浙江理工大学学报，2019，41(2)：208-216
Journal of Zhejiang Sci-Tech University
DOI:10.3969/j.issn.1673-3851(n).2019.02.011

基于注意力反馈机制的深度图像标注模型

邓远远，沈　炜
（浙江理工大学信息学院，杭州 310018）

摘　要：针对图像标注任务提出了一种基于注意力反馈机制的深度图像标注模型。该模型采用编码器-解码器框架；编码器采用VGG-16的网络结构，以提取图像的特征信息；在解码器部分设计了一种堆叠方式自上而下的处理注意力信息，使网络的每一层都可以获得额外的特征信息。然后从生成的标注语句中提取特征，将关注特征和图像的关注区域结合，增强和图像关注区域的匹配性，使生成的标注语句近似真实语境。在Flickr8k、Flickr30k和MSCOCO等数据集进行实验，实验结果显示，所提出模型的识别率比经典图像识别模型高5%～9%。

关键词：卷积神经网络；深度学习；图像识别；注意力机制
中图分类号：TP181　　**文献标志码**：A　　**文章编号**：1673-3851 (2019) 03-0208-09

Depth image caption model based on attention feedback mechanism

DENG Yuanyuan，*SHEN Wei*
(School of Information Science and Technology，Zhejiang Sci-Tech University，Hangzhou 310018，China)

Abstract：A depth image caption model based on attention feedback mechanism is proposed for image caption tasks. The model uses the encoder-decoder framework. The encoder adopts the VGG-16 network structure to extract the feature information of images. A stacking method is designed in the decoder part to handle the attention information from top to bottom，so that additional feature information is available for each layer of network. Then，the feature is extracted from the generated annotation statement，and the attention feature is combined with the attention area of the image to enhance the matching with the image attention area，so that the generated annotation statement approximates the real context. Experiments were carried out on data sets such as Flickr8k，Flickr30k and MSCOCO. The experimental results show that the recognition rate of the proposed model is 5%～9% higher than that of the classical image recognition model.

Key words：convolutional neural network；deep learning；image recognition；attention mechanism

0　引　言

随着移动互联网技术应用的快速发展，图像已成为传递信息的重要媒介。有效地管理和识别所需要的图像是一个重要并且具有现实意义的研究课题，如通过图像标注任务为医生提出建议，减少医生的工作量[1-2]。

图像标注任务不仅需要识别出图像中有哪些目标，还需要理解图像中目标之间的关系，生成一段近似真实语境的标注语句。Bernardi等[3]

收稿日期：2018－09－08　　网络出版日期：2018－12－28
作者简介：邓远远(1992－)，男，河南安阳人，硕士研究生，主要从事图像识别方面的研究。
通信作者：沈　炜，E-mail：120259565@qq.com

提出了图像标注技术包含图像理解和语言生成两大挑战。Hodosh 等[4]建立了基于句子的图像描述和检索的排序框架，通过排序比较来查找关系。Gong 等[5]通过分析在训练集合中的信息，为映射方法提供有用信息。Cho 等[6]通过对原始的递归循环网络（Recurrent neural network，RNN）[7-8]结构进行分析，解决了编码器-解码器模型长度不一致问题。Vinyals 等[9]利用卷积神经网络（Convolutional neural network，CNN）[10-11]提取的特征信息输入 LSTM（Long short-term memory，LSTM）[12]循环神经网络，获得一段对图像的语句标注，标注语句接近真实语境。Xu 等[13]提出一种注意力机制，有效提高了模型的性能。在基于注意力机制的图像标注模型中，采用注意力机制的模型可以轻松处理图像中存在的对象信息。符合真实语境的标注语句存在一些修饰词语，用于客观地描述图像的场景内容。然而图像中不存在有修饰词语的关注区域，如果强行将某个关注区域和修饰词语相关联，将导致关注区域分散。在预测修饰词语时，模型不能确定当前关注区域是对象，还是修饰词语，无法做到准确预测，使标注语句发生错乱。

针对传统图像标注模型的不足和缺陷，本文提出一种基于注意力反馈机制的深度图像标注模型。该模型首先利用 CNN 的最后一个卷积层作为图像的特征信息，将特征向量输入到 RNN 中，然后设计一种网络结构来处理注意力信息，最后引入基于注意力机制的文本反馈结构，以有效保证输入和输出注意力描述信息的匹配性，使生成的标注语句更加符合真实语境。

1　注意力机制

注意力机制源于对人类视觉的研究。人类首先观察全局图像，然后对感兴趣的图像区域进行关注，最后详细地了解关注区域，以获取图像信息。研究人员在进行图像标注训练时，将注意力机制应用到其中，利用图像中的每个关注区域对应的单词来预测生成标注语句。传统基于注意力机制的图像标注模型如图 1 所示。该模型利用 CNN 提取图像的特征信息，然后将特征信息输入到 LSTM 中，注意力机制在 LSTM 解码的不同时刻关注图像的不同区域，进而生成更合理的词，LSTM 依次预测生成语句中的每一个单词，最后生成一段对图像的描述。

