文章编号：2095-2163(2020)01-0012-10 中图分类号：TP391 文献标志码：A

融合多标签和双注意力机制的图像语义理解模型

吴 倩[1]，应 捷[2]，黄影平[1]，杨海马[3]，胡文凯[1]
（1 上海理工大学 光电信息与计算机工程学院，上海 200093；2 上海理工大学 测试技术与信息工程研究所，上海 200093；
3 上海市现代光学系统重点实验室（上海理工大学），上海 200093）

摘 要：针对现有图像语义理解模型存在描述不充分以及视觉属性冗余的问题，提出了一种带有视觉三元组标签且能够挖掘潜在信息的图像语义理解模型 VT-BLSTM。首先，使用卷积神经网络提取图像的全局特征和视觉三元组标签；其次，构建双向长短期神经网络，并利用改进的双注意力模型分别获得动态视觉特征和动态文本特征，融合该两者特征得到视觉语义上下文；最后，融合视觉语义上下文、视觉三元组和神经网络隐含层特征，比较前向和后向长短期神经网络的输出结果，得到对应时刻的单词。在 Flickr8K 和 Flickr30K 数据集上的实验结果表明，本文提出的双注意力模型 VT-BLSTM 能够自主地选择文本特征和视觉特征参与生成单词的比例，并且结合历史时刻和未来时刻获得更丰富的视觉信息，在少量视觉属性的前提下也能生成质量较高的句子，并在多个统计指标上超过同类其他方法。

关键词：图像语义理解；双向长短期记忆网络；视觉属性；注意力机制

Image captioning with multi-label and dual-attention

WU Qian[1], YING Jie[2], HUANG Yingping[2], YANG Haima[3], HU Wenkai[1]
(1 School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China; 2 Institute of Testing Technology and Information Engineering(University of Shanghai for Science and Technology, Shanghai 200093, China; 3 Shanghai Key Laboratory of Modern Optical System(University of Shanghai for Science and Technology), Shanghai 200093, China)

【Abstract】Aiming at the problem that the existing image caption models based on attention are inadequately described and have redundant visual attributes, this paper proposes a VT-BLSTM model with visual triples which can mine potential information. Firstly, the convolutional neural network is used to extract the global features of the image and visual triples. Then, a bi-directional long-short term memory network is constructed, the improved attention model is used to obtain dynamic visual features and dynamic text features respectively, and the visual semantic context is obtained by integrating the two features. Finally, combining visual semantic context, visual triples and hidden layer features of neural network, the output results of forward and backward long short-term memory network are compared to obtain words at the corresponding moment. Results on Flickr8K and Flickr30K datasets show that the VT-BLSTM can select the proportion of text features and visual features, combine historical moments and future moments to obtain abundant visual information. It can also generate high-quality sentences with a small number of visual attributes, and surpass other similar methods in multiple statistical indicators.

【Key words】image captioning; bi-directional long short-term memory model; visual attributes; attention mechanism

0 引 言

图像语义理解融合了计算机视觉与自然语言处理两个方向，是利用人工智能算法解决多模态、跨领域问题的典型代表。通过计算机实现图像语义理解，再通过语音播报系统传送语义信息，不仅能够帮助视觉障碍者获取图像信息，而且能辅助驾驶，便利人们的日常生活。另外，图像语义理解的研究成果可用于图像检索、智能儿童早教机、智能机器人等方面[1-20]，具有较为广泛的应用前景和现实意义。

随着深度学习的发展，人们提出了基于编码-解码的图像语义理解模型[5-19]，这种模型通常采用卷积神经网络（Conventional Neural Networks，CNN）提取图像特征，再通过循环神经网络（Recurrent Neural Network，RNN）构建语言模型生成文本。Vinyals 等人[7]提出了 NIC 模型，该模型在 m-RNN[5]的基础上将 RNN 部分替换成性能更好的长短期记忆网络（Long Short-Term Memory，LSTM），并且图像的特征只在 LSTM 的第一个时刻输入。这种改进不仅能够减少模型参数，而且还能够解决神经网络训练中的梯度消失和梯度爆炸问题。

基金项目：国家自然科学基金（61701296）；上海市自然科学基金（17ZR1443500）。

作者简介：吴 倩（1996-），女，硕士研究生，主要研究方向：深度学习、图像理解；应 捷（1973-），女，博士，副教授，主要研究方向：图像处理；黄影平（1966-），男，博士，教授，主要研究方向：计算机视觉、模式识别；杨海马（1979-），男，博士，副教授，主要研究方向：模式识别；胡文凯（1998-），男，本科生，主要研究方向：图像语义理解。

收稿日期：2019-10-13

单纯的编码-解码模型不能很好地解释生成的单词与图像中对象的位置关系，因此基于注意力机制的图像语义理解模型随之产生。Kelvin 等人[8]首次将注意力机制引入图像语义理解模型，提出了 soft-Attention 和 hard-Attention，并取得了较大的成果，为后续注意力的发展提供了基础。Qu 等人[9]使用 CNN 提取了更加丰富的特征，如颜色，轮廓等，并且将这些额外的视觉特征加入注意力机制中，使得生成的描述更加准确。Chen 等人[10]提出的 SCA-CNN 模型不仅考虑图像平面中物体的位置，而且将提取到的图像特征通道信息也加入其中，并且验证了先使用空间信息再使用通道信息的模型效果更好。Marcella 等人[11]提出了一种结合显著图和上下文的注意力机制，并且验证了显著图在生成文本中的重要性。吕凡等人[12]引入注意力反馈机制，不断强化图像与文本的匹配关系，在一定程度上解决了注意力分散的问题。Jiang 等人[13]改进了 LSTM 的内部结构，分别在输入门和输出门中增加存储单元，使得 LSTM 单元在必要的时候能够读取这些之前没有计算到网络中信息。此外，该模型还将前一时刻的 attention 信息引入下一时刻，解决了模型重复关注的问题。

为了解决传统编码-解码可能出现预测物体不正确的问题，You 等人[14]首次提出了基于属性预测的图像语义理解模型，即首先使用多标签分类获取图像的视觉属性，再结合视觉属性和图像特征生成文本。随后，Zhe 等人[15]基于标签检测提出了 SCN-LSTM 模型，该模型中 LSTM 每个时刻的权值矩阵被扩展为一个与标签相关的权值矩阵的集合，根据标签与图像的相关性来决策生成的单词。这种模型在语言概念被修改时，仍然能够生成正确且流畅的句子，但是整个模型的参数会随着标签数量的增大而增加，训练十分困难。Zhao 等人[16]提出了一种多模型融合的图像语义理解方法，采用 CNN 提取图像的全局特征与图像的视觉属性，在 LSTM 的 0 时刻输入全局特征向量，在后续的每个时刻均输入前一时刻产生的单词编码向量、图像的视觉属性与时变的语句特征向量。这种方法解决了因时间的推移造成的信息丢失问题，但存在大量冗余信息。He 等人[17]提出了一种融合视觉属性和注意力机制的图像语义理解模型，使用 DenseNet 提取图像的全局特征和视觉属性，并且将 DenseNet 网络的参数共享到注意力模块以及文本生成模块，该模型增强了属性预测模块与文本生成模块的关系，减小了模型的复杂度。

