DOI:10.13878/j.cnki.jnuist.2019.03.007

李金轩[1]　杜军平[1]　周南[1]

# 基于注意力特征提取网络的图像描述生成算法

**摘要**

针对解决图像描述生成中对浅层图像特征利用不充分、图像目标间关系提取不足的问题,提出一种基于注意力图像特征提取的图像描述生成算法.通过语言模型上下文信息对不同深度图像特征进行自适应注意力权重分配,使带有注意力的图像特征参与指导图像描述生成,提升了图像描述生成的效果.在MSCOCO测试集中所提算法的BLEU-1和CIDEr得分分别达到0.752和0.934,从而验证了所提算法的有效性.

**关键词**

注意力机制;图像描述;长短期记忆网络;图像特征提取

**中图分类号** TP391.1

**文献标志码** A

## 0　引言

随着移动互联网的快速发展,移动社交平台丰富了人们的日常生活.社交网络中热点话题数据中包含着大量的图像数据,使用人工方法对每张图像进行内容标注的成本也越来越高.因此,使用智能方法自动提取出图像特征,并对图像表达内容进行描述,已经成为计算机视觉领域的研究热点.通过对输入图像进行特征提取,自动生成相应的文字描述图像内容是进行跨媒体信息关联的有效手段.图像描述自动提取的质量取决于对图像中目标的识别能力以及图像中目标间相对关系的识别能力.图像描述自动提取的方法可以分为三种:1)基于已有描述语句特征与图像特征进行匹配的方法,这种方法缺乏对特定图像的细节描述;2)基于固定文本描述模板添加关键词的方法,该方法依赖于模板类型,描述缺乏多样性;3)基于神经网络的方法,通过编码器对输入图像进行编码并将其作为图像特征,再通过解码器对图像特征进行解码,生成最终的文本描述.基于神经网络的方法通常使用卷积神经网络作为编码器,循环神经网络作为解码器.通过端对端的方法训练模型,利用卷积神经网络(Convolutional Neural Networks,CNN)特征提取的优势与循环神经网络(Recurrent Neural Network,RNN)自然语言处理的优势,共同指导图像描述的生成.这种方法使得文字描述解决了类型单一、缺乏细节的问题,也使得这种模型在近几年研究中频繁出现.

注意力机制在视觉特征描述等任务中已经被证明是有效的,这些注意力机制常常被应用于语言模型中,语言模型可以决定应该关注图像特征的哪个位置.本文工作将注意力机制引入特征提取模型中,根据语言模型的反馈自适应地全局调整不同尺度图像特征的注意力权重.

社交网络图像数据由于受拍摄设备、拍摄场景等条件限制,所得图像往往存在分辨率较低、图像主题与背景划分不清晰的问题,因此传统图像描述生成算法存在以下不足:1)使用卷积神经网络提取图像特征时,仅提取了图像的深度特征,而忽略了底层的语义特征,深度特征对图像中物体识别准确率有较大贡献,但是对不同物体间关系的识别准确率贡献不高,单纯提取深度语义在图像特征利用方面具有不足;2)传统的特征提取方法提取的图像特征对特征大小进行了固定,图像特征无法对语言模型的上下文信息进行自适应调整,图

**收稿日期** 2019-05-16

**资助项目** 国家自然科学基金(61772083,61532006,61877006,61802028);广西科技重大专项(桂科AA18118054)

**作者简介**

李金轩,男,硕士生,研究方向为机器学习、图像处理和模式识别.1206656459@qq.com

杜军平(通信作者),女,博士,教授,博士生导师,研究方向为人工智能、图像处理和模式识别.junpingdu@126.com

1　北京邮电大学　智能通信软件与多媒体北京市重点实验室/计算机学院,北京,100876

像特征在图像描述工作中利用得不充分;3)传统的图像特征提取与语言模型训练过程是分开的,没有考虑到使用语言模型反向训练图像提取,难以将两种网络结合起来.

