孙韩科. 基于跨模态自适应学习的医学影像报告生成技术[J]. 智能计算机与应用,2025,15(4):108-114. DOI:10.20169/j.issn.2095-2163.24091601

基于跨模态自适应学习的医学影像报告生成技术

孙韩科
(上海理工大学 健康科学与工程学院, 上海 200093)

摘 要: 在医学影像自动化分析领域,对精确诊断和治疗规划的需求推动了放射学报告自动生成技术的发展。传统方法在协调视觉与文本信息、处理多模态数据方面显示出局限。本研究提出了综合放射学报告与自适应学习网络(Integrated Radiology Report and Adaptive Learning Network, IRRAL-Net),该网络通过跨模态自适应记忆网络和多尺度自适应注意力机制,加强视觉和文本信息的交互,优化视觉提取性能,并实现了特征间的紧密融合。IRRAL-Net 在 IU X-Ray 和 MIMIC-CXR 数据集上的实验显示,相比传统 CMN 模型,其 *BLEU* - 4 得分分别提升了约 10% 和 4.72%,在临床相关性和诊断精度上均优于现有方法,为放射科医生提供了更准确、可靠的诊断工具。

关键词: 放射学报告生成; 医学影像分析; 深度学习网络; 多模态数据融合; 自适应学习机制
中图分类号: TP18 **文献标志码**: A **文章编号**: 2095-2163(2025)04-0108-07

Cross-modal adaptive learning for radiology report generation

SUN Hanke
(School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: In the field of automated medical imaging analysis, the demand for precise diagnostics and treatment planning has spurred the development of radiology report generation technologies. Traditional methods show limitations in synchronizing visual and text information and handling multimodal data. This study introduces the Integrated Radiology Report and Adaptive Learning Network (IRRAL-Net), which employs a cross-modal adaptive memory network and multi-scale adaptive attention mechanisms to enhance the interaction between visual and text information, optimize visual extractor performance, and achieve tight integration of features. Experiments on the IU X-Ray and MIMIC-CXR datasets demonstrate that IRRAL-Net outperforms the conventional CMN model, with *BLEU* - 4 scores increasing by approximately 10% and 4.72% respectively, surpassing existing methods in clinical relevance and diagnostic accuracy, thus providing radiologists with more accurate and reliable diagnostic tools.

Key words: radiology report generation; medical image analysis; deep learning networks; multimodal data fusion; adaptive learning mechanisms

0 引 言

随着医学诊断领域对精细度和描述性要求的提升,整合图像字幕技术与放射学报告生成(Radiology Report Generation, RRG)已显得尤为重要,标志着技术进步的新阶段。Vinyals 等学者[1]利用神经网络(Neural Networks, NN)生成描述性字幕,为人工智能(AI)在图像理解领域设定了新的标准。Xu 等学者[2]引入的注意力机制不仅显著提高了文本与图像内容的相关性,而且在精确的医学成像分析中发挥了重要作用。

然而,尽管存在这些技术进展,现有模型在捕捉医学图像中的细节变化上仍然存在不足,这常常导致生成的报告偏重于整体观察而忽略具体细节。为应对这一挑战,TieNet 模型通过融合注意力编码的文本嵌入和显著性加权池化技术,显著提升了分类和报告生成的准确性[3]。最新的记忆网络技术,如 Chen 等学者[4]的 R2Gen 模型和 Banino 等学者[5]的 MEMO 模型,通过在解码器中整合关系记忆或将结构化记忆整合到 AI 任务中,有效地提高了报告的逻辑性和信息准确性。

作者简介: 孙韩科(1999—),男,硕士研究生,主要研究方向: 自然语言处理,多模态医学影像处理。Email:withhksun@gmail.com。

收稿日期: 2024-09-16

本研究旨在通过开发一种名为综合放射学报告与自适应学习网络(IRRAL-Net)的新型架构,旨在进一步推动自动化放射学报告生成技术的发展。IRRAL-Net 融合了医学变压器和交互式区域引导报告生成方法,不仅提高了临床报告的相关性和诊断精度,还通过跨模态记忆网络增强了视觉与文本间的协同,预期将显著提升报告质量和准确性,从而优化医疗诊断支持[6-9]。

研究给出的胸部 X 光图及其相关报告如图 1 所示。在图 1 中,通过不同颜色高亮显示了视觉和文本特征,直观地揭示了视觉信息与文本内容生成之间的关联性,进一步说明了本研究的创新点与应用前景。