现有的研究成果表明，基于注意力机制的方法和基于属性预测的方法均能提高模型的预测效果。但现有的注意力模型只关注图像特征而没有考虑到之前生成的文本，也没有考虑前后注意力区域之间的联系，并且多采用单向 LSTM，不能很好地结合未来的信息生成文本。此外，基于属性预测的方法均采用 CNN 进行多标签分类，选取概率最大的 K 个属性标签作为生成文本的前提，这种方法能够获得丰富的视觉信息但存在大量冗余信息。

为了解决上述问题，本文提出了一种带有视觉属性三元组且能够挖掘潜在信息的图像语义理解模型 VT-BLSTM(Visual-text Bi-directional Long Short-Term Memory)。该算法使用 VGG19 的卷积层提取图像的全局特征，并且为每个图像选定<物体，动作，场景>三元组，训练多标签分类模型得到每张图像的属性三元组信息，减少额外视觉属性的引入。其次，构建双向 LSTM 神经网络，在传统视觉注意力的基础上，加入文本注意力机制，通过变量b控制当前输入单词参与生成文本的比例，决定此时生成的单词更多地依靠图像视觉信息还是之前生成的文本，得到视觉语义上下文。另外，视觉注意力模型的输入向量在原有基础上扩展一个上一时刻注意力模型的中间状态，使得模型在选择关注区域时充分考虑历史信息。最后，融合视觉三元组属性、视觉语义上下文和 LSTM 隐含层特征，预测文本信息。经过实验研究证明，本文提出的 VT-BLSTM 模型不仅引入未来信息，改善了单向 LSTM 生成文本仅依靠之前生成单词的缺陷，使之生成更加丰富、更符合图像的描述，而且考虑之前关注的区域，增强了前后时刻关注区域的联系。同时，解决了视觉属性冗余的问题，在不降低模型正确率的同时减小了模型的复杂度。

1　基于注意力机制的图像语义理解模型 VT-BLSTM

1.1　VT-BLSTM 模型整体框架

本文提出的基于视觉注意力的图像语义理解模型 VT-BLSTM 包括图像视觉概念提取模块、视觉语义上下文提取模块和文本生成模块三个部分，如图 1 所示。视觉概念提取模块使用 VGG19 的卷积层提取图像的全局特征 V，并且修改 VGG19 的最后一个全连接层，利用多示例学习获得<物体，动作，场景>三元组 V_{attr}。视觉语义上下文提取模块由双向 LSTM 组成，图 1 中 LSTM _f 和 LSTM_b 分别表示前向 LSTM 和后向 LSTM，att_unit 表示注意力单元，用于获取图像的视觉语义上下文。文本生成模块由一

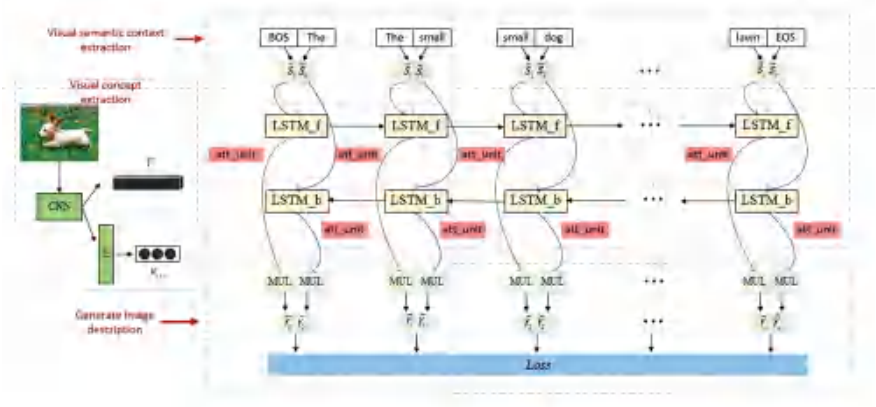


图 1 VT-BLSTM 模型框架图

Fig. 1 VT-BLSTM model framework

个多模态融合层组成,其输入包括视觉属性三元组 V_{attr}、LSTM 的隐含状态 h、视觉语义上下文 C^O 和输入单词 S, 根据多模态层的输出计算模型 VT-BLSTM 的误差,利用反向传播更新模型权重。

1.2 图像视觉概念提取模块

本文的图像视觉概念是指图像的全局特征和图像的视觉属性三元组<物体,动作,场景>。图像语义理解模型中常使用在 ImageNet 上预训练的 VGG19 进行特征提取,将调整为 224 × 224 × 3 的图像输入 VGG19,使用其第一个全连接层之前的网络提取图像全局视觉特征 V, 那么 $V = \{V_1, V_2, V_3, \cdots, V_L\}$, $V_i \in R^D$,即将图像 i 划分为 $L(196)$ 块区域,每个区域 V_i 都是一个 $D(512)$ 维的向量。

图像的视觉属性提取通常由多示例学习完成[18-19],为了得到图像的视觉属性三元组,本文沿用文献[18]的思想,并在其基础上做相应的修改。文献[18]选取了出现次数最多的1 000个属性词构建属性词典,当测试图片对应的某个属性概率大于 0.5 时,该属性被认为是图片拥有的属性。这种方法能够获得丰富的图像视觉属性,但存在属性之间关联度过高,比如“green”和“grass”,以及仅依靠语法而不需要看图就能生成的单词,比如“in”、“the”。针对上述问题,本文提出了使用视觉属性<物体,动作,场景>作为属性前提,参与后续文本的生成。该方法不仅能够在视觉属性预测时极大地减少模型的训练参数,而且在生成文本时也减小了模型的复杂度。

