结合视觉特征和场景语义的图像描述生成

李志欣[1)] 魏海洋[1)] 黄飞成[1)] 张灿龙[1)] 马慧芳[1),2)] 史忠植[3)]

[1)](广西师范大学广西多源信息挖掘与安全重点实验室 桂林 541004)

[2)](西北师范大学计算机科学与工程学院 兰州 730070)

[3)](中国科学院计算技术研究所智能信息处理重点实验室 北京 100190)

摘 要 现有的图像描述生成方法大多只使用图像的视觉信息来指导描述的生成,缺乏有效的场景语义信息的指导，而且目前的视觉注意机制也无法调整对图像注意的聚焦强度. 针对这些问题，本文首先提出了一种改进的视觉注意模型，引入聚焦强度系数自动调整注意强度. 在解码器的每个时间步，通过模型的上下文信息和图像信息计算注意机制的聚焦强度系数，并通过该系数自动调整注意机制的“软”、“硬”强度，从而提取到更准确的图像视觉信息. 此外，本文利用潜在狄利克雷分布模型与多层感知机提取出一系列与图像场景相关的主题词来表示图像场景语义信息，并将这些信息添加到语言生成模型中来指导单词的生成. 由于图像的场景主题信息是通过分析描述文本获得，包含描述的全局信息，所以模型可以生成一些适合图像场景的重要单词. 最后，本文利用注意机制来确定模型在解码的每一时刻所关注的图像视觉信息和场景语义信息，并将它们结合起来共同指导模型生成更加准确且符合场景主题的描述. 实验评估在 MSCOCO 和 Flickr30k 两个标准数据集上进行，实验结果表明本文方法能够生成更加准确的描述，并且在整体的评价指标上与基线方法相比有 3%左右的性能提升.

关键词 图像描述生成；注意机制；场景语义；编码器-解码器框架；强化学习

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2020.01624

Combine Visual Features and Scene Semantics for Image Captioning

LI Zhi-Xin[1)] WEI Hai-Yang[1)] HUANG Fei-Cheng[1)] ZHANG Can-Long[1)]
MA Hui-Fang[1),2)] SHI Zhong-Zhi[3)]

[1)] (*Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University, Guilin* 541004)

[2)] (*College of Computer Science and Engineering, Northwest Normal University, Lanzhou* 730070)

[3)](*Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

Abstract Most of the existing image captioning methods only use the visual information of the image to guide the generation of the captions, lacking the guidance of effective scene semantic information. In addition, the current visual attention mechanism cannot adjust the focus intensity on the image effectively. In order to solve these problems, this paper firstly proposes an improved visual attention model, which introduces a focus intensity coefficient so as to adjust attention intensity automatically. Specifically, the focus intensity coefficient of the attention mechanism is a learnable scaling factor. It can be calculated by the image information and the context information of the model at each time step of the language model decoding procedure. When using the attention mechanism to calculate the attention weight distribution on

收稿日期：2019-08-27；在线发布日期：2020-02-07. 本课题得到国家自然科学基金(61966004, 61663004, 61866004, 61762078)、广西自然科学基金(2019GXNSFDA245018, 2018GXNSFDA281009, 2017GXNSFAA198365)、广西多源信息挖掘与安全重点实验室基金(16-A-03-02, MIMS18-08, MIMS 19-02)资助. **李志欣(通信作者)**，博士，教授，博士生导师，中国计算机学会(CCF)高级会员，主要研究领域为图像理解、机器学习、跨媒体计算. E-mail: lizx@gxnu.edu.cn. **魏海洋**，硕士研究生，主要研究领域为图像理解、机器学习. **黄飞成**，硕士研究生，主要研究领域为图像理解、机器学习. **张灿龙**，博士，教授，中国计算机学会(CCF)会员，主要研究领域为模式识别、目标跟踪. **马慧芳**，博士，教授，中国计算机学会(CCF)会员，主要研究领域为机器学习、数据挖掘. **史忠植**，研究员，博士生导师，中国计算机学会(CCF)会士，主要研究领域为人工智能、机器学习、神经计算、认知科学.

the image, the “soft” or “hard” intensity of attention mechanism can be adjusted automatically by adaptively scaling the input value of softmax function through the focus intensity coefficient. Then the concentration and dispersion of the visual attention can be achieved. Therefore, the proposed attention model can make the extracted image visual information more accurate. Furthermore, we combine unsupervised and supervised learning methods to extract a series of topic words related to the image scene to represent scene semantic information of the image, which is added to the language model to guide the generation of captions. We believe that each image contains several scene topic concepts, and each topic concept can be represented by some topic words. Specifically, we use the latent Direchlet allocation (LDA) model to cluster all the caption texts in the dataset. Then the topic category of the caption text is used to represent the scene category of corresponding image. What is more, we train a multi-layer perceptron (MLP) to classify the image into topic concepts. As a result, each topic category is represented by a series of topic words obtained from clustering. Then the scene semantic information of each image can be represented by these topic words, which are very relevant to the image scene. We add these topic words to the language model so that it can obtain more prior knowledge. Since the topic information of the image scene is obtained through analyzing the captions, it contains some global information of the captions to be generated. Therefore, our model can predict some important words that suitable for image scene. Finally, we use the attention mechanism to determine the visual information of the image and the semantic information of the scene that the model pays attention to at each time step of the decoding procedure, and use the gating mechanism to control the proportion of the input of these two information. Afterwards, both information is combined to guide the model to generate more accurate and scene-specific captions. In the experimental section, we evaluate our model on two standard datasets, i.e. MSCOCO and Flickr30k. The experimental results show that our approach can generate more accurate captions than many state-of-the-art approaches. In addition, compared with the baseline approach, our approach achieves about 3% improvement on overall evaluation metrics.

Keywords image captioning; attention mechanism; scene semantics; encoder-decoder framework; reinforcement learning

1 引 言

图像描述是指根据给定图像的内容，为其生成合理的自然语言描述，是人工智能的一个重要研究领域，主要应用于图像和文本的相互检索、残障人士的生活辅助等方面. 图像描述包含对图像内容的识别以及自然语言生成方面的工作. 首先要求模型能够理解图像的内容，识别图像中的对象，并推理对象之间的关系等；其次是要求模型能够生成被人类理解的自然语言描述. 这是一个结合计算机视觉和自然语言处理的跨领域任务.

通常图像中包含有大量显式和隐式的视觉语义信息,而图像和文本两种模态信息之间本身存在语义鸿沟,图像中的视觉信息实际上很难直接用自然语言完全表征. 最近大量关于图像描述的研究表明[1-3]，基于深度学习的方法可以很好地处理这一复杂任务. 这些方法通常基于一种来自机器翻译的编码器-解码器框架[4-6]，其主要思想是将图像描述任务看作是将一幅图像翻译成一段文本描述. 该框架一般使用卷积神经网络（Convolutional Neural Network，CNN）作为编码器进行图像编码，使用循环神经网络（Recurrent Neural Network，RNN）作为解码器生成文本描述. 这种方法在图像描述任务上取得了突破性的进展，因此目前编码器-解码器框架已经成为图像描述生成的基本方法.

尽管取得了一些进展，但图像描述仍然是一项具有挑战性的任务，面临着若干需要重点考虑和解决的问题. 例如，如何更好地利用图像信息？如何更好地建立图像视觉特征与生成文本之间的联系？如何更准确地生成图像的描述语句？目前视觉注意机制已经被证明在图像描述任务中能够发挥很好的作用. 注意机制可以根据解码器的上下文信息，来

重点关注在图像的一些显著区域，从而为词汇的生成提供精准有效的视觉信息指导. 目前的注意机制主要分为“软”注意和“硬”注意，在“硬”注意机制中模型只关注在图像最显著的一个区域上，去除了大量不必要信息的干扰，注意力权重分布是一个 one-hot 向量. 但由于其不能直接通过反向传播来进行训练，需要通过采样方法来估计梯度，相对来说比较复杂，因此目前大多数图像描述系统采用的是“软”注意机制[7-8]，即通过 softmax 函数来计算所有图像区域上的注意力权重分布，将所有图像区域的加权和作为视觉注意特征. 但由于 softmax 函数自身计算方式的缘故，如果传入 softmax 函数的数值区间较大，则生成的注意力分布就会相对集中，模型则变得比较“硬”，反之则较“软”. 而在生成描述的过程中，对于不同单词的生成，模型需要对图像施加不同聚焦强度的关注. 例如在生成一些名词时模型可能需要更“硬”的关注，集中地关注在图像的某个显著对象上；而在生成一些连词和介词时，可能需要较“软”的关注，分散地关注图像的所有区域. 此外，大多数图像描述系统缺乏场景语义信息的指导. 目前的图像描述系统对于图像的利用，大部分是直接通过 CNN 提取图像特征[1-3]或者是通过目标检测器提取图像上一些候选区域的特征[9-10]来表示图像信息，图像中潜在的场景语义信息却很少被利用. 而对于图像描述生成任务来说，场景语义信息对语句的生成至关重要. 因为对于同样的视觉特征，可能会有不同的场景含义，例如同样是草地，场景可能是公园的草地也可能是足球场的草地.

