词包括句子的开始符号 BOS 和句子的结束符号 EOS.

本文使用 BLEU1, BELU2, BELU3, BELU4[24], METEOR (Metric for evaluation of translation with explicit ordering)[25], 以及 CIDER[26], ROUGE (Recall-oriented understudy for gisting evaluation)[27], SPICE (Semantic propositional image caption evaluation)[28] 作为评价标准. 对于这些评价标准的计算, 使用的是 MS COCO 图像描述评价工具.

3.2 全局显著性特征提取器性能分析

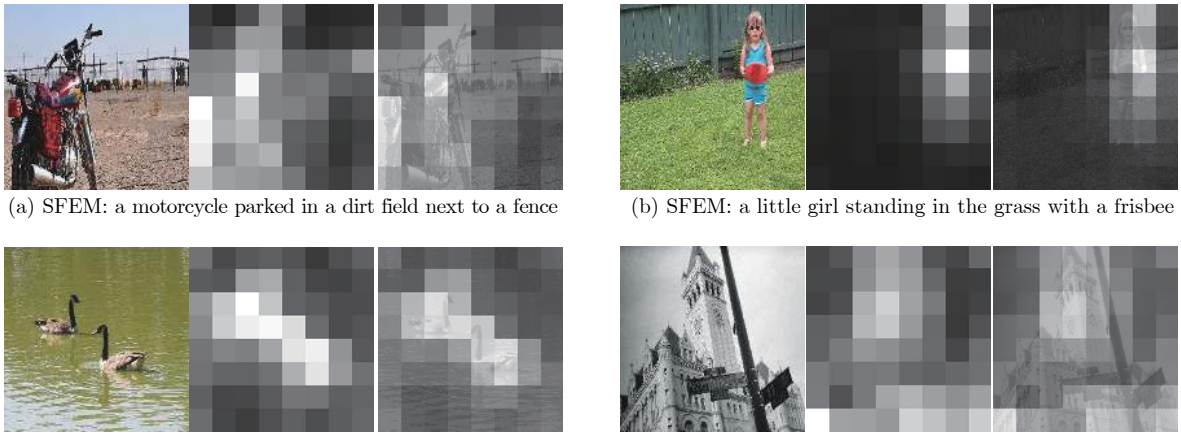
本文的 GE 能够从局部视觉向量中获取到显著性视觉特征, 但是 GE 是通道级别的注意力, 人类很难去理解每个通道表示的是什么, 所以本文采用了一种间接的方式, 可视化出显著性特征在图像上的分布, 以此来展示显著性特征与图像中的哪些内容能够对应起来.

本文通过 W_i 来衡量 GE 从 $\boldsymbol{v}_i$ 提取的显著性视觉特征的量, 具体表示为

$$W_i = \frac{\|\boldsymbol{d}_i^{\phi} \odot \boldsymbol{v}_i\|_1}{K} \tag{12}$$

其中, $\|\boldsymbol{v}\|_1$ 表示向量的 L1 范数, K 是 $\boldsymbol{v}_i$ 的维度. 结合第 2.3.1 节对 GE 的介绍, 可以看出当 W_i 为 0 时, GE 未从 $\boldsymbol{v}_i$ 上提取到任何视觉特征. W_i 越大, 说明 GE 从 $\boldsymbol{v}_i$ 上提取的显著性视觉特征越多.

图 4 是 W_i 的可视化结果, 每个子图中左边是原图, 中间是 W_i 的可视化图, 右边是原图和 W_i 的可视化图的叠加, 文字为本文的 SFEM 生成的图像描述. 在本文的实验中, 视觉特征提取器会从图像中提取出 64 个局部视觉向量, 按顺序对应于图像的 64 个区域. 本文将每个 $\boldsymbol{v}_i$ 对应的 W_i 平铺于对应的区域, 得到 W_i 的可视化图, 其中灰度值越大表示 GE 从该区域的显著性视觉特征越多, 反之则越少. 从图 4 中可以发现, GE 更加关注图像上与周围环境差异比较大的区域, 对于形状、纹理、颜色相似的区域则会适当降低关注. 由此可以推测, 通过 GE 在训练过程中的学习可以得知, 这些相似的区域能够向语言模型提供的视觉特征基本一致, 并且这些区域大概率是背景. 为了避免这部分视觉特征在 $\hat{\boldsymbol{v}}$ 所占比重过高, GE 通常认为这些区域的视觉特征的显著性程度低; 而与周围环境差异比较大的区域通常会存在模型感兴趣的实体对象, 所以 GE 认为这些区域的视觉特征显著性程度高. 需要注意的是模型不会将图像上所有实体对象所在区域都作为感兴趣区域. 由于一句话所能表达的内容有限, 因此模型会与人的表达行为相似, 通常只表达自己感兴趣的内容, 从这一点来看, 本文的 GE 其实是一个内容注意力模块, 对照图 5 可以看出 GE 能够决定语言模型将要描述的图像内容.