首先构建属性词典,经过数据分析,本文分别选择数据集中出现最多的前 80 种物体、50 种动作以及 40 个场景作单词为图像的视觉属性词典 W, 那么 $W = \{w_1, w_2, w_3, \cdots, w_c\}$, $c = 170$。其次,修改 VGG19 的最后一个全连接层,设置结点数为 170,计算图像包含单词 w 的概率。给定一张图像 i, 假设每个区域 j 包含单词 w 的概率为 p_{ij}^w, 那么图像 i 包含单词 w 的概率 p_i^w 为:

$$p_i^w = 1 - \prod_{j \in i} (1 - p_{ij}^w), \tag{1}$$

其中, p_{ij}^w 的计算公式为:

$$p_{ij}^w = \frac{1}{1 + \exp(-(V_w \varphi(b_{ij}) + u_w))}, \tag{2}$$

其中, $\varphi(b_{ij})$ 表示图像 i 中,第 j 区域的 $fc7$ 输出向量,$\varphi(b_{ij}) \in R^{4096}$, V_w 和 u_w 都是与单词 w 相关的权重和偏置。

最后,计算模型的交叉熵损失函数并更新权重,利用训练好的模型得到每张图片的<物体,动作,场景>属性。假设数据集含有 N 张图片,每张图片 i 的标签 $y_i = \{y_{i1}, y_{i2}, y_{i3}, \cdots\cdots, y_{ic}\}$, 当 y_i 包含属性 k 时, $y_{ik} = 1$, 否则, $y_{ik} = 0$。那么,本模块中属性预测损失函数 $L(I)$ 的计算公式为:

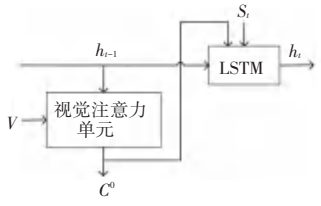
$$L(I) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{c} (y_{ik} \log(p_i^{w_k}) + (1 - y_{ik}) \log(1 - p_i^{w_k})). \tag{3}$$

1.3 视觉语义上下文提取模块

传统的注意力机制单指视觉注意力[8,17]即前一时刻 LSTM 的隐含层状态决定当前关注图像的哪些区域,该机制极大地提高了生成文本的准确率。但是这种机制存在以下两点问题。一方面,某些单词的生成并不依靠图像,而是依靠之前生成的文本。比如已经生成某句子片段“A girl is sitting in front”,那么后面的“of”就可以根据前面生成的单词决定。另一方面,关注的图像区域之间可能存在某种联系。

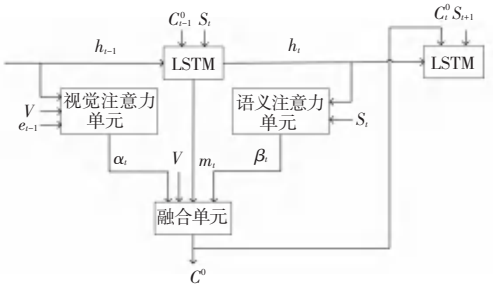
针对上述问题,本文在文献[8]的基础上提出了一种新的注意力机制提取上下文信息,包含视觉注意力单元、语义注意力单元和融合单元,其生成上

下文的结构与传统注意力的结构对比如图 2 所示。视觉注意力单元用于获取当前时刻每个图像区域的权重 $\alpha, \alpha = \{\alpha^1, \alpha^2, \alpha^3, \cdots, \alpha^L\}, \alpha \in R$。语义注意力单元用于获取当前时刻保留文本信息的概率 $\beta, \beta \in (0,1)$。融合单元先根据 α 和图像全局特征向量 $\boldsymbol{V}$ 得到视觉上下文 C^V，再根据“语义门” g 和 LSTM 存储单元 m 得到语义上下文 C^T，最后融合 C^V、C^T 以及 β 得到视觉语义上下文 C^O。这种新的注意力机制不仅考虑了历史关注的区域，而且引入文本保留概率 β，使得模型能够自主地选择多关注图像还是之前生成的文本。



（a）传统方法

（a）Traditional method



（b）本文方法

（b）The proposed in the paper

图 2　传统及本文注意力视觉上下文提取图

Fig. 2　Traditional and the proposed attention visual context extraction graph

在传统的视觉注意力单元中，t 时刻视觉上下文 C_t^O 计算公式如下：

$$e_t^i = f_{att}(V_i, h_{t-1}); \tag{4}$$

$$\alpha_t^i = \frac{\exp(e_t^i)}{\sum_{k=1}^{L} \exp(e_t^i)}; \tag{5}$$

$$C_t^O = \sum_{i=1}^{L} \alpha_t^i V_i. \tag{6}$$

根据公式(4)获得 t 时刻图像区域权重的中间状态 e_t^i，其中 f_{att} 是一个多层感知机，并且按照公式(5)进行归一化，得到权重 $\alpha, \alpha = \{\alpha^1, \alpha^2, \alpha^3, \cdots, \alpha^L\}$。将每个区域的权重 α^i 与局部特征相乘，得到视觉上下文 C_t^O。

在本文的视觉注意力单元中，t 时刻图像区域权重 a_t 计算公式如下：

$$e_t = \begin{cases} W_0 V, & t = 0; \\ W_V V + W_h h_{t-1} + e_{t-1} + b, & t > 0. \end{cases} \tag{7}$$

$$a_t = softmax(W_e e_t) \tag{8}$$

其中，e_t 表示 t 时刻关注区域权重的中间状态，$t = 0$ 时刻的状态由全局图像特征得到，$t > 0$ 时，该中间状态由全局图像特征 V、前一时刻 LSTM 隐含层 h_{t-1} 和前一时刻的中间状态决定。这种改进的视觉注意力机制将中间状态 e_t 看成一种记忆单元，能够记忆历史时刻关注的区域信息。公式(7)中的 W_0、W_1 和 b，以及公式(8)中的 W_e，均是可训练参数。在 $t = 0$ 时刻，LSTM 的隐含层状态 h_0 和存储细胞状态 c_0 由输入图像的平均全局特征决定：

$$h_0 = f_{init,h}\left(\frac{1}{L}\sum_{i=1}^{L} F_i\right), \tag{9}$$

$$c_0 = f_{init,c}\left(\frac{1}{L}\sum_{i=1}^{L} F_i\right), \tag{10}$$

语义注意力单元在 t 时刻生成文本保留概率 β_t 的计算公式如公式(11)所示：

$$\beta_t = softmax(W_e(W_s S_t + W_h h_t)), \tag{11}$$

其中，W_e 与公式(8)中的 W_e 相同；W_h 与公式(7)中的 W_h 相同；W_s 是与语句相关的可训练参数。

融合单元中，t 时刻的视觉上下文 C_t^V 的计算可参考公式(6)，语义上下文 C_t^T 的计算公式如公式(12)所示：

$$C_t^T = g_t \cdot \tanh(m_t), \tag{12}$$

其中，m_t 表示当前时刻 LSTM 的存储单元；σ 表示 sigmoid 激活函数；g_t 表示一个“语义门”，用来控制当前单词遗留下来的概率，计算时会用到如下公式：

$$g_t = \sigma(W_x S_t + W_h h_{t-1} + W_c C_{t-1}^O), \tag{13}$$

根据前面得到的视觉上下文 C_t^V，语义上下文 C_t^T 和文本保留概率 β_t，可根据公式(14)计算视觉语义上下文 C_t^O，即：

$$C_t^O = \beta \cdot C_t^T + (1 - \beta) \cdot C_t^V. \tag{14}$$

本文提出的注意力机制引入了视觉注意力存储单元和语义注意力模块，视觉注意力存储单元能够保留历史时刻关注的区域，语义注意力模块能够自主地选择关注之前生成的文本，使得模型的在每个时刻关注的区域更加准确，生成更符合图像的描述。