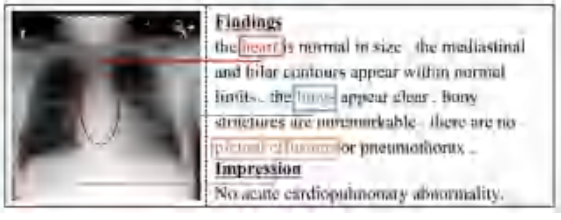

图 1　胸部 X 光图及其相关报告

Fig. 1　Chest X-ray and its related report

1　IRRAL-Net 概述

1.1　任务及问题定义

每幅图像均附有一个类别标签 Y, 该标签显示图像中是否存在特定的放射学特征,其中 $Y_i = 0$ 或 1 代表类别。此外,模型还依据一组由专家撰写的放射学报告 $\boldsymbol{W}^* = \{\boldsymbol{w}_1^*, \boldsymbol{w}_2^*, \cdots, \boldsymbol{w}_R^*\}$ 作为训练的基准。

1.2　特征提取与整合

1.2.1　视觉特征提取

利用卷积神经网络(CNN)作为视觉编码器,从放射学图像中提取视觉特征。这些特征通过可学习的仿射变换参数 $\boldsymbol{W}^E$ 做进一步处理,最终形成视觉特征向量 $\boldsymbol{V}$:

$$\boldsymbol{V} = CNN(Img)\,\boldsymbol{W}^E \tag{1}$$

此流程确保从图像中抽取关键信息,为特征融合步骤提供必要的输入[10-12]。

1.2.2　文本特征提取

采用 BERT 模型从专家撰写的参考报告中提取文本特征。BERT 模型利用其双向编码器功能,深入挖掘语义信息,此信息随后用于指导视觉特征的学习与整合[13]。通过这种方式,文本特征能够与视觉特征协同工作,以增强模型生成的放射学报告在临床相关性及语言精确性方面的表现。文本特征 $\boldsymbol{T}$ 由 BERT 模型输出,计算公式如下:

$$\boldsymbol{T} = BERT(\boldsymbol{W}^*) \tag{2}$$

其中, $\boldsymbol{W}^*$ 是参考报告的文本。

接着,文本特征向量 $\boldsymbol{E}_{\text{txt}}$ 通过处理 BERT 输出的特定标记(如 CLS 标记)得到,计算公式为:

$$\boldsymbol{E}_{\text{txt}} = \sigma(\boldsymbol{U} \cdot \boldsymbol{T}_{[CLS]} + \boldsymbol{C}) \tag{3}$$

其中, $\boldsymbol{U}$ 表示权重矩阵; $\boldsymbol{C}$ 表示偏置项。激活函数 σ 被用来进一步处理这些特征,从而提取出更为丰富的语义层面信息。

1.3　多模态特征对齐

多模态特征对齐技术(Multimodal Feature Alignment, MFA)旨在优化和调整放射学图像与文本特征之间的交互,增强模型在生成放射学报告时的准确性和一致性。此技术通过动态调整特征间的交互性来优化信息整合,确保视觉与文本数据之间的有效融合[14]。为了确保模型可以有效地融合来自图像和文本的信息,引入了一种改进的损失函数来优化特征对齐过程。该损失函数 L_{Align} 定义如下:

$$L_{\text{Align}} = \frac{1}{N_w}\sum_{i=1}^{N_w}[\log p(w_i \mid \boldsymbol{V}; \boldsymbol{M}^S) + \lambda \max(0, T - \cos(\boldsymbol{Z}_{\text{img}} \cdot \boldsymbol{Z}_{\text{txt}}))] \tag{4}$$

该损失函数包括 2 部分:首先,计算给定视觉输入 $\boldsymbol{V}$ 和模型状态 $\boldsymbol{M}^S$ 条件下,每个单词 w_i 生成概率的对数 $\log p(w_i \mid \boldsymbol{V}; \boldsymbol{M}^S)$,直接关联到模型在特定输入下的文本生成能力;其次,包括一个由 λ 加权的正则化项 $\lambda_{\max}(0, T - \cos(\boldsymbol{Z}_{\text{img}} \cdot \boldsymbol{Z}_{\text{txt}}))$, 其中 λ 是权重因子, T 是设定的最低余弦相似度阈值。此项确保当图像与文本特征相似度低于时,激活以增强特征间的对齐[15]。通过精确调整 λ 和 T, 可以精确控制模型性能,实现最佳的生成效果[16]。