针对以上问题，本文提出了结合视觉特征和场景语义的图像描述方法，图 1 是模型的整体结构. 首先，本文引入了聚焦强度系数来改进传统的视觉注意机制. 在每个时间步，通过模型的上下文信息和图像信息来计算视觉注意机制的聚焦强度系数，并通过该系数来自动地调控注意机制在每个时间步对图像区域的聚焦强度，从而使模型能够捕捉到更准确的图像视觉特征. 此外，本文将图像的场景语义信息添加到解码器中来指导单词的生成. 具体地说，对于图像中存在的场景主题概念，本文通过潜在狄利克雷分布（Latent Dirichlet Allocation，LDA）模型[11]来对数据集中的所有描述语句进行聚类分析，以描述语句的主题类别来表示图像的场景类别. 这里主题类别由聚类所得的一系列主题词汇来表示. 由于图像的场景主题信息是通过分析描述语句所得，因此可以获取到生成描述的一些全局信息，并且可以预知图像的描述语句可能包括哪些重要词汇，从而可对描述的生成提供很大的帮助. 在每个时间步中，通过注意机制来确定解码器重点关注的主题词，并结合图像的视觉特征，共同引导解码器生成更加准确和符合场景的描述. 最后，在标准数据集上对本文方法进行了测试，并在各项评估指标上与其他先进方法进行了对比. 结果表明，本文方法明显优于其他图像描述生成方法.

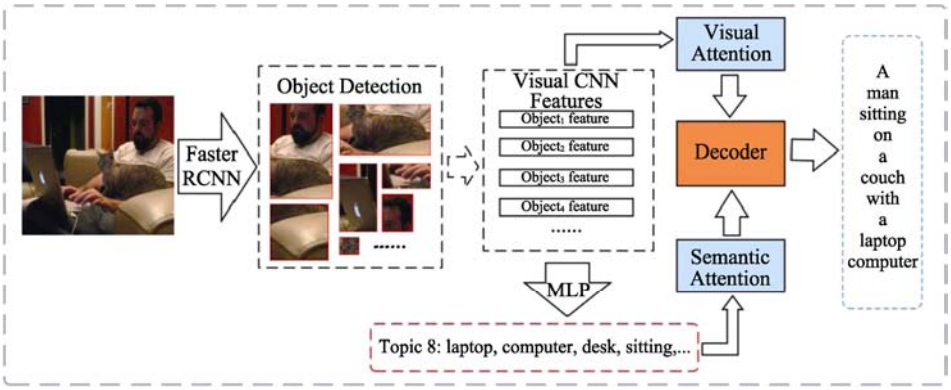


图 1 模型整体结构

本文的主要贡献包括以下几个方面：

(1) 提出了一种改进的视觉注意模型，通过自动调节注意机制的聚焦强度来提取更精准的图像视觉特征.

(2) 通过将无监督和有监督方法相结合来进行图像场景语义信息的抽取，并使用注意机制将场景语义作为先验知识融合到解码器中.

(3) 将图像视觉特征和场景语义相结合，来指导解码器生成更加准确和符合场景的描述.

(4) 在 MSCOCO 和 Flickr 30k 两个标准数据集上进行了充分的实验验证，实验结果从定量和定性两方面证明了本文方法的有效性.

2　相关工作

目前大多数图像描述系统由图像编码器和语言解码器两部分组成. Mao 等人[1]首次创造性地将机器翻译中的编码器-解码器框架应用于图像描述任务，并使用 CNN 作为图像编码器. 为了获得更好的解码能力，Vinyals 等人[2]使用长短期记忆网络（Long Short-Term Memory，LSTM）替换普通 RNN 来为图像生成描述. 这种结构在图像描述任务上取得了突破性的进展，因此后续大量的研究人员一直基于此框架进行研究，并试图对编码器和解码器这两个部分进行改进.

早期的图像描述编解码框架[1-3]只是简单地将 CNN 和 RNN 进行连接，将 CNN 最后的全连接层输出的向量作为图像的编码特征，在解码的初始时间步被输入到解码器中. 但这种方法只能得到图像的全局信息，不仅会丢失一部分图像特征，而且不能随着解码的进行有针对性地对图像进行解析. 为了能够更好地编码图像信息，Xu 等人[7]在图像描述系统中引入了视觉注意机制，去除 CNN 最后的全连接层，将卷积层输出的特征向量作为图像的空间特征，并且通过“软”、“硬”两种注意机制来计算图像的空间注意力分布. 在每个时间步图像的注意特征与词嵌入向量拼接起来共同输入到解码器中来预测词汇的生成. 在生成不同词汇时，注意机制针对性地关注到图像的不同区域. 然而，在语句生成的过程中有些单词（例如：a, of）可能不需要关注图像的视觉信息，于是 Lu 等人[12]针对这个问题提出了一种基于视觉哨兵的自适应注意模型. 视觉哨兵可以确定何时关注图像信息，何时关注语言生成模型. 注意机制在图像描述任务中的运用，使解码器能够更好地整合图像信息，从而建立了更好的视觉信息与生成文本间的联系，使得图像描述任务取得了极大的改进. 目前大部分图像描述系统也都是基于注意机制和编解码框架相结合的结构.

随着目标检测技术的发展，研究人员开始使用基于目标检测的编码器来提取图像特征，并且现在的语言解码器也变得越来越复杂. You 等人[13]首先使用目标检测器从图像中提取一些视觉属性，然后将这些属性整合到语言模型中来增强视觉信息. Anderson 等人[9]提出了一种自底向上和自顶向下的注意模型，使用 Faster R-CNN[14]目标检测器来挑选一组具有高置信度的候选区域，并将这些区域的平均卷积特征作为图像的视觉特征. Lu 等人[10]提出的神经婴儿谈话方法将早期的槽填充方法与基于神经网络的方法相结合，首先生成一个具有插槽的语句“模板”，然后通过目标检测器在图像区域中识别的视觉概念来填充这些槽. Gu 等人[15]提出了由粗到细的多级预测框架，使用多个解码器来生成描述，每个解码器在前一级的输出上执行，产生越来越精细的描述语句. Jiang 等人[16]提出了一种循环融合网络 RFNet，利用多个编码器来提取图像特征，并通过多个 LSTM 进行信息间的循环融合，设计了非常复杂的信息交互.

此外，强化学习和生成对抗网络（Generative Adversarial Net, GAN）[17]也逐渐应用到图像描述系统中，用以优化模型的生成结果. Ranzato 等人[18]首先提出了一种基于 RNN 的策略梯度强化学习方法，直接在评价指标上优化模型的生成结果. Rennie 等人[19]提出了一种自批评的强化学习方法，将模型在推理时生成的语句作为训练的基线，鼓励模型生成相对于基线更好的描述. Dai 等人[20]则通过条件生成对抗网络[21]来生成多样化的图像描述. 强化学习主要针对模型的评价指标进行优化，使得模型的整体评估指标得分可以有较大的提高. 而基于 GAN 的图像描述，大多关注的是描述语句的自然性和多样性，评价指标的得分可能反而较低.

3　模　型

本文采用统一的编码器-解码器框架来构建模型. 给定图像 $\boldsymbol{I}$，首先使用图像编码器来提取图像特征 $\boldsymbol{V}$，然后使用语言解码器进行逐步解码. 在每个时间步注意机制为解码器提供信息引导，最终解码器输出单词序列 $Y=\{y_1,y_2,...,y_T\}$（T 是生成语句的最大长度）.

3.1　模型概述

图像编码器：本文使用预训练的 Faster R-CNN 从输入图像中提取一组候选区域特征作为图像的编码特征. 编码后的图像特征可以表示为 $\boldsymbol{V}=\{\boldsymbol{v}_1,\boldsymbol{v}_2,...,\boldsymbol{v}_L\}$，其中 L 是图像中候选区域的数量. 对于每个图像区域 i，$\boldsymbol{v}_i\in\mathbb{R}^C$ 表示该区域的全局平均卷积特征. Faster R-CNN 可以被认为是一种“硬”注意机制，它从整幅图像中挑选出相对少量的图像区域，可以去除一些不必要区域的干扰，与直接通过 CNN 提取特征相比，这种方法更具有针对性，具有明显的优势.

语言解码器：在经典的编解码模型中，LSTM 通常作为解码器，用于构建语言生成模型. 在每个时间步 $t\in[1,\ T]$，将图像的注意特征 $\boldsymbol{V}_t$ 和上一时间步生成的单词 y_{t-1}，一同输入到 LSTM 中，输出

LSTM 的隐状态 $\boldsymbol{h}_t$，然后通过 $\boldsymbol{h}_t$ 来预测单词的生成，LSTM 以此来逐步解码生成最终的描述序列.

$$\boldsymbol{V}_t = Att(\boldsymbol{V}, \boldsymbol{h}_{t-1}) \tag{1}$$

$$\boldsymbol{x}_t = \boldsymbol{W}_e y_{t-1} \tag{2}$$

$$\boldsymbol{h}_t = LSTM([\boldsymbol{x}_t; \boldsymbol{V}_t], \boldsymbol{h}_{t-1}) \tag{3}$$

$$y_t \sim \boldsymbol{p}_t = \text{softmax}(\boldsymbol{W}_p \boldsymbol{h}_t) \tag{4}$$

其中 $Att(\cdot)$表示视觉注意机制，用于计算图像的注意特征，$\boldsymbol{x}_t$ 是在 t 时间步输入的词嵌入向量，$\boldsymbol{h}_t \in \mathbb{R}^H$ 是 t 时间步 LSTM 的隐状态，$\boldsymbol{W}_p \in \mathbb{R}^{|D| \times H}$ 用于将隐状态 $\boldsymbol{h}_t$ 映射到词典大小维度，$\boldsymbol{p}_t \in \mathbb{R}^{|D|}$ 是预测单词的概率向量（D 是包含所有单词的词典）.

