面向图像自动语句标注的注意力反馈模型

吕 凡[1,5], 胡伏原[1,2)*], 张艳宁[3)], 夏振平[1)], 盛胜利[4,6)]

[1)] (苏州科技大学电子与信息工程学院 苏州 215009)
[2)] (苏州科技大学苏州市虚拟现实智能交互及应用技术重点实验室 苏州 215009)
[3)] (西北工业大学计算机学院 西安 710029)
[4)] (Department of Computer Science, University of Central Arkansas, Conway AZ 72035)
[5)] (天津大学智能与计算学部 天津 300072)
[6)] (江苏省建筑智慧节能重点实验室 苏州 215009)
(fuyuanhu@mail.usts.edu.cn)

摘 要: 图像自动语句标注利用计算机自动生成描述图像内容的语句, 在服务机器人等领域有广泛应用. 许多学者已经提出了一些基于注意力机制的算法, 但是注意力分散问题以及由注意力分散引起的生成语句错乱问题还未得到较好解决. 在传统注意力机制的基础上引入注意力反馈机制, 利用关注信息的图像特征指导文本生成, 同时借助生成文本中的关注信息进一步修正图像中的关注区域, 该过程不断强化图像和文本中的关键信息匹配、优化生成的语句. 针对常用数据集 Flickr8k, Flickr30k 和 MSCOCO 的实验结果表明, 该模型在一定程度上解决了注意力分散和语句顺序错乱问题, 比其他基于注意力机制方法标注的关注区域更加准确, 生成语句更加通顺.

关键词: 图像自动语句标注; 注意力机制; 注意力反馈
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2019.17505

Feedback Attention Model for Image Captioning

Lyu Fan[1,5], Hu Fuyuan[1,2)*], Zhang Yanning[3)], Xia Zhenping[1)], and Victor S Sheng[4,6)]

[1)] (*School of Electronic & Information Engineering, Suzhou University of Science and Technology, Suzhou* 215009)
[2)] (*Virtual Reality Key Laboratory of Intelligent Interaction and Application Technology of Suzhou, Suzhou* 215009)
[3)] (*School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an* 710029)
[4)] (*Department of Computer Science, University of Central Arkansas, Conway AZ* 72035)
[5)] (*College of Intelligence and Computing, Tianjin University, Tianjin* 300072)
[6)] (*Jiangsu Province Key Laboratory of Intelligent Building Energy Efficiency, Suzhou* 215009)

Abstract: The image captioning problem aims to let machine generate relevant sentence of a given image, which has been applied to the service robot. To improve the performance of image captioning effectively, some researchers propose to leverage the attention mechanism. However, the mechanism often suffers from distraction and sentence-disorder. In this paper, we propose an image captioning model based on a novel feed-back attention mechanism. In generating the corresponding language for a given image, the proposed model uses the attention feedback from the generated language. With the feedback, the attention heatmap of the original image will be revised, and the generated sentence will also be better. We evaluate the proposed method on three benchmark data-

收稿日期: 2018-08-27; 修回日期: 2019-04-03. 基金项目: 国家自然科学基金(61876121, 61472267, 61728205, 61502329); 江苏省重点研发计划(BE2017663). 吕 凡(1993—), 男, 博士研究生, 主要研究方向为图像处理、深度学习、多标签分类、图像自动语句标注; 胡伏原(1978—), 男, 博士, 教授, 硕士生导师, CCF 会员, 论文通讯作者, 主要研究方向为图像处理、模式识别、深度学习; 张艳宁(1967—), 女, 博士, 教授, 博士生导师, CCF 理事, 主要研究方向为图像处理、模式识别、计算机视觉与智能信息处理; 夏振平(1985—), 男, 博士, 讲师, 主要研究方向为立体显示图像质量测量、评价和优化; 盛胜利(1969—), 男, 博士, 副教授, 博士生导师, 主要研究方向为数据挖掘与机器学习、人工智能、数据安全和决策支持及其在商业、生物信息学、医疗信息学、软件工程等方面的基础和应用研究.

sets, i.e., Flickr8k, Flickr30k and MSCOCO, and the experimental results show the superiority of the proposed method.

Key words: image captioning; attention mechanism; attention feedback

图像自动语句标注问题属于计算机视觉和自然语言处理的结合问题, 是新一代人机交互系统的研究的重点之一, 具有很强的实际意义, 如应用在聊天机器人、残疾人助理系统、机器人视觉系统等. 近年来, 图像自动语句标注问题引起了国内外学者的关注, 并获得了巨大的发展.