(a) SFEM: a motorcycle parked in a dirt field next to a fence

(b) SFEM: a little girl standing in the grass with a frisbee

(c) SFEM: two ducks swimming in a body of water

(d) SFEM: a large building with a clock on it

图 4　显著性特征在空间上的分布

Fig. 4　Spatial distribution of salient features

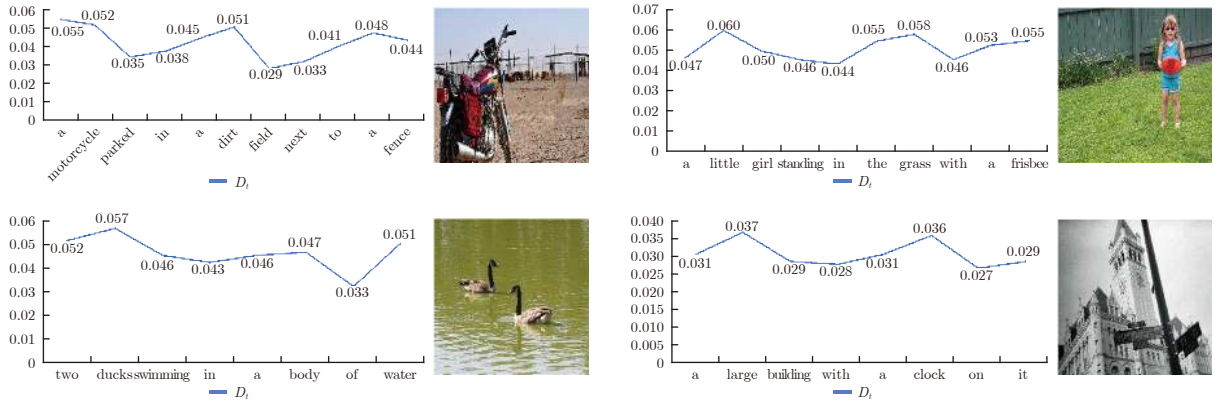


图 5　即时显著性特征随预测单词的变化

Fig. 5　The change of instant salient features with predicted words

3.3　即时显著性特征提取器性能分析

全局显著性视觉向量 $\hat{\boldsymbol{v}}$ 中包含了整幅图像中的显著性视觉特征, 但是语言模型预测不同单词时需要的视觉特征并不相同, 每个单词只与 $\hat{\boldsymbol{v}}$ 中部分视觉特征相关, 所以 IE 需要向语言模型提供 t 时刻最需要的视觉特征, 这些视觉特征和模型在 t 时刻预测出的单词具有较强的相关性, 这些视觉特征称为 t 时刻预测单词对应的显著性视觉特征.

本文通过 D_t 来衡量 IE 在 t 时刻提取显著性视觉特征的量, D_t 等于 $\boldsymbol{d}_t^{\gamma} \odot \hat{\boldsymbol{v}}$ 的 L1 范数除以 $\hat{\boldsymbol{v}}$ 的维数, 即

$$D_t = \frac{\|\boldsymbol{d}_t^{\gamma} \odot \hat{\boldsymbol{v}}\|_1}{K} \tag{13}$$

结合第 2.3.2 节对 IE 的介绍, 可以看出当 D_t 为 0 时, IE 不会从全局显著性视觉向量中提取到任何视觉特征, 此时解码器对单词的预测完全参考 LSTM 在 t 时刻之前累积的语义特征. 当 D_t 越大时, 说明 LSTM 在 t 时刻参考的视觉特征越多.