1.4　文本生成模块

文本生成模块由一个多模态融合层组成，根据输入单词 S、LSTM 的隐含状态 h、视觉语义上下文 C^O 和视觉属性三元组 V_{attr}，预测图像对应的文本。

由于本文采用了双层 LSTM,因此会从 2 个方向生成文本。在训练时,根据前向 LSTM 和后向 LSTM 的预测损失之和更新权重。在测试时,分别计算前后向 LSTM 生成文本的概率和,选择概率和最大的文本作为模型最终的输出。以训练过程中前向 LSTM 为例,文本生成模块的实现流程如下:

(1)计算输入单词的编码向量。图 1 中$\overrightarrow{S_1}$,$\overrightarrow{S_2}$,……,$\overrightarrow{S_{n-1}}$表示一个句子中每个单词对应的词向量,$\overrightarrow{S_0}$和$\overrightarrow{S_n}$分别表示 EOS(开始)标签和 BOS(结束)标签对应的向量。假设 t 时刻输入的单词为 w_t,由 w_t 生成 $\boldsymbol{S}_t$ 的过程如下:

$$w_t \xrightarrow{one-hot} \boldsymbol{o}_t = \begin{bmatrix} 0 \\ 1 \\ \cdots \\ 0 \end{bmatrix} \xrightarrow{embedding} \boldsymbol{S}_t = \begin{bmatrix} e_1 \\ e_2 \\ \cdots \\ e_m \end{bmatrix}$$

过程中,$\boldsymbol{o}_t$ 表示 w_t 的独热编码,$\boldsymbol{o}_t \in R^V$,V 表示词汇库的大小;使用 word embedding 降维,得到 $\boldsymbol{S}_t$,$\boldsymbol{S}_t \in R^M$,其中 M 表示词嵌入向量的维度。

(2)LSTM 的隐含层状态更新。其中 $\boldsymbol{h}_t$ 表示 t 时刻 LSTM 单元的隐含层向量,$\boldsymbol{h}_t \in R^H$。用公式表示为:

$$f_t = \sigma(\boldsymbol{W}_f[\boldsymbol{h}_{t-1}, \boldsymbol{S}_t, C^O_{t-1}] + b_f); \quad (15)$$

$$i_t = \sigma(\boldsymbol{W}_i[\boldsymbol{h}_{t-1}, \boldsymbol{S}_t, C^O_{t-1}] + b_i); \quad (16)$$

$$m_i = \tanh(\boldsymbol{W}_c[\boldsymbol{h}_{t-1}, \boldsymbol{S}_t, C^O_{t-1}] + b_c); \quad (17)$$

$$m_t = f_t * m_{t-1} + i_t * m_i; \quad (18)$$

$$o_t = \sigma(\boldsymbol{W}_o[\boldsymbol{h}_{t-1}, \boldsymbol{S}_t, C^O_{t-1}] + b_o); \quad (19)$$

$$h_t = o_t * \tanh(m_t). \quad (20)$$

其中,m_{t-1} 和 m_t 分别表示 $t-1$ 时刻和 t 时刻的细胞状态;σ 表示 sigmoid 函数;f_t 表示遗忘门函数,用于控制前一时刻细胞保留下来的信息;i_t 表示输入门函数,用于更新当前时刻的信息;o_t 表示输出门函数,控制更新后细胞状态的输出。$\boldsymbol{W}_f$、$\boldsymbol{W}_i$、$\boldsymbol{W}_o$ 分别表示遗忘门、输入门、输出门的参数矩阵,b_f、b_i、b_o 分别为对应的偏置。$*$ 表示矩阵点乘。

(3)根据上述得到的单词编码向量 $\boldsymbol{S}$、LSTM 的隐含状态 h、视觉语义上下文 C^O 和视觉属性三元组 V_{attr} 生成文本 Y_t。其过程如下:

$$\begin{matrix} \boldsymbol{h}_t \\ C^0_t \\ \boldsymbol{S}_t \\ V_{attr} \end{matrix} \xrightarrow{mul} \boldsymbol{r}_t = \begin{bmatrix} e_1 \\ e_2 \\ \cdots \\ e_m \end{bmatrix} \xrightarrow{f_c} \boldsymbol{y}_t = \begin{bmatrix} k_1 \\ k_2 \\ \cdots \\ k_m \end{bmatrix} \xrightarrow{softmax} Y_t$$

在单词编码阶段,先采用独热编码得到 V 维向量,后采用 embedding 得到 M 维向量,因此生成 Y_t 需要经过 3 个步骤。首先,将上述 4 个特征向量输入多模态层,得到一个 M 维向量 $\boldsymbol{r}_t$,其计算方法如公式(21)所示:

$$\boldsymbol{r}_t = \boldsymbol{W}_{w\_h}\boldsymbol{h}_t + \boldsymbol{S}_t + \boldsymbol{W}_{c2t}\boldsymbol{C}^O_t + \boldsymbol{W}_{attr}V_{attr}, \quad (21)$$

其中,W_{w_h}、W_{c2t} 和 W_{attr} 都是可学习参数。

其次,经过一个全连接层 f_c 得到一个 V 维的向量 $\boldsymbol{y}_t$。最后,$\boldsymbol{y}_t$ 经过归一化得到每个单词的概率 $p_t, p_t \in R^V$,在词汇表中取 p_t 的最大值所对应单词作为最终输出 Y_t。

