



面向医学影像报告生成的门归一化编解码网络

谭立玮, 张淑军, 韩琪, 郭淇, 王鸿雁

引用本文:

谭立玮,张淑军,韩琪,郭淇,王鸿雁. 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(2): 411–419.
TAN Liwei, ZHANG Shujun, HAN Qi, et al. Gate normalized encoder–decoder network for medical image report generation[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(2): 411–419.

在线阅读 View online: <https://dx.doi.org/10.11992/tis.202207013>

您可能感兴趣的其他文章

基于二进制生成对抗网络的视觉回环检测研究

Visual loop closure detection based on binary generative adversarial network
智能系统学报. 2021, 16(4): 673–682 <https://dx.doi.org/10.11992/tis.202007007>

融合迁移学习的AlexNet神经网络不锈钢焊缝缺陷分类

Welding defect classification of stainless steel based on AlexNet neural network combined with transfer learning
智能系统学报. 2021, 16(3): 537–543 <https://dx.doi.org/10.11992/tis.202005013>

基于生成对抗网络的机载遥感图像超分辨率重建

Super-resolution reconstruction of airborne remote sensing images based on the generative adversarial networks
智能系统学报. 2020, 15(1): 74–83 <https://dx.doi.org/10.11992/tis.202002002>

基于小样本学习的LCD产品缺陷自动检测方法

An automatic small sample learning-based detection method for LCD product defects
智能系统学报. 2020, 15(3): 560–567 <https://dx.doi.org/10.11992/tis.201904020>

基于双向消息链路卷积网络的显著性物体检测

Salient object detection based on bidirectional message link convolution neural network
智能系统学报. 2019, 14(6): 1152–1162 <https://dx.doi.org/10.11992/tis.201812003>

词边界字向量的中文命名实体识别

Chinese named entity recognition via word boundary based character embedding
智能系统学报. 2016, 11(1): 37–42 <https://dx.doi.org/10.11992/tis.201507065>

DOI: 10.11992/tis.202207013

网络出版地址: <https://link.cnki.net/urlid/23.1538.TP.20231116.1207.004>

面向医学影像报告生成的门归一化编解码网络

谭立玮¹, 张淑军², 韩琪², 郭淇¹, 王鸿雁³(1. 青岛科技大学信息科学技术学院, 山东 青岛 266061; 2. 青岛科技大学数据科学学院, 山东 青岛 266061;
3. 青岛市干部保健服务中心, 山东 青岛 266071)

摘要: 医学影像报告的自动生成可以减轻医生的工作强度, 减少误诊或漏诊的情况发生。由于医学影像的独特性, 通常病灶比较小, 与正常区域灰度差异难以分辨, 导致文本生成时关键词的缺失, 报告不够准确。对此提出一种面向医学影像报告生成的门归一化编解码网络, 通过门控通道变换单元优化视觉特征提取, 加强特征间的差异, 自动筛选关键特征; 提出门归一化算法, 沿通道维度整合上下文信息, 在浅层网络激活、深层网络抑制通道间神经元活性, 过滤无效特征, 使文本和视觉语义充分交互, 提高报告生成质量。在 2 种广泛使用的基准数据集 IU X-Ray 和 MIMIC-CXR 上的试验结果表明, 模型能够取得先进的性能, 生成的影像报告也具有更好的视觉语义一致性。

关键词: 医学影像处理; 文本处理; 特征提取; 信息融合; 通道编码; 深度学习; 报告生成器; 灰度差异

中图分类号: TP391.4; R445 **文献标志码:** A **文章编号:** 1673-4785(2024)02-0411-09

中文引用格式: 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(2): 411-419.

英文引用格式: TAN Liwei, ZHANG Shujun, HAN Qi, et al. Gate normalized encoder-decoder network for medical image report generation[J]. CAAI transactions on intelligent systems, 2024, 19(2): 411-419.

Gate normalized encoder-decoder network for medical image report generation

TAN Liwei¹, ZHANG Shujun², HAN Qi², GUO Qi¹, WANG Hongyan³

(1. School of Data Science, Qingdao University of Science and Technology, Qingdao 266061, China; 2. College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China; 3. Qingdao Cadre Health Care Service, Qingdao 266071, China)

Abstract: Automatic generation of medical image reports can alleviate the workload of doctors and reduce the rate of misdiagnosis or missed diagnosis. Because of the uniqueness of medical images, lesions are usually small, and the gray difference between them and normal areas is hard to differentiate, resulting in loss of keywords in text generation and inaccurate reporting. Herein, a gated normalized encoder-decoder network for medical image report generation is developed, which optimizes visual feature extraction through the gated channel transformation unit, enhances the difference between features, and automatically screens key features. A gate normalization algorithm is designed to combine contextual information along with the channel dimensions, activate the neurons between channels in the shallow network, inhibit the neuron activity in the deep network, and filter invalid features, allowing full interaction between text and visual semantics to enhance the quality of report generation. Experimental results on two widely used reference datasets, IU X-Ray and MIMIC-CXR, reveal that the model can achieve advanced performance and generate image reports with better visual semantic consistency.

Keywords: medical image processing; text processing; feature extraction; information fusion; channel coding; deep learning; report generator; gray difference

随着计算机和现代医学成像技术的高速发

展, 出现了包含各种不同类型信息的图像, 如 CT (computer tomography)、MRI (magnetic resonance imaging) 等。医学影像与报告可以为临床医生提供病人详细的病情描述, 在诊断与治疗过程中起着至关重要的作用。影像形成之后, 通常先由影

收稿日期: 2022-07-11. 网络出版日期: 2023-11-17.

基金项目: 山东省高等学校青创人才引育计划“人工智能与医学影像分析创新团队”建设项目.

通信作者: 张淑军. E-mail: zhangsj@qust.edu.cn.

