DOI: 10.16452/j.cnki.sdkjzk.2023.03.010 文章编号:1672-3767(2023)03-0085-09

自动生成影像学报告的混合特征提取无卷积结构深度学习模型

王 瑞,花 嵘,仪秀龙,韩承磊
(山东科技大学 计算机科学与工程学院,山东 青岛 266590)

摘 要:在影像学报告的生成中,由于正常区域和异常区域的数据不平衡,描述疾病的关键词经常被描述图像正常区域的句子掩盖,导致异常图像特征的误判和漏判,严重影响医疗报告的质量。本研究提出混合特征提取无卷积深度学习模型,首次将 Swin Transformer 引入放射学报告中,设计了一个混合特征提取器,以提取更加细粒度的图像特征,准确地捕捉生成影像学报告所需要的异常特征;设计一个名为视觉-语义协同注意力的注意力机制,在生成报告时突出图像重点特征信息,对非关键信息进行过滤,有效提升生成异常报告的质量;使用具有记忆机制的解码器模块生成影像学报告。最后,在流行的影像学报告 IU X-Ray 数据集上与当前的主流模型进行对比表明,本模型在语言生成指标和临床评估方面都达到较理想的效果。

关键词:影像学报告; 混合特征;多标签;标签特征;深度学习模型
中图分类号:TP391.41 **文献标志码**:A

A hybrid feature extraction deep learning model without convolutional architecture for automatic imaging report generation

WANG Rui, HUA Rong, YI Xiulong, HAN Chenglei
(College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China)

Abstract: In the generation of imaging reports, due to the imbalance of data between normal areas and abnormal areas, keywords describing diseases are often covered by sentences describing normal areas of images, leading to the misjudgment and missing judgment of abnormal image features, which seriously affects the quality of medical reports. This study proposed a convolution-free deep learning model for hybrid feature extraction. In this model, Swin Transformer was introduced into radiology reports for the first time, and a hybrid feature extractor was designed to extract more fine-grained image features and accurately capture the abnormal features required for generating imaging reports. An attention mechanism named visual-semantic collaborative attention was designed to highlight the key feature information of the image and filter the non-critical information during the report generation, thus effectively improve the quality of the abnormal report generation. Imaging reports were generated by using decoder modules with memory mechanisms. Finally, an experimental comparison was made between the proposed model and the current mainstream models in the IU X-Ray dataset of popular imaging report. The results show that the proposed model achieves better results in language generation index and clinical evaluation.

Key words: image report; mixed feature; multi-tag; tag feature; deep learning model

收稿日期:2023-03-07
基金项目:山东省自然科学基金项目(ZR2022MF274)

作者简介:王 瑞(1999—),男,山东青岛人,硕士研究生,研究方向为深度学习与分布式计算.
花 嵘(1969—),男,江苏常州人,副教授,博士,研究方向为高性能计算、深度学习,本文通信作者.
E-mail:huarong@sdust.edu.cn

医疗报告自动生成是图像描述任务的一个分支,图像描述[1]旨在理解给定的图像,并生成相应的描述性语句,近年来已进行了一些研究并取得了较多成果。放射学图像属于医疗图像的一种,在医生诊断过程中,通常会参考放射学医生给出的影像学报告来辅助判断患者的具体情况。图像生成影像学已成为计算机科学在医学领域应用的一个重点研究方向。

一份完整的肺部放射学医疗报告由作出诊断的印象(Impression)、描述放射学观察的发现(Findings)以及机器根据描述生成的标签(Tags)三部分组成,如图 1 所示(数据来源:https://openi.nlm.nih.gov/gridquery?q=Indiana chest X-ray collection&it=xg)。目前医疗报告的生成主要是通过深度学习模型模拟人类书写报告的过程来实现,大多采用编码器-解码器(encoder-decoder)结构[2-3],大量实验证明这种结构可以得到较为满意的效果。在该结构中,编码器通常由一系列卷积神经网络(convolutional neural networks,CNN)组成,例如 RestNet[4]、DenstNet[5]等模型;解码器通常基于循环神经网络(recurrent neural network,RNN)的模型,例如分层长短期记忆网络(long short term memory,LSTM),可以在句子级和单词级进行解码,生成规范报告。为了进一步提高对图片重点区域的关注,Xu 等[3]将注意力机制(attention mechanism)引入图像字幕,使图片不同区域具有不同权重;Anderson 等[6]提出一种协调工作的注意力机制,使模型可以准确判断需要重点注意的区域并且在生成不同单词时关注相应的区域。此外还有通过优化编码器中注意力机制来提高报告质量的方法[7]。

由于 RNN 结构不适应长依赖关系,在医疗报告这种由多个句子组成的长叙述中难以达到预期目标,人们开始尝试使用 Transformer[8]作为解码器,用来解决循环神经网络的弊端。但目前将 Transformer 结构直接用作编码器的研究还较少,特别是在图像描述领域,纯 Transformer 架构的研究尚不成熟。

Impression: No acute cardiopulmonary abnormality.