针对一张图像 I,训练模型的最终目的是得到模型中的最优参数 θ^*,即:

$$\theta^* = \arg\max_\theta \sum_{(I,S)} \log p(S \mid I;\theta), \quad (22)$$

其中,S 表示图像 I 的标定描述,θ 是模型中的自学习参数,训练时模型的损失函数为:

$$L(I,S) = -\sum_{t=0}^{N}\sum_{t=0}^{T}\log(p(\overrightarrow{w_t} \mid I, V_{attr}, \overrightarrow{w_0}, \cdots, \overrightarrow{w_{t-1}})) - \sum_{t=0}^{N}\sum_{t=0}^{T}\log(p(\overleftarrow{w_t} \mid I, V_{attr}, \overleftarrow{w_0}, \cdots, \overleftarrow{w_{t-1}})) + \lambda \|\theta\|^2, \quad (23)$$

公式(23)的第一部分和第二部分分别表示前向和后向 LSTM 的交叉熵损失函数,其中 N 为训练图像的大小,w_t 表示 t 时刻生成发单词,最后一项表示视觉语义上下文提取模块中前向和后向 LSTM 每个时刻 α 的正则项。

本文训练好的图像语义理解模型能分别根据前向 LSTM 和后向 LSTM 生成文本,最终生成的句子由最大概率和决定,如式(24)所示:

$$p(w_{1:T} \mid I, V_{attr}) = \max\left(\sum_{t=1}^{T}(p(\overrightarrow{w_t} \mid I, V_{attr})), \sum_{t=1}^{T}(p(\overleftarrow{w_t} \mid I, V_{attr}))\right). \quad (24)$$

2 实验结果与分析

2.1 实验参数设置和评价指标

本文中使用 Flickr8K 和 Flickr30K 数据集进行实验。Flickr8K 图像集中含有 6 000 张训练图像,1 000张验证图像,1 000 张测试图像,其中每张图像对应有 5 个人工标定的描述信息。Flickr30K 数据集包含 31 783 张图片,每张图片也对应 5 条标注文本,本文参考文献[6]的分割方法,将数据集分割为 29 000张训练图像、1 000张验证图像和1 000张测试图像。使用 VGG 提取图像特征时,得到特征的维度是 196 * 512,即 $L = 196$,$D = 512$。在文本预处理部分,首先删除标定语句中所有的标点符号,其次将单

词全部转换成小写字母，设置句子最大长度是 20。使用 Flickr8K 数据集时，得到有效的 29 318 条语句，其中包含 7 224 个单词，根据这些单词建立词汇库。使用 Flickr30K 数据集时，词汇库包含 18 344 个单词。在训练时使用 Adam 优化算法更新参数，Adam 的参数设定为 *alpha* = 0.001、*beta*1 = 0.9、*beta*2 =0.999，同时使用 *dropout* 防止模型过拟合。

实验中采用了 BLEU（Bilingual Evaluation understudy）和 METEOR 两种方法对生成的语句进行评价。BLEU 表示候选语句与标定语句中 n 元组共同出现的程度，是一种基于精确度的评估方法，包括 BLEU－1、BLEU－2、BLEU－3 和 BLEU－4。METEOR 指标同时考虑了整个语料库上的准确率和召回率，其结果和人工判断的结果有较高相关性。这两种评价指标得分越高表示模型能够对图像进行更加准确的语义理解，生成的语句质量越好。

2.2　视觉属性三元组提取结果

使用 NLTK 处理文本数据，选取出现次数最多的属性单词组成数据集的属性词典，其<物体，动作，场景>三元组属性词云如图 3 所示。

（a）物体词云　（b）动作词云　（c）场景词云

（a）The word cloud of object　（b）The word cloud of action　（c）The word cloud of the scene

图 3　三元组属性词云

Fig. 3　Triple attribute word cloud

根据上述属性词构建每个图像的属性标签，通过多示例学习得到图像的视觉三元组属性如图 4 所示。图 4 中右上角的单词分别表示图像对应的属性，属性后的数字表示为该属性的概率。如图 4（a）所示，该图像生成的<物体，动作，场景>三元组为<people，working，snow>，其概率分别为 0.83，0.81，0.92。图 4（b）的视觉三元组属性为<dog，running，grass>。图 4（c）的视觉三元组属性为<girl，running，grass>。图 4（d）的视觉三元组属性为<girl，playing，water>。这些生成的视觉三元组属性在文本生成时，编码为单词向量输入多模态模块，引导其生成更准确的句子表达。

（a）视觉三元组 1　（b）视觉三元组 2　（c）视觉三元组 3　（d）视觉三元组 4

（a）Visual triple 1　（b）Visual triple 2　（c）Visual triple 3　（d）Visual triple 4

图 4　视觉属性生成示意图

Fig. 4　Visual attributes schematic

为了验证视觉三元组的有效性，本文分别对比使用视觉三元组、传统视觉属性提取[18]和不使用视觉属性三种方法，实现图像的语义理解，并使用 BLEU 和 METEOR 评价指数对生成的句子进行评价。表 1 记录了上述三种不同的方法在 Flickr8K 数据集和 Flickr30K 数据集上生成文本的结果，$B@1$ ~ $B@4$ 分别表示 BLEU－1 ~ BLEU－4 评价指标，M 表示 METEOR 评价指标。

通过对比可以看出，使用了属性特征的模型效果最好，另外，在 Flickr8K 数据集下，使用传统属性提取方法和使用视觉属性三元组方法生成文本的正确率相差不大，但是在 Flickr30K 数据集下，带有视觉属性三元组的模型效果更好。不带属性特征的模型是端到端的模型，仅根据文本预测得到的句子可能会导致图像主体预测出错，在模型中引入视觉属性能够在一定程度上减小这样的错误。而传统的属性预测往往包含十几个单词，这些单词之间可能存在包含关系或者单词意思相近，会引入冗余属性，而

且使得模型的参数增加。视觉属性三元组不仅能够准确预测出图像主要属性,还能减小模型的参数,降低模型训练的复杂度。

综合以上分析,得到结论:使用视觉属性三元组的模型能够在减小模型参数的情况下,有效地引导文本生成,且不降低模型生成文本的正确率。

表 1 不同视觉属性生成文本的结果

Tab. 1 Evaluation score of the generated text under different visual attributes

评价指标	Flickr8K					Flickr30K				
	B@1	*B*@2	*B*@3	*B*@4	*M*	*B*@1	*B*@2	*B*@3	*B*@4	*M*
WithoutAttr	63.5	43.8	30.9	21.1	20.5	64.2	45.8	33.1	23.6	19.0
TraditionalAttr	**68.9**	**48.6**	33.5	23.0	21.2	**68.3**	47.8	34.5	24.1	19.7
TripleAttr	68.4	48.1	**33.7**	**23.2**	**22.6**	68.1	**48.6**	**34.8**	**24.2**	**19.8**

