软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2018,29(Suppl.(2)):30−43

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

一种基于生成式对抗网络的图像描述方法*

薛子育[1], 郭沛宇[1], 祝晓斌[2], 张乃光[1]

[1](国家广播电视总局 广播科学研究院 信息技术研究所,北京 100866)

[2](北京工商大学 计算机与信息工程学院,北京 100048)

通讯作者: 郭沛宇, E-mail: guopeiyu@abs.ac.cn 祝晓斌, E-mail: zhuxiaobin@btbu.edu.cn

摘 要: 近年来,深度学习在图像描述领域得到越来越多的关注.现有的深度模型方法一般通过卷积神经网络进行特征提取,递归神经网络对特征拼接生成语句.然而,当图像较为复杂时,特征提取不准确且语句生成模型模式固定,部分语句不具备连贯性.基于此,提出一种结合多频道特征提取模型与生成式对抗网络框架的图像描述方法——CACNN-GAN.此方法在卷积层加入频道注意力机制在各频道提取特征,与COCO图像集进行近似特征比对,选择排序靠前的图像特征作为生成式对抗网络的输入,通过生成器与鉴别器之间的博弈过程,训练句法多样、语句通顺、词汇丰富的语句生成器模型.在实际数据集上的实验结果表明,CACNN-GAN 能够有效地对图像进行语义描述,相比其他主流算法,显示出了更高的准确率.

关键词: 图像描述;生成式对抗网络;频道注意力模型;卷积神经网络

中文引用格式: 薛子育,郭沛宇,祝晓斌,张乃光.一种基于生成式对抗网络的图像描述方法.软件学报,2018,29(Suppl.(2)):30−43. http://www.jos.org.cn/1000-9825/18015.htm

英文引用格式: Xue ZY, Guo PY, Zhu XB, Zhang NG. Image description method based on generative adversarial networks. Ruan Jian Xue Bao/Journal of Software, 2018,29(Suppl.(2)):30−43 (in Chinese). http://www.jos.org.cn/1000-9825/18015.htm

Image Description Method Based on Generative Adversarial Networks

XUE Zi-Yu[1], GUO Pei-Yu[1], Zhu Xiao-Bin[2], ZHANG Nai-Guang[1]

[1](Information Technology Institute, Academy of Broadcasting Science, National Radio and Television Administration, Beijing 100866, China)

[2](School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China)

Abstract: In recent years, deep learning has gained more and more attention in image description. The existing deep learning methods using CNNs to extract features and RNNs to fold into one sentence. Nevertheless, when dealing with complex images, the feature extraction is inaccurate. And the fixed mode of sentence generation model leads to inconsistent sentences. To solve this problem, this study proposes a method combine channel-wise attention model and GANs, named CACNN-GAN. The channel-wise attention mechanism is added to each conv-layer to extract features, compare with the COCO dataset, and select the top features to generate sentence. Using GANs to generate the sentences, which is generated by the game process between the generator and the discriminator. After that, we can get a sentence generator contains the varied syntax, smooth sentence, and rich vocabulary. Experiments on real datasets illustrates that CACNN-GAN can effectively describe images, and get higher accuracy compared with the state-of-art.

Key words: image description; generative adversarial networks; channel-wise attention model; convolutional neural network

随着科技的快速发展,图像、视频等媒体数据大量出现在互联网中,成为信息传播的主要介质,且呈现出爆炸性增长的趋势.无标记和错误标记的图像散落在网络当中无法被检索和利用,造成大量资源的浪费.利用机器

* 基金项目: 国家广播电视总局广播科学研究院基本科研业务费课题(130016018000123)

Foundation item: Basal Research Fund of Academy of Broadcasting Science, National Radio and Television Administration (130016018000123)

收稿时间: 2018-04-16; 采用时间: 2018-10-24

学习进行图像处理,在图像的分类、识别、描述等应用场景下都取得了很好的效果.将图像和语义相匹配成为研究的重点和难点,初期的图像语义转换主要利用图像分类和图像识别等方法.图像分类是根据图像信息的不同特征区分类别信息,分配关键词作为语义关联[1].图像识别是根据图像不同模式的目标和对象,生成描述性单词或短语作为语义关联[2].初期的图像语义转换方法对图像中对象之间存在的位置、距离、比例分配等关系并无涉及,对图像信息的挖掘不够深入.随着图像语义转换方法的发展,从图像中获取更多有效的知识成为最需要解决的问题.近年来已成为研究热点的图像描述方法[3],为解决该问题提供了新的思路.

图像描述是运用机器学习方法对图像进行语句表述的图像理解方法.图像描述作为机器学习的热门研究方向,在很多领域已经得到了广泛应用,比如影片摘要描述,影片字幕生成,图像检索系统等.图像描述主要包含两个执行过程:图像特征提取过程和语句拼接生成过程,可以划分为两种:传统机器学习算法与深度模型结合的方法,以及深度模型实现方法.Chen 等人[4]提出的视觉递归描述模型、Fang 等人[5]提出的反馈视觉概念模型,都属于利用传统的机器学习方法提取特征,利用递归神经网络(recurrent neural networks,简称 RNN)进行语句拼接.Kiros 等人[6]提出的统一视觉语义嵌入模型,属于利用卷积神经网络(convolutional neural network,简称 CNN)提取图像特征,利用传统的编码器对特征进行句法拼接.为了适应大量的语句拼接以产生更连贯语句,Venugopalan 等人[7]提出的利用 RNN 的视频描述转换模型、Karpathy 等人[8]提出的 NeuralTalk、Vinyals 等人[9]提出的利用 GoogLeNet 的图像描述模型、Mao 等人[10]提出的多模态的 RNN、Chen 等人[11]提出的 SCA-CNN,均利用 CNN 提取图像特征,利用 RNN 进行句法拼接,属于直接利用深度模型进行图像描述的方法.

以上图像描述方法的目标识别过程面向整幅图像,没有考虑 CNN 提取特征的准确程度,目标物体的准确性难以保障.在句法生成过程中,大多利用 RNN 生成语法固定的语句,且不具连贯性和平滑性.针对上述图像描述方法的不足,本文提出了一种基于生成式对抗网络(generative adversarial networks,简称 GAN)的多频道特征提取模型,能够有效地从多个频道获取图像特征,通过特征排列的先后顺序选择特征,利用生成式对抗网络训练语句生成器,以生成平滑连贯的语句.本文主要做了如下 3 个方面的工作.

(1) 提出一种多频道注意力机制的特征提取方法.

(2) 提出一种基于生成式对抗网络的语句生成方法.利用(1)中提出的特征提取模型提取图像特征,将特征合并成为一组向量作为生成式对抗网络生成器的输入.

(3) 从理论和实验的角度对结合多频道特征提取模型和生成式对抗网络模型的合理性和有效性进行分析.

本文第 1 节介绍图像描述的相关工作.第 2 节给出多频道视觉机制的特征提取方法的合理性解释、介绍基于生成式对抗网络进行句法生成的基本思路.第 3 节对图像特征提取模型、语句生成模型利用 COCO 图像集进行实验,与主流算法进行对比,以验证该图像描述方法的合理性和有效性.第 4 节进行总结.

