Journal of Computer Applications
计算机应用,2020,40(4):985－989
ISSN 1001-9081
CODEN JYIIDU
2020-04-10
http://www.joca.cn

文章编号:1001-9081(2020)04-0985-05
DOI:10.11772/j.issn.1001-9081.2019091569

基于多空间混合注意力的图像描述生成方法

林贤早*,刘　俊,田　胜,徐小康,姜　涛
(杭州电子科技大学 通信信息传输与融合技术国防重点学科实验室,杭州 310018)
(*通信作者电子邮箱lilcore_lxz@163.com)

摘　要:针对近海船舶监测系统中自动化情报生成的空缺,为了构建智能化船舶监测系统,提出基于多空间混合注意力的图像描述生成方法,对近海船舶图像进行描述。图像描述生成方法就是让计算机通过符合语言学的文字描述出图像中的内容。首先使用图像的感兴趣区域的编码特征预训练出多空间混合注意力模型,然后加入策略梯度改造损失函数对预训练好的解码模型继续进行微调,得到最终的模型。在MSCOCO(MicroSoft Common Objects in COntext)图像描述数据集上的实验结果表明,所提模型较以往的注意力模型提升了图像描述生成的评价指标,比如CIDEr分数。使用该模型在自建船舶描述数据集中能够自动描述出船舶图像的主要内容,说明所提方法能为自动化情报生成提供数据支持。

关键词:图像描述;深度学习;注意力机制;情报生成;多空间混合注意力
中图分类号:TP389.1　　**文献标志码**:A

Image description generation method based on multi-spatial mixed attention

LIN Xianzao*, LIU Jun, TIAN Sheng, XU Xiaokang, JIANG Tao
(*Fundamental Science on Communication Information Transmission and Fusion Technology Laboratory*, *Hangzhou Dianzi University*, *Hangzhou Zhejiang* 310018, *China*)

Abstract: Concerning the vacancy of automatic information generation in offshore ship monitoring system, and aiming to build an intelligent ship monitoring system, an image description generation method based on multi-spatial mixed attention was proposed to describe the offshore ship images. The image description generation task is designed to let the computer describe the content of the image with words satisfying linguistics. Firstly, the multi-spatial mixed attention model was trained by the encoding features of the region of interest on the image, then the pretrained decoding model was fine-tuned by reconstructing the loss function with gradient policy, and the final model was obtained. Experimental results on MSCOCO (MicroSoft Common Objects in COntext) image description dataset show that the proposed model is better than the previous attention model on the evaluation index of image description generation, such as CIDEr score. The main content of ship image can be automatically described by the model on the self-constructed ship description dataset, demonstrating that the method can provide the data support for automatic information generation.

Key words: image description; deep learning; attention mechanism; information generation; multi-spatial mixed attention

0　引言

随着近些年人工智能的高速发展,近海地区也在跟进构建智能化船舶监测系统。而自动化的情报生成就是其中至关重要的一环,也是极为困难的一环。船舶监测系统中关于情报的生成不仅需要船舶类别、位置等信息,还需要描述船舶图像内容的语义信息作为数据支撑。得益于深度学习在计算机视觉中的广泛应用,计算机通过训练可以自动生成对图像的文本描述,同样可以对船舶图像的运动状态和四周场景进行描述。

视觉作为人类的主要感官,发挥着巨大的作用。人们通过在短时间快速地浏览图片就能在脑海中生成符合语言学且与内容相符合的图像描述。由此可知,图像描述生成领域关联两个基础问题,也就是视觉理解和语言处理。换而言之,解决图像描述生成问题需要连接计算机视觉和自然语言处理两个社区,这项任务不仅需要高度理解图像语义内容,还需要用人类化的语言表达出该信息。从以往的研究得知,确定图片中的物体的存在、属性还有之间的关系本身就不是一个轻松的工作,进一步用符合语法的语句去描述此类信息则更加提升了这项工作的难度。