2.3 VT-BLSTM 模型结果与分析

设置 *batch* 为 64,迭代次数为 15K,随着迭代次数的增加,本文提出的 VT-BLSTM 模型在 Flickr30K 上的 *loss* 变化曲线如图 5 所示,由 *loss* 的变化趋势可知,VT-BLSTM 是可训练的、并且有效的。BLEU 和 METEOR 评价分数随着迭代次数的变化曲线如图 6 所示。这五种评价分数总体来说都随着迭代次数的增加呈现上升的趋势,并且在迭代到10 000次左右时,能达到稳定值。

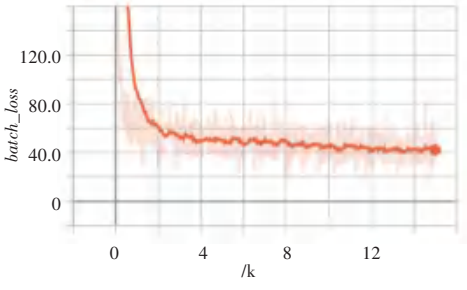


图 5 训练时模型的损失变化

Fig. 5 Training loss of the model

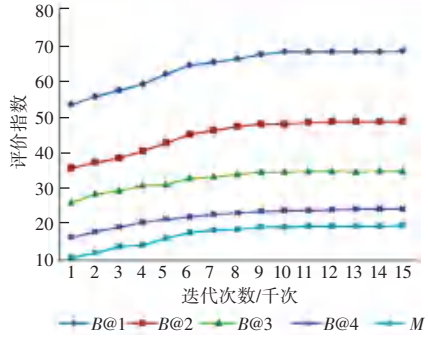


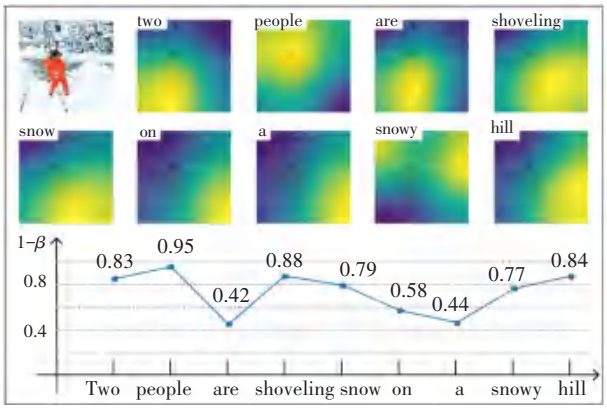
图 6 不同迭代次数下 VT-BLSTM 的性能

Fig. 6 Performance of VT-BLSTM under different iteration

图 7 是采用 VT-BLSTM 模型实现图像的语义理解,并且可视化每个时刻聚焦的区域以及视觉语义上下文模块中"视觉门"1-β 的示意图。图 7(a)~(d)的上半部分均表示每个时刻生成的文本以及对应的关注区域,高亮的部分表示关注度更高,下半部分表示生成文本时"视觉门"的变化曲线。

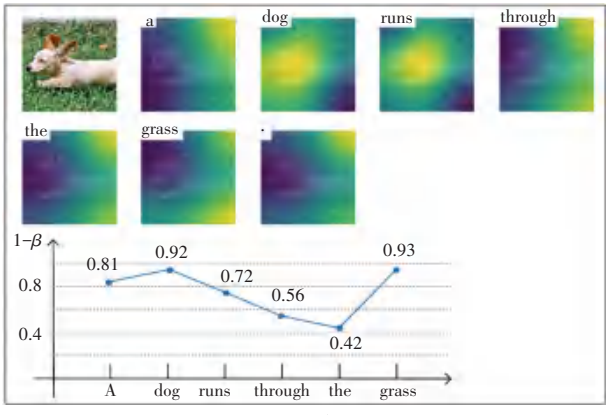
图 7(a)生成文本描述"Two people are shoveling snow on a snowy hill.",在生成单词"people"、"snow"的时候能够准确定位到人和雪地的部位。并且在"people"、"shoveling"、"snow"和"hill"对应时刻,1-β 数值较大,表明该模型更多地关注图像而不是文本,而在生成"are"这种无视觉的信息时,1-β 较小。另外,图 7(a)显示,该图像生成的属性单词为<people,working,snow>,VT-BLSTM 模型沿用了属性词 people 和 snow,但是生成了更加符合图像描述的动词"shoveling",还正确预测数量"two",表明了模型生成文本的准确性。图 7(b)输出的句子是"A dog runs through the grass.",生成"dog"的时候能聚焦狗在的位置,并且生成"grass"的时候能够定位草地区域。与视觉相关的单词"dog"、"runs"和"grass"对应时刻的 1-β 数值较大,与视觉无关的单词"the"的 1-β 较小。图 7(c)能够正确输出文本"A little girl is running on a grassy area.",定位女孩和草地的位置,并在"girl"前加上修饰词"little",表明模型能够根据图像以及未来的"girl"和"grass"生成潜在的视觉单词。与图 7(b)相比,同样是草地,一个生成"grass",一个生成"grassy area",表明了模型生成单词的多样性。在生成第一个"a"时,1-β 值比较大,而在生成第二个"a"时,1-β 数值较小,是因为在 $t=0$ 时刻,没有过多的语义信息指导,所以模型选择更多的关注图像。图 7(d)表明,在出现了不完整的脸时,该模型也能输出正确的描述,并指出主体"girl"。一方面该模型重点关注了图像特征,一方面视觉三元组输出的"girl"对文本的生成也有一定的指导作用。

由此可见,VT-BLSTM 模型能够很好地识别出图中的主要目标和属性信息,选择性地关注图像或之前生成的文本,并且结合未来信息生成更丰富的、具有多样性的图像描述,实现图像的语义理解。