1 相关工作

与传统的图像语义获取方法不同,图像描述方法具有信息量大、关系复杂、表达丰富等显著特性,很多学者针对图像描述展开了深入研究.目前,主流的图像描述方法大多包括两个主要步骤,利用传统的特征计算方法或 CNN 进行图像特征提取,以及利用具有时序信息的递归神经网络进行语句拼接.Show and tell[9]以及 NeuralTalk[8]作为最经典的图像描述方法,其策略为:利用 CNN 进行特征提取,利用 RNN 模型或利用解决了梯度消失问题的长短时记忆模型(long short-term memory,简称 LSTM)进行语句拼接.很多学者在此基础上进行了改进,如 Chen 等人[1]提出的视觉递归描述模型,Fang 等人[5]提出的反馈视觉概念模型,均采用传统的机器学习方法提取图像特征,利用深度模型进行语句拼接.传统方法在数据量较小的数据集上表现较好,方法难以适应大规模图像集.并且传统模型在应对复杂图像时,表现出了特征提取准确度低、硬编码方式对机器要求高、特征提取时间长等问题.

深度模型适用于大规模复杂图像集,在图像描述中应用较为广泛.Jeff 等人[12]通过在时空上构建更为复杂的双深度模型和增加 CNN 模型的层数,弥补了传统模型在特征提取过程中关键目标缺失等问题,提升了特征提取的准确率.Kai 等人[13]利用基于区域的 CNN 进行特征提取,通过对视频关键帧位置的跟踪和目标大小变化的

影响训练特征提取模型,利用视频推进的方式训练图像之间的词语关联.方法虽然提升了准确率,但对图像中目标的位置选择以及目标与生成短语的对应关系没有讨论,而位置信息在生成语句当中至关重要.Karpathy 等人[8]提出的模型利用双向的 RNN 进行了关键目标与短语的匹配工作,方法通过已经生成的短语和将要生成的短语共同影响本次的短语选择,并确定前 19 类作为语句生成的建议词汇.该方法获得了很多认可和成功,但是特征提取过程并未考虑到空间中的信息.Krause 等人[14]借鉴 Faster R-CNN[15]模型思想,利用 CNN 结合区域推荐网络进行目标选择,对空间信息的合理利用使得语句的关键词准确率有了提高,但是该方法没有考虑到卷积过程中同样可以获得知识,从而提高特征提取的准确率.Chen 等人[11]提出的 SCA-CNN 对卷积过程加入频道注意力机制,使生成语句在单词的准确度上有了保证,但是该方法利用传统的 LSTM 模型进行语句拼接,在语句的平滑度、完整度以及准确度上仍有很大的提升空间.

在图像描述方法中,句法生成模型的调整和优化被用于生成平滑、完整、表达清晰的语句和段落.Mao 等人[10]提出了利用 CNN 提取图像特征,结合 RNN 进行语句生成的方法.该方法通过增加模型层数,提高了模型的准确度,但是方法并未考虑语句生成过程具有的时序性,单词需要被“记住”和“遗忘”.Vinyals 等人[9]、Kiros 等人[6]、Jeff 等人[12]、Venugopalan 等人[7]、Shekhar 等人[16]均利用 LSTM 进行语句生成,该模型加入了“遗忘门”,相比 RNN 模型,也在梯度上进行了优化,但是模型中并未检查词汇嵌入的合理性,导致以上方法难以取得较好的表达效果.Krause 等人[14]提出的模型,在语句生成器后加入词汇 LSTM,通过对个别词汇的优化,使语句更具有关联性.但是算法本质仍是利用固定句型生成固定语句,在语句的合理表达上仍有提升空间.Yang 等人[17]和 Yao 等人[18]提出的图像描述算法利用多层 CNN 进行特征获取,但是在确定目标主体的过程中,通过区域的权重大小确定区域的重要性,仍然具有很强的不确定性[19].

随着生成式对抗网络的发展,通过零和博弈过程训练和产生语句生成器的方法逐渐被利用和认可.GAN 模型在图像生成领域应用较为成功,2017 年初部分学者尝试利用 GAN 模型进行语句生成.Liang 等人[20]利用 WGAN 的思想提出 RTT-GAN,采用类似于 Krause 等人[14]的词语 LSTM 的思想,通过段落生成器、语句生成器、词语生成器共同训练,取得较好的语句生成器.但是由于该模型对特征提取过程采用简单的 CNN 模型,导致语料拼接过程中知识较少,生成语句描述不够充分,并且论文围绕图像和段落展开,对于单个语句的优化讨论较少.Press 等人[21]将 CNN 和 RNN 模型嵌入 GAN 模型,用以合作生成语句.该方法可以取得连贯通顺,词汇量丰富的语句描述,但是方法利用简单的 CNN 模型提取图像特征,并未对注意力机制有所提及和利用.Liang 等人[22]提出的双向 GAN 模型,一端用于视频关键帧抽取,另一端用于关键帧语句生成.以上方法通过对语句生成模型进行优化和调整,利用 GAN 模型进行段落生成,在段落内语句并未得到优化,待拼接特征的准确度考虑因素较少,因此以上方法生成的描述语句较为平滑连贯,但是语料丰富度可以进一步提升.针对以上算法的不足,受 Chen 等人[11]的 SCA-CNN 模型和 Liang 等人[20]提出的 RTT-GAN 模型的启发,本文提出了一种结合多频道特征提取模型和生成式对抗网络的图像描述方法——CACNN-GAN.该方法利用频道注意力模型加入到 CNN 卷积层中,并利用语句 LSTM、词语 LSTM 和鉴别器构建 GAN 模型.相比其他图像特征提取方法[8−10,13,16−18],CACNN-GAN 通过在频道层面对比图像库进行特征提取,获取到更准确的主体目标.相比现有的应用于语义描述领域的 GAN 模型方法[14,20,21],CACNN-GAN 创新地利用词语 LSTM 对语句中出现的词语进行了优化,在语句层面获得了通顺连贯的图像描述.

2 算法简介

2.1 研究动机

图 1(a)是图像物体识别过程,通过识别生成图 1(b)中的词语,根据词语的关联性进行排序,选择出图 1(c)中的有效词汇,以此连接生成图 1(d)中的语句, 图 1(a)~图 1(d)即是图像描述的整体过程.图像的语句描述可以获取更多的知识和信息,贴近人类的表达习惯.本文主要利用深度模型方法生成正确通顺的语句描述.研究动机包括:(1) 获取更准确的图像特征.在目前的方法当中,图 1(a)、图 1(b)的提取过程均面向整幅图像,特征提取粒度较粗,从而导致提取结果不准确,大量的无关词语出现在语句当中,生成语句与原图描述产生很大的偏差;(2) 获

取更具有关联性的特征.每幅图像之间的目标物体都具有一定的关联,在图 1(b)、图 1(c)的选择中,需考虑图 1(a)中词语间的位置关系,如①是整幅图像“后面(behind)”的背景,②中的船只在湖泊的“上面(on)”,③中的树木在岸的“上面(on)”等.在生成语句中对词汇的选择和排列,关系到语句的合理性.若被选择的词语之间关联性低或位置不正确,将导致语句连接不够合理;(3) 生成语句连贯通顺.图 1(c)、图 1(d)的语句生成过程按照固定的模型进行语句生成,句型单一固定,与人类表达方式差别大.需要减少模型的句法约束,增强各个词语间的连贯性.本文通过提取更准确的特征,选择合理的词汇,遵循自由的语法规则,生成准确、连贯、有效的语句,更好地实现图像描述.