一般来说, 图像自动语句标注问题需要对图像进行特征提取, 并利用所提取的图像特征指导文本标注的生成. 随着深度神经网络研究的发展和应用推广, 现有的工作大多数是基于卷积神经网络(convolutional neural network, CNN)结合循环神经网络(recurrent neural network, RNN)的模型结构[1-2]. 其中, 利用 AlexNet[3], VGG[4], ResNet[5]等的深度 CNN 提取图像特征, 利用深度 RNN[6]对序列数据的处理特性生成对应的文字描述. 但是, 由于在图像自动语句标注问题中提取出的文本描述往往只和图像中最为突出的目标相关, 而传统的 CNN-RNN 模型考虑的是整个图像, 这使得其并不能非常好地理解图像中的内容并准确地生成对应的表述语句.

图像自动语句标注问题中普遍存在着注意力(显著性)的区分. 例如, 对于一般图像而言, 前景中的物体相对背景的重要性要高; 对于文本, 关键词相对修饰语的重要性要高. 目前, 研究人员开始关注注意力机制, 并在许多方面得到了应用[7-8]. 神经机器翻译中首先采用注意力机制[6], 并取得了一定突破. 还有一些学者将注意力机制应用到图像自动语句标注上[2,9], 利用每个单词在图像中对应的关注区域来指导生成标注语句.

1　相关工作

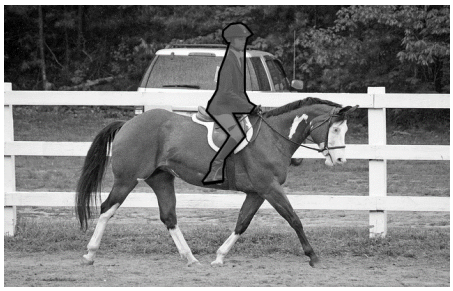
1.1　图像自动语句标注

图像自动语句标注问题近年来得到了广泛关注[10-11]. 作为连接计算机视觉和自然语言的关键问题之一, 基于该方向的研究对于新一代人工智能有望打破机器与人类交流的壁垒, 做到更加智能的人机交互.

已有一些关于图像自动语句标注问题的研究成果. Vinyals 等[1]首先利用 CNN 提取图像的特征, 接着利用该特征初始化 RNN, 利用 RNN 的特点序列化生成文本, 大大降低了传统图像自动语句标注模型的设计复杂度. Mao 等[12]利用多模的 RNN 读取从 CNN 中提取的图像信息, 分步预测语句中的每一个单词. Wu 等[13]利用 CNN 取出图像中的高级概念, 对于图像自动语句标注具有十分有效. Karpathy 等[14]提出一种多模的 RNN 结构, 可以同时生成图像中对应的子区域. Johnson 等[15]提出一种密集标注(DenseCap)的方法, 可以联合生成密集的图像标注并定位物体所在区域边界框.

1.2　注意力机制

注意力机制最初始于人类的注意力机制理论[16]. 该理论中认为, 人类的观察是有关注点的, 例如, 看图像的时候会集中注意力在关注的物体上. 如图 1 所示, 对于生成语句“a woman is riding a horse within a fence”, 在生成单词“woman”时, 图像中的关注区域应当是和“女人”有关的区域.



标注语句: a **woman** is riding a horse within a fence

图 1　图像注意力

近年来, 注意力机制在图像自动语句标注中逐渐受到关注. 基于注意力机制的图像自动语句标注分为 2 类: 基于全图的注意力机制, 基于显著目标的注意力机制. 基于全图的注意力机制将图像整体作为输入, 寻找图像中的关注区域. Xu 等[2]提出一种 soft 的注意力模型和一种 hard 的注意力模型, 利用 RNN 记录语句中的信息, 并计算图像中的关注点来指导接下来的单词预测. 在 Xu 等[2]研究的基础上, You 等[9]利用一系列的属性检测来获得视觉的属性特征标签, 然后将其融入 RNN 的

hidden state 中. 基于显著目标的注意力机制利用目标检测方法提取图像中的目标, 重点关注该目标以生成对应的文本. Liu 等[17]提出利用数据中关键区域的标注(边界框或分割图)对注意力进行监督学习. Liu 等[18]将图像检测目标看做序列, 结合注意力机制进行语句生成. Li 等[19]提出一种联合全局和局部注意力的图像自动语句标注方法. 基于全图的注意力机制直接利用全图信息指导文本的生成, 预测关注区域的准确性无法保证. 基于显著目标的注意力机制虽然利用局部信息来指导生成语句, 但是目标提取的算法提高了运算成本.

