Journal of Computer Applications
计算机应用,2021,41(9):2504 - 2509
ISSN 1001-9081
CODEN JYIIDU
2021-09-10
http://www.joca.cn

文章编号:1001-9081(2021)09-2504-06
DOI:10.11772/j.issn.1001-9081.2020111838

基于注意力机制的多层次编码和解码的图像描述模型

李康康,张　静*
(华东理工大学 信息科学与工程学院,上海 200237)
(*通信作者电子邮箱 jingzhang@ecust.edu.cn)

摘　要:图像描述任务是图像理解的一个重要分支,它不仅要求能够正确识别图像的内容,还要求能够生成在语法和语义上正确的句子。传统的基于编码器-解码器的模型不能充分利用图像特征并且解码方式单一。针对这些问题,提出一种基于注意力机制的多层次编码和解码的图像描述模型。首先使用Faster R-CNN(Faster Region-based Convolutional Neural Network)提取图像特征,然后采用Transformer提取图像的3种高层次特征,并利用金字塔型的融合方式对特征进行有效融合,最后构建3个长短期记忆(LSTM)网络对不同层次特征进行层次化解码。在解码部分,利用软注意力机制使得模型能够关注当前步骤所需要的重要信息。在MSCOCO大型数据集上进行实验,利用多种指标(BLEU、METEOR、ROUGE-L、CIDEr)对模型进行评价,该模型在指标BLEU-4、METEOR和CIDEr上相较于Recall(Recall what you see)模型分别提升了2.5个百分点、2.6个百分点和8.8个百分点;相较于HAF(Hierarchical Attention-based Fusion)模型分别提升了1.2个百分点、0.5个百分点和3.5个百分点。此外,通过可视化生成的描述语句可以看出,所提出模型所生成的描述语句能够准确反映图像内容。

关键词:图像描述;卷积神经网络;长短期记忆网络;多层次编码;多层次解码;注意力机制
中图分类号:TP391　　**文献标志码:**A

Multi-layer encoding and decoding model for image captioning based on attention mechanism

LI Kangkang, ZHANG Jing*
(*School of Information Science and Engineering, East China University of Science and Technology, Shanghai* 200237, *China*)

Abstract: The task of image captioning is an important branch of image understanding. It requires not only the ability to correctly recognize the image content, but also the ability to generate grammatically and semantically correct sentences. The traditional encoder-decoder based model cannot make full use of image features and has only a single decoding method. In response to these problems, a multi-layer encoding and decoding model for image captioning based on attention mechanism named MLED was proposed. Firstly, Faster Region-based Convolutional Neural Network (Faster R-CNN) was used to extract image features. Then, Transformer was employed to extract three kinds of high-level features of the image. At the same time, the pyramid fusion method was used to effectively fuse the features. Finally, three Long Short-Term Memory (LSTM) Networks were constructed to decode the features of different layers hierarchically. In the decoding part, the soft attention mechanism was used to enable the model to pay attention to the important information required at the current step. The proposed model was tested on MSCOCO dataset and evaluated by BLEU, METEOR, ROUGE-L and CIDEr. Experimental results show that on the indicators BLEU-4, METEOR and CIDEr, the model is increased by 2.5 percentage points, 2.6 percentage points and 8.8 percentage points compared to the Recall what you see (Recall) model respectively, and is improved by 1.2 percentage points, 0.5 percentage points and 3.5 percentage points compared to the Hierarchical Attention-based Fusion (HAF) model respectively. The visualization of the generated description sentences show that the sentence generated by the proposed model can accurately reflect the image content.

Key words: image captioning; Convolutional Neural Network (CNN); Long Short-Term Memory (LSTM) network; multi-layer encoding; multi-layer decoding; attention mechanism

0　引言

图像描述任务是图像理解中的重要研究内容,它结合了计算机视觉和自然语言处理两大任务。图像描述需要利用计算机视觉相关的技术准确地识别出图像中的内容,也需要利用自然语言处理中的文本生成的方法生成语法和语义上正确的句子。图像描述任务的关键在于如何充分利用图像特征以及如何有效地解码。

近年来大多数图像描述任务[1-2]都是基于编码器-解码器(Encoder-Decoder)的模型,编码器部分利用VGG(Visual Geometry Group)[3]、ResNet[4]等卷积神经网络(Convolutional