Findings: There are no focal areas of consolidation. No suspicious pulmonary opacities. Heart size within normal limits. No pleural effusions. There is no evidence of pneumothorax. Degenerative changes of the thoracic spine.

Tags: degenerative change

图 1 IU X-Ray 数据集中一个标准的肺部放射学医疗报告

Fig. 1 Medical reports of lung radiology in the IU X-Ray dataset

受图像字幕领域研究的启发,许多自动生成医疗报告的框架被提出[9-18]。但是,医疗领域严重的数据偏差导致许多使用分层 LSTM 模型[12-14]生成了多个重复描述正常区域的句子,不能满足医疗领域中需要强调异常描述的要求。文献[10,13,15]和文献[12,16]分别引入了医学先验知识图和强化学习。文献[17]对正常图像和异常图像之间的差别进行了对比,从而更好地识别异常情况。文献[18]根据某些医生书写报告的场景,引入模板生成医疗报告。

尽管目前自动生成影像学报告的研究取得了一定的成果,但是由于图像中正常区域占很大比例,且正常图片占比高,这种数据不平衡导致描述疾病的单词通常会被正常的单词覆盖,在某些情况下还会导致异常区域被生成正常的描述,出现错判、漏判,影响生成的影像学报告质量。针对上述问题,文献[9]使用医学标签来提高对疾病的判断能力,文献[10]使用知识图谱进行预编码,使模型能够了解不同疾病之间的关联,提高判断能力。但上述文献均未使用 Transformer 框架作为解码器,导致后生成的序列单词过分依赖先生成的单词,不适于长叙述的医疗报告生成。文献[11]使用内存模块和归一化层来增强特征信息,可以平衡生成序列的前后依赖,但由于未对标签语义特征进行提取,无法针对数据不平衡问题进行相应的优化。

为了更好地解决自动生成影像学报告中出现的误判、错判等问题,本研究提出一个混合特征提取无卷积结构深度学习模型,设计了一个基于 Swin Transformer 的混合特征提取器(hybrid feature extractor,HFE),对影像学图像进行更细粒度的特征提取,通过分层特征映射,提高提取图片细粒度特征的能力和标签特征质量,从而进一步提高疾病单词的准确性;提出一个新的视觉-语义协同(visual-semantic cooperative,VSC)注意力机制,可以有效处理混合特征,根据视觉-语义混合特征来优化所需解码的特征;使用具有记忆功能的归一化层,通过记忆矩阵来增强模型学习过的特征。

1　混合特征提取无卷积结构深度学习模型

1.1　模型总体结构

本研究提出的混合特征提取无卷积结构深度学习模型总体结构如图 2 所示。其中,Encoder 表示模型的编码器,Decoder 表示模型的解码器,Visual Feature 表示混合特征提取器提取到的视觉特征,Sematic Feature 表示混合特征提取器提取到的语义特征,Ground Truth Feature 表示从真实报告中提取到的文字特征。首先将需要生成报告的图像划分为多个不同区域,然后对这些区域使用视觉特征(Visual Feature)提取器批量地提取视觉特征,同时标签特征提取器提取标签的语义特征(Sematic Feature)。视觉特征经过编码器编码后,两种特征同时输入视觉-标签协同注意力模块,进入解码器,此时注意力模块处理后的特征进入记忆驱动的归一化层(memory-driven conditional layer normalization,MCLN)进行修饰。最后经过一个全连接网络(full-connected network,FC)进行概率预测,自动产生相应报告。

