陈嘉鸿, 黄国恒, 谭喆. 基于跨模态差异注意力的医学报告生成[J]. 广东工业大学学报, 2025, 42(1): 70–78. doi: 10.12052/gdutxb.240002.
Chen Jiahong, Huang Guoheng, Tan Zhe. Cross-modal discrepancy attention network for medical report generation[J]. Journal of Guangdong University of Technology, 2025, 42(1): 70–78. doi: 10.12052/gdutxb.240002.

基于跨模态差异注意力的医学报告生成

陈嘉鸿，黄国恒，谭　喆
（广东工业大学　计算机学院，广东　广州　510006）

摘要: 医学报告自动生成技术对辅助诊断起着重要作用，能够极大减轻医护工作者的工作量。随着深度学习在医学领域不断发展，医学报告自动生成技术已成为智慧医疗领域里的研究热点之一。目前，医学报告生成的主要挑战是图像中的病灶区域难以被模型捕捉，以及视觉和语言语义之间存在较大的语义鸿沟，其一致性问题仍没有很好地解决。因此，本文提出了跨模态差异注意力网络拉近不同模态之间的语义，该网络包括反向注意力模块和语义一致模块：反向注意力模块更全面探索医学图像中的重要区域；语义一致模块利用大语言模型的特征作为参考，引导视觉特征不断靠近参考文本特征，使得视觉语义更准确地转化成一致的语言语义。实验表明，跨模态差异注意力网络在IU X-Ray和MIMIC-CXR两个公开数据集上的表现均优于之前的模型，在BLEU4上的指标分数分别达到17.9%和10.9%，相比于基线模型，本文模型性能有较大的提高，证明了本文所提模型能生成准确和流畅的医学报告。

关键词: 医学报告生成；语义一致；注意力机制
中图分类号: TP391.4　　**文献标志码:** A　　**文章编号:** 1007–7162(2025)01–0070–09

Cross-modal Discrepancy Attention Network for Medical Report Generation

Chen Jiahong, Huang Guoheng, Tan Zhe
(School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: Automatic medical report generation technology plays an important role in auxiliary diagnosis and can greatly reduce the workload of medical workers. As deep learning continues to develop in the medical field, automatic medical report generation technology has become one of the research hotspots. Currently, the main challenges in medical report generation are (1) the difficulty of capturing lesion regions in images by models, and (2) the large semantic gap between visual and language semantics, whose consistency problem is still not well solved. Therefore, in order to solve the above problems, a Cross-Modal Discrepancy Attention Network (CDAN) is proposed to bring closer the semantics between different modalities. The network includes a Reverse Attention (RA) module and a Semantic Consistency (SC) module: (1) the Reverse Attention module explores important areas in medical images more comprehensively, and (2) the Semantic Consistency module utilizes the features of the large language model as a reference to guide the visual features to continuously approach the reference language features, so that the visual semantics can be more accurately converted into language semantics. Experiments show that the Cross-Modal Discrepancy Attention Network is better than the previous model on both IU X-Ray and MIMIC-CXR public datasets, with BLEU4 scores reaching 17.9% and 10.9% respectively. Compared with the baseline model, improvement is significant in performance, which proves that the proposed model is capable of generating accurate and fluent medical reports.

Key words: medical report generation; semantic consistency; attention mechanism

收稿日期：2024-01-02　　**录用日期：**2024-03-27　　**网络首发日期：**2024-07-04
基金项目：广东省重点领域研发计划项目(2019B010153002)；佛山市重点领域科技攻关项目(2020001006832)
作者简介：陈嘉鸿(1998–)，男，硕士研究生，主要研究方向为医学报告生成、图像描述，E-mail：jhchan003@gmail.com
通信作者：黄国恒(1985–)，男，副教授，博士，主要研究方向为计算机视觉、模式识别和人工智能，E-mail：kevinwong@gdut.edu.cn

医学报告往往需要经验丰富的医护工作者撰写，不仅耗费医护工作者的大量时间，而且容易出现误写。为了解决以上问题，医学报告自动生成技术应运而生，医学报告自动生成任务是计算机模仿医护工作者分析医学图像的诊断过程，最终输出一段诊断性描述。由于医学报告自动生成能够有效辅助医护工作者诊断疾病，所以该方向研究在医学领域中具有重大意义。

随着医学图像处理技术逐渐成熟[1-2]和图像描述任务的快速发展[3-4]，医学领域的图像描述任务也引起了研究者的广泛关注，并且逐渐发展成医学报告生成任务。目前，医学报告生成任务沿用与图像描述任务相同的编码器–解码器结构，这不可避免地存在编码器和解码器之间语义不一致的问题，即图像语义转化到语言语义存在着语义鸿沟，导致生成的报告难以准确地描述图像中的病灶区域。已有许多医学报告生成工作致力于生成准确和流畅的报告[5-8]，但生成报告的效果仍不满足人们的需求。目前，该任务存在以下挑战：(1) 模型难以发现图像病灶区域。部分工作利用注意力机制让模型更准确捕捉图像病灶[9-10]。另一部分工作则利用诊断关键词辅助模型发现图像中的病灶区域。上述的研究都是利用正向关注的方式去捕捉图像中的重要区域，缺乏对正向关注区域之外的区域探索，导致模型容易忽略图像中的潜在重要区域。(2) 视觉和语言语义的潜在关系没有被模型较好地理解，模型难以对图像的病灶区域生成准确描述。大多数方法采用注意力机制融合不同模态之间的语义[6,9]，但它们对不同模态之间的语义转化是有限的，在模态交互过程中忽略了有效信息的指导，并且在损失函数层面缺乏语义一致的约束。