收稿日期:2020-11-23;**修回日期:**2021-02-21;**录用日期:**2021-02-22。　**基金项目:**国家自然科学基金资助项目(61402174)。
作者简介:李康康(1995—),男(蒙古族),河南洛阳人,硕士研究生,CCF会员,主要研究方向:计算机视觉、图像描述;　张静(1978—),女,河南三门峡人,副教授,博士,CCF会员,主要研究方向:深度学习、计算机视觉、图像检索、图像描述、视觉问答。

Neural Network, CNN)得到图像的特征,然后利用长短期记忆(Long Short-Term Memory, LSTM)网络[5]对特征进行解码从而得到生成的语句。Vinyals 等[1]首先将机器翻译[6]中的编码器-解码器结构引入图像描述任务,他们首先利用预训练的卷积神经网络得到图像特征,并将其输入到 LSTM 的第一个时间步骤,然后依次输入上一时间步骤生成的单词,得到描述语句的每个单词。但这种方式仅在第一个时间步骤输入图像特征,使得不是每个时间步骤都关注到图像的重点,而且图像特征信息会随着时间步骤的增加而逐渐减少。注意力机制[7]的使用使得解码器在 LSTM 的每个时间步骤都能关注到重要的信息,显著提升了模型的有效性。一些工作[8-10]对编码器部分作出改进,使用 Faster R-CNN (Faster Region-based Convolutional Neural Network)[11]提取对象特征或卷积神经网络的多层次特征;也有一些工作在解码器部分使用了 LSTM 的变体[9,12],例如双向 LSTM (Bidirectional LSTM, Bi-LSTM)[13]、双 LSTM 等。然而现有工作缺乏将多层次结构的编码器和解码器有效结合,且现有方法提取特征的方式单一。例如,DSEN(Dense Semantic Embedding Network)[9]使用了两个双向的 LSTM 进行特征解析,第一个 LSTM 输入为单词,第二个 LSTM 输入为图像特征和属性,然而第二个 LSTM 要对图像特征和属性进行同时解析,这在一定程度上限制了模型的能力,此外该模型的图像特征仅为卷积神经网络最后一个卷积层输出的特征,缺少对特征的再加工与提炼过程。与 DSEN[9]类似,Bi-LSTM[13]的改动在于在编码部分对不同卷积层的特征进行级联作为第二个 LSTM 的输入,然而本质上该模型仍然没有将不同层次的特征进行层次化解码。DHEDN-3(Deep Hierarchical Encoder-Decoder Network)[14]使用了多个 LSTM 进行特征解析,该模型中使用的特征来自卷积神经网络最后两个卷积层的输出,很显然该模型缺少对特征的再次提炼过程,且其第一个 LSTM 仅输入单词,缺少使模型在初始时获得图像整体感知的过程。针对上述的问题,本文提出了一种基于注意力机制的多层次编码和解码的图像描述模型 MLED(Multi-Layer Encoding and Decoding model),如图 1 所示。该模型首先利用 Faster R-CNN 提取对象特征,并使用 Transformer[15]进行高层次特征提取,Transformer 内部的多头注意力机制能够建立起特征之间隐含的复杂联系从而形成高层次特征,并同时能够将特征映射到不同的子空间。借鉴卷积神经网络中的特征金字塔网络(Feature Pyramid Network, FPN)[16]模型,用高层次的特征依次和低层次特征融合得到新的特征,这将有助于改善低层特征的质量。由于单个 LSTM 解析能力有限,本文使用多个 LSTM 的结构,将不同层次的特征分别输入到对应层次的 LSTM 中层次化解码,最终生成单词。

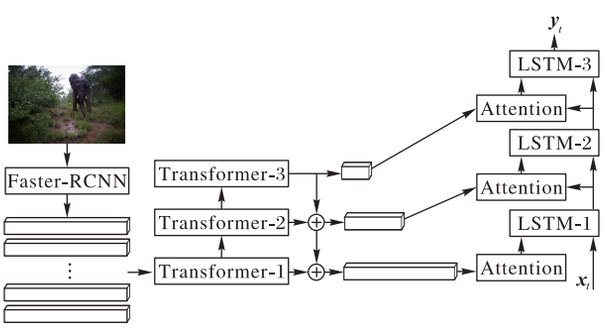


图 1　本文模型框架

Fig. 1　Framework of proposed model

本文的工作主要如下:1)提出了新颖的层次化编码方式,利用 Transformer 以及残差的思想,得到不同层次的特征;2)将不同层次的特征以特征金字塔的方式进行融合,有效提升了特征的准确性;3)提出了层次化解码的 LSTM 结构以及组合方式,对不同层次特征进行有效解码。