1.4　自适应计算层

1.4.1　基于医学知识的自适应特征组合

在自适应计算层(Adaptive Computational Layer, ACL)中,研究采用了基于医学知识的权重调整机制,使视觉与文本特征的融合更为细致和有针对性。图 2 展示了自适应计算层在实际应用中的流程及关键步骤。此机制通过以下公式调整特征权重:

$$\boldsymbol{F}_{\text{combined}} = \alpha_{\text{med}}(\boldsymbol{C}) \cdot \boldsymbol{V} + (1 - \alpha_{\text{med}}(\boldsymbol{C})) \cdot \boldsymbol{T} \tag{5}$$

其中, $\boldsymbol{V}$ 表示从放射学图像中提取的视觉特征; $\boldsymbol{T}$ 表示从文本报告中提取的文本特征; $\boldsymbol{C}$ 表示临床情景(如病理、症状等); $\alpha_{\text{med}}(\boldsymbol{C})$ 表示一个基于医学知识动态调整的权重。

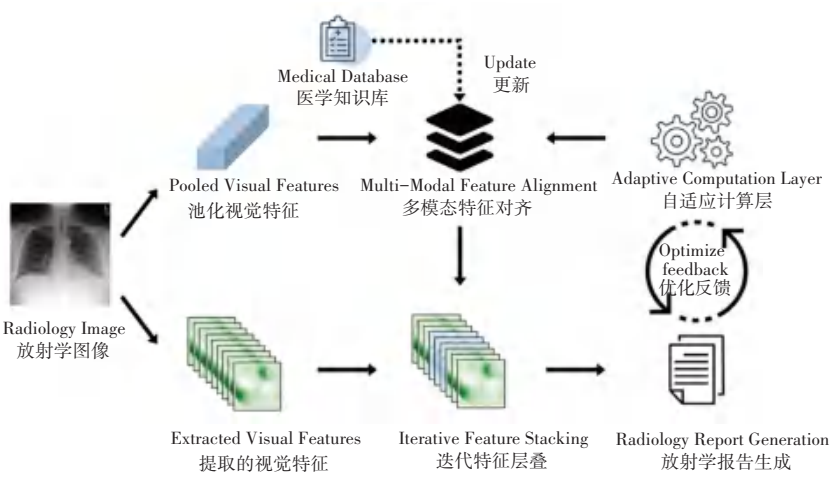

图 2 基于多模态特征对齐和自适应计算的放射学报告自动生成流程

Fig. 2 Automatic generation process of radiology reports based on multimodal feature alignment and adaptive computation

1.4.2 动态医学参数优化

为了进一步提高模型的适应性和精准度,采用了一种动态医学参数优化机制,该机制根据外部临床条件和内部性能反馈调整模型参数,数学公式如下:

$$\theta_{new} = g(\theta_{old}, \boldsymbol{\Delta}_{med}(\boldsymbol{C}), \boldsymbol{\Delta}_{internal}) \quad (6)$$

其中,θ_{old} 表示模型既有参数;$\boldsymbol{\Delta}_{med}(\boldsymbol{C})$ 表示根据临床情况调整的医学知识差异;$\boldsymbol{\Delta}_{internal}$ 表示模型内部的性能反馈。

2 实验方法与结果分析

2.1 数据集介绍

本研究使用了 2 个主要的数据集:MIMIC-CXR 和 IU-Xray。其中,MIMIC-CXR 数据集包含 377 110 张胸部 X 光图像及 227 827 份放射学报告,覆盖了 65 379 名患者。特别选取了包含正面和侧面视图的图像,所有图像均通过 CheXpert 工具进行了详细的病理学标注,涵盖了 14 种常见的胸部病理学观察[17]。IU-Xray 数据集由 3 955 份放射学报告和 7 470 张胸部 X 光图像组成,同样包括正位和侧位视图。每份报告详细记录了患者的 MeSH 术语、适应症、对比、发现和印象等关键信息,并且同样利用 CheXpert 工具进行了疾病标签的标注,确保数据的一致性和可比性[18-19]。

