arXiv:2303.00534v1 [cs.CV] 1 Mar 2023

RAMM：基于检索增强的生物医学视觉问答多模态预训练

Zheng Yuan[1], Qiao Jin[12], Chuanqi Tan[1], Zhengyun Zhao[2],Hongyi Yuan[12],Fei Huang[1],Songfang Huang[1]阿里巴巴集团， [2]清华大学
{yuanzheng.yuanzhen,qiao.jqa,chuanqi.tcq,f.huang,songfang.hsf}@alibaba-inc.com{
zhengyun21,yuanhy20}@mails.tsinghua.edu.cn

摘要

视觉与语言多模态预训练和微调已在视觉问答（*VQA*）领域取得巨大成功。与通用领域的 *VQA* 相比，生物医学视觉问答的性能受限于数据匮乏。本文提出一种名为 ***RAMM*** 的检索增强预训练与微调范式，以解决生物医学视觉问答中的数据限制问题。具体而言，我们收集了一个名为 ***PMCPM*** 的新生物医学数据集，其中包含来自 *PubMed*、涵盖多种患者情境的基于患者的图文对。随后，我们对生物医学多模态模型进行预训练，使其学习图文对的视觉和文本表征，并通过图文对比目标（*ITC*）对齐这些表征。最后，我们提出一种检索增强方法，以更充分地利用有限的数据。我们基于 *ITC* 从预训练数据集中检索相似的图文对，并引入一种新颖的检索注意力模块，将图像和问题的表征与检索到的图像和文本融合。实验表明，我们提出的检索增强预训练与微调范式在 *Med-VQA2019*、*Med-VQA2021*、*VQARAD* 和 *SLAKE* 数据集上取得了最先进的性能。进一步分析表明，与以往的资源和方法相比，所提出的 *RAMM* 和 *PMCPM* 能够提升生物医学视觉问答的性能。我们将开源数据集、代码和预训练模型。

1. 引言

生物医学视觉问答（VQA）旨在根据给定图像回答与生物医学相关的问题。生物医学视觉问答中的图像通常是放射学图像（如计算机断层扫描、磁共振成像和X射线），问题通常涉及相应图像中的模态、身体部位、成像平面和异常部位。回答这些问题需要丰富的临床和医学影像知识——

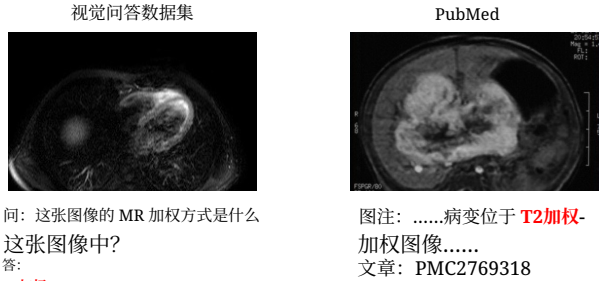


图1. 一个相似图像的示例，其文本包含回答生物医学问题的有用信息。左侧是一个与生物医学相关的问题及其提供的图像。右侧是 PubMed 中的相似图像，其相关说明文字给出了该问题的答案（红色文本）。

影像知识 [17],这对自动化方法而言颇具挑战。

随着通用领域视觉问答的成功 [10,26,27],，一些研究遵循预训练与微调范式，尝试针对生物医学视觉问答进行领域特定的视觉语言建模 [4, 20]。然而，与拥有充足标注视觉问答对用于微调的通用领域不同，生物医学领域的微调数据仍然有限，这使模型容易过拟合，并阻碍模型学习全面的领域知识以回答复杂问题。

与生物医学视觉问答数据集中的有限数据相比，生物医学文献包含大量高质量图像-文本对。这些文本通常会从各个方面描述图像，包括成像模态、身体部位、成像平面和异常情况，而这些正是生物医学视觉问答关注的信息。图1 展示了来自 PubMed 的图像-文本对[1] 如何有助于回答视觉问答数据集中的问题。基于这一观察，我们认为检索是生物医学视觉问答的一种可行方案。这一过程模拟了医生解决复杂患者病例的方式：

[1]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2769318/

医生会搜索与给定病例相似的病例，并根据以往经验提出解决方案。因此，我们可以从文献中检索相似的图像-文本对，以帮助回答问题。要从生物医学文献中检索相关图像-文本对，需要回答三个关键问题：**从哪里**检索？**如何**检索？以及**如何**利用检索到的图像-文本对？

为此，我们提出一种检索增强的生物医学多模态预训练与微调范式（**RAMM**），用于生物医学视觉问答。首先，已有可用于检索的生物医学图文对，包括ROCO [36]和 MIMIC-CXR[18] 。然而，ROCO 受限于数据量，而 MIMIC-CXR 仅关注胸部X光片。要实现更好的检索，需要一个规模庞大、多样且高质量的数据集。因此，我们提出 PMC-Patients-Multi-modal（**PMCPM**），这是一个从 PubMed Central 筛选而来的患者级图文数据集，其规模大于 ROCO 和 MIMIC-CXR，涵盖多种模态和图像状况。其次，由于生物医学视觉问答中的问题具有同质性，我们提出使用图像进行检索。我们结合 PMCPM 数据集以及此前广泛使用的 ROCO 和 MIMIC-CXR 数据集，使用三项任务对生物医学多模态模型进行预训练：掩码语言建模、图文对比学习和图文匹配。由于预训练阶段的图文对比目标旨在将视觉和文本表征对齐到同一嵌入空间中， [27, 37], 因此，图文对比相似度可用于检索相似的图文对。最后，我们将上述数据集作为资源，提出检索与给定图像相关的图文对。为充分利用检索到的图文对，我们提出一种检索增强微调方法，在多模态编码器层中引入新颖的检索注意力模块，并使用融合表示预测最终答案。

我们在四个生物医学视觉问答数据集上进行了实验。我们的 RAMM 在 VQA-Med 2019[2], VQA-Med 2021 [16], VQARAD [22], 和 SLAKE（英文版） [30], 上分别取得了 82.13、39.20、78.27 和 86.05 分的最新最佳成绩。我们还进行了消融研究，以验证我们收集的 PM- CPM 数据集以及检索增强微调方法的有效性。案例研究表明，检索到的图文对有助于回答问题。

我们的贡献总结如下：

•我们从 PubMed 收集了 PMCPM，这是一个基于患者的图文数据集，可用于生物医学多模态预训练和检索。

•我们提出了 RAMM，这是一种用于生物医学视觉问答的检索增强多模态预训练与微调范式，并包含一个新颖的检索注意力模块。

•实验表明，所提出的 RAMM 配合额外的 PMCPM 数据集，在 VQA-Med 2019、VQA-Med 2021、VQARAD 和 SLAKE 数据集上的表现优于以往的视觉问答方法。