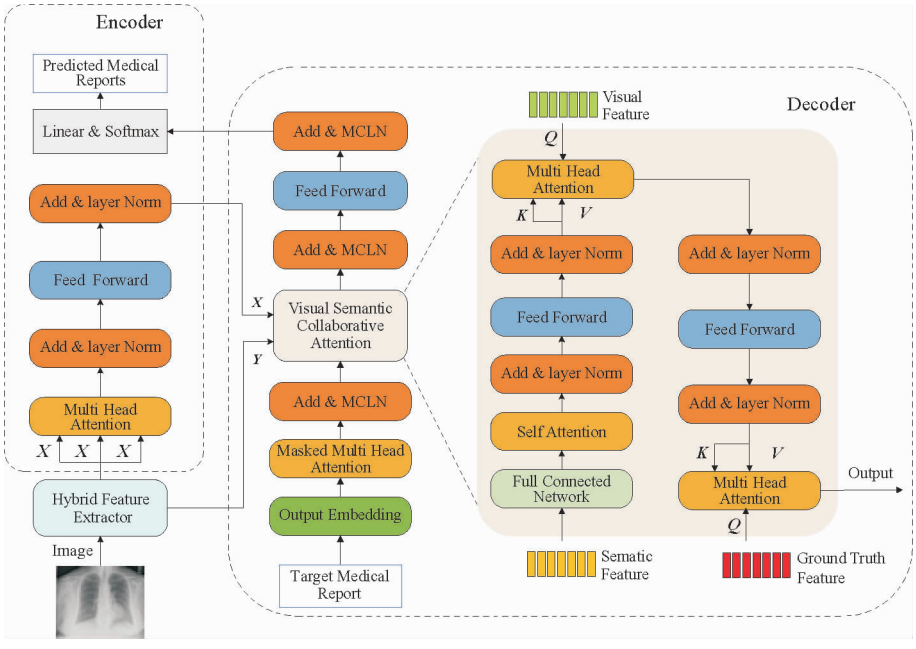

图 2　混合特征提取无卷积结构深度学习模型总体结构

Fig. 2　Overall structure of hybrid feature extraction deep learning model without convolutional architecture

1.2　基于 Transformer 的混合特征提取器

本研究提出的模型采用基于 Transformer 的新型混合特征提取器(结构如图 3 所示)。对图片进行特征提取,通过滑动窗口机制更好地适用于影像学图像。其中,标签语义特征提取器(tag semantic feature extractor)可以尽可能地提取深层次的图像信息,LFE(label feature extractor)表示标签特征提取器。

用 $I \in \mathbf{R}^{C\times H\times W}$ 表示 1 张输入图像(C 表示图像特征图的深度,H 表示图像高度,W 表示图像宽度),图像经过分块后经过 4 个阶段处理。首先进入由 1 个线性嵌入(Linear Embeding)层和两个 Swin Transformer Block 组成的模型阶段Ⅰ,另外 3 个阶段均由 1 个补丁合并(Patch Merging)层和不同数量的 Swin Transformer Block 组成。基于 Transformer 的新型混合特征提取器的各部分工作过程如下。

1) 补丁分割(Patch Partition)层。对输入图像 I 进行分块操作。若使用的补丁大小是 $n\times n$,则每 $n\times n$ 相邻的像素形成一些大小相同的二维补丁,每个补丁的特征维度为 $n\times n\times C$,补丁的数量为 $H/n\times W/n$。

2) 线性嵌入层。对每个像素的通道数进行线性变换,将补丁的特征维度投射到合适的维度(记为 D),即图像特征图维度由[H/n,W/n,$n\times n\times C$]转换为[H/n,W/n,D]。

3) 补丁合并层。在阶段Ⅱ、Ⅲ和Ⅳ中，图像特征图通过 Patch Merging 模块进行下采样。以阶段Ⅱ为例，输入 Patch Merging 大小为[H/n,W/n,D]的图像特征图，Patch Merging 将每个大小为 $H/2n\times W/2n$ 的相邻像素划分为 1 个补丁，然后将每个补丁中相同位置像素拼接在一起，得到 4 个图像特征图。然后将这 4 个图像特征图在深度方向进行拼接，再通过 1 个归一化层(Layer Norm，LN)和 1 个全连接层在图像特征图的深度方向进行线性变换，将图像特征图的深度由 D 变成 $2D$。通过 Patch Merging 层后，图像特征图的高和宽减半，深度翻倍，即阶段Ⅱ中图像维度由[H/n,W/n,D]转换为[$H/2n$,$W/2n$,$2D$]。