为了解决以上问题，本文提出了跨模态差异注意力网络(Cross-Modal Discrepancy Attention Network, CDAN)去生成语义准确的医学报告。针对图像病灶区域难以捕捉的问题，设计反向注意力模块指导模型更全面地探索图像，关注不同区域的差异，提供丰富的视觉语义信息生成准确文本。另外，针对跨模态语义不一致的问题，借助医学大模型的语言理解能力，设计视觉语言一致性模块，减少视觉和语言模态的转换差异，解决视觉和语言模态之间的语义鸿沟问题。在2个公开数据集上的相关实验证明了跨模态差异注意力网络的有效性。本文的贡献有以下几点：

(1) 本文提出跨模态差异注意力网络去捕捉图像病灶区域以及缩小视觉和语言模态之间的语义鸿沟。在充分发掘视觉语义信息的前提下进行模态交互，模型进一步有效理解模态之间的潜在关系。

(2) 提出反向注意力模块提高视觉特征中的病灶语义，加强了模型对图像中不同区域的探索，得到更丰富的视觉语义。

(3) 提出语义一致模块缩小视觉和语言之间的语义距离，视觉和语言特征能够更深层次地交互，减轻了不同模态转化之间的难度。

(4) 在IU X-Ray和MIMIC-CXR两个公开数据集上展开实验，相关指标分数均高于之前的模型，验证了本文提出的跨模态差异注意力网络有充分的能力生成语义一致的报告。

1　相关工作

1.1　图像描述

图像描述是对图像生成简短描述的任务，许多研究者对此进行研究[3,9-11]。起初，基于模板[12-13]和检索[14-15]的方法被广泛用于图像描述任务，但这些方法生成的描述效果有限，模板和检索的方式难以灵活地生成报告。随着深度学习的发展，视觉编码器–文本解码器的深度学习架构成为研究热点。注意力机制被提出用来关注图像中的关键对象[4,11]，生成描述的效果得到有效提高，但在生成长文本任务上仍有巨大提升空间。因此，用于生成段落描述的层级生成方法被提出[16]，然而，生成长文本任务需要模型去捕捉句子间的关联。由于Transformer[17]模型在处理长文本任务中取得巨大成功，多模态Transformer模型[18]被用于加强模态内交互，后续基于Transformer的改进方法在此任务上取得了优异的表现。

但是，图像描述任务的方法不能很好地迁移到医学报告生成任务中，医学领域的文本生成任务更复杂，不仅要求模型生成长报告，而且要求模型能够捕捉图像中的异常区域以生成准确的描述。

1.2　医学报告生成

随着编码器–解码器的架构在图像描述任务中取得显著效果，医学领域的图像描述任务，即医学报告生成任务吸引了众多研究者的注意。Li等[19]应用强化学习，结合模板和生成的方式生成报告，但模板需要精心挑选，因此很难泛化到其他数据集。Jing等[6]提出了共同注意力机制关联不同模态的语义。Wang等[20]提出了多层级注意力机制去发现关键单词和重要图像区域，但模型不是聚焦在报告生成任

务上。Yin等[21]介绍了一种主题匹配机制生成更多样和准确的报告。由于生成报告需要涉及的专业术语和知识较多，Li等[22]将视觉特征转换为图结构去生成更准确的报告。Zhang等[23]构建疾病知识图谱去发现疾病的关系。由于自注意力机制能够并行处理文本，所以最近一些研究都是采用Transformer架构[5,24]。Chen等[5]设计了一种记忆驱动的Transformer模型记录报告中的关键信息，You等[24]介绍了一种AlignTransformer模型对齐视觉特征和病灶特征。但以上注意力方法对图像重要区域的探索和不同区域差异的关注有限，并且少有针对性地拉近图像和报告两种不同模态的语义。因此本文提出了跨模态差异注意力的医学报告生成模型，生成语义一致的医学报告。

1.3 跨模态语义一致

跨模态语义一致性问题一直是医学报告生成的研究重点，视觉编码器和文本解码器表征着不同模态的语义，导致它们之间存在着较大的语义鸿沟问题。之前的大部分工作是通过注意力机制的方式融合不同模态，Jing等[6]提出了共同注意力机制关联视觉、疾病和语言语义。You等[24]设计层级注意力将视觉特征和病灶标签特征对齐，以得到更好的视觉病灶表征。Chen等[5]利用Transformer的注意力机制将视觉和语言语义对齐，但是注意力机制解决视觉和语言语义一致问题仍有限。因此，本文提出语义一致模块提高视觉病灶语义转化成准确语言语义的能力，通过提供视觉特征一个可靠的语言特征作为参考，结合损失的方式约束网络关注视觉和语言语义不一致的问题。