1　相关工作

近年来,基于编码器-解码器的框架被广泛地应用于图像描述任务中。本章将主要从注意力机制、编码器结构、解码器结构三方面介绍图像描述相关工作。

Xu 等[7]改进了 Vinyals 等[1]的工作,他们利用软注意力机制,在每个时间步骤用 LSTM 的隐藏层状态对图像特征加权,将加权后的特征和单词共同输入 LSTM 得到结果。由于 LSTM 的隐藏层状态包含上下文信息,因此这种加权在一定程度上能够选择到当前时间步骤的重要信息。Chen 等[8]提出了一种空间和通道的注意力机制,它充分利用 CNN 的空间和通道特征,以便在每个时间步骤进行解码。考虑到生成的单词并不总是取决于视觉信息,Lu 等[12]提出了带有视觉哨兵的自适应注意力模型,该模型可以决定何时关注视觉信息和语言模型。

在编码器方面,Zhang 等[17]提出了一种全卷积神经网络(Fully Convolutional Network, FCN),该网络能够生成细粒度网格化的注意力特征图,从而得到更为精细的图像特征。Yao 等[18]提出了五种不同的模型来结合图像及其属性以探索图像特征和属性特征的结合方式。Wu 等[19]不使用视觉特征,仅使用经过预训练的特定检测器提取图像中的属性,并将其输入 LSTM 中以生成描述。Yu 等[20]提出了一种给定主题的图像描述模型,主题是从训练的分类器中获得。该模型从主题、描述和图像中学习了一种跨模式的嵌入空间,在生成单词时可从嵌入空间中检索信息。

在解码器方面,Anderson 等[21]提出了一种 Bottom-up 模型,该模型包括 Faster R-CNN 和一个双 LSTM 结构。他们首先利用 Faster R-CNN 提取图像的对象特征,然后将其均值化后的特征和对象特征分别输入到第一个语言 LSTM 和第二个注意力 LSTM。相较于单 LSTM,这种双 LSTM 结构有更好的解析能力。Xiao 等[14]建立了一个由三个不同的 LSTM 组成的深层次的编解器-解码器框架,最底层的 LSTM 用于对单词进行编码,中间的 LSTM 和最顶部的 LSTM 用于接收来自 CNN 不同卷积层的特征。Wu 等[22]使用 GridLSTM 在生成单词时能够有选择地包含视觉特征,其中 GridLSTM 由 Temporal LSTM 和 Depth LSTM 组成,Depth LSTM 能够将视觉信息作为潜在的记忆保留,该解码器能够使得每一步动态地接收视觉信息,且无需添加额外的参数。

受上述工作启发,本文提出了一种基于注意力机制的多层次编码和解码的图像描述模型,通过多层次编码得到不同层次的对象特征,并利用多层次 LSTM 分别解析。

2　本文模型

本章将首先使用公式介绍图像描述问题,并给出其目标损失函数,然后分别介绍本文所提出模型的编码器和解码器部分。

2.1　问题描述

给定一张图像 I 和其对应描述单词的序列 $\boldsymbol{Y}=$

$\{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n\}$，生成序列的分布可由式(1)表示：

$$\log p(\boldsymbol{Y}|\boldsymbol{I}) = \sum_{t=1}^{T} \log p(\boldsymbol{y}_t|\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_{t-1}, \boldsymbol{I}) \tag{1}$$

任务的目标是最小化式(2)的损失函数：

$$Loss = -\log p(\boldsymbol{Y}|\boldsymbol{I}) \tag{2}$$

2.2 模型

本文提出了一种基于注意力机制的多层次编码和解码的图像描述模型MLED，该模型框架如图1所示，包括编码层的Faster R-CNN、Transformer、金字塔特征融合以及解码层的层次LSTM。

2.2.1 编码器

Faster R-CNN[8]主要用于目标识别，本文利用Faster R-CNN得到每张图像N个感兴趣区域(Region of Interest, ROI)特征，并将其转化为N个2048维特征$\boldsymbol{A} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_N\}$。