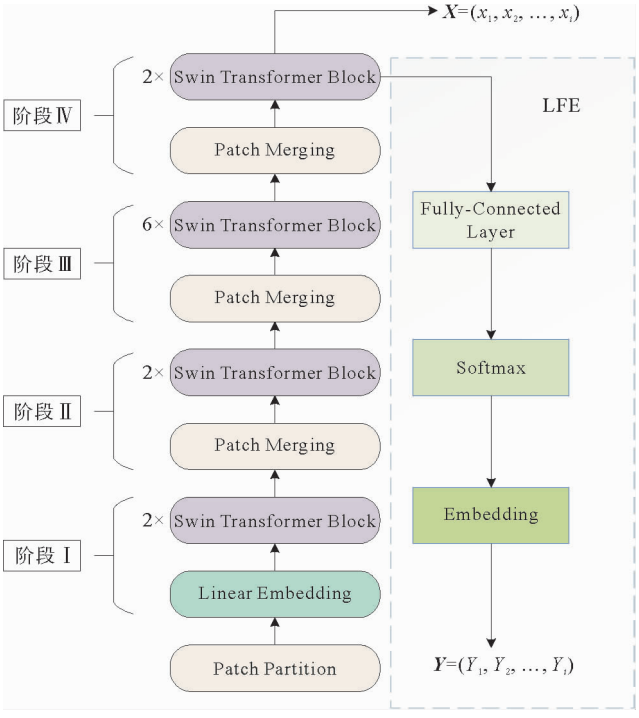

图 3 混合特征提取器结构

Fig. 3 Hybrid feature extractor structure

4) Swin Transformer 块。图像数据经过线性嵌入层或补丁合并层后被送入 Swin Transformer Block 模块。如图 4 所示，每个 Swin Transformer Block 包含 1 个 W-MSA (windows multi-head self-attention)模块或 SW-MSA(shifted windows multi-head self-attention)模块，之后有 2 层 MLP。在每个 MSA 模块和每个 MLP 之前有 1 个归一化层。W-MSA 与 SW-MSA 模块成对使用，先使用 1 个 W-MSA 模块，再使用 1 个 SW-MSA 模块。与常见的卷积模块相比，W-MSA 模块可以深化细粒度特征，使模型更易于捕捉影像学图像中的异常细节，同时减少计算量。与普通的 MSA 模块将特征图里的每个像素和全部像素进行计算不同，使用 W-MSA 模块时，首先将图像特征图划分成若干 $M\times M$($M\in\mathbf{N}$)大小的窗口，然后对每个窗口内部单独进行自注意力计算。MSA 与 W-MSA 模块的计算量分别为：

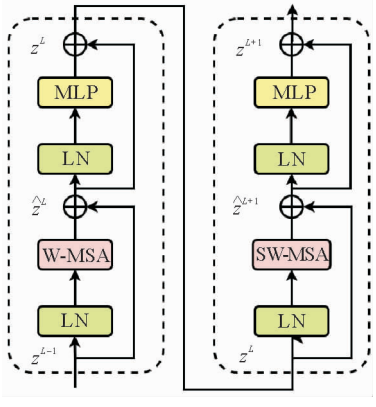

图 4 Swin Transformer Block 模块内部细节

Fig. 4 Internal details of Swin Transformer Block

$$\Omega(\mathrm{MSA})=4HWC^2+2(HW)^2C\ , \tag{1}$$

$$\Omega(\mathrm{W\text{-}MSA})=4HWC^2+2M^2HWC\ 。 \tag{2}$$

在 W-MSA 模块中，特征图只会在每个窗口内进行自注意力计算。为了在滑动窗口中进行计算，本研究使用 SW-MSA 模块。如图 5 所示，窗口从左上角分别向右侧和下方各偏移 $M/2$ 个像素，偏移后的窗口($L+1$)再进行运算时可以使用偏移前的窗口(L)信息。连续的 Swin Transformer Block 可表示为：

$$\hat{z}^L=\mathrm{W-MSA}(\mathrm{LN}(z^{L-1}))+z^{L-1}\ , \tag{3}$$

$$z^L=\mathrm{MLP}(\mathrm{LN}(\hat{z}^L))+\hat{z}^L\ , \tag{4}$$

$$\hat{z}^{L+1}=\mathrm{SW-MSA}(\mathrm{LN}(z^L))+z^L\ , \tag{5}$$

$$z^{L+1}=\mathrm{MLP}(\mathrm{LN}(\hat{z}^{L+1}))+\hat{z}^{L+1}\ 。 \tag{6}$$

式中，$\hat{z}^L$ 和 z^L 分别表示块 L 的 W-MSA/SW-MSA 模块和 MLP 模块的输出。最后，阶段Ⅳ输出的图像特征 $\boldsymbol{X}$ 的维度为[$H/8n$,$W/8n$,$8D$]。

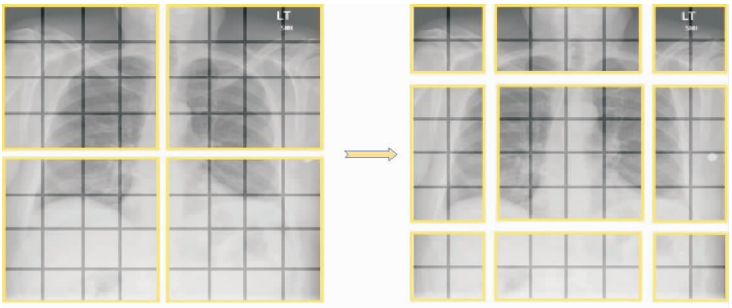

图 5　窗口滑动过程

Fig. 5　Window sliding process

通过 Swin Transformer Block 模块可以得到视觉特征矩阵 $\boldsymbol{X}$($\boldsymbol{X}\in\mathbf{R}^{Lx\times S}$)。为了得到标签语义特征,本研究将视觉特征矩阵 $\boldsymbol{X}$ 输入深度分类特征提取器(图 3)来提取语义特征。LFE(图 3)可以表示为:

$$\mathrm{LFE}=\mathrm{Em}[S_{\max}(\mathrm{FC}(\boldsymbol{X}))]\ 。\tag{7}$$

式中:Em 表示 Embedding 操作,$S_{\max}$表示 softmax 激活函数,FC 表示全连接网络。

1.3　特征编码器

本研究使用 Transformer[8] 的编码器进行编码操作,其中前馈(Feed Forward)模块挖掘视觉特征得到深层特征,可以更深入分析视觉特征的细节差异。然后通过归一化层加速模型收敛,提升模型运行速度。本模块可以表示为:

$$\boldsymbol{Q}=\mathrm{Fe}(\boldsymbol{X})\ 。\tag{8}$$

式中:Fe 表示编码操作,输入为混合特征提取器提取的视觉特征矩阵 $\boldsymbol{X}$,输出为视觉特征的隐藏状态 $\boldsymbol{Q}$。

1.4　特征解码器

本研究对 Transformer 进行改进,设计了全新的视觉-语义协同注意力模块(图 2),并且采用 Chen 等[11] 提出的 MCLN 增强生成过程中信息的记忆。首先输入正确的文本信息给掩盖多头注意力(masked multi-head attention)层进行有掩盖的注意力计算,再进入 MCLN。MCLN 中集成了文献[8]提出的关系内存模块,以充分利用之前的记忆信息,同时也避免其影响生成文本的核心信息。存储器驱动条件层规范化表达式为:

$$\hat{\gamma}=\gamma+f_{\mathrm{mlp}}(m)\ ,\tag{9}$$

$$\hat{\beta}=\beta+f_{\mathrm{mlp}}(m)\ ,\tag{10}$$

$$f_{\mathrm{mcln}}(r)=\hat{\gamma}\odot\frac{r-\mu}{\upsilon}+\hat{\beta}\ 。\tag{11}$$

式中:γ 和 β 分别为初始的缩放参数和平移参数,$\hat{\gamma}$ 和 $\hat{\beta}$ 分别为当前关系内存修饰后的缩放参数和平移参数,m 为关系内存的输出,f_{mlp}代表多层感知器,r 为前一模块的输出,μ 和 υ 分别为 r 的均值和标准差,$\odot$代表哈达玛积。根据图 2 中 MCLN 的位置,本研究将来自 MCLN 的结果 $f_{\mathrm{mcln}}(r)$输入到下一个模块或作为最终输出。

与此同时,将解码器的输出作为视觉特征、LFE 的输出作为语义特征以及经过记忆功能的归一化层的文本信息一并输入视觉-语义协同注意力模块进行解码。视觉-语义协同注意力模块的具体细节在 1.5 节中详细说明。本模块可以表示为:

$$e_t=\mathrm{Fd}(\mathrm{VSC}(\boldsymbol{Q};\boldsymbol{Y};e_1,e_2,\cdots,e_{t-1}))\ ,i\in\mathbf{R}\ 。\tag{12}$$

式中:Fd 表示解码操作;VSC 表示视觉-语义协同注意力模块;$\boldsymbol{Y}$ 为语义特征;$\boldsymbol{E}=(e_1,e_2\cdots,e_t)$为目标文本序列,$t$ 为序列数。文本特征经解码器解码后输入线性层输出最终报告。

1.5 视觉-语义协同注意力模块

由于混合特征提取器产生了语义特征,根据本模型的特性提出新的注意力模块——视觉-语义协同注意力模块,如图 2 所示。视觉-语义协同注意力模块有 3 个输入接口,分别接收混合特征提取器提取到的视觉特征、语义特征以及来自真实报告的特征。为解决视觉特征与语义特征之间的张量对齐问题,将语义特征送入全连接网络进行维度转换,学习不同维度之间的映射关系,然后对其进行自注意力编码和前馈层提取,得到深度信息,在每一层后添加归一化层加快模型收敛。