©《智能系统学报》编辑部版权所有

像科医生对其进行解读并撰写相应的报告,由于一位患者会得到数百张甚至上千张CT或MRI图像,撰写影像报告工作量大,费时乏味容易倦怠而导致误诊或漏诊。另一方面,由于医生个人经验等主观因素的影响,不同的医生可能对同一患者的影像产生不同诊断结果^[1]。随着硬件设备的发展,医学影像的爆炸式增长明显增加了影像医师的阅片量,这无疑加大了医生的工作强度^[2]。为了减轻医生繁重的工作负担,有效提高医生的工作效率,医学影像报告的自动生成成为临床实践中亟待突破的一项重要任务。

近年来,深度学习在各种计算机视觉任务中的突出表现,推动了其在医学影像识别中的发展,使深度学习技术处理医学影像成为一个重要的研究方向。目前,医学影像报告自动生成最流行的方法是使用编码器解码器结构,编码器通过卷积神经网络(convolutional neural network, CNN)产生视觉特征,解码器分析得到的视觉特征,结合文本输入,自动生成报告。解码器通常采用循环神经网络(recurrent neural network, RNN)。由于医学影像的特殊性,这种方法所生成的报告质量是不够的,因为这种方法旨在用简短的句子简要描述视觉场景,视觉特征提取不充分,语义信息不丰富,视觉与语义特征无法进行很好的对齐,生成的描述不清晰。然而在医学领域中,提供准确的临床描述的能力是最重要的,这对报告生成过程提出了更高的要求。

近年来,归一化层被广泛应用于神经网络中,使每个神经层的输入分布在训练过程中保持一致,加速网络收敛。局部响应归一化(local response normalization, LRN)^[3]计算每个像素的通道间小邻域的统计数据,批处理归一化(batch normalization, BN)在批次维度计算全局空间信息,层归一化(layer normalization, LN)沿着通道维度而不是批次维度进行归一化,组归一化(group normalization, GN)^[4]将通道分为不同的组,并在每组内计算归一化的均值和方差。与LRN、LN和GN类似,门控通道变换单元(gated channel transformation, GCT)^[5]也采用归一化的方法来建立通道间的竞争与合作关系,深入挖掘视觉特征隐含的本质。

本研究设计了一种面向医学影像报告生成的门归一化编解码网络,基于门控通道变换单元改进卷积神经网络,更好地从医学影像中提取关键视觉特征;提出门归一化算法来改进Transformer编解码网络,构建不同模态特征间的合作与竞争关系,最终生成更加准确、完整的报告。

在2个基准数据集IU X-Ray^[6]和MIMIC-CXR^[7]上都达到了先进的性能。

1 相关工作

1.1 基于图像描述的报告生成方法

基于深度学习的医学报告生成技术最初采用图像生成文本描述的方法,大多使用编码器-解码器框架。主要步骤是通过使用卷积神经网络(常用的网络结构如ResNet、DenseNet等)提取图像的特征,再将得到的特征输入循环神经网络,通过解码生成文本描述。

Kisilev等^[8]最早尝试医学影像到报告的自动生成,通过半自动分割方法确定病灶边界,随后使用支持向量机(support vector machine, SVM)为每个病灶生成语言标签,最后将标签嵌入到诊断句子的模板中。在随后的工作中,Kisilev等^[9]使用卷积神经网络取代支持向量机为每个病灶排序并生成语言标签,并套用诊断句子模板。Shin等^[10]率先将编码器-解码器框架方法应用在医学影像注释生成,模型使用GoogleNet卷积网络作为图像编码器,结合循环神经网络中的长短期记忆(long short-term memory, LSTM)和门控循环单元(gate recurrent unit, GRU)作为解码器,可以改善图像注释结果,但是生成的注释最长只有5个单词,更接近于术语集,而非流畅连贯的影像报告。

因此,普通图像描述生成侧重于用少量词语描述图像,而医学文本报告需要生成序列、上下文相关的段落描述,这就需要更深入地挖掘视觉特征与生成文本之间的深层联系。

1.2 基于注意力机制的报告生成方法

现有计算机辅助诊断方法普遍的缺点是无法用语义和视觉上直观的方式解释模型的预测,而这又是医生和患者都非常关注的问题。为此,注意力机制被用于医学影像报告生成的研究中。Zhang等^[11]提出MDNet模型是最先应用注意力机制来改进编解码框架的,旨在从医学影像和对应的报告中,学习句子词语和图像像素的直接映射。MDNet模型在膀胱癌病理图像及其诊断报告数据集上的表现优于基准模型。

Wang等^[12]提出了一种多任务学习的编解码器框架,可以联合完成影像标签的预测和报告的生成。编码器采用在ImageNet数据集上预训练的VGG-19模型;在解码器过程,即报告的生成过程中,协同注意力机制辅助定位异常区域,层次LSTM为异常区域生成文本报告,停止模块控制句子的数量,在IU X-Ray和PEIR Gross数据集上验证了该方法的有效性。而Liao等^[13]尝试从

解码器的隐藏层状态中提取额外的图像特征,模型使用基于 ResNet 的卷积神经网络对图像进行编码,并使用结合多级注意力机制的单层 LSTM 进行解码。该模型检测出影像中的病灶区域进而生成诊断报告,并使用影像和报告预测疾病分类。通过在胸部 CT 数据集上的试验验证了该方法的有效性。而 Xue 等^[14]优化了解码器部分,采用递归注意力机制的编解码器框架。解码器采用分层的循环神经网络,先根据图像特征生成报告的总结,再使用递归注意力机制结合图像特征产生报告的具体内容。Yang 等^[15]提出了自适应多模态注意网络在乳腺癌数据集上生成了较高质量的影像报告。

