arXiv:2304.01611v2 [cs.CV] 5 Jun 2023

Q2ATransformer：通过答案查询解码器改进医学视觉问答

Yunyi Liu[1],Zhanyu Wang[1],Dong Xu[2],和 Luping Zhou[1]

[1] 悉尼大学，悉尼，新南威尔士州，澳大利亚
{yunyi.liu,zhanyu.wang,luping.zhou}@sydney.edu.au[2]
香港大学，香港特别行政区 dongxu@cs.hku.hk

摘要。 医学视觉问答（VQA）系统在理解医学图像所承载的临床相关信息方面发挥着辅助作用。针对医学图像的问题分为两类：封闭式问题（如是/否问题）和开放式问题。为获得答案，现有的大多数医学视觉问答方法依赖分类方法，而少数研究尝试采用生成方法，或将两者结合，分别处理这两类问题（封闭式问题采用分类，开放式问题采用生成）。分类方法相对简单，但在处理较长的开放式问题时表现不佳；生成方法则面临生成大量并不存在的答案这一挑战，导致准确率较低。为弥合这一差距，本文提出一种用于医学视觉问答的新型Transformer框架（称为Q2ATransformer），该框架兼具分类方法和生成方法的优势，并为封闭式问题和开放式问题提供统一的处理方式。具体而言，我们引入一个额外的Transformer解码器，并配备一组可学习的候选答案嵌入，用于查询给定图像—问题对中各答案类别是否存在。通过Transformer注意力，候选答案嵌入与图像—问题对的融合特征进行交互，从而作出判断。这样一来，尽管我们的方法基于分类，却能像基于生成的方法一样，通过与答案信息交互来进行预测。另一方面，通过采用分类方法，我们缩小了答案的搜索空间，从而降低了任务难度。我们的方法在两个医学视觉问答基准上取得了新的最优性能。尤其是在开放式问题上，我们在VQA-RAD和PathVQA上分别取得79.19%和54.85%的成绩，绝对提升分别为16.09%和41.45%。

关键词： 医学视觉问答 · 注意力机制 · 分类。

1 引言

视觉问答（VQA）是一项众所周知的具有挑战性的人工智能任务，需要根据图像内容回答图像相关问题。这一过程涉及图像和

自然语言处理技术，通常包含四个关键组成部分：提取图像特征、提取问题特征、整合特征以及作答。近年来，这一领域取得了显著进展 [8,15]。医学视觉问答是视觉问答在医学图像及其临床相关问题上的自然延伸。通过提问和作答，它为辅助临床决策提供了一种用户友好的方式。医学视觉问答中的问题既可以是封闭式问题（例如是/否问题），也可以是开放式问题。

医学视觉问答仍处于发展的早期阶段，目前的性能远未达到令人满意的程度。大多数现有方法 [13,4,7,6,5] 可被称为封闭式方法，如图 1(a)所示；这类方法将每个答案视为一个类别，并直接对输入图像—问题对的融合特征应用分类模型来预测答案。这类方法的优势在于，将视觉问答视为分类任务后，可以降低任务的复杂度，并缩小答案搜索空间。尽管封闭式方法在是/否问题上表现良好，但对于开放式问题，它们难以准确预测答案，因为开放式问题的答案通常比封闭式问题的答案更长、变化更多。另一方面，少数研究[2,9] 将视觉问答视为生成任务，并采用基于生成的方法逐词生成答案。在图 1(b)中，这些方法被称为开放式方法。在这类方法中，当前的词语生成通常依赖于答案中先前生成的词语。因此，这些方法允许图像-问题特征与答案信息进行交互以完成预测，从而有望提升长答案的预测效果。然而，由于生成答案的搜索空间极其庞大，这些方法往往会生成许多不存在的答案，导致准确率较低，因此目前并不是医学视觉问答的主流方法。尽管已有研究尝试结合这两类方法 [14], ，但它们只是直接将封闭式问题和开放式问题区别处理，例如对封闭式问题采用分类方法，对开放式问题采用生成方法。