针对以上研究中存在的问题,本文提出了一种基于注意力图像特征融合的图像描述生成算法.该算法通过 CNN 提取不同尺度的图像特征并输入多层长短期记忆力网络(Long and Short Term Memory network,LSTM)中,根据语言模型中上下文信息的反馈自适应选取图像中各区域的特征.本文的主要贡献如下:1)提出一种融合卷积与注意力机制的图像特征提取算法,提取了多个尺度的图像特征;2)可以动态地提取随时间变化的上下文注意力信息,根据语言模型的预测目标自适应决定输出的注意力图像特征;3)在语言解码过程中,使用多层 LSTM 网络结构,充分利用图像所蕴含的内容,提高图像描述的精度.

## 1 相关工作

随着深度学习的发展,以 Mao 等[1]提出的多模态循环神经网络(multimodal RNN,m-RNN)为代表的基于神经网络的方法开始被广泛应用.m-RNN 将图像描述的工作转化成两个任务:采用 CNN 进行图像提取、RNN 建立语言模型将特征转化成文本输出.m-RNN 中的 CNN 使用 AlexNet 结构,RNN 使用两层编码层将文本转换成 One-hot 向量表示,输入到循环层中,传入 Softmax 层得到输出.尽管 m-RNN 第一次将编码-解码方式引入图像描述工作中,但是由于 RNN 网络结构限制,对于较长的网络序列容易出现梯度消失的问题.Vinvals 等[2]使用 LSTM 网络替代 RNN,提高对长期依赖信息的学习效率,并使用了带有标准化层的 CNN 提取图像特征,算法准确率和速度均有提升.在视频语言描述工作中,Liang 等[3]使用 AlexNet 模型与 VGG 模型[4]分别提取视频空间特征,使用多特征融合的方法,将空间特征与运动特征及视频时间特征融合后输入 LSTM 语言描述模型,提升了视频语言描述的准确性.

Xu 等[5]使用 Bahdanau 等[6]提出的注意力机制将自然语言处理领域中的注意力机制引入到图像描述工作中,使得每个 LSTM 模块在训练隐含层时可以自主决定关注图像特征中的位置,使得描述结果的准确性和细节丰富程度都有所提高.Lu 等[7]提出了自适应注意力结构,LSTM 的每一层可以通过图像和语言模型自适应输出文字描述.Vaswani 等[8]使用基于注意力机制的翻译模型,提高了深度学习网络的并行能力.

目前,深度神经网络可以学习到更加丰富的特征,计算机视觉领域的快速发展得益于更深的网络模型.但是若深度神经网络的深度突破限制,准确率提升速度反而会饱和甚至下降[9].He 等[10]提出的残差网络(Residual Learning Network,ResNet)有效地解决了网络退化问题.

## 2 基于注意力特征提取网络的图像描述生成算法

本文提出了一种结合注意力机制对图像特征提取的图像描述生成算法.通过对不同深度图像特征进行自适应权重分配,使得输出的图像特征的目标区域得到增强,同时使图像背景区域对前景特征的影响有所限制.本文的算法由基于注意力的图像特征提取和语言生成两部分构成,如图 1 所示.

### 2.1 基本定义

对要处理的问题进行基本定义,令 $I=(I_1,I_2,\cdots,I_n)$ 为一个有 $n$ 个图像-文本实例的数据集,其中 $I_i=(x_i,T_i)$,$x_i$ 表示原始图像,$T=(t_{i,1},t_{i,2},\cdots,t_{i,m})$ 表示第 $i$ 张图像的 $m$ 个人工描述构成的描述集.定义 $X=(x_1,x_2,\cdots,x_n)$ 为图像特征提取网络的输入,$V=(v_1,v_2,\cdots,v_n)$ 为图像特征提取网络的输出.定义 $Y=(y_1,y_2,\cdots,y_n)$ 为语言模型输入的词向量.