2 方法

跨模态差异注意力网络主要由3个核心部分组成，分别是CDAN编码器、CDAN解码器和视觉语义一致模块。CDAN编码器包含了本文设计的反向注意力模块。整个模型的架构如图1所示，CDAN编码器提取疾病语义的视觉表征，较为全面地发现图像中的病灶区域，CDAN解码器融合视觉和语言语义，将视觉特征解码成报告文本，视觉语义一致模块则通过准确的语言语义指导视觉语义，约束整个模型视觉和语言语义一致，使视觉语义更容易和更准确地被解码器理解，最终模型能够生成准确的医学报告。对于给定的医学图像I，首先被视觉提取器转化为视觉分片特征，然后经过CDAN解码器得到视觉语义特征$X=\{x_1,x_2,x_3,\cdots,x_s\}$，其中$x_s$是第$s$个视觉序列特征，图像对应的报告序列文本$Y=\{y_1,y_2,y_3,\cdots,y_T\}$，其中$y_T$是生成报告的某一字符，$T$是报告的长度。整个报告生成过程可以被描述为式(1)。

$$p(Y|I)=\prod_{t=0}^{T}p(y_t|y_1,\cdots,y_{t-1},I) \tag{1}$$

然后在给定I的情况下，通过Y的负条件对数似然，将$p(Y|I)$最大化，如式(2)所示。

$$\theta^{*}=\arg\max_{\theta}\sum_{t=1}^{T}\log p(y_t|y_1,\cdots,y_{t-1},I;\theta) \tag{2}$$

式中：θ是模型的参数。

另外，交叉熵被作为损失函数，模型生成的报告序列$R=\{r_1,r_2,r_3,\cdots,r_T\}$和报告标签$Y$被交叉熵函数处理，如式3所示。

$$L_{\text{text}}=\text{CrossEntry}(R,Y) \tag{3}$$

2.1 反向注意力模块

视觉特征的语义表征能力决定报告生成的质量，因此，本文提出反向注意力模块使网络能对图像有更全面和细致的观察。对于医学图像I，经过网络中的CNN特征提取器，得到初步的视觉特征，如式(4) 所示。

$$u=\text{VisualExtract}(I) \tag{4}$$

接着，经过CDAN编码器得到语义丰富的视觉特征。初步的视觉特征u首先会被CDAN编码器中的反向注意力模块处理，该模块由自注意力和反向注意力组成，能够提高模型发现病灶的能力。反向注意力由自注意力机制优化而来，为了区分反向注意力和自注意力，本文把自注意力机制定义为PositiveAttention，反向注意力为NegativeAttention，自注意力机制如式(5)~(8)所示。

$$\text{PositiveAttention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Softmax}\left(\frac{\boldsymbol{Q}_{\text{w}}\boldsymbol{K}_{\text{w}}^{\text{T}}}{\sqrt{d}}\right)\boldsymbol{V}_{\text{w}} \tag{5}$$

$$\boldsymbol{Q}_{\text{w}}=\boldsymbol{W}_{\text{q}}\boldsymbol{Q} \tag{6}$$

$$\boldsymbol{K}_{\text{w}}=\boldsymbol{W}_{\text{k}}\boldsymbol{K} \tag{7}$$

$$\boldsymbol{V}_{\text{w}}=\boldsymbol{W}_{\text{v}}\boldsymbol{V} \tag{8}$$

式中：$\boldsymbol{Q}_{\text{w}}$、$\boldsymbol{K}_{\text{w}}$和$\boldsymbol{V}_{\text{w}}$是矩阵，$d$是矩阵的维度。$\boldsymbol{W}_{\text{q}}$、$\boldsymbol{W}_{\text{k}}$和$\boldsymbol{W}_{\text{v}}$是可学习矩阵。反向注意力机制与自注意力不同，它被用来关注自注意力机制中分数较低的特征，如式(9) 所示。

$$\text{NegativeAttention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=(1-\text{Softmax}\left(\frac{\boldsymbol{Q}_{\text{w}}\boldsymbol{K}_{\text{w}}^{\text{T}}}{\sqrt{d}}\right))\boldsymbol{V}_{\text{w}} \tag{9}$$