由于 MHA 的特殊机制可以用于计算多特征的关联权重,因此本研究将语义特征和视觉特征输入 MHA 进行计算,对不重要的视觉特征进行过滤,结果经新的前馈层后与真实报告的特征进行新的 MHA 计算。由此,VSC 模块可以表示为:

$$\mathrm{VSC}(\boldsymbol{Q},\boldsymbol{Y},\boldsymbol{E})=\mathrm{MHA}(\mathrm{MHA}(\boldsymbol{Q},\mathrm{FC}(\boldsymbol{Y})),\boldsymbol{E})\ 。 \tag{13}$$

通过这种方式,VSC 模块将解码中间结果输出给解码器以生成最终结果。

2 实验设计

为验证所提模型的有效性,进行消融实验验证本模型两个模块的作用,并与当前的主流模型进行性能对比。为更好地了解本模型生成报告的质量,将模型用于具体医学案例进行进一步分析。

2.1 数据集

实验采用印第安纳大学收集的公共放射学 IU X-Ray 数据集,其中包括 7 470 张胸部 X 光图像和 3 955 份报告。本研究统计了数据集中出现频次最高的单词(图 6),发现抽象单词的频次较多。在对数据的处理中,根据当前主流研究惯例,去除没有报告的图片,并进行相应的数据划分。采用与主流模型相同的 7∶1∶2 比例划分训练/验证/测试集,并将所有的字母转换为小写、剔除特殊符号等影响因素。

2.2 基线和评价指标

本研究使用的 Baseline 模型为 Vanilla Transformer,包含 3 层、8 个头和 512 个隐藏单元,没有进行其他扩展和修改,详细描述见文献[8]。Base+HFE 模型是本研究提出模型的简单替代方案,使用混合特征提取器替换常规的视觉提取器,其中对于新增加的语义特征,将其与视觉特征进行相同的处理,使用编码后的文本特征分别进行多头注意力操作,最后两者数据特征相加进入解码生成报告。

利用传统的自然语言生成(natural language generation,NLG)评价指标,如 BLEU[19]、METEOR[20]和 ROUGE-L[21]等,将本模型与当前的主流模型进行对比。

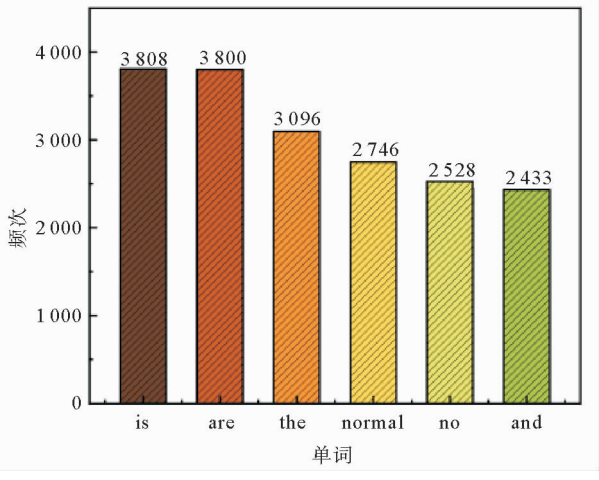

图 6 IU X-Ray 数据集出现频次排行前 5 的单词

Fig. 6 Top 5 words in the IU X-Ray dataset

2.3 实现细节

对于 IU X-Ray 数据集,与当前主流模型的实验设置一致,使用 1 位患者的正面和侧面图像作为输入。消融实验中所有模型参数均采取随机初始化。本模型主体部分在 ADAM 优化器的交叉熵损失下进行训练,LFE 部分采用二元交叉熵损失进行训练。实验将视觉提取器、LFE 及其他参数的学习率分别设置为 1×10^{-4}、5×10^{-4}和 5×10^{-5},每个 epoch 的速率衰减 0.8 倍,Batchsize 大小设置为 8,Beam 大小设置为 3,以提升训练效率。

3 结果和分析

消融实验结果如表 1 所示。其中,Base+HFE+VSC 代表本研究的最终模型,B1、B2、B3、B4 分别是 n-gram 为 1、2、3、4 时的 BLEU 指标。由表 1 可见,与基础模型相比,本模型每项的得分都明显提高。

表 1　基础模型与本研究模型的效果比较

Table 1　Effect comparison between the basic model and the model in this study

模型	B1	B2	B3	B4	METEOR	ROUEGE	提升比例/%
Base	43.9	26.7	18.6	14.0	16.9	35.4	—
Base+HFE	46.2	29.6	21.7	17.0	18.5	35.3	10.6
Base+HFE+VSC	48.8	31.6	23.0	17.8	19.7	37.2	17.0

为研究模型层数对性能的影响,使用不同的 Transformer 层数进行实验,结果如表 2 所示。由表 2 可见,不同的模型层数模型性能不同,且当模型层数为 3 时性能最佳,相较其他层数性能提升较大。