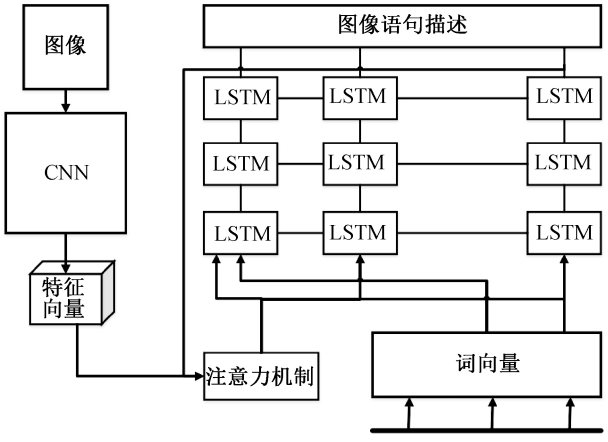


图 1　基于注意力机制的图像标注模型

注意力机制中有两个比较关键的信息，即上一时刻 LSTM 生成的隐藏状态和输入的图像区域。获取图像的关注区域在某一时刻的打分函数 e_{ti}，该函数可以用公式表示为：$e_{ti}=f_{\text{att}}(\boldsymbol{V}_i,\boldsymbol{h}_{t-1})$，其中：$\boldsymbol{h}_{t-1}$ 是上一时刻生成的隐藏状态，$\boldsymbol{V}_i$ 是输入的图像区域，f_{att} 是打分函数。

各个图像区域的权值计算和获取关注区域的上下文向量 Z_t，可以公式表示为：$Z_t=\sum_{i=1}^{L}a_{ti}\boldsymbol{V}_i$，其中：$L$ 是参考语句单词的数量；a_{ti} 为各个图像区域的注意力信息，$a_{ti}=\exp(e_{ti})\Big/\sum_{k=1}^{L}\exp(e_{tk})$。

2　基于注意力反馈机制的深度图像标注模型

2.1　模型框架

本文提出改进的图像标注模型采用编码器-解码器结构。编码器采用 VGG-16 的网络结构，以提取图像的特征信息。解码器部分采用一种网络结构 Q-LSTM，它由 LSTM 和注意力机制组成，从图像中获取的特征信息通过选择特征向量的子集，选择性地聚焦于图像的某些部分，捕捉特定区域视觉信息的上下文向量，解码生成一段语句信息。然后从生成的标注语句中提取特征，将关注特征和图像的关注区域结合，增强和图像关注区域的匹配性，使生成的标注语句更加接近真实语境。模型框架如图 2 所示，其中：$\boldsymbol{f}_{\text{conv}}$ 表示卷积神经网络从图像中提取的特征向量，N 为生成标注语句长度，$\boldsymbol{g}$ 表示 LSTM 生成的注意力，Q-LSTM 由 LSTM 和注意力机制组成，S_i 表示参考的语句；$\log p(s_i)$ 表示预测单词最大概率。

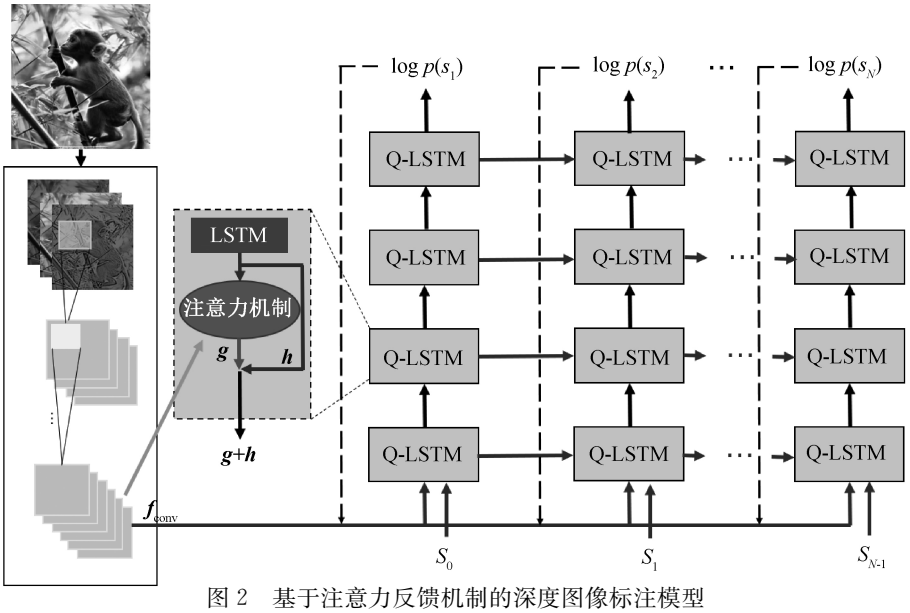


图 2 基于注意力反馈机制的深度图像标注模型

2.2 编码器结构

CNN 编码器部分,先对图像进行预处理,对每个尺度下的图像进行随机裁剪、反转、颜色调整等数据扩充方法,使数据进行归一化,每张图像都转变为 224×224 的图像。在 CNN 提取图像特征采用在大规模单标签图像分类数据集 ImageNet 上预训练的 VGG-16 的网络结构,VGG-16 模型的每层参数设置如表 1 所示,其中包括 1 个 Softmax 输出层、2 个全连接层和 13 个卷积层。提取卷积层 Conv-5_3 的卷积特征,维度为 14×14×512,其中特征图 Feature map 的大小为 14×14=196,通道数为 512。

表 1 VGG-16 模型的层参数设置

层名称	卷积核大小	步长	填充	通道数
Conv-1_1	3×3	1	1	64
Conv-1_2	3×3	1	1	64
Maxpooling-1	2×2	2	—	—
Conv-2_1	3×3	1	1	128
Conv-2_2	3×3	1	1	128
Maxpooling-2	2×2	2	—	—
Conv-3_1	3×3	1	1	256
Conv-3_2	3×3	1	1	256
Conv-3_3	3×3	1	1	256
Maxpooling-3	2×2	2	—	—
Conv-4_1	3×3	1	1	512
Conv-4_2	3×3	1	1	512
Conv-4_3	3×3	1	1	512
Maxpooling-4	2×2	2	—	—
Conv-5_1	3×3	1	1	512
Conv-5_2	3×3	1	1	512
Conv-5_3	3×3	1	1	512
Maxpooling-5	2×2	2	—	—
FC-6	—	—	—	4096
FC-7	—	—	—	4096
Softmax	—	—	—	—