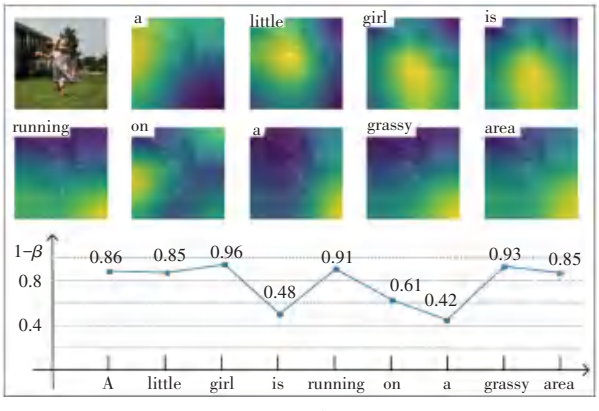
（a）生成文本描述 1

（a）Generate text description 1



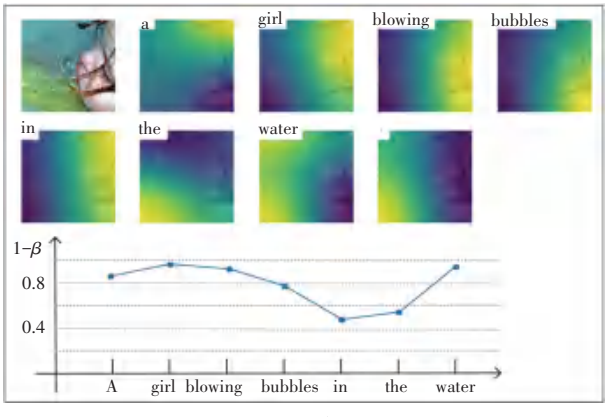
（b）生成文本描述 2

（b）Generate text description 2



（c）生成文本描述 3

（c）Generate text description 3



（d）生成文本描述 4

（d）Generate text description 4

图 7　可视化注意力机制

Fig. 7　Visual attention mechanism

2.4　对比实验设置

为了验证模型 VT-BLSTM 的有效性，本文选择了近 3 年图像语义理解领域中具有代表性且性能优越的模型进行对比实验。具体对比模型如下：

（1）DeepVS[6]。Karpathy 等人将图像中的多个局部区域和文本描述片段相对应，再使用 RNN 生成图像的文本描述。

（2）GoogleNIC[7]。Vinyals 等人提出基于 CNN 和 LSTM 的图像语义描述模型，为了减少模型参数，图像的特征只在 LSTM 的第一个时刻输入，但性能并没有降低。

（3）Hard-Attention[8]。Kelvin 等人提出了 Soft Attention 和 Hard Attention 机制。前者是对每个局部区域计算注意力权值，不会设置筛选条件。后者会在得到权重后筛选出不符合条件的权重，并将其置 0，使得每个时刻只注意到更小的区域。文献[12]中指出 Hard-Attention 比 Soft-Attention 生成的句子准确率更高，因此本文将 Hard-Attention 作为对比模型。

（4）VA-LSTM[9]。Qu 等人建立了一个大小为 6 的颜色词典，并且预测物体边框，在注意力机制中加入图像的颜色信息和物体的轮廓信息，从而生成更加准确的文本描述。

（5）Saliency+Context Att[11]。Marcella 等人提出了一种结合显著图和上下文的注意力机制。首先使用 SAM[20]获得图像的显著图，再将显著图与图像全局特征结合，分别输入显著图注意力模型和上下文注意力模型，最后根据 LSTM 得到图像对应的文本。

（6）CNN－RNN－f3[12]。吕凡等人在传统的 CNN-RNN 结构上加入了生成文本的反馈信息，并且配合循环迭代结构，通过不停循环使得关注信息得到强化。

（7）LSTM-EM-DA[13]。Jiang 等人在 LSTM 的输入门和输出门中分别增加存储单元和门控信号，并且使用上一时刻的 attention 信息来获得当前的关注区域。文献[13]指出，这种使用额外记忆信息的

注意力模型比 soft - attention 更能从不同的角度生成图像的文本描述。

(8) ATT-FCN[14]。You 等人提出了一种将图像视觉属性参与 LSTM 更新以及单词解码的方法,并且对比分析了 3 种产生视觉属性的方式和 3 种属性与模型的结合方式,文献[14]指出,ATT-FCN 模型效果最好,因此本文将 ATT-FCN 作为对比模型。

(9) SCN-LSTM[15]。Gan 等人为每个视觉属性都扩展了一个 LSTM 参数,使得每个属性都对应一个概率,根据加权视觉属性计算每个时刻生成的单词。

(10) $LSTM_p+ATT_{ssd}+CNN_m$[16]。Zhao 等人提出了一种多模型融合的图像语义理解模型,该模型分为图像特征提取模块、属性预测模块、语句特征提取模块和文本生成模块。文本生成模块的输入包含前一时刻生成的单词、预测的属性以及时变的语句特征。

(11) VD-SAN[17]。He 等人提出了一种包含共享参数的图像语义理解模型,DenseNet 中的参数不仅用于图像特征提取、视觉属性预测,而且用于注意力模型和 LSTM。

2.5 对比实验结果与分析

在 Flickr8K 和 Flickr30K 数据集中,VT-BLSTM 与该领域其它模型的性能对比结果见表 2。在 Flickr8K 数据集中,除了 BLEU-1 分数比 LSTM-EM-DA 低,其余的评价得分与对比模型相比均有较大的提升,尤其是在 METEOR 上,比其它基于注意力的模型性能平均提升了 7.7%。在 Flickr30K 数据集下,BLEU-3 和 BLEU-4 指数相比于其它模型有较大提升。

表 2 实验结果

Tab. 2 Experimental results

评价指标	Flickr8K					Flickr30k				
	B@1	*B*@2	*B*@3	*B*@4	*M*	*B*@1	*B*@2	*B*@3	*B*@4	*M*
Deep VS	57.9	38.3	24.5	16.0	-	57.3	36.9	24.0	15.7	15.3
Google NIC	63.0	41.0	27.0	-	-	66.3	42.3	27.7	18.3	-
Hard-Attention	67.0	45.7	31.4	21.3	20.3	66.9	43.9	29.6	19.9	18.4
VA-LSTM	67.6	47.0	32.6	22.1	-	67.5	45.1	30.2	21.1	-
Saliency+Context Attention	63.6	45.6	31.5	21.2	21.1	61.5	43.8	30.5	21.3	20.0
CNN-RNN-f3	68.3	46.5	32.1	22.1	22.0	67.5	44.5	30.0	20.3	**20.1**
LSTM-EM-DA	**69.0**	47.7	33.3	22.8	20.6	**68.4**	46.4	31.8	21.6	18.8
ATT-FCN	-	-	-	-	-	64.7	46.0	32.4	23.0	18.9
SCN-LSTM	-	-	-	-	-	67.2	48.2	34.0	23.8	-
$LSTM_p+ATT_{ssd}+CNN_m$	64.5	46.2	32.7	22.7	20.6	66.1	47.2	33.4	23.2	19.4
VT-BLSTM	68.4	**48.1**	**33.7**	**23.2**	**22.6**	68.1	**48.6**	**34.8**	**24.2**	19.8