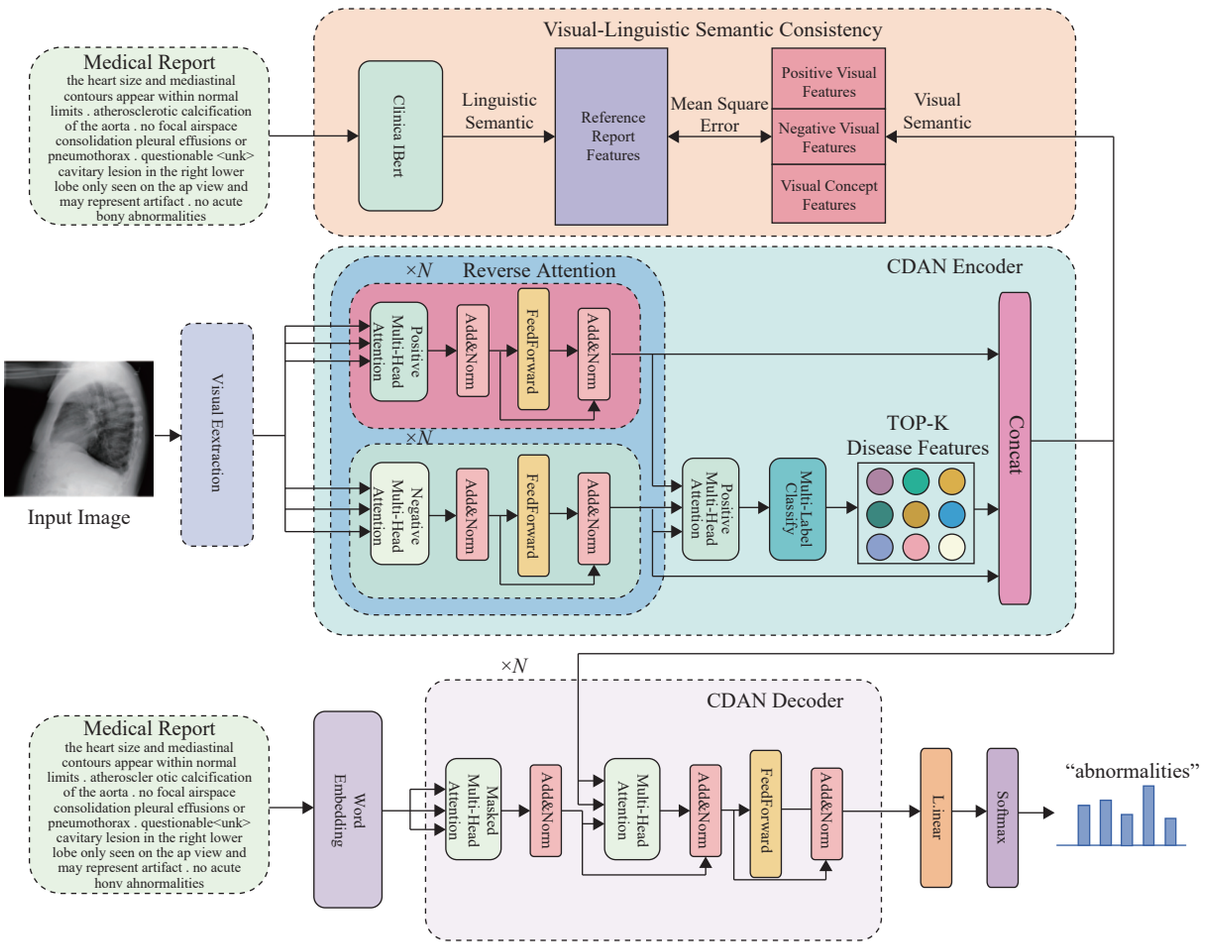


图 1　跨模态差异注意力网络架构图

Fig.1　Architecture of Cross-modal discrepancy attention network

初步的视觉特征被自注意力和反向注意力机制处理的过程如式(10)~(11)所示：

$$u_{\mathrm{p}} = \mathrm{PositiveAttention}(u,u,u) \tag{10}$$

$$u_{\mathrm{n}} = \mathrm{NegativeAttention}(u,u,u) \tag{11}$$

对于增强视觉特征的语义，仅有自注意力和反向注意力机制还是不够，本文希望模型能获得图像中的疾病关键词语义。首先，通过自注意力，模型能学习到视觉特征u_{p}和u_{n}的差异，得到语义较为丰富的视觉特征u_{pn}，分类网络处理该特征后，输出u_{c}，最后经过Topk函数，得到分数较高的Top-K个疾病词语义u_{k}。如式(12) ~(14) 所示。

$$u_{\mathrm{pn}} = \mathrm{PositiveAttention}(u_{\mathrm{p}},u_{\mathrm{n}},u_{\mathrm{n}}) \tag{12}$$

$$u_{\mathrm{c}} = \mathrm{Classify}(u_{\mathrm{pn}}) \tag{13}$$

$$u_{\mathrm{k}} = \mathrm{Topk}(\mathrm{Softmax}(u_{\mathrm{c}})) \tag{14}$$

式中：Classify是分类网络。此外，二元交叉熵被应用于多标签分类网络中：

$$L_{\mathrm{mlc}} = -\frac{1}{C}\sum_{i\in[1,2,\cdots,C]} y_i \cdot \log((1+\exp(-x_i))^{-1}) + (1-y_i)\log\left(\frac{\exp(-x_i)}{1+\exp(-x_i)}\right) \tag{15}$$

式中：x_i是被预测的疾病关键词，C是疾病词的类别数目，疾病词标签来自Chexpert-Label[25]，$y_i \in \{0,1\}$是对应的类别标签，$y_i = 0$表示该医学图像没有该疾病，$y_i = 1$则表示图像存在该疾病。

最后，将u_{p}、u_{n}和u_{k}拼接成视觉语义u_{s}，特征拼接能够让网络同时注意到这3种特征的差异和获得更丰富的视觉语义。

$$u_{\mathrm{s}} = \mathrm{Concat}([u_{\mathrm{p}},u_{\mathrm{n}},u_{\mathrm{k}}]) \tag{16}$$

2.2　语义一致模块

虽然解码器生成的视觉特征u_{s}有丰富的视觉语义，但是如果没有外部信息指导视觉语义靠近语言语义，模型生成准确的语言语义的能力有限，因此，本文利用Clinical BERT[26]获取报告的语言语义e_{r}，通过平均方差误差损失约束视觉语义趋向准确的语言

语义，如式(17)~(18)所示。

$$e_{\mathrm{r}} = \mathrm{ClinicalBert}(Y) \tag{17}$$

$$L_{\mathrm{mse}}(u_{\mathrm{s}}, e_{\mathrm{r}}) = \frac{1}{N}\sum_{i=1}^{N}(u_{\mathrm{s}_i} - e_{\mathrm{r}_i})^2 \tag{18}$$