注:Conv 表示卷积层;Maxpooling 表示池化层;FC 表示全连接层;Softmax 表示分类层。

2.3 解码器结构

传统的基于注意力机制的图像标注模型,利用 CNN 从图像中提取特征向量,然后将特征信息输入到 RNN 中,RNN 部分通过选择图像特征向量的子集,选择性地聚焦于图像的某些部分,捕捉特定区域视觉信息。最后,使用 LSTM 计算出来的图像局部语义信息进行解码。给定一张图像 $\boldsymbol{I}$,其特征映射数据集用 $\boldsymbol{A}$ 表示,目标图像的参考语句如式:

$$\boldsymbol{S}=\{s_1,s_2,\cdots,s_L\}。$$

获取图像标注的最大概率为:

$$\log p(\boldsymbol{S}\mid\boldsymbol{I})=\sum_{t=1}^{L}\log p(s_t\mid\boldsymbol{I},s_{1:t-1}),$$

其中:$s_{1:t-1}$ 代表参考语句中第 1 个到第 $t-1$ 个元素,$p(s_t\mid\boldsymbol{I},s_{1:t-1})$ 是图像 $\boldsymbol{I}$ 和之前预测的文本 $s_{1:t-1}$ 来预测 s_t 的最大可能性。其概率为:

$$p(s_t\mid\boldsymbol{I},s_{1:t-1})=\boldsymbol{u}(\boldsymbol{h}_t),$$

其中 $\boldsymbol{u}(\cdot)$ 是输出 s_t 概率的非线性函数,循环神经网络生成的隐藏状态可用公式表示为:

$$\boldsymbol{h}_t=LSTM(s_t,\boldsymbol{g}_t,\boldsymbol{h}_{t-1},\boldsymbol{c}_{t-1}),$$

其中 $\boldsymbol{c}_{t-1}$ 是 $t-1$ 时刻的记忆单元,由上层隐藏状态和特征映射集计算当前注意力可以公式表示为:

$$\boldsymbol{g}_t=\varphi(\boldsymbol{A},\boldsymbol{h}_t),$$

其中 $\varphi(\cdot)$ 代表注意力函数,当给定隐藏状态和特征映射集时,该函数返回一个特征信息。这两个输入首先用于计算注意力,然后通过将特征图与相应的权重进行内积生成注意力。

传统的基于注意力机制的图像标注模型中,隐藏状态用于提取更多有用信息并产生注意力,当前隐藏状态与下一个要预测的词有关。如果预测的单

词是一个具体的对象时，可以容易地学习其信息。如果预测的单词是动词或形容词这样的修饰词语，图像中不存在有修饰词语的关注区域，如果强行预测某个关注区域和修饰词语相关，将导致关注区域分散。因此本文设计一种堆叠的网络结构 Q-LSTM，自上而下地处理注意力信息。深度 Q-LSTM 网络流程图如图 3 所示。Q-LSTM 由 LSTM 和注意力机制组成，网络结构如图 4 所示。该模型的注意力可用公式表示为：

$$\boldsymbol{g}_t^k = \varphi(\boldsymbol{A}, \boldsymbol{h}_t^{k-1}),$$

其中：k 表示 Q-LSTM 的第 k^{th} 层网络，$\boldsymbol{h}_t^{k-1}$ 表示当前 t 时刻的 $(k-1)^{th}$ 层的隐藏状态，$\boldsymbol{g}_t^k$ 是图 4 中 Q-LSTM 单元的一部分，表示在当前 t 时刻从 $(k-1)^{th}$ 层到 k^{th} 层的注意力。

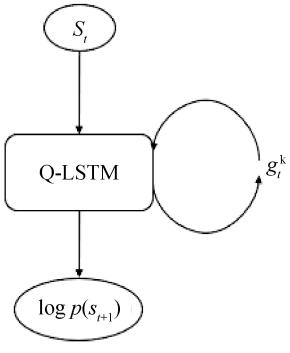


图 3　深度 Q-LSTM 网络流程图

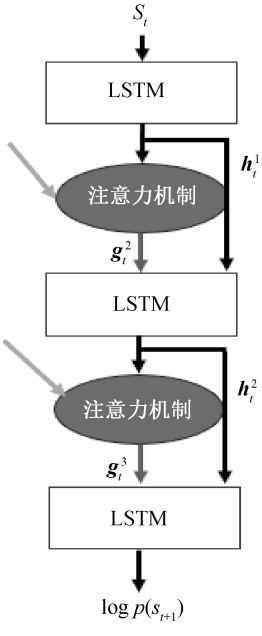


图 4　深度 Q-LSTM 网络结构图

本文使用堆叠方式依次输入注意力，使得 Q-LSTM 的每一层可以获得更多的额外特征信息。Q-LSTM 各层的隐藏状态可用公式表示为：

$$\boldsymbol{h}_t^1 = LSTM^1(s_t, \boldsymbol{h}_{t-1}^1, \boldsymbol{c}_{t-1}^1),$$

$$\boldsymbol{h}_t^k = LSTM^k(\boldsymbol{h}_{t-1}^k, \boldsymbol{g}_t^k, \boldsymbol{h}_{t-1}^k, \boldsymbol{c}_{t-1}^k),$$