深度学习在计算机视觉和自然语言处理等人工智能领域表现优越,可知深度神经网络能同时为视觉模型和语言模

收稿日期:2019-09-16;**修回日期**:2019-10-28;**录用日期**:2019-10-30。
基金项目:国家自然科学基金资助项目(61673146);国家自然科学基金重大仪器专项(61427808);浙江省重点研发计划项目(2019C05005)。
作者简介:林贤早(1994—),男,浙江温州人,硕士研究生,主要研究方向:自然语言处理、计算机视觉、强化学习;　刘俊(1971—),男,贵州安顺人,教授,博士,主要研究方向:模式识别、智能系统、目标检测、信息融合;　田胜(1994—),男,安徽铜陵人,硕士研究生,主要研究方向:目标检测、目标跟踪;　徐小康(1996—),男,安徽滁州人,硕士研究生,主要研究方向:深度学习、目标检测;　姜涛(1995—),男,江苏常州人,硕士研究生,主要研究方向:目标检测、信息融合。

型[1]提供支撑。受到神经机器翻译中编解码框架的启发，图像描述生成任务也可以分解成两个步骤：对图像内容和语义进行编码，使用语言模型对该特征进行解码。卷积神经网络（Convolutional Neural Network，CNN）[2]现如今已成为目标检测和识别的主流方法，而循环神经网络（Recurrent Neural Network，RNN）在自然语言处理也拥有着卓越表现，两者的有机结合刚好为图像描述生成提供了有效的解决方案。

1 相关工作

早期在图像描述生成方面的工作主要集中在基于检索的方法和基于模板的方法。这些方法要么通过关键词直接套用现有的描述文字[3]，要么依靠严格编码的语言结构完成文字描述[4]，因此早期工作中这两种方法产生的图像描述在很大程度上十分晦涩而又低效。现如今，许多基于循环神经网络的深度学习模型已经广泛应用于图像描述生成。而这些使用深度学习的方法大多数采用编码/解码框架。这个框架的流程是先通过预训练好的卷积神经网络将图像编码成能够表征图像内容的特征，然后结合部分完整描述文字提供的语义输入到循环神经网络中将该特征解码成句子。这是Vinyals等[5]率先提出的，该模型是受到最近神经机器翻译[6]在序列生成中的成功应用所启发，与神经机器翻译的区别就是图像描述生成的输入不是句子而是卷积网络得到的特征，特征进行解码时采用了长短时记忆（Long Short-Term Memory，LSTM）单元。LSTM作为RNN的变种，由于其门控单元的设计，能够很大程度改善RNN在长时间序列上的梯度弥散，因此后续的模型大多都是用LSTM或其变种来解决句子生成这类序列结构问题。后续的研究则分别在编码和解码上对其进行改良，近来备受关注的注意力机制就广泛应用于该任务。Xu等[7]使用带有空间信息的卷积图像特征作为输入，在二维空间上使用注意力对位置进行选择，他采取了两种注意力方式，分别为只选取固定数量位置的“硬”注意力和给所有的空间位置分配不同权重的“软”注意力。这种空间注意力能够有效地对特征再编码，从而提高了语言模型生成句子的正确性。You等[8]将注意力转向语义集合中，基于语义特征集合解码生成图像描述。Chen等[9]甚至还对不同的特征通道使用了注意力，将注意力延伸到三维空间。

图像描述生成方法在解码阶段一般使用交叉熵函数进行训练，但是测试阶段评价使用的是不可微的自然语言评价指标，比如BLEU（Bilingual Evaluation Understudy）[10]、ROUGE（Recall-Oriented Understudy for Gisting Evaluation）[11]、CIDEr（Consensus-based Image Description Evaluation）[12]等，因此使用交叉熵函数无法直接优化评价指标，而只能拟合模型去生成与数据集相近的语言描述，容易在解码阶段过拟合，无法对语言表达进行有效的学习。不止于此，测试阶段的图像描述生成是通过已训练好的模型生成的单词结合图像特征，迭代地预测后续的单词，所以这种预测方式容易对错误进行积累，这种现象叫作exposure bias，Rennie等[13]1提出加入强化学习策略可以弥补交叉熵损失函数无法优化指标的缺陷，该策略可以在训练中通过采样的方式计算奖励期望的梯度，进而更新模型权重，使得评价指标作为直接优化的目标。

