基于空间信息融合的肺部诊断报告生成方法

摘要

医学影像常被用作一种临床辅助工具。通过医学影像撰写放射学报告对于缺乏经验的放射科医生来说，既费时又容易出错。深度学习技术可以帮助医生减少手工撰写诊断报告的工作量，提高诊断效率和准确性。本文旨在研究基于深度学习的 X 光胸片诊断报告自动生成方法，利用卷积块注意力模块来改进 Memory-driven Transformer，模型生成的报告使用自然语言生成指标进行评估。项目所使用的数据集是目前被广泛使用的放射学报告数据集 IU X-Ray。在该数据集上的实验结果表明，相比于 Memory-driven Transformer，在加入了卷积块注意力模块后，在自然语言生成各指标上都有所提高。

关键词：医学影像；深度学习；卷积块注意力模块；自然语言生成

1 引言

近年来，随着我国经济与科学技术的发展，医疗水平也在不断提高。当前，各个行业的数字化转型如火如荼，在“健康中国 2030”，“人工智能 + 医疗”等国家战略的驱动下，医疗信息化，也进入了智能化的全面阶段。人工智能使智慧医疗朝着更加智能，更加安全，更加便捷的方向迈进，对社会发展起着积极推动作用。

X 光胸片作为最常见的医学影像之一，是医生进行肺部、心脏等相关疾病诊断的重要工具，被广泛应用于医学临床诊断和治疗中。传统的 X 光胸片诊断需要医生仔细阅读 X 光胸片，并准确地描述 X 光图像中所显示的异常情况，然后再给出相应的诊断结果。然而，这种方式存在着时间成本高、主观性强等问题，因此，如何快速而准确地对 X 光胸片进行诊断成为了医学界和计算机科学界共同关注的课题。

自动化诊断报告生成是一项需要计算机视觉和自然语言处理等多个前沿研究领域相结合的研究任务，是推动智慧医疗建设的重要方法，本文基于近年来迅速发展的深度学习方法，利用卷积神经网络得到 X 光胸片特征，将得到的特征输入序列模型生成报告，依此构建整个编解码模型框架。在一定程度上提高医生撰写报告的效率，有效地区分出正常、异常还是潜在异常，给医生提供有效的参考，尤其在降低城乡、地区间医疗水平的差距有着重大的意义。

传统的 Memory-driven Transformer 会在特征图被分割成各个 patch 之后做注意力计算，但此时已经损失了特征图的空间信息。本文利用卷积块注意力模块来保留和学习更多有助于诊断的信息。通过将其添加在视觉提取器之后，编码器之前来改进传统的 Memory-driven Transformer。改进后的 Memory-driven Transformer 生成的报告在各项自然语言生成指标上都有所提高。报告生成网络使用预训练的 ResNet101 提取 X 光胸片的视觉特征，然后将提取到的特征依次输入卷积块注意力模块、编码器和解码器，以生成诊断报告。

2 相关工作

2.1 国内外研究进展

在早期，一些研究者将传统的图像说明方法应用在了这一领域，如 2015 年 Vinyals 等人 [16] 提出的 ST 模型，2017 年 Rennie 等人 [14] 提出的 ATT2IN 模型和 2018 年 Anderson 等人 [1] 提出的 TOPDOWN 模型。这些方法的目的是用短句子简要描述图像内容。但实际上，放射学报告生成的一个重大挑战是放射学报告是由多个句子组成的长描述。如图1所示，报告通常包含发现部分和印象部分，发现部分是对 X 光胸片诊断情况的一段描述，其中包含正常和异常的情况。印象部分是对 X 光胸片诊断情况的总结。因此，传统的图像说明方法可能不足以生成满足要求的放射学报告，这对报告的生成过程提出了更高的要求。

近几年，在放射学报告生成领域提出了许多新的方法。例如 Wang 等人 [17] 在 2018 年提出了一种对 X 光胸片常见疾病分类并生成诊断报告的多任务模型。同年，Baoyu Jing 等人 [7] 提出了利用 CNN 提取图像特征，再使用序列模型生成报告。张宇等人 [19] 提出了在报告生成时采用自适应注意力，为模型提供了更多的灵活性，从而提升模型性能。2020 年，Chen 等人 [3] 提出了 Memory-driven Transformer，采用记忆驱动单元来记录报告中的模式信息，进一步提升了模型生成报告的质量。上述方法都是基于生成的方法。其中，Memory-driven Transformer 通过将 X 光片特征通过编解码器架构来生成放射学报告 [12]。总的来说，模型由三个部分组成：视觉提取器、编码器和解码器。这三个部分紧密联系：视觉提取器负责提取 X 光片特征，编码器将这些特征转换为隐藏状态，隐藏状态中包含丰富的信息；解码器对隐藏状态进行解码，输出文字报告。

放射学报告值得注意的一个重要特性是其具有高度模式化的性质。如图1所示，发现部分中有些句子会多次出现在类似的不同报告中。利用这种高度模式化的性质，有研究者提出了基于检索的方法。例如 Liu 等人 [11] 在 2019 年提出了一种检索方法可以达到不错的性能。Li 等人 [9] 在 2018 年将基于检索和基于生成的方法与手动提取的模板相结合。总的来说，就是定义好模板，然后从定义好的模板数据库中进行检索，以生成标准化的放射学报告。虽然基于检索的方法拥有不错的性能，但在面对大数据量时，其前期的准备工作太过繁杂。因此，该方法也有很多局限性。