其中 $k>1$，图像区域的注意力从第二层开始输入。假设深度 Q-LSTM 中存在 n 层，首先在第一层输入单词 s_t，然后将上一层获取的隐藏状态 $\boldsymbol{h}_t$ 和图像区域注意力 $\boldsymbol{g}_t$ 输入到下一个 Q-LSTM 层中。以 $n=4$ 为例，在 t 时刻，首先将单词 s_t 输入到第一层，然后使用隐藏状态 h_t^1 和特征映射集计算当前图像区域注意力 $\boldsymbol{g}_t^2$。之后将 $\boldsymbol{g}_t^2$ 和 $\boldsymbol{h}_t^1$ 输入到第二层可以得到输出 $\boldsymbol{h}_t^2$，然后 $\boldsymbol{h}_t^2$ 输入到第三层，通过相同的方法计算注意力 $\boldsymbol{g}_t^3$。同理第四层的计算与前两层的计算相同。

2.4　基于注意力机制的文本反馈结构

图像标注是一个跨模态数据转换的过程。在模态的转变过程中，关注区域的特征信息被用来预测单词，该过程一直持续到生成一个完整的句子，是一种单向传播的操作。该操作将导致关注区域不集中和生成标注语句错乱问题。基于以上问题，本文提出了基于注意力机制的文本反馈方法，与传统的 CNN-RNN 结构最大区别在于引入基于注意力机制的文本反馈结构，循环迭代地更新图像的关注区域。文本的生成与反馈流程示意图如图 5 所示。

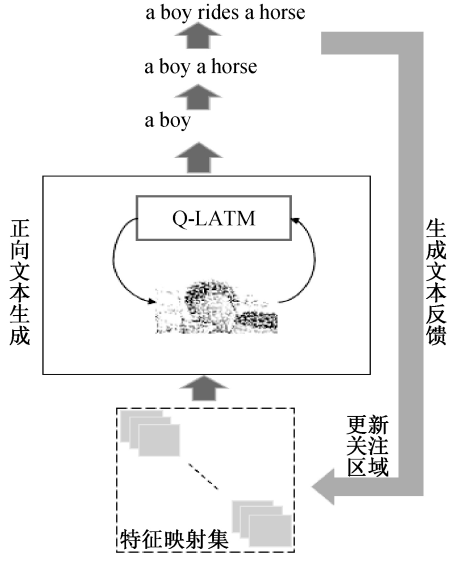


图 5　文本的生成与反馈流程示意图

在标注语句上的注意力 $\beta=\{\beta_1,\beta_2,\cdots,\beta_T\}$，注意力 β_t 可用公式表示为：

$$\beta_t = \frac{\exp(d_t)}{\sum_{k=1}^{T}\exp(d_k)},$$

其中：$d_t = g_d(\boldsymbol{h}_{t-1})$ 为每个单词的注意力，T 为生成文本的单词数量。$g_d(\cdot)$ 为一个多层感知机。从文本中提取的关注特征可以公式表示为：

$$\boldsymbol{r} = \sum_{t=1}^{T}\beta_t \boldsymbol{h}_t。$$

图像标注模型生成一段标注语句,提取标注语句的关注特征,关注特征和预测单词相关,若图像标注模型认为是图中的对象则加大权重,若图像标注模型认为是修饰词语则减小权重。使用从标注语句提取的关注特征修正图像注意力,图像和标注语句的联合关注特征可以公式表示为:

$$\boldsymbol{h}=\boldsymbol{W}_{\mathrm{h}\alpha}\boldsymbol{Z}_i+\boldsymbol{W}_{\mathrm{hr}}\boldsymbol{r},$$

$$\boldsymbol{\lambda}^k=\mathrm{softmax}(\tanh(\boldsymbol{h})),$$

其中:$\boldsymbol{W}_{\mathrm{h}\alpha}$ 为图像关注特征的权重,$\boldsymbol{W}_{hr}$ 为文本关注特征的权重。

3 实验和结果分析

为评估本文提出的模型,在三个常用的数据集 Flickr8k、Flickr30k 和 MSCOCO[14] 上进行实验和分析,使用两个评价标准 BLEU[15] 和 Meteor[16] 进行评测,并和经典的图像标注模型进行大量的对比实验,来评估所提出模型的性能。

3.1 数据集

实验使用三个常用的数据集 Flickr8k、Flickr30k 和 MSCOCO。Flickr8k 数据集有 8000 张图像,其中训练图像有 6000 张、验证图片和测试图片各 1000 张。Flickr30k 数据集有 31784 张图像,有 29000 张图像用于训练,另外 1000 张图像用于测试,其余用于验证。MSCOCO 具挑战性,数据集由大约一半的图像组成,分为训练、验证和测试集,以及人工注释。每个注释是一个大约 10～20 个单词的句子,有 5～7 个注释。

3.2 评价指标

由于数据集的数据量巨大,需要一种方法来量化模型在整个数据集上的平均精确度。使用 BLEU 和 Meteor 作为评价指标。BLEU 是机器翻译的评测标准,是一种十分常见的评测指标。Meteor 也是来评测机器翻译的,对模型给出的译文与参考译文进行词对齐,计算词汇完全匹配、词干匹配和同义词匹配等各种情况的准确率。

3.3 结果分析

本文在三个常用的数据集 Flickr8k、Flickr30 K 和 MSCOCO 上进行实验。从 VGG-16 的卷积层 Conv-5_3 中提取卷积特征。在训练过程中采取改进的动量优化算法[17],参数 *momenteum* 设为 0.9,并采用分段式学习率。卷积层和全连接层的学习率分别为 0.01 和 0.02,其余为 0.1。学习率每 10 个 epoch 衰减一次,衰减系数为 0.1,训练次数(epoch)为 100。由于设计了一种堆叠的方式处理注意力机制,需要评估 Q-LSTM 设置不同层数的性能,将模型层数分别设为 2 层、3 层、4 层、5 层和 6 层。分别在三个数据集上进行实验。首先模型设置不同层数在 Flickr8k 数据集上进行性能对比,在 Flickr8K 数据集上的实验结果如表 2 所示。