上述 2 类方法都会对语句进行按单词顺序解析, 在生成每个单词时估计图像上的关注区域(与预测单词最相关的图像子区域), 此信息被用来指导单词的预测. 该过程是一个单向传播的操作, 一直持续至生成一个完整的句子. 这类方法存在 2 个问题: (1) 注意力分散问题; (2) 生成的语句错乱问题. 该过程中, 注意力机制在图像中能反映仅有关键词的区域, 对于一些修饰语的区域无法定位, 使得图像中的注意力极易发散(无法找到单词对应的确切物体所在区域), 生成的语句产生错乱. 基于此, 本文引入注意力反馈机制, 将生成语句上的注意力反馈回图像, 迭代地修正图像中的关注区域、强化图像和文本中的关键信息匹配、优化生成语句. 基于注意力反馈机制的图像自动语句标注方法基于 CNN-RNN 结构, 但与传统的 CNN-RNN 结构的最大不同点在于: (1) 加入来自生成文本的反馈; (2) 引入循环结构迭代地更新图像的关注区域. Yang 等[20]提出一种 Review 网络来生成图像的文本标注, 该方法构建一系列包含注意力的 review 步骤, 用以“复习”来自编码器的信息. 文献[20]同样包含强化信息的思想, 但与本文模型具有以下不同: (1) 不包含反馈过程; (2) 不包含循环过程. 本文从生成文本中提取注意力来循环反馈并强化图像中的注意力, 前提是基于文本和图像的关注信息应当一致; 文献[20]则通过将多个 review 步骤获得的信息融合以强化编码器的全局表征能力, 该过程不存在反馈和循环, 是一个全前传结构.

2 本文模型

2.1 问题描述

本文模型分为训练和测试 2 部分. 训练时, 给定一系列的图像 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$, 其中, N 是样本数量; 图像 $\boldsymbol{x}_i$ 对应的语句表述为 $\boldsymbol{S}_i=\{\boldsymbol{s}_{i,1},\boldsymbol{s}_{i,2},\cdots,\boldsymbol{s}_{i,T}\}$, T 代表句子 $\boldsymbol{S}_i$ 的长度. 训练过程的主要目标是让构建的模型学习一种映射, 即 $h:\boldsymbol{X}\to\mathcal{S}$, 其中, $\mathcal{S}$ 代表所有的训练语句. 本文将每一个生成语句的过程看成是一个序列产生的过程, 即

$$\log p(\boldsymbol{S}_i\mid\boldsymbol{x}_i)=\sum_{t=1}^{T}\log\ p(\boldsymbol{s}_{i,t}\mid\boldsymbol{x}_i,\boldsymbol{s}_{i,1},\cdots,\boldsymbol{s}_{i,t-1}) \quad (1)$$

测试时, 利用训练好的模型 h, 生成来自测试图片的语句标注.

2.2 反馈式 CNN-RNN 结构

在基于深度方法的图像自动语句标注方法中, 由图像生成的文本和原图像是相互孤立的, 模型仅能够在训练过程中通过损失函数间接判断生成文本和真实文本的差距; 然而由于同义文本之间仍然存在较大差异性, 图像和生成文本中的关键信息不能很好地匹配. 利用反馈机制[21]将生成文本中的关键信息反向传给图像, 将有利于在提取图像特征的过程中更加关注文本中的信息所对应的显著目标, 从而有利于使得图像关键信息和文本关键信息更加匹配.

借鉴基于注意力机制的图像自动语句标注方法[2,9], 本文引入 3 点假设: (1) 图像中存在重点关注区域; (2) 生成文本中存在关键词; (3) 图像中的关注区域和生成文本中的关键词存在匹配关系. 根据上述假设, 针对图像自动语句标注问题, 本文提出一种注意力反馈机制: 模型结构基于经典的 CNN-RNN 设计, 符合“编码器-解码器”的设计模式. CNN-RNN 结构分为 CNN 部分和 RNN 部分, 其中, CNN 部分接收图像输入, 提取图像的深层特征; RNN 部分分析图像特征生成序列输出(如文本标注). 如图 2 所示, 每次迭代分成 3 个步骤:

Step1. 正向文本生成.

Step2. 生成文本反馈.

Step3. 图像关注区域更新.

其中, Step1 加持了注意力的图像特征, 被作为先验知识来初始化长短期记忆模型(long short-term memory, LSTM)[22]; Step2 生成文本中存在关键词上的注意力, 该信息和图像中的关键区域是匹配的, 用来指导图像矫正其关注区域; Step3 矫正后的图像特征继续用来初始化 LSTM, 指导文本的生成; Step1~Step3 称为一个循环, 该过程通过不停循环使得关注信息得到强化, 输出可以是任意循环中生成的文本.