本文认为一个完全符合图像内容的句子, 它的每一个单词都应该在图像上有据可查, 所以单词对应的 D_t 值通常不会为 0, 而 D_t 值的高低主要取决于 3 个因素: 首先是单词的抽象程度 (抽象程度越低则 D_t 值越高); 其次是单词对应的图像内容应大致位于 GE 给出的显著性程度较高的区域 (GE 会过滤掉大量视觉特征); 最后是单词在数据集中出现的频数 (频数越高则 D_t 值越高). 通常情况下不考虑单词出现的频数, 只有当单词出现的频数过低时, 频数才会成为主因. 本文对单词的抽象程度进行了简单定义: 可以从图像上直接观察到, 不需要根据图像内容做出推理的单词我们认为其抽象程度比较低, 需要根据图像内容进行推理或者需要根据英语语法进行推理的单词我们认为其抽象程度比较高 (注意单词的抽象程度与词性没有直接关系, 实体对象的名称、数量和属性通常都可以从图像上直接观察到, 所以它们的抽象程度一般都比较低).

我们使用本文模型为测试集中所有图像生成对

应句子, 然后统计该单词在所有句子中 D_t 的均值, 从而得到 $\overline{D}_t$, 在表 1 中给出了 $\overline{D}_t$ 值最高的 20 个单词. 可以看出这些单词包括实体对象的名称以及属性, 通常情况下可以从图像上直接观察到. 我们对图 5 第 1 张图中每个单词进行详细分析: 第 1 个单词 "a"表示摩托车的数量, 可以直接从图像中观察到, 所以其抽象程度较低; 第 2 个单词 "motorcycle"表示摩托车的类别名称, 可以直接从图像中观察到, 所以其抽象程度较低; 第 3 个单词 "parked"抽象程度比较高, 因为模型需要从摩托车上没有人来推测它的状态是停放的; 第 4 个单词 "in"抽象程度比较高, 因为模型需要根据语法和图像内容进行推理才能得到; 第 5 个单词 "a"抽象程度比较高, 因为这一个 "a"并不是很直观, 它需要从语法和图像内容进行推理才能得到; 第 6 个单词 "dirt"表示地面的属性, 但是模型不需要识别出 "field", 模型从 "field"所在的显著性程度较高的单块区域就可以判断出泥地面是脏的; 第 7 个单词 "field"对应的图像内容大部分位于 GE 给出的显著性程度较低的区域, 从仅剩的几块显著性较高的区域模型很难识别出 "field"; 第 8 个单词 "next"抽象程度比较高, 因为模型需要从摩托车和栅栏的位置关系推理得到, 以此类推后面的几个单词的 D_t 值.

表 1 $\bar{D}_t$ 值最高的 20 个单词

Table 1 The top-20 words with $\bar{D}_t$ value

单词	$\overline{D}_t$	单词	$\overline{D}_t$	单词	$\overline{D}_t$
hood	0.0592	ducks	0.0565	doughnut	0.0546
cats	0.0589	pug	0.0564	baby	0.0546
teddy	0.0576	rug	0.0561	bird	0.0545
little	0.0573	hummingbird	0.0556	pen	0.0543
duck	0.0571	pasta	0.0549	motorcycle	0.0543
bananas	0.0569	horse	0.0547	colorful	0.0542
seagull	0.0565	panda	0.0546	—	—

3.4 SFEM 评估

1) Encoder-Decoder + SFEM. 本文使用第 2.1 节的视觉特征提取器作为编码器, 以第 2.2 节提到的语言模型作为解码器, 搭建出编码器–解码器模型作为实验的基准模型, 在基准模型上面分别添加空间注意力模型和本文提出的 SFEM 进行对比实验. 如表 2 所示, 本文模型比基准模型在 BLEU-4 值上提升了 8.63%, 在 CIDER 值上提升了 11.24%. 本文模型比空间注意力模型在 BLEU-4 值上提升了 4.29%, 在 CIDER 上提升了 5.79%.