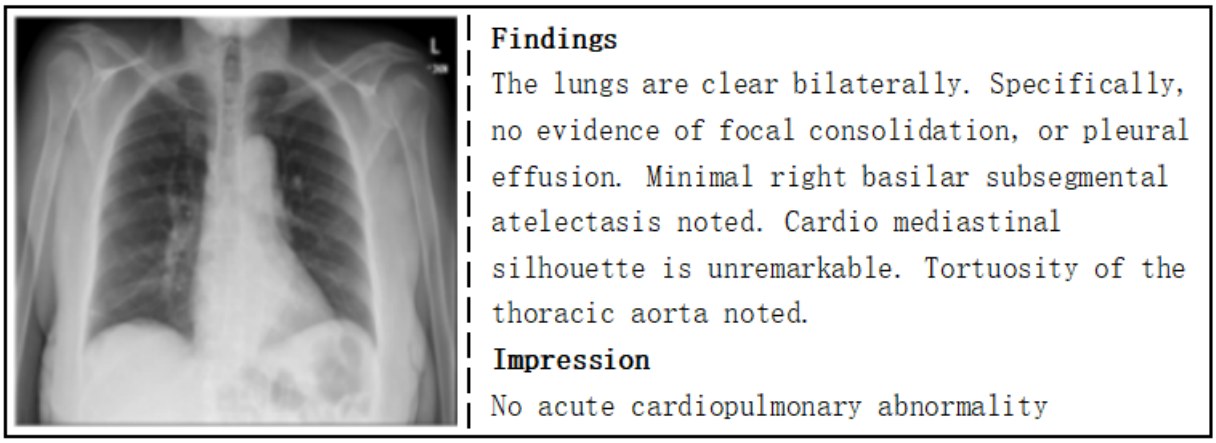

图 1. IU X-Ray 数据集中的一个样本

2.2 数据集

目前使用最广泛的数据集是 IU X-Ray [4] 和 MIMIC-CXR [8]。IU X-Ray 包含 3955 位患者的报告，数据集大小为 1GB 左右。MIMIC-CXR 是迄今为止最大的放射学数据集，包括 473057 张胸部 X 光图像和来自 63478 名患者的 206563 份报告，数据集大小为 557.6GB 左右。

3 本文方法

3.1 ResNet

在 ResNet [5] 出现之前，训练深层网络会出现训练损失先下降后上升的现象，这种现象被称为退化现象。退化与过拟合有着本质的不同。而 ResNet 主要是由残差块组成，解决了上述的退化问题。ResNet 提出者做了对比实验。实验结果表明：对于普通的网络结构，34 层网络的训练和验证误差要高于 18 层的网络。而当采用 ResNet 网络结构时，34 层的网络训练和验证误差要小于 18 层的网络。因此，残差结构确实能够有助于训练更深的网络，避免退化现象的发生。

ResNet 改进了网络结构，使误差表面更易于优化，将一定数量的层称为一个 block，即 ResNet 包含多个 block。对于一个 block，假设其拟合的函数为 $F(x)$，我们希望的潜在映射为 $H(x)$，通过残差连接前向传播就变成了：$H(x) = F(x) + x$。用 $F(x) + x$ 来拟合 $H(x)$。作者认为相比于让 $F(x)$ 学习恒等映射，让 $F(x)$ 学习成 0 要更加容易一些，即网络结构更易于优化。因此，在网络加深时，如果存在没有意义的 block，只要让这些 block 学习到 $F(x) \to 0$ 就能实现恒等映射，不会降低网络性能，同时也可以适当的缓解梯度消失。

形如图2构成的 block 称为残差块，通过构建前后层之间的“短路连接”，加强信息反向流动的能力，便于训练时梯度的传播，因此能够训练出很深的神经网络。

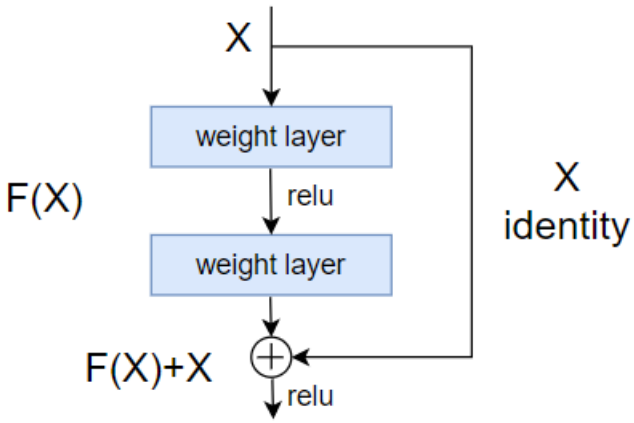

图 2. 残差块示例

ResNet 是由多个残差块串联起来的，常见的有 ResNet18、ResNet34、ResNet101 和 ResNet152。可见引入了 block 后，即使网络结构变得复杂，也不用担心网络性能受到影响。只要训练的样本足够多，就不用担心欠拟合、过拟合和退化问题，从而使网络拥有更强的表达能力和更优的性能.