图像描述生成还受益于图像描述生成数据集不断扩大，比如原先的Flicker 8K、Flicker 30K到现在MSCOCO（MicroSoft Common Objects in COntext）caption提供十几万张图片和对应的文字描述，使得深度神经网络的训练得到了有效的数据集支撑。为了将该方法应用于船舶监测中，本文自建船舶描述数据集对船舶的运动状态和四周场景进行标注。

2 本文算法

本文提出的基于多空间混合注意力的图像描述方法，使用预训练好的检测网络提取感兴趣区域的特征编码，在解码阶段对该特征施加多空间注意力和视觉选择，引入强化学习的策略梯度对优化目标进行重塑，从而使得训练和测试阶段的解码统一，直接针对评价指标进行优化。整体框架如图1所示，这种模式本质上属于端到端的设计，但是由于实际训练中无法同时优化卷积神经网络和LSTM，图像和文字虽然能表征同样的事件或者事物，但是在表达形式上存在着鸿沟。本文将编码解码分成两个步骤分开训练，在得到丰富的语义特征之后，将该特征作为解码模型的输入。如图1所示，为了得到图像的感兴趣区域特征，算法总体框架中的卷积编码器选用的是目标检测网络。具体采用的感兴趣区域特征提取方案是以ResNet-101[14]为卷积骨干的Faster-RCNN[15]。为了感兴趣区域特征能够表征图像中的相关属性，在损失函数中添加属性分类交叉熵损失。训练数据集使用的是带有属性、坐标、类别标签的Visual Genome数据集。编码采用的具体卷积结构如图2所示。

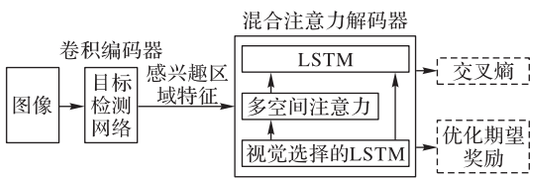

图1 算法总体框架

Fig. 1 Overall framework of the proposed algorithm

Layer name	parameters
conv1 pool1	7×7×64, stride 2, 3×3 max pool, stride 2
conv2.x (totally 3 layers)	1×1×64, 3×3×64, 1×1×256
conv3.x (totally 4 layers)	1×1×128, 3×3×128, 1×1×512
conv4.x (totally 23 layers)	1×1×256, 3×3×256, 1×1×1024
conv5.x (totally 23 layers)	1×1×512, 3×3×512, 1×1×2048

图2 卷积网络结构

Fig. 2 Convolution network structure

沿用Faster-RCNN的框架，网络的改动部分如下：首先将区域候选网络（Region Proposal Network，RPN）结构接在分类网络的第4个卷积模块之后，得到候选区域；然后将候选区域与第4个卷积模块的特征结合，得到感兴趣区域特征；最后利用第5个卷积模块接的图像特征分别对401个属性进行分类，对于1 601个目标种类进行目标检测。该目标检测网络的设计方式是为了与ResNet-101分类网络结构保持一致，提高网络迁移的稳定性，使得网络可训练。

除此之外，当引入强化学习目标作为训练的优化函数之后，增加了模型的不稳定性，通过实验可知，直接优化平均期望奖励这一目标，会使得模型无法训练。而交叉熵损失函数往往能构成凸函数，使得模型易于收敛，所以本文先通过交叉熵模型得到性能较好的解码模型，再使用策略梯度优化模型时就可以稳定地提高评价指标。

2.1　多空间注意力

在人类的视觉系统中,注意力信号大致可以划分为两种:一种是自顶向下的注意力,这类信号受当前的任务的驱动,由人的主动意识所控制;另外一种是外界新奇或者显著的激励因子组成的自底向上的信号,一般是被动地接收。这两种注意力信号都与视觉元素的内容相关联。

由于卷积操作本身的特性,特征图的每一通道都由一组卷积核对上一层特征块卷积后得到,可将其对应为自底向上的局部空间特征提取器,因此特征块的通道可以认为是图像的不同语义部分。换而言之,卷积核能够在局部感受野中融合空间和通道信息。既然卷积的作用是对图像进行特征编码,那么注意力编码的设计可以认为是对不同位置、不同通道的特征进行解耦。添加注意力后得到的特征图,可以看作是对空间、通道信息的重新校准,可以对后续的解码过程产生积极的影响。本文在解码阶段采用多空间注意力,如图 3 所示。