即使加入注意力机制,以上工作仍主要基于图像描述的方法,因此,训练的模型没有充分考虑医学影像和报告的独特性。

1.3 基于 Transformer 的报告生成方法

2017 年, Google 团队提出 Transformer^[16], 随后在自然语言处理领域表现出优异的性能, Transformer 在报告生成领域也取得了良好的效果。Alfarghaly 等^[17]提出了一种名为 CDGPT2 的基于条件 Transformer 的模型,使用预先训练的变压器来消除词汇选择和标点处理等问题。Chen 等^[18]提出用内存驱动的 Transformer 生成放射学报告,其中设计了一个关系存储器来记录生成过程的关键信息,并应用内存驱动的条件层标准化将存储器纳入 Transformer 的解码器。Chen 等^[19]还提出了一个跨模态记忆网络,以增强编解码器框架的放射学报告生成,其中共享记忆被设计来记录图像和文本之间的对齐,以便促进跨模式的交互和生成。

Liu 等^[20]提出了对比注意(contrastive attention, CA)模型和后验与先验知识探索与蒸馏方法

(posterior-and-prior knowledge exploring-and-distilling approach, PPKED)^[21]。其中, CA 模型将当前输入图像与正常图像进行比较,提取对比信息,因为这些信息更能代表异常区域的视觉特征,而 PPKED 模仿了放射科医生的工作模式,首先检查异常区域,并为异常区域分配疾病主题标签,然后依靠多年的医学知识和工作经验撰写报告。

综上,本研究在 Transformer 架构之下,通过门归一化算法进行优化改进,以深层建模视觉特征与文本序列的关联,进一步提高报告生成的质量和效率。

2 试验方法

本研究提出的网络总体结构如图 1 所示,主要包含 2 个阶段:视觉特征提取阶段及编解码阶段。视觉提取阶段完成“读图分析”功能,提取出特征向量送入编解码阶段,对图像视觉信息及训练数据的报告文本(对应图 1 中的 Ground Truth)进行编码和解码,最后生成一份预测的报告。

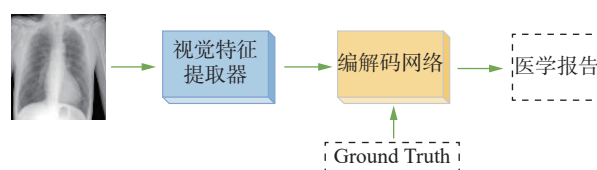


图 1 网络结构

Fig. 1 Network structure

该网络详细结构如图 2 所示,选取 ResNet101 作为视觉特征提取阶段的主干网络,通过门控通道变换改进 ResNet,提出 GCT-ResNet(gated channel transformation ResNet),充分提取视觉特征。在编解码阶段,使用多个所提出的门归一化模块优化 Transformer 模型,使编码端能够获取更详细的上下文信息,解码端更好地对齐视觉与文本语义。

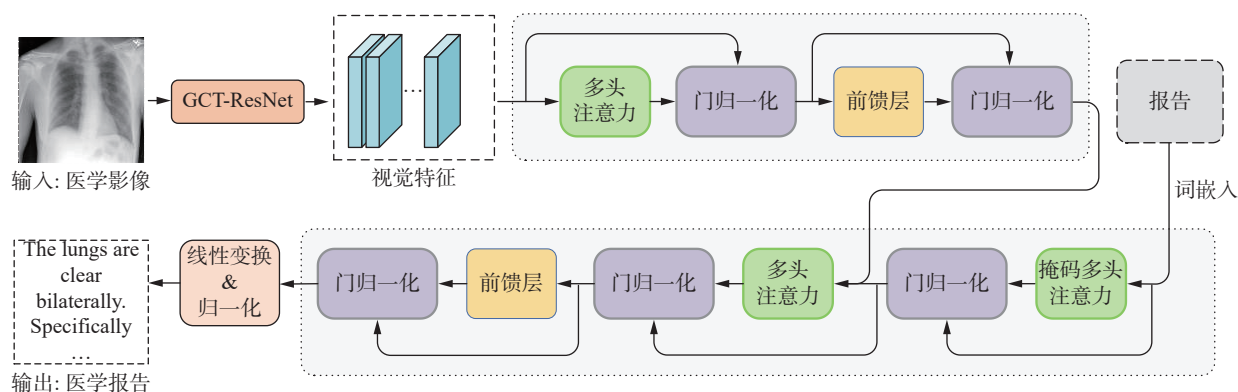


图 2 门归一化编解码网络

Fig. 2 Gate normalized encoder-decoder network

2.1 基于门控通道变换的特征提取

通常医学影像报告自动生成采用在自然图像

数据上预训练好的深层卷积神经网络提取视觉特征,例如 ResNet、DenseNet 等,这忽略了医学图像

的特殊性,效果不尽人意。不同于自然图像中各种类别间特征差异明显,医学图像中病灶关系复杂,且不易区分。

为了进一步建模视觉特征间的关系,提升深层卷积神经网络的性能,本研究采用文献[5]中 GCT 的思想,提出了一种基于门控通道变换改进的 ResNet 网络(GCT-ResNet),并在此网络上使用胸部数据集进行训练作为特征提取模块。GCT-ResNet 结构如图 3 所示,其中 ResNet 核心由多个残差块(residual block)组成,每个 Block 包含 Conv、Batch Norm、Relu 激活函数。ResNet 使用残差连接结构避免网络层数深导致的梯度消失或梯度爆炸问题。