3.2 Transformer

3.2.1 背景

循环神经网络常被用来处理时序信息，如像机器翻译这样的领域，其通常会有时间的维度，需要利用当前时刻 t 的输出去更新隐藏状态 $h_{t-1} \rightarrow h_t$，再由更新后的隐藏状态去预测下一时刻的输出。这使得网络只能一步一步计算。直接导致了模型训练难以并行化，并且当输入序列较长时，会导致早期输入的重要信息很难被保留下来。从另一角度来看，即使重要的信息被保留了下来，随着网络继续向后传递，也会造成内存的冗余，导致开销过大。

Transformer 是 Google 的研究团队在 2017 年提出的模型 [15]。在这之前，NLP 使用的最主流的模型通常是循环神经网络和长短期记忆网络。而 Transformer 与上述这些模型有着本质区别，其完全摒弃了之前的循环操作，相较于之前的模型，其拥有更强的并行能力，并且能够利用注意力机制，更多地关注对当前生成过程更有利的信息，避免过早的输入信息被丢弃，从而提高模型训练的效率，具有更好的长期依赖性建模能力。Transformer 的强大建模能力和高效的并行计算机制，使其成为了当前 CV 和 NLP 领域的研究热点之一。

3.2.2 架构

绝大多数的序列处理模型都采用编解码器架构，Transformer 也不例外。先将输入转换为中间表示，解码器利用中间表示和当前解码器的输入，预测下一时刻的输出。图3展示了 Transformer 的传统架构，公式表示如下：

$$\{\mathrm{h}_1, \mathrm{h}_2, ..., \mathrm{h}_\mathrm{s}\} = f_e\left(\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_\mathrm{s}\right) \tag{1}$$

$$y_t = f_d\left(\mathrm{h}_1, \mathrm{h}_2, ..., \mathrm{h}_\mathrm{s}, y_1, ..., y_{t-1}\right) \tag{2}$$

公式1中 $f_e(\cdot)$ 为编码器，$\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_\mathrm{s}$ 为编码器的输入，$\mathrm{h}_1, \mathrm{h}_2, ..., \mathrm{h}_\mathrm{s}$ 为中间表示，即通常所说的隐藏状态。公式2中 $f_d(\cdot)$ 为编码器，$\mathrm{h}_1, \mathrm{h}_2, ..., \mathrm{h}_\mathrm{s}$ 为中间表示，$y_1, ..., y_{t-1}$ 为解码器在 t 时刻之前的所有预测输出。

从图3可以看出 Transformer 主要包括：输入编码、位置编码、编码器、解码器和输出层这几个部分。

3.2.3 位置编码

通常要将词嵌入加上位置信息后再输入到编码器中，因为 Transformer 与传统序列模型不同，其无法根据单词的输入顺序学习到与位置有关的信息。但是这种顺序信息在 NLP 中十分重要。为了让模型能够学习到位置信息，就需要显式将位置信息输入给模型，这个位置信息通常是一个小的向量，它被加到词向量中，以形成一个新的表示。Transformer 中位置编码的计算公式如下：

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d_{\text{model}}}\right) \tag{3}$$

$$PE_{(pos,2i+1)} = \cos\left(pos/10000^{2i/d_{\text{model}}}\right) \tag{4}$$

其中 pos 是该单词在句子中的位置，$2i$ 是偶数维度，$2i+1$ 是奇数维度。使用上述公式计算位置编码，可以使位置编码能够适应不同的句子长度，并且能够很方便计算相对位置。

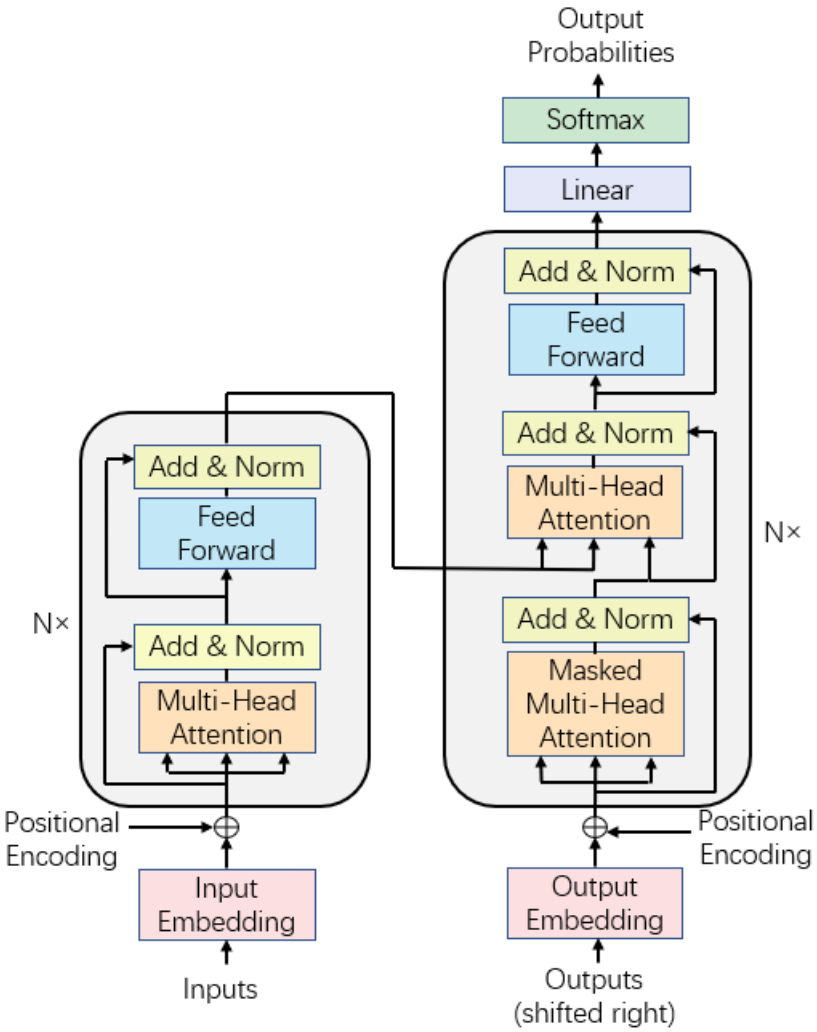

图 3. Transformer 传统架构

3.2.4 注意力机制

注意力机制就是首先将输入通过线性变换成一组查询 q 和键值对 k、v。然后计算当前 q 和所有 k 的相似度得到每一个 v 对当前 q 的重要程度，从而得到一组新的输入 [15]。通常使用缩放点积来计算注意力，简单地说，就是用两个向量的点积来表示相似度。为加快计算的效率，将 n 个 q 记为 Q，m 个 k 记为 K，m 个 v 记为 V，缩放点积注意力计算如图4所示。