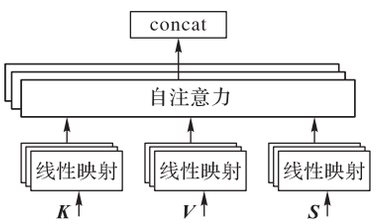

图3　多空间注意力

Fig. 3　Multi-spatial attention

这种注意力也同样属于自注意力。此自注意力本质是对特征进行重新编码。回顾之前的框架,本文通过卷积神经网络得到了图像的特征向量表达,这一环节就是结合解码输出构成的上下文语境引导特征的重新编码。具体的操作为:

$$\boldsymbol{att} = \mathrm{softmax}(\alpha \boldsymbol{h}^{\mathrm{T}} \boldsymbol{V}) \boldsymbol{V} \tag{1}$$

其中:$\boldsymbol{h}$ 为循环神经网络的隐层状态;$\boldsymbol{V}$ 为感兴趣区域特征。与一般注意力不同的是,本文将这种注意力扩展到了多个空间中。假设隐层状态长度为 k 维,每个空间位置的图像特征也为 k 维,先将其扩展成 N 个子空间后,通过式(1)计算子空间注意力的权重,然后将其重新拼接成最后的注意力特征。

2.2　视觉选择

因为评价标准依据的是生成句子的内容和流畅性,因此仅仅关注图像的视觉部分还不够,还需要考虑将图像内容串联起来的一些非视觉词语,所以本文在原有的 LSTM 中加入视觉选择门控机制。带有视觉选择的解码模型可以自动决定什么时候关注视觉信号,什么时候依赖语言模型。当依赖视觉信号时,模型同样会决定对视觉区域的选择作出判断。一般的 LSTM 模型如下:

$$\boldsymbol{h}_t = \mathrm{LSTM}(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}, \boldsymbol{m}_{t-1}) \tag{2}$$

其中:$\boldsymbol{x}_t$ 是输入向量;$\boldsymbol{m}_{t-1}$ 是 $t-1$ 时刻的记忆细胞向量。通过在该向量上进行扩展,得到可供非视觉词产生的信息,形成视觉选择门控机制。

$$\boldsymbol{g}_t = \sigma(\boldsymbol{W}_x \boldsymbol{x}_t + \boldsymbol{W}_h \boldsymbol{h}_{t-1}) \tag{3}$$

$$\boldsymbol{s}_t = \boldsymbol{g}_t \odot \tanh(\boldsymbol{m}_t) \tag{4}$$

其中:$\boldsymbol{W}_x$ 和 $\boldsymbol{W}_h$ 是需要被学习的权重;$\boldsymbol{x}_t$ 是 LSTM 在 t 时刻的输入;$\boldsymbol{g}_t$ 向量对记忆细胞施加影响;$\boldsymbol{m}_t$ 包含了时刻 t 及其之前的语义信息;"⊙"是点乘操作。

基于非视觉词的信息 $\boldsymbol{s}_t$ 和注意力的特征 $\boldsymbol{att}_t$ 来重新组合得到自适应语义向量 $\boldsymbol{c}'$。

$$\boldsymbol{c}' = \mu_t \boldsymbol{s}_t + (1 - \mu_t) \boldsymbol{att}_t \tag{5}$$

其中 μ_t 是一个标量,它决定了对视觉信息的选择,它的取值是先将 $\boldsymbol{s}_t$ 和 $\boldsymbol{h}_t$ 映射到嵌入空间,将其进行组合后再投射到一维空间得到标量值,具体实现如下:

$$\mu_t = \lambda \boldsymbol{w}_h^{\mathrm{T}} \tanh(\boldsymbol{W}_s \boldsymbol{s}_t + \boldsymbol{W}_g \boldsymbol{h}_t) \tag{6}$$

视觉选择与多空间注意力构成了多空间混合注意力,既能关注视觉方面的信息,也能对图像中的非视觉信息进行选择。多空间混合注意力同时还得益于编码特征中将图像之间的属性关系融合到优化目标中,使得感兴趣区域特征融合进了图像的属性信息。