视觉注意机制：视觉注意机制源于对人类视觉的研究，其本质是对图像区域特征 $\boldsymbol{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_L\}$ 进行加权整合，即 $\boldsymbol{V}_t = \sum_{i=1}^{L} \alpha_{t,i} \boldsymbol{v}_i$，在每个时间步将加权后的特征作为视觉信息输入到解码器中，在生成不同单词时，注意机制针对性地关注图像的不同区域. 权值分布是根据先前解码器的隐状态 $\boldsymbol{h}_{t-1}$ 和图像视觉特征 $\boldsymbol{V}$ 来进行计算：

$$e_{t,i} = \boldsymbol{W}_a^T \tanh(\boldsymbol{W}_v \boldsymbol{v}_i + \boldsymbol{W}_h \boldsymbol{h}_{t-1}) \tag{5}$$

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{i=1}^{L} \exp(e_{t,i})} \tag{6}$$

其中 $\boldsymbol{W}_v \in \mathbb{R}^{K \times C}$ 和 $\boldsymbol{W}_h \in \mathbb{R}^{K \times H}$ 将图像特征 $\boldsymbol{v}_i$ 和隐状态 $\boldsymbol{h}_{t-1}$ 映射到同一维度，$\boldsymbol{W}_a \in \mathbb{R}^K$. $\alpha_{t,i}$ 是第 i 个图像区域在 t 时间步的注意权重.

3.2 改进的视觉注意模型

在生成描述的过程中，由于不同词汇的特性不同，对于不同单词的生成，模型应该给予图像不同聚焦强度的关注来提取图像的视觉信息. 如式(6)所示，传统视觉注意机制通过 softmax 函数来计算图像各个区域的注意分布，如果输入 e 的数值区间较大，则经过指数化后，会进一步拉大数值间的差距，最终输出的权重分布就会相对集中；反之，权重分布则相对分散. 有鉴于此，本文提出了自适应聚焦强度的视觉注意机制，通过自动调整 e 的数值区间来使模型能够自动调控视觉注意力的聚焦强度，从而能够针对不同的生成单词更好地提取图像的视觉信息.

本文设置了一个聚焦强度系数 η 来控制视觉注意机制的聚焦强度. 在每个时间步随着上下文信息的变化，$\eta_t = \lambda^{\beta_t}$ 可以自动的调整模型的聚焦强度，其中 λ 是设定的超参数，β_t 通过模型自身学习得到，具体地说是通过图像信息和模型的上下文信息来进行计算：

$$\beta_t = \tanh(\boldsymbol{W}_b^T \overline{\boldsymbol{V}} + \boldsymbol{W}_d^T \boldsymbol{h}_{t-1}) \tag{7}$$

其中 $\boldsymbol{W}_b \in \mathbb{R}^C$，$\boldsymbol{W}_d \in \mathbb{R}^H$. $\overline{\boldsymbol{V}}$ 表示图像的所有区域的平均特征，$\boldsymbol{h}_{t-1}$ 是解码器上一时间步的隐状态. 将聚焦强度系数添加到式(6)中，可得

$$\alpha_{t,i} = \frac{\exp(\eta_t e_{t,i})}{\sum_{i=1}^{L} \exp(\eta_t e_{t,i})} \tag{8}$$

改进后的视觉注意机制和编解码框架相结合可构建基于视觉注意的图像描述模型. 这里采用文献[9]提出的解码器结构作为语言解码器，其构造如图 2 所示. 与传统解码器不同的是，它由两个 LSTM 组成，其中 V-LSTM 为表示视觉注意 LSTM，用于整合当前信息，并为注意机制提供上下文信息输入；L-LSTM 表示语言 LSTM，用于预测单词生成，IV-Att 是改进后的视觉注意机制，具体的解码操作如下所示：

$$\boldsymbol{h}_t^V = LSTM^V([\overline{\boldsymbol{V}}; \boldsymbol{x}_t; \boldsymbol{h}_{t-1}^L], \boldsymbol{h}_{t-1}^V) \tag{9}$$

$$\boldsymbol{V}_t = Att\left(\boldsymbol{W}_u \boldsymbol{V}, \boldsymbol{h}_t^V\right) \tag{10}$$

$$\boldsymbol{h}_t^L = LSTM^L([\boldsymbol{V}_t; \boldsymbol{h}_t^V], \boldsymbol{h}_{t-1}^L) \tag{11}$$

$$y_t \sim \boldsymbol{p}_t = \text{softmax}(\boldsymbol{W}_p \boldsymbol{h}_t^L) \tag{12}$$

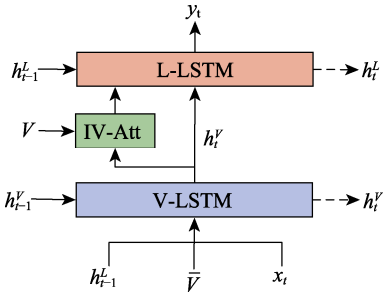


图 2 改进的视觉注意模型

在每个时间步 V-LSTM 接收图像的平均特征 $\overline{\boldsymbol{V}}$ 和词嵌入向量 $\boldsymbol{x}_t$，以及模型的历史信息，并将这些信息进行整合输出隐状态 $\boldsymbol{h}_t^V$. 然后，$\boldsymbol{h}_t^V$ 与图像特征 $\boldsymbol{V}$ 一同输入到改进后的注意力模块 IV-Att，通过注意机制来计算图像的注意特征 $\boldsymbol{V}_t$，其中 $\boldsymbol{W}_u \in \mathbb{R}^{H \times C}$. 得到的视觉注意特征 $\boldsymbol{V}_t$ 与上下文信息 $\boldsymbol{h}_t^V$ 一同被输入到 L-LSTM 并输出 $\boldsymbol{h}_t^L$，最终通过 $\boldsymbol{h}_t^L$ 来预测当前时间步要生成的单词 y_t.

3.3 场景语义信息提取

目前，大多数图像描述系统缺乏场景语义信息的指导. 而对于图像描述生成任务来说，场景语义信息对语句的生成至关重要. 本文的图像场景语义信息提取方法可分为三个步骤，如图 3 所示.

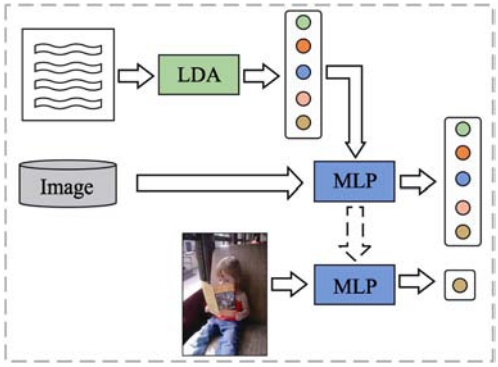


图 3　场景语义信息提取方法

首先，使用无监督的方法对数据集中所有的描述文本进行聚类. 具体的，将数据集中每张图像的多个描述语句合并为一个文档，然后使用 LDA 模型来对所有的描述文档进行聚类分析. 数据集中的所有描述文档被划分为 N 个主题类别，每个主题类别都通过一系列主题词汇来表示，这里选取具有最高概率的 M 个词来表示一个主题类别，即：$U_i=\{w_{i,1},w_{i,2},...,w_{i,M}\},i\in(1,N)$. 分类后，描述文档的主题类别可以看作其对应图像的场景主题类别，这样就可以得到了数据集中所有图像的场景类别标注.

然后，利用标注后的图像数据，可通过监督学习来训练一个多层感知器（Multi-Layer Perceptron，MLP）. 输入图像的视觉特征，MLP 可输出图像对应的场景主题类别.

第三步是通过训练好的多层感知器来对没有描述语句的图像进行场景主题分类. 分类后的每张图像都有一个对应的主题类别 U_i，主题类别 U_i 通过一系列主题词 w 来表示，这样就得到了图像的场景语义信息. 而在这些语义信息中包含了该图像对应描述的一些重要词汇，这些信息可以使解码器能够预知到一些需要生成的词汇，并对生成描述的全局信息有一定的掌控，对于提升模型性能会有很大的帮助.

3.4　结合视觉特征和场景语义

为了使模型能够生成更加准确且符合图像场景的描述，本文提出了结合图像视觉特征和场景语义的图像描述生成方法. 通过一系列与图像场景密切相关的主题词汇来表示图像的场景语义信息，并将其添加到语言模型中，与图像的视觉信息相结合，从而使模型能够得到更丰富的图像信息，并提前预知到一些需要生成的重要词汇，以此来共同引导模型生成更加准确且符合场景的描述.

本文方法是基于图 2 的结构进行拓展，将图像的场景语义信息添加到解码器中. 图 4 是模型解码器的结构图，其中 S-LSTM 表示场景 LSTM，S-Att 表示场景语义注意机制. 具体的解码操作如下：

$$\boldsymbol{S}=\boldsymbol{W}_e U_i \tag{13}$$

$$\boldsymbol{h}_t^S=LSTM^S([\bar{\boldsymbol{S}};\boldsymbol{x}_t;\boldsymbol{h}_{t-1}^L],\boldsymbol{h}_{t-1}^S) \tag{14}$$

$$\boldsymbol{S}_t=Att(\boldsymbol{S},\boldsymbol{h}_t^S) \tag{15}$$

其中 U_i 表示图像的场景主题类别，它由 M 个主题词汇组成，首先将其转化成词嵌入向量形式 $\boldsymbol{S}\in\mathbb{R}^{M\times H}$，$\bar{\boldsymbol{S}}$ 表示 M 个主题词汇的平均特征. 与 V-LSTM 类似，S-LSTM 用于整合当前时刻输入的信息和历史信息，并为注意机制提供上下文信息输入 $\boldsymbol{h}_t^S$. 然后将 $\boldsymbol{h}_t^S$ 和 $\boldsymbol{S}$ 共同输入到 S-Att 模块中，通过 S-Att 输出当前时间步模型关注的场景语义信息.