2.2 实现细节

在实验中,研究使用了 ImageNet 上预训练的 ResNet-101 作为视觉编码器[20]。文本编码器则采用专为放射学报告生成设计的基于 Transformer 的架构[21]。视觉编码器采用 Adam 优化器,初始学习率设为 5e-5。文本编码器和报告生成模块使用 AdamW 优化器[22-23],初始学习率设为 1e-4。所有输入图像调整至 224×224 分辨率并进行标准化处理[24]。文本数据经过预处理,包括转换为小写、去除停用词、词干提取和词形还原,并通过 Byte Pair Encoding (BPE) 处理词汇表外的单词[25]。训练的 *batch size* 为 16,以平衡内存使用和训练效率。所有训练步骤均运行在配备 NVIDIA Tesla V100 GPU 的服务器上[26-27]。

2.3 模型性能评估指标

在本研究中,采用了若干自然语言生成(NLG)标准性能评估指标来评价本文的模型性能。这些指标包括 *BLEU - n*,*METEOR* 和 *ROUGE - L*,广泛应用于评估机器翻译和文本生成任务的质量。

首先,*BLEU - n* 指标是通过计算机器翻译输出与一组参考翻译之间的 *n - gram* 重叠程度来评估翻译质量的[28]。其计算公式为:

$$BLEU - n = BP \cdot \exp\left(\sum_{i=1}^{N} w_i \log p_i\right) \quad (7)$$

其中,p_i 表示机器翻译输出与参考翻译之间的 n-gram 精确匹配的比率,w_i 表示相应的权重,使得所有 n-gram 的权重之和为 1。而 *BP* 是短句惩罚因子(Brevity Penalty),用于惩罚过短的翻译输出,其公式为:

$$BP = \begin{cases} 1, & \text{if } c > r \\ e^{(1-r/c)}, & \text{if } c \leqslant r \end{cases} \quad (8)$$

其中,c 表示候选翻译的长度,r 表示参考翻译的最接近长度。

接着,*METEOR* 指标通过综合考虑词汇的精确

度和句子的结构匹配，以实现更有效地模仿人类的翻译评价过程[29]。*METEOR* 的计算公式为：

$$METEOR = \frac{P \cdot R}{\alpha \cdot P + (1-\alpha) \cdot R} \tag{9}$$

其中，P 表示精确率；R 表示召回率；α 表示平衡参数，通常设为 0.9，以偏重召回率。

最后，*ROUGE - L* 指标通过计算最长公共子序列（LCS）来评估自动生成文本的质量。*ROUGE - L* 的计算方式如下[30]：

$$ROUGE - L = \frac{(1+\beta^2) \cdot R_{lcs} \cdot P_{lcs}}{R_{lcs} + \beta^2 \cdot P_{lcs}} \tag{10}$$

其中，P_{lcs} 和 R_{lcs} 分别表示基于 LCS 的精确度和召回率，β 是一个平衡精确率与召回率的参数，通常赋予更高的值以偏重召回率。

2.4　实验结果

2.4.1　模型比较与性能分析

在模型性能比较中，IRRAL-Net 在多个自然语言生成（NLG）指标上显著优于当前领先的医学报告生成模型，尤其在 *BLEU - 4* 指标上的表现最为突出。此指标评估了生成文本的连贯性和语法完整性，是衡量模型性能的关键指标之一。在 IU X-Ray 数据集上，与 CMN 模型进行比较后可知，IRRAL-Net 的 *BLEU - 4* 得分从 0.170 提升到 0.187，实现了约 10% 的提升。同样，在 MIMIC-CXR 数据集上，IRRAL-Net 的 *BLEU - 4* 得分从 CMN 模型的 0.106 提升至 0.111，增幅达到 4.72%。

此外，IRRAL-Net 在 IU X-Ray 和 MIMIC-CXR 两个数据集上均展现了卓越的性能，明显优于其他图像描述模型及专门的医学报告生成模型。表 1 清晰展示了 IRRAL-Net 在各个自然语言生成指标上相比其他模型的出色性能。

2.4.2　MIMIC-CXR 数据集临床效能指标

在深入分析 IRRAL-Net 在 MIMIC-CXR 数据集上的自然语言生成（NLG）性能之后，进一步评估了该模型在 MIMIC-CXR 数据集临床效能指标的表现。这些指标，如精确度（*Precision*）、召回率（*Recall*）和 *F1* 分数能够共同评估模型在识别和报告医学影像中特定特征的准确性和效率。

