聚知刊出版社
JZK publishing

基于视觉——语言模型的医学影像描述文献综述

宋莲君

天津财经大学珠江学院，天津，301811；

摘要：医学图像辅助诊断是指使用计算机技术来分析医学成像数据，旨在从医学图像中提取有用的特征信息，以帮助诊断疾病和评估治疗效果。自动理解医学图像是新型的人工智能的一个分支，结合了医学、计算机视觉和自然语言处理领域相关知识。本篇文章将回顾医学影像描述的各种方法，将其分类为基于模板的方法、基于检索的方法、基于生成的方法其中包括编码器-解码器模型和基于注意力机制的模型等。

关键词：深度学习；计算机视觉；医学影像描述；自动图像字幕；报告生成

DOI:10.69979/3041-0673.24.6.024

引言

医生通过医学数字成像和通信系统来存储医疗影像中的病理信息并根据标准自动生成结构化报告。随着图像及视频描述、视觉问答等多模态理解和生成任务的不断发展，以及智能医疗的兴起，医学图像描述技术开始用于自动生成医学图像的描述，例如由 CT 图像自动生成放射学诊断或诊断草稿等，计算机辅助生成报告的方法可以帮助医生定位图中感兴趣的区域或是给予医生基本的判断，减少医生的错误。

在通用图像处理中视觉-语言模型首先会用某些现成的大规模训练数据对深度神经网络进行预训练，然后用特定任务的带注释训练数据对预训练模型进行微调，这种模型训练的方式可以为各种下游任务做好前期工作。目前所存在的医学领域的图像描述的综述研究[1]中也提到了这些方法。下文将梳理近五年医学影像领域结合人工智能领域的研究最新进展。

1 应用于医学图像处理的视觉-语言模型

1.1 基于模板和检索的视觉-语言模型

最早出现 MIC (Medical images captioning) 是 Shin 等人[2] 利用医学主题词表标签发表的有关于 X 光胸片的一篇研究。首先图像中的单一疾病标签来训练 CNN 模型，再使用 RNN 来获得位置信息，再生成有单个单词或词组组合成的候选标签，利用最后候选标签重新送入 CNN-RNN 模型，生成完整标签，例如右上叶钙化肉芽肿。但生成结果是一系列术语，例如 X 光片对应的疾病名称、位置、严重性和影响器官等，而不是连贯流畅的医学报告。

由于对信息准确描述的标准越来越高，单一的短句描述不足满足当下的医疗现状，后续的研究者使用了模板的方式，将生成短句嵌入到已有的模板中，获得大篇幅的描述。其代表性的创新有 Gale W [3] 的专注于对骨盆 X 光片中的髋部骨折的研究。 基于检索的模型是依赖于相似图像的假设具有相同的标题。Ayesha 等人[4] 对于每个新颖的图像，一组在视觉上类似于在第一步中，从大型数据集中检索查询图像。然后，做出两个选项可用时，要么将最相似图像的标题分配给新颖图像，或者利用候选字幕并组合以生成基于一些预先定义的规则和方案。

1.2 基于生成式的视觉-语言模型

编码器-解码器的架构是研究医学图像描述的主干架构。一般来说，CNN 被用来图像编码器来产生固定长度的矢量表示，应用 RNN 模型来解码该表示并生成描述性的语言。大多文章中都采用了目标检测算法对超声图像中的病变区域进行检测，并将病变区域编码为编码向量，利用语言生成模型 B 对编码向量进行解码，得到诊断报告。这种方法可以实现从患者的图片信息转换为文本描述。

医学影像报告任务中，当生成文本时运用注意力机制可以在生成不同单词时自动学习为不同的图像区域分配权重。Wang [5] 等人提出了 Tienet，是一种多级注意力文本图像嵌入模型，将临床自由文本放射学报告作为先验知识，该网络融合了生成文本的注意分布图和嵌入信息，提高了分类的准确性。

2020 年以前的研究者普遍关注于生成文本是否符合人类语言习惯，在 Zeng 等人[6]的工作中笔者自称是第一篇考虑到生成的诊断报告中病理信息准确性的工作。在 2020 年以后的工作中，更加关注于疾病标签、文本报告和图像之间的一致性，使不同模态的信息表示在同一个维度中，这其实是一个从粗放照搬到针对于任