如图2所示，为了得到复杂的高层次的特征，利用Transformer隐含式地对特征进行聚合，以此来得到更高层次的特征。

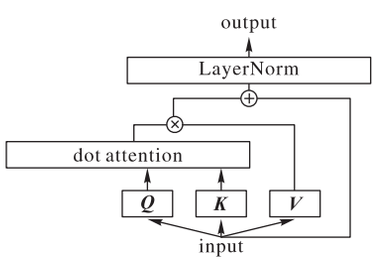


图2 提取高层次特征的模型

Fig. 2 Model of extracting high-level features

Transformer中的多头注意力机制能够让模型关注到特征的不同位置，并将来自较低层的特征映射到不同的子空间。同时多个Transformer串联使用，可以获得不同层次的特征，高层次特征的语义性更强，但包含的信息少，低层次的特征包含的信息内容多，语义较弱。该过程可由式(3)~(5)表示：

$$f_{\text{att}}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Concat}(\boldsymbol{head}_1, \boldsymbol{head}_2, \cdots, \boldsymbol{head}_n) \tag{3}$$

$$\boldsymbol{head}_i = f_{\text{dot-att}}(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i) \tag{4}$$

$$f_{\text{dot-att}}(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i) = \text{Softmax}\left(\frac{\boldsymbol{Q}_i \boldsymbol{K}_i^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V}_i \tag{5}$$

其中：$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$是特征$\boldsymbol{A}$的三个独立线性映射；f_{att}是由$n = 8$个$f_{\text{dot-att}}$函数组成的多头注意力函数；d为常数。对于每对$(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i)$，模型都能将输入映射到不同的子空间，多个Transformer串联又进一步增加了不同子空间的组合，从而进一步增强了模型的表达能力。深度学习网络的问题之一是梯度消失，这会使网络出现饱和现象，因此本文采用残差和归一化的方式，避免在Transformer层出现梯度消失现象，特征$\boldsymbol{A}$经过一层Transformer后的输出记为$\boldsymbol{A}'$，其过程如式(6)所示：

$$\boldsymbol{A}' = \text{LayerNorm}(\boldsymbol{A} + \text{Transformer}(\boldsymbol{A})) \tag{6}$$

其中：LayerNorm为归一化层；Transformer函数为式(3)~(5)所示。本文设置3个不同的Transformer层，根据上述公式，可依次得到维度为512的特征$\boldsymbol{A}'$、维度为256的特征$\boldsymbol{A}''$、维度为128的特征$\boldsymbol{A}'''$。与传统的Transformer只利用最后一层输出不同，本文利用了多层Transformer的输出，并且使用了残差的思想，将Transformer处理后的特征与输入特征相融合，最后使用LayerNorm对特征进行归一化。使用式(6)的目的一方面在于使得梯度能够直接回传到原始输入本身，使得网络能够更深，有更强的表现能力。另一方面归一化的使用使得结果能够落入非线性函数的线性区，缓解梯度消失问题，以此使得模型更加稳定，加速收敛过程。

在获取到不同维度的高层次的特征后，本文使用金字塔型的多尺度特征融合方式对特征进行有效的融合。这种自上而下的融合方式能够使得高层特征得到加强，低层特征得到补充，如图1所示，该融合方式可由式(7)~(9)表示：

$$\boldsymbol{P}''' = \boldsymbol{W}_{33}\boldsymbol{A}''' + b_3 \tag{7}$$

$$\boldsymbol{P}'' = \boldsymbol{W}_{22}\boldsymbol{A}'' + \boldsymbol{W}_{23}\boldsymbol{A}''' + b_2 \tag{8}$$

$$\boldsymbol{P}' = \boldsymbol{W}_{11}\boldsymbol{A}' + \boldsymbol{W}_{12}\boldsymbol{A}'' + b_1 \tag{9}$$

其中：$\boldsymbol{W}$为参数；b为偏置项。经过融合后$\boldsymbol{P}'$、$\boldsymbol{P}''$、$\boldsymbol{P}'''$的维度均为512。低层特征$\boldsymbol{P}'$在融合时仅使用$\boldsymbol{A}'$和$\boldsymbol{A}''$，而不加入$\boldsymbol{A}'''$的信息，一方面是防止与$\boldsymbol{A}'$和$\boldsymbol{A}''$的维度差异过大，导致升维效果不理想，使得模型表现能力受限；另一方面，$\boldsymbol{A}''$中包含了$\boldsymbol{A}'''$部分信息，不添加$\boldsymbol{A}'''$在一定程度上能够防止重复信息的叠加和信息冗余。