表 2　不同 Transformer 层数的模型实验结果

Table 2　Experimental results of the model with different Transformer layers

层数	B1	B2	B3	B4	METEOR	ROUEGE
1	45.8	29.5	21.5	16.6	18.8	36.7
2	43.2	28.1	20.2	15.3	17.8	35.0
3	**48.8**	**31.6**	**23.0**	**17.8**	**19.7**	**37.2**
4	41.8	26.9	19.8	15.5	17.1	34.8
5	42.9	27.1	19.8	15.4	17.2	34.6

此外,将本模型与当前的主流模型 SentSAT+KG[10]、CMAS-RL[16]、CMCL[17]、R2Gen[11]和 CMN[22]进行自然语言生成指标效果对比,如表 3 所示。对比结果表明,本模型除 ROUEGE 之外的 5 个指标均优于其他主流模型。

表 3　本研究模型和已有模型的自然语言生成指标效果对比

Table 3　Comparison of NLG index effect between the proposed model and the existing models

模型	B1	B2	B3	B4	METEOR	ROUEGE
SentSAT+KG	44.1	29.1	20.3	14.7	—	36.7
CMAS-RL	46.4	30.1	21	15.4	—	36.2
CMCL	47.3	30.5	21.7	16.2	18.6	37.8
R2GEN	47.0	30.4	21.9	16.5	18.7	37.1
CMN	47.5	30.9	22.2	17.0	19.1	37.5
本模型	48.8	31.6	23.0	17.8	19.7	37.2

为更好地了解本模型生成的报告质量,对一些具体医学案例进行测试分析。图 7 展示了 1 位病人的前胸部和侧面胸部图像,图中 Ground-truth report 表示正确的报告,Generated report 表示本模型生成的报告。由图 7 可以看出,Base+HFE+VSC 模型可以提取更加细粒度的图像特征,准确地捕捉生成影像学报告所需要的异常特征,对随机分布的小型病灶,如肺结节、肺结核钙化灶的辨别较好,能够生成与放射科医生基本一致的描述。图 7 中红色部分为本模型成功预测到的医疗词汇,能更好地与指示疾病或部位对齐,表明本模型不仅增强了生成的影像学报告的精度和准确度,而且改善了图像和生成文本之间的关系。

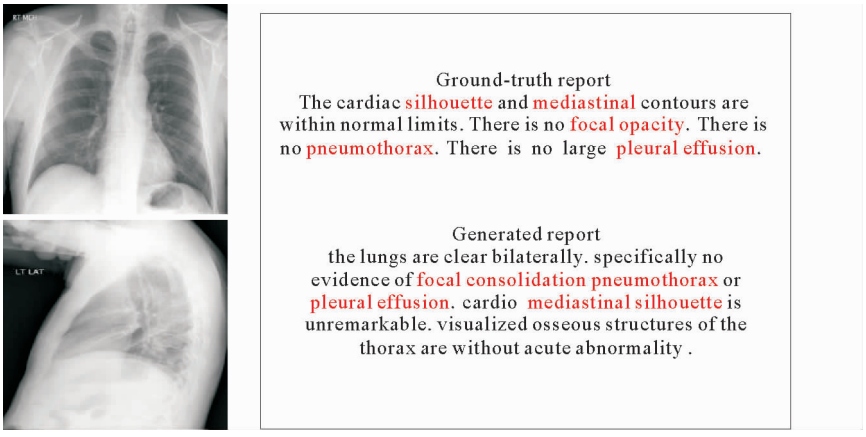

图 7 本研究所提模型生成的医疗报告示例

Fig. 7 An example of a medical report generated by the proposed model

4 结论

本研究提出一种自动生成影像学报告的混合特征提取无卷积结构深度学习模型,设计了一种混合特征提取器和一种协同注意机制来生成影像学报告,可以实现对影像学图像更细粒度的特征提取,并有效过滤不重要的特征,降低正常区域对报告质量的影响,减少错报、漏报。通过在 IU X-Ray 数据集上验证表明,本研究所提模型可以生成较好的医学报告。但是,本研究所提模型仍具有一定局限性,如关键词的预测准确度仍需进一步提高,未来将着力于解决混合特征与图像区域对齐问题,尝试使用注意力机制进行标签加权。

参考文献:

[1] CHEN X,FANG H,LIN T Y,et al.Microsoft COCO captions:Data collection and evaluation server[J].Computer Science,2015(5):1-7.

[2] VINYALS O,TOSHEV A,BENGIO S,et al.Show and tell:A neural image caption generator[C/OL]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.IEEE,2015.DOI:10.1109/CVPR.2015.7298935.