式中：N为u_{s}和e_{r}的维度大小。e_{r}相对于u_{s}是训练标签。在损失函数的约束下，模型减小视觉和语言之间的语义跨度，病灶区域的视觉特征有效地转化为文本特征，最终生成准确的医学报告。另外，语义一致模块仅在训练过程引导视觉特征，在测试阶段无需参与。

2.3 解码器

视觉特征和语言特征被解码器充分融合后，生成语义一致的医学报告。报告文本经过词嵌入后得到文本特征z，文本特征之间的内在关系被自注意力捕捉，得到丰富的语言语义特征z_{a}。接着，通过自注意力机制融合文本特征z_{a}和解码器生成的视觉特征u_{s}，输出用于生成文本的上下文特征u_{ctx}。最后利用Softmax函数计算生成单词的概率分布。主要过程如式(19)~(21)所示。

$$z_{\mathrm{a}} = \mathrm{PositiveAttention}(z, z, z) \tag{19}$$

$$u_{\mathrm{ctx}} = \mathrm{PositiveAttention}(z_{\mathrm{a}}, u_{\mathrm{s}}, u_{\mathrm{s}}) \tag{20}$$

$$u_{\mathrm{w}} = \mathrm{Softmax}(\mathrm{Linear}(u_{\mathrm{ctx}})) \tag{21}$$

式中：Linear是线性网络层。

2.4 损失函数

本文所提出的模型由L_{text}、L_{mlc}和L_{mse}损失函数共同约束，总损失如式(22) 所示。

$$L_{\mathrm{total}} = L_{\mathrm{text}} + L_{\mathrm{mlc}} + L_{\mathrm{mse}} \tag{22}$$

3 实验结果及讨论

3.1 数据集

本文在IU X-Ray[27]和MIMIC-CXR[28]两个公开数据集上开展相关实验。

IU X-Ray是一个胸片图像数据集，包含7 470张图像和3 955份报告。本文遵循文献[5]的做法，将数据集按7:1:2比例划分成训练集、验证集和测试集。

MIMIC-CXR是一个数据量巨大的胸片数据集，包含473 057张胸片图像和206 563份报告。本文采用MIMIC-CXR官方的数据划分比例，区分训练集、验证集和测试集。

3.2 实验细节

本文采用Resnet101预训练模型作为视觉特征提取器，得到维度是2 048的分片特征。所有的实验均在Nvidia 2080ti GPU上进行。通过Clinical Bert获取的语言特征维度是768。反向注意力机制输出特征的维度是512。解码器和编码器层数N均是3。IU X-Ray数据集的Batch Size是16，每个epoch所需运行时间大约1 min，由于MIMIC-CXR数据集数据量较大，Batch Size设置成32。每个epoch所需运行时间大约102 min。另外，Adam被用作模型的优化器。模型总参数量是82.25 M。

3.3 Baseline和验证指标

本文将提出的跨模态差异注意力网络与最新的方法开展对比实验。基线方法包括经典的图像描述方法：ST[3]、ATT2IN[9]、ADAATT[10]和TOPDOWN[11]。还包括最新的医学报告生成方法：R2GEN[5]、COATT[6]、HRGR[19]、CMAS-RL[29]、CMCL[30]、R2GenCMN[31]、PPKED[32]、GNEDNET[33]和XPRONET[34]。此外，本文采用文本生成自动评价指标衡量模型的性能，包括BLEU[35]、METEOR[36]和ROUGE-L[37]。

3.4 对比实验

实验结果如表1所示，由于文献[5]提出的R2GEN[5]方法在医学报告生成领域中有着较大影响力，所以本文ST、ATT2IN、ADAATT、COATT、HEGR、CMAS-RL和R2GEN方法的结果均来自文献[5]，“—”表示相关指标分数在文献[5]中未记录，CMCL，R2GenCMN，PPKED，GNEDNET和XPRONET的结果来自原文献。本文模型无论在MIMIC-CXR大数据集，还是在IU X-RAY小数据集上均取得优异表现。此外，本文模型不仅分数超越图像描述模型，而且在各项指标上与之前的医学报告生成模型相比均有一定的提高，表明了本文模型能够生成较准确的医学报告。具体地，R2GEN模型提出了一个记录生成过程中关键信息的模块，以此生成流畅的长报告，但在捕捉视觉病灶和语义一致方面的能力有限，而本文模型考虑了这些问题，因此在IUX-RAY数据集上比R2GEN平均每个指标提升4.76%，在MIMIC-CXR上平均提升2.97%。BLEU4指标反映了报告生成的流畅度，BLEU1则反映了报告的准确度。在MIMIC-CXR中，本文模型在BLEU4上比R2GEN提高了6%，并且在BLEU1上比R2GEN提升了4%，进一步证明了本文模型能够生成准确和流畅的报告。

3.5 消融实验

实验结果如表2所示，本文提出了反向注意力模

块(Reverse Attention，RA) 和语义一致模块(Semantic Consistency, SC) , Transformer作为模型的主要架构，简称BASE。在IU X-RAY数据集中，BASE+RA和BASE+SC大部分指标分数都比BASE高，反映了本文设计的反向注意力模块和语义一致模块在提高报告质量方面的有效性。BASE+RA+SC结合了两个模块的能力，由于RA模块提高了视觉特征的病灶语义，SC更好地将视觉特征和语言特征的语义拉近，因此，BASE+RA+SC大部分指标分数都比BASE+RA和BASE+SC高，BASE+RA+SC相对BASE平均每个指标提升了9.27%。在MIMIC-CXR数据集上，BASE+RA+SC比BASE+RA和BASE+SC平均各指标提升6.81%和1.66%，表明了在RA和SC模块的共同作用下，模型的整体性能有较大的提高。另外BASE+SC相比于BASE+RA，分数提高显著，反映了SC模块在模型语义一致方面有着重要的贡献。