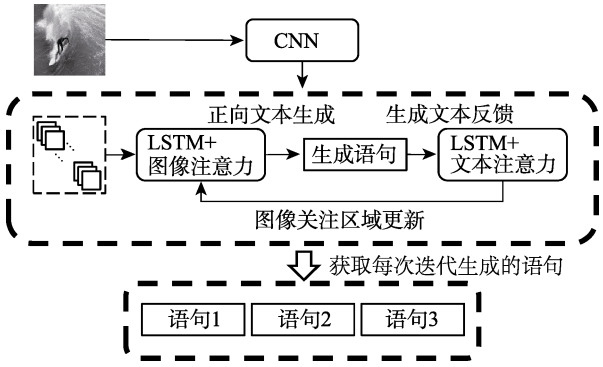


图 2　本文模型

2.2.1　正向文本生成

正向文本生成过程中，图像 $\boldsymbol{x}_i$ 被调整为固定大小，并分割成 L 个同等大小的切片．如图 3 所示，每个切片对应 CNN 从图像中提取出的特征图的一个区域特征．利用此信息，可以全局估计图像的关注区域．

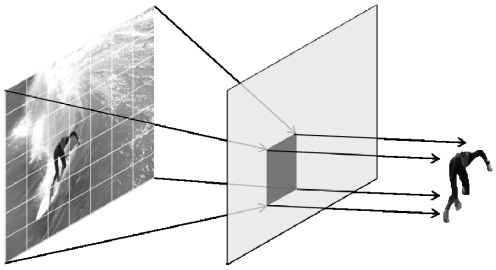


图 3　图像关注区域

利用 RNN 指导每个单词的生成，从而生成完整的句子，该过程如图 4 所示．对于图像，本文采用文献[23]中 LSTM 表达，在 LSTM 的 t 时刻，该过程可以表述为

$$\boldsymbol{s}_{i,t+1} = \mathrm{LSTM}(\boldsymbol{h}_t, \boldsymbol{c}_t, z_t) \tag{2}$$

其中，$\boldsymbol{c}_t$ 和 $\boldsymbol{h}_t$ 分别表示 LSTM 的记忆单元和隐藏状态；z_t 表示利用基于文本内容的注意力的图像关注特征，即

$$z_t = \sum_{j=1}^{L} \alpha_{t,j} \boldsymbol{a}_j,\ \ j \in \{1, \cdots, L\} \tag{3}$$

利用 CNN 从图像中提取出的特征为 $\boldsymbol{f}_{\mathrm{conv}} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_L\}$，conv 代表 CNN．$t$ 步骤时，图像中的注意力 $\boldsymbol{\alpha}_t = \{\alpha_{t,1}, \alpha_{t,2}, \cdots, \alpha_{t,L}\}$，第 i 个元素表示为

$$\alpha_{t,j} = \frac{\exp(e_{t,j})}{\sum_{k=1}^{L} \exp(e_{t,k})} \tag{4}$$

其中，

$$e_{t,j} = g(\boldsymbol{a}_j, \boldsymbol{h}_{t-1}) \tag{5}$$

$e_t = \{e_{t,1}, e_{t,2}, \cdots, e_{t,L}\}$ 表示 $\boldsymbol{\alpha}_t$ 中每个元素的能量大小，同时反映了来自图像特征 $\boldsymbol{f}_{\mathrm{conv}}$ 和上一个隐藏状态 $\boldsymbol{h}_{t-1}$ 的信息；$g(\cdot)$ 表示一个简单的多层感知机 (multi-layer perception, MLP).

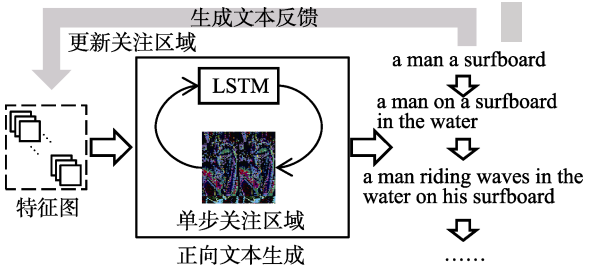


图 4　文本的生成与反馈

本文利用包含关注信息的图像特征来初始化模型，包括初始化图像中的关注区域和初始化 LSTM．因为图像中初始关注区域是未知的，所以初始化图像中的关注区域需对每个区域进行注意力均匀初始化

$$\boldsymbol{\lambda}^1 = \{\lambda_1^1, \lambda_2^1, \cdots, \lambda_L^1\} \tag{6}$$

其中，$\lambda_j^1 = 1/L$，$j \in \{1, \cdots, L\}$，1 代表第一个循环．利用初始化的注意力，对从图像中提取的特征 $\boldsymbol{f}_{\mathrm{conv}}$ 进行加权，可得

$$\boldsymbol{f}_{\mathrm{conv}} = \{\hat{\boldsymbol{a}}_1, \hat{\boldsymbol{a}}_2, \cdots, \hat{\boldsymbol{a}}_L\} = \{\lambda_1 \boldsymbol{a}_1, \lambda_2 \boldsymbol{a}_1, \cdots, \lambda_L \boldsymbol{a}_L\} \tag{7}$$

在每一次循环中利用加了关注的图像特征，可以初始化 LSTM 的记忆单元 $\boldsymbol{c}$ 和隐藏状态和 $\boldsymbol{h}$，分别为

$$\boldsymbol{c}_0 = I_c\left(\frac{1}{L}\sum_{i=1}^{L} \hat{\boldsymbol{a}}_i\right) \tag{8}$$

$$\boldsymbol{h}_0 = I_h\left(\frac{1}{L}\sum_{i=1}^{L} \hat{\boldsymbol{a}}_i\right) \tag{9}$$

对于图像 $\boldsymbol{x}_i$，经由正向文本生成过程，LSTM 的每一步都将产生一个单词，生成句子 $\boldsymbol{S} = \{\boldsymbol{s}_{i,1}, \boldsymbol{s}_{i,2}, \cdots, \boldsymbol{s}_{i,L}\}$.