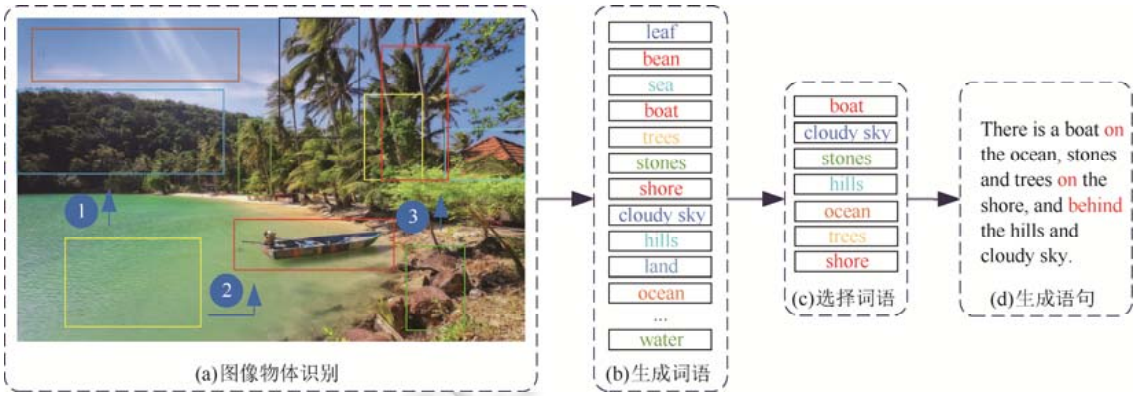

Fig.1 The main process flow-chart in image description

图 1 图像描述主要过程流程图

2.2 问题描述

本文按照[23]中的方法将整幅图像 I 分割为目标区块,每个区块包含有一个目标物体.在目标物体内利用 CNN 进行特征提取,其中 CNN 是包含了频道注意机制的特征提取多层模型,将在第 2.3 节进行描述.本文按照[24]的实验结果,在整幅图像中按经验选取前 19 个目标物体,并计算各个目标物体在像素区域 I_p 内的特征化表示 v:

$$v = W_M[CNN_{\theta_c}(I_p)] \tag{1}$$

其中,θ_c 是 CNN 的模型参数,该参数包含有 6 000 万的参数量,W_M 表示 h×4096 维度的矩阵,其中 h 是模型嵌入空间的尺寸大小,本文按照文献[24]的实验结果将 h 设定为 1000~1600,每个图像都是一个 h 维的向量表示.设图像 I 将被转换成为 N 个词汇,每个词汇嵌入都用一个向量进行表示,每幅图像可以表示为一组 h 维的向量 $\{v^i|i=1,...,19\}$,语句输出序列为 $\{y_1,y_2,...,y_L\}$,按照图像特征描述方式生成语句:

$$y = W_{TL}[RNN(v)],$$

其中,W_{TL} 表示图像区域特征描述向量 v 与语句输出序列 y 之间的转换矩阵,RNN 影响着输出序列中的下一个词汇,影响因素包括以往的生成词汇以及新的目标区域特征,将在第 2.4 节进行描述.其中单个词汇 y_1 在第 1 步的语句选择式生成结果可以表示为

$$y_t = soft\max\{W_1 \times f(W_2 x_t + W_3 h_{t-1} + v_a)\} + b_o,$$

其中,W_1、W_2、W_3 与 b_o 均是向量和偏置量,用于调节和匹配大小,其中 b_o 是线性参数.v_a 是公式 1 中特征提取模型网络的最后一层输出特征,x_t 为新输入的区域特征向量,h_{t-1} 隐藏层状态向量,表达了以往的词汇对新词汇的影响值.方法 f 表示了词汇选择过程中的分数选择方法.

2.3 多频道图像特征提取算法

如图 2 所示,多频道图像特征提取算法按照加解密机原理进行设计,加密机对图像进行特征提取,并在在卷积层增加频道注意力机制[25].设公式(1)中第 i 个区域的视觉特征向量为 v_i,在公式(1)的 CNN 当中,利用多频道

图像特征提取算法,通过频道模型转换将视觉特征向量 V 转换为 U,则 $u_i \in R^{W \times H}$ 对应视觉特征向量 v_i,其中 W 和 H 表示图像的宽度和高度,V 和 U 具有相同的视觉尺寸 $W \times H$.

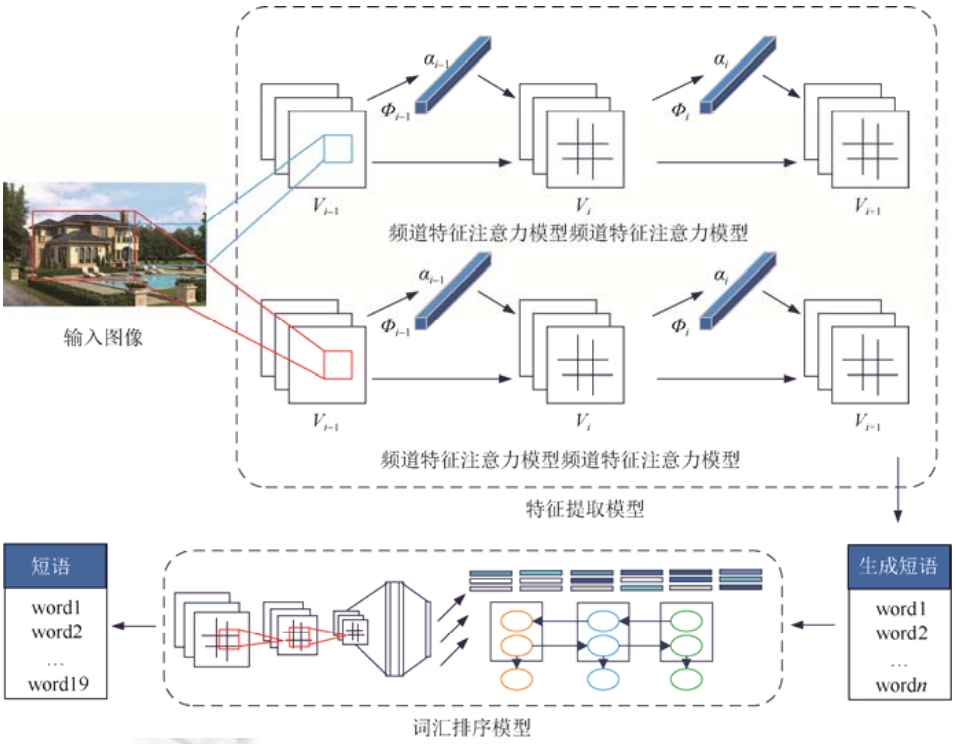

Fig.2 The flow chart of image feature extraction algorithm using multi-channel

图 2 多频道图像特征提取算法流程图

本文在 CNN 卷积过程中,在各层加入频道特征注意力模型.设在卷积层 t 加入了基于频道注意力机制,ϕ_c 是频道注意力机制模型,用于获得多层的频道注意力权重特征图 β^t.根据前一个隐藏层的结果为 h_{t-1},利用单层的神经网络生成图像区域的注意力贡献值为∂.频道注意力机制模型 ϕ_c 可以用公式(2)进行表示:

$$\left.\begin{aligned} \alpha &= \tanh\left(\left(W_C \otimes v + b_c\right) \oplus W_{hc} h_{t-1}\right) \\ \partial &= soft\max\left(W_i \alpha + b_i\right) \end{aligned}\right\} \tag{2}$$

其中矩阵 $W_c \in R^k$、$W_{hc} \in R^{k \times d}$、$W_i \in R^k$ 用于将图像的视觉特征处理为隐藏层相同的维数.⊗表示与向量外部元素的运算,⊕表示矩阵与向量之间的运算,b_c 与 b_i 为偏置值.