2) Up-Down-SFEM. 为了充分对比 SFEM 和空间注意力模型的性能, 并验证显著性目标检测方法能否提高 SFEM 的性能, 本文以文献 [20] 中提出的 Up-Down 模型作为基准模型进行实验. Up-Down 模型包含自底向上注意力模型和自顶向下注意力模型, 其中自底向上注意力模型也是视觉特征提取器, 由一个 Faster-RCNN (Region-based convolutional neural network)[29] 构成, 自顶向下注意力模型就是空间注意力模型. Up-Down 模型使用 Faster-RCNN 从图像上检测出显著性目标, 并提取出显著性目标对应的视觉向量, 每个显著性目标对应一个视觉向量, 所以视觉特征提取器输出的也是一个局部视觉向量集合, 接下来这些局部视觉向量会送给空间注意力模型用来获得语言模型预测每个单词时所需的视觉特征. 由于文献 [20] 中训练 Up-Down 模型使用了额外的 VG (Visual genome)[30] 数据集, 以及强化学习[31], 所以本文对 Up-Down 模型的实现细节以及训练方式可参考文献 [21]. 实验中使用 SFEM替换掉空间注意力模型来对比 SFEM和空间注意力的性能, 表 3 中 Up-Down-Spatial Attention 表示按照文献 [21] 方法实现的 Up-Down模型, Up-Down-SFEM 表示用 SFEM 替换空间注意力模块后的模型. 我们取 Faster-RCNN 中 (Region proposal network) 之前的视觉特征提取网络作为编码器, 构造了一个编码器–解码器模型并为其添加 SFEM 模块, 以此来验证用显著性目标检测方法替换掉编码器能否提高 SFEM的性能, 在表 3 中将该模型表示为 Encoder-Decoder$_\star$ + SFEM. 对比 Encoder-Decoder$_\star$ + SFEM 和 Up-Down-SFEM 的结果, 可以看出使用显著性目标检测方法并没有明显提高 SFEM的性能, 其中 BLEU-4 和 ROUGE-L 值有轻微的下降, 我们认为有两方面的原因, 首先是 SFEM 中 GE 本身就具有选取显著性区域的能力, 所以显著性目标检测方法对 SFEM 的增益有限; 其次是显著性目标检测方法会将实体对象分割开来, 可能会丧失表示实体对象相互关系的特征. 另外对

表 2 Encoder-Decoder + SFEM 在 MS COCO 数据集上的表现 (%)

Table 2 The performance of Encoder-Decoder + SFEM on MS COCO dataset (%)

模型名称	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDER	SPICE
Encoder-Decoder[7, 19]	72.2	55.4	41.7	31.3	24.6	53.0	95.5	17.2
Encoder-Decoder + Spatial Attention[7, 19]	73.4	57.0	43.2	32.6	25.3	54.0	100.1	18.5
Encoder-Decoder + SFEM	**75.1**	**58.8**	**44.9**	**34.0**	**26.3**	**55.2**	**105.9**	**19.5**

表 3　Up-Down + SFEM 在 MS COCO 数据集上的表现 (%)
Table 3　The performance of Up-Down + SFEM on MS COCO dataset (%)

模型名称	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDER	SPICE
Encoder-Decoder $\star$ + SFEM	74.3	55.8	42.1	**33.2**	25.7	**54.5**	105.2	19.4
Up-Down-Spatial Attention[20-21]	74.2	55.7	42.3	**33.2**	25.9	54.1	105.2	19.2
Up-Down-SFEM	**74.6**	**56.0**	**42.4**	33.1	**26.0**	54.2	**106.1**	**19.7**

比 Encoder-Decoder $\star$ + SFEM 和 Up-Down-Spatial Attention 的结果, 可以看出在显著性目标检测方法的辅助下, 空间注意力模型的性能才能够接近本文的 SFEM, 但这样进一步降低了空间注意力模型的时间性能.

3) SFEM 的时间性能. 在表 4 中给出了空间注意力模型和本文的 SFEM 的时间性能对比, 对于 Karpathy 划分下的测试集中的 5 000 个样本, 本文模型在单块 Nvidia TITAN X GPU 环境下测试得到 FPS 值比空间注意力模型高 17.34%, 在 Intel Xeon CPU 环境下本文模型的 FPS 值比空间注意力模型高 43.80%. 事实上, 现有的很多图像方法都是在空间注意力模型的基础上添加模块得到的, 所以这些方法的计算复杂度都要比空间注意力模型高, 相应在速度上都要比空间注意力模型慢. 所以, 本文方法相比这些方法在速度上的优势明显.