2.2.2 解码器

为了对编码器输出的3种不同层次的特征进行有效解码，本文使用层次LSTM的结构处理，该结构包括3层LSTM。其中低层LSTM处理低层次特征，高层LSTM处理高层次特征，这种结构化的层次处理，其目的在于对不同层次的特征进行递进式地解码。此外，使用软注意力机制，在每层特征输入LSTM之前，根据上下文信息对特征进行加权，关注当前步骤应该关注的重要信息。LSTM的隐藏层状态包含丰富的上下文信息，利用隐藏层状态和特征得到特征的权重向量，并对特征做加权，以此来获得经过加权的特征。

第一层LSTM的输入为前一时刻生成的单词的词嵌入向量$\boldsymbol{x}_{t-1}$和本层对应的编码器特征$\boldsymbol{P}'$，其中使用注意力机制对编码器的特征进行加权处理，该过程如式(10)~(13)：

$$\boldsymbol{v}'_{i,t} = \boldsymbol{W}_1^{\mathrm{T}} \tanh(\boldsymbol{W}_{1p}\boldsymbol{p}'_i + \boldsymbol{W}_{1h}\boldsymbol{h}^1_{t-1} + b_1) \tag{10}$$

$$\boldsymbol{\alpha}_t = \text{Softmax}(\boldsymbol{v}'_t) \tag{11}$$

$$\widehat{\boldsymbol{p}'_t} = \sum_{i=1}^{N} \alpha_{i,t}\boldsymbol{p}'_i \tag{12}$$

$$\boldsymbol{h}^1_t = \text{LSTM-1}(\boldsymbol{x}_{t-1}, \widehat{\boldsymbol{p}'_t}, \boldsymbol{h}^1_{t-1}, \boldsymbol{m}^1_{t-1}) \tag{13}$$

其中：$\boldsymbol{W}$为参数；b为偏置项；t表示解码器第t个时间步骤；i表示当前图像N个特征中的第i个；$\boldsymbol{m}^1_{t-1}$是LSTM的细胞状态，在t时刻第一层LSTM的输出为$\boldsymbol{h}^1_t$。第一层LSTM的输入包括单词和编码器特征，其中编码器特征经过软注意力机制加权，重要信息得到加强。第一层LSTM的输出的隐藏层状态包含了上下文特征，能为第二层LSTM提供语义信息。

第二层LSTM的输入为第一层LSTM输出的隐藏层状态$\boldsymbol{h}^1_t$和本层对应的编码器特征$\boldsymbol{P}''$，同样使用软注意力机制对编码器的特征进行处理，该过程如式(14)~(17)：

$$\boldsymbol{v}''_{i,t} = \boldsymbol{W}_2^{\mathrm{T}} \tanh(\boldsymbol{W}_{2p}\boldsymbol{p}''_i + \boldsymbol{W}_{2h}\boldsymbol{h}^2_{t-1} + b_2) \tag{14}$$

$$\boldsymbol{\alpha}_t = \text{Softmax}(\boldsymbol{v}''_t) \tag{15}$$

$$\widehat{\boldsymbol{p}''_t} = \sum_{i=1}^{N} \alpha_{i,t}\boldsymbol{p}''_i \tag{16}$$

$$\boldsymbol{h}^2_t = \text{LSTM-2}([\widehat{\boldsymbol{p}''_t}, \boldsymbol{h}^1_t], \boldsymbol{h}^2_{t-1}, \boldsymbol{m}^2_{t-1}) \tag{17}$$

第三层LSTM的输入为第二层LSTM输出的隐藏层状态$\boldsymbol{h}^2_t$和本层对应的编码器特征$\boldsymbol{P}'''$，该过程同前两层LSTM相同，

最终第三层LSTM输出为式(18)所示:

$$\boldsymbol{h}_t^3 = \text{LSTM-3}([\widehat{\boldsymbol{p}_t'''}, \boldsymbol{h}_t^2], \boldsymbol{h}_{t-1}^3, \boldsymbol{m}_{t-1}^3) \tag{18}$$

在第t时间步骤,生成单词的概率分布函数由多层感知机f_{MLP}和Softmax函数组成,得到的最大值即为生成单词在词典中的索引,具体如式(19)所示:

$$p(\boldsymbol{y}_t) = \text{Softmax}(f_{\text{MLP}}(\boldsymbol{h}_t^3)) \tag{19}$$