2.2.2　生成文本反馈

在正向文本生成的过程中，图像中的某些区域会对应文本中的某几个单词，同时也存在无法很好配对的问题，即存在注意力分散和生成语句错乱问题．在此过程中，本文从生成的文本中提取关键词注意力反向矫正图像注意力．

在利用 LSTM 生成文本的过程中，每一步的隐藏状态的集合记为 $\boldsymbol{H} = \{\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_L\}$，那么利用类似于正向生成文本时图像中注意力的计算方式，

在文本上的注意力 $\boldsymbol{\beta}=\{\beta_1,\beta_2,\cdots,\beta_T\}$，其中

$$\beta_t=\frac{\exp(c_t)}{\sum_{k=1}^{T}\exp(c_k)} \tag{10}$$

$c_t=g_c(\boldsymbol{h}_{t-1})$，表示每个单词的能量，$g_c(\cdot)$ 表示一个浅层的 MLP. 由上述计算可以得到 LSTM 中每一步的隐藏状态的关注特征集合

$$\boldsymbol{r}=\sum_{t=1}^{T}\beta_t\boldsymbol{h}_t \tag{11}$$

本文模型从生成文本中提取出文本上的关注特征，该特征包含对每个单词的重要性分析，对模型认为是关键词的单词加大权重，反之减小权重.

2.2.3 图像关注区域更新与文本选择

利用从文本中提取出的关注特征对原本的图像关注特征进行注意力矫正，该过程如图 5 所示. 利用图像的关注特征 $\hat{\boldsymbol{f}}_{\text{conv}}$ 和来自文本中的关注特征 $\boldsymbol{r}$，可以计算出图像和文本的联合关注特征

$$\boldsymbol{H}=\boldsymbol{W}_{\text{ha}}\boldsymbol{a}_i+\boldsymbol{W}_{\text{hr}}\boldsymbol{r}+\boldsymbol{b}_{\text{h}} \tag{12}$$

重新计算图像中的注意力可得

$$\boldsymbol{\lambda}^k=\text{softmax}(\tanh(\boldsymbol{H})) \tag{13}$$

利用式(7)更新关注特征 $\hat{\boldsymbol{f}}_{\text{conv}}$，利用该信息在下一个循环中初始化 LSTM 以指导文本的生成. 本文分别观察 1~3 次迭代效果，分别标记为本文-f_1，本文-f_2 和本文-f_3.

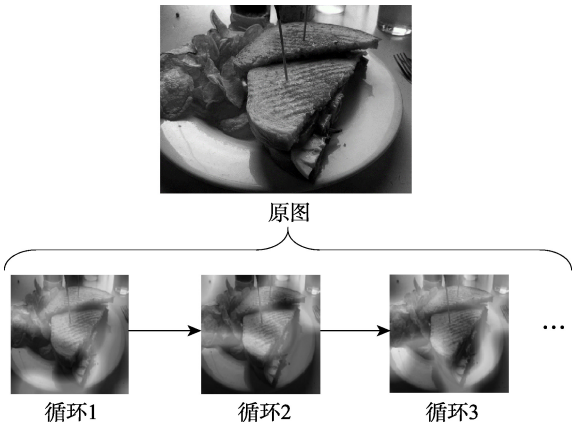


图 5 关注区域更新

3 实 验

3.1 数据集

为了衡量本文模型的表现能力，采用 3 个在图像自动语句标注中最常用的大规模数据集：Flickr8K[24], Flickr30K[25]和 MSCOCO[26].

Flickr8K 和 Flickr30K 收集于雅虎公司旗下的大规模图像分享网站 Flickr. Flickr8K 数据集共包含 8 000 幅图像，其中包含 6 000 幅训练图像、1 000 幅验证图像和 1 000 幅测试图像. Flickr30K 数据集共包含 31 784 幅图像，按照文献[14]可分为 28 000 幅训练图像，1 000 幅验证图像和 1 000 幅测试图像. 由于利用了众包系统，Flickr8K 和 Flickr30K 中的每幅图像都对应 5 个文本描述. MSCOCO 数据集包含 123 000 幅图像，且按照文献[20]中的数据集分割进行训练和测试.