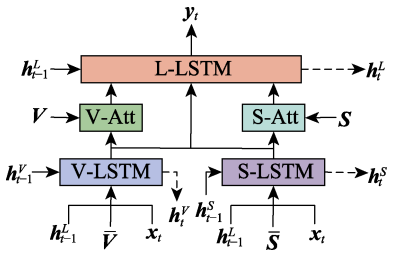


图 4　结合视觉特征和场景语义的图像描述模型

此外，这里设置了一个控制门，来控制当前信息的输入. 在提取到的场景语义信息中实际上包含了一些重要的视觉语义和描述中可能存在的一些重要词汇，这些词汇可以看作是视觉信息额外的补充，可以和视觉信息共同指导单词的生成. 与文献[12]的方法不同，本方法的控制门主要用来控制信息的输入量，以避免引入过多的重复信息或者其他的干扰信息. 具体操作如下：

$$\boldsymbol{g}_t=\sigma(\boldsymbol{W}_g[\boldsymbol{h}_t^V;\boldsymbol{h}_t^S;\boldsymbol{V}_t;\boldsymbol{S}_t]+\boldsymbol{b}_g) \tag{16}$$

$$\boldsymbol{V}_t=\boldsymbol{g}_t\odot\boldsymbol{V}_t,\boldsymbol{S}_t=(1-\boldsymbol{g}_t)\odot\boldsymbol{S}_t \tag{17}$$

$$\boldsymbol{h}_t^L=LSTM^L([\boldsymbol{V}_t;\boldsymbol{S}_t;\boldsymbol{h}_t^V;\boldsymbol{h}_t^S],\boldsymbol{h}_{t-1}^L) \tag{18}$$

$$y_t\sim\boldsymbol{p}_t=\mathrm{softmax}(\boldsymbol{W}_p\boldsymbol{h}_t^L) \tag{19}$$

其中 $\boldsymbol{g}_t\in\mathbb{R}^H$，$\odot$ 表示对应位置元素相乘的操作. 本方法最后将图像的视觉注意信息 $\boldsymbol{V}_t$，场景语义注意信息 $\boldsymbol{S}_t$，隐状态 $\boldsymbol{h}_t^V$ 和 $\boldsymbol{h}_t^S$ 一同输入到 L-LSTM 中，输出 $\boldsymbol{h}_t^L$ 用于预测当前时刻生成的单词.

在本方法中，图像的场景语义信息作为先验知识添加到语言模型中，使模型可以获取更多的图像信息，并且具有更强的全局建模能力. 同时在每个时间步，场景语义注意机制可以使模型获取一些重要词汇信息，结合图像的视觉注意信息，以此来使模型生成更加准确且符合场景的描述.

3.5 训练目标

在模型的训练中，首先使用交叉熵损失来进行训练. 通过给定训练图像的正确参考语句的单词序列 $Y=\{y_1,y_2,...,y_T\}$ 来最小化模型的交叉熵损失：

$$L(\theta)=-\sum_{t=1}^{T}\log p(y_t) \tag{20}$$

即在每个时间步最大化正确参考单词的概率.

然而基于交叉熵损失的训练存在严重的问题. 首先，模型在训练时每个时刻输入的是正确的参考单词，通过交叉熵损失来最大化下一个正确单词的概率，但在推理阶段模型依赖自身之前生成的单词来预测接下来单词，这样在训练和推理之间就存在一种“暴露偏差”[18]. 也就是说，在推理过程中一旦模型前面的单词生成的不好，就会导致误差累计，使得后面的单词也不会很好地生成. 此外，评估模型的评估指标也与交叉熵损失不相关，即模型训练和测试过程的目标不一致. 为了解决这些问题，可以将图像描述生成过程转化为一个强化学习问题，通过强化学习方法直接在评价指标上优化模型的生成效果. Rennie 等人[19]的研究表明，优化模型的评价指标 CIDEr[22]可以使模型所有的评估指标得分都会有所提升. 因此，为了能够得到更好的生成结果，本文也使用了强化学习方法针对评价指标 CIDEr 对模型进行了进一步的优化. 基于强化学习方法的训练目标是最大限度地减小负奖励期望：

$$L(\theta)=-E_{Y\sim p_\theta}[r(Y)]\approx -r(Y) \tag{21}$$

其中 $r(Y)$表示模型生成语句 Y 的 CIDEr 得分. 梯度 $\nabla_\theta L(\theta)$ 可以通过蒙特卡罗方法近似估计：

$$\begin{aligned}\nabla_\theta L(\theta)&=-E_{\mathbf{Y}\sim\mathbf{p}_\theta}[r(Y)\nabla_\theta\log p_\theta(Y)]\\&\approx -r(Y)\nabla_\theta\log p_\theta(Y)\end{aligned} \tag{22}$$

本文遵循 Rennie 等人[19]提出的 SCST 训练方法，使用模型在推理时生成的语句 $\hat{Y}$ 作为基线，来强迫模型生成相对于基线语句更好的描述. 即：

$$\nabla_\theta L(\theta)\approx -(r(Y)-r(\hat{Y}))\nabla_\theta\log p_\theta(Y) \tag{23}$$

基于强化学习的训练方法使得模型在训练和推理过程中保持一致，解决了图像描述模型中存在的“暴露偏差”[18]问题. 更重要的是，它直接在评价指标上优化了描述的生成，从而使模型在训练目标和测试指标上也保持一致，极大地提升了模型的整体性能.

4 实 验

为了证明本文提出方法的有效性，在 MSCOCO 和 Flickr30k 两个标准数据集上进行了充分的实验验证，将其与当前先进的模型进行了对比，并从定量和定性两个方面进行了结果分析.

4.1 数据集和评估指标

本文在 MSCOCO 和 Flickr30k 数据集上进行实验来评估提出的模型，数据集划分方式如表 1 所示. 由于 MSCOCO 数据集的测试集中没有标注语句，因此采用 Karpathy[3]对 MSCOCO 数据集的划分方式，从验证集中挑选出 5000 张图像用于验证，5000 张图像用于测试，验证集剩余的图像与训练集一起用作训练数据，数据集的每张图像包含 5 个人工标注的描述语句. 本文通过对数据集的图像描述文本分析来对文本进行预处理，用“UNK”替换掉出现次数少于 5 次的低频单词.

表 1 数据集划分

数据集	训练集	验证集	测试集
Flickr30k	28000	1000	1000
MSCOCO	113287	5000	5000

图 5 展示了不同语句长度在各自数据集中所占的比重，可以看出 MSCOCO 数据集中的语句长度大多集中在 8 到 15 个单词之间，因此在 MSCOCO

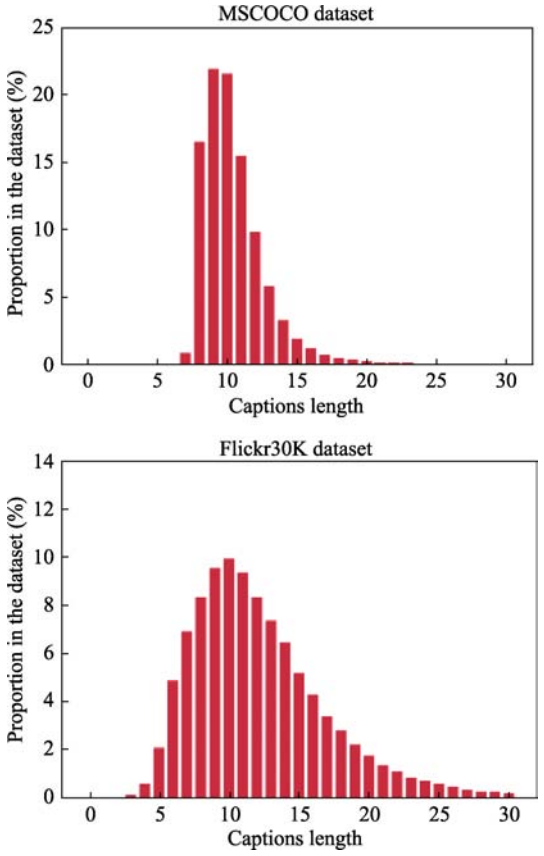


图 5 数据集中语句长度分布图

数据集的实验中，语句的最大长度设置为 16；而 Flickr30k 数据集中的语句长度比较分散，大多集中在 5 到 20 个单词之间，因此在 Flickr30k 数据集的实验中，语句最大长度设置为 20. 为了评估模型的性能，本文使用 BLEU(1-4)[23]，METEOR[24]，ROUGE-L[25]，CIDEr[22]和 SPICE[26]作为评估指标来评估生成语句的质量.