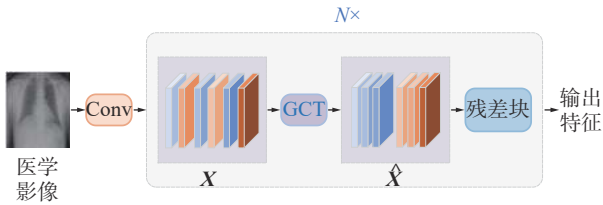


图 3 GCT-ResNet 网络结构

Fig. 3 GCT-ResNet network structure

在每个残差块之前都设置一个 GCT 单元,以便在残差块之间对通道间特征进行调整转换,显式地建模通道间特征的竞争与合作关系。设输入视觉特征 X , 经过 GCT 作用输出 \hat{X} , 在浅层残差块, GCT 倾向于减小通道之间特征的差异,由图 3 可知, \hat{X} 中具有相近颜色的特征聚合在一起,这种行为有利于促进通道间的合作关系,避免激活值过多或有用特征的丢失,缓解过拟合。在深层残差块, GCT 倾向于增加通道间的差异,图 3 表现为 \hat{X} 中色差明显的特征之间距离较远,因为在靠近输出时,通道间特征差异大,可以使每个特征具有更强的辨别性和任务相关性。因此,本研究提出的 GCT-ResNet 可以充分提取医学图像中的病灶特征,任务导向性更强。

2.2 门归一化算法

通常归一化可以缓解深层神经网络训练中的内部协变量偏移问题,即由于浅层网络中神经元参数的更新导致深层网络输入数据分布发生变化。其中,层归一化可以有效解决批归一化中依赖批大小的问题,但由于通道与通道间交互信息不足,易导致无效特征数量增加、语义信息冗余问题,应用在医学影像报告生成中,会出现重复语句,降低辅助诊断的有效性。为解决上述问题,本研究提出一种门归一化(Gate-Norm)算法,由图 4 可知,通过设置全局上下文嵌入、L2 正则化和门控机制,加强语义通道间信息交互,过滤无效特征,并使用残差思想,以替代编解码结构

中的层归一化处理。

门归一化算法如图 4 所示,设 $Y \in \mathbf{R}^{C \times D}$ 是注意力机制输出的语义特征,其中 C 为通道数、 D 为词向量维度, $Y = [y_1 \ y_2 \ \cdots \ y_c]$, y_c 对应于 Y 的单个通道, $c \in \{1, 2, \dots, C\}$ 。

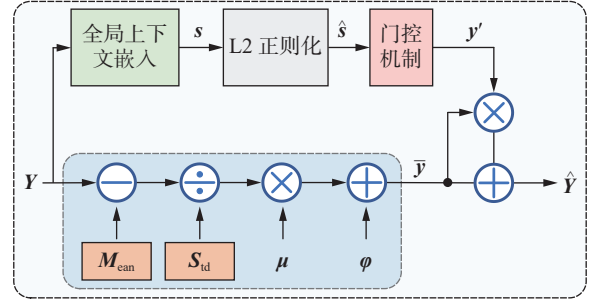


图 4 门归一化算法模块

Fig. 4 Module of gate normalization algorithm

使用一种全局上下文嵌入将全局上下文信息聚合到每个通道中。给定嵌入权值 $\alpha = [\alpha_1 \ \alpha_2 \ \cdots \ \alpha_c]$ 定义如下

$$s_c = \alpha \parallel y_c \parallel_2 = \alpha \left\{ \left[\sum_{i=1}^D (y_c^i)^2 \right] + \varepsilon \right\}^{\frac{1}{2}} \quad (1)$$

式中: s_c 是聚合后的全局上下文信息向量; i 是词向量维度; ε 是一个很小的常数,以避免在零点处求导的问题。在聚合通道信息时,使用正则化避免输出永远是一个常量的极端情况发生。

在获取全局上下文信息向量 (s_c) 后对其进行 L2 正则化操作, L2 正则化具有计算资源轻量化、训练性能稳定的特点,可跨通道操作,公式如下

$$\hat{s}_c = \frac{\sqrt{C} s_c}{\parallel s_c \parallel_2} = \frac{\sqrt{C} s_c}{\left[\left(\sum_{c=1}^C s_c^2 \right) + \varepsilon \right]^{\frac{1}{2}}} \quad (2)$$

式中: \hat{s}_c 是 L2 正则化后的输出; ε 是一个很小的常数; 标量 \sqrt{C} 是为了避免通道数 C 太大时 \hat{s}_c 太小。

门归一化设置门控机制自动调整原始特征,促进通道特征间的竞争合作关系,设权值 $\gamma = [\gamma_1 \ \gamma_2 \ \cdots \ \gamma_c]$, 偏置 $\beta = [\beta_1 \ \beta_2 \ \cdots \ \beta_c]$, 门控适应机制公式为

$$y'_c = \tanh(\gamma_c \hat{s}_c + \beta_c) \quad (3)$$

式中: y'_c 是输出的门控权值,每个原始通道 y_c 的大小由其对应的门控来调整; 权值 γ 和偏置 β 是可训练的,用于学习控制门通道的激活,通过结合归一化方法和门控适应机制,对不同通道之间更多类型的关系(如竞争与合作)进行建模。当一个通道的权值 γ_c 被正向激活时,会加大该通道与其他通道间的神经元差异,处于竞争关系。当门控权值被负向激活时,各通道神经元之间差异减小,用于提取完整的全局信息。

同时为了缩短网络收敛时间,防止梯度消失

和梯度爆炸的情况发生, 原始语义特征 y_c 经过层归一化得到 \bar{y}_c , 公式如下

$$\bar{y}_c = \mu \left[\frac{y_c - M_{\text{can}}}{S_{\text{id}} + \varepsilon} \right] + \varphi \quad (4)$$

式中: ε 是一个很小的常数; M_{can} 代表通道均值; S_{id} 代表通道标准差; μ 、 φ 是设置的可学习的参数。

为了避免输入输出流之间没有直流通路, 梯度流会被层归一化模块阻断, 出现顶层梯度消失的情况。结合残差思想, 将归一化后的特征 \bar{y}_c 与门控权值 y'_c 相乘, 输出转换后的特征 \hat{y}_c , 公式如下

$$\hat{y}_c = \bar{y}_c [1 + y'_c] \quad (5)$$

门归一化算法结合门机制与归一化操作在训练过程中整合特征分布空间, 构建神经元之间的多种关系, 利用门控权值过滤无效特征, 加强有效特征间的差异, 使文本视觉信息充分交互, 完整映射视觉语义关系。