表 4　本模型和空间注意力模型的时间性能对比 (帧/s)
Table 4　Time performance comparison between our model and the spatial attention model (frame/s)

模型名称	帧速率 (GPU)	帧速率 (CPU)
Encoder-Decoder + Spatial Attention[7, 19]	69.8	36.3
Encoder-Decoder + SFEM	81.9	52.2

空间注意力模型可以表示为

$$\hat{\boldsymbol{v}}_t = \sum_{i=1}^{N} \frac{\exp\left(\alpha\left(\boldsymbol{v}_i, \boldsymbol{h}_{t-1}\right)\right)}{\sum_{j=1}^{N} \exp\left(\alpha\left(\boldsymbol{v}_j, \boldsymbol{h}_{t-1}\right)\right)} \boldsymbol{v}_i$$

$$\alpha\left(\boldsymbol{v}_i, \boldsymbol{h}_{t-1}\right) = \boldsymbol{w}_\alpha^{\mathrm{T}} \tanh\left(W_{v\alpha}\boldsymbol{v}_i + W_{h\alpha}\boldsymbol{h}_{t-1}\right) \quad (14)$$

实验中, $W_{h\alpha}$ 与式 (9) 中 W_{hv} 参数量相同, 所以 $W_{h\alpha}\boldsymbol{h}_{t-1}$ 与式 (9) 的计算量是一致的, 式 (10) 是两个向量对应元素相乘, 它的计算量相比矩阵乘法可以忽略不记, 所以 $W_{h\alpha}\boldsymbol{h}_{t-1}$ 的计算量几乎等同于整个 IE 的计算量, 另外由于生成每个单词时 $\alpha(\boldsymbol{v}_i, \boldsymbol{h}_{t-1})$ 需要计算 N 次, 所以 IE 的计算量远远小于空间注意力模型.

表 5 中是各个模块单次执行时平均花费的时间, 其中 GE 单次执行花费的时间和空间注意力模型相当, 但是由于 GE 对于单幅图像只需要执行一次就可以用来生成任意长度、任意数量的图像描述, 所以在生成图像描述的完整过程中 GE 花费的时间小于空间注意力模型. IE 与空间注意力模型类似, 在生成每个单次时都要执行一次, 但是 GPU 环境下空间注意力模型花费的时间是 IE 的 4.79 倍, CPU 环境下空间注意力模型花费的时间是 IE 的 21.84 倍.

表 5　各个模块单次执行平均花费时间 (s)
Table 5　The average time spent by each module in a single execution (s)

模型名称	单次执行时间 (GPU)	单次执行时间 (CPU)
Spatial Attention[7, 19]	0.00035	0.0019
GE	0.00034	0.0020
IE	0.000073	0.000087

4) SFEM 与其他注意力模型的对比. 表 6 中列出了近几年在图像描述领域常用的一些注意力模型, 其中 Soft-Attention、Hard-Attention 与本文中对比的空间注意力模型基本一致, 本文的 SFEM 性能优于这两种注意力模型, SCA-CNN 和 Up-Down

表 6　本文模型在 MS COCO 数据集上的表现 (%)
Table 6　The performance of our model on MS COCO dataset (%)

模型名称	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDER	SPICE
Soft-Attention[16]	70.7	49.2	34.4	24.3	23.9	—	—	—
Hard-Attention[16]	71.8	50.4	35.7	25.0	23.0	—	—	—
Semantic Attention[9]	70.9	53.7	40.2	30.4	24.3	—	—	—
SCA-CNN[18]	71.9	54.8	41.1	31.1	25.0	—	—	—
Up-Dwon[20]	74.2	55.7	42.3	33.2	25.9	54.1	105.2	19.2
本文: SFEM	**75.1**	**58.8**	**44.9**	**34.0**	**26.3**	**55.2**	**105.9**	**19.5**

是在空间注意力模型上添加新的模块改进得到的, 可以看出本文的 SFEM 与这些改进过的空间注意力模型也是具有可比性的, 所以我们认为本文提出的 SFEM 能够作为一种新的注意力模型应用在图像描述领域.