为弥补研究空白并推动医学视觉问答的发展，我们提出了一种新的模型框架 Q2ATransformer，并将其称为半开放式，如图 1（c）所示。所谓半开放式，是指我们继续采用基于分类的方法来缩小答案搜索空间，同时引入答案语义的学习，使融合后的图像-问题特征与答案语义能够相互作用，以实现更好的预测，类似于基于生成的方法。我们的模型弥补了基于分类的封闭式框架的不足，同时具备基于生成的开放式框架的优势。为此，我们引入一组可学习的候选答案嵌入，并将其输入 Transformer解码器，使图像-问题特征与候选答案嵌入相互作用。在解码器中，候选答案嵌入作为查询，用于计算其与融合后的图像-问题特征之间的关系，从而判断答案类别是否存在。借此，我们的分类方法考虑了答案信息与融合后的图像-问题特征之间的相互作用，这与现有的基于分类的方法不同。与基于生成的方法相比，

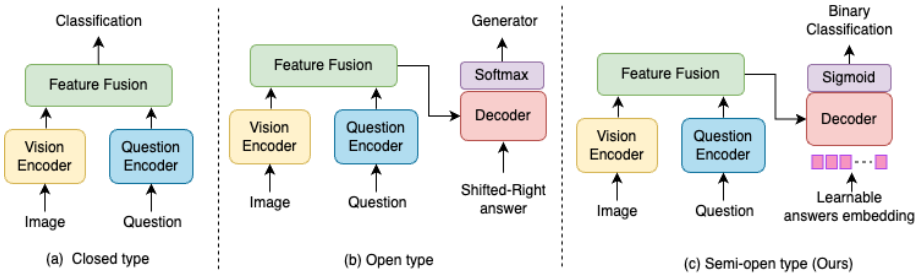


图1. 医学视觉问答框架范式。(a) 封闭式框架将视觉问答视为答案类别预测，并直接在融合后的图像-问题特征之上构建分类器。(b) 开放式框架基于生成，融合后的图像-问题特征与答案的前序词语交互，并通过文本解码器生成答案的下一个词语。(c) 我们提出的半开放式框架通过解码器学习候选答案嵌入，并使其与融合后的图像-问题特征交互，从而改进答案类别预测。

开放式方法下，我们的模型降低了任务难度，并显著提高了准确率。

最后但同样重要的是，我们的模型对封闭式问题和开放式问题一视同仁。

本文的主要贡献可总结如下。

首先，我们为医学视觉问答提出了一种半开放式框架，它融合了医学视觉问答文献中基于分类的封闭式框架和基于生成的开放式框架各自的优势。该框架通过一种精心设计的机制，利用 Transformer 解码器学习并使用候选答案嵌入，同时通过分类限制答案的搜索范围。

其次，我们提出了跨模态融合网络（CMAN），以有效融合图像和问题特征。它直接拼接两种模态的特征，而不是通过矩阵乘法或求和进行特征融合，从而减轻信息损失。随后，通过在拼接后的特征上计算自注意力来捕获图像特征与问题特征之间的关系，生成融合特征。正如我们的消融实验所示，CMAN 在医学视觉问答中优于常用的图像-问题融合方法。

第三，我们的模型在两个大型医学视觉问答基准上，无论是封闭式问题还是开放式问题，都展现出了优越的性能。尤其是，我们在开放式问题回答方面的提升十分显著，在 VQA-RAD 和 PathVQA 上的绝对提升幅度分别达到 16% 和 41%，验证了我们提出的半开放式框架的有效性。

2 方法

本节将介绍 Q2ATransformer，这是一个用于医学视觉问答的半开放式结构化模型。我们首先概述模型，然后在第 3.1 节介绍视觉-问题编码器，并在第 3.2 节介绍答案查询解码器。