缩放点积注意力计算的公式如下：

$$\text{Attention}\left(\mathrm{Q},\mathrm{K},\mathrm{V}\right)=\text{softmax}\left(\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}}\right) \tag{5}$$

其中 d 为 K 的列数，从公式中可看出，若 d 越大，Q 和 K 的内积就越大。故除以 d 的平方根，提升网络训练的稳定性。

我们通常通过对输入做线性变换得到 Q，K，V。以自注意力为例，得到 Q，K，V 的过程如图5所示。

在 CNN 中，使用分组卷积可以提取到更多有用的特征，Transformer 的多头注意力与此有着类似的思想：相比于只用一个注意力，不如使用多头注意力，让每个头独立地去学习不

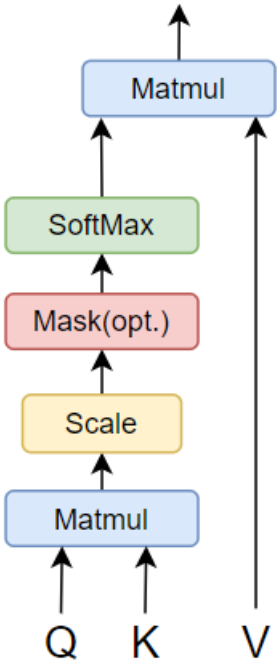

图 4. 缩放点积注意力

同的线性变换，最后将每个头的输出进行拼接，如图6所示。多头注意力的公式表示如下：

$$\text{MultiHead}\left(\mathrm{Q},\mathrm{K},\mathrm{V}\right)=\text{Concat}\left(\text{head}_1,...,\text{head}_\mathrm{h}\right)\mathrm{W}^\mathrm{O} \tag{6}$$

$$\text{where head}_\mathrm{i}=\text{MultiHead}\left(\mathrm{QW}_\mathrm{i}^\mathrm{Q},\mathrm{KW}_\mathrm{i}^\mathrm{K},\mathrm{VW}_\mathrm{i}^\mathrm{V}\right) \tag{7}$$

其中 $\mathrm{W}_\mathrm{i}^\mathrm{Q}\in\mathrm{R}^{\mathrm{d}_{\text{model}}\times\mathrm{d}_\mathrm{k}},\mathrm{W}_\mathrm{i}^\mathrm{K}\in\mathrm{R}^{\mathrm{d}_{\text{model}}\times\mathrm{d}_\mathrm{k}},\mathrm{W}_\mathrm{i}^\mathrm{V}\in\mathrm{R}^{\mathrm{d}_{\text{model}}\times\mathrm{d}_\mathrm{v}},\mathrm{W}^\mathrm{O}\in\mathrm{R}^{\mathrm{hd}_\mathrm{v}\times\mathrm{d}_{\text{model}}}$

Transformer 中有多头自注意力、多头交叉注意力和多头掩码注意力三种 [15]。其中需要特别说明的是掩码注意力，由于在预测当前时刻的输出时，只能看到当前时刻之前的所有信息。故通常将 softmax 的输入中当前时刻之后的所有值设为负无穷来屏蔽当前时刻之后的信息。

3.2.5 位置前馈网络

位置前馈网络类似于一个多层感知机，其中包含了两次线性变换和一个 ReLU 激活函数。不同层中的位置前馈网络参数是不共享的。位置前馈网络的公式表示如下：

$$\text{FFN}\left(x\right)=\max\left(0,xW_1+b_1\right)W_2+b_2 \tag{8}$$

3.2.6 层归一化

层归一化 (LN) 和批归一化 (BN) 是非常相似的归一化方法。如果我们把一批句子组成一个 batch，BN 就是将每句话对应位置的词做归一化，而 LN 就是对某一句子里的所有词做归一化。在 NLP 领域，文本的复杂性是很高的，任何一个词都可能会出现在一句话的任何位置，且单词顺序的变化有时候对理解句子并不重要。由此可见，BN 的归一化方式并不符合 NLP 的规律，使用 LN 更为合适。