[3] XU K,BA J,KIROS R,et al.Show,attend and tell:Neural image caption generation with visual attention[C/OL]// 32nd International conference on machine learning:ICML.Lile,Jul. 6-11,2015,3(3).DOI:10.48550/arXiv.1502.03044.

[4] HE K M,ZHANG X Y,REN S Q,et al.Deep residual learning for image recognition[C/OL]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.IEEE,2016:770-778.DOI:10.1109/CVPR.2016.90.

[5] HUANG G,LIU Z,VAN DER MAATEN L,et al.Densely connected convolutional networks[C/OL]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.IEEE,2017:4700-4708.DOI:10.1109/CVPR.2017.243.

[6] ANDERSON P,HE X,BUEHLER C,et al.Bottom-up and top-down attention for image captioning and visual question answering[C/OL]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.2018:6077-6086.DOI:10.48550/arXiv.1707.07998.

[7] DEVLIN J,CHANG M W,LEE K,et al.BERT:Pre-training of deep bidirectional transformers for language understanding [C/OL]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies.2019:4171-4186.DOI:10.48550/arXiv.1810.04805.

[8] VASWANI A,SHAZEER N,PARMAR N,et al.Attention is all you need[C/OL].31st Conference on Neural Information Processing Systems (NIPS 2017).Long Beach,2017,30.DOI:10.48550/arXiv.1706.03762.

[9] JING B,XIE P,XING E.On the automatic generation of medical imaging reports[C/OL]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics.2018:2577-2586.DOI:10.18653/v1/P18-1240.

[10] ZHANG Y,WANG X,XU Z,et al.When radiology report generation meets knowledge graph[C/OL]// Proceedings of the AAAI Conference on Artificial Intelligence.2020.DOI:10.1609/aaai.v34i07.6989.

[11] CHEN Z,SONG Y,CHANG T H,et al.Generating radiology reports via memory-driven transformer[C/OL]// Confer-

ence on Empirical Methods in Natural Language Processing (EMNLP).2020:1439-1449.DOI:10.18653/v1/2020.emnlp-main.112.

[12] XUE Y,XU T,LONG L R,et al.Multimodal recurrent model with attention for automated radiology report generation[C]//Proceedings of 21st International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2018),Part I.Springer,2018:457-466.

[13] LIU F L,WU X,GE S,et al.Exploring and distilling posterior and prior knowledge for radiology report generation[C/OL]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021: 13748-13757. DOI: 10.48550/arXiv.2106.06963.

[14] LIU F L,YIN C C,WU X,et al.Contrastive Attention for Automatic Chest X-ray Report Generation[C/OL]//ACL-IJCNLP (Findings) 2021.DOI:10.18653/v1/2021.findings-acl.23.

[15] LI C Y,LIANG X,HU Z,et al.Knowledge-driven encode,retrieve,paraphrase for medical image report generation[C/OL]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(1): 6666-6673. DOI: 10.1609/aaai.v33i01.33016666.

[16] JING B,WANG Z,XING E.Show,describe and conclude:On exploiting the structure information of chest X-ray reports[C/OL]//57th Annual Meeting of the Association-for-Computational-Linguistics (ACL). 2019: 6570-6580. DOI: 10.18653/v1/P19-1657.

[17] LIU F,GE S,WU X.Competence-based multimodal curriculum learning for medical report generation[C/CD]//Joint Conference of 59th Annual Meeting of the Association-for-Computational-Linguistics (ACL).2021:3001-3012.

[18] LI C Y,LIANG X D,HU Z T,et al.Hybrid retrieval-generation reinforced agent for medical image report generation[C/OL]//32nd Conference on Neural Information Processing Systems (NIPS).2018,31.DOI:10.48550/arXiv.1805.08298.

[19] PAPINENI K,ROUKOS S,WARD T,et al.BLEU:A method for automatic evaluation of machine translation[C/OL]//Proceedings of the 40th annual meeting of the Association for Computational Linguistics.2002:311-318.DOI:10.3115/1073083.1073135.

[20] DENKOWSKI M,LAVIE A.Meteor 1.3:Automatic metric for reliable optimization and evaluation of machine translation systems[C/CD]//Proceedings of the Sixth Workshop on Statistical Machine Translation.2011:85-91.

[21] LIN C Y.ROUGE:A package for automatic evaluation of summaries[C/CD]//Proceedings of Workshop on Text Summarization of Branches Out.Barcelona,2004.

[22] CHEN Z,SHEN Y,SONG Y,et al.Cross-modal memory networks for radiology report generation[J/OL].Computers & Electrical Engineering,2022,98:1879-0755.DOI:10.18653/v1/2021.acl-long.459.

(责任编辑:齐敏华)