此外，为了更好地对比本文每个模块的性能，在图2展示了IU X-Ray数据集的报告样例。由于R2GEN模型在报告的流畅性和准确性方面都做出了巨大贡献，所以，本文模型将与R2GEN生成的报告作对比。与之前类似，反向注意力模块简称RA，语义一致模块简称SC，Ours由RA+SC组成，Reference Report代表医学图像的参考报告。相比于R2GEN，在Example 1样例中，Ours生成的报告更加接近Reference

表 1　对比实验

Table 1　Comparative experiment

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
IU X-RAY	ST [3]	0.216	0.124	0.087	0.066	—	0.306
	ATT2IN[9]	0.224	0.129	0.089	0.068	—	0.308
	ADAATT[10]	0.220	0.127	0.089	0.068	—	0.308
	COATT[6]	0.455	0.288	0.205	0.154	—	0.369
	HRGR[19]	0.438	0.298	0.208	0.151	—	0.322
	CMAS-RL[29]	0.464	0.301	0.210	0.154	—	0.362
	R2GEN[5]	0.470	0.304	0.219	0.165	0.187	0.371
	CMCL [30]	0.473	0.305	0.217	0.162	0.186	0.378
	R2GenCMN[31]	0.475	0.309	0.222	0.170	0.191	0.375
	Ours	**0.472**	**0.315**	**0.230**	**0.179**	**0.200**	**0.386**
MIMIC-CXR	ST[3]	0.299	0.184	0.121	0.084	0.124	0.263
	ATT2IN[9]	0.325	0.203	0.136	0.096	0.134	0.276
	ADAATT[10]	0.299	0.185	0.124	0.088	0.118	0.266
	TOPDOWN[11]	0.317	0.195	0.130	0.092	0.128	0.267
	R2GEN[5]	0.353	0.218	0.145	0.103	0.142	0.277
	CMCL [30]	0.344	0.217	0.140	0.097	0.133	0.281
	R2GenCMN[31]	0.353	0.218	0.148	0.106	0.142	0.278
	PPKED[32]	0.360	0.224	0.149	0.106	0.149	0.284
	GNEDNET[33]	0.361	0.223	0.150	0.108	0.150	0.287
	XPRONET[34]	0.344	0.215	0.146	0.105	0.138	0.279
	Ours	**0.367**	**0.225**	**0.151**	**0.109**	**0.143**	**0.277**

表 2　消融实验

Table 2　Ablation experiment

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
IU X-RAY	BASE	0.458	0.290	0.208	0.159	0.180	0.352
	BASE+RA	0.452	0.292	0.210	0.160	0.184	0.354
	BASE+SC	0.474	0.298	0.215	0.166	0.187	0.360
	BASE+RA+SC	**0.472**	**0.315**	**0.230**	**0.179**	**0.200**	**0.386**
MIMIC-CXR	BASE	0.307	0.189	0.126	0.099	0.125	0.270
	BASE+RA	0.343	0.208	0.139	0.100	0.135	0.271
	BASE+SC	0.360	0.221	0.148	0.106	0.142	0.275
	BASE+RA+SC	**0.367**	**0.225**	**0.151**	**0.109**	**0.143**	**0.277**

Report，体现了本文所提出的模型能够生成流畅和完整的报告。在Example 2样例中，R2GEN没有该病灶的相关描述，而SC生成的报告与Reference Report有着同样的病灶描述。最后在Example 3样例中，尽管RA生成的报告与Reference Report不完全相同，但都是针对肺部的右侧病灶进行描述，而R2GEN与Reference Report对于病灶描述的准确性不足，相近的语义有限。

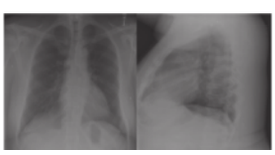
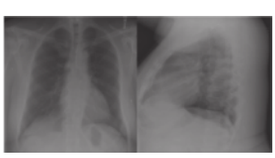
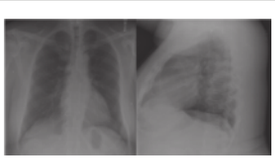
Example 1	Ours	R2GEN	Reference Report
	the heart size is normal . the lungs are clear . there is no pleural effusion or pneumothorax . there is no pneumothorax .	the lungs are clear bilaterally . specifically no evidence of focal consolidation pneumothorax or pleural effusion . cardio mediastinal silhouette is unremarkable . visualized osseous structures of the thorax are without acute abnormality	the lungs are clear . there is no pleural effusion or pneumothorax . the heart and mediastinum are normal . the skeletal structures are normal .
Example 2	LLMGVR	R2GEN	Reference Report
	the heart size and pulmonary vascularity appear within normal limits . the lungs are free of focal airspace discasc . no pleural effusion or pneumothorax is seen . degenerative changes are present in the spine .	the heart size and pulmonary vascularity appear within normal limits . the lungs are free of focal airspace disease . no pleural effusion or pneumothorax is seen .	the heart size and pulmonary vascularity appear within normal limits . the lungs are free of focal airspace disease . no pleural effusion or pneumothorax is seen . calcified granuloma are present . degenerative changes are present in the spine .
Example 3	RA	R2GEN	Reference Report
	heart size within normal limits . no focal alveolar consolidation no definite pleural effusion seen . no typical findings of pulmonary edema . mediastinal calcification and dense right upper lung nodule suggest a previous granulomatous process .	the cardiomediastinal silhouette and pulmonary vasculature are within normal limits in size . the lungs are clear of focal airspace disease pneumothorax or pleural effusion . there are no acute bony findings .	the heart is normal in size and contour . there is a vague arca of airspace disease identified within the right midlung on the pa view . this is not <unk> on the lateral view . there is no pneumothorax or effusion .