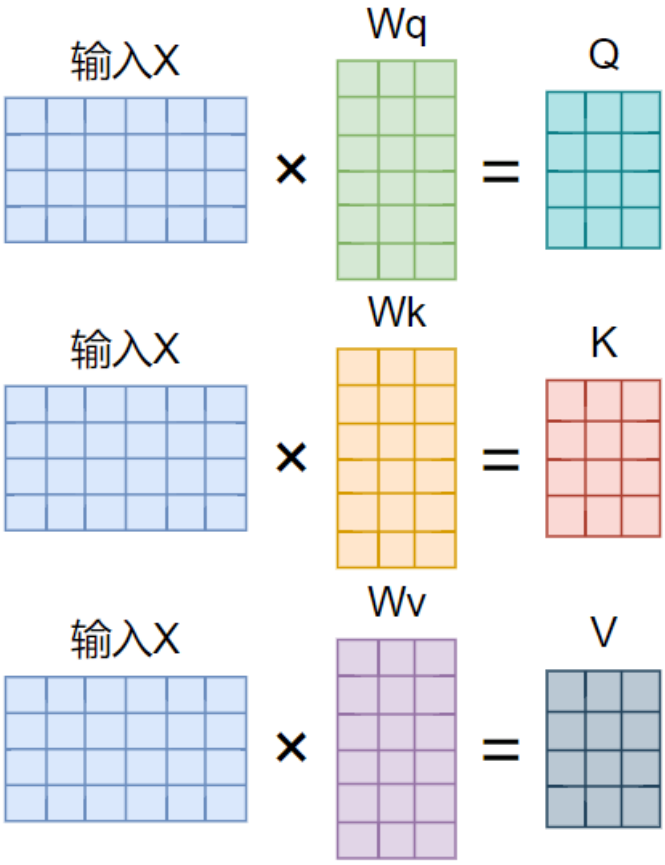

图 5. 通过线性变换得到 Q，K，V

3.3 Memory-driven Transformer

3.3.1 模型结构

我们曾提到在不同的诊断报告中具有相似的模式。Memory-driven Transformer 作为 Transformer 的改进，提出了一种关系内存 (RM) 来记录上一次生成过程的信息，并设计了一种新的内存驱动条件层归一化 (MCLN) 来将 RM 合并到 Transformer 中。因此，在生成过程中，可以隐式地对不同报告中的相似模式进行建模和记忆，从而为 Transformer 的解码过程提供更丰富的信息，使其能够生成更加真实的报告，整个报告生成模型可分为：视觉提取器、编码器和解码器三个部分，结构如图7所示。其中视觉提取器、编码器和解码器显示在灰色的破折号框中，视觉提取器和编码器的细节被省略。RM 和 MCLN 用灰色实框和蓝色虚线表示。报告生成过程的公式表示如下：

$$\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_\mathrm{S} = f_v(\mathrm{Img}) \tag{9}$$

$$\mathrm{h}_1, \mathrm{h}_2, ..., \mathrm{h}_\mathrm{S} = f_e(\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_\mathrm{S}) \tag{10}$$

$$y_t = f_d\left(\mathrm{h}_1, ..., \mathrm{h}_\mathrm{S}, \mathrm{MCLN}\left(\mathrm{RM}\left(y_1, ..., y_{t-1}\right)\right)\right) \tag{11}$$

公式9中 Img 为 X 光胸片，其视觉特征 X 由视觉提取器 ResNet101 提取，提取的结果用作所有后续模块的源序列。公式10中 X 被编码器映射为中间表示 H。在实现中采用 Transformer

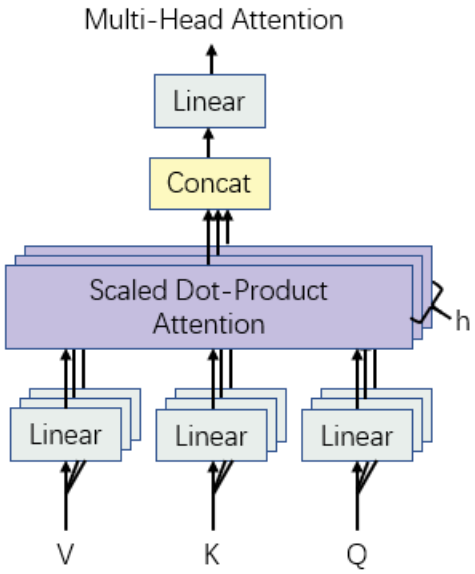

图 6. 多头注意力结构

中的标准编码器公式11中骨干解码器来自标准 Transformer，引入了 RM 和 MCLN，RM 接收 t 时刻之前的输出，并通过 MCLN 为解码过程提供信息。

3.3.2 关系内存 RM

对于任何相似的 X 光胸片，它们的报告可能会共享某些类似的模式，它们可以相互参考，帮助报告的生成。为了利用这一性质，作者提出了关系内存 RM 来增强 Transformer 学习模式的能力。

关系内存通过一个矩阵在生成步骤中转移其状态，我们称矩阵的每一行为内存槽，其记录了重要的模式信息。在报告生成过程中，矩阵会结合之前步骤的输出来更新自己。在时间步 t 处，上一个时间步的矩阵 M_{t-1} 被当作查询，它与上一个时间步的输出 y_{t-1} 连接起来作为键和值提供给多头注意力模块。即对于每个头，都会得到一组 Q，K，V。

$Q = M_{t-1} \cdot W_q, K = [M_{t-1}; y_{t-1}] \cdot W_k, V = [M_{t-1}; y_{t-1}] \cdot W_v$。其中 y_{t-1} 是最后一个输出的嵌入表示，$[M_{t-1}; y_{t-1}]$ 是 M_{t-1} 和 y_{t-1} 的行级连接。