2.3　策略梯度优化指标

如图 4 所示,循环神经网络模型可以看作一个智能体与外部环境(单词和图像特征)进行交流。这个网络模型的参数 θ 定义了策略 π。策略 π 会产生一种动作,对应的就是句子的预测。在每个动作之后,这个智能体即 LSTM 会更新它的状态。这个状态指的是 LSTM 中的记忆细胞状态和隐层状态。这个过程迭代生成句子描述,直到生成句子结束标识符。智能体通过观测环境可以获得回报,动作的选择就是通过最小化这个回报的负期望得到的。回报的产生就是依赖常用的评价指标,比如 CIDEr-D,计算生成句子的得分值,本文将这种回报记作 r。

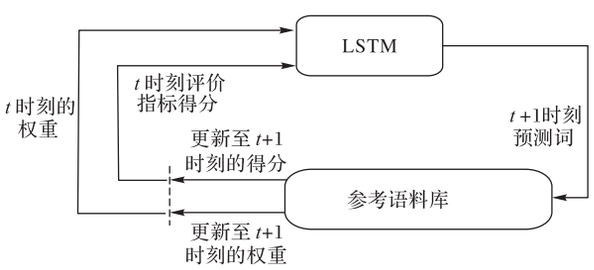

图4　强化学习优化过程

Fig. 4　Optimization process of reinforcement learning

目标函数就从原来的交叉熵函数重新塑造成回报的期望:

$$J(\theta) = -\mathrm{E}_{\boldsymbol{w}^s \sim \pi_\theta}[r(\boldsymbol{w}^s)] \approx -\frac{1}{N} \sum_i \sum_t r(\boldsymbol{s}_{i,t}, \boldsymbol{a}_{i,t}) \tag{7}$$

由于无法得知回报的分布,一般常用蒙特卡洛方法经验平均来作为模型期望的无偏估计。此方法主要的限制是在强化学习下使用小批量样本会使需要优化的回报这一随机变量产生高的方差,从而使得训练过程十分不稳定,难以收敛,并且无法选择学习率。除了适当地增加批尺寸外,为了稳定性的需要还可以加入合适的偏差修正 baseline。

$$\nabla_\theta J(\theta) \approx -\frac{1}{N} \sum_{i=1}^{N} (r(\boldsymbol{w}^s) - b) \nabla_\theta \ln(\pi_\theta(\boldsymbol{w}^s)) \tag{8}$$

baseline 的设置为当前模型在测试阶段得到回报。那么式(8)可改写为:

$$\nabla_\theta J(\theta) \approx -\frac{1}{N} \sum_{i=1}^{N} (r(\boldsymbol{w}^s) - r(\hat{\boldsymbol{w}})) \nabla_\theta \ln(\pi_\theta(\boldsymbol{w}^s)) \tag{9}$$

因为 baseline 是一个常数,所以并不影响梯度的大小。除此之外本文还使用限定采样方式为多项式分布来加速训练过程。

3　实验与结果分析

3.1　评价指标

针对图像描述生成任务,本文主要使用 CIDEr-D 进行评

分，其他评价指标有机器翻译工作中基于精确度的BLEU和自动摘要工作中基于召回率的ROUGE。以下是CIDEr的计算公式：

$$CIDEr_n(c_i,S_i)=\frac{1}{m}\sum_j\frac{g^n(c_i)\cdot g^n(s_{ij})}{\|g^n(c_i)\|\|g^n(s_{ij})\|} \tag{10}$$

$$CIDEr(c_i,S_i)=\sum_{n=1}^{N}w_n CIDEr_n(c_i,S_i) \tag{11}$$

其中：c_i是生成候选句子；s_{ij}是参考的句子；$g^n(c_i)$是一个向量，它的长度为候选句子和真实句子中n元词组的个数之和，每个元素是计算n元语法在候选生成句子中的TF-IDF（Term Frequency-Inverse Document Frequency）；$\|\cdot\|$是取模操作。同理$g^n(s_{ij})$即是将生成候选句子替换为参考句子后进行计算。w_n一般设为$1/N$（N一般设为4）。为了评价的公平性，微软官方重新对CIDEr进行修改，加上了句子长度的差异的高斯惩罚和对大于参考句子的TF-IDF元素进行截断，记为CIDEr-D，重写为：