2.3 基于门归一化的编解码处理

Transformer 模型主要由多头注意力机制 (multi-head attention, MHA)、前馈层 (feed forward, FF)、层归一化 (layer norm, LN) 及残差结构组成, 其以 MHA 为核心, 可以获取全局及局部信息, 并行化处理有助于提升训练速度。但由于 Transformer 过于关注全局信息, 局部信息的获取不如 CNN 与 RNN 有效, 对报告生成任务而言, 只用层归一化整合数据不够充分, 通道特征关系散乱, 无效特征扰乱正确输出结果。

因此, 本研究用提出的门归一化算法对 Transformer 编解码网络进行优化 (图 2), 在每个多头注意力机制层以及前馈层之后设置门归一化, 构建多模态多通道间特征的竞争与合作关系, 关注视觉与文本通道特征间的局部信息, 改善 Transformer 局部信息获取不充分的问题, 过滤无效特征。同时归一化重新整合视觉及语义信息在特征空间的分布, 使多头注意力对视觉与语义信息完整映射, 让医学影像中的病灶信息对齐报告中的句子信息, 实现更精准的文本预测。

3 试验分析

本节将详细阐述所用的公共数据集以及试验结果, 对提出的方法进行详细的对比和分析。

3.1 数据集与试验参数

使用 2 个基准数据集进行试验: 印第安纳大学的 IU X-Ray^[6] 和 Beth Israel Deaconess 医疗中心的 MIMIC-CXR^[7]。前者是一个相对较小的数据集, 有 7470 张胸片和 3955 份相关报告, 每位患者的 2 张胸片对应一份报告; 后者是目前最大的公共放射学数据集, 有 473 057 张胸片和 206 563 份

报告。每份数据都包含图像与报告 2 部分。对于 IU X-Ray, 使用与 Chen 等^[18] 所描述的相同的划分 (即训练/验证/测试集的 70%/10%/20%), 而对于 MIMIC-CXR, 采用其官方的数据分割方法。

视觉提取器 (GCT-ResNet) 以 ResNet101 作为基础网络, 通过在每个残差 Block 前设置一个 GCT 单元, 重新整合输入分布, 最终提取出维度为 (49×512) 的 patch 特征, 具体网络参数见表 1。

表 1 GCT-ResNet 网络结构参数
Table 1 Parameters of GCT-ResNet

GCT-ResNet	卷积尺寸	输入尺寸	输出尺寸
Conv	7×7,64	(3,224,224)	(64,112,11)
Block1	$\begin{bmatrix} \text{GCT} \\ 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 128 \end{bmatrix} \times 3$	(64,112,112)	(256,56,56)
Block2	$\begin{bmatrix} \text{GCT} \\ 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 128 \end{bmatrix} \times 4$	(256,56,56)	(512,28,28)
Block3	$\begin{bmatrix} \text{GCT} \\ 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 128 \end{bmatrix} \times 23$	(512,28,28)	(1024,14,14)
Block4	$\begin{bmatrix} \text{GCT} \\ 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 128 \end{bmatrix} \times 3$	(1024,14,14)	(2048,7,7)
Reshape & Linear	(2048,512)	(49,2048)	(49,512)

对于编码器-解码器主干, 使用具有 4 层和 8 个注意头的 Transformer 结构, 隐藏状态有 512 维, 并对其进行随机初始化。使用 Adam 优化器在交叉熵损失下训练本研究的模型。视觉提取器的学习速率设置为 5×10^{-5} , 对所有数据集, 设置以 0.8 的速率每轮衰减, 在报告生成过程中, 为了增加生成句子的多样性, 使用集束搜索策略 (beam search), 将 beam size 设置为 3, 以平衡所有模型的有效性和效率。共训练 100 个 Epoch, BatchSize 值设置为 16, 训练设备选取具有 24 GB 显存的 NVIDIA RTX A5000 显卡。

3.2 评价指标

BLEU- n (bilingual evaluation understudy) 双语评估替补^[22] 是报告生成领域最常用的评价指标, 用来测量计算生成的报告和 GT 报告之间的精确率, 包括 BLEU-1、BLEU-2、BLEU-3、BLEU-4 等 4 种, 其中 n 指的是连续的单词个数。BLEU-1 衡量的是单词级别的准确性, 更高阶的 BLEU- n 可

以衡量句子的流畅度。其公式如下：

$$B_n = t \times \exp \left(\sum_{n=1}^N W_n \times \log P_n \right) \quad (6)$$

式中： B_n 代表 BLEU- n ； t 是惩罚因子； P_n 指 n 个连续单词的精确率； W_n 为其权重，一般设置为均匀权重，即 $W_n = 1/N$ 。

METEOR^[23] 建立在 BLEU-1 之上，更突出召回率的重要性，使用词干和同义词匹配，更灵活地评价文本的流畅性，公式如下：

$$M_{\text{METEOR}} = (1 - P_{\text{en}}) \cdot \frac{P \cdot R}{\alpha \cdot P + (1 - \alpha) \cdot R} \quad (7)$$

式中： P_{en} 为碎片性惩罚因子，越小越好； P 与 R 分别表示单词组的准确率和召回率； α 是超参数。

ROUGE-L^[24] 计算最长公共子序列的重合率，反映句子级别的准确率，公式如下：

$$R_L = \frac{(1 + \beta^2) R_{\text{LCS}} P_{\text{LCS}}}{R_{\text{LCS}} + \beta^2 P_{\text{LCS}}} \quad (8)$$

式中： R_L 表示 ROUGE-L； R_{LCS} 表示召回率； P_{LCS} 表示精确率， β 为超参数。

3.3 试验结果

为证明本研究所提出方法的有效性，对比了各种先进的医学报告生成模型，包括 HRNN^[25]、CoAtt^[26]、CMAS-RL^[27]、Transformer^[18]、R2Gen^[18]、CMN^[19]、CA^[20]、PPKED^[21] 和 Relation-paraNet^[28]。