### 2.2 图像特征提取

通过多个注意力结构进行堆叠来构建图像特征,每个注意力结构由两个分支构成:采样分支与主干分支.主干分支可以适应多种前沿网络结构.将 VGG 网络的一个卷积层作为注意力结构的主干分支,对于输入的一个原始图像 $x_i$,主干分支输出特征图 $F(x_{i,c})=\mathrm{conv3}(x_{i,c})$.

通过提取浅层特征如 SIFT 特征[11]等,可以对图像细节进行表示,如局部边缘、颜色等,深度特征如深度卷积特征,可以对包含在图像的全局语义中的信息进行表示,这意味着深度特征相比于浅层特征更关注图像的显著性区域.因此采样分支任务提取输入特征图的深度特征,关注特征图中语义相关的区域,每个注意力结构的输出为 $X_{l-1}=[x_1,x_2,\cdots,x_n]$,$n=W\times H$.其中,$x_i$ 是第 $i$ 个区域的特征,采样分支结构如图 2 所示.计算过程“×”表示向量笛卡尔积.采样分支的目的可为当前输入的特征图计算每

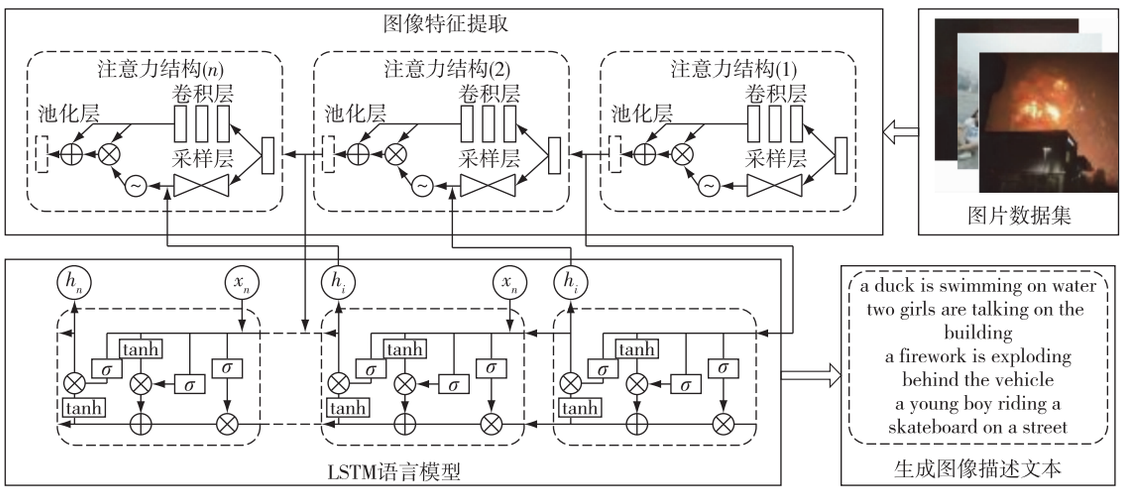


图 1　基于注意力特征提取网络的图像描述生成算法

Fig. 1　Image description generation algorithm based on attention feature extraction network

个像素的注意力权重,图像的深度特征中可以体现出与图像中重要目标关联性高的区域,因此采样分支需要通过多次卷积提取输入特征图 $x_i$ 的深度特征.由于卷积操作使得特征图大小降低,需要在提取深度特征后通过反卷积层,将深度特征放大到与输入特征图大小一致.在得到 LSTM 网络前一个时刻的隐含层状态 $h_{t-1}$ 后,使用单层神经网络将隐含层状态与卷积生成特征图进行融合.Sigmoid 激活函数层将特征图归一化到(0,1)之间.输出结果如式(1)、(2)、(3)所示:

$$V_l = \mathrm{CNN}(X_{l-1}), \tag{1}$$

$$a_c = \tanh((W_s \times V_l + b_s) \oplus W_{hs} \times h_{t-1}), \tag{2}$$

$$M(x_{i,c}) = \frac{1}{1 + \exp(-(a_c))}, \tag{3}$$

其中 $x_{i,c}$ 表示输入特征图,$c$ 表示注意力结构层数,$W_s$,$W_{hs}$,$b_s$ 为待学习的线性变换参数,$V_l$ 表示对前一个注意力结构输出特征进行卷积,作为下一个注意力结构的输入.