结合公式(1),按照公式(3)将 v_i 利用频道注意机制卷积神经网络转化特征提取结果为 u_i,其中 h 为上一隐藏层状态.

$$u_i = \phi_c\left(h_{i-1}, v_i\right) \tag{3}$$

2.4 利用生成式对抗网络的语义描述算法

本文利用 GAN 模型进行语句生成,图 3 是本文提出的 GAN 模型的生成器示意图.该生成器由语句 LSTM、词汇 LSTM 以及两个视觉注意力机制模型构成,LSTM 分别用于处理语句级别拼接和词汇级别的嵌入.

其中视觉注意力模型用于选择词汇对应区域,语句 LSTM 由一个单层的 LSTM 构成,用于嵌入词汇到语句当中,在获取语句特征向量之后,本文按照文献[26]所提方法获取词语中包含的主成分,去除无关词语对语句生成造成的噪声影响.词汇 LSTM 通过语料库学习知识,根据语句生成情况调整嵌入到语句中的词汇.根据 2.2 节的总体描述,设区域描述序列为$(v_1, v_2, \ldots, v_T)$,隐藏层状态为$(h_1, h_2, \ldots, h_T)$,输出序列为$(y_1, y_2, \ldots, y_T)$,在第 t 步输出到

语句的下个词汇预测结果可以表示为

$$y_t = soft\max\left\{W_{oh} \times f\left(W_{hx}x_t + W_{hh}h_{t-1} + u_i\right) + b_o\right\} \tag{4}$$

其中,W_{oh}、W_{hx}、W_{hh}均是学习参数,b_o是偏置量,u_i为公式(3)在第i个位置的视觉特征向量输出.为获得更为平滑的语句,词汇 LSTM 根据语序对词语的关系进行调整.根据词语和词语在图像中的概率进行选择,词汇 LSTM 在语句中被重复利用,直到语句全部输出完成.图 4 是语句鉴别器模型,实现对语句生成器产生语句的鉴别工作.语句鉴别器是一个单层的 LSTM,该模型根据词语嵌入语句的顺序和真实语句进行对比,按照 BLEU 评价标准[27]输出语句的真实程度.

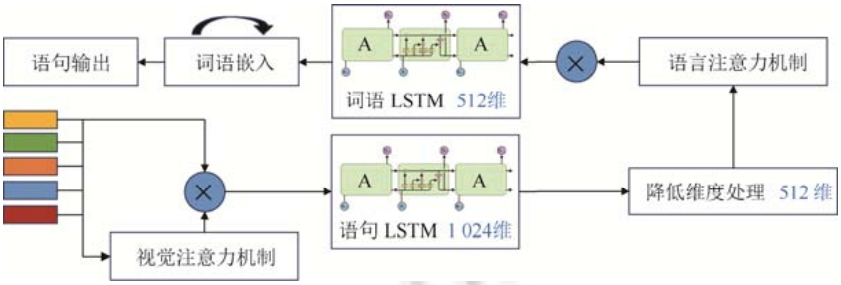

Fig.3 Model of the sentence generator

图 3 语句生成器模型图

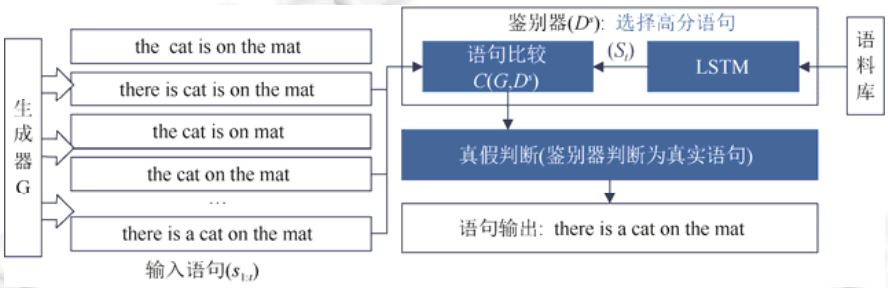

Fig.4 Model of the sentence discriminator

图 4 语句鉴别器模型图

在生成器和鉴别器相互博弈的过程中,模型会出现梯度消失的现象.为避免这一问题,采用确定性近似方法,在单词生成过程中根据单词的 softmax 扩散贡献,利用词汇进行梯度扩散.本文跟随 WGAN 模型[20]的训练方式对生成器进行训练:

$$\min\max C\left(G,D^s\right)_{G,D^s} = E_{\hat{s}\sim S}\left[D^s\left(\hat{s}\right)\right] - E_{\hat{s}\sim S_{1:t}}\left[D^s\left(\hat{s}\right)\right] \tag{5}$$

其中,$\hat{s}$ 表示真实语句,S 和 $S_{1:t}$ 代表生成语句和真实语句,真实语句来自于训练集中的语料.本文按照公式(6)计算梯度更新鉴别器.

$$\nabla_{\theta_D}\frac{1}{t}\sum_{i=1}^{t}\left[\log D^s\left(S_{1:t}^{(i)}\right) + \log\left(1 - D^s\left(\left(S^{(i)}\right)\right)\right)\right] \tag{6}$$

鉴别器调节过程需要伴随生成器的更新,本文按照公式(7)计算梯度更新生成器.

$$\nabla_{\theta_G}\frac{1}{t}\sum_{i=1}^{t}\left[\log\left(1 - D^s\left(\left(S^{(i)}\right)\right)\right)\right] \tag{7}$$

其中,$\nabla_{\theta D}$和$\nabla_{\theta G}$分别表示鉴别器和生成器的梯度,公式(6)和公式(7)的更新趋于平缓时,可以结束训练.

2.5 结合多频道特征提取模型和生成式对抗网络的图像描述方法

如图 5 所示,特征提取模型与生成式对抗网络相结合进行语义描述的方法框架图.在图像特征提取完成后,

语句生成器进行语句生成,语句鉴别器用于判断语句的真假,判断为假的语句,通过反向传播对生成器做出调整,其中 γ 是平衡因子:

$$G' = \arg\min\max_{G,D^s} \gamma C\left(G, D^s\right).$$

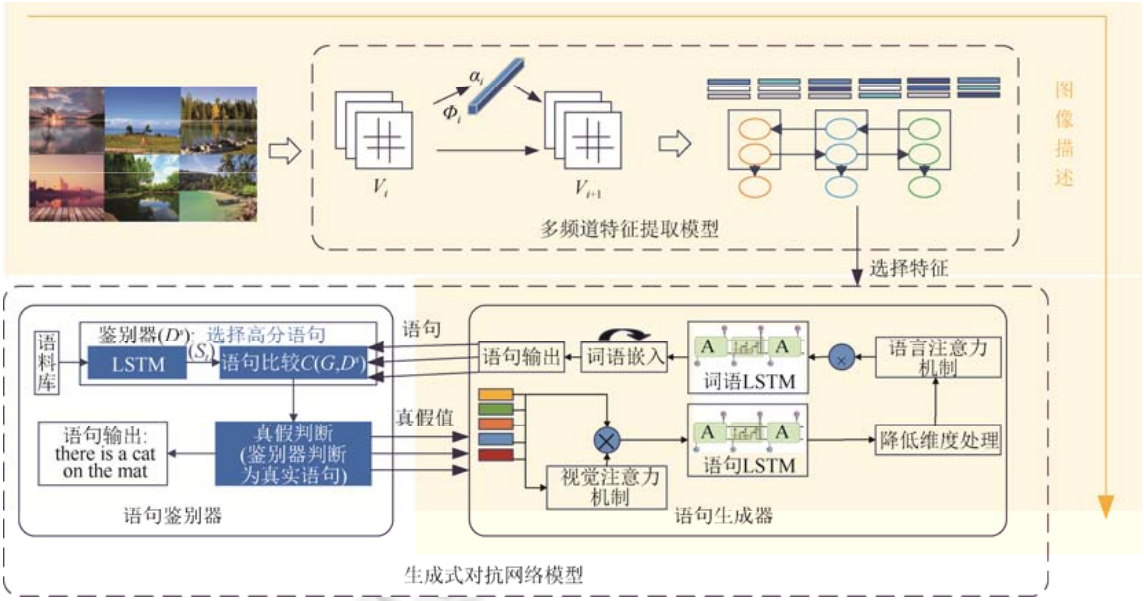

Fig.5 The framework of the image description method based on multi-channel feature extraction model and generative adversarial networks

图 5 结合多频道特征提取模型和生成式对抗网络的图像描述方法框架图

图 5 描述了模型的整体训练过程,有底色部分为模型训练完成后的图像描述过程.鉴别器是为了训练语句生成器,待生成器训练完成后,鉴别器将不再语句生成框架中使用.图像描述过程的运行流程包括特征提取、特征选择和语句生成等主要部分.

3 实验结果及分析

为了验证本文提出方法的合理性和有效性,设计如下 3 组实验:(1) 多频道特征提取算法结合递归神经网络的图像描述;(2) 卷积神经网络结合生成式对抗网络的图像描述;(3) 结合多频道特征提取模型和生成式对抗网络的图像描述.

3.1 实验数据和测评标准介绍

为了精确地控制源领域和目标领域的相似度,以探讨在各种相似度情形下本文提出方法的准确性,本文利用了 Flickr8k 数据集、Flickr30k 数据集和来自微软的 COCO 数据集.

Flickr8k 数据集[28]包含 8 000 张图像,根据官方发布结果,该图像集选择 6 000 张图像构成训练集,选择 1 000 张图像构成验证集,其余 1 000 张图像构成测试集.Flickr30k 数据集[29]包含 31 000 张图像,该图像集并未规定分布在各个集合中的图像数量,本文按照文献[11]的分配方案,选择 29 000 张图像构成训练集,选择 1 000 张图像构成验证集,其余 1 000 张图像构成测试集.MSCOCO 数据集[11]包含 82 783 张图像训练图像,40 504 张验证图像和 50 775 张测试图像,在部分研究当中,验证图像集可以作为测试图像集进行使用.本文按照文献[11]的分配方案,分配 82 783 张图像作为训练数据集,各选取 5 000 张图像作为验证集和测试集.

本文的测评指标包括 BLEU(B@1,B@2,B@3,B@4)[27]、METEOR(MT)[30]、CIDEr(CD)[31]、ROUGE-L(RG)[32]这 4 个指标都是通过比较生成语句和真实语句之间的相似程度获得的百分数,其中语句的相似程度判断标准

是计算相邻词语之间连续重合的词数.

3.2 实验方案介绍和实验结果分析

本文按照上文描述的方式进行对比试验,其中第 1 部分主要对改进的图像特征提取模型进行合理性验证,第 2 部分主要对句法生成模型进行合理性验证,第 3 部分对两个模型利用端到端方式进行拼接,对该图像描述方法做合理性验证.

在实验设置方面,本文提出方法所利用的 CNN 模型为 VGG-19 和 ResNet-152 基础模型,频道注意机制计算权重为 512 维.利用的 GAN 模型所包含的单层长短时记忆模型隐藏层维度设置为 512 维,隐藏层和记忆元素初始元素设定为 0.生成器和鉴别器的梯度更新空间限定为[–0.01,0.01],模型训练轮数设定为 40.在数据集设定上,Flickr8k 最小尺寸设置为 16,Flickr30k 和 MSCOCO 最小尺寸设置为 64.在语句限制设定上,最长语句限定为 30 词,超长即发出 STOP 指示.本文方法基于 Torch7 平台,使用 2 块英伟达泰坦 X 系列 GPU 进行运算.

3.2.1 多频道特征提取算法结合递归神经网络的图像描述

在本实验中,本文对两个基础模型(VGG-19 以及 ResNet-152)增加频道注意力机制,并利用实验证明新模型的合理性.见表 1,本文在 Flickr8k、Flickr30k、MSCOCO 图像集上进行实验.

Table 1 Performances of multi-channel attention mechanism.

表 1 多频道注意力机制效果分析

数据集	基础模型	实验方法	B@4	MT	RG	CD
Flickr8k	VGG-19	原始模型	21.3	20.3	–	–
		频道注意力模型	22.6	20.2	48.5	58.7
	ResNet-152	原始模型	21.7	20.1	48.4	55.5
		频道注意力模型	24.5	21.4	50.1	65.5
Flickr30k	VGG-19	原始模型	19.9	18.5	–	–
		频道注意力模型	20.2	18.2	42.6	38.1
	ResNet-152	原始模型	20.1	17.8	42.9	36.3
		频道注意力模型	21.5	18.3	43.8	42.1
MS COCO	VGG-19	原始模型	25.0	23.0	–	–
		频道注意力模型	28.0	23.1	50.7	84.9
	ResNet-152	原始模型	28.4	23.2	51.2	84.9
		频道注意力模型	29.6	23.7	51.9	91.1

本文提出的方法在增加多频道注意力机制之后,较普通方法有了一定提高.为了更直观地反映本文方法的合理性,图 6 列出 MSCOCO 中的几个实例,第 1 列为原图,第 2 列为利用 VGG-19 模型框取的实验结果,后三列分别为利用 VGG-19 模型、VGG-19 模型加入频道注意力机制的特征提取结果以及图像对应的真实标记,所有标记都经过了人工的关键字提取以方便比较,本文只列出了排序前 5 的短语和词汇.

如图 6 所示,加入频道注意力机制的方法,可以获取更多更有效的关键词.如第 1 幅图像左下角处趴着的长颈鹿,利用 VGG-19 原模型检测到的为“鹿”,而加入频道注意力机制的方法,经过与原图像集的对比,检测到的目标为“长颈鹿”;以及第 2 幅图像中,加入注意力机制检测到跳起的人是“运动员”,而不是普通的“男人”,背景是“房子”而不仅仅是“树木”;第 3 幅和第 4 幅中的示例同样证明了加入了频道注意机制的方法.通过在频道内的特征对比,取得了更贴近于真实值的目标.