表 2 在 Flickr8K 数据集上的实验结果

模型	按不同评测标准的评估准确率/%				
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Q-LSTM2	71.6	48.7	33.8	23.9	20.5
Q-LSTM3	71.7	49.0	34.6	24.6	21.0
Q-LSTM4	72.4	49.3	34.9	24.8	21.5
Q-LSTM5	72.2	49.1	34.7	24.5	21.1
Q-LSTM6	71.9	49.3	34.3	24.9	21.0

在数据集 Flickr8k 上进行大量的实验对比后进行数据可视化分析,如图 6 所示。

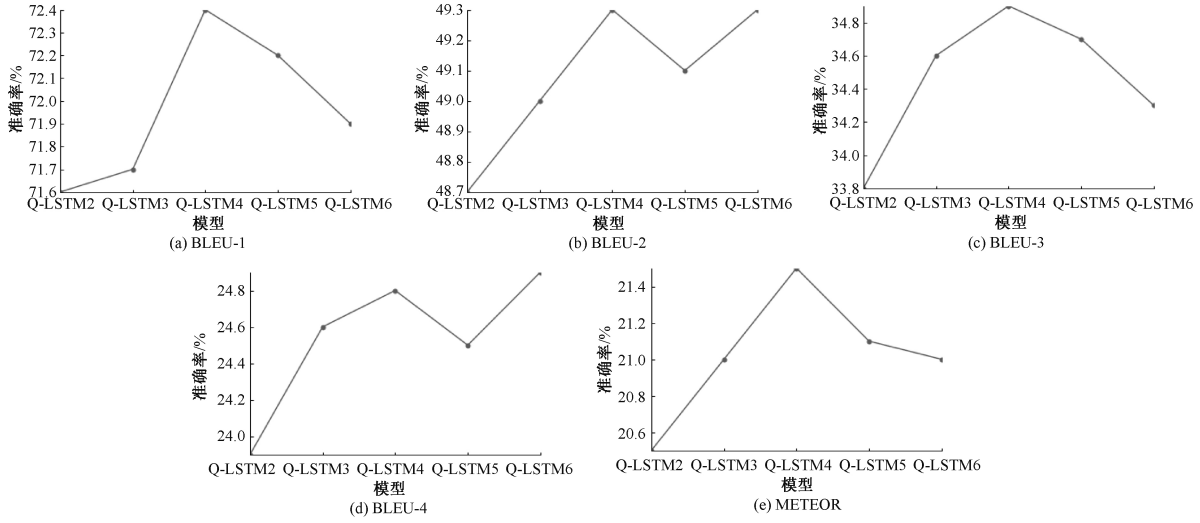


图 6 在 Flickr8k 上的对比实验结果

从图 6 可以观察到，当评测标准为 BLEU-1、BLEU-3 和 METEOR，模型层数设置为 4 时，图像标注的准确率达到最大，结果分别为 72.4%、34.9%和 21.5%。当评测指标为 BLEU-2 时，模型层数设为 4 层和 6 层，其结果都为 49.3%，图像标注的准确率最大。当评测标准为 BLEU-4，模型层数设置 6 时，图像标准的准确率达到最大。从数据集 Flickr8k 上的对比实验结果可知，模型层数设为 4 层或 6 层，图像标注的准确率达到最大。下面将在 Flickr30 K 数据集上的进行大量对比实验，来验证模型设为多少层，图像标注达到最好的效果，实验对比结果如表 3 所示。

表 3　在 Flickr30 K 数据集上的实验结果

模型	按不同评测标准的评估准确率/%				
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Q-LSTM2	69.2	46.7	31.8	22.6	18.5
Q-LSTM3	69.7	47.0	32.6	23.6	18.8
Q-LSTM4	70.3	47.1	33.5	23.8	19.0
Q-LSTM5	70.0	47.8	33.5	23.4	18.8
Q-LSTM6	70.2	47.5	33.3	23.4	18.9

在数据集 Flickr30 K 上进行大量对比实验后进行可视化分析，如图 7 所示。从图 7 可以观察到，当评测标准为 BLEU-1、BLEU-2、BLEU-4 和 METEOR，模型层数设置为 4 时，图像标注的准确率达到最大，结果分别为 70.3%、47.1%、23.8%和 19.0%。当评测指标为 BLEU-3 时，模型层数设为 4 层和 5 层，其结果都为 33.5%，图像标注的准确率最大。从数据集 Flickr30 K 上的对比实验结果可知，模型层数设为 4 层或 5 层，图像标注的准确率达到最大。在 MSCOCO 数据集上的进行对比实验，以选择图像标注达到最好的效果的模型层数，实验对比结果如表 4 所示。

表 4　在 MSCOCO 数据集上的实验结果

模型	按不同评测标准的评估准确率/%				
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Q-LSTM2	74.1	53.2	43.2	33.1	26.2
Q-LSTM3	74.8	54.1	43.7	33.7	26.4
Q-LSTM4	75.4	54.9	44.5	34.6	26.5
Q-LSTM5	74.2	54.3	44.1	34.0	26.4
Q-LSTM6	75.0	54.5	43.8	33.8	26.3