将采样分支的输出 $M(x_{i,c})$ 与主干分支的输出 $F(x_{i,c})$ 进行对位相乘,主干分支输出的每个像素都经过了注意力权重处理,注意力结构的输出 $A(x_{i,c})$ 如式(4)所示:

$$A(x_{i,c}) = M(x_{i,c}) \otimes F(x_{i,c}), \tag{4}$$

其中,⊗表示对位相乘.

注意力模块有利于增强每一层特征图中的重要的部分,多层注意力结构叠加会导致模型的性能大幅下降,原因是采样分支的输出经过了 Sigmoid 函数进行归一化,再与主干分支进行对位相乘,使得该层中部分特征值遭到抑制,当多个注意力结构进行堆叠计算后,可能造成最终输出的特征图中每个像素的特征值都很低.为了解决上述问题,注意力结构输出在采样分支与主干分支对位相乘的基础上,再与主干分支进行对位相加,注意力结构输出结果如式(5)所示:

$$A(x_{i,c}) = (M(x_{i,c}) \otimes F(x_{i,c})) \oplus F(x_{i,c}), \tag{5}$$

其中,⊕表示对位相加.

在主干分支卷积神经网络进行拟合得到的特征图基础上,结合采样分支输出的注意力特征,可以使

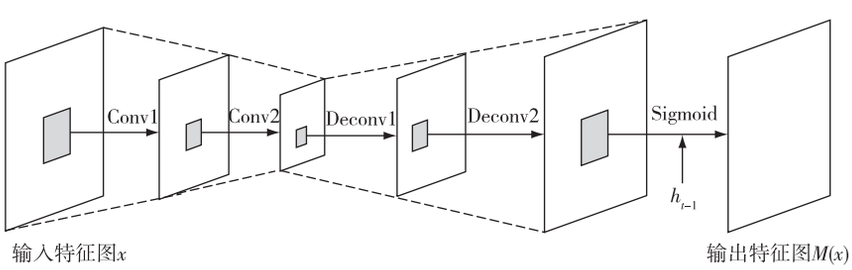


图 2　采样分支结构

Fig. 2　Sampling branching structure

得主干分支输出的特征中重要的特征被增强,不重要的特征被抑制,这使得特征 $A(x_{i,c})$ 与 $F(x_{i,c})$ 构成了一种恒等映射,使每层注意力结构输出特征中包含的语义信息不会比主干分支输出特征中包含的语义信息差.随着注意力结构的增多,模型更加关注对目标提取有帮助的目标.

### 2.3 语言生成模型

使用 LSTM 作为语言生成模型的基本单元.语言模型结构如图 3 所示.

初始化 0 时刻第一层 LSTM 的隐含层网络,该层输入为第一个注意力结构输出的图像特征 $A(x_{i,1})$,通过一次线性变换与 ReLU 激活函数,将输入的图像特征投影为维度为 $d$ 的初始化隐含层.

$$h'_0 = \mathrm{ReLU}(W_0 \times A(x_{i,1}) + b_0), \tag{6}$$

其中 $W_0$ 与 $b_0$ 是待学习的线性变换的参数.式(6)得到的结果即为初始化后的隐含层.每一层 LSTM 语言模型的输入包括三部分,分别是 $W_{\mathrm{input}}$,$V_1$,$h^n_{t-1}$.$h^n_{t-1}$ 表示上一时刻第 $n$ 层(最终层)LSTM 的隐含层状态,$t-1$ 表示上一时刻;$W_{\mathrm{input}}$ 表示经过编码后的词向量;$v_i$ 表示经过第 $i$ 个注意力结构提取的图像特征.输入数据同时包含了图像特征和文字相关的上下文特征.