务精细化打磨的过程。

在 Li 等人[7]的文章中，引用到了图结构（图 1），文中定义医疗标签图即图中的每个节点代表一个检测到的医疗标签；每个节点的特征是医疗标签的分类概率；每对节点之间的相关性表示为边权重。使用图结构应用非常广泛，但由于目前在医学报告自动生成领域所能采集到的数据是有限的，因此构建的图网络也通常仅仅包含几十个节点，目前可用的数据集无论从疾病种类还是病例数量来说都不足以支持在该任务上更大规模医学知识图谱的建立。

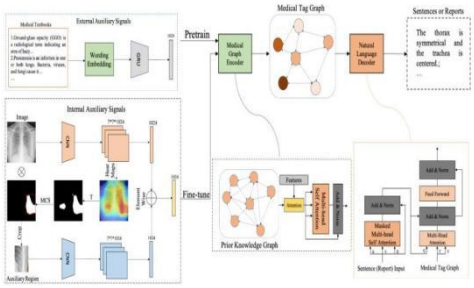

图. 1. 使用图结构存储疾病标签

此外，在对与以往工作进行回顾，发现使用因果模型的思想解决疾病标签预测这一任务的实验有若干篇。比如在 Chen 等人[8]的工作中，不同颜色的实线框表示一种器官，在图 2.(a)中 A 是心脏，B 是肺，在图 2.(b)中是另一位患者的心肺位置。使用因果模型能够判断是 A 引起 B 的问题、B 引起 A 的问题或者 C 共同使得 AB 致病。

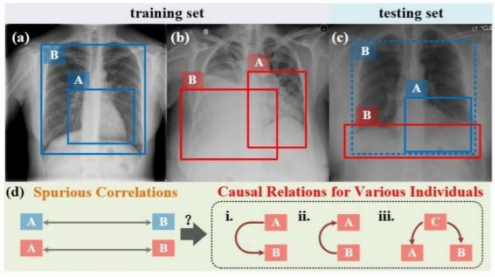

图. 2. 在 IU X-ray 数据集上的使用视觉语言因果干预（VLCI） 框架检测到的结果

2 结论与展望

随着近年来视觉-语言模型技术的快速发展，其在医学影像描述方面的应用展现出巨大潜力。尽管存在一些尚未解决的问题，但通过加强跨领域协作、优化算法结构并重视模型可解释性的提升。本文针对国内医学图像描述领域综述不足的问题，总结了最新的研究进展和结果，按照模型架构的不同，将医学图像描述模型分为 5 个部分，总结各类方法的优劣。尽管近年来医疗图像描述已经被越来越多的研究者所研究，并取得了不错的效果，但相关技术仍处在实验室研究阶段，尚未达到临床落地水平，未来深度医疗图像描述还有很多挑战。

参考文献

[1] 朱翌，李秀，医学图像描述综述:编码、解码及最新进展[DOI：10.11834/jig.21102]

[2]Shin H C,Roberts K,Lu L, Demner-Fushman D, Yao J H and Summers R M. 2016 Learning to read chest X-rays: recurrent neural cascade model for automated image annotation

[3]Gale W, Oakden-Rayner L, Carneiro G, Palmer L J and Bradley A P. 2019. Producing radiologistquality reports for interpretable deep learning//The 16th IEEE International Symposium on Biomedical Imaging. Venice, Italy: IEEE: 1275-1279 [DOI：10.1109/ISBI.2019.8759236]

[4]Ayesha H, Iqbal S, Tariq M, et al (2021) Automatic medical image interpretation: state of the art and future directions. Pattern Recognition, p 107856

[5]Wang. X, Peng. Y, Lu. L, Tienet: text-image embedding network for common thorax disease classification and reporting in chest x-rays, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, p. 9049-9058.

[6] Zeng X, Wen L, Liu B et al (2020) Deep learning for ultrasound image caption generation based on object detection. Neurocomputing 392: 132 - 141

[7]Li M, Wang F, Chang X, Liang X Auxiliary Signal-Guided Knowledge Encoder-Decoder for Medical Report Generation

[8]Chen W, Liu Y, Wang C, Zhu J, Li G, Lin G Cross-Modal Causal Intervention for Medical Report Generation

作者简介：宋莲君（1998 年 1-），女，汉，北京，天津财经大学珠江学院数据工程学院信息科学与技术系教师，医学影像处理。