因为 RM 与解码过程一起以循环的方式执行，可能会出现梯度消失和梯度爆炸的问题。所以引入了残差连接和门机制。残差连接的公式如下：

$$\tilde{M}_t = f_{mlp}(Z + M_{t-1}) + Z + M_{t-1} \tag{12}$$

其中 $f_{mlp}(\cdot)$ 为多层感知机。门机制的详细结构如图8所示，其中遗忘门和输入门分别用于平衡来自 M_{t-1} 和 y_{t-1} 的输入。为了保证 y_{t-1} 和 M_{t-1} 能够一起计算，通过将 y_{t-1} 复制多行扩展成矩阵 Y_{t-1}。门机制的公式表示如下：

$$G_t^f = Y_{t-1}W^f + \tanh(M_{t-1}) \cdot U^f \tag{13}$$

$$G_t^i = Y_{t-1}W^i + \tanh(M_{t-1}) \cdot U^i \tag{14}$$

$$M_t = \sigma\left(G_t^f\right) \odot M_{t-1} + \sigma\left(G_t^i\right) \odot \tanh\left(\tilde{M}_t\right) \tag{15}$$

其中 W^f 和 W^i 是每个门中 Y_{t-1} 的可训练权重，U^f 和 U^i 是每个门中 M_{t-1} 的可训练权重。$\odot$ 指的是 Hadamard 乘积，M_t 是时间步 t 处 RM 的输出。

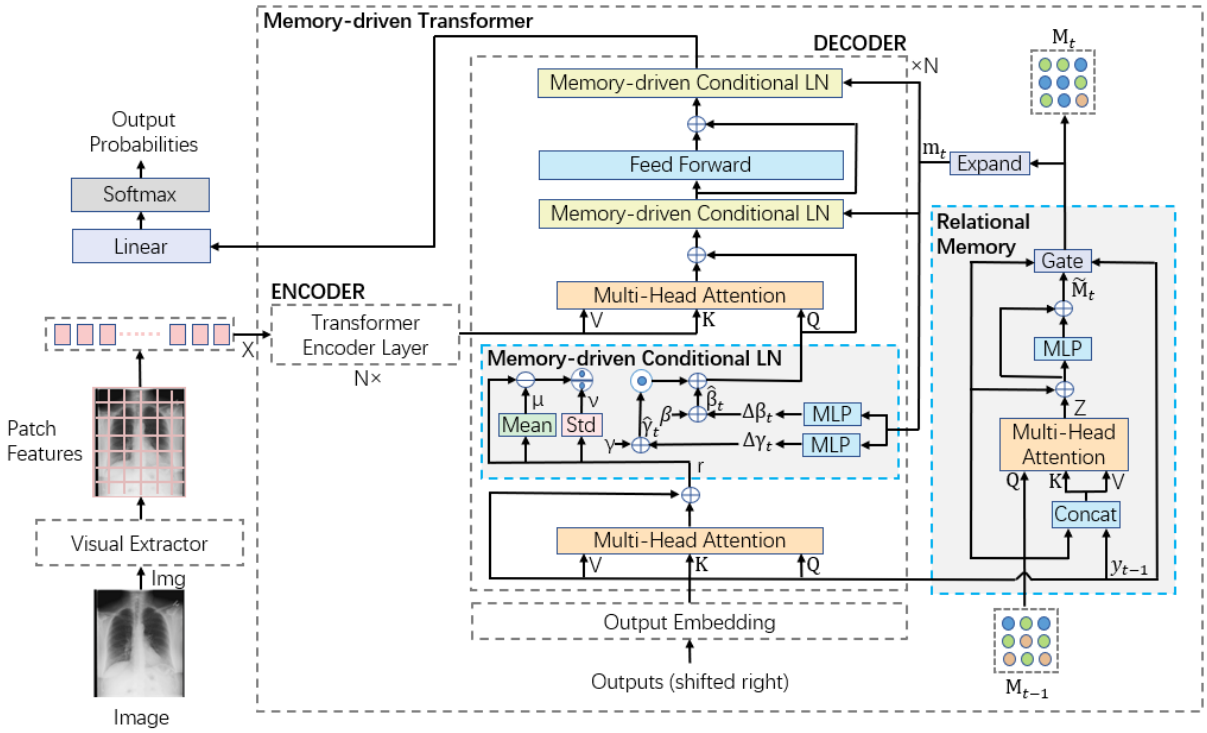

图 7. Memory-driven Transformer 模型整体结构

3.3.3 内存驱动条件层归一化 MCLN

在传统 Transformer 的层归一化中，为了增加泛化能力，引入了 γ 和 β 两个参数。MCLN 将 RM 的输出 M_t 提供给这两个参数来合并 RM 记录的模式信息。这样既合并了 RM 中的信息，又避免了 RM 对 Transformer 太多参数产生影响，使一些用于生成的核心信息不受影响。如图7所示，在每个 Transformer 解码层使用了三个 MCLN。第一个 MCLN 的输出被用于后面多头注意力模块的查询。在时间步 t 处 RM 的输出 M_t 通过简单地将所有行连接起来，扩展成一个向量 m_t 提供给每一个 MCLN。在 MCLN 中首先通过一个 MLP 预测 γ 和 β 的变化，然后利用预测的变化来做更新。MCLN 的公式表示如下：

$$\Delta\gamma_t = f_{\text{mlp}}(m_t) \tag{16}$$

$$\hat{\gamma}_t = \gamma + \Delta\gamma_t \tag{17}$$

$$\Delta\beta_t = f_{\text{mlp}}(m_t) \tag{18}$$

$$\hat{\beta}_t = \beta_+ \Delta\beta_t \tag{19}$$

$$f_{\text{mcln}}(r) = \hat{\gamma}_t \odot \frac{r-\mu}{\nu} + \hat{\beta}_t \tag{20}$$