图 2 展示了我们提出的 Q2ATransformer 模型概览。与大多数医学视觉问答方法一样，该模型预测答案类别，但利用候选答案嵌入进行预测。Q2ATransformer 由视觉-问题编码器和答案查询解码器组成。视觉-问题编码器以医学图像和临床相关问题为输入，输出融合了图像与问题信息的融合特征。它由三个部分组成：视觉编码器、问题编码器和融合网络。我们采用 Swin Transformer 作为视觉编码器，采用 BERT 作为问题编码器。对于融合网络，我们提出了跨模态注意力网络（CMAN），用于整合图像特征和问题特征。答案查询解码器以融合的图像-问题特征和可学习的候选答案嵌入为输入，输出每个候选答案的概率。我们的答案查询解码器由两层 Transformer 解码器和一个用于预测的分类器组成。

2.1 视觉-问题编码器

视觉-问题编码器由图像编码器、问题编码器和特征融合模块组成，具体如下。

图像编码器。 我们的编码器使用 Swin Transformer [12]，而非基于卷积神经网络的模型作为图像特征提取器。Swin Transformer 有三大优势。首先，Swin Transformer 将视觉 Trans former 转变为类似卷积神经网络的层次结构，使视觉 Transformer 能够更灵活地处理不同尺度，并且其计算复杂度会随着图像尺寸增加而呈线性增长。其次，Swin Transformer 通过窗口移位考虑跨窗口连接，以获取长程依赖关系，从而增强网格之间的交互。因此，与卷积神经网络相比，它能提供更多区域特征和交互，更适合医学图像的细粒度特性。第三，Swin Transformer 在大型数据集上进行了预训练，因此是一个非常稳健的特征提取器。基于这些特性，我们选择 Swin Transformer 对输入图像进行编码。

给定输入图像$\mathbf{I} \in \mathbb{R}^{H \times W \times C}$，其中 C 是通道数，H 和 W 分别表示图像的高度和宽度，图像嵌入$\mathbf{F}_i \in \mathbb{R}^{N \times D_f}$可表示为$\mathbf{F}_i = \mathbf{W}_i \times \text{SwinTransformer}(\mathbf{I}) + \mathbf{b}_i$，其中$\mathbf{W}_i$和$\mathbf{b}_i$是可学习参数，用于将 Swin Transformer 的输出投影到与问题嵌入相同的维度 D_f。它们还为调整 Swin Transformer 以适应我们任务中的数据集提供了一定的灵活性。这里， N 表示提取的图像区域特征的数量。

问题编码器。对于输入问题，我们使用预训练的 BERT 模型 [3] 作为编码器来提取文本特征。BERT [3] 是一种成功的自然语言处理模型。它在对问题进行嵌入时，会结合句子两个方向的上下文。它已应用于问答任务，并达到当前最先进的水平。