表 1　IRRAL-Net 与其他模型在 IU-Xray 和 MIMIC-CXR 数据集上的比较

Table 1　Comparison of IRRAL-Net with other models on the IU-Xray and MIMIC-CXR datasets

Dataset	Model	NLG Metrics					
		BL-1	BL-2	BL-3	BL-4	MTR	RG-L
IU X-Ray	ST	0.216	0.124	0.087	0.066	-	0.306
	ADAATT	0.220	0.127	0.089	0.068	-	0.308
	ATT2IN	0.224	0.129	0.089	0.068	-	0.308
	ContrAttn	0.455	0.288	0.205	0.154	-	0.369
	HRGR	0.438	0.298	0.208	0.151	-	0.322
	CMAS-RL	0.464	0.301	0.210	0.154	-	0.362
	SentSAT+KG	0.441	0.291	0.203	0.147	-	0.367
	R2Gen	0.470	0.304	0.219	0.165	0.187	0.371
	CMN	0.475	0.309	0.222	0.170	0.191	0.375
	IRRAL(Ours)	**0.478**	**0.312**	**0.229**	**0.187**	**0.197**	**0.376**
MIMIC-CXR	ST	0.299	0.184	0.121	0.084	0.124	0.263
	AdaAtt	0.311	0.178	0.111	0.075	0.118	0.246
	ATT2IN	0.325	0.203	0.136	0.096	0.134	0.276
	TopDown	0.280	0.169	0.108	0.074	0.128	0.250
	R2Gen	0.353	0.218	0.145	0.103	0.142	0.277
	CMCL	0.344	0.217	0.140	0.097	0.133	0.281
	CMN	0.353	0.218	0.148	0.106	0.142	0.278
	IRRAL(Ours)	**0.362**	**0.223**	**0.153**	**0.111**	**0.147**	**0.286**

MIMIC-CXR 数据集上临床效能指标的结果见表 2。分析表 2 可知,IRRAL-Net 在所有指标上均表现优异,尤其在 *F*1 分数上,其值达到 0.311,比最接近的 CMN 模型的 0.278 高出约 11.87%。这一显著提升展示了 IRRAL-Net 在整合精确度和召回率方面的优异性能表现。

表 2 MIMIC-CXR 数据集上临床效能指标的结果

Table 2 Results of clinical efficacy indicators on the MIMIC-CXR dataset

Model	*Precision*	*Recall*	*F*1
S&T	0.084	0.066	0.072
SA&T	0.181	0.134	0.144
AdaAtt	0.265	0.178	0.197
ATT2IN	0.322	0.239	0.249
TopDown	0.166	0.121	0.133
R2Gen	0.333	0.273	0.276
CMN	0.334	0.275	0.278
IRRAL (Ours)	**0.343**	**0.282**	**0.311**

2.4.3 消融研究

本研究深入探讨了多模态特征对齐(MFA)和自适应计算层(ACL)对 IRRAL-Net 模型性能的具体影响。研究目的是通过比较不同模型配置,验证这些组件单独及联合使用时对整体模型性能的影响。

4 种不同配置的模型被设计并比较:未集成 MFA 或 ACL 的基线模型;仅集成 MFA 的 Model-1;仅集成 ACL 的 Model-2;同时集成 MFA 和 ACL 的 IRRAL 模型。实验结果汇总在表 3 中,展示了各配置模型在自然语言生成(NLG)指标上的表现。

表 3 消融研究中不同模型的性能比较

Table 3 Performance comparison of different models in ablation studies

Model	MFA	ACL	*BL* - 3	*BL* - 4	*MTR*	*RG* - *L*
Baseline			0.105	0.079	0.103	0.244
Model-1	√		0.132	0.093	0.128	0.249
Model-2		√	0.146	0.106	0.117	0.271
IRRAL	√	√	**0.153**	**0.111**	**0.147**	**0.286**

从表 3 中可见,所有带有 MFA 或 ACL 的模型配置都显示出性能提升。基线模型在所有评估指标上的得分最低。相较于基线模型,Model-1 在 *BLEU* - 4 指标上提升了 17.72%,而 Model-2 在 *BLEU* - 4 指标上提升了 34.18%,显示了集成自适应计算层后在语义准确性和结构一致性上的显著改进。Model-1 和 Model-2 的比较表明,虽然每种技术单独使用均带来性能提升,但 Model-2 在 *ROUGE* - *L* 上的提升尤为显著,提升了 11.07%。