4.2　实施细节

本文使用在 Visual Genome 数据集[27]上预训练过的 Faster R-CNN 对输入图像进行编码，每张图片提取 36 个候选区域，每个候选区用 2048 维的向量表示，输入图像被编码成 2048×36 维的向量. 在场景语义信息抽取过程中，数据集所有描述文档被聚类成 60 个主题类别，每个类别挑选出概率最大的前 20 个词来表示该类别. 在模型的解码器部分，所有 LSTM 的神经元数量被统一设置为 1024，注意层模块神经元数量设置为 1024，词嵌入层的大小同样也是 1024，其他网络参数采用随机初始化.

在训练过程中，首先使用 Adam 优化器[28]在交叉熵损失下训练模型. 初始学习率为 4×10^{-4}，动量参数为 0.9，批量大小为 100. 学习率在训练 15 轮后，每 5 轮衰减一次，衰减率为 0.8. 在交叉熵损失下训练 35 轮后，运行基于强化学习的训练方法，来优化模型的 CIDEr 评估指标. 在这个阶段，学习率设置为 5×10^{-5}. 在每轮训练结束后，在验证集上评估模型的性能. 最后，选择在验证集上具有最高 CIDEr 得分的模型用于测试. 在测试期间，使用波束搜索来生成语句，波束大小设置为 5.

5　实验结果分析

5.1　参数 λ 的选取

本文设置了一个聚焦强度系数 $\eta_t=\lambda^{\beta_t}$ 来控制视觉注意机制的聚焦强度，从而能够针对不同的生成词汇更好地提取图像视觉信息，其中 β 通过模型自身学习而得到，λ 是超参数. 实验在 Flickr30k 数据集上实施，在实验中首先验证了不同的 λ 值对模型生成结果的影响，具体实验结果如图 6 所示，其中横坐标是 λ 的取值，纵坐标是评价指标得分.

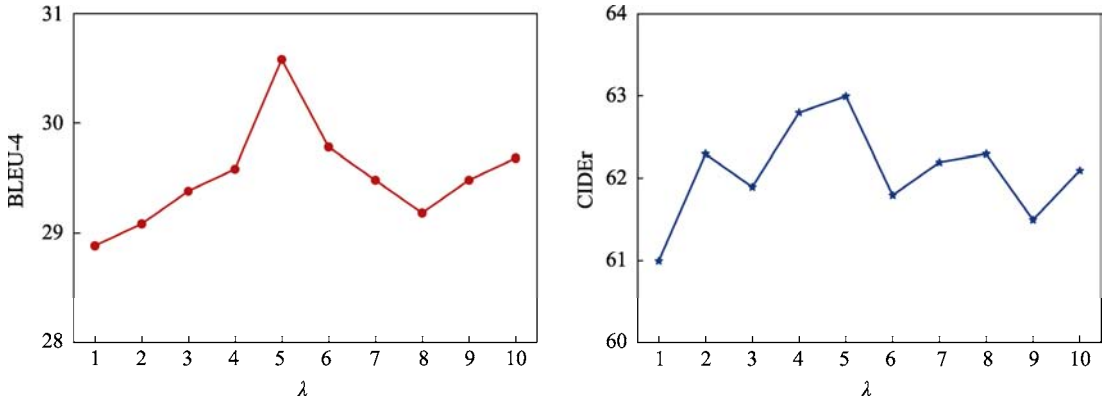


图 6　不同的 λ 值对模型性能的影响

从图 6 中可以明显看出当 λ=1 时，模型的整体性能表现是最差的，因为当 λ=1 时，η=1，即模型的聚焦强度不会变化，相当于没有设置聚焦强度系数. 而当 λ>1 时，无论是模型的 BLEU-4 评分还是 CIDEr 评分都有明显的提升，这也充分说明了改变注意机制的聚焦强度，可以改善模型的整体性能表现. 同时可以看出在 λ=5 时模型整体性能表现最好，因此在后续的实验中将 λ 设置为 5.

5.2　场景语义提取结果分析

对于场景语义信息的提取，首先使用 LDA 模型来对所有的描述文本进行聚类，为图像数据打上主题类别标签，然后根据标记后的数据来训练多层感知器. 由于数据集的图像没有真正的场景类别标签，所以无法从定量的角度来展示本文方法对场景类别分类的好坏. 因此本文从定性的角度展示场景分类效果. 如图 7 所示，从定性分析可以看出，本文方法能够很好地对图像进行分类，并且提取出准确的场景主题词. 在这些主题词中，不仅包含了一些实体名词，还包含了一些描述场景动作的动词，甚至还包含了一些描述场景的形容词，而这些词汇可以很好地帮助模型来生成准确的描述.

5.3　生成结果定量分析

5.3.1　MSCOCO 实验结果分析

本文的方法是在 Up-down[9]的模型上进行的改进，为了展示更加公平的对比，本文首先实现了 Up-down 模型，并使用完全相同的数据和模型参数

图 7 场景主题分类效果

来进行后续的实验对比. 表 2 展示了本文模型在 MSCOCO 数据集上与基线模型的性能比较，其中 Our-Up-down 是本文实现的 Up-down 模型，IVAIC 表示在 Our-Up-down 基础上添加聚焦强度系数后改进的视觉注意模型，VASS 表示结合图像视觉特征和场景语义信息的模型. 表中 B、M、R、C、S 分别表示评价指标 BLEU、METEOR、ROUGE-L、CIDEr、SPICE.

对于改进的视觉注意模型 IVAIC，可以看出，在交叉熵损失的训练下，添加聚焦强度系数后模型的整体性能有略微的改进，在所有评价指标上都略高于基线模型. 由于在交叉熵损失训练下，模型存在“暴露偏差”以及训练目标和评估指标的不匹配问题，模型很难得到完全的优化，所以性能改进的不太明显，但经过强化学习优化后，模型的整体性能得到很大的提升，各项评价指标得分均明显优于基线方法. 其中 BLEU-4 评分提高了 1.3，ROUGE-L 评分提高了 1.1，CIDEr 评分提高 2.2. 这主要是因为模型添加聚焦强度系数后，可以自适应地调整注意机制的聚焦强度来提取更准确的视觉信息，并通过强化学习的优化后，进一步扩大了模型的性能优势. 在结合场景语义信息后，可以看出模型的评价指标得分有了显著的提高. 这充分证明了场景语义信息对描述的生成是非常有帮助的. 在交叉熵损失训练下，模型的整体评价指标得分都已明显高于基线模型. 经过强化学习优化后，模型的性能又得到了进一步的提升，在各项评价指标上都显著优于基线模型. 其中 BLEU-4 评分提高 2.3 分，CIDEr 评分提高 6.6 分. 最终模型的 BLEU-4/CIDEr 评分达到了 38.9/126.7 的性能表现. 这充分表明，在模型中添加场景语义信息后，模型可以获得更多的先验知识，从而生成更准确的描述.

表 2 与基线模型在 MSCOCO 数据集上的性能对比

Approach	Cross-Entropy Loss						CIDEr Optimization					
	B-1	B-4	M	R	C	S	B-1	B-4	M	R	C	S
Our-Up-down	76.3	36.0	27.2	56.3	113.5	20.2	79.3	36.6	27.7	57.0	120.1	21.2
IVAIC	76.5	36.3	27.6	56.4	113.7	20.5	79.9	37.9	27.8	58.1	122.3	21.5
VASS	**76.9**	**36.5**	**27.9**	**56.5**	**114.0**	**20.8**	**80.5**	**38.9**	**28.3**	**58.8**	**126.7**	**21.7**

表 3 展示了本文模型与一些现有先进模型的性能比较，其中(XE)表示交叉熵损失训练后的结果，(RL)表示强化学习优化后的结果. 可以看出，本文的模型依然具有很强的竞争优势. Stack-Cap[15]模型虽然运用了多级的 LSTM，并通过强化学习方法来逐级地优化，但它并没有引入新的知识来指导模型的生成；RFNet[16]同样是通过多个编码器和解码器来进行知识的融合；CVAP[29]方法通过强化学习来优化图像上下文信息对生成描述语句的影响，取得了很好的效果；EICP[30]倾向于生成有吸引力的个性描述，反而生成结果的评价指标得分并不是很高. 这些方法大多都只是利用图像的视觉特征，并通过强

化学习来优化模型的生成，而没有额外辅助信息的引入. 本文的方法则是从数据集的描述语句入手，针对图像的描述语句进行分析，获得图像的语义信息，从而能够预知到生成的语句中可能包含的一些重要词汇，通过将这些重要的词汇添加进模型，使模型获得更多有用的先验知识，从而指导模型生成更加符合标注语句的描述，最终的实验结果也充分证明了本文方法的有效性.

5.3.2　Flickr30k 实验结果分析

表 4 展示了本文的模型在 Flickr30k 数据集上的性能表现. 可以明显的看出，与其他方法相比本文的模型在各项评价指标上都有更好的表现. 在交叉熵损失训练下，添加了场景语义信息的 VASS 模型相比于 IVAIC 有了很大的提升，其中 BLEU-1 评分提升了 2.4 分，CIDEr 评分提高了 3 分，这充分说明，无论数据集的大小，添加场景语义信息都可以显著提升模型的性能. 同时通过强化学习优化(RL)后，可以看出相比于交叉熵损失(XE)训练的结果，VASS 模型的性能提升更加明显，BLEU-1 评分提升了 3.1 分，BLEU-4 评分提升了 3 分，特别是 CIDEr 评分提升了 9.3 分. 这可以说明强化学习的优化可以进一步扩大 VASS 模型的性能优势. 最终的实验结果证明了本文模型在 Flickr30k 这样的小型数据集上依然能取得良好的性能表现，并且结合场景语义信息可以显著改进模型的性能.