3　实验与结果分析

3.1　实验数据及配置

本文主要使用2014年发布的MSCOCO(Microsoft COCO: Common Objects in Context)[23]大型数据集进行模型验证。MSCOCO包括12 387张图像,每张图像有5句描述。本文使用BLEU(BiLingual Evaluation Understudy)[24]、METEOR(Metric for Evaluation of Translation with Explicit ORdering)[25]、ROUGE-L(Recall-Oriented Understudy for Gisting Evaluation with Longest common subsequence)[26]、CIDEr(Consensus-based Image Description Evaluation)[27]四种指标验证模型的有效性。

实验平台为Ubuntu16.04,使用Pytorch深度学习框架,GPU为显存为8 GB的Geforce RTX 2080,CUDA版本为10.0,CPU为英特尔E5-2630,内存为32 GB。

编码器采用的Faster R-CNN使用ResNet101为基础框架,输出的对象特征维度为2 048,限制每幅图像最大对象数量上限为50。对于每条描述,限制其最长单词序列长度为20。选取在数据集中至少出现过5次的单词组成词典并忽略大小写,得到的单词词典大小为10 201,词典使用"BOS"作为单词序列的起始输入标志,"END"作为生成单词序列的结束标志,"PAD"作为补位标志,"UNKNOW"作为不在词典中的单词标志。

训练阶段,本文使用Adam[28]优化器,损失函数为交叉熵,首先使用小的学习率1×10^{-6}进行热身学习10 000次,然后使用1×4^{-4}的学习率训练模型20轮。为了防止在反向传播阶段出现梯度消失或梯度爆炸,本文在编码器部分使用了残差及归一化的方式,如式(6)所示,在解码器部分设置了梯度截断,其范围为[-0.1,0.1]。在测试阶段,使用beam search策略每次选择3个得分最高的候选。

3.2　实验分析

将本文提出的基于注意力机制的多层次编码和解码的图像描述模型MLED与当前主流模型进行对比,实验结果如表1所示。实验对比了NIC(Neural Image Caption)[1]、SA(Soft Attention)[7]、DSEN(Dense Semantic Embedding Network)[9]、Adaptive(Adaptive attention)[12]、Bi-LSTM[13]、DHEDN-3[14]、FCN[17]、Recall(Recall what you see)[22]、基于注意力特征自适应矫正(Attention Feature Adaptive Recalibration, AFAR)模型[29]、HAF(Hierarchical Attention-based Fusion)模型[30]。

由表1可知,本文提出的算法在指标BLEU-1(BLEU with 1-gram)、BLEU-2(BLEU with 2-grams)、BLEU-3(BLEU with 3-grams)、BLEU-4(BLEU with 4-grams)、METEOR、ROUGE-L和CIDEr上得分分别为75.9%、59.7%、46.1%、35.6%、27.3%、56.2%、112.5%,结果优于其他模型,相较于DSEN和Adaptive模型,指标平均提升约2个百分点,相对Recall模型和HAF模型分别提升约3个百分点和1个百分点。以DSEN为例,该模型利用了多层次特征和双LSTM,本文模型相较于DSEN,利用Transformer提取了高层次特征,并对特征做了进一步融合,最后利用层次LSTM对特征层次化解析,因此结果优于DSEN。相较于其他利用注意力机制的模型Attention、FCN以及Adaptive,本文提出的模型在Transformer中使用的多头注意力机制以及对每个LSTM的输入使用软注意力机制,都进一步对特征进行了重要信息提取,因此获得了优异的结果。此外,这种金字塔型的特征以及融合方式在一定程度上体现了全局特征、局部特征以及对象特征,符合人类对图像的整体认知的顺序。

表2为对比实验结果,包括单Transformer和单LSTM的结构模型SLED(Single-Layer Encoder and Decoder)、无特征融合模型MLED-NF(No Fusion)实验结果。

由表2可知,相较于MLED,SLED指标均低于MLED,单编码单解码的结构一定程度上限制了对特征的提取和解析能力,验证了MLED模型多层次编码和解码的有效性。MLED-NF无特征融合过程,而MLED的自上而下的融合过程有一定程度的特征加强作用,最顶层的特征最接近单词输出,反向传播后的梯度能多次影响下层特征,使得模型能达到更优的解。