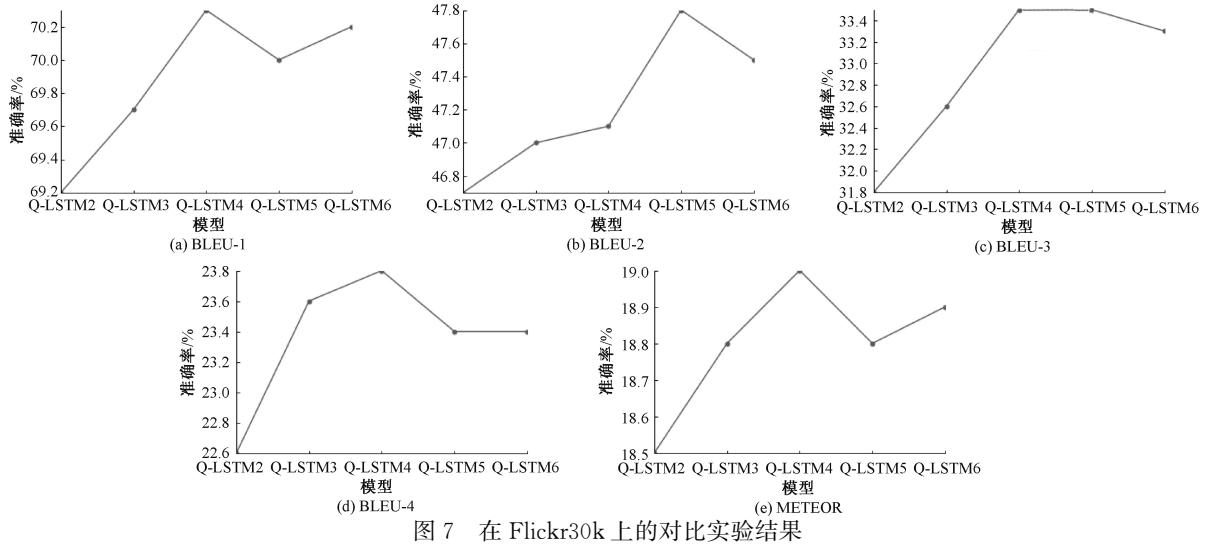


图 7　在 Flickr30k 上的对比实验结果

在数据集 MSCOCO 上进行大量的对比实验后进行数据可视化分析，如图 8 所示。从图 8 可以观察到，当评测标准为 BLEU-1、BLEU-2、BLEU-3、BLEU-4 和 METEOR，模型层数设置为 4 时，图像标注的准确率达到最大，结果分别为 75.4%、54.9%、44.5% 和 34.6% 和 26.5%。从数据集 MSCOCO 上的对比实验结果可知，模型层数设为 4 层，图像标注的准确率达到最大。

从以上分别对 Flickr8k、Flickr30k 和 MSCOCO 数据集进行实验分析可知，在 Flickr8k 数据集上，模型设为 4 层或 6 层效果达到最好，然而模型设为 6 层只在评测标准 BLEU-4 上效果最好，模型为 4 层在其他评测标准上效果最好。Flickr30 K 数据集实验结果和 Flickr8k 数据集上的实验结果相似。在 MSCOCO 数据集上，模型设为 4 层在所有的评测标准上都达到了最好的效果。本文将采用 Q-LSTM4 模型与经典的图像识别模型进行大量的对比实验，经典的图像识别模型如 BRNN、Google NIC、Log Bilinear、Soft-Attention、Hard-Attention，对比实验结果如表 5 所示。