5.4　实验结果定性分析

5.4.1　视觉注意的可视化

图 8 展示了改进的视觉注意模型的注意权重可视化效果，其中图 8(a)是添加聚焦强度系数的可视化效果，图片的左上角是每个时间步生成的单词，右上角是每个时间步的注意机制的聚焦强度系数，图 8(b)是没有添加聚焦强度系数的可视化效果.

表 3　在 MSCOCO 数据集上与现有先进模型的性能比较

Approach	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Google NIC[2]	66.6	46.1	32.9	24.6	–	–	–	–
Soft-Attention[7]	70.7	49.2	34.4	24.3	23.9	–	–	–
Adaptive[12]	74.2	58.0	43.9	33.2	26.6	–	108.5	–
SCST[19]	–	–	–	34.2	26.7	55.7	114.0	–
Stack-Cap[15]	78.6	62.5	47.9	36.1	27.4	56.9	120.4	20.9
Up-down[9]	79.8	–	–	36.3	27.7	56.9	120.1	21.4
RFNet[16]	79.1	63.1	48.4	36.5	27.7	57.3	121.9	21.2
CVAP[29]	80.1	64.7	50.0	38.6	28.3	58.5	126.3	21.6
EICP[30]	79.3	–	–	36.4	–	57.5	124.0	21.2
IVAIC(XE)	76.5	59.7	46.3	36.3	27.6	56.4	113.7	20.5
VASS(XE)	76.9	60.1	46.5	36.5	27.9	56.5	114.0	20.8
IVAIC(RL)	79.9	63.9	49.6	37.9	27.8	58.1	122.3	21.5
VASS(RL)	**80.5**	**65.3**	**51.0**	**38.9**	**28.3**	**58.8**	**126.7**	**21.7**

表 4　在 Flickr30k 数据集上的性能比较

Approach	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEO	ROUGE-L	CIDEr	SPICE
Google NIC[2]	66.3	42.3	27.7	18.3	–	–	–	
Soft-Attention[7]	66.7	43.4	28.8	19.1	18.5	–	–	
ATT[13]	64.7	46.0	32.4	23.0	18.9	–	–	
SCA-CNN[8]	66.2	46.8	32.5	22.3	19.5			
RA+SS[31]	64.9	46.2	32.4	22.4	19.4	45.1	47.2	–
CNN+GRU[32]	71.4	54.0	39.5	28.2	21.1	–	–	–
Att-RegionCNN[33]	73.0	55.0	40.0	28.0	–	–	–	–
IVAIC(XE)	70.8	54.1	40.7	30.6	22.5	49.8	63.0	16.8
VASS(XE)	73.2	56.0	41.5	30.6	22.7	50.8	66.0	16.8
IVAIC(RL)	73.3	55.7	42.0	31.6	22.3	50.6	66.5	16.9
VASS(RL)	**76.3**	**58.9**	**44.5**	**33.6**	**23.7**	**52.5**	**75.3**	**17.6**

从图 8(a)中可以看出，注意机制对于不同单词的生成可以自动地调整对图像的聚焦强度，从而提取出更准确的图像视觉特征. 此外，从右上角的聚焦强度系数中也可以看出，在生成描述的大部分时间步中，聚焦强度系数都大于 1，这说明在大部分时间步注意机制都变得更加集中，特别是关注在一些小区域时更加明显(例如在生成“baseball”，“bat”时). 而在一开始聚焦强度系数小于 1，说明模型还不清楚需要生成什么内容，需要将注意分散地关注到更多的图像区域. 在与图 8(b)的可视化对比中也可以明显地看出，没有添加聚焦强度系数的注意机制产生的注意分布相对比较分散(例如在生成“baseball”和“on”这些单词时). 这同时也证明了稍“硬”的注意机制对于视觉信息的提取更有帮助，这一结论也与文献[7]中的结论相一致.

5.4.2 场景语义注意的可视化

图 9 展示了在单词生成过程中场景语义注意权重的可视化以及公式(16)中控制门权重 g_t 的可视化. 其中的折线图表示在单词生成的每个时间步 g_t 均值的变化，其纵坐标为 $1-g_t$ 的均值，表示模型在 t 时间步保留场景语义信息的程度. 从图中可以看出在每个时间步模型还是更多的依赖于图像的视觉信息，场景语义信息只是作为图像的额外信息来指导单词生成，同时也可以看出在生成大部分的非实体名词时，模型会较多的关注在场景语义信息上. 图 9 中的注意力分布图展示了场景语义权重的可视化，从图中可以发现，在单词的生成过程中，注意机制在生成某个单词时并没有把注意力集中在场景词中的这个单词上. 例如，在左图中生成单词“tie”时，模型并没有关注主题词中的“tie”，在右图中当生成“soccer”时，模型也没有关注在“soccer”上，但是可以看出，在生成描述时，模型会关注在一些与场景整体相关的重要词汇上，例如左图中的“man”、“wearing”、“black”和右图中的“children”、“game”、“people”、“team”. 这主要是因为在生成这些实体名词时，图像的视觉信息起到了主导作用，已经提供足够的信息量，而场景语义能够提供一些其他的辅助信息，从而共同引导模型生成更好的语句.

(a) 添加聚焦强度系数的可视化

(b) 没有添加聚焦强度系数的可视化

图 8 视觉注意的可视化

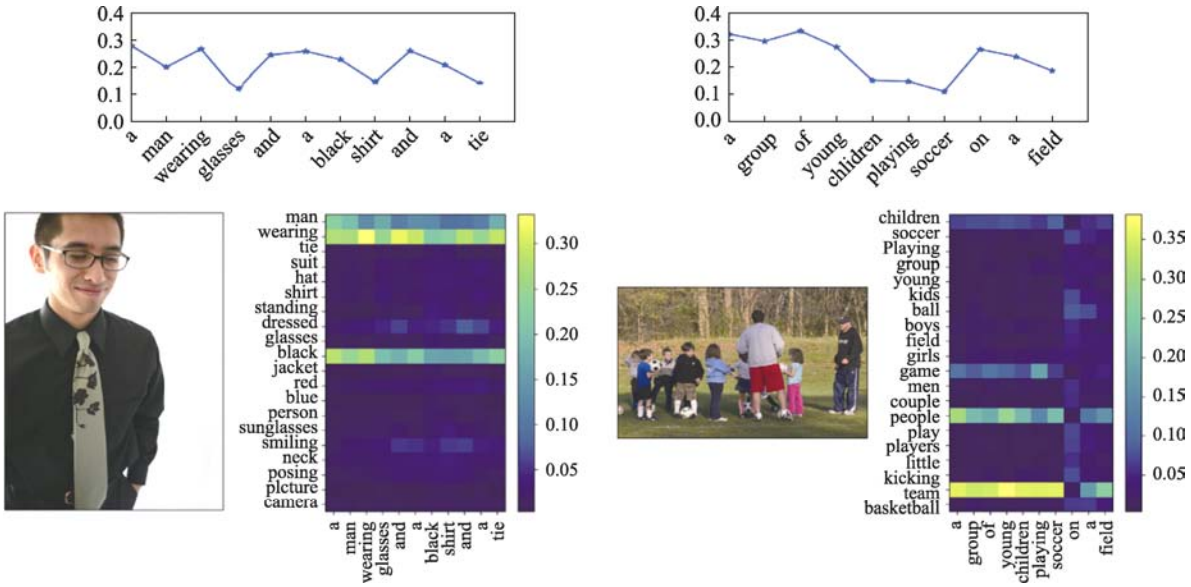


图 9 场景语义注意的可视化

5.4.3　生成描述示例

为了能够更加直观地证明本文的模型能够生成良好的描述语句，图 10 展示了本文模型生成的一些描述示例. 可以直观地看到，本文的两个模型在各种的图像场景下都可以生成良好的描述. 并且相比之下，结合场景语义信息的 VASS 模型生成的语句更加准确.

当然在引入场景语义信息后也可能带来一些其他的干扰信息而造成生成不好的结果，图 11 展示了一些生成错误或者不好的示例. 但从各项评价指标得分中可以看出，添加场景语义信息后在整体上能够生成更加准确的描述.

此外，在附录中列出了更多的可视化示例和本文模型生成的描述来证明本文模型的有效性.