图3为随机挑选的样例,包括图片、标注的5句描述以及模型生成的对应描述。可以看出,本文所提出的MLED模型能够很好地捕捉图像内容和细节,例如图3中样例1的"large"和"lush"、样例2的"metal fence"以及样例3的"covered slope",这些生成的语句能够涵盖图像的内容和细节,并且模型生成的描述基本符合语法和语义。

表1　不同图像描述算法模型指标对比　　单位:%

Tab. 1　Comparison of different image captioning algorithm models on indicators　　unit:%

模型	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
NIC	66.6	46.1	32.9	24.6	—	—	—
SA	71.8	50.4	35.7	25.0	23.0	—	—
AFAR	69.4	52.3	38.6	28.5	23.3	—	83.6
FCN	71.2	51.4	36.8	26.5	24.7	—	88.2
DSEN	75.3	59.6	45.7	34.7	26.2	55.6	106.8
Adaptive	74.2	58.0	43.9	32.2	26.6	—	108.5
Bi-LSTM	74.5	59.3	45.5	**36.3**	**28.7**	55.8	99.7
DHEDN-3	75.5	59.5	**46.1**	35.6	27.1	56.0	109.7
Recall	75.8	—	—	33.1	24.7	—	103.7
HAF	**75.9**	59.5	45.4	34.4	26.8	—	109.0
MLED	**75.9**	**59.7**	**46.1**	**35.6**	**27.3**	**56.2**	**112.5**

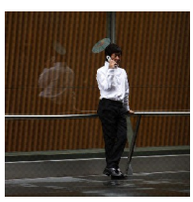
表 2 消融实验结果

Tab. 2 Ablation experimental results

模型	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
SLED	74.7	57.5	43.6	34.0	26.7	55.1	107.5
MLED-NF	75.4	58.9	45.5	35.2	27.1	55.8	110.0
MLED	**75.9**	**59.7**	**46.1**	**35.6**	**27.3**	**56.2**	**112.5**

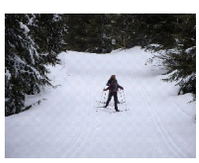
1.An elephant walking through a clearing in a jungle.
2.A elephant standing in the clearing of a wooded area.
3.An African elephant walks through a patch of swampy brush.
4.A large elephant standing on a muddy grass covered path.
5.A tusked elephant is walking among the greenery.
生成描述：a large elephant walking through a lush green forest.

(a) 样例1



1.A man talking on a phone standing in front of a building.
2.The man is standing on the sidewalk, talking on his phone with a small umbrella.
3.A man leaning on a railing talking on a phone with an attached tiny umbrella.
4.A man in white shirt on cellphone with a tiny umbrella extension.
5.A man is holding a cellular phone against the rail.
生成描述：a man talking on a cell phone next to a metal fence.

(b) 样例2



1.A couple of people riding a pair of skis down a snow coveredslope.
2.a couple of people skiing down a snowy slope.
3.there are people that are on the snow slope skiing.
4.two snow skiers are coming down a hill.
5.The cross countryskiers are enjoying their run.
生成描述：a coupleof people riding skis down a snow covered slope.

(c) 样例3

图 3 生成描述样例

Fig. 3 Examples of generated description

4 结语

本文设计了一种基于注意力机制的多层次编码和解码的图像描述模型，通过 Faster R-CNN 得到对象特征，然后使用 Transformer 进行高层次特征的提取并使用金字塔型的融合方式对特征进行融合，最后使用层次 LSTM 对不同层次的特征分别解析。该模型在不同层次充分利用了对象特征，并且使用金字塔型的融合方式对特征进行了有效融合。模型对不同层次特征的层次化解码促进了网络的解码能力。实验结果和分析表明，本文提出的模型在各项指标均有优异的表现，所生成的描述也符合语法和语义，并且优于其他模型。未来将结合卷积神经网络的不同层次特征、对象的不同层次特征、特征融合方式以及 Transformer 方式的层次化解码等方面进一步研究。

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 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2015: 3156-3164.

[2] 李文惠，曾上游，王金金. 基于改进注意力机制的图像描述生成算法[J]. 计算机应用，2021，41(5)：1262-1267.（LI W H, ZENG S Y, WANG J J. Image description generation algorithm based on improved attention mechanism [J]. Journal of Computer Applications, 2021, 41(5): 1262-1267.）

[3] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2015-04-10) [2021-01-20]. https://arxiv.org/pdf/1409.1556.pdf.

[4] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 770-778.