IRRAL 模型,同时集成了 MFA 和 ACL,展现了在所有指标上最优的性能,尤其是在 *BLEU* - 4 上,与基线模型相比提升了 40.51%。这表明 2 种技术的联合使用可显著提高报告生成的准确性和流畅性,有效地增强了模型在复杂医学报告任务中的性能。

2.4.4 图文映射可视化对比

本节通过可视化分析,展示了不同模型配置对胸部 X 光图像的处理效果及其对特定医学术语的关注差异。图 3 中,展示了原始胸部 X 光图像及 3 种模型配置:基线模型(Base)、基线模型加多模态特征对齐(Base+MFA)、以及基线模型加自适应计算层(Base+ACL)的处理结果。

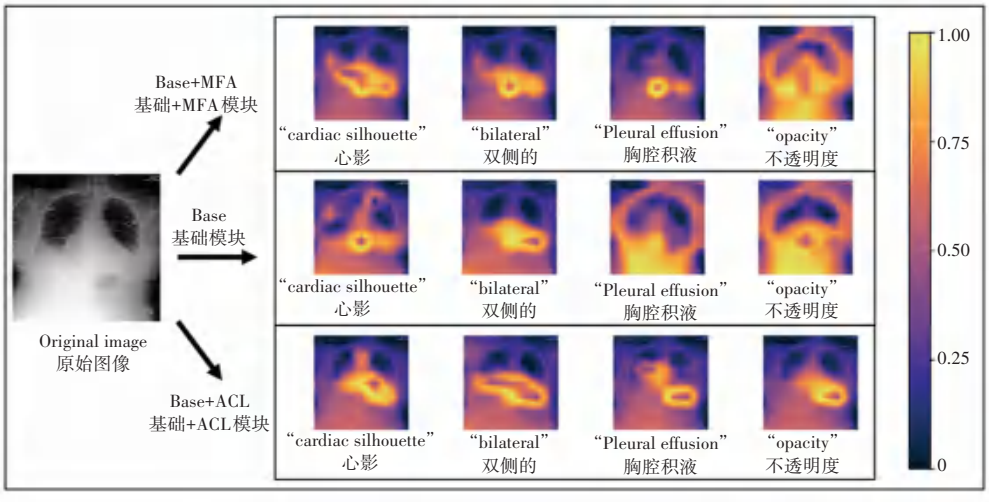

图 3 胸部 X 光图像与诊断术语关联的模型可视化对比

Fig. 3 Visual comparison of models linking chest X-ray images with diagnostic terms

从图 3 中可以看出，各模型配置针对 4 个关键医学术语："心影（cardiac silhouette）"、"双侧的（bilateral）"、"胸腔积液（Pleural effusion）"和"不透明度（opacity）"的热力图表示。热力图的颜色强度代表模型关注的程度，从深蓝色（最低关注）到亮黄色（最高关注）。

基线模型（Base）显示了较为基础的图文映射能力，对上述术语的关注较为均匀，但不够突出。引入 MFA 后（Base+MFA），关注度有显著提升，尤其是在"心脏轮廓"和"胸膜积液"两个术语上，热力图显示出更高的关注强度，这表明 MFA 有效地增强了模型对于图像特定区域的解析能力。加入 ACL 的模型（Base+ACL）进一步增强了这一效果，尤其在"双侧"和"不透明度"上，显示出更为集中和明确的关注区域，证明 ACL 在提高模型的空间解析精度和语义关联性上的有效性。

3　结束语

本研究开发的综合放射学报告与自适应学习网络（IRRAL-Net）显著促进了放射学报告自动生成技术的发展。通过整合跨模态自适应记忆网络与多尺度自适应注意力机制，IRRAL-Net 成功解决了视觉与文本信息协调的传统难题。广泛的测试证明，该网络在自然语言生成指标上优于现有技术，尤其是在 IU X-Ray 和 MIMIC-CXR 数据集上表现突出。此外，通过图文映射的可视化对比和消融研究，进一步证明了模型在实用性和诊断精度方面的显著优势。尽管 IRRAL-Net 已展示出卓越性能，自动化医学影像报告生成仍面临一系列挑战。未来研究将专注于提高模型对各种医学影像类型的适应性，增强其泛化能力，并探索多语言支持，以满足全球医疗需求的多样性。

参考文献

[1] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: A neural image caption generator [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 3156-3164.