2. 相关工作

生物医学多模态预训练 在通用领域，多模态预训练模型在各种图文相关下游任务中均展现出有效性 [10,26,27,41]。在生物医学自然语言处理领域，已有研究表明，在领域内语料库上继续训练可以提升预训练语言模型的性能 [1, 14, 23, 44, 45]。结合上述两点，直接考虑为生物医学视觉问答预训练一个生物医学领域内多模态模型是合理的。以往的生物医学多模态预训练模型 [4, 5, 11, 20, 34]主要利用 ROCO [36] 和 MIMIC-CXR [18] 作为预训练数据集。这些模型预测掩码词元 [4,5]或图像块 [4] ，并将图文对齐 [11]作为预训练任务。与以往模型相比，我们的 RAMM 使用我们额外提出的 PMCPM 数据集进行预训练，该数据集包含来自不同患者病情的图像。在预训练任务方面，我们的模型采用掩码语言建模、图文对比学习和图文匹配。RAMM 使用图文对比相似度进行检索。

生物医学视觉问答 视觉问答要求回答图像相关问题。与通用领域的视觉问答不同， [10, 26, 27], 生物医学视觉问答的训练样本要少得多。为克服生物医学视觉问答中的数据限制，研究者采用元学习，以便快速适应视觉问答问题 [9, 35]。随着预训练与微调范式取得成功，使用大规模无监督数据集预训练多模态模型，可以学习到更好的视觉和文本表征，从而减轻数据限制的影响，并显著提升生物医学视觉问答的性能 [3,4,11]。本文也使用基于患者的图文对 PMCPM，对生物医学多模态模型 RAMM 进行预训练。在微调视觉问答数据集时，RAMM 会检索相似的图文对，并尝试利用这些图像和文本中的有用信息。

检索增强已成功应用于自然语言处理 [15,24,25]和计算机视觉 [32,39]领域的知识密集型任务。本文中，我们决定采用检索增强来解决知识密集型且标注不足的生物医学视觉问答任务。检索器可以是 BM-25 [24, 38]这样的稀疏检索器，也可以是利用预训练编码器 [15, 39]的密集检索器。对于我们的任务，通过图文对比（ITC）任务预训练的模型可直接用作多模态检索器。因此，我们通过 ITC 预训练了一个生物医学多模态模型，用于检索增强。我们还提出了一种检索-

注意力模块，用于融合这些检索样本。据我们所知，这是首项在生物医学视觉问答中利用检索增强的工作。

3. 方法

本节首先介绍 PM-CPM 数据集的构建方法，然后说明模型架构和预训练任务。最后，介绍 RAMM 如何从预训练数据集中检索图文对，以及在生物医学视觉问答微调过程中如何结合检索样本。整体工作流程如图 2所示。

3.1. PMCPM 数据集

为获取高质量、大规模的领域专属图文对，我们首先考虑使用 PubMed Central（PMC）的开放获取子集（PMC OA）[2], 。这是一个免费的生物医学文献档案，收录了超过4百万篇出版物，这些文章的全文（包括图注）和图片均可通过官方 FTP 服务公开获取[3]。然而，PMC OA 全集中的图片可能相当嘈杂：除临床图像外，研究文章中还会出现许多类型的图表，包括流程图、直方图和散点图，这些内容对多模态预训练提供的信息很少。Pelka等人 [36] 使用基于人工标注数据集训练的神经网络，自动筛除复合图、多面板图和非放射影像，收集了一个包含 8.1 万个图文对的干净数据集。为进一步探索并利用 PMC 中的图片资源，我们转而采用受 Zhao 启发的“基于患者”的收集流程等人 [47]。“病例报告”是一种特定类型的医学文献，通常描述患者的整个住院过程。遵循 Zhao等人 [47], 的方法，我们使用正则表达式识别包含患者病历的章节，提取相应文本，并应用多种筛选方法去除噪声，最终收集到 16.7 万份涵盖多种病况的高质量患者病历。这些患者病历附带的图片与患者的临床状况密切相关，且比单纯的放射学图像包含更丰富的信息。因此，我们收集这些图片及其对应图注，构建 PMCPM 数据集。图3展示了 PMCPM（PMCID：509249）中的两个示例，包括图像及其对应图注。图3A 展示了一张显微照片，描述了筛状结构及神经周围侵犯的腺样囊性癌。图3B 展示了一张放射影像，其中双侧肺叶均散布着多个转移性病灶。这些示例表明，我们基于患者的收集方法能够针对临床相关图片，获取更多图像模态和更详尽的文本描述。

[2]https : / / www . ncbi . nlm . nih . gov / pmc / tools / openftlist/[3]https://www.ncbi.nlm.nih.gov/pmc/tools/ftp/

3.2. 模型架构

对于一个图文对（I， T），RAMM 首先分别提取单模态特征，然后将其融合为多模态表征。融合后的表征用于视觉问答分类。RAMM 由文本编码器、图像编码器以及用于融合两种不同模态表征的多模态编码器组成。文本T 被分词为子词$\{x_{cls}, x_1, ..., x_n\}$，并由文本编码器转换为隐藏表征$\mathbf{w}_0 = \{\mathbf{w}_{cls}, \mathbf{w}_1, ..., \mathbf{w}_n\}$。图像$I$被划分为图像块 $\{y_{cls}, y_1, ..., y_m\}$，并由图像编码器表示为$\mathbf{v}_0 = \{\mathbf{v}_{cls}, \mathbf{v}_1, ..., \mathbf{v}_m\}$。多模态编码器是一种协同注意力式双流 Transformer。

多模态融合层 我们采用协同注意力式多模态融合层 [10] 处理文本特征 $\mathbf{w}_0$ 和视觉特征 $\mathbf{v}_0$ ，层数为 L。具体而言，协同注意力融合层由双流 Transformer 层组成，其中每个流对应一种模态。每个 Transformer 层结合了一个自注意力模块和一个交叉注意力模块[4]。对于第 i层，我们将文本隐藏表示记为 $\mathbf{w}_i = \{\mathbf{w}^i_{cls}, \mathbf{w}^i_1, ..., \mathbf{w}^i_n\}$，并将图像隐藏表示记为 $\mathbf{v}_i = \{\mathbf{v}^i_{cls}, \mathbf{v}^i_1, ..., \mathbf{v}^i_m\}$。Transformer 中的注意力模块定义如下：

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}(\frac{QK^\top}{\sqrt{d_k}})V \tag{1}$$