公式20中 r 指的是前面模块的输出，μ 和 ν 分别是 r 的均值和标准差。

4 复现细节

4.1 与已有开源代码对比

Memory-driven Transformer 论文的代码已经开源，我们直接使用开源的代码进行实验。在改进的部分，我们在模型中添加了 CBAM 模块，该模块的实现也是参考了已开源的代码。

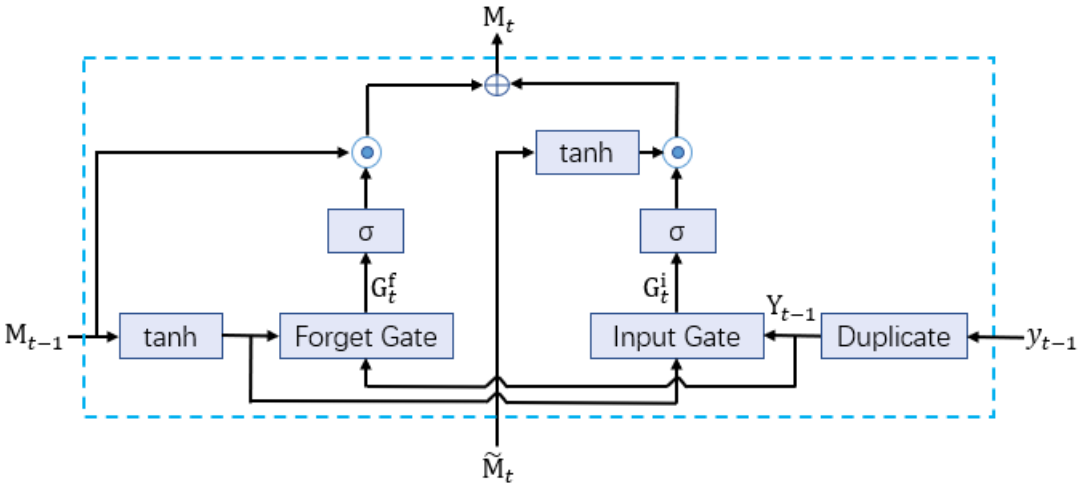

图 8. RM 中的门机制

本文的主要工作主要是发现 Memory-driven Transformer 中未使用 ViT，而是将一块块小区域特征直接伸缩变换成一维向量特征。而且即使是在 ViT 中，也需要给各 patch 一个位序信息，这里没有使用 ViT，所以损失了大量的空间信息，于是我们添加了 CBAM 模块来加强空间信息的学习，提高后续报告生成的准确性。

4.2 实验环境搭建

torch==1.7.1
torchvision==0.8.2
opencv-python==4.4.0.42

4.3 创新点

Memory-driven Transformer 会在特征图被分割成各个 patch 之后做注意力计算，因为这里使用了传统的 Transformer，不像 ViT 一样提供位序信息，所以此时已经大量损失了特征图的空间信息。对此，我们通过将卷积块注意力模块 CBAM 添加在视觉提取器的尾部来促使模型学习到更多与空间相关的信息，提高后续报告生成的准确性。改进后报告生成网络结构如图9所示。

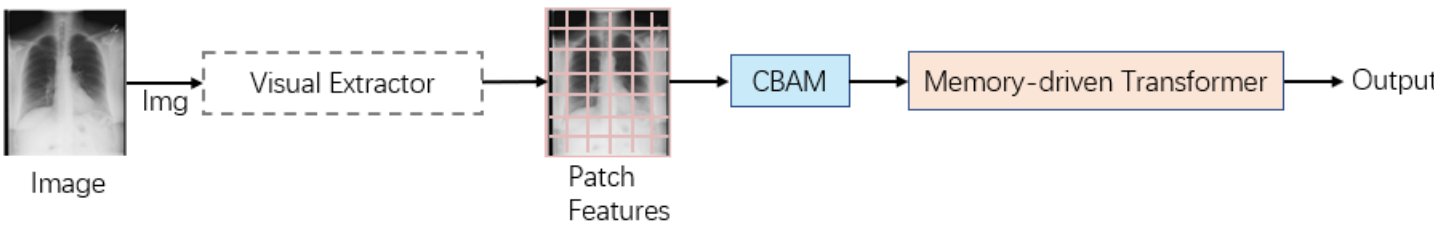

图 9. 报告生成网络结构

卷积块注意力模块 [18] 由通道注意力和空间注意力组成，如图10所示。

通道注意力首先会对输入的每一个通道分别应用一次全局最大池化和平均池化，然后利用共享的全连接层对最大池化和平均池化的结果处理后相加，最后经过 sigmoid 映射得到每