表 2 中的 Deep VS 和 Google NIC 都只在文本生成模块的开始输入图像信息,后面的任何一个时刻,均不输入与图像信息有关的数据。因此随着时间的推移,图像中的有效信息会逐渐减小,即使生成的句子通顺,但有时候与图像关系不大,从而导致模型效果不好,准确率不高。而基于时序信息的 VT-BLSTM 能够定位到具体区域,结合图像深层特征对图像进行理解,与 Deep VS 和 Google NIC 相比,在 Flickr8K 数据集下,BLEU-1、BLEU-2 和 BLEU-3 指数平均提升 13.4、21.5、31.2,在 Flickr30K 数据集下,BLEU-1、BLEU-2 和 BLEU-3 指数平均提升了 10.8、23.3、35.3,证明注意力机制能够大幅度提升语义理解模型的准确性。

在具有注意力机制的模型中,本文提出的 VT-BLSTM 模型除了 BLEU-1 得分略低于 LSTM-EM-DA,在 METEOR 指数上略低于 CNN-RNN-f3 之外,在其它得分上均超过这五个模型。Hard-Attention 模型通过前一时刻 LSTM 隐含层的全连接层来预测当前关注的区域,即是通过语义特征预测图像特征,并且通过设置阈值得到更小的注意力区域,但文献[8]的可视化结果显示,注意的区域常与生成的单词没有直接关系。VA-LSTM 在原本的 LSTM 单元中,融入图像的颜色特征和轮廓,更新了原有的输入门、遗忘门和输出门的信息,但并没有考虑当前生成单词与前文的关系。Saliency+Context Attention 模型通过两路注意力机制预测文本,显著图注意力机制能够聚焦图像局部区别,上下文注意力机制使得生成的文本更加自然,但该模型不能用未来信息预测

潜在的视觉单词。CNN-RNN-f3 模型通过引入循环结构更新网络参数，能够在一定程度上解决注意力分散的问题，但是对于复杂的图像，该模型不能进行准确矫正。LSTM-EM-DA 模型通过门控信号使用额外的记忆信息，能够更加全面地关注图像，但存在聚焦区域分散的问题。而本文提出的 VT-BLSTM 不仅使用图像局部特征和之前生成的文本作为视觉语义信息，而且使用双向 LSTM，关注历史信息与未来信息，改善了注意力机制中没有考虑时序这一缺陷，使得生成的注意力区域更加准确，并且能够预测图像中的潜在视觉单词。

在具有属性特征的 ATT-FCN、SCN-LSTM、$LSTM_p+ATT_{ssd}+CNN_m$ 和 VT-BLSTM 中，除了本文提出的 VT-BLSTM 用到了注意力机制，其它的均没有使用注意力机制。并且除了评价指数 BLEU3 和 BLEU4 指数比 SCN-LSTM 低，其它指数均比这三个模型高，说明了模型在属性特征的基础上增加注意力机制能在一定程度上提高模型的性能。

由以上实验数据以及分析可知，本文提出的 VT-BLSTM 模型在考虑关注已生成文本和聚焦区域的时序基础上，能够联系历史和未来的信息使得模型定位更加准确，并且生成更加丰富多样的句子，有效地提升图像语义理解的准确率。

3　结束语

本文分析了现有图像语义理解模型存在的问题，提出了一种新的注意力模型 VT-BLSTM，首先通过 VGG19 提取图像的特征和视觉属性三元组，再使用双向 LSTM 获得当前时刻的视觉语义上下文，该视觉语义上下文和视觉属性三元组引导模型实现图像的语义理解。VT-BLSTM 在传统注意力机制的基础上，充分考虑了对生成文本的关注和聚焦的时序信息，使得每个时刻聚焦区域更加准确，从而生成更加符合人类描述的文本。

参考文献

[1] 郭聪. 基于关注度机制的图像理解 [D]. 合肥：中国科学技术大学，2018.

[2] CHANG Yanshuo. Fine-grained attention for image caption generation[J]. Multimedia Tools and Applications, 2018, 77(3): 2959-2971.

[3] ZHU Xinxin, LI Lixiang , LIU Jing , et al. Image captioning with word gate and adaptive self-critical learning [J]. Applied Sciences, 2018, 8(6): 909-922.

[4] MICAH H, PETER Y, JULIA H. Framing image description as a ranking task: Data, models and evaluation metrics[J]. Journal of Artificial Intelligence Research, 2014, 47(1): 853-899.

[5] MAO J, XU W, YANG Y, et al. Deep captioning with multimodal Recurrent Neural Networks (m-RNN) [C]// International conference on learning representations. San Diego, CA, USA: dblp, 2015: 11-301.

[6] KARPATHY A, LI Feifei. Deep visual-semantic alignments for generating image descriptions [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(4): 664-676.

[7] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: A neural image caption generator [J]. arXiv preprint arXiv: 1411.4555, 2014.

[8] KELVIN X, BA J, KIROS R, et al. Show, attend and tell: Neural image caption generation with visual attention [J]. arXiv preprint arXiv: 1502.03044, 2015.

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

[10] CHEN Long, ZHANG Hanwang, XIAO Jun, et al. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning [C]// Proc of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 6298-6306.

[11] MARCELLA C, BARALDI L, SERRA G, et al. Paying more attention to saliency: Image captioning with saliency and context attention [J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2018, 14(2): 48: 1-48: 21.

[12] 吕凡，胡伏原，张艳宁，等. 面向图像自动语句标注的注意力反馈模型[J]. 计算机辅助设计与图形学学报，2019，31(7)：1122-1129.

[13] JIANG Teng, ZHAN Chengjun, YANG Yupu. Long short-term memory network with external memories for image caption generation [J]. Journal of Electronic Imaging, 2019, 28(2): 023022.

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

[15] GAN Zhe, GAN Chuang, HE Xiaodong, et al. Semantic compositional networks for visual captioning [C]//IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017: 1141-1150.

[16] ZHAO Dexin, CHANG Zhi, GUO Shutao. A multimodal fusion approach for image captioning[J]. Neurocomputing, 2019, 329: 476-485.

[17] HE Xinwei, YANG Yang, SHI Baoguang, et al. VD-SAN: Visual densely semantic attention network for image caption generation[J]. Neurocomputing, 2019, 328: 48-55.

[18] FANG Hao, GUPTA S, IANDOLA F N, et al. From captions to visual concepts and back [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA: IEEE, 2015: 1473-1482.

[19] YAO Ting, PAN Yingwei, LI Yehao, et al. Boosting image captioning with attributes [C]//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 4894-4902.

[20] MARCELLA C , LORENZO B , GIUSEPPE S , et al. Predicting human eye fixations via an LSTM-based saliency attentive model [J]. IEEE Transactions on Image Processing, 2017, 27(10): 5142-5154.