将从图像中提取的多尺度特征依次输入每一层 LSTM 中,将语言模型第 $n-1$ 层隐含层,即 $h_{n-1}$ 词向量 $W_{\mathrm{input}}$ 与最后一层注意力结构输出的图像特征 $A(x_{i,c})$ 相结合,输入最后一层 LSTM 语言模型中.

$$h_t = \mathrm{LSTM}(h_{n-1}, A(x_{i,c}), W_{\mathrm{input}}). \tag{7}$$

将最后一层 LSTM 输出的维度为 $d$ 的隐含层映射到维度为 $m$ 的向量中,其中 $m$ 表示语义字典中单词的数量.通过 Softmax 层选出 LSTM 模型每一时刻的输出中概率最大的词连接成描述句,作为模型的最终输出结果.

$$\overline{y_t} \sim p_t = \mathrm{Softmax}(h_t, W_{\mathrm{input}}). \tag{8}$$

采用图像描述生成任务常用的交叉熵作为损失函数进行模型训练,损失函数形式如式(9)所示:

$$\mathrm{Loss}(\theta) = -\sum_{t=1}^{T} \ln(p_\theta(\overline{y_t} \mid \overline{y_{1:t-1}})), \tag{9}$$

其中:$y_{1:T}$ 与 $\theta$ 分别表示目标描述的真实词序列和图像描述生成模型中解码器的参数;$(p_\theta(\overline{y_t} \mid \overline{y_{1:t-1}})$ 是 LSTM 语言模型输出单词 $\overline{y_t}$ 的概率.

基于注意力图像特征提取网络的图像描述生成过程如算法 1 所示.

**算法 1:基于注意力图像特征提取网络的图像描述生成**

输入:图像数据集、Wiki 文本数据集
输出:图像特征描述文本
对于数据集中每张图像采取如下步骤:
Step1.提取第 1 层图像特征 $V_1$;
Step2.将该层图像特征传入第 1 层 LSTM 初始化 $h_0$;
Step3.提取第 $i$ 层图像特征 $V_i$;
Step4.将词向量 $W_{\mathrm{input}}$、前一层 LSTM 隐含层 $h^n_{t-1}$、图像特征 $V_i$ 输入下一层 LSTM,计算下一个输出的单词;
Step5.通过交叉熵计算损失 Loss,反馈调整参数;
Step6.返回 Step3,直到输出为<END>或达到句子最大长度;
Step7.返回图像描述文本.

## 3 实验及结果分析

### 3.1 数据集及实验环境

在 MSCOCO(Microsoft Common Objects in Context)和微博突发事件图像数据集上进行验证,在微博突发事件图像数据集上进行了图像描述.MSCOCO 数据集为每一张图像的描述至少有 5 种,对微博突发事件图像数据集进行人工标注时,对每张图像标注了 5 种描述.将 MSCOCO 数据集分为训练集、验证集、测试集,分别包含 113 287、5 000、5 000 张图像.微博突发事件图像数据集分为训练集、测试集,分别包含 66 792、1 801 张图像.分析这些数据集中所有

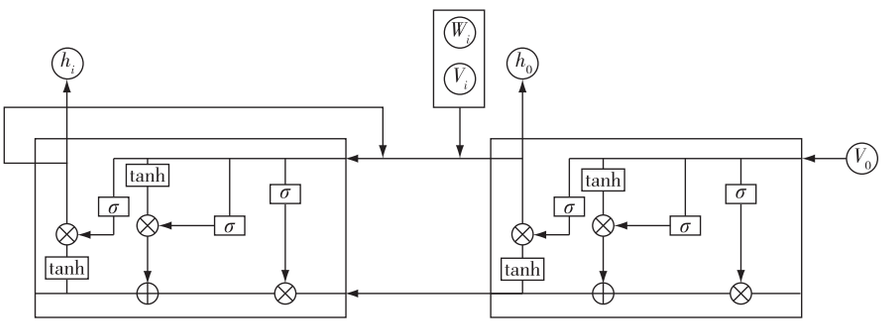


图 3 语言模型结构

Fig. 3 Language model structure

图像描述句子的长度,描述句子的长度集中在 11 至 18 个单词之间,在实验中设定输出描述的长度为 20,若描述文本长度小于 20,则使用<PAD>进行占位.