使用此记号，自注意力定义如下：

$$Q^{i,w} = W^w_Q \mathbf{w}_i, K^{i,w} = W^w_K \mathbf{w}_i, V^{i,w} = W^w_V \mathbf{w}_i \tag{2}$$

$$Q^{i,v} = W^v_Q \mathbf{v}_i, K^{i,v} = W^v_K \mathbf{v}_i, V^{i,v} = W^v_V \mathbf{v}_i \tag{3}$$

$$\mathbf{w}'_i = \mathrm{Attn}(Q^{i,w}, K^{i,w}, V^{i,w}) \tag{4}$$

$$\mathbf{v}'_i = \mathrm{Attn}(Q^{i,v}, K^{i,v}, V^{i,v}) \tag{5}$$

我们沿用 $Q^{i,\{w,v\}}, K^{i,\{w,v\}}, V^{i,\{w,v\}}$ 的记号，将 $\mathbf{w}_i$ 替换为 $\mathbf{w}'_i$，并将 $\mathbf{v}_i$ 替换为 $\mathbf{v}'_i$，交叉注意力定义如下：

$$\mathbf{w}_{i+1} = \mathrm{Attn}(Q^{i,w}, K^{i,v}, V^{i,v}) \tag{6}$$

$$\mathbf{v}_{i+1} = \mathrm{Attn}(Q^{i,v}, K^{i,w}, V^{i,w}) \tag{7}$$

文本和图像的融合表示分别记为$\mathbf{w}_L$ 和$\mathbf{v}_L$。

[4]为简洁起见，省略前馈网络。

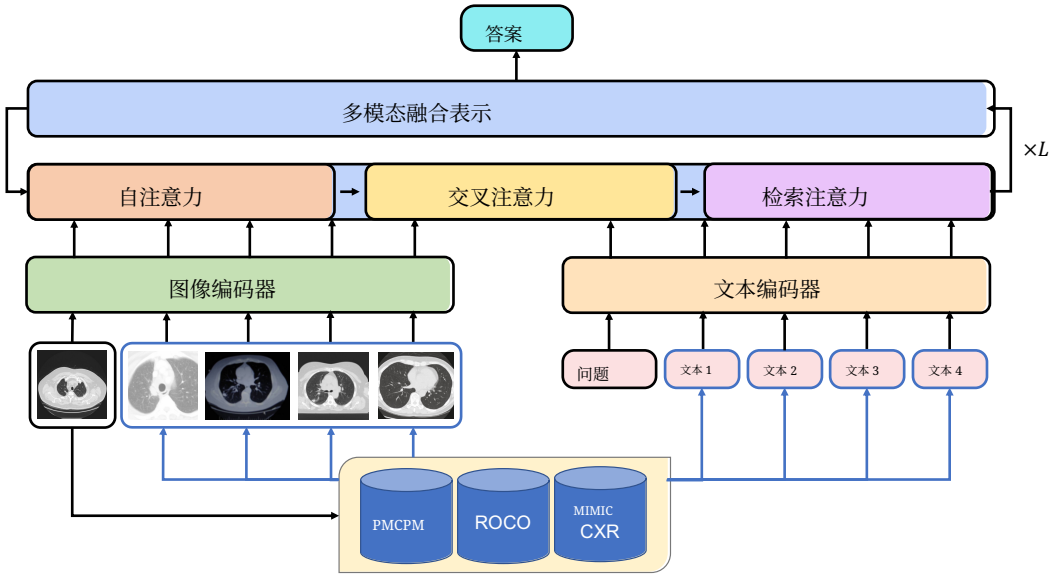
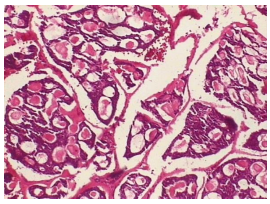


图 2。我们提出的用于生物医学视觉问答的检索增强范式 RAMM 的工作流程。原始图像来自视觉问答数据集，其他图像则通过检索获得。蓝线表示基于图文对比相似度，从 PMCPM、ROCO 和 MIMIC-CXR 数据集中检索图文对（PMCID:4061445、2278144、4725742 和 3616602）。检索增强样本被送入单模态和多模态编码器。多模态编码器包含检索注意力，用于融合检索到的信息。多模态编码器获得的表示用于视觉问答分类。



（A）说明：
显示小型细胞的显微照片
深染细胞，且细胞质稀少
细胞质呈巢状排列
被圆形或椭圆形的窗孔状结构分隔
间隙——筛状结构，H & E（× 100）



(B) 图注：
腹部超声显示
多个低回声区（短箭头）
和高回声病灶（长
箭头）位于肝脏中

图 3。来自 PMCPM 的图文对示例。

3.3. 预训练任务

为了对 RAMM 进行预训练，我们采用了三个预训练任务，包括图文对比学习（ITC） [27],图文匹配（ITM） [27], 和掩码语言建模（MLM） [8]。RAMM 的预训练目标函数定义为这三个任务之和。ITC 将单模态的视觉和文本表征$\mathbf{w} = g_w(\mathbf{w}_{cls})$ 和$\mathbf{v} = g_v(\mathbf{v}_{cls})$ 映射到同一嵌入空间，其中 g_w 和 g_v 是经过 L^2 归一化的两个线性投影。ITM 根据融合表示$\mathbf{w}_L$ 和$\mathbf{v}_L$判断一幅图像和一段文本是否匹配。MLM 根据融合T 进行文本中掩码词元的预测文本表示 $\mathbf{w}_L$。由于 ITC 旨在通过余弦相似度对齐图像和文本，因此检索相关图像和文本成为一种自然选择。

3.4. 检索增强微调

我们使用图文对比相似度检索相关图文对，并在微调过程中通过检索注意力模块将其融合。

检索相关图文对设 $\mathcal{D} = \{(I_j, T_j)|j \in \mathcal{A}\}$ 表示所有图文对的集合，其中 $\mathcal{A}$ 是索引集。对于视觉问答数据集中的图像-问题对 (I, T)，我们根据单模态图像表示$\mathbf{v}_{cls}$ 从 $\mathcal{D}$ 中检索 r 个增强样本。这里我们不使用问题 T 的表示进行检索，因为视觉问答数据集中的问题同质且缺乏信息。对于不同的图像，问题可能类似于这张图像中有什么异常？或这张图像中是否存在异常？对于具有单模态表示$\mathbf{v}_{cls}^j$ 和$\mathbf{w}_{cls}^j$的图文对 $(I_j, T_j) \in \mathcal{D}$，我们使用 ITC 相似度分数计算图像$I$ 之间的相似度：

$$s_w^j = g_v(\mathbf{v}_{cls})^\top g_w(\mathbf{w}_{cls}^j) \tag{8}$$

$$s_v^j = g_v(\mathbf{v}_{cls})^\top g_v(\mathbf{v}_{cls}^j) \tag{9}$$

ITC 在预训练期间优化 $g_v(\mathbf{v}_{cls}^j)^\top g_w(\mathbf{w}_{cls}^j) \rightarrow 1$，这使得 s_w^j 和 s_v^j 趋于接近。因此，我们使用 s_w^j 和 s_v^j 中的最大值来定义 (I_j, T_j) 之间的相似度

其中 I:

$$s_j = \max(s_w^j, s_v^j) \tag{10}$$

计算 s_j时，我们冻结单模态编码器。首先，ITC 分数在微调过程中可能会发生显著变化。其次，预训练数据集中的表示可以预先计算，并可通过 Faiss [19]快速进行相似度计算。

我们首先使用 Faiss，在 $\{s_w^j|j \in \mathcal{A}\}$ 中计算前 r $\{s_w^j|j \in \mathcal{R}_w\}$ 个，在 $\{s_v^j|j \in \mathcal{A}\}$ 中计算前 r $\{s_v^j|j \in \mathcal{R}_v\}$ 个。$\mathcal{R}_w \cup \mathcal{R}_v$ 包含 $[r, 2r]$ 个图文对。训练时，我们从 $\mathcal{R}_w \cup \mathcal{R}_v$中随机选取 r 个图文对。每个图文对被选取的概率与相似度 s^j成正比。推理时，我们根据 s^j从 $\mathcal{R}_w \cup \mathcal{R}_v$ 中选取前 r 个相似图文对。我们将检索到的 r 个图文对记作 $\{(I_j,\ T_j)|1 \leq j \leq r\}$。