3.5 组合模型评估

1) 全局显著性特征提取器+空间注意力模型. 本文的全局显著性特征提取器与空间注意力模型配合使用时需要做以下两个改变: 首先是使用全局显著性视觉信息向量 $\hat{\boldsymbol{v}}$ 替代全局视觉向量 $\boldsymbol{g}$; 其次是重新构建一个局部显著性视觉信息向量集合 $\{\boldsymbol{v}'_1, \boldsymbol{v}'_2, \cdots, \boldsymbol{v}'_N\}$ 替代局部视觉向量集合, 具体表示为

$$\boldsymbol{v}'_i = \boldsymbol{d}_i^{\phi} \odot \boldsymbol{v}_i \tag{15}$$

2) 即时显著性特征提取器+空间注意力模型. 本文的即时显著性特征提取器和空间注意力模型配合使用时有两种方式: 第 1 种是空间注意力模型在前, 而即时显著性特征提取器在后; 第 2 种则是即时显著性特征提取器在前, 而空间注意力模型在后. 这两种方式都能够提升空间注意力模型的性能, 但是第 2 种方式的提升更加明显, 所以本文仅对第 2 种组合方式进行介绍. 第 2 种组合方式将即时显著性特征提取器作用于每一个局部视觉向量上, 相当于重新构建了一个局部显著性视觉信息向量集合 $\{\boldsymbol{v}''_1, \boldsymbol{v}''_2, \cdots, \boldsymbol{v}''_N\}$ 替代最初的局部视觉向量集合, 具体表示为

$$\boldsymbol{v}''_i = \boldsymbol{d}_t^{\gamma} \odot \boldsymbol{v}_i \tag{16}$$

在表 7 中给出了全局显著性特征提取器结合空间注意力模型的实验结果以及即时显著性特征提取器结合空间注意力模型的实验结果, 可以看出空间注意力模型添加了全局显著性特征提取器和即时显著性特征提取器之后, 在各个评估标准上都能取得一定程度的提高.

4 结束语

目前空间注意力模型结合编码器–解码器框架在图像描述领域得到了广泛的应用, 但是空间注意力模型有 3 个主要的缺陷. 本文按照语言模型对图像上视觉信息的需求, 将每条视觉向量上的特征分为显著性视觉特征和非显著性视觉特征, 在提取显著性视觉特征过滤非显著性特征的思想指导下, 本文尝试提出一种新的显著性特征提取机制 (SFEM)

表 7 组合模型在 MS COCO 数据集上的表现 (%)

Table 7 Performance of the combined model on MS COCO dataset (%)

模型名称	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDER	SPICE
Spatial Attention[7, 19]	73.4	57.0	43.2	32.6	25.3	54.0	100.1	18.5
GE+Spatial Attention	**74.5**	**57.9**	**44.0**	33.1	**25.9**	54.4	**103.6**	**19.0**
IE+Spatial Attention	74.3	57.8	**44.0**	**33.3**	**25.9**	**54.7**	102.7	18.9


A little girl holding a cell phone in her hand

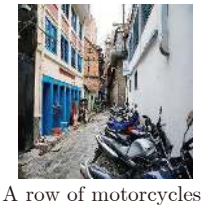

A group of people flying kites in a park


A baseball player throwing a baseball on a field


A group of young men playing a game of soccer


A man talking on a cell phone in a city


A young boy holding a plate of food


A row of motorcycles parked in front of a building


A bedroom with a television and a book shelf

图 6 本文模型生成的图像描述展示

Fig. 6 Image descriptions generated by the model of this paper

用来替代空间注意力模型, 实验表明, 本文的 SFEM 在图像描述的各个评价指标上均优于空间注意力模型, 并且时间性能明显优于空间注意力模型.

References

1 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

2 Mao J H, Xu W, Yang Y, Wang J, Yuille A L. Deep captioning with multimodal recurrent neural networks (m-RNN). In: Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: ICLR, 2015.

3 Tang Peng-Jie, Wang Han-Li, Xu Kai-Sheng. Multi-objective layer-wise optimization and multi-level probability fusion for image description generation using LSTM. *Acta Automatica Sinica*, 2018, **44**(7): 1237−1249
(汤鹏杰, 王瀚漓, 许恺晟. LSTM逐层多目标优化及多层概率融合的图像描述. 自动化学报, 2018, **44**(7): 1237−1249)

4 Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [Online], available: https://arxiv.org/pdf/1406.1078v3.pdf, September 3, 2014

5 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: ICLR, 2015.

6 Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: NIPS, 2014.