图 2 生成报告对比

Fig.2 Comparison of generated reports

为了更好地证明本文所提出的模型能够生成视觉语言语义一致的报告，可视化了经过PositiveAttention和NegativeAttention得到的u_p和u_n特征，如图3所示，图像中红色区域是特征中的重要区域，蓝色区域则是关注度较低的区域，标黄的文字表示预测的报告和参考报告所描述的异常一致。对于Example 1 和Example 3，u_p和u_n各自关注区域是互补的，模型能够对图像的各区域有不同程度的关注。在Example 1中，u_p和u_n的关注区域都集中在肺部区域，利于生成“low lung volumes”，并且u_n的关注区域偏向图像的右边，辅助模型生成“right hemidiaphragm”描述。在Example 2中，u_p和u_n都集中在正面影像图的“lung”区域，但u_n相对于u_p，对“lung”区域关注得更全面，加强了模型对病灶区域的注意，辅助模型生成“lungs

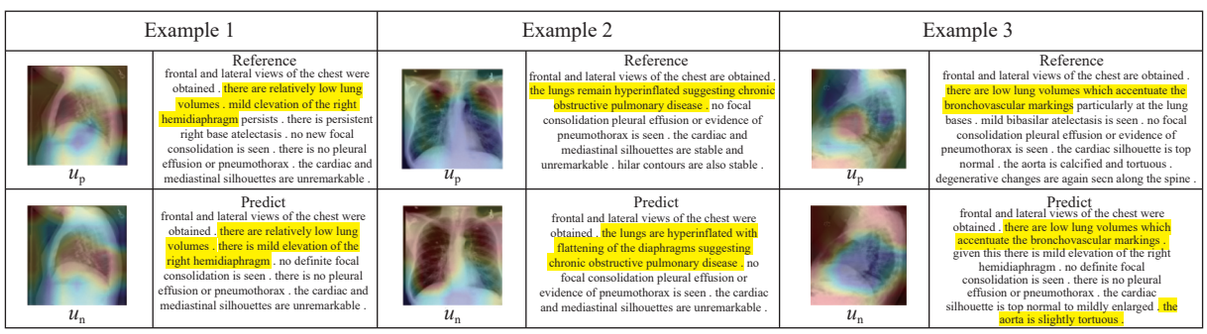


图 3 语义一致可视化

Fig.3 Visualisation of semantic consistency

are hyperinflated”相关异常描述。在Example 3中，u_p和u_n关注的重要区域不同，这加大了模型对图像区域的探索，从而生成对“lung”和“aorta”区域的异常描述。通过特征可视化，进一步表明了模型生成的描述与模型探索的图像区域一致，具有视觉语言语义一致的能力。另外，通过可视化u_p和u_n，展示了u_p和u_n关注的区域侧重不同，表明反向注意力模块探索了图像的不同潜在病灶区域，从u_p和u_n的不同视角，为模型提供丰富视觉的语义。

4　结论

本文提出了基于跨模态差异注意力的医学报告生成网络，针对图像中的病灶区域难以被模型捕捉和视觉和语言语义之间的一致性问题，分别设计了反向注意力模块和语义一致模块。反向注意力模块提高模型探索视觉特征中病灶区域的能力，语义一致模块拉近视觉和语言语义，缩小语义鸿沟。在IUX-RAY和MIMIC-CXR上的实验结果表明，与现有的模型相比，本文的模型在各指标分数上均取得有效提升；消融实验表明，本文设计的两个模块对模型的性能提高都有较大的贡献，进一步证明了本文模型的优越性，能够生成较准确和流畅的医学报告。

参考文献：

[1] 李小雷. 人工智能在医学影像图像处理中的研究进展[J]. 中国医学计算机成像杂志, 2023, 29(4): 454-457.
LI X L. Advances in the application of artificial intelligence in medical image processing [J]. Chinese Computed Medical Imaging, 2023, 29(4): 454-457.

[2] 丛超. 基于多模态医学影像智能分析的深度学习算法研究与应用 [D]. 重庆: 中国人民解放军陆军军医大学, 2023.

[3] 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. New York: IEEE, 2015: 3156-3164.

[4] XU K, BA J, KIROS R, *et al*. Show, attend and tell: neural image caption generation with visual attention[C]//International Conference on Machine Learning. San Diego: ACM , 2015: 2048-2057.

[5] CHEN Z, SING Y, CHANG T H, *et al*. Generating radiology reports via memory-driven transformer[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) . Stroudsburg: ACL, 2020: 1439-1449.

[6] JING B, XIE P, XING E. On the automatic generation of medical imaging reports[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) . Stroudsburg: ACL, 2018: 2577-2586.