3.2 实验结果与分析

实验中，本文模型的 CNN 部分利用 VGG-16 网络结构，RNN 部分利用 LSTM 结构. 其中，VGG-16 中的网络参数采用在大规模单标签图像数据集 ImageNet 上预训练的参数进行初始化. 图像特征从 VGG-16 的最后一个卷积层 conv5_3 中提取; LSTM 的 state size 设为 512. 训练过程中，采取带冲量的梯度下降法，并采用分段式学习率. VGG-16 的卷积层的学习率为 0.01，全连接层学习率为 0.02，其余层的学习率为 0.1. 所有学习率每 10 个 epoch 降为原先的 1/10，共采用 40 个 epoch. 采用本文-f_n 表示基于注意力反馈机制的图像自动语句标注模型的结果，n 表示第 n 次迭代过程. 本文中，对 $n\in\{1,2,3\}$ 进行评估.

本文中采用的评估指标包括 BLEU-1, BLEU-2, BLEU-3, BLEU-4 和 Meteor. BLEU[27]是机器翻译中一种十分常用的衡量指标，通过分析生成句子和一系列真实语境句子的 n-gram[28]同现性来判断句子的优劣；其中，BLEU-n($n=\{1,2,3,4\}$)表示 n-gram 的长度. BLEU 的优势在于采用了 n-gram，从而考虑了更长的匹配信息，但是随着 n 的增加，在句子层次的匹配会越来越差. Meteor[29]通过计算生成语句和真实语句的单词匹配校准来衡量优劣. Meteor 考虑基于整个语料库上的准确率和召回率，得出最终结果. 本文在 Flickr8K 和 Flickr30K 上验证模型的有效性. 进行对比的模型包括 Mind's eye[30], BRNN[31], Google NIC[1], Multimodal[32], Soft-attention, Hard-attention[2], Mix6v4[33], Bi-LSTM[34], phi-LSTM[35]和(sf+ra)-beam[36]. Mind's eye 提出一种双向表达方式，可以从图像生成语句，也可以从语句中生成图像信息表达. BRNN 利用一种多模态 RNN，结合文本图像匹配信息来生成新的文本. Google NIC 利用 CNN 从图像中提取特征，结合 RNN 生成新的语句. Multimodal 利用多模态信息生成语句. Soft-attention 和 Hard-attention 都

利用了注意力机制, 使得在生成句子时能集中在关键区域. Mix6v4, Bi-LSTM 和 phi-LSTM 则通过不同的方式修改 LSTM 在模型中的运用来提高效果. (sf+ra)-beam 引入了 region 信息强化运算.

从表 1 可以看出, 本文模型比其他模型有更好的效果, 在 BLEU-1 和 Meteor 指标上的循环过程中, 本文-f_3的结果分别是 68.3 和 23.0, 所有指标都比其他模型有所提升. 同样, 从表 2 中可以看出, 在 BLEU-1 和 Meteor 指标上的循环过程中, 本文-f_3结果分别是 67.5 和 20.1, 所有指标都比其他模型有所提升. BLEU-2~BLEU-4 相比(sf+ra)-beam[36]效果稍显不足, 这是由于该模型中引入了 region 信息并采用 beam search 的方式. 每一次反馈循环的效果都有一定提升, 说明本文模型在经过多次迭代更新图像的关注区域, 同时使得图像自动语句标注过程中可以优化生成的语句.

表 1　在 **Flickr8K** 上的实验结果

模型	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Meteor
Mind's eye[30]				13.1	16.9
BRNN[31]	57.9	38.3	24.5	16.0	
Google NIC[1]	63.0	41.0	27.0		
Multimodal[32]	65.6	42.4	27.7	17.7	17.3
Soft-attention[2]	67.0	44.8	29.9	19.5	18.9
Hard-attention[2]	67.0	45.7	31.4	21.3	20.3
Mix6v4[33]	64.7				18.8
Bi-LSTM^V[34]	65.5	46.8	32.0	21.5	
phi-LSTM[35]	63.6	43.6	27.6	16.6	
(sf+ra)-beam[36]	66.5	**47.8**	**33.2**	**22.4**	20.8
本文-f_1	67.2	46.1	30.9	21.2	21.5
本文-f_2	67.5	46.9	31.2	21.8	21.9
本文-f_3	**68.3**	46.5	32.1	22.1	**22.0**

表 2　在 **Flickr30K** 上的实验结果

模型	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Meteor
Mind's eye[30]				12.0	15.2
BRNN[31]	57.3	36.9	24.0	15.7	
Google NIC[1]	66.3	42.3	27.7	18.3	
Multimodal[32]	60.0	38.0	25.4	17.1	16.8
Soft-attention[2]	66.7	43.4	28.8	19.1	18.5
Hard-attention[2]	66.9	43.9	29.6	19.9	18.5
Bi-LSTM^V[34]	62.1	42.6	28.1	19.3	
phi-LSTM[35]	66.6	45.2	28.2	17.0	
(sf+ra)-beam[36]	67.0	**47.5**	**33.0**	**24.3**	19.4
本文-f_1	66.3	44.3	28.9	19.9	18.2
本文-f_2	66.8	44.1	30.1	19.7	19.1
本文-f_3	**67.5**	44.5	30.0	20.3	**20.1**

从表 3 可以看出, 通过 3 次迭代, BLEU-4 可以和 Review 网络[20]达到一致, 且 Meteor 指标比 Review 网络效果要更好一些. 说明通过循环反馈模型, 能够提高单词匹配的准确率.