由于本文提出模型是在各卷积层增加频道注意力机制,为了测量卷积过程层数对模型的影响,设置以下实验.表 2 中,本文利用 VGG-19 以及 ResNet-152 模型在 MSCOCO 数据集进行实验.按照 VGG-19 模型和 ResNet-152 模型在卷积层加入注意力模型的方法[11],分别在 VGG-19 模型卷积层中的 conv5_4、conv5_3、conv5_2; ResNet-152 模型卷积层中的 res5c、res5c_branch2b、res5c_branch2a 加入频道注意力模型,本文采用以上模型预训练结果.

图像 / 方法				
框取结果				
原模型 排序前5名	**deer** branch tree wood ground	**man** **man** bat lawn tree	**building** **wall** vehicle automobile house	train railway **brand** tree automobile
新模型 排序前5名	**giraffe** branch tree wood ground	**athlete** **athlete** bat lawn **building**	**mountain** **brand** vehicle automobile house	train railway **building** tree pole
真实标记 排序前5名	**giraffe** branch tree wood ground	**athlete** **athlete** bat lawn **building**	**mountain** **brand** vehicle automobile house	train railway **building** tree pole

Fig.6 The diagram of multi-channel attention mechanism

图 6 多频道注意力机制合理性展示图

Table 2 The performance of multi-layer in CNN models

表 2 卷积神经网络的层数效果分析

网络	层数	B@4	MT	RG	CD
VGG-19	conv5_4	27.2	22.7	49.8	83.8
	conv5_3	**28.9**	**23.2**	**50.9**	**88.9**
	conv5_2	28.4	22.9	50.8	87.3
ResNet-152	res5c	29.5	23.3	51.5	90.6
	res5c_branch2b	**30.0**	**24.1**	**52.0**	**94.3**
	res5c_branch2a	29.8	23.9	52.0	93.6

本文按照表 2 中最有效的实验结果作为后续实验的实验条件,选取 VGG-19 模型的 conv5_3 层以及 ResNet-152 模型的 res5c_branch2b 层添加频道注意力模型.图 7 表示了原图在 VGG-19 网络中的两个卷积层进行特征提取,与 COCO 数据库中特征进行比对之后,排名靠前的结果.

Fig.7 The comparison results in each layer when using VGG-19 network to extract feature.

图 7 利用 VGG-19 网络在卷积层提取特征对比的结果图

3.2.2 卷积神经网络结合生成式对抗网络的图像描述

在本实验中,利用 Deep VS[8]中的特征提取方法进行图像特征提取,方法中基于区域的卷积神经网络利用 ImageNet 进行预训练,采用 2015 年 ILSVRC 比赛的目标检测冠军算法[8],该算法可以分辨 200 个物体类别.本节算法规定选取排名靠前的 19 个区域进行句法生成.采用双向递归神经网络模型对目标区域与生成短语进行关联操作,确保选择区域在图像中的位置正确.

图像特征提取结果如图 8 所示,方法在测试图像数量为 1 000 和 5 000 时的召回率见表 3,其中 K 表示前 K 个结果中检索到正确短语的数量,Med 表示中位秩.

Fig.8 Image feature extraction and selection results

图 8 图像特征提取和选取结果

Table 3 The recall rate and median rank of image area match with phrase

表 3 图像区域与短语匹配正确的召回率和中位秩

	R@1	R@5	R@10	Med r
1 000 张测试图像	38.4	69.9	80.5	1.0
5 000 张测试图像	16.5	39.2	52.0	9.0

本节利用基于 Wasserstein 距离的生成式对抗网络[33]进行语义生成,其中生成器和鉴别器分别采用单层和两层的长短时记忆模型作为网络设置,实验结果见表 4.

Table 4 The performance of multi-layer in both generator and discriminator

表 4 生成式对抗网络生成器鉴别器层数设置方式表

鉴别器	生成器							
	单层				双层			
	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4
单层	62.7	45.3	32.5	19.7	62.7	45.5	32.8	20.1
双层	62.7	45.1	32.5	19.8	62.7	45.6	32.9	19.9

从表 4 可以看出,在生成器的长短时记忆模型层数固定之后,鉴别器长短时记忆模型层数的变化对实验结果影响不大.生成器采用双层的长短时记忆模型生成语句的结果与采用单层模型在 B-n($n\geqslant 1$)时差别不大,在 B-n($n\geqslant 2$)时略有增加,但增幅不大.本文在下文采用单双层长短时记忆模型作为生成器分别进行实验.

相比 RCNN 进行图像特征提取,LSTM 进行语句生成的方式,本文进行了如下实验,实验本文方法中利用 GAN 模型博弈训练 LSTM 的合理性,实验数据见表 5.

Table 5 The comparision between using GAN model to train LSTM and using original model

表 5 利用 GAN 模型训练 LSTM 和原始模型比较结果表格

模型	B-1	B-2	B-3	B-4	MT	CIDEr
RCNN+LSTM	62.5	45.0	32.1	23.0	19.5	66.0
RCNN+GAN(LSTM)	62.7	45.3	32.5	23.5	19.7	66.7

表 5 利用 GAN 模型对 LSTM 进行训练,直到生成器 LSTM 足以蒙骗鉴别器,获取此刻的生成器模型.该生成器相比原生成器,可以获得更优的语句生成.如图 9 所示,是本文按表 5 方法的选取的 5 个示例,其中前 4 个为较为明显的转换结果,最后 1 个为改变较小的转换结果.本文以 BLEU 值作为评价标准,标记为粗体部分为与真实语句的重叠量,黑体越多的表明重叠部分越多,即对应的 BLEU 值越优.

图像 / 内容					
MSCOCO 图像编号	140603	242298	491992	34439	321717
RCNN+ LSTM	a wooden door and a white bathroom.	an airplane stopped on the road.	a woman and **a dog** paddle.	a man with a **skateboard on** a high brick.	there are several **sandwiches on** pan.
RCNN +GAN (LSTM)	**a wooden door to a bathroom that is white inside.**	an airplane **is parked on** the way.	a woman **paddle boarding with a dog.**	a **man riding a skateboard on** a high brick.	there are **several sandwiches on** pan.
真实语句	**a wooden door to a bathroom that is white inside.**	a big airplane that **is parked on** a runway.	a couple of people **paddle boarding with a dog** on one board.	a young **man riding a skateboard on** top of a cement brick.	person grilling **several** ham **sandwiches on** white bread.

Fig.9 The diagram of GAN

图 9 GAN 模型合理性展示图

3.2.3 结合多频道特征提取模型和生成式对抗网络的图像描述

本文将提出方法 CACNN-GAN 与主流算法[4,5,8,9,25,31,32]进行对比,选取对比的数据集是 Flickr8k、Flickr30k 以及 MS COCO 数据集.表 6、表 7 是 CACNN-GAN 与其他主流方法的比较结果,根据文献[11]将本文词汇嵌入最大尺寸设置为 100,LSTM 隐藏层尺寸设置为 1 000,视觉注意力模型最大权重设置为 512.Flickr8k 的批尺寸设置为 16,Flickr30k 和 MS COCO 数据集的批尺寸设置为 64.本文提出的所有实验模型均按照 Adadelta 方法[34],采用端到端的训练方式,该方法使用自适应学习率方法进行随机梯度下降.语句在生成过程当中,遇到语句结束符 END 或者到达语句最大限制长度,即停止语句生成.本文利用 BeamSearch 方法[9]进行测试,该方法各候选集中选择最佳生成语句,最大尺寸设置为 5.