多模态编码器中的检索融合 为了融合检索样本中的知识，我们在多模态融合层中提出了一个检索注意力模块。检索注意力模块应用于每个多模态融合层中的自注意力和交叉注意力之后。图 4 展示了应用于文本模态的检索注意力模型架构。具体而言，我们使用单模态编码器对检索到的图像-文本对 $\{(\mathrm{I}_j, \mathrm{T}_j)|1 \leq \mathrm{j} \leq \mathrm{r}\}$ 进行编码，得到 $\{(\mathrm{w}_0^j, \mathrm{v}_0^j)|1 \leq j \leq r\}$。为统一符号，我们将 $(\mathrm{w}_0^0, \mathrm{v}_0^0)$ 记作 $(\mathrm{w}_0, \mathrm{v}_0)$。对于 i^{th}多模态融合层，文本隐藏表示的输入为 $\{\mathrm{w}_i^j|0 \leq j \leq r\}$，图像隐藏表示的输入为 $\{\mathrm{v}_i^j|0 \leq j \leq r\}$。通过应用第 3.2 节定义的自注意力和交叉注意力，我们得到文本表示 $\{\mathrm{w}_i^{j\prime}|0 \leq j \leq r\}$ 和图像表示 $\{\mathrm{v}_i^{j\prime}|0 \leq j \leq r\}$。检索注意力独立应用于每种模态。以文本表示为例，检索注意力使用 $\mathrm{w}_i^{0\prime}$ 的 [CLS]表示作为注意力查询，并使用 $\{\mathrm{w}_i^{j\prime}|0 \leq j \leq r\}$ 的 [CLS] 表示作为注意力键和值。检索注意力的输出仅添加到原始表示 $\mathrm{w}_i^{0\prime}$ 中，而不添加到检索表示$\{\mathrm{w}_i^{j\prime}|1 \leq j \leq r\}$中。

$$Q^{i,r} = W_Q^r \mathbf{w}_{i,cls}^{0\ \prime} \tag{11}$$

$$K^{i,r} = W_K^r [\mathbf{w}_{i,cls}^{0\ \prime}, \mathbf{w}_{i,cls}^{1\ \prime}, ..., \mathbf{w}_{i,cls}^{r\ \prime}] \tag{12}$$

$$V^{i,r} = W_V^r [\mathbf{w}_{i,cls}^{0\ \prime}, \mathbf{w}_{i,cls}^{1\ \prime}, ..., \mathbf{w}_{i,cls}^{r\ \prime}] \tag{13}$$

$$\mathbf{w}_{i+1,cls}^j = \begin{cases} \mathbf{w}_{i,cls}^{j\ \prime} + \mathrm{Attn}(Q^{i,r}, K^{i,r}, V^{i,r}), & j = 0 \\ \mathbf{w}_{i,cls}^{j\ \prime}, & j > 0 \end{cases} \tag{14}$$

图像模态上的检索注意力的计算方式类似。

$$Q^{i,r} = W_Q^r \mathbf{v}_{i,cls}^{0\ \prime} \tag{15}$$

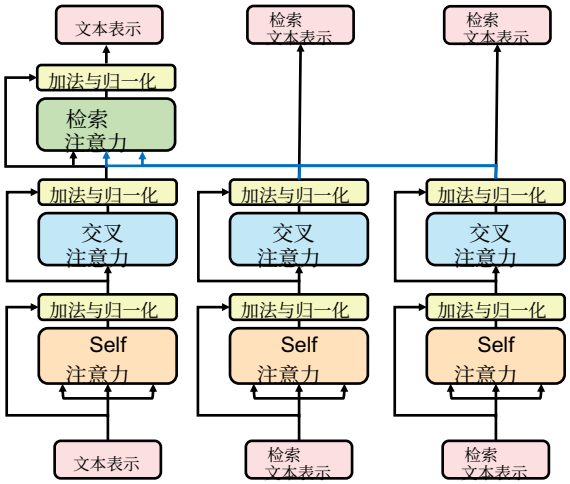


图4. 文本模态的检索注意力架构。

$$K^{i,r} = W_K^r [\mathbf{v}_{i,cls}^{0\ \prime}, \mathbf{v}_{i,cls}^{1\ \prime}, ..., \mathbf{v}_{i,cls}^{r\ \prime}] \tag{16}$$

$$V^{i,r} = W_V^r [\mathbf{v}_{i,cls}^{0\ \prime}, \mathbf{v}_{i,cls}^{1\ \prime}, ..., \mathbf{v}_{i,cls}^{r\ \prime}] \tag{17}$$

$$\mathbf{v}_{i+1,cls}^j = \begin{cases} \mathbf{v}_{i,cls}^{j\ \prime} + \mathrm{Attn}(Q^{i,r}, K^{i,r}, V^{i,r}), & j = 0 \\ \mathbf{v}_{i,cls}^{j\ \prime}, & j > 0 \end{cases} \tag{18}$$

我们对原始表示 $\mathbf{w}_i^{0\prime}$ 或$\mathbf{v}_i^{0\prime}$ 执行检索注意力，以融合来自$\{\mathbf{w}_i^{j\prime}|1 \leq j \leq r\}$ 或 $\{\mathbf{v}_i^{j\prime}|1 \leq j \leq r\}$的额外信息。我们不会对检索样本应用该操作，因为无需在它们之间交换信息。我们使用原始样本表示的最后一层（$\mathbf{w}_L^0$，$\mathbf{v}_L^0$）进行视觉问答分类。

4. 实验

4.1. 数据集

我们在四个生物医学视觉问答数据集上进行实验，即 VQA-Med 2019 [2]，VQA-Med 2021 [16],VQARAD [22]，和 SLAKE（英文版） [30]。为与以往工作进行公平比较，对于 VQA-Med 2019、VQARAD 和 SLAKE，我们遵循 Chen等人提出的相同数据划分。 [4] ；对于 VQA-Med 2021，则遵循 Gong 等人提出的划分。 [13] 我们使用准确率衡量模型性能。这些数据集的详细统计信息列于附录 A。

4.2. 基线方法

我们将 RAMM 与生物医学视觉问答中的非预训练方法 [7,13,21,42,43]以及预训练方法 [3,29,46]进行比较。由于以下方法与我们的方法更相关，因此予以介绍:

PubmedCLIP [11] 使用 ROCO 对 CLIP 进行微调，以获得生物医学图像和文本表示。

数据集	数量	来源
ROCO	70,306	PubMed
MIMIC-CXR	225,829	BIDMC
PMCPM	398,109	病例报告

表 1. 预训练数据集统计，其中数量表示训练集中图文对的数量。BIDMC 是贝斯以色列女执事医疗中心的缩写。

MMBert [20] 输入图像的卷积神经网络特征和掩码词元，以进行掩码语言模型预训练。**MTPT** [12] 提出了一种多任务预训练方法，包含图像理解任务和问题-图像匹配任务。**M3AE** [4] 通过重建掩码词元和图像，对生物医学多模态模型进行预训练。