此外还与一般的图像字幕方法进行了比较，如 S&T^[29]、SA&T^[30]、AdaAtt^[31]、Att2in^[32] 和 Up-Down^[33]，结果如表 2 所示，所有指标值越大说明效果越好。由表 2 可见，与一般的图像字幕方法（表 2 的前 4 行 1/2/3/4）相比，放射学报告生成模型（表 2 其余行所给出的方法）有明显的改进，这表明进行领域特定的放射学报告生成研究是必要且重要的。与最先进的模型 Relation-paraNet^[28] 相比，在 IU X-Ray 数据集上，本研究模型的评价指标 BLEU-4 分数提升了 7.1%、ROUGE-L 分数上提升了 3.8%。本研究模型采用 Transformer 作为基本结构，既可以同时编码双向的语义还可以抽取长距离的特征，所以在上下文特征抽取方面强于 LSTM，模型中的门归一化使通道间特征联系更加紧密，所以生成的报告具有更好的流畅性、完整性。BLEU-1 和 BLEU-2 分数略低于 Relation-paraNet，是因为这 2 个指标侧重单个词语和双词这种较短词句的生成，是局部指标，而本研究的基本框架 Transformer 模型在全局语义方面更有优势。在 MIMIC-CXR 数据集上的效果与最先进模型 CA 与 PPKED 相近，因为 CA 使用了模板化生成报告，PPKED 嵌入了多种先验知识提高了报告生成效果，将医生经验作为先验知识加入网络也是下一步的研究工作。

表 2 各模型试验结果对比

Table 2 Comparative experimental results of different models

数据集	年份	模型	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
IU X-Ray	2015	S&T ^[29]	0.216	0.124	0.087	0.066	—	0.306
	2015	SA&T ^[30]	0.399	0.251	0.168	0.118	—	0.323
	2017	AdaAtt ^[31]	0.220	0.127	0.089	0.068	—	0.308
	2017	Att2in ^[32]	0.224	0.129	0.089	0.068	—	0.308
	2017	HRNN ^[25]	0.439	0.281	0.190	0.133	—	0.342
	2018	CoAtt ^[26]	0.455	0.288	0.205	0.154	—	0.369
	2019	CMAS-RL ^[27]	0.464	0.301	0.210	0.154	—	0.362
	2020	Transformer ^[18]	0.396	0.254	0.179	0.135	0.164	0.342
	2020	R2Gen ^[18]	0.470	0.304	0.219	0.165	0.187	0.371
	2021	CMN ^[19]	0.475	0.309	0.222	<u>0.170</u>	0.191	0.375
	2021	CA ^[20]	0.492	0.314	0.222	0.169	0.193	<u>0.381</u>
	2021	PPKED ^[21]	0.483	0.315	0.224	0.168	0.190	0.376
MIMIC-CXR	2021	Relation-paraNet ^[28]	0.505	0.329	<u>0.230</u>	0.168	—	0.372
	2022	本研究模型	<u>0.498</u>	<u>0.319</u>	0.231	0.181	0.212	0.386
	2015	S&T ^[29]	0.256	0.157	0.102	0.070	—	0.249
	2015	SA&T ^[30]	0.304	0.177	0.112	0.077	—	0.249
	2017	AdaAtt ^[31]	0.299	0.185	0.124	0.088	0.118	0.266
	2017	Att2in ^[32]	0.325	0.203	0.136	0.063	0.134	0.276
	2018	Up-Down ^[33]	0.317	0.195	0.130	0.092	0.128	0.267
	2020	Transformer ^[18]	0.314	0.192	0.127	0.090	0.125	0.265
	2020	R2Gen ^[18]	0.353	0.218	0.145	0.103	0.142	0.277

续表 2

数据集	年份	模型	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
MIMIC-CXR	2021	CMN ^[19]	0.353	0.218	0.148	0.106	0.142	0.278
	2021	CA ^[20]	0.350	0.219	0.152	0.109	0.151	0.283
	2021	PPKED ^[21]	<u>0.360</u>	0.224	0.149	0.106	0.149	<u>0.284</u>
	2022	本研究模型	0.361	<u>0.223</u>	<u>0.150</u>	<u>0.108</u>	<u>0.150</u>	0.287

注: 黑色加粗表示最高, 黑色下划线表示次高。

3.4 消融试验分析

为了探究门归一化算法的作用, 以 Chen 等^[18]复现的 ResNet101 + Transformer 的结果作为 Baseline^[18], 使用 GCT-ResNet(GR)、门归一化(Gated-Norm)作为 2 个替代模块, 在 IU X-Ray 与 MIMIC-CXR 2 个数据集进行了消融试验, 结果如表 3 所示。其中, 行 1、行 5 为 Baseline 的结果, 行 2、行 6 表示在 Baseline 的基础上, 只使用 GCT-ResNet 替代原始的特征提取网络, Transformer 不做更改。行 3、行 7 表示仅使用门归一化来改进原始 Transformer, 行 4、行 8 代表提出的完整模型, 即视觉提取模块使用 GCT-ResNet, 编解码使

用门归一化改进的 Transformer。

由表 3 中可以看出: 1) 仅使用 GCT-ResNet(行 2、行 6)与仅使用门归一化(行 3、行 7)的表现结果都优于 Baseline 模型, 这是因为前者改善了 ResNet 提取视觉特征的能力, 后者优化了视觉语义信息交互对齐过程; 2) 同时使用 GCT-ResNet 和门归一化(行 4、行 8)可以达到最好的结果, 这是因为门归一化具有强大的灵活性, 在视觉分析与文本生成中起到相辅相成、互相促进的作用, 可以建模多模态多通道特征间的关系, 充分提取视觉与语义信息, 优化视觉语义对齐过程, 使生成的报告更加准确、流畅、完整。