$$CIDEr_n\text{-}D(c_i,S_i)=10\times\frac{1}{m}\sum_j e^{\frac{-(l(c_i)-l(s_{ij}))^2}{2\sigma^2}}\times\frac{\min(g^n(c_i),g^n(s_{ij}))\cdot g^n(s_{ij})}{\|g^n(c_i)\|\|g^n(s_{ij})\|} \tag{12}$$

$$CIDEr\text{-}D(c_i,S_i)=\sum_{n=1}^{N}w_n CIDEr_n\text{-}D(c_i,S_i) \tag{13}$$

一般使用$\sigma=6$，乘以10是为了让这个分数与其他的评价的指标相近。

3.2 数据集和参数

本文选用在MSCOCO caption数据集上验证算法的有效性。MSCOCO是微软公开的图像描述数据集，包含着82 783张训练集、40 504张验证集和40 775张测试集。相对于其他小规模的图像描述生成数据集，COCO caption数据集更有挑战力，也更加具有公信力，其中一张图片对应5句描述，由json格式提供。本文采取的验证模型优劣的方式分为两个步骤：先通过训练集和验证集在线下调节模型的参数，然后提交测试集的结果到服务器上获取对应指标的分数。最终的解码模型获取分为两轮，区别在于第一轮是对交叉熵损失函数进行优化，第二轮是通过策略梯度对模型进行调节。第一轮设置为学习率0.000 1，选用Adam优化器降低交叉熵损失，收敛至平稳后，再降低学习率，直至交叉熵损失无法进一步优化，最大迭代轮数为30。得到较稳定的交叉熵解码模型后，再使用策略梯度替换交叉熵损失函数，采取相同的超参数进行优化，两轮训练的总迭代周期为70。沿用Karpathy等[16]的数据集设置，分别使用5 000张图片用于线下的验证和测试。表1列出训练时候的超参数设置。词嵌入向量设为1 024，LSTM的隐藏层向量大小设置为1 024。为了防止过拟合对加入dropout，设为0.5。

3.3 结果与分析

为了使实验结果有说服力，本文将COCO测试集在本地得出的图像描述提交到后台验证算法设计的有效性，并与近些年带有注意力机制的算法进行比较。主要实验内容如表2所示。

通过表2可以得知，相比在解码阶段单纯使用LSTM，现今的方法都会加上注意力机制，注意力机制能够在解码阶段对于卷积得到的整体特征再次重新编码，使得特征得以映射到能与语言空间容易转换的嵌入空间，提升特征的表达能力。而本文使用的混合注意力，则首先将特征映射到不同的空间中，扩展注意力的表达，再使用视觉选择机制分配视觉信息与语言信息的权重，不仅提升了特征的表征能力，还能联系生成单词的语义，从而获得较好的指标结果。

在线下验证实验中，本文叠加多空间注意力和视觉选择模块进行训练，融合成本文所提出的混合注意力进行优化模型。从表3的结果来看，在没有使用策略梯度微调模型的情况下，还是能够使结果达到比较好的效果。当加上策略梯度优化时能够极大地提升混合注意力模型解释特征的能力。这里的强化学习算是一种优化手段，本质上也是在复杂模型提供的参数空间中寻找最优的参数优化指标，最终还是混合注意力起到了作用，使得该模型的图像描述能力提升，获得了较高的评价分数。同时实验统计了编解码模型在前向的耗时，编码前向平均每帧平均耗时200 ms，解码前向每帧平均耗时40 ms。

除了在权威的COCO数据集上进行模型验证实验之外，本文还自建船舶描述数据集，将船舶在海上航行的情况进行描述，为情报生成打下基础。如图5所示，给出带有船舶的图片，可以自动输出语句来描述出其船舶明显的主体颜色及其在海上航行或岸边停靠等内容，并且语句的表述能够合乎语法规则。