4.3. 实现细节

预训练数据 如第3.1节所述，我们从 PMC OA 收集了 39.8 万对出版质量高、类型多样的生物医学图文对，并将其称为 PMCPM。随后，我们将 PMCPM 与两个广泛使用的生物医学预训练数据集 ROCO [36] 和 MIMIC-CXR [18]结合起来进行模型预训练。遵循先前的工作，我们仅使用它们的训练集作为预训练数据集，并且对于 MIMIC-CXR，仅使用正面视图的图像。训练数据的统计信息如表1所示。

预训练设置 对于单模态编码器，我们使用 PubmedBert 初始化文本编码器 [14]，并使用 Swin-Base 初始化图像编码器 [31]。对于多模态编码器，我们采用一个 6 层 Transformer [40]，其隐藏维度为 $d = 768$，注意力头数为 $n_{head} = 12$，并采用随机初始化。将图像转换为分辨率 224×224，并使用 RandAugment [6]。我们采用动量蒸馏 [27]，动量参数设为 0.995。模型训练 30 个 epoch，批量大小为 512。我们使用 AdamW [33] 优化模型，并采用线性学习率衰减。详细的预训练超参数请参阅附录 B。

微调设置 与以往工作一致，我们将生物医学视觉问答视为分类问题，并使用交叉熵损失优化模型。对于所有数据集，我们对预训练模型进行 100 个周期的微调。如第3.4节所述，我们使用 RAMM 中冻结的单模态编码器检索图像- $r = 4$文本对。我们从 ROCO、MIMIC-CXR 和 PMCPM 的组合中为每个实例检索 224×224 对。微调期间，图像会被转换为分辨率 [6]。在 VQA-Med2019、VQARAD 和 SLAKE 中，我们对图像进行中心裁剪；在 VQA-Med 2021 中，则使用 RandAugment。指数移动平均 $\theta = 0.999$ 和 R-Drop [28] $\alpha = 0.6$ 微调期间也会使用这些方法。详细的超参数列于附录 C。

模型	VQAMed2019	VQARAD	SLAKE
无预训练			
MFB [43]	-	50.60	73.30
SAN [42]	-	54.30	76.00
BAN [21]	-	58.30	76.30
使用预训练			
MEVF-BAN [3]	77.86	66.10	78.60
CPRD-BAN [29]	-	67.80	81.10
COND-RE [46]	-	71.60	-
PubmedCLIP [11]	-	72.10	80.10
MMBert [20]	77.90	72.00	-
MTPT [12]	-	73.20	-
M3AE [4]	79.87	77.01	83.25
RAMM	**82.13**	**78.27**	**86.05**
不使用检索	80.27	76.72	85.58

表 2. VQA-Med 2019、VQARAD 和 SLAKE 数据集上的结果。

模型	VQA-Med 2021
VilMedic * [7]	37.80
SYSU-HCP * [13]	38.20
RAMM	**39.20**

表 3. VQA-Med 2021 上的结果。* 表示集成结果。

4.4. 主要结果

我们在四个数据集上将所提出的方法与基线方法进行了比较。如表 2 所示，采用预训练模型的方法明显优于未采用预训练模型的方法。我们提出的 RAMM 在 VQA-Med 2019、VQARAD 和 SLAKE 上分别取得了 82.13、78.27 和 86.05 的成绩，分别比此前的最佳结果高出 1.86、1.26 和 2.80。表 4 展示了在 VQA-Med 2021 数据集上的结果。RAMM 的准确率达到 39.20，甚至以 1.00 的优势超过了 VQA-Med 2021 挑战赛中的最先进集成方法。

我们还报告了 RAMM 在 VQA-Med 2019、VQARAD 和 SLAKE 数据集上不检索额外图文对时的结果。如表 2 所示，其在 VQA-Med 2019、VQARAD 和 SLAKE 上分别取得了 80.27、76.72 和 85.58 的成绩。即使没有检索增强，RAMM 在 VQA-Med 2019 上仍以 0.4 的优势、在 SLAKE 上以 2.33 的优势超越了此前的最先进方法 M3AE，在 VQARAD 上的表现也不相上下（76.72 对比 77.01）。这表明，利用额外的 PMCPM 数据集进行预训练，能够为生物医学视觉问答任务学习到更优的视觉和文本表征。相比之下，

与未采用检索增强的模型相比，我们的 RAMM 在 VQA-Med 2019、VQARAD 和 SLAKE 上分别取得了 1.90、1.55 和 0.47 的稳定提升，这表明我们提出的检索增强方法能够提升生物医学视觉问答的性能。

4.5. 消融研究

为进一步分析所收集的 PMCPM 数据集和检索策略的影响，我们进行了详细的消融研究，如表4所示。

首先，我们分析了在预训练阶段使用不同数据的影响。ROCO 和 MIMIC-CXR 是广泛使用的生物医学多模态预训练数据集。ROCO 包含数量相对较少的带标注放射学图像。MIMIC-CXR 的数据量很大，但仅包含胸部 X 射线图文对。PMCPM 数据集涵盖了患者丰富多样的病情，规模远大于 ROCO 和 MIMIC-CXR。借助 PMCPM 中大量且多样的图文对，模型有望获得更好的视觉和文本表征。通过比较 (a) 和 (b)，我们展示了在所收集的 PMCPM 数据集上进行预训练的有效性。(b) 和 (c) 的结果表明，PMCPM 数据与此前使用的 ROCO 和 MIMIC-CXR 可以相互补充。ROCO 和 MIMIC-CXR 在 VQA-Med-2019 和 VQARAD 数据集上的表现更好，而 PMCPM 在 SLAKE 数据集上的表现更佳。将这两种设置与三种数据源相结合后，设置 (d) 的表现大幅优于单独使用其中任一种设置。上述结果证明，PMCPM 能够增强生物医学视觉问答任务中的模型预训练效果。

其次，在本研究中，上述数据集也被视为检索资源。因此，我们还分析了微调阶段使用不同数据的影响。设置 (e) 仅使用先前的 ROCO 和 MIMIC-CXR 数据集进行检索。通过与设置 (c) 比较，我们表明，所提出的检索策略在此前广泛使用的资源上同样有效。随后，我们进一步证明了所收集的 PMCPM 数据的实用性。与设置 (f) 和 (g) 相比，我们发现，检索额外的 PMCPM 数据分别带来了 1.06、1.11 和 0.38 的提升。此外，设置 (g) 与 (d) 的结果体现了我们提出的检索方法带来的显著提升。上述消融实验表明，我们收集的 PMCPM 数据集和提出的检索方法，既有助于预训练，也能使生物医学视觉问答任务中的检索受益。