表 3 消融试验结果

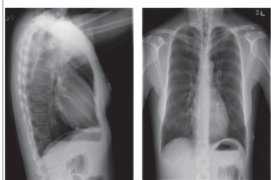
Table 3 Ablation results

数据集	行	模型	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
IU X-Ray	1	Baseline ^[18]	0.396	0.254	0.179	0.135	0.164	0.342
	2	+ GR	0.490	0.310	0.220	0.169	0.197	0.370
	3	+ Gated-Norm	0.487	0.307	0.222	0.175	0.205	0.375
	4	+ GR+ Gated-Norm	0.498	0.319	0.231	0.181	0.212	0.386
MIMIC-CXR	5	Baseline ^[18]	0.314	0.192	0.127	0.090	0.125	0.265
	6	+ GR	0.353	0.215	0.146	0.101	0.142	0.278
	7	+ Gated-Norm	0.350	0.210	0.146	0.103	0.144	0.283
	8	+ GR+ Gated-Norm	0.361	0.223	0.150	0.108	0.150	0.287

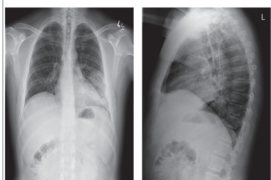
3.5 案例对比分析

选取两组图像, 使用 Baseline 模型与本研究模型生成相应的报告, 如图 5 所示, Ground Truth

代表真实的报告, Generated Report 代表模型生成的报告。斜体加粗代表报告中的异常描述, 正体加粗表示正常描述, 下划线代表预测错误的句子。

	<p>Ground Truth</p> <p>There is <i>enlargement of the cardiac silhouette</i>. There is a <i>focal opacity within the right upper lung</i>. There is <i>dense calcification of the thoracic aorta</i>. There is no pneumothorax. There is no large pleural effusion.</p>	<p>Generated Report (Baseline)</p> <p>The heart size is normal. <i>Focal opacity in the right lung</i>. <i>Calcification in the thoracic aorta</i>. No pneumothorax. No large pleural effusion.</p>	<p>Generated Report (本文)</p> <p>The heart size is mildly enlarged. There is no pneumothorax or pleural effusion. <i>Focal opacity is seen at the upper right of the lung</i>. <i>Calcifications occurs in the thoracic aorta</i>.</p>
---	---	--	---

(a) 案例 1

	<p>Ground Truth</p> <p>There is a <i>rounded dense opacity in the lateral left midlung zone</i> probably the <i>left upper lobe most suggestive of a rounded pneumonia</i>. There is no pleural effusion. The heart and mediastinum are normal. The skeletal structures are normal.</p>	<p>Generated Report (Baseline)</p> <p>The heart and mediastinum are normal. No skeletal structures abnormality. The <i>dense opacity in the middle and lateral left lung</i>. No pleural effusion.</p>	<p>Generated Report (本文)</p> <p>The heart and mediastinum are normal. There a <i>rounded dense opacity in the middle and lateral part of the left lung, in the upper left lobe</i>. There is no pleural effusion. The skeletal structures are normal.</p>
---	--	--	--

(b) 案例 2

图 5 案例对比

Fig. 5 Case comparisons

在图5中,由第1个案例可知,Baseline模型生成的报告句子短,语句不流畅,且存在错误,原始报告中描述 enlargement of the cardiac silhouette(心脏轮廓增大)为异常部分,但是Baseline模型生成的报告中描述 heart size is normal(心脏大小正常),明显与真实报告不同。而本研究模型生成的报告与真实报告句子长度接近,语句流畅,对异常部分描述准确。第2个案例,Baseline模型生成的报告中 The dense opacity in the middle and lateral left lung(左肺中外侧致密影),对比真实报告,描述 dense opacity(密集阴影)时缺少了 rounded 这样的形容词,并且并未生成 probably the left upper lobe most suggestive of a rounded pneumonia(可能在左上叶,最可能提示圆形肺炎)这样的推理描述,而病灶位置对医学诊断是非常重要的词汇。相比来说,本研究模型生成的报告中,可以推理出 the left upper lobe(左上叶)这种病灶位置信息,模型生成的报告与真实报告相接近,语句完整流畅,保持了较好的视觉文本一致性。

4 结束语

为增强医学辅助诊断的智能性和高效性,本研究提出了一种面向医学影像报告生成的门归一化编解码网络,通过门控转换单元改进卷积神经网络,设计门归一化算法,加强模型数据处理过程中各通道特征间的关系,过滤无用特征,使得提取出的视觉特征及文本特征更加详细,视觉与语义信息充分对齐。在MIMIC-CXR和IU X-Ray数据集上进行的大量的试验结果表明了本研究方法的有效性,详细的消融试验又深入分析了不同模块对模型性能的影响。下一步将引入更多医学领域的先验知识及医学相关的评价指标,进一步提升报告生成模型的可靠性与通用性。

参考文献:

- [1] 姜婷, 裴肖明, 岳厚光. 基于分布先验的半监督FCM的肺结节分类[J]. 智能系统学报, 2017, 12(5): 729-734.
JIANG Ting, XI Xiaoming, YUE Houguang. Classification of pulmonary nodules by semi-supervised FCM based on prior distribution[J]. CAAI transactions on intelligent systems, 2017, 12(5): 729-734.
- [2] 杨晓兰, 强彦, 赵涓涓, 等. 基于医学征象和卷积神经网络的肺结节CT图像哈希检索[J]. 智能系统学报, 2017, 12(6): 857-864.
YANG Xiaolan, QIANG Yan, ZHAO Juanjuan, et al. Hashing retrieval for CT images of pulmonary nodules based on medical signs and convolutional neural networks[J]. CAAI transactions on intelligent systems, 2017, 12(6): 857-864.
- [3] ROBINSON A E, HAMMON P S, DE SA V R. Explaining brightness illusions using spatial filtering and local response normalization[J]. Vision research, 2007, 47(12): 1631-1644.
- [4] WU Yuxin, HE Kaiming. Group normalization[C]//European Conference on Computer Vision. Cham: Springer, 2018: 3-19.
- [5] YANG Zongxin, ZHU Linchao, WU Yu, et al. Gated channel transformation for visual recognition[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 11791-11800.
- [6] 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, 2015, 23(2): 304-310.
- [7] JOHNSON A E W, POLLARD T J, GREENBAUM N R, et al. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs[EB/OL]. (2019-11-14)[2022-07-11]. <https://arxiv.org/abs/1901.07042.pdf>.
- [8] KISILEV P, WALACH E, BARKAN E, et al. From medical image to automatic medical report generation[J]. IBM journal of research and development, 2015, 59(2/3): 1-7.
- [9] KISILEV P, SASON E, BARKAN E, et al. Medical image description using multi-task-loss CNN[M]//Deep Learning and Data Labeling for Medical Applications. Cham: Springer International Publishing, 2016: 121-129.
- [10] SHIN H C, ROBERTS K, LU Le, et al. Learning to read chest X-rays: recurrent neural cascade model for automated image annotation[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 2497-2506.
- [11] ZHANG Zizhao, XIE Yuanpu, XING Fuyong, et al. MD-Net: a semantically and visually interpretable medical image diagnosis network[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 3549-3557.
- [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]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 9049-9058.
- [13] LIAO Fangzhou, LIANG Ming, LI Zhe, et al. Evaluate the malignancy of pulmonary nodules using the 3-D deep leaky noisy-OR network[J]. IEEE transactions on neural networks and learning systems, 2019, 30(11): 3484-3495.
- [14] XUE Yuan, XU Tao, RODNEY LONG L, et al. Multimodal recurrent model with attention for automated radiology report generation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2018: 457-466.
- [15] YANG Shaokang, NIU Jianwei, WU Jiyan, et al. Auto-

- matic ultrasound image report generation with adaptive multimodal attention mechanism[J]. *Neurocomputing*, 2021, 427: 40–49.
- [16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all You need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000–6010.
- [17] ALFARGHALY O, KHALED R, ELKORANY A, et al. Automated radiology report generation using conditioned transformers[J]. *Informatics in medicine unlocked*, 2021, 24: 100557.
- [18] CHEN Zhihong, SONG Yan, CHANG T H, et al. Generating radiology reports via memory-driven transformer[EB/OL]. (2022–04–28)[2022–07–11]. <https://arxiv.org/abs/2010.16056.pdf>.
- [19] CHEN Zhihong, SHEN Yaling, SONG Yan, et al. Cross-modal memory networks for radiology report generation[EB/OL]. (2022–04–28)[2022–07–11]. <https://arxiv.org/abs/2204.13258.pdf>.
- [20] LIU Fenglin, YIN Changchang, WU Xian, et al. Contrastive attention for automatic chest X-ray report generation[EB/OL]. (2022–01–09)[2022–07–11]. <https://arxiv.org/abs/2106.06965.pdf>.
- [21] LIU Fenglin, WU Xian, GE Shen, et al. Exploring and distilling posterior and prior knowledge for radiology report generation[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 13748–13757.
- [22] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics-ACL '02. Morristown: Association for Computational Linguistics, 2001: 311–318.
- [23] 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 Press, 2011: 85–91.
- [24] LIN C Y. Rouge: A package for automatic evaluation of summaries[C]//Proceedings of Workshop on Text Summarization Branches Out. Barcelona: ACL, 2004: 74–81.
- [25] KRAUSE J, JOHNSON J, KRISHNA R, et al. A hierarchical approach for generating descriptive image paragraphs[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 3337–3345.
- [26] JING Baoyu, XIE Pengtao, XING E. On the automatic generation of medical imaging reports[EB/OL]. (2018–07–20)[2022–07–11]. <https://arxiv.org/abs/1711.08195.pdf>.
- [27] JING Baoyu, WANG Zeya, XING E. Show, describe and conclude: on exploiting the structure information of chest X-ray reports[EB/OL]. (2020–07–23)[2022–07–11]. <https://arxiv.org/abs/2004.12274.pdf>.
- [28] WANG Fuyu, LIANG Xiaodan, XU Lin, et al. Unifying relational sentence generation and retrieval for medical image report composition[J]. *IEEE transactions on cybernetics*, 2022, 52(6): 5015–5025.
- [29] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: a neural image caption generator[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 3156–3164.
- [30] XU K, BA J, KIROUS R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]//International Conference on Machine Learning. New York: ACM Press, 2015: 2048–2057.
- [31] LU Jiasen, XIONG Caiming, PARIKH D, et al. Knowing when to look: adaptive attention via A visual sentinel for image captioning[EB/OL]. (2017–06–07)[2022–07–11]. <https://arxiv.org/abs/1612.01887.pdf>.
- [32] RENNIE S J, MARCHERET E, MROUEH Y, et al. Self-critical sequence training for image captioning[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 1179–1195.
- [33] ANDERSON P, HE Xiaodong, BUEHLER C, et al. Bottom-up and top-down attention for image captioning and visual question answering[EB/OL]. (2018–03–14)[2022–07–11]. <https://arxiv.org/abs/1707.07998.pdf>.

作者简介:



谭立玮, 硕士研究生, 主要研究方向为计算机视觉。E-mail: 2020-110009@qust.edu.cn。



张淑军, 副教授, 主要研究方向为计算机视觉、虚拟现实技术。以第一作者发表学术论文 27 篇。E-mail: zhangsj@qust.edu.cn。



韩琪, 硕士研究生, 主要研究方向为计算机视觉。E-mail: hanqi@mails.qust.edu.cn。