实验环境为运行于 Ubuntu 环境的 Pytorch 深度学习框架,配置 NVIDIA CUDA 9.0,cuDNN 7.5 深度学习库加速 GPU 计算,Python 环境为 Python3.5.

已有的图像描述生成的评测标准包括人工评价与客观量化评分.主观评价即人工观测图像,对生成的图像描述进行评价.目前普遍使用的客观量化评分包括 BLEU(BiLingual Evaluation Understudy)[12]、METEOR(Metric for Evaluation of Translation with Explicit Ordering)[13]、CIDEr(Consensus-based Image Description Evaluation)[14].

### 3.2 主要参数设置

在进行语言模型训练之前,使用 Wiki 数据集构建语料库,通过预训练的 Word2Vec 将单词表示为[1,300]维词向量形式.

为了让语言模型取得较好的图像描述效果,训练时将模型中每一个 LSTM 的隐含层节点数设置为 1 000.为使最终得到的文字描述合理,在训练中采用了集束搜索的方法,并将 Beam 的大小设置为 3.使用 Dropout 方法,将 LSTM 中的单元按照一定的概率进行屏蔽来防止过拟合,实验中 Dropout 设置为 0.5.

### 3.3 实验方法

在训练模型时采用 Adam 作为优化器,学习率设置为 0.000 1,权重衰减设置为 0.000 01,Batchsize 大小设置为 192,进行 20 轮训练.训练结果如图 4、5 所示,其中 Batch 表示训练次数,Loss 表示训练误差,Perplexity 表示训练生成的描述文本的混淆度,ATTIC_COCO 表示本文算法在 MSCOCO 数据集上的训练结果,VGGIC_COCO 表示使用 VGG 网络进行图像特征提取的图像描述生成算法在 MSCOCO 数据集上的训练结果,ATTIC_WEIBO 与 VGGIC_COCO 则分别表示两种算法在社交网络微博数据集上的训练结果.

对模型进行训练,在 MSCOCO 数据集的 CIDEr 的得分达到 0.947,BLEU-1 的得分达到 0.752.使用客观量化评分方法对本文算法结果进行评价,在使用相同数据集、相同训练条件下,本文算法生成的图像描述的客观量化评分较高.

为了验证本文算法中基于注意力特征提取网络对图像描述生成的效果影响,对比了不含注意力结构的卷积神经网络 VGG 与本文特征提取网络,使用 BeamSearch 方法来验证算法.随着集束搜索的值增加,模型生成的图像描述评分会增大,当集束搜索的值设置为 3 时,模型的客观评价指标最高.

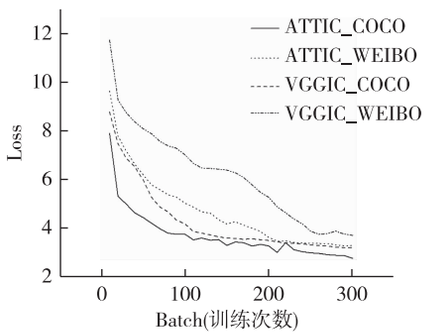


图 4 不同数据集下 Loss 随 Batch 次数变化趋势

Fig. 4 The variation trend of Loss with Batch times in different data sets

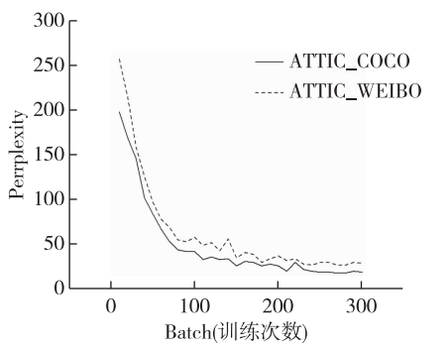


图 5 不同数据集下 Perplexity 随 Batch 次数变化趋势

Fig. 5 The variation trend of Perplexity varies with Batch times in different data sets