4.6. 分析与讨论

4.6.1 检索

检索分布 为了了解 RAMM 如何从不同资源中检索内容，以及检索到的文本如何发挥作用于生物医学视觉问答任务，我们统计了检索样本的来源，以及检索到的文本中有多少包含答案。统计结果列于表5。在 VQA-Med 2019、VQARAD 和 SLAKE 中，分别有 29.6%、28.4% 和 38.7% 的答案可以在检索到的文本中找到，这一比例相当可观。这些检索到的文本可能有助于 RAMM 直接回答问题。我们发现，在这些任务中，RAMM 检索到的图文对有一半以上来自 PMCPM，这表明与 ROCO 和 MIMIC-CXR 相比，PMCPM 与生物医学视觉问答图像的相关性更高。在 VQA-Med 2019 和 VQARAD 中，PMCPM 与 *PMCPM* 之外的来源所检索到的、包含答案的文本比例几乎相同。在 SLAKE 中，PMCPM 提供了更多包含答案的文本。总之，PMCPM 在检索过程中非常重要，能够提供额外的包含答案的文本。

图文对检索数量 为探究图文对检索数量如何影响生物医学视觉问答微调，我们搜索了检索图像的数量范围 $\{0, 1, 2, 4, 8\}$。VQARAD 和 SLAKE 的结果如表6所示。与原始微调（即检索数量为 0）相比，检索 $\{1, 2, 4\}$ 个图文对可以提升性能。根据经验，我们发现检索 2 或 4 个图文对效果较好。此外，使用 8 个图文对会降低模型性能，这表明检索过多图文对可能会引入无关的图文对及其中的噪声知识。

4.6.2 问题类型

在生物医学视觉问答中，问题可分为封闭式问题和开放式问题。封闭式问题的答案仅限于选择题，例如 是或否，而开放式问题的答案则不受此限制。由于题型形式不同，封闭式问题比开放式问题更容易回答。表7 列出了不同方法在开放式问题和封闭式问题上的对比。RAMM 在封闭式问题和开放式问题上的表现均优于 M3AE。与开放式问题相比，RAMM 在封闭式问题上对 M3AE 的提升更大。这是因为封闭式问题的答案选项有限，检索到的文本更容易覆盖这些选项。

4.6.3 与 ROCO 的比较

ROCO 和 PMCPM 都是 PubMed 的子集。ROCO 使用训练好的神经网络筛选放射学图像，而 PMCPM 则通过 PubMed 中的“病例报告”查找图像。我们统计了 ROCO 和 PMCPM 训练集所涉及的文章数量。ROCO 使用了 PubMed 中的 40,857 篇文章，而 PMCPM 使用了 PubMed 中的 127,455 篇文章。两个数据集中共有 17,189 篇文章重叠。尽管 ROCO 和 PMCPM 使用不同的方法筛选图文对，

	预训练		检索		视觉问答数据集		
	PMCPM	ROCO+胸部X光片	PMCPM	ROCO+胸部X光片	VQAMed19	VQARAD	SLAKE
(a)	✗	✗	✗	✗	77.60	74.06	81.43
(b)	✓	✗	✗	✗	78.40	74.94	83.60
(c)	✗	✓	✗	✗	79.73	75.17	82.94
(d)	✓	✓	✗	✗	80.27	76.72	85.58
(e)	✗	✓	✗	✓	80.00	76.72	85.20
(f)	✓	✓	✗	✓	81.07	77.16	85.67
(g)	✓	✓	✓	✓	**82.13**	**78.27**	**86.05**

表 4. 结果　　RAMM 在 VQA-Med 2019、VQARAD 和 SLAKE 数据集上采用不同预训练和检索设置的结果　gs.

Task	VQAMed19		VQARAD		SLAKE	
数据集	检索率	有答案率	检索率	有答案率	检索率	有答案率
PMCPM + ROCO + MIMIC-CXR	100.0%	29.6%	100.0%	28.4%	100.0%	38.7%
PMCPM	52.2%	20.8%	51.9%	**21.3%**	54.9%	**30.5%**
ROCO + MIMIC-CXR	47.8%	**21.1%**	48.1%	20.0%	45.1%	26.9%

Tab图 5. VQA-Med 2019、VQARAD 和 SLAKE 数据集上检索来源及质量分析　　ts.

# 检索	VQARAD	SLAKE
0	76.72	85.58
1	77.38	**86.33**
2	78.05	**86.33**
4	**78.27**	86.05
8	76.50	84.54

表 6. 使用不同 r 检索到的图像-文本对时，VQARAD 和 SLAKE 数据集上的结果。

模型	VQARAD		SLAKE	
	Open	封闭式	Open	已关闭
MEVF-BAN	49.20	77.20	77.80	79.80
CPRD-BAN	52.50	77.90	79.50	83.40
M3AE	67.23	83.46	80.31	87.82
RAMM	**67.60**	**85.29**	**82.48**	**91.59**

表7. VQARAD 和 SLAKE 数据集上开放式与封闭式结果分析。

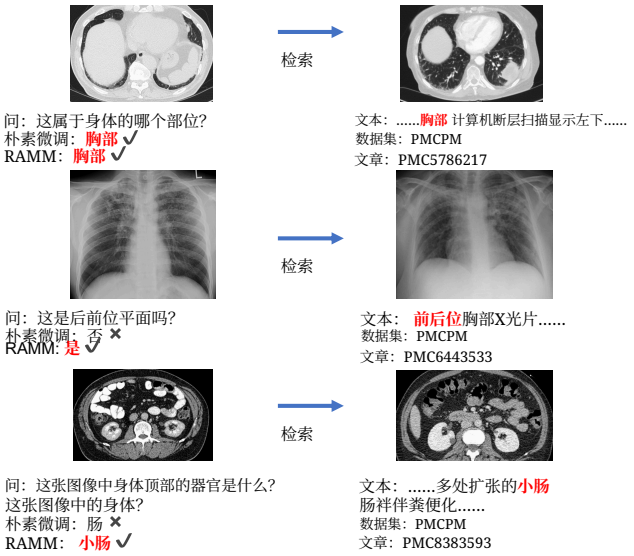


图5. RAMM 检索到的图像-文本对示例。

ROCO 中近一半的图文对也由 PMCPM 收集。鉴于 ROCO 是经过人工校正的数据集，因此它可以从侧面反映规模更大的 PMCPM 数据集的质量。

4.6.4 案例研究

我们在图5中展示了 RAMM 检索到的图文对。检索样本均来自 PMCPM。检索到的图像与原始图像属于相同模态，且器官相同。这些案例表明，检索到的图像包含有助于回答问题的信息。虽然第二个示例的答案（即 是/否）未在检索到的文本中提及，但检索到的文本仍有助于回答问题。第二个和第三个示例进一步表明，RAMM 借助检索增强能够正确作答，而朴素微调则无法做到。

5. 结论

本文提出了一种用于生物医学视觉问答的检索增强预训练与微调范式 RAMM，其中包括一个高质量图像-文本对数据集 PMCPM、一个预

训练多模态模型，以及一种用于微调的新型检索增强注意力模块。RAMM 使用 ITC 检索相关图文对，并通过检索注意力融合检索到的信息。实验表明，我们的方法在各种生物医学视觉问答数据集上优于最先进的方法。消融实验表明，所提出的 RAMM 和 PMCPM 在预训练和检索方面均有效。