从表 6、表 7 中可以看到,CACNN+LSTM 是本文利用 VGG-19 网络加入频道注意力机制的实验结果,其中文献[8,9]中介绍的方法也是采用部分 CNN 结合 LSTM 进行的实验,本文方法在卷积过程中,在各频道进行特征提取,语句生成结果明显高于其他方法.特别是在 B@1 的计算上,相比文献[8]方法提高了 7.3 个百分点,相比文献[9]方法也提高了 0.5 个百分点.由于其他方法对语句生成器模型有所更改,没有做到单一变量原则,没有进行比较.RCNN+GAN 是本文利用[8]中的特征提取方法,结合 LSTM 做的图像描述方法.相比文献[8]方法,本文方法在 B@n 上有所提升,总体提升 0.5 个百分点左右.

Table 6 Performances compared with the state-of-art in Flickr8k and Flickr30k dataset

表 6 CACNN-GAN 方法与主流方法在 Flickr8k、Flickr30k 数据集中的比较表格

模型	Flickr8k					Flickr30k				
	B@1	B@2	B@3	B@4	MT	B@1	B@2	B@3	B@4	MT
Deep VS[8]	57.9	38.3	24.5	16.0	–	57.3	36.9	24.0	15.7	–
Google NIC[9]	63.0	41.0	27.0	–	–	66.3	42.3	27.7	18.3	–
m-RNN[35]	–	–	–	–	–	60.0	41.0	28.0	19.0	–
Soft-Attention[36]	67.0	44.8	29.9	19.5	18.9	66.7	43.4	28.8	19.1	18.5
Hard-Attention[36]	67.0	45.7	31.4	21.3	20.3	66.9	43.9	29.6	19.9	18.5
emb-gLSTM[37]	64.7	45.9	31.8	21.2	20.6	64.6	44.6	30.5	20.6	17.9
ATT[38]	–	–	–	–	–	64.7	46.0	32.4	23.0	18.9
CACNN+LSTM[8,9]	65.2	46.3	32.3	22.6	20.3	64.3	45.1	31.5	20.1	18.0
RCNN[8]+GAN	58.2	39.1	25.2	17.0	–	58.2	37.4	24.5	16.2	–
CACNN-GAN(VGG-19)	66.3	47.7	33.5	23.9	21.1	65.2	46.3	32.2	21.3	19.2
CACNN-GAN (ResNet-152)	69.3	49.9	35.8	25.9	22.3	68.3	49.1	34.6	23.5	20.1

Table 7 Performances compared with the state-of-art in MS COCO dataset

表 7 CACNN-GAN 方法与主流方法在 MS COCO 数据集中的比较表格

模型	MS COCO				
	B@1	B@2	B@3	B@4	MT
Deep VS[8]	62.5	45.0	32.1	23.0	19.5
Google NIC[9]	66.6	46.1	32.9	24.6	–
m-RNN[35]	67.0	49.0	35.0	25.0	–
Soft-Attention[36]	70.7	49.2	34.4	24.3	23.9
Hard-Attention[36]	71.8	50.4	35.7	25.0	23.0
emb-gLSTM[37]	67.0	49.1	35.8	26.4	22.7
ATT[38]	70.9	53.7	40.2	29.2	24.3
Review Net[17]	70.9	53.9	40.1	29.2	24.2
MSM[18]	71.2	54.1	40.3	29.3	24.2
GLA-BEAM3[16]	72.5	55.6	41.7	31.2	24.9
CACNN+LSTM[8,9]	69.3	52.1	37.5	27.3	22.7
RCNN[8]+GAN	62.7	45.3	32.5	23.5	19.7
CACNN-GAN(VGG-19)	71.2	53.7	38.9	27.9	23.9
CACNN-GAN (ResNet-152)	74.3	55.9	41.3	30.9	25.1

结合以上两种方法,本文利用 VGG-19 以及 ResNet-152 作为基础模型,在此基础上利用频道注意力模型进行特征提取.利用本文提出的 GAN 模型训练生成的 LSTM 作为语句生成器,本文提出 CACNN-GAN 模型,从表 6、表 7 中可以看到,本文提出方法较其他主流方法均有提高.图 10 是利用以上方法,是本文选取的部分图像输出的语句作为主观评价结果.

图像 / 方法					
CACNN +LSTM	a chimney building on the lawn.	the man on city street ride bike.	a plane flies in cloudy sky.	there is a bridge near the lake, and a red leaf tree is on the edge of the bridge.	there are waterfalls under the green tree.
RCNN +GAN	a large building with a clock tower on top of it.	a man riding a bike on a city street.	a plane flies in the sky.	a park bench in a wooded area with trees in the background.	water ran down the mountain.
CACNN-GAN (VGG-19)	a chimney building on the lawn.	a man riding a bike on a city street.	a plane flies in cloudy sky.	a bridge near the lake, with a tree is on the edge of the bridge.	there are waterfalls under the green tree.
CACNN-GAN (ResNet)	a chimney building on the lawn under the cloudy sky.	a man riding a black bike on a city street.	a plane flies in cloudy sky.	a bridge near the lake, with a red leaf tree is on the edge of the bridge.	there are waterfalls under the green tree.
真实语句	a building with a chimney on the lawn.	a man ride a bike on the street.	a plane is flying in cloudy sky.	there is a bridge near the lake, with a tree is on the edge of the bridge.	there are waterfalls under the green tree.

Fig.10 Examples of visualization results

图 10 利用本文提出方法的图像描述结果

从图 10 的实验结果,可以总结以下 3 点.

(1) 本文提出的 CACNN 模型相比其对应的 CNN 模型,可以检测出更准确的细节.如第 1 幅图当中,RCNN+GAN 方法检测到的是一个塔钟,而不是烟囱;第 3 幅图中,真实语句是“多云的天空”,CNN 方法只检测到天空;第 4 幅图像更为明显,未加频道注意力机制的 CNN,把石桥的特征提取并匹配成为木制的椅子;第 5 幅图中 CNN 检测的是“山”,而利用频道的方式,可以检测到“瀑布”和“绿色的树”.

(2) 本文提出的 GAN 模型相比未经博弈训练获得的 LSTM 方法,生成语句在主观上感觉更平滑和连贯.如

第 1 幅中的“一个带着烟筒的房子”以及未经博弈训练的“一个大房子和钟楼”相比前者更加连贯.

(3) 利用 ResNet 比 VGG-19 可以获得更丰富的词汇,具有更强的语句表现力.如第 1 幅图当中的草坪、第 2 幅图中自行车的颜色、第 4 幅图中树的颜色.

最后一幅图展示了一个不好的例子,虽然本文提出的利用 GAN 模型博弈生成 LSTM 进行语句生成,但是由于真实语句是由人类主观生成,所以虽然在检测过程中所有特征提取正确、语句拼接正确,但是在 B@4 测量度上,仍然降低了模型总体的测试结果.