[7] LI Y, LIANG X, HU Z, *et al*. Hybrid retrieval-generation reinforced agent for medical image report generation[C]//Advances in Neural Information Processing Systems. California: NIPS, 2018: 1537-1547.

[8] LIU G, HSU T M H, MCDERMOTT M, *et al*. Clinically accurate chest x-ray report generation[C]//Machine Learning for Healthcare Conference. New York: Proceedings of Machine Learning Research (PMLR) , 2019: 249-269.

[9] RENNIE S J, MARCHERET E, MROUEH Y, *et al*. Self-critical sequence training for image captioning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2017: 7008-7024.

[10] 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 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2017: 375-383.

[11] 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 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 6077-6086.

[12] FARHADI A, HEJRATI M, SADEGHI M A, *et al*. Every picture tells a story: generating sentences from images[C]// Computer Vision–ECCV 2010: 11th European Conference on Computer Vision. Berlin Heidelberg: Springer, 2010: 15-29.

[13] MITCHELL M, DODGE J, GOYAL A, *et al*. Midge: generating image descriptions from computer vision detections [C]//Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg: ACL, 2012: 747-756.

[14] GONG Y, WANG L, HODOSH M, *et al*. Improving image-sentence embeddings using large weakly annotated photo collections[C]//Computer Vision–ECCV 2014: 13th European Conference. Berlin Heidelberg: Springer, 2014: 529-545.

[15] GUPTA A, VERMA Y, JAWAHAR C. Choosing linguistics over vision to describe images[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI, 2012, 26(1) : 606-612.

[16] KRAUSE J, JOHNSON J, KRISHNA R, *et al*. A hierarchical approach for generating descriptive image paragraphs[C]// Proceedings of the IEEE Conference on Computer Vision and Ppattern Recognition. New York: IEEE, 2017: 317-325.

[17] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need[C]//Advances in Neural Information Processing Systems. La Jolla: NIPS, 2017: 30.

[18] YU J, LI J, YU Z, *et al*. Multimodal transformer with multi-view visual representation for image captioning [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(12): 4467-4480.

[19] LI Y, LIANG X, HU Z, *et al*. Hybrid retrieval-generation reinforced agent for medical image report generation[J]. Advances in Neural Information Processing Systems. La Jolla: NIPS, 2018, 31.

[20] WANG X, PENG Y, LU L, *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. New York: IEEE, 2018: 9049-9058.

[21] YIN C, QIAN B, WEI J, *et al*. Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network[C]//2019 IEEE International Conference on Data Mining (ICDM) . New York: IEEE, 2019: 728-737.

[22] LI C Y, LIANG X, HU Z, *et al*. Knowledge-driven encode, retrieve, paraphrase for medical image report generation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2019, 33: 6666-6673.

[23] ZHANG Y, WANG X, XU Z, *et al*. When radiology report generation meets knowledge graph[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2020, 34: 12910-12917.

[24] YOU D, LIU F, GE S, *et al*. Aligntransformer: hierarchical alignment of visual regions and disease tags for medical report generation[C]//Medical Image Computing and Computer Assisted Intervention-MICCAI 2021. Berlin Heidelberg: Springer, 2021: 72-82.

[25] IRVIN J, RAJPURKAR P, KO M, *et al*. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2019, 33(1) : 590-597.

[26] ALSENTZER E, MURPHY J, BOAG W, *et al*. Publicly available clinical BERT embeddings[C]//Proceedings of the 2nd Clinical Natural Language Processing Workshop. Stroudsburg: ACL, 2019: 72-78.

[27] 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: 304-310.

[28] JOHNSON A E W, POLLARD T J, GREENBAUM N R, *et al*. MIMIC-CXR:a large publicly available database of labeled chest radiographs[J]. arXiv: 1901.07042 (2019-11-14) [2024-03-18].https://arxiv.org/abs/1901.07042.

[29] JING B, WANG Z, XING E. Show, describe and conclude: on exploiting the structure information of chest x-ray reports[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2019: 6570-6580.

[30] LIU F, GE S, WU X. Competence-based multimodal curriculum learning for medical report generation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) . Stroudsburg: ACL, 2021: 3001-3012.

[31] CHEN Z, SHEN Y, SONG Y, *et al*. Cross-modal memory networks for radiology report generation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) . Stroudsburg: ACL, 2021: 5904-5914.

[32] LIU F, WU X, GE S, *et al*. Exploring and distilling posterior and prior knowledge for radiology report generation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE, 2021: 13753-13762.

[33] 谭立玮, 张淑军, 韩琪等. 面向医学影像报告生成的门归一化编解码网络[J]. 智能系统学报, 2024, 19: 411-419.
TAN L W, ZHANG S J, HAN Q. Gate normalized encoder-decoder network for medical image report generation [J]. CAAI Transactions on Intelligent Systems, 2024, 19: 411-419.

[34] WANG J, BHALERAO A, HE Y. Cross-modal prototype driven network for radiology report generation[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland. Berlin Heidelberg: Springer, 2022: 563-579.

[35] 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. Somerset: ACL, 2002: 311-318.

[36] 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. Stroudsburg: ACL, 2011: 85-91.

[37] LIN C Y. Rouge: a package for automatic evaluation of summaries[C]//Text Summarization Branches Out. Barcelona: ACL, 2004: 74-81.

（责任编辑：张玮欣　英文审核：熊荣斌）