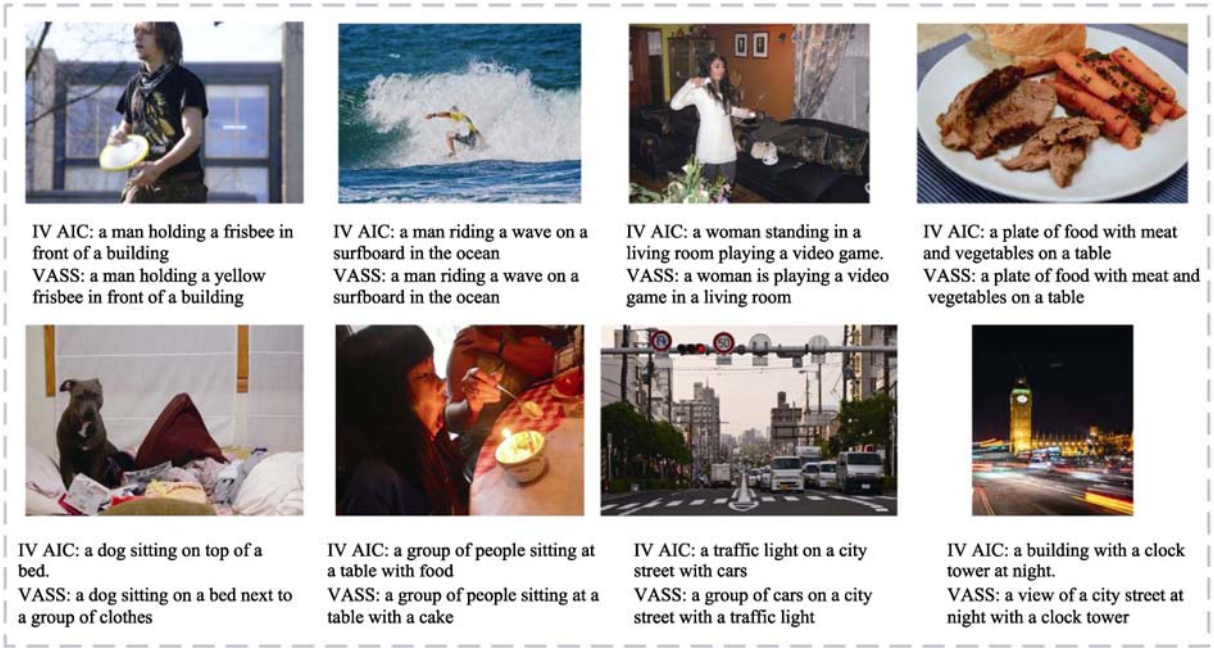


图 10　生成描述示例

图 11　添加场景信息后生成结果不好的示例

6　结　论

本文提出了一种结合图像视觉特征和场景语义信息的图像描述生成方法. 首先，通过引入聚焦强度系数，来改进视觉注意机制，自动地调节注意机制的聚焦强度. 然后，将图像的场景语义信息整合到模型中，通过注意机制来确定每个时间步模型关注的视觉信息和场景语义信息，以此来引导模型生成更加准确且符合场景主题的描述. 最后，在 MSCOCO 和 Flickr30k 数据集上进行了实验评估. 结果表明本文方法在各项评估指标上都明显优于基线方法，并且与最近的许多先进模型相比，也展现出了明显的性能优势.

参　考　文　献

[1] Mao J, Xu W, Yang Y, et al. Deep captioning with multimodal recurrent neural networks (m-RNN). arXiv preprint arXiv: 1412.6632, 2014

[2] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator//Proceedings of the IEEE Conference on

Computer Vision and Pattern Recognition (CVPR). Boston, USA, 2015: 3156-3164

[3] Karpathy A, Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA, 2015: 3128-3137

[4] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv: 1406.1078, 2014

[5] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv: 1409.0473, 2014

[6] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks//Advances in Neural Information Processing Systems (NIPS). Montreal, Canada, 2014: 3104-3112

[7] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention//Proceedings of the International Conference on Machine Learning (ICML). Lille, France, 2015: 2048-2057

[8] Chen L, Zhang H, Xiao J, et al. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA, 2017: 6298-6306

[9] Anderson P, He X, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA, 2018: 6077-6086

[10] Lu J, Yang J, Batra D, et al. Neural baby talk//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA, 2018: 7219-7228

[11] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. Journal of Machine Learning Research, 2003, 3(1): 993-1022

[12] Lu J, Xiong C, Parikh D, et al. Knowing when to look: Adaptive attention via a visual sentinel for image captioning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA, 2017: 3242-3250

[13] You Q, Jin H, Wang Z, et al. Image captioning with semantic attention//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA, 2016: 4651-4659

[14] Dai J, Li Y, He K, et al. R-FCN: Object detection via region-based fully convolutional networks//Advances in Neural Information Processing Systems (NIPS). Barcelona, Spain, 2016: 379-387

[15] Gu J, Cai J, Wang G, et al. Stack-captioning: Coarse-to-fine learning for image captioning//Proceedings of 32nd AAAI Conference on Artificial Intelligence (AAAI). New Orleans, USA, 2018: 6837-6844

[16] Jiang W, Ma L, Jiang Y G, et al. Recurrent fusion network for image captioning//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 499-515

[17] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets//Advances in Neural Information Processing Systems (NIPS). Montreal, Canada, 2014: 2672-2680

[18] Ranzato M A, Chopra S, Auli M, et al. Sequence level training with recurrent neural networks. arXiv preprint arXiv:1511.06732, 2015

[19] Rennie S J, Marcheret E, Mroueh Y, et al. Self-critical sequence training for image captioning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA, 2017: 1179-1195

[20] Dai B, Fidler S, Urtasun R, et al. Towards diverse and natural image descriptions via a conditional GAN//Proceedings of the IEEE International Conference on Computer Vision (CVPR). Honolulu, USA, 2017: 2970-2979

[21] Mirza M, Simon O. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784, 2014

[22] Vedantam R, Lawrence Zitnick C, Parikh D. CIDEr: Consensus-based image description evaluation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA, 2015: 4566-4575

[23] 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 (ACL). Philadelphia, USA, 2002: 311-318

[24] 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. Beijing, China, 2005: 65-72

[25] Lin C Y. ROUGE: A package for automatic evaluation of summaries//Proceedings of the ACL Workshop on Text Summarization Branches Out. Baltimore, USA, 2004: 74-81

[26] Anderson P, Fernando B, Johnson M, et al. SPICE: Semantic propositional image caption evaluation//Proceedings of the European Conference on Computer Vision (ECCV). Amsterdam, The Netherlands, 2016: 382-398

[27] Krishna R, Zhu Y, Groth O, et al. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision, 2017, 123(1): 32-73

[28] Kingma D P, Ba J. ADAM: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014

[29] Liu D, Zha Z J, Zhang H, et al. Context-aware visual policy network for sequence-level image captioning. arXiv preprint arXiv: 1808.05864, 2018

[30] Shuster K, Humeau S, Hu H, et al. Engaging image captioning via personality//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA, 2019: 12516-12526

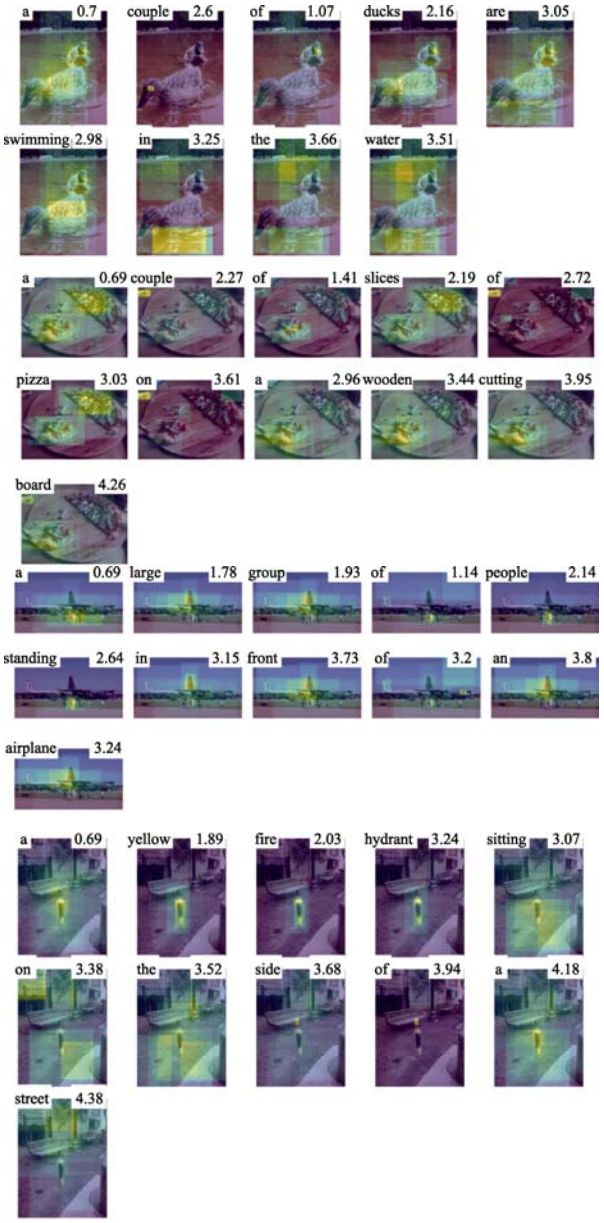
[31] Fu K, Jin J, Cui R, et al. Aligning where to see and what to tell: Image caption with region-based attention and scene-specific contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2321-2334

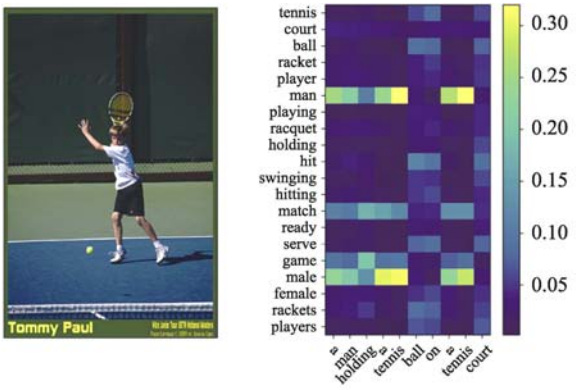
[32] Gu J, Wang G, Cai J, et al. An empirical study of language cnn for image captioning//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy, 2017: 1222-1231

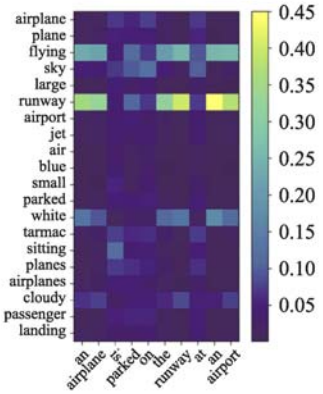
[33] Wu Q, Shen C, Wang P, et al. Image captioning and visual question answering based on attributes and external knowledge. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(6): 1367-1381

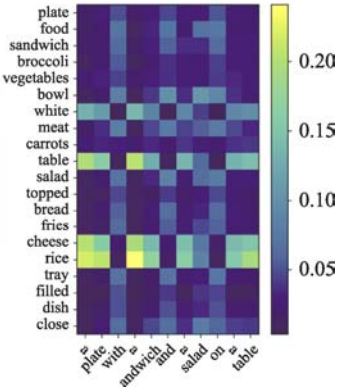
附录.