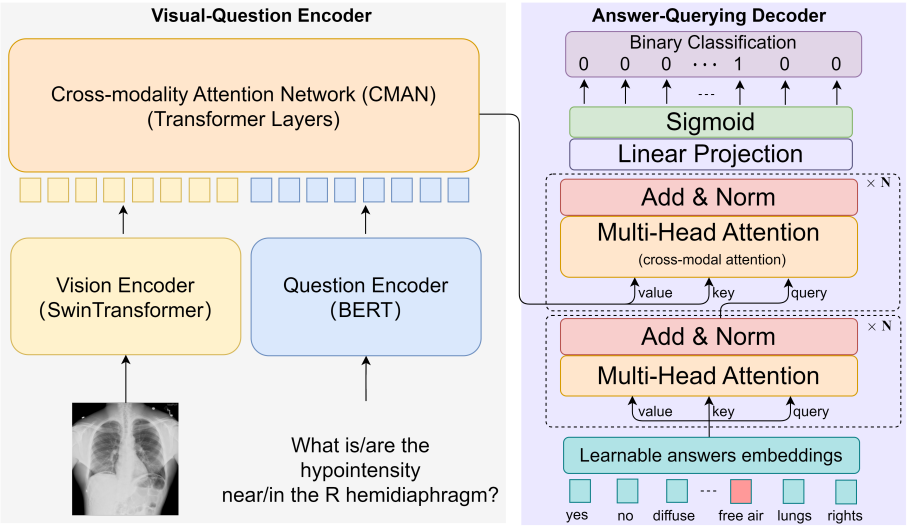


图 2. Q2ATransformer 概述。输入的图像—问题对会送入视觉-问题编码器，以提取并融合图像和问题特征。视觉-问题编码器由基于 Swin Transformer 的视觉编码器、基于 BERT 的问题编码器，以及用于特征融合的跨模态注意力网络组成。融合特征随后传递至答案查询解码器，其中输入的可学习候选答案嵌入用作查询，以计算注意力图，并根据受注意力作用的融合图像-问题特征进行细化，从而预测查询的答案是否存在。

目前最先进的结果，因此在我们的任务中被选作问题编码器。问题嵌入 $\mathbf{F}_q \in \mathbb{R}^{M \times D_f}$ 通过 $\mathbf{F}_q = \text{BERT}(\mathbf{Q}_e)$获得，其中$\mathbf{Q}_e$ 表示输入问题， M 表示问题特征数量， D_f表示问题特征维度。

特征融合机制。 分别提取图像和问题特征后，我们提出跨模态注意力网络（CMAN），以融合这两种模态的信息。由于医学图像具有细粒度特征，且具有临床重要性的视觉差异往往十分细微，因此我们通过探究图像区域特征与问题特征之间的交互，探索一种复杂的特征融合方法。在所提出的融合模块 CMAN 中，我们首先将图像特征$\mathbf{F}_i$ 与问题特征$\mathbf{F}_q$ 拼接起来。与通常用于特征融合的矩阵乘法或求和相比，拼接能够减少信息损失，并有助于后续在模块中计算图像与问题之间的交互。随后，将拼接后的特征输入两个 Transformer 编码器层，通过 Transformer 的自注意力机制计算每一对图像问答特征之间的关系。由此，我们能够获得融合特征，其中保留了图像问答特征之间的关系，且信息损失最小

。数学上，融合特征$\mathbf{F}_f$ 的计算方式如下。

$$
\begin{aligned}
&\mathbf{F}_c = [\mathbf{F}_i; \mathbf{F}_q] \\
&\mathbf{Q}_{F_c} = \mathbf{W}_q \mathbf{F}_c, \quad \mathbf{K}_{F_c} = \mathbf{W}_k \mathbf{F}_c, \quad \mathbf{V}_{F_c} = \mathbf{W}_v \mathbf{F}_c \\
&\mathbf{F}_{att} = \mathbf{Att}(\mathbf{Q_{F_c}}, \mathbf{K_{F_c}}, \mathbf{V_{F_c}}) = \mathbf{softmax}(\frac{\mathbf{Q_{F_c}} \mathbf{K_{F_c}^T}}{\sqrt{\mathbf{d_k}}}) \mathbf{V_{F_c}} \\
&\mathbf{F}_f = \mathbf{W}_f \mathbf{F}_{att} + \mathbf{b}_f
\end{aligned}
\tag{1}
$$

此处$\mathbf{W}_q$、$\mathbf{W}_k$、$\mathbf{W}_v$、$\mathbf{W}_f$和$\mathbf{b}_f$ 是可学习参数，“;”表示拼接操作。矩阵$\mathbf{Q}_{F_c}$、$\mathbf{K}_{F_c}$、$\mathbf{V}_{F_c}$ 在自注意力计算中分别称为查询、键和值值，此处它们是拼接特征$\mathbf{F}_c$的线性变换。