4 总 结

本文结合多频道特征提取模型和生成式对抗网络,提出一种新的图像描述方法——CACNN-GAN.本文首先提出了一种利用多频道的图像特征提取方法,并从理论上分析了在各卷积层引入多频道的合理性.CACNN-GAN 利用多频道获取特征,与 COCO 数据集产生对比,选取结果靠前的方法作为频道特征.生成式对抗网络根据排序结果进行语句生成,通过鉴别器和生成器的不断博弈,产生语法自然的语句.进一步的实验结果表明,多频道特征提取模型与生成式对抗网络相结合进行图像描述,可以有效提高描述的准确度,提升语句的连贯性和完整度.未来的工作包括:研究其他特征提取模型方法;研究确定特征提取模型与语句生成模型特征传递的方法;研究特征提取模型与语句生成模型的流水线训练方式.

References:

[1] Huang KQ, Ren WQ, Tan TN. A review on image object classification and detection. Chinese Journal of Computers, 2014,37(6):1225–1240 (in Chinese with English abstract).

[2] Huang WG, Gu C, Shang L, *et al*. Hierarchical representation method for object recognition. Chinese Journal of Electronics, 2015,43(5):854–861 (in Chinese with English abstract).

[3] Chang L, Deng XM, Zhou MQ, *et al*. Convolutional neural networks in image understanding. Acta Automatica Sinica, 2016,42(9):1300–1312 (in Chinese with English abstract).

[4] Chen XL, Zitnick CL. Mind’s eye: A recurrent visual representation for image caption generation. In: Proc. of the IEEE Conf. on CVPR. 2015. 2422–2431.

[5] Fang H, Gupta S, Iandola F, *et al*. From captions to visual concepts and back. In: Proc. of the IEEE Conf. on CVPR. 2015. 1473–1482.

[6] Kiros R, Salakhutdinov R, Zemel R. Multimodal neural language models. In: Proc. of the Int’l Conf. on Machine Learning. 2014. 595–603.

[7] Venugopalan S, Xu H, Donahue J, *et al*. Translating videos to natural language using deep recurrent neural networks. arXiv preprint arXiv:1412.4729, 2014.

[8] Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions. In: Proc. of the IEEE Conf. on CVPR. 2015. 3128−3137.

[9] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator. In: Proc. of the IEEE Conf. on CVPR. 2015. 3156–3164.

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

[11] Chen L, Zhang H, Xiao J, *et al*. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In: Proc. of the IEEE Conf. on CVPR. IEEE Computer Society, 2017. 6298–6306.

[12] Donahue J, Hendricks LA, Guadarrama M, *et al*. Long-Term recurrent convolutional networks for visual recognition and description. In: Proc. of the IEEE Conf. on CVPR. 2015. 2625–2634.

[13] Kang K, Ouyang WL, Li HS, *et al*. Object detection from video tubelets with convolutional neural networks. In: Proc. of the CVPR. IEEE, 2016. 817–825.

[14] Krause J, Johnson J, Krishna R, *et al*. A Hierarchical Approach for Generating Descriptive Image Paragraphs. In: Proc. of the CVPR. IEEE, 2017. 3337–3345.

[15] Ren S, He K, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. on Pattern Anal Mach Intell, 2017,39(6):1137–1149.

[16] Shekhar R, Pezzelle S, Klimovich Y, *et al*. FOIL it! Find one mismatch between Image and Language caption. arXiv preprint arXiv:1705.01359, 2017.

[17] Yang Z, Yuan Y, Wu Y, *et al*. Review networks for caption generation. In: Advances in Neural Information Processing Systems. 2016. 2361–2369.

[18] Yao T, Pan Y, Li Y, *et al*. Boosting image captioning with attributes. In: Proc. of the ICCV. IEEE, 2017. 22–29.

[19] Anderson P, He X, Buehler C, *et al*. Bottom-Up and top-down attention for image captioning and visual question answering. In: Proc. of the IEEE Conf. on CVPR. 2018,3(5):6.

[20] Liang X, Hu Z, Zhang H, *et al*. Recurrent topic-transition GAN for visual paragraph generation. arXiv preprint arXiv:1703.07022, 2017.
[21] Press O, Bar A, Bogin B, *et al*. Language generation with recurrent generative adversarial networks without pre-training. arXiv preprint arXiv:1706.01399, 2017.
[22] Liang X, Lee L, Dai W, *et al*. Dual motion GAN for future-flow embedded video prediction. In: Proc. of the ICCV. IEEE, 2017,1.
[23] Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2014. 580–587.
[24] Karpathy A, Joulin A, Fei-Fei L F. Deep fragment embeddings for bidirectional image sentence mapping. In: Advances in neural information processing systems. 2014. 1889–1897.
[25] Zhuo J, Wang S, Zhang W, *et al*. Deep unsupervised convolutional domain adaptation. In: Proc. of the 2017 ACM on Multimedia Conference. ACM, 2017. 261–269.
[26] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. arXiv preprint arXiv:1709.01507, 2017.
[27] Papineni K, Roukos S, Ward T, *et al*. BLEU: A method for automatic evaluation of machine translation. In: Proc. of the 40th Annual Meeting on ACL. 2002. 311–318.
[28] Hodosh M, Young P, Hockenmaier J. Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research, 2013,47:853–899.
[29] Young P, Lai A, Hodosh M, *et al*. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In: Proc. of the TACL. 2014,2:67–78.
[30] Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proc. of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. 2005. 65–72.
[31] Vedantam R, Zitnick CL, Parikh D. Cider: Consensus based image description evaluation. In: Proc. of the IEEE Conf. on CVPR. 2015. 4566–4575.
[32] Lin CY. Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out, 2004.
[33] Arjovsky M, Chintala S, Bottou L. Wasserstein gan. arXiv preprint arXiv:1701.07875, 2017.
[34] Zeiler MD. Adadelta: An adaptive learning rate method. arXiv preprint arXiv:1212.5701, 2012.
[35] Mao J, Xu W, Yang Y, *et al*. Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint arXiv:1412.6632, 2014.
[36] Xu K, Ba J, Kiros R, *et al*. Show, attend and tell: Neural image caption generation with visual attention. In: Proc. of the Int'l Conf. on Machine Learning. 2015. 2048–2057.
[37] Jia X, Gavves E, Fernando B, *et al*. Guiding the long-short term memory model for image caption generation. In: Proc. of the IEEE ICCV. 2015. 2407–2415.
[38] You Q, Jin H, Wang Z, *et al*. Image captioning with semantic attention. In: Proc. of the IEEE Conf. on CVPR. 2016. 4651–4659.

附中文参考文献:

[1] 黄凯奇,任伟强,谭铁牛.图像物体分类与检测算法综述.计算机学报,2014,37(6):1225–1240.
[2] 黄伟国,顾超,尚丽,杨剑宇,朱忠奎.基于轮廓分层描述的目标识别算法研究.电子学报,2015,43(5):854–861.
[3] 常亮,邓小明,周明全,武仲科,袁野,杨硕,王宏安.图像理解中的卷积神经网络.自动化学报,2016,42(9):1300–1312.

薛子育(1992－),男,北京人,助理工程师,CCF 学生会员,主要研究领域为机器学习,计算机图形学,图像理解.

郭沛宇(1973－),男,教授级高级工程师,主要研究领域为数字媒体,数字版权保护,广播电视人工智能应用.

祝晓斌(1981－),男,副教授,CCF 专业会员,主要研究领域为多媒体数据挖掘,机器学习.

张乃光(1978－),男,高级工程师,主要研究领域为数字媒体,数字版权保护,广播电视人工智能应用.