[2] XU K, BA Lei, KIROS R, et al. Show, attend and tell: Neural image caption generation with visual attention[J]. arXiv preprint arXiv, 1502.03044, 2015.

[3] JING Baoyu, XIE Pengtao, XING E. On the automatic generation of medical imaging reports[J]. arXiv preprint arXiv, 1711.08195, 2017.

[4] CHEN Zhihong, SONG Yan, CHANG T H, et al. Generating radiology reports via memory-driven transformer [J]. arXiv preprint arXiv, 2010.16056, 2020.

[5] BANINO A, BADIA A P, KÖSTER R, et al. Memo: A deep network for flexible combination of episodicmemories[J]. arXiv preprint arXiv, 2001.10913, 2020.

[6] HOU B, KAISSIS G, SUMMERS R M, et al. Ratchet: Medical transformer for chest x-ray diagnosis and reporting [C]// 24th International Conference on Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI 2021). Cham: Springer, 2021: 293-303.

[7] WANG Z, TANG M, WANG L, et al. A medical semantic-assisted transformer for radiographic report generation [C]// International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2022: 655-664.

[8] TANIDA T, MÜLLER P, KAISSIS G, et al. Interactive and explainable region-guided radiology report generation [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 7433-7442.

[9] CHEN Zhihong, SHEN Yaling, SONG Yan, et al. Cross-modal memory networks for radiology report generation [J]. arXiv preprint arXiv, 2204.13258, 2022.

[10] ANTOL S, AGRAWAL A, LU J, et al. VQA: Visual question answering [C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 2015: 2425-2433.

[11] JING Baoyu, WANG Zeya, XING E. Show, describe and conclude: On exploiting the structure information of chest x-ray reports [J]. arXiv preprint arXiv, 2004.12274, 2020.

[12] WANG Xiaosong, PENG Yifan, LU Le, et al. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 9049-9058.

[13] DEVLIN J. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint arXiv, 1810.04805, 2018.

[14] BALTRUŠAITIS T, AHUJA C, MORENCY L P. Multimodal machine learning: A survey and taxonomy [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(2): 423-443.

[15] LECUN Y, BENGIO Y, HINTON G. Deep learning [J]. Nature, 2015, 521: 436-444.

[16] HUANG G, LIU Z, MAATEN V D L, et al. Densely connected convolutional networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 4700-4708.

[17] JOHNSON A E W, POLLARD T J, GREENBAUM N R, et al. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs [J]. arXiv preprint arXiv, 1901.07042, 2019.

[18] DEMNER-FUSHMAN D, KOHLI M D, ROSENMAN M B, et al. Preparing a collection of radiology examinations for distribution and retrieval [J]. Journal of the American Medical Informatics Association, 2016, 23(2): 304-310.

[19] RVIN J, RAJPURKAR P, KO M, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 590-597.

[20] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Piscataway, NJ: IEEE, 2016: 770-778.
[21] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Proceedings of the 31st International Conference on Neural Information Processing System. Long Beach, USA: NIPS Foundation, 2017: 6000 - 6010.
[22] KINGMA D P. Adam: A method for stochasticoptimization[J]. arXiv preprint arXiv, 1412.6980, 2014.
[23] LOSHCHILOV I. Decoupled weight decayregularization[J]. arXiv preprint arXiv, 1711.05101, 2017.
[24] KRIZHEVSKY A, SUTSKEVER I, HINTON G E, et al. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.
[25] SENNRICH R. Neural machine translation of rare words withsubword units[J]. arXiv preprint arXiv, 1508.07909, 2015.
[26] BENGIO Y. Practical recommendations for gradient - based training of deep architectures [M]//Neural networks: Tricks of the trade. Cham: Springer, 2012: 437-478.
[27] NICKOLLS J, BUCK I, GARLAND M, et al. Scalable parallel programming with CUDA: Is CUDA the parallel programming model that application developers have been waiting for? [J]. Queue, 2008, 6(2): 40-53.
[28] 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. ACL, 2002: 311-318.
[29] DENKOWSKI M, LAVIE A. METEOR 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems [C]// Proceedings of the Sixth Workshop on Statistical Machine Translation. New York: ACM, 2011: 85-91.
[30] LIN C Y. ROUGE: A package for automatic evaluation of summaries [C]// Text Summarization Branches Out. ACL, 2004: 74-81.