7 Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: A neural image caption generator. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015. 3156−3164

8 Zhang Xue-Song, Zhuang Yan, Yan Fei, Wang Wei. Status and development of transfer learning based category-level object recognition and detection. *Acta Automatica Sinica*, 2019, **45**(7): 1224−1243
(张雪松, 庄严, 闫飞, 王伟. 基于迁移学习的类别级物体识别与检测研究与进展. 自动化学报, 2019, **45**(7): 1224−1243)

9 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 (CVPR). Las Vegas, USA: IEEE, 2016. 4651−4659

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

11 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 (ICCV). Santiago, Chile: IEEE, 2015. 2407−2415

12 Wu Q, Shen C H, Liu L Q, Dick A, Van Den Hengel A. What value do explicit high level concepts have in vision to language problems? In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 203−212

13 Yang Z L, Yuan Y, Wu Y X, Cohen W W, Salakhutdinov R R. Review networks for caption generation. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: NIPS, 2016.

14 Xi Xue-Feng, Zhou Guo-Dong. A survey on deep learning for natural language processing. *Acta Automatica Sinica*, 2016, **42**(10): 1445−1465
(奚雪峰, 周国栋. 面向自然语言处理的深度学习研究. 自动化学报, 2016, **42**(10): 1445−1465)

15 Hou Li-Wei, Hu Po, Cao Wen-Lin. Automatic Chinese abstractive summarization with topical keywords fusion. *Acta Automatica Sinica*, 2019, **45**(3): 530−539
(侯丽微, 胡珀, 曹雯琳. 主题关键词信息融合的中文生成式自动摘要研究. 自动化学报, 2019, **45**(3): 530−539)

16 Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, et al. Show, attend and tell: Neural image caption generation with visual attention. In: Proceedings of the 32nd International Conference on Machine Learning. Lille, France: JMLR.org, 2015. 2048−2057

17 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 (CVPR). Honolulu, USA: IEEE, 2017. 3242−3250

18 Chen L, Zhang H W, Xiao J, Nie L Q, Shao J, Liu W, et al. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 6298−6306

19 Chen X P, Ma L, Jiang W H, Yao J, Liu W. Regularizing RNNs for caption generation by reconstructing the past with the present. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 7995−8003

20 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/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 6077−6086

21 Lu J S, Yang J W, Batra D, Parikh D. Neural baby talk. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 7219−7228

22 Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: Common objects in context. In: Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014. 740−755

23 Karpathy A, L F F. Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015. 3128−3137

24 Papineni K, Roukos S, Ward T, Zhu W J. BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, USA: ACL, 2002. 311−318

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

26 Vedantam R, Zitnick C L, Parikh D. CIDEr: Consensus-based image description evaluation. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015. 4566−4575

27 Lin C Y. ROUGE: A package for automatic evaluation of summaries. In: Proceedings of the Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004. Barcelona, Spain: Association for Computational Linguistics, 2004.

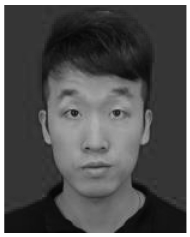
28 Anderson P, Fernando B, Johnson M, Gould S. SPICE: Semant-

ic propositional image caption evaluation. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer, 2016.

29 Ren S Q, He K M, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: NIPS, 2015. 91−99

30 Krishna R, Zhu Y K, Groth O, Johnson J, Hata K, Kravitz J, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 2017, **123**(1): 32−73

31 Rennie S J, Marcheret E, Mroueh Y, Ross J, Goel V. Self-critical sequence training for image captioning. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 1179−1195



王 鑫 西安交通大学软件学院硕士研究生. 主要研究方向为图像内容描述. E-mail: 18991371026@163.com
(**WANG Xin** Master student at the School of Software Engineering, Xi′an Jiaotong University. His main research interest is image captioning.)

宋永红 西安交通大学人工智能学院研究员. 主要研究方向为图像与视频内容理解、智能软件开发. 本文通信作者. E-mail: songyh@xjtu.edu.cn
(**SONG Yong-Hong** Researcher at the College of Artificial Intelligence, Xi′an Jiaotong University. Her research interest covers image and video content understanding, intelligent software development. Corresponding author of this paper.)

张元林 西安交通大学人工智能学院副教授. 主要研究方向为计算机视觉及机器学习.
E-mail: ylzhangxian@xjtu.edu.cn
(**ZHANG Yuan-Lin** Associate professor at the College of Artificial Intelligence, Xi′an Jiaotong University. His research interest covers computer vision and machine learning.)