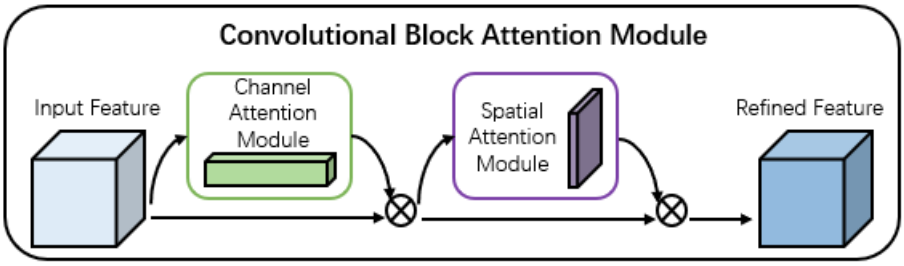

图 10. 卷积块注意力模块

一个通道对应的权值 [6]。将这些权值乘上各自对应的原输入通道就完成了通道注意力的计算，如图11所示。

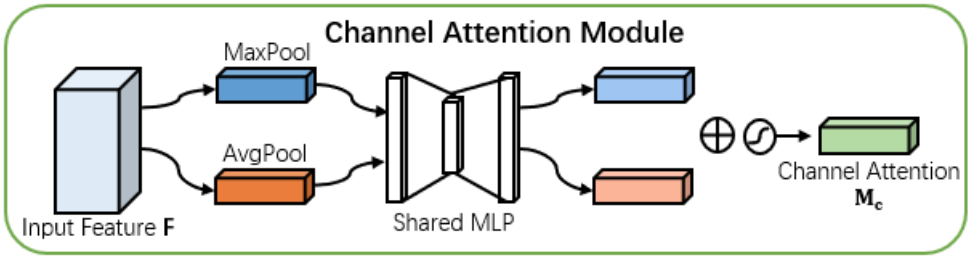

图 11. 通道注意力机制

空间注意力在每一个特征点的通道上取最大值和平均值，将两者堆叠之后利用卷积把通道数降为 1，最后经过 sigmoid 得到每一个特征点对应的权值 [18]。将权值乘上每个原通道就完成了空间注意力的计算，即完成了卷积块注意力模块的计算，如图12所示。

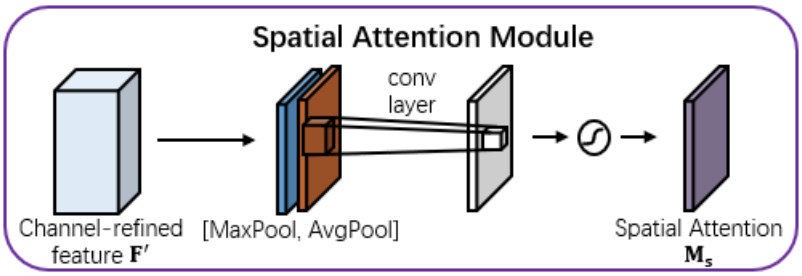

图 12. 空间注意力机制

5 实验结果分析

我们对比了 Memory-driven Transformer 和 Memory-driven Transformer+CBAM 在 NLG 指标上的评估结果，后者将 CBAM 添加到视觉提取器的尾部。在 NLG 指标上的结果如表1所示，其中 MDT 指的是 Memory-driven Transformer，* 表示我们用原论文的代码复现的结果。实验结果表明：在视觉提取器尾部添加 CBAM 使得其在 NLG 各项指标中都得到了不错的提升。模型评估共用了六个指标：

BL-n [13]：通过比较标签句子和生成句子中 n 个连续词的出现情况来评估性能。

MTR [2]：通过单精度的加权调和平均数和单字召回率来评估性能。

RG-L [10]：通过标签句子和生成句子的最长共有子句来评估性能。

该实验对应的超参数设置如下：

单词切除频率：3

词嵌入维度：512

注意力头数：8

编解码器层数：3

批量大小：8

gpu 数：1

训练迭代次数：100

学习率：5e-5

权重衰减：5e-5

随机种子：9233

表 1. 改进前后性能对比

Methods	IU X-Ray					
	BL-1	BL-2	BL-3	BL-4	MTR	RG-L
	Batch Size = 8					
MDT	0.470	0.304	0.219	0.165	0.187	0.371
MDT*	0.463	0.296	0.214	0.163	0.189	0.362
MDT+CBAM	**0.492**	**0.318**	**0.228**	**0.169**	**0.205**	**0.376**

6 总结与展望

医生撰写 X 光胸片报告是一项既费时费力又容易出错的任务。随着深度学习技术的发展和人们生活质量的提高，人们对 X 光胸片诊断报告自动生成技术的要求会越来越高。目前 X 光胸片诊断报告自动生成技术还存在很大的提升空间，主要体现在以下方面：

一是缺少一个衡量医学正确性的评价指标。目前所使用的评价指标是 NLG 方面的评价指标，这些指标更多的是衡量生成报告的流利性，而不是从医学正确性的角度来评估报告，生成报告的流利性可能并没有那么重要，所以在这些指标上得分高的报告也不一定能够满足人们的要求。

二是报告生成模型的可解释性。相比将深度学习应用在其它领域时的可解释性，应用在医学领域时模型的可解释性更为重要。因为这直接关系到患者的健康状况，因此可能需要加入一些工作来揭示模型学习到的特征，比如对于某一种疾病，模型最关注的是 X 光胸片的哪些部分，是否符合专业医生的预期。只有得到了专业医生的认可，这项技术才更有价值。

参考文献

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and

visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[2] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[3] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020.

[4] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[7] Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*, 2017.

[8] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

[9] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Hybrid retrieval-generation reinforced agent for medical image report generation. *Advances in neural information processing systems*, 31, 2018.

[10] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[11] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, pages 249–269. PMLR, 2019.

[12] Pablo Messina, Pablo Pino, Denis Parra, Alvaro Soto, Cecilia Besa, Sergio Uribe, Marcelo Andia, Cristian Tejos, Claudia Prieto, and Daniel Capurro. A survey on deep learning and explainability for automatic report generation from medical images. *ACM Computing Surveys (CSUR)*, 54(10s):1–40, 2022.

[13] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[14] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024, 2017.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[16] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[17] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. 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*, pages 9049–9058, 2018.

[18] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[19] 张宇. 基于深度学习的医学影像报告自动生成. Master's thesis, 北京协和医学院, 2020.