1. 本文方法生成图像描述的视觉注意可视化实例（每幅图像左上角为当前时间步生成的单词，右上角为当前时间步的聚焦强度系数）.

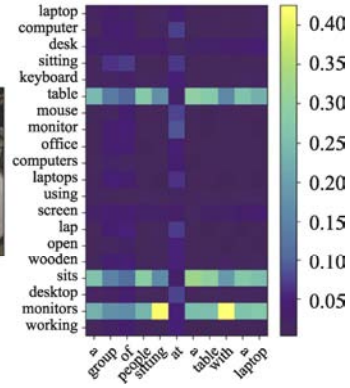


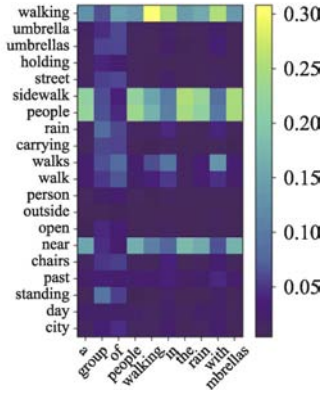


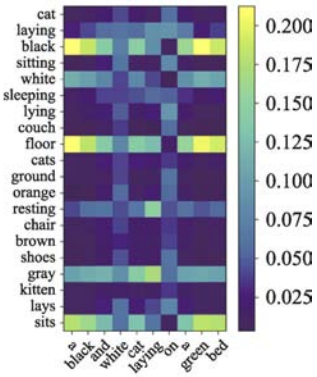


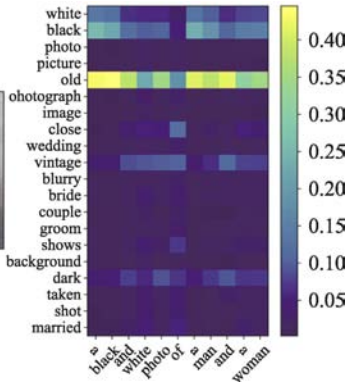


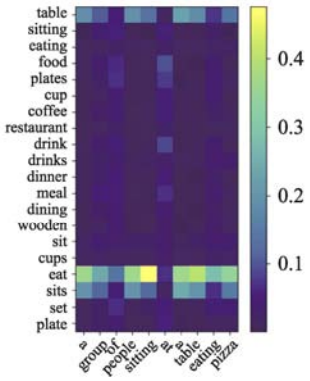
2. 本文方法生成图像描述的语义注意可视化实例.





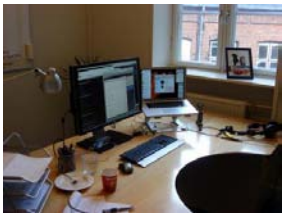



3. 本文方法生成的描述与原始参考描述对比的实例.

图像	本文方法生成的描述	参考描述
	**IVAIC:** a group of men playing a video game in a living room. **VASS:** a group of men playing a video game in a living room.	1. a man standing in a living room holding a Nintendo Wii game controller. 2. three people with cups on the couch and one with remote standing. 3. a man holding a motion controlled video game controller. 4. man with video game controller in living room with onlooker seated nearby. 5. a man playing a video game while two men sit on a couch.
	**IVAIC:** a man riding a wave on a surfboard in the ocean. **VASS:** a man riding a wave on a surfboard in the ocean.	1. a surfer is moving through a small wave. 2. a person riding a surf board on a wave. 3. a man surfs the waves in the ocean. 4. an image of a guy in the water on surfboard. 5. a man on a surfboard riding a wave.
	**IVAIC:** a group of colorful umbrellas in front of a building. **VASS:** a group of colorful umbrellas sitting in front of a building.	1. a lot of blue and yellow umbrellas sitting under a clock. 2. blue and white umbrellas outside building with similar awning. 3. a group of yellow and blue umbrellas near a building clock. 4. many blue and yellow umbrellas are shown next to a building. 5. a bunch of umbrellas sitting in front of a building.
	**IVAIC:** a couple of women sitting on a bench looking at their cell phones. **VASS:** two women sitting on a bench looking at their cell phones.	1. two women sitting and looking at a cell phone. 2. two people siting on a curb with a cell phone. 3. two women are sitting outside and looking at a phone. 4. two women discussing a cell phone while sitting on a garden edge. 5. a woman shows a man her cellphone while sitting.

续表

图像	本文方法生成的描述	参考描述
	**IVAIC:** a group of people sitting at a table with wine glasses. **VASS:** a group of people sitting at a table with wine glasses.	1. a man and two women standing around a wooden table. 2. a woman pours a glass of wine for a man. 3. a man and two women standing near a table with wine glasses. 4. a women pours wine into a glass for a man and another woman. 5. two people are watching as a woman pours a glass of wine.
	**IVAIC:** a desk with a laptop computer sitting on top of it. **VASS:** a desk with a laptop computer sitting on top of it.	1. a desk with a cup plate laptop monitor and keyboard. 2. a laptop sitting next to a monitor, keyboard and a mouse. 3. a laptop and a desktop monitor are displayed on top of the desk. 4. large office desk with computers near a window. 5. a desk with a laptop, second monitor and keyboard.
	**IVAIC:** a bike parked next to a bridge over the water. **VASS:** a bike parked next to a bridge over a body of water.	1. a bike is parked alongside the lake shore. 2. a bike is parked on the grass in front of the lake. 3. a bicycle is parked on the lawn across from a bridge. 4. mountain bike parked on grass near edge of water. 5. a bike sits parked next to a body of water.
	**IVAIC:** a person sitting at a table with a plate of pizza. **VASS:** a man sitting at a table with a plate of food.	1. a person in a silly shirt sits at a table so they can eat some pizza. 2. a person sitting at a table with some food. 3. two plates of pizza are served with waters at a wooden table. 4. two plates of pizza with someone about to take a slice. 5. a German guy in a funny shirt eats pizza with friends.

LI Zhi-Xin, Ph.D., professor, Ph.D. supervisor. His research interests include image understanding, machine learning and cross-media computing.

WEI Hai-Yang, M.S. candidate. His research interests include image understanding and machine learning.

HUANG Fei-Cheng, M.S. candidate. His research interests include image understanding and machine learning.

ZHANG Can-Long, Ph.D., professor. His research interests include pattern recognition and target tracking.

MA Hui-Fang, Ph.D., professor. Her research interests include machine learning and data mining.

SHI Zhong-Zhi, professor, Ph.D. supervisor. His research interests include artificial intelligence, machine learning, neural computing and cognitive science.

Background

Image captioning is to generate a reasonable natural language description for a given image according to its content. It is an important research field of artificial intelligence, which is mainly used in image and text retrieval, disabled people's life assistance, and so on. This is a task that combines computer vision and natural language processing. Inspired by machine translation work, the current image captioning models are based mainly on the combination of the encoder-decoder framework and the attention mechanism. This type of method generally includes the following four steps:

(1) Encoding the input image using Convolutional Neural Network (CNN).

(2) Generating words with a Recurrent Neural Network

(RNN) based on the output of step (1).

(3) The attention mechanism focus on the salient regions of the image in each time step of the RNN.

(4) Dynamically updating the generated sentence until the end of the RNN decoding.

This type of method generally uses backpropagation for end-to-end training and has achieved good results in the task of image captioning.

Although some progress has been made, there are still some problems in the existing image captioning system. First, in the process of captions generation, the model needs to pay different focus intensity attention to the images for the generation of different words. However, the current visual attention mechanism cannot adjust the focus intensity on the image. Second, most image captioning systems lack guidance of scene semantic information. Although the scene semantic information is very important for image captioning, it is rarely utilized in the state-of-the-art methods. In this paper, we introduce a focus intensity coefficient into the attention mechanism. It can automatically adjust the focus intensity of the attention mechanism to obtain more accurate visual information. In addition, we incorporate the scene semantic information of the image into the model, and combine the visual information and the scene semantic information to guide the model to generate captions. The experimental results show that ours approach can generate accurate captions and achieve superior performance.

This work is supported by the National Natural Science Foundation of China (Nos. 61966004, 61663004, 61866004, 61762078), the Guangxi Natural Science Foundation (Nos. 2019GXNSFDA245018, 2018GXNSFDA281009, 2017GXNS FAA198365), the Research Fund of Guangxi Key Lab of Multi-source Information Mining and Security (16-A-03-02, MIMS18-08, MIMS19-02), the Guangxi “Bagui Scholar” Teams for Innovation and Research Project, Guangxi Collaborative Innovation Center of Multi-source Information Integration and Intelligent Processing.