参考文献

[1] Iz Beltagy、Kyle Lo 和 Arman Cohan。《SciBERT：用于科学文本的预训练语言模型》。载于《2019年自然语言处理经验方法会议暨第九届国际自然语言处理联合会议（EMNLP-IJCNLP）论文集》，第3615–3620页，中国香港，2019年11月。计算语言学协会。2[2]Asma Ben Abacha、Sadid A Hasan、Vivek V Datla、Dina Demner-Fushman 和 Henning M¨uller。《Vqa-med：imageclef 2019医学视觉问答任务概述》。载于CLEF（评估论坛会议与实验室）2019年工作笔记论文集。9月9日至12日 2019, 2019 2, 5 。D. Nguyen Binh、[3] Do Thanh-Toan、X. Nguyen Binh、Do Tuong、Tjiputra Erman 和 D. Tran Quang。《克服医学视觉问答中的数据限制》。载于MICCAI，2019年。2、5、6[4] Zhihong Chen、Yuhao Du、Jinpeng Hu、Yang Liu、Guanbin Li、Xiang Wan 和 Tsung-Hui Chang。《用于医学视觉语言预训练的多模态掩码自编码器》。载于《医学图像计算与计算机辅助干预国际会议》。施普林格，2022年。1、2、5、6[5] Zhihong Chen、Guanbin Li 和 Xiang Wan。《对齐、推理与学习：利用知识增强医学视觉语言预训练》。载于《第30届ACM国际多媒体会议论文集》，第5152 5161页，2022年2 – 。[6] Ekin Dogus Cubuk、Barret Zoph、Jon Shlens 和 Quoc Le。《Randaugment：以缩小的搜索空间实现实用的自动化数据增强》。载于 H. Larochelle、M. Ranzato、R. Hadsell、M.F. Balcan 和 H. Lin 编，《神经信息处理系统进展》第33卷，第18613–18624页。Curran Associates公司，2020年。6[7] Jean-benoit Delbrouck、Khaled Saab、Maya Varma、Sabri Eyuboglu、Pierre Chambon、Jared Dunnmon、Juan Zambrano、Akshay Chaudhari 和 Curtis Langlotz。《ViLMedic：一个致力于医学人工智能中视觉与语言交叉领域研究的框架》。载于《第60届计算语言学协会年会：系统演示论文集》，第23–34页，爱尔兰都柏林，2022年5月。计算语言学协会。5、6[8] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。《BERT：用于语言理解的深度双向变换器预训练》。载于《2019年北美计算语言学协会分会年会论文集：人类语言技术，第1卷（长短篇论文），第4171–4186页，明尼阿波利斯，明尼苏达州，2019年6月。计算语言学协会。 4[9] Tuong Do、Binh X. Nguyen、Erman Tjiputra、Minh Tran、Quang D. Tran 和 Anh Nguyen。用于医学视觉问答的多种元模型量化方法。发表于 MIC- CAI，2021年。 2[10] Zi-Yi Dou、Yichong Xu、Zhe Gan、Jianfeng Wang、Shuohang Wang、Lijuan Wang、Chenguang Zhu、Pengchuan Zhang、Lu Yuan、Nanyun Peng、Zicheng Liu 和 Michael Zeng。端到端视觉语言变换器训练的实证研究。发表于计算机视觉与模式识别会议（CVPR），2022年。 1, 2, 3[11] Sedigheh Eslami、Gerard de Melo 和 Christoph Meinel。CLIP 在医学领域中对视觉问答的助益是否与其在通用领域中的助益相当？arXiv电子预印本，2021年12月。 2, 5, 6[12] Haifan Gong、Guanqi Chen、Sishuo Liu、Yizhou Yu 和 Guanbin Li。通过多任务预训练实现医学视觉问答的跨模态自注意力。发表于 ICMR '21：国际多媒体检索会议，台北，台湾，2021年8月21日至24日，第456–460页。ACM，2021年。 6[13] Haifan Gong、Ricong Huang、Guanqi Chen 和 Guanbin Li。SYSU-HCP 在 VQA-Med 2021 中的方案：一种采用高效训练方法的以数据为中心的医学视觉问答模型。发表于 CLEF 2021——评测论坛会议与实验室，2021年9月21–24日，布加勒斯特，罗马尼亚，CEUR研讨会论文集，2021年。 5, 6[14] Yuxian Gu、Robert Tinn、Hao Cheng、Michael R. Lucas、Naoto Usuyama、Xiaodong Liu、Tristan Naumann、Jianfeng Gao 和 Hoifung Poon。面向生物医学自然语言处理的领域专用语言模型预训练。ACM医疗保健计算汇刊（HEALTH），3:1–23，2022年。 2, 6[15] Kelvin Guu、Kenton Lee、Zora Tung、Panupong Pasupat 和 Mingwei Chang。检索增强语言模型预训练。发表于国际机器学习会议，第3929–3938页。PMLR，2020年。 2[16] Bogdan Ionescu、Henning M¨uller、Renaud P´eteri、Asma Ben Abacha、Mourad Sarrouti、Dina Demner-Fushman、Sadid A Hasan、Serge Kozlovski、Vitali Liauchuk、Yashin Dicente Cid 等。ImageCLEF 2021 综述：医学、自然、互联网和社交媒体应用中的多媒体检索。发表于欧洲语言跨语言评测论坛国际会议，第345–370页。施普林格，2021年。 2, 5[17] Qiao Jin、Zheng Yuan、Guangzhi Xiong、Qianlan Yu、Huaiyuan Ying、Chuanqi Tan、Mosha Chen、Songfang Huang、Xiaozhong Liu 和 Sheng Yu。生物医学问答：方法与挑战综述。ACM计算机评论，55(2)，2022年1月。 1[18] Alistair E. W. Johnson、Tom J. Pollard、Seth J. Berkowitz、Nathaniel R. Greenbaum、Matthew P. Lungren、Chih ying Deng、Roger G. Mark 和 Steven Horng。MIMIC-CXR：一个去标识化且公开可用的胸部X线片数据库，附有自由文本报告。Scientific Data，6，2019年。 2, 6