2.2 答案查询解码器

给定一个输入图像-问题对，在一组感兴趣的答案中，我们的答案查询解码器会预测每个候选答案是否与相应的图像-问题对匹配，并将概率最高的候选答案作为最终答案。为此，我们采用一个由两层 Transformer 解码器和一个线性投影器组成的分类器，并引入一组可学习的候选答案嵌入，同时将融合后的图像-问题特征$\mathbf{F}_f$ 作为解码器的输入。假设总共有 C 个答案类别，我们需要 C 个候选答案嵌入，并使其与 C 个答案类别一一对应。这些答案嵌入共同表示为矩阵$\mathbf{A}$，其初始值为随机值，并会在训练过程中依次经过自注意力模块、交叉注意力模块和前馈网络（FFN）更新。自注意力模块和交叉注意力模块均实现多头自注意力（MSA（查询、键、值）），但使用的键、查询和值各不相同。自注意力模块使用$\mathbf{A}$ 构造所有键、查询和值 矩阵，以计算不同答案嵌入之间的关系。交叉注意力模块关注答案嵌入$\mathbf{A}$ 与融合后的图像-问题特征$\mathbf{F}_f$之间的关系。因此，它使用答案嵌入$\mathbf{A}$ 作为查询，使用融合后的图像-问题特征$\mathbf{F}_f$ 作为键 和值 来计算注意力，并通过结合经注意力处理的图像-问题特征进一步更新答案嵌入。数学上，将第 l层的答案嵌入记为$\mathbf{A}_l$，它将根据上一层的输出$\mathbf{A}_{l-1}$ 按如下方式更新:

$$
\begin{aligned}
&\mathbf{A}_l = MSA(\mathbf{A}_{l-1}, \mathbf{A}_{l-1}, \mathbf{A}_{l-1}) \\
&\mathbf{A}_l = MSA(\mathbf{A}_l, \mathbf{F}_f, \mathbf{F}_f) \\
&\mathbf{A}_l = FFN(\mathbf{A}_l),
\end{aligned}
\tag{2}
$$

其中，$l = 1 \cdots L$和 L是 Transformer解码器层的数量。在此过程中，图像-问题特征被注入答案嵌入中

并用于对后者进行细化。细化后的 C 答案嵌入被送入最终的线性投影层，随后经过 Sigmoid 函数 $\sigma(\cdot)$，以预测答案类别的概率。即：

$$\mathbf{p} = \sigma(\mathbf{W}_A \mathbf{A}_L + \mathbf{b}), \tag{3}$$

其中 $\mathbf{W}_A$ 和 $\mathbf{b}$ 是可学习参数，$\mathbf{p}$ 是一个向量，包含与 C 答案类别相对应的 C 概率。概率最高的答案类别将被选为预测答案。

2.3 损失函数

医学视觉问答面临着显著的类别不平衡问题：是/否答案类别的规模远大于长开放式答案类别。为了更有效地解决样本不平衡问题，我们选择了一种简化的非对称损失，它是焦点损失的一种变体，其中正类和负类的超参数 γ 设置不同，如公式 4 所示：

$$\mathcal{L} = \frac{1}{C} \sum_{c=1}^{C} \begin{cases} (1-p_c)^{\gamma+} \log\left(p_c\right), y_c = 1 \\ (p_c)^{\gamma-} \log\left(1-p_c\right), y_c = 0 \end{cases} \tag{4}$$

其中，y_c 是真实二元标签，表示输入的图像—问题对是否具有答案类别 c；p_c 是类别 c 的预测概率。总损失通过对训练数据集中的所有样本的该损失取平均值计算得出。默认情况下，我们将超参数设为 $\gamma+ = 1$ 和 $\gamma- = 4$。

3 实验与结果

3.1 数据集

我们在两个医学视觉问答基准上进行实验：VQA-RAD [11]和 PathVQA [7], ，具体介绍如下。

VQA-RAD 是迄今为止最常用的放射学数据集，包含 315 张图像和 3515 个问答对，每张图像至少对应一个问答对。问题类型包括 11 类：“异常”、“属性”、“颜色”、“数量”、“形态”、“器官类型”、“其他”和“切片”。58% 的问题是封闭式问题，其余为开放式问题。图像涉及人体的头部、胸部和腹部。训练集和测试集需要手动划分。为便于比较，我们按照 MMQ 方法[4]划分数据集。