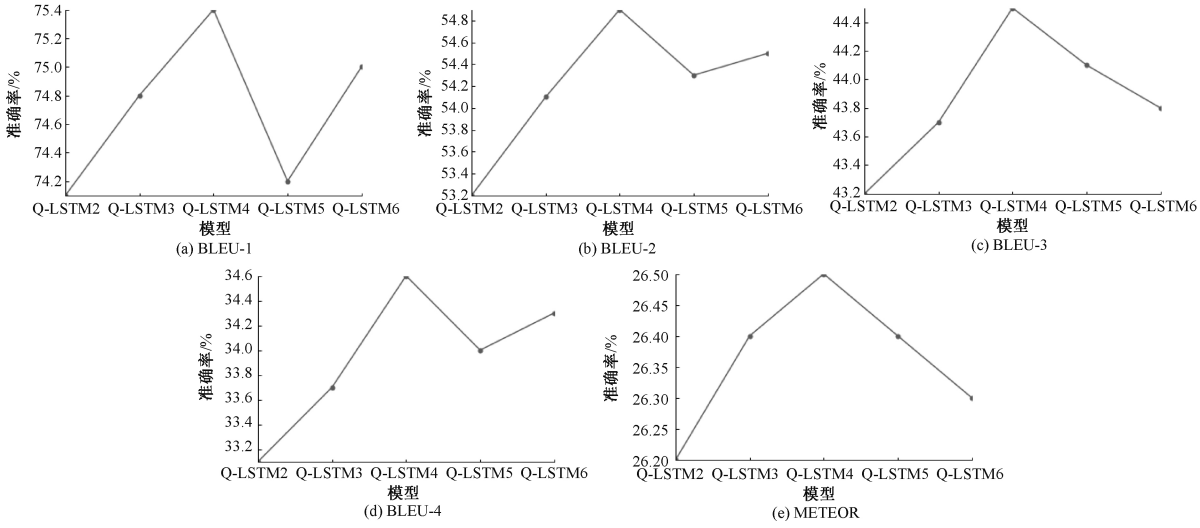


图 8 在 MSCOCO 上的对比实验结果

表 5 与其他模型对比结果

数据集	模型	按不同评测标准的评估准确率/%				
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Flickr8k	Google NIC	63.0	41.0	27.0	17.0	17.2
	Log Bilinear	65.6	42.4	27.7	17.7	17.3
	Soft-Attention	67.0	44.8	29.9	19.5	18.9
	Hard-Attention	67.0	45.7	31.4	21.3	20.3
	Q-LSTM4	72.4	49.3	34.9	24.8	21.5
Flickr30k	Google NIC	66.3	42.3	27.7	18.3	17.5
	Log Bilinear	60.0	38.0	25.4	17.1	16.9
	Soft-Attention	66.7	43.4	28.8	19.1	18.5
	Hard-Attention	66.9	43.9	29.6	19.9	18.5
	Q-LSTM4	70.3	47.1	33.5	23.8	19.0
MSCOCO	BRNN	64.2	45.1	30.4	20.3	20.1
	Google NIC	66.6	46.1	32.9	24.6	20.3
	Log Bilinear	70.8	48.9	34.4	24.3	20.0
	Soft-Attention	70.7	49.2	34.4	24.3	23.9
	Hard-Attention	71.8	50.4	35.7	25.0	23.0
	Q-LSTM4	75.4	54.9	44.5	34.6	26.5

在三个常用数据集上的实验结果进行数据可视化分析如图 9—图 11 所示。

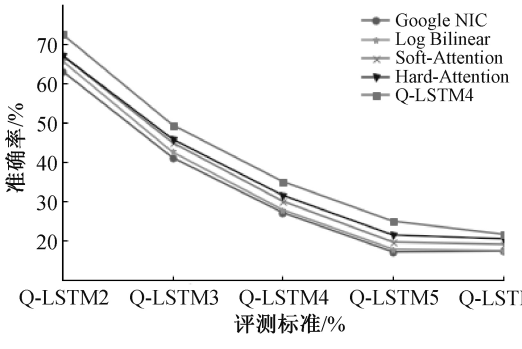


图 9 Flickr8k 对比实验结果

从图 9 可以看出,将五种模型分别在 Flickr8k 数据集上进行实验,所提出的模型优于其他几种模型,

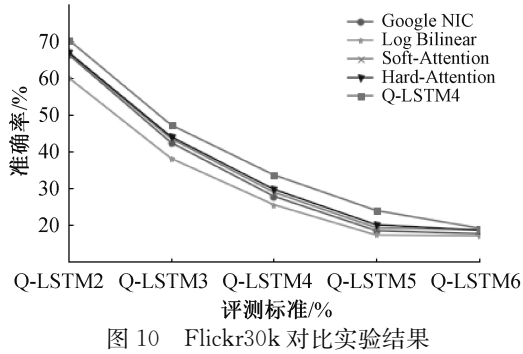


图 10 Flickr30k 对比实验结果

所提出的模型在五个评价指标 BLEU-1、BLEU-2、BLEU-3、BLEU-4 和 METEOR 上的结果分别为 72.4%、49.3%、34.9%、24.8%和 21.5%。由于这五个评价指标的数值越大代表模型的性能越好,图中显示了所提出模型的各项性能指标都优于对比模型。

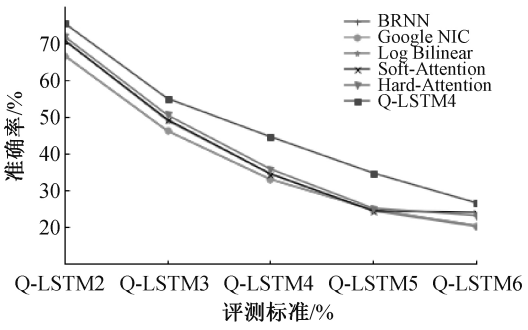
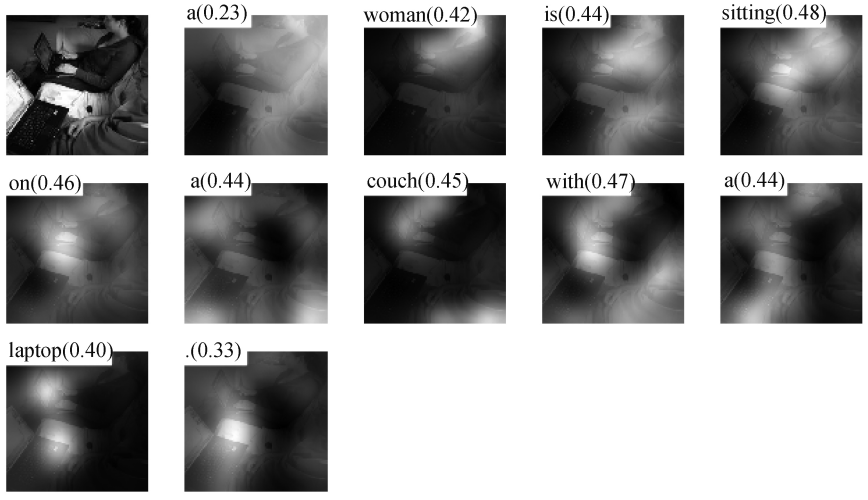