### 3.4 实验结果及分析

使用客观量化评分方法对本文算法生成的图像描述进行评价.本文算法评分与 mRNN[1]、GoogleNIC(Google Neural Image Caption)[2]、DeepVS(Deep Visual-Semantic alignments)[15]、ERD(Encode Review and Decode)[16]、Hard-Attention[17]、VGG_LSTM、ATTIC 在两种数据集上对比结果如表 1、2 所示,其中 VGG_LSTM 表示使用 VGG 网络进行图像特征提取并使用 LSTM 网络作为语言生成模型的算法,ATTIC 表示本文提出的基于注意力特征提取网络的图像描述生成算法.

图像描述生成工作分为两个部分:一是图像特征提取,二是结合图像特征来建立语言模型.本文算法在 BLEU[12]、METEOR[13]、CIDEr[14]指标上对比其他算法都有一定的提升.在图像特征提取中引入注

意力机制与使用传统卷积神经网络提取图像特征相比各个指标均有提升.本文算法图像描述的结果提升主要是由于以下原因:使用基于注意力机制的改进卷积神经网络,使得对由浅入深每一层的特征都突出了重要的部分,抑制了不重要的部分,提取得到的图像特征可以更好地表征图像中的目标与目标间的相关关系,为训练语言模型提供了较好的初始化条件,可以提高最终图像描述效果;增加 LSTM 语言模型中隐含层节点数量,让隐含层可以携带更多的语义信息,能够有效提升本文算法的效果.在此基础上,加入集束搜索可以提升模型的效果;采用多层注意力结构堆叠,并让语言模型上下文信息参与指导各个尺度图像的特征提取,可以提升图像描述的效果.

**表 1　不同算法在 MSCOCO 数据集下评价指标**

Table 1　Different algorithms evaluate indexes under MSCOCO data sets

| 算法 | BLEU-1 | BLEU-4 | METEOR | CIDEr |
|---|---|---|---|---|
| mRNN[1] | 0.670 | 0.240 | — | — |
| GoogleNIC[2] | — | 0.277 | 0.237 | 0.855 |
| Hard-Attention[17] | 0.718 | 0.250 | 0.230 | — |
| DeepVS[15] | 0.625 | 0.230 | 0.195 | 0.660 |
| ERD[16] | — | 0.288 | 0.240 | 0.895 |
| VGG_LSTM | 0.746 | 0.284 | 0.242 | 0.925 |
| ATTIC(本文) | 0.752 | 0.287 | 0.247 | 0.934 |

**表 2　不同算法在微博数据集下评价指标**

Table 2　Evaluation indexes of different algorithms under Weibo data sets

| 算法 | BLEU-1 | BLEU-4 | METEOR | CIDEr |
|---|---|---|---|---|
| mRNN[1] | 0.623 | 0.218 | — | — |
| CoocleNIC[2] | — | 0.252 | 0.225 | 0.803 |
| Hard-Attention[17] | 0.689 | 0.208 | 0.216 | — |
| DeepVS[15] | 0.602 | 0.198 | 0.178 | 0.618 |
| ERD[16] | — | 0.259 | 0.231 | 0.875 |
| VGG_LSTM | 0.732 | 0.252 | 0.235 | 0.902 |
| ATTIC(本文) | 0.737 | 0.255 | 0.239 | 0.914 |

使用 BLEU、METEOR、CIDEr 分别对生成的图像描述的准确性和相关性进行评价,结果表明本文算法(ATTIC)通过将语言模型输出的隐含层与不同深度的图像特征结合,输出对不同深度图像特征的注意力关注区域,增加了图像提取的网络深度,同时语言模型对不同关注区域的图像信息给出了相应的文本描述,在三种评价指标上均取得较高得分.

## 4　结论

本文提出了基于注意力特征提取网络的图像描述生成算法,通过堆叠注意力结构,使语言模型上下文信息参与指导图像特征,提取网络提取图像特征,将不同尺度图像特征输入 LSTM 语言模型中生成图像描述.本文算法的图像描述效果在 BLEU、METEOE、CIDEr 等评价指标上都取得较高得分.实验结果表明基于注意力图像特征提取网络可以根据上下文信息优化图像特征提取结果,并因此提升最终的描述效果.