[19] Jeff Johnson、Matthijs Douze 和 Herv´e J´egou。利用图形处理器进行十亿级相似性搜索。IEEE Transactions on Big Data，7(3):535–547，2019年。 5[20] Yash Khare、Viraj Bagal、Minesh Mathew、Adithi Devi、U Deva Priyakumar 和 CV Jawahar。MMBert：用于改进医学视觉问答的多模态 BERT 预训练。发表于2021 年第 18 届 IEEE 国际生物医学成像研讨会（ISBI），第1033–1036页。IEEE，2021年。 1, 2, 6[21] Jin-Hwa Kim 、 Jaehyun Jun 和 Byoung-Tak Zhang。双线性注意力网络。发表于第 31 届神经信息处理系统进展，第1571–1581页，2018年。 5, 6[22] Jason J Lau、Soumya Gayen、Asma Ben Abacha 和 Dina Demner-Fushman。一个关于放射学图像的临床生成视觉问答数据集。Scientific Data，5(1):1–10，2018年。 2, 5[23] Jinhyuk Lee、Wonjin Yoon、Sungdong Kim、Donghyeon Kim、Sunkyu Kim、Chan Ho So 和 Jaewoo Kang。BioBERT：用于生物医学文本挖掘的预训练生物医学语言表征模型。Bioinformatics，36 1240，2020 4 1234 2 ( ): – 。[24] Kenton Lee、Ming-Wei Chang 和 Kristina Toutanova。用于弱监督开放域问答的潜在检索。发表于第 57 届计算语言学协会年会论文集，第6086–6096页，意大利佛罗伦萨，2019 年 7 月。计算语言学协会。 2[25]Patrick Lewis、Ethan Perez、Aleksandra Piktus、Fabio Petroni、Vladimir Karpukhin、Naman Goyal、Heinrich K¨uttler、Mike Lewis、Wen-tau Yih、Tim Rockt¨aschel 等。面向知识密集型自然语言处理任务的检索增强生成。神经信息处理系统进展，33:9459–9474，2020年。 2[26] Chenliang Li、Haiyang Xu、Junfeng Tian、Wei Wang、Ming Yan、Bin Bi、Jiabo Ye、Hehong Chen、Guohai Xu、Zheng Cao 等。mPLUG：通过跨模态跳跃连接实现高效的视觉-语言学习。arXiv预印本 arXiv:2205.12005，2022年。 1, 2[27] Junnan Li、Ramprasaath R. Selvaraju、Akhilesh Deepak Got- mare、Shafiq Joty、Caiming Xiong 和 Steven Hoi。先对齐再融合：通过动量蒸馏进行视觉与语言表征学习。发表于 NeurIPS，2021年。 1, 2, 4, 6[28] Xiaobo Liang、Lijun Wu、Juntao Li、Yue Wang、Qi Meng、Tao Qin、Wei Chen、Min Zhang 和 Tie-Yan Liu。R-drop：用于神经网络的正则化随机失活。发表于 NeurIPS，2021年。 6[29] Bo Liu、Li-Ming Zhan 和 Xiao-Ming Wu。基于放射学图像的医学视觉问答对比预训练与表征蒸馏。发表于国际医学图像计算与计算机辅助干预会议，第210–220页。施普林格，2021年。 5, 6[30] Bo Liu、Li-Ming Zhan、Li Xu、Lin Ma、Yan Yang 和 Xiao-Ming Wu。SLAKE：一个语义标注、知识增强的医学视觉问答数据集。发表于2021 年第 18 届 IEEE 国际生物医学成像研讨会（ISBI），第1650–1654页。IEEE，2021年。 2, 5[31] Ze Liu、Yutong Lin、Yue Cao、Han Hu、Yixuan Wei、Zheng Zhang、Stephen Lin 和 Baining Guo。Swin Transformer：采用移位窗口的层次化视觉Transformer。载于IEEE/CVF国际计算机视觉会议论文集，第10012–10022页，2021年。6[32] Alexander Long、Wei Yin、Thalaiyasingam Ajanthan、Vu Nguyen、Pulak Purkait、Ravi Garg、Alan Blair、沈春华和Anton van den Hengel。面向长尾视觉识别的检索增强分类。载于IEEE/CVF计算机视觉与模式识别会议论文集，第6959–6969页，2022年。2[33] Ilya Loshchilov 和 Frank Hutter。解耦权重衰减正则化。载于第7届国际表征学习会议（ICLR 2019），美国路易斯安那州新奥尔良，2019年5月6日至9日，2019年。6[34] Jong Hak Moon、Hyungyung Lee、Woncheol Shin、Young-Hak Kim 和 Edward Choi。通过视觉语言预训练实现医学图像与文本的多模态理解与生成。《IEEE生物医学与健康信息学杂志》，2022年。2[35] Binh D Nguyen、Thanh-Toan Do、Binh X Nguyen、Tuong Do、Erman Tjiputra 和 Quang D Tran。克服医学视觉问答中的数据限制。载于国际医学图像计算与计算机辅助干预会议论文集，第522–530页。施普林格，2019年。2[36] Obioma Pelka、Sven Koitka、Johannes R¨uckert、Felix Nensa 和 Christoph M Friedrich。语境中的放射学对象（ROCO）：多模态图像数据集。载于《血管内成像与计算机辅助支架置入，以及生物医学数据的大规模标注与专家标签合成》，第180–189页。施普林格，2018年。2, 3, 6[37]Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。通过自然语言监督学习可迁移视觉模型。载于国际机器学习会议论文集，第8748–8763页。PMLR，2021年。2[38] Stephen E. Robertson 和 Hugo Zaragoza。概率相关性框架：BM25及其拓展。《信息检索基础与趋势》，3：333–389，2009年。2[39] Sara Sarto、Marcella Cornia、Lorenzo Baraldi 和 Rita Cucchiara。用于图像字幕生成的检索增强Transformer。载于国际基于内容的多媒体索引会议论文集，第1–7页，2022年。2[40] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。注意力机制就是一切。神经信息处理系统进展，30，2017年。6[41] Wenhui Wang、Hangbo Bao、Li Dong、Johan Bjorck、Zhil-iang Peng、Qiang Liu、Kriti Aggarwal、Owais Khan Mo-hammed、Saksham Singhal、Subhojit Som 等。将图像视为外语：面向所有视觉及视觉语言任务的BEiT预训练。arXiv预印本 arXiv:2208.10442，2022年。2[42] Zichao Yang、He Xiaodong、Gao Jianfeng、Li Deng 和 Alex Smola。用于图像问答的堆叠注意力网络。载于 IEEE计算机视觉与模式识别会议论文集，第21–29页，2016年。5, 6

[43] Zhou Yu、Jun Yu、Jianping Fan 和 Dacheng Tao。通过协同注意力学习实现用于视觉问答的多模态因子化双线性池化。*IEEE*国际计算机视觉会议（*ICCV*），第1839–1848页，2017年。5，6[44] Hongyi Yuan、Zheng Yuan、Ruyi Gan、Jiaxing Zhang、Yutao Xie 和 Sheng Yu。BioBART：生物医学生成式语言模型的预训练与评估。载于第*21*届生物医学语言处理研讨会论文集，第97–109页，爱尔兰都柏林，2022年5月。计算语言学协会。2[45] Zheng Yuan、Yijia Liu、Chuanqi Tan、Songfang Huang 和 Fei Huang。利用知识改进生物医学预训练语言模型。载于第*20*届生物医学语言处理研讨会论文集，第180–190页，线上，2021年6月。计算语言学协会。2[46] Li-Ming Zhan、Bo Liu、Lu Fan、Jiaxin Chen 和 Xiao-Ming Wu。通过条件推理实现医学视觉问答。载于第*28*届*ACM*多媒体国际会议论文集，第2345–2354页，2020年。5，6[47] Zhengyun Zhao、Qiao Jin 和 Sheng Yu。PMC-Patients：从PubMed Central中的病例报告提取的患者病历与关系大规模数据集。*arXiv*预印本 *arXiv:2202.13876*，2022年。3