本文在 Flickr8K 上对图像中的关注区域和文本中的关键字进行可视化. 计算图像的关注区域时, 本文利用高斯滤波器进行上采样, 放大因子为 16. 在计算文本上的注意力中, 本文通过比较 $\boldsymbol{\beta}=\{\beta_1,\beta_2,\cdots,\beta_T\}$, 即每个单词的权重, 用红色代表权重大的单词, 蓝色为较大, 其余为黑色. 如图 6 所示, 经过 3 次反馈迭代过程, 图像的关注区域

表 3　在 **MSCOCO** 上的实验结果

模型	BLEU-4	Meteor
Google NIC[1]	58.7	34.6
Review 网络[20]	**59.7**	34.7
本文-f_1	58.8	34.2
本文-f_2	59.2	34.7
本文-f_3	59.5	**35.1**

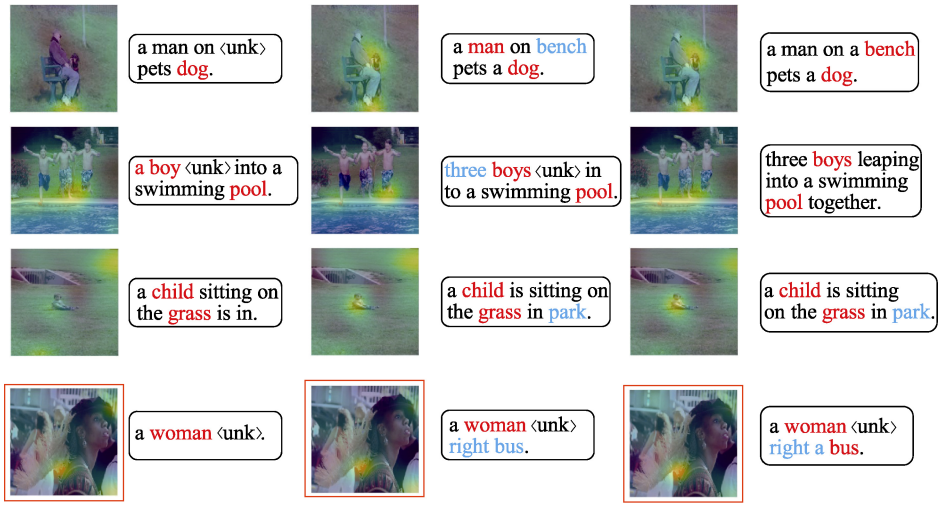


图 6 图像中的关注区域和文本中的关键字可视化

明显更加准确，生成语句更加通顺. 图 6 中，本文同时给出了一个失败例子(标红框)，可以看出，当图像较为复杂的时候，反馈信息不能保证对图像的注意力进行有效的矫正更新.

本文模型在单块 NVIDIA GeForce 1080 GPU 上运行实验，由于采用了端到端的结构，在 Flickr8K 上训练 40 个 epoch 仅需耗时 3~4 h，对比 Google NIC 等采用大量训练技巧的方法缩短了时间.

4 结 语

本文提出一种注意力反馈机制的图像自动语句标注模型，迭代地修正图像中的关注区域、强化图像和文本中的关键信息匹配、优化生成语句. 本文在 Flickr8K、Flickr30K 和 MSCOCO 这 3 个常用的数据集上进行实验，验证结果显示了模型的优越性.

参考文献(References):

[1] 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. Los Alamitos: IEEE Computer Society Press, 2015: 3156-3164

[2] Xu K, Ba J, Kiros R, *et al*. Show, attend and tell: neural image caption generation with visual attention[C] //Proceedings of International Conference on Machine Learning. Madison: Omnipress, 2015: 2048-2057

[3] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C] //Proceedings of the 25th International Conference on Neural Information Processing Systems. New York: Curran Associates Press, 2012, 1: 1097-1105

[4] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[OL]. [2018-08-27]. https://arxiv.org/abs/1409.1556

[5] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 770-778

[6] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[OL]. [2018-08-27]. https://arxiv.org/abs/1409.0473

[7] Sun Feng, Qin Kaihuai, Sun Wei, *et al*. Image saliency detection based on region merging[J]. Journal of Computer-Aided Design & Computer Graphics, 2016, 28(10): 1679-1687(in Chinese)
(孙 丰, 秦开怀, 孙 伟, 等. 基于区域合并的图像显著性检测[J]. 计算机辅助设计与图形学学报, 2016, 28(10): 1679-1687)