## 参考文献

References

[1] Mao J,Xu W,Yang Y,et al.Explain images with multimodal recurrent neural networks[J].arXiv Preprint arXiv:1410.1090,2014

[2] Vinyals O,Toshev A,Bengio S,et al.Show and tell:a neural image caption generator[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2015.Boston,MA,USA.New York,USA:IEEE,2015:3156-3164

[3] Liang R,Zhu Q X,Liao S J,et al.Deep natural language description method for video based on multi-feature fusion[J].Journal of Computer Applications,2017,37(4):1179-1184

[4] Karen S,Andrew Z.Verydeep convolutional networks for large-scale image recognition[J].arXiv Preprint arXiv:1409.1556,2014

[5] Xu K,Ba J L,Kiros R,et al.Show,attend and tell:neural image caption generation with visual attention[EB/OL].[2018-05-08].https://arxiv.org/pdf/1502.03044.pdf

[6] Bahdanau D,Cho H,Bengio Y.Neural machine translation by jointly learning to align and translate[EB/OL].[2018-05-10].https://arxiv.org/pdf/1409.0473.pdf

[7] Lu J S,Xiong C M,Parikh D,et al.Knowing when to look:adaptive attention via a visual sentinel for image captioning[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2017.Honolulu,HI,USA.New York,USA:IEEE,2017:3242-3250

[8] Vaswani A,Shazeer N,Parmar N,et al.Attention is all you need[EB/OL].[2018-05-10].https://arxiv.org/pdf/1706.03762.pdf

[9] Gong C,Tao D C,Liu W,et al.Label propagation via teaching-to-learn and learning-to-teach[J].IEEE Transactions on Neural Networks and Learning Systems,2017,28(6):1452-1465

[10] He K M,Zhang X Y,Ren S Q,et al.Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2016.Las Vegas,NV,USA.New York,USA:IEEE,2016:770-778

[11] David G.Lowedistinctive image featuresfromscale-

invariant keypoints[EB/OL].[2018-05-10].https://www.cs.ubc.ca/~lowe/papers/ijcv04.pdf
[12] 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,PA:ACL,2002:311-318
[13] Banerjee S,Lavie A.METEOR:an automatic metric for MT evaluation with improved correlation with human judgments[C]//Proceedings of the 2005 ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization,Stroudsburg,PA:ACL,2005:65-72
[14] Vedantam R,Zitnick C L,Parikh D.CIDEr:consensus-based image description evaluation[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition(CVPR),2015. Boston,MA,USA. New York,USA:IEEE,2015:4566-4575
[15] Karpathy A,Li F F.Deep visual-semantic alignments for generating image descriptions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition(CVPR),2015. Boston,MA,USA. New York,USA:IEEE,2015:3128-3137
[16] Yang Z,YuanY,Wu Y,et al. Encoder,review,anddecode:reviewer module for caption generation[J].arXiv Preprint arXiv:1605.07912,2016
[17] 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,2015:2048-2057

# Image caption algorithm based on an attention image feature extraction network

LI Jinxuan[1] DU Junping[1] ZHOU Nan[1]
1 Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia,
School of Computer Science,Beijing University of Posts and Telecommunications,Beijing 100876

**Abstract** To solve the problem of the lack of use of shallow image features in image captions and insufficient extraction of image objects,an image caption generation algorithm based on attention image feature extraction is proposed.Through context information of a language model,adaptive attention weight assignment is performed on different depth image features to ensure that the attention-grabbing image features guide the image caption generation,thereby improving the image caption effect.In the MSCOCO test set,the BLEU-1 and CIDEr scores of the proposed algorithm reached 0.752 and 0.934,respectively,thus verifying the effectiveness of the proposed method.

**Key words** attention mechanism;image caption;long and short term memory network;image feature extraction