[5] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[6] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate [EB/OL]. (2016-05-19) [2019-09-01]. https://arxiv.org/pdf/1409.0473.pdf.

[7] XU K, BA J, KIROS R, et al. Show, attend and tell: neural image caption generation with visual attention [C]// Proceedings of the 32nd International Conference on Machine Learning. New York: JMLR.org, 2015: 2048-2057.

[8] CHEN L, ZHANG H W, XIAO J, et al. SCA-CNN: spatial and channel- wise attention in convolutional networks for image captioning [C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 6298-6306.

[9] XIAO X Y, WANG L F, DING K, et al. Dense semantic embedding network for image captioning [J]. Pattern Recognition, 2019, 90: 285-296.

[10] ZHANG M X, YANG Y, ZHANG H W, et al. More is better: precise and detailed image captioning using online positive recall and missing concepts mining [J]. IEEE Transactions on Image Processing, 2019, 28(1): 32-44.

[11] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[12] LU J S, XIONG C M, PARIKH D, et al. Knowing when to look: adaptive attention via a visual sentinel for image captioning [C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 3242-3250.

[13] 赵小虎，李晓. 基于多特征提取的图像语义描述算法[J]. 计算

机应用, 2021, 41(6): 1640-1646. (ZHAO X H, LI X. Image captioning algorithm based on multi-feature extraction[J]. Journal of Computer Applications, 2021, 41(6): 1640-1646.)

[14] XIAO X Y, WANG L F, DING K, et al. Deep hierarchical encoder-decoder network for image captioning [J]. IEEE Transactions on Multimedia, 2019, 21(11): 2942-2956.

[15] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc., 2017: 6000-6010.

[16] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection [C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 936-944.

[17] ZHANG Z J, WU Q, WANG Y, et al. High-quality image captioning with fine-grained and semantic-guided visual attention [J]. IEEE Transactions on Multimedia, 2019, 21(7): 1681-1693.

[18] YAO T, PAN Y W, LI Y H, et al. Boosting image captioning with attributes [C]// Proceedings of the 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 4904-4912.

[19] WU Q, SHEN C H, WANG P, et al. Image captioning and visual question answering based on attributes and external knowledge [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(6): 1367-1381.

[20] YU N G, HU X L, SONG B H, et al. Topic-oriented image captioning based on order-embedding[J]. IEEE Transactions on Image Processing, 2019, 28(6): 2743-2754.

[21] ANDERSON P, HE X D, BUEHLER C, et al. Bottom-up and top-down attention for image captioning and visual question answering [C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 6077-6086.

[22] WU L X, XU M, WANG J Q, et al. Recall what you see continually using GridLSTM in image captioning [J]. IEEE Transactions on Multimedia, 2020, 22(3): 808-818.

[23] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context [C]// Proceedings of the 2014 European Conference on Computer Vision, LNCS 8693. Cham: Springer, 2014: 740-755.

[24] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation [C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2002: 311-318.

[25] 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: Association for Computational Linguistics, 2005: 65-72.

[26] LIN C Y. ROUGE: a package for automatic evaluation of summaries [C]// Proceedings of the ACL 2004 Workshop on Text Summarization. Stroudsburg, PA: Association for Computational Linguistics, 2004: 74-81.

[27] VEDANTAM R, ZITNICK C L, PARIKH D. CIDEr: consensus-based image description evaluation [C]// Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2015: 4566-4575.

[28] KINGMA D P, BA J M. ADAM: a method for stochastic optimization [EB/OL]. (2017-01-30) [2020-04-22]. https://arxiv.org/pdf/1412.6980.pdf.

[29] 韦人予,蒙祖强. 基于注意力特征自适应校正的图像描述模型[J]. 计算机应用, 2020, 40(S1): 45-50. (WEI R Y, MENG Z Q. Image caption model based on attention feature adaptive recalibration [J]. Journal of Computer Applications, 2020, 40(S1): 45-50.)

[30] WU C L, YUAN S Z, CAO H W, et al. Hierarchical attention-based fusion for image caption with multi-grained rewards [J]. IEEE Access, 2020, 8: 57943-57951.

This work is partially supported by the National Natural Science Foundation of China (61402174).

LI Kangkang, born in 1995, M. S. candidate. His research interests include computer vision, image caption.

ZHANG Jing, born in 1978, Ph. D., associate professor. Her research interests include deep learning, computer vision, image retrieval, image caption, visual question answering.