图 11　MSCOCO 对比实验结果

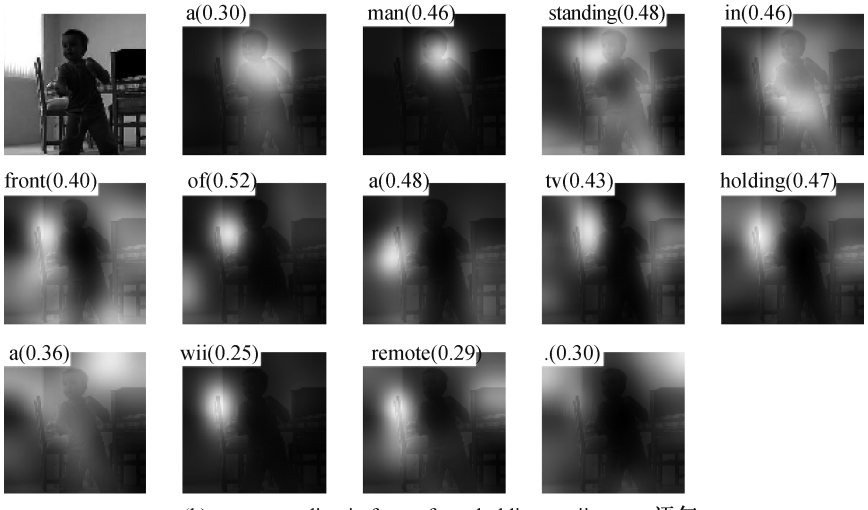
在图 10 中，通过使用 Flickr30 K 数据集对所提出的模型和经典的图像识别模型进行实验对比分析。通过折线图可以看出所提出的模型性能都高于其他模型。在实验结果的评价指标 BLUE-1 中，模型的准确率达到 70.3%，在实验结果的评价指标 BLUE-2 中，模型的准确率达到 47.1%，在实验结果的评价指标 BLUE-3 中，模型的准确率达到 33.5%，在实验结果的评价指标 BLUE-4 中，模型的准确率达到 23.8%，在实验结果的评价指标 METEOR 中，模型的准确率达到 19.0%。由以上分析可知，所提出的模型比经典的图像识别模型识别率高。

在图 11 中，通过使用 MSCOCO 数据集对所提出的模型和经典的图像识别模型进行实验对比分析。通过折线图可以看出所提出的模型性能都高于其他模型。

从三个数据集的对比实验结果可以看出，所提出模型的识别率比经典图像识别模型高出 5%～9%，能够有效提升图像标注的效果，从而提升语句标注生成的效果。如图 12 所示，列举了几个在 MSCOCO 数据集上的语句标注结果。从结果中可以得出，当关注的区域是一个女人的时候，模型输出 woman 关键词。本文提出的基于注意力反馈机制的深度图像标注模型能够形象地描述图像中的信息，生成一段近似真实语境的标注语句。



(a) a woman is sitting on a couch with a laptop语句



(b) a man standing in front of a tv holding a wii remote语句

图 12　MSCOCO 数据集上的语句标注结果示例

4 结 论

本文提出了基于注意力反馈机制的深度标注模型,实现了图像到语句标注的转变。在转变的过程中,采用堆叠的网络结构获取额外的特征信息,从生成文本中提取注意力信息,循环地修正图像中的关注区域,使得关注点集中,能够有效地预测语句中的非关键词。通过实验对比分析,本文提出的模型能够合理地描述图像场景中的内容,生成一段合理的语句,有效地提升图像标注的效果。对比经典的图像识别模型,在生成语句的效果上有所提高。在后续的工作中,将继续深入研究图像标注这一领域,提升图像标注的识别率。

参考文献:

[1] Plis S M, Hjelm D R, Salakhutdinov R, et al. Deep learning for neuroimaging: a validation study [J]. Frontiers in neuroscience, 2014, 8(8): 00229.

[2] Roth H R, Lu L, Seff A, et al. A new 2.5 D representation for lymph node detection using random sets of deep convolutional neural network observations [C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2014: 520-527.

[3] Bernardi R, Cakici R, Elliott D, et al. Automatic description generation from images: A survey of models, datasets, and evaluation measures[J]. Journal of Artificial Intelligence Research, 2016, 55: 409-442.

[4] Hodosh M, Young P, Hockenmaier J. Framing image description as a ranking task: Data, models and evaluation metrics[J]. Journal of Artificial Intelligence Research, 2013, 47: 853-899.

[5] Gong Y, Wang L, Hodosh M, et al. Improving image-sentence embeddings using large weakly annotated photo collections [C]//European Conference on Computer Vision. Springer, 2014: 529-545.

[6] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [EB/OL]. (2014-09-03) [2018-12-06]. https://arxiv.org/abs/1406.1078.

[7] Fang F, Wang H, Chen Y, et al. Looking deeper and transferring attention for image captioning [J]. Multimedia Tools and Applications, 2018(8): 1-17.

[8] Chang Y S. Fine-grained attention for image caption generation [J]. Multimedia Tools and Applications, 2018, 77(3): 2959-2971.

[9] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, 2015: 3156-3164.

[10] Mojoo J, Kurosawa K, Kurita T. Deep CNN with graph Laplacian regularization for multi-label image annotation [C]//International Conference Image Analysis and Recognition. Springer, 2017: 19-26.

[11] 庞超,尹传环. 基于分类的中文文本摘要方法[J]. 计算机科学,2017, 45(1):144-147.

[12] You Q, Jin H, Wang Z, et al. Image captioning with semantic attention [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, 2016: 4651-4659.

[13] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention [C]//International Conference on Machine Learning, Lille. JMLR: W&CP, 2015, 37: 2048-2057.

[14] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//European Conference on Computer Vision 2014. Springer, 2014: 740-755.

[15] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation [C]// Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Stroudsburg. Association for Computational Linguistics, 2002: 311-318.

[16] Lavie A, Agarwal A. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments[C]//Proceedings of the Second Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2007: 228-231.

[17] Sutskever I, Martens J, Dahl G, et al. On the importance of initialization and momentum in deep learning[C]// Proceedings of the 30th International Conference on International Conference on Machine Learning. JMLR.org, 2013, 28: 1139-1147.

(责任编辑:康 锋)