图5 自动生成船舶图像描述

Fig. 5 Automatic generation of ship image descriptions

表1 超参数设置

Tab. 1 Hyperparameter setting

名称	值
初始学习率	0.000 1
批量大小	50
预设置最大训练周期	30
总训练周期	70
LSTM的隐藏层单元	1 024
词嵌入向量长度	1 024
最大序列长度	16
优化器	Adam

表2 不同注意力机制的算法比较

Tab. 2 Comparison of algorithms with different attention mechanisms

算法	BLEU4	ROUGE	CIDEr-D
Adaptive[17]	0.336	0.550	1.042
SCST[13]7	0.352	0.563	1.147
LSTM-A[18]	0.356	0.564	1.160
Up-Down[19]	0.369	0.571	1.179
本文算法	0.373	0.579	1.225

表3　叠加不同模块的效果

Tab. 3　Effect of adding different modules

改进模块	BLEU4	ROUGE	CIDEr-D
使用混合注意力	0.356	0.565	1.155
加入策略梯度	0.378	0.581	1.269

4　结语

本文深入研究了图像描述生成方案,提出了基于多空间混合注意力的图像描述生成模型,并将该方法应用于船舶图像上,以填补近海船舶监测系统的情报生成的缺失。但是该模型还是有局限性,比如句子的长度是被限制在16个单词,所以对于语义内容多的图片可能无法进行有效的描述。值得一提的优化方法有增大语料库来提高生成句子的丰富性,这种方式是最直接有效的提升指标,但是工作量较大。

参考文献(References)

[1] 奚雪峰,周国栋. 面向自然语言处理的深度学习研究[J]. 自动化学报, 2016, 42(10): 1445-1465. (XI X F, ZHOU G D. A survey on deep learning for natural language processing [J]. Acta Automatica Sinica, 2016, 42(10): 1445-1465.)

[2] 周飞燕,金林鹏,董军. 卷积神经网络研究综述[J]. 计算机学报, 2017, 40(6): 1229-1251. (ZHOU F Y, JIN L P, DONG J. Review of convolutional neural network [J]. Chinese Journal of Computers, 2017, 40(6): 1229-1251.)

[3] FARHADI A, HEJRATI M, SADEGHI M A, et al. Every picture tells a story: generating sentences from images [C]// Proceedings of the 2010 European Conference on Computer Vision, LNCS 6314. Berlin: Springer, 2010: 15-29.

[4] YANG Y, TEO C, DAUMÉH III, et al. Corpus-guided sentence generation of natural images [C]// Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics: Association for Computational Linguistics, 2011: 444-454.

[5] 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.

[6] KALCHBRENNER N, BLUNSOM P. Recurrent continuous translation models [C]// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2013: 1700-1709.

[7] XU K, BA J L, 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] YOU Q, JIN H, WANG Z, et al. Image captioning with semantic attention [C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 4651-4659.

[9] CHEN L, ZHANG H, 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.

[10] 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: Association for Computational Linguistics, 2002: 311-318.

[11] LIN C Y. Rouge: a package for automatic evaluation of summaries [M]// Text Summarization Branches Out. Stroudsburg, PA: Association for Computational Linguistics, 2004: 74-81.

[12] VEDANTAM R, ZITNICK C L, PARIK 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.

[13] RENNIE S J, MARCHERET E, MROUEH Y, et al. Self-critical sequence training for image captioning [C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 1179-1195.

[14] HE K, ZHANG X, REN S, 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.

[15] REN S, HE K, 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.

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

[17] LU J, XIONG C, 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.

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

[19] ANDERSON P, HE X, 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.

This work is partially supported by the National Natural Science Foundation of China (61673146), the National Natural Science Foundation of China Major Instrument Project (61427808), the Key Research and Development Project of Zhejiang Province (2019C05005).

LIN Xianzao, born in 1994, M. S. candidate. His research interests include natural language processing, computer vision, reinforcement learning.

LIU Jun, born in 1971, Ph. D., professor. His research interests include pattern recognition, intelligent system, object detection, information fusion.

TIAN Sheng, born in 1994, M. S. candidate. His research interests include object detection, target tracking.

XU Xiaokang, born in 1996, M. S., candidate. His research interests include deep learning, object detection.

JIANG Tao, born in 1995, M. S. candidate. His research interests include object detection, information fusion.