[8] Gao Sihan, Zhang Lei, Li Chenglong, *et al*. Image saliency detection via graph representation with fusing low-level and high-level features[J]. Journal of Computer-Aided Design & Computer Graphics, 2016, 28(3): 420-426(in Chinese)
(高思晗, 张 雷, 李成龙, 等. 融合低层和高层特征图表示的图像显著性检测算法[J]. 计算机辅助设计与图形学学报, 2016, 28(3): 420-426)

[9] You Q Z, Jin H L, Wang Z W, *et al*. Image captioning with semantic attention[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 4651-4659

[10] Gu J X, Cai J F, Wang G, *et al*. Stack-captioning: coarse-to-fine learning for image captioning[C] //Proceedings of AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2018: 6837-6844

[11] Gan Z, Gan C, He X D, *et al*. Semantic compositional networks for visual captioning[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 5630-5639

[12] Mao J H, Xu W, Yang Y, *et al*. Deep captioning with multimodal recurrent neural networks (m-RNN)[OL]. [2018-08-27]. https://arxiv.org/abs/1412.6632

[13] Wu Q, Shen C H, Liu L Q, *et al*. What value do explicit high level concepts have in vision to language problems?[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 203-212

[14] Karpathy A, Li F F. Deep visual-semantic alignments for generating image descriptions[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 3128-3137

[15] Johnson J, Karpathy A, Li F F. DenseCap: fully convolutional localization networks for dense captioning[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 4565-4574

[16] Rensink R A. The dynamic representation of scenes[J]. Visual Cognition, 2000, 7(1-3): 17-42

[17] Liu C X, Mao J H, Sha F, *et al*. Attention correctness in neural image captioning[C] //Proceedings of the 31st AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2017: 4176-4182

[18] Liu C, Sun F C, Wang C H, *et al*. MAT: a multimodal attentive translator for image captioning[OL]. [2018-08-27]. https://arxiv.org/abs/1702.05658

[19] Li L H, Tang S, Deng L X, *et al*. Image caption with global-local attention[C] //Proceedings of the 31st AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2017: 4133-4139

[20] Yang Z L, Yuan Y, Wu Y X, *et al*. Review networks for caption generation[C] //Proceedings of Advances in Neural Information Processing Systems. New York: Curran Associates, 2016: 2369-2377

[21] Cavana R Y. Modeling the environment: an introduction to system dynamics models of environmental systems[J]. System Dynamics Review, 2003, 19(2): 171-173

[22] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780

[23] Zaremba W, Sutskever I. Learning to execute[OL]. [2018-08-27]. https://arxiv.org/abs/1410.4615

[24] Young P, Lai A, Hodosh M, *et al*. From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions[J]. Transactions of the Association for Computational Linguistics, 2014, 2: 67-78

[25] 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

[26] Lin T Y, Maire M, Belongie S, *et al*. Microsoft coco: common objects in context[C] //Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2014: 740-755

[27] 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: ACL Press, 2002: 311-318

[28] Brown P F, Desouza P V, Mercer R L, *et al*. Class-based *n*-gram models of natural language[J]. Computational Linguistics, 1992, 18(4): 467-479

[29] Lavie A, Agarwal A. Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments[C] //Proceedings of the 2nd Workshop on Statistical Machine Translation. Stroudsburg: ACL Press, 2007: 228-231

[30] Chen X L, Lawrence Zitnick C. Mind's eye: a recurrent visual representation for image caption generation[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 2422-2431

[31] Karpathy A, Li F F. Deep visual-semantic alignments for generating image descriptions[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015, 1: 3128-3137

[32] Kiros R, Salakhutdinov R, Zemel R. Multimodal neural language models[C] //Proceedings of the 31st International Conference on Machine Learning. Madison: Omnipress, 2014, 32: II-595-II-603

[33] Wang M S, Song L, Yang X K, *et al*. A parallel-fusion RNN-LSTM architecture for image caption generation[C] //Proceedings of the IEEE International Conference on Image Processing. Los Alamitos: IEEE Computer Society Press, 2016: 4448-4452

[34] Wang C, Yang H J, Bartz C, *et al*. Image captioning with deep bidirectional LSTMs[C] //Proceedings of the 24th ACM International Conference on Multimedia. New York: ACM Press, 2016: 988-997

[35] Tan Y H, Chan C S. phi-LSTM: a phrase-based hierarchical LSTM model for image captioning[C] //Proceedings of Asian Conference on Computer Vision. Heidelberg: Springer, 2016: 101-117

[36] Fu K, Jin J Q, Cui R P, *et al*. Aligning where to see and what to tell: image captioning with region-based attention and scene-specific contexts[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39: 2321-2334