PathVQA 是用于探索病理学视觉问答的数据集。带有说明文字的图像取自数字资源（电子教科书和在线图书馆）。开放式问题占所有问题的 50.2%。对于封闭式是/否问题，答案数量均衡，其中有 8,145 个“是”问题和 8,189 个“否”问题。PathVQA 由 32,799 个问答对、从两本病理学教科书中收集的 1,670 张病理图像，以及从 PEIR 数字图书馆中收集的 3,328 张病理图像组成 [1]。为便于比较，我们还根据 MMQ 方法 [4] 对数据集进行了划分。

3.2 与当前最先进方法的比较

我们将所提出的模型与 7 种最先进水平（SOTA）的医学视觉问答方法进行比较，包括 StAn [7], 、BiAn [7], 、MAML [6], 、MEVF [13],、MMQ [4], 、PubMedCLIP [5], 和 MMBERT [9]。前 6 种方法均为基于分类的方法。选择这些方法，是因为它们在 VQA-RAD 和 PathVQA 这两个基准测试中表现出色。最后一种方法 MMBERT [9] 被选作基于生成的方法的代表，其在 VQA-RAD 上的性能已有报告。除 PubMedCLIP [5] 和 MMBERT [9] 的结果引自其原始论文外，其他方法的结果均引自 MMQ [4]。值得注意的是，与我们的 Q2ATransformer 相同，PubMedCLIP [5] 和 MMBERT [9] 采用了与 MMQ [4] 相同的数据划分。因此，这些结果可以进行严格比较。

如表 1 所示，在两个数据集上，我们的 Q2ATransformer 始终优于所比较的模型。具体而言，与表现第二好的模型相比，在 VQA-RAD 上，我们在开放式问题上的准确率达到 79.19%（绝对提升 16.09%），在封闭式问题上的准确率达到 81.2%（绝对提升 1.2%），在所有问题上的准确率达到 80.48%（绝对提升 8.48%）；在 PathVQA 上，我们在开放式问题上的准确率达到 54.85%（绝对提升 41.45%），在是/否问题上的准确率达到 88.85%（绝对提升 4.85%），在所有问题上的准确率达到 74.61%（绝对提升 25.81%）。如我们的消融实验所示，如果增大候选答案嵌入的维度，结果还可能进一步提升。从这些结果可以看出，我们的 Q2ATransformer 在开放式问题上展现出显著优势，这支持了我们的分析：通过将答案信息与融合后的图像-问题特征进行交互，我们的模型能够更好地处理长答案问题。我们的模型也优于基于生成的方法 MMBERT [9],，因为我们减少了答案的搜索空间，而 MMBERT [9] 可能生成不存在的答案。

表1。 不同方法的性能比较。 † 和 ‡ 分别表示这些方法是基于分类（封闭式）还是基于生成（开放式）。

参考文献 方法	融合 方法	PathVQA			VQA-RAD		
		自由形式	是/否	总体	开放式	封闭式	总体
StAn†[7]	SAN	1.6	59.4	30.5	24.2	57.2	44.2
BiAn†[7]	BAN	2.9	68.2	35.6	28.4	67.9	52.3
MAML†[6]	SAN	5.4	75.3	40.5	38.2	69.7	57.1
	BAN	5.9	79.5	42.9	40.1	72.4	59.6
MEVF†[13]	SAN	6.0	81.0	43.6	40.7	74.1	60.7
	BAN	8.1	81.4	44.8	43.9	75.1	62.7
MMQ†[4]	SAN	11.2	82.7	47.1	46.3	75.7	64.0
	BAN	13.4	84.0	48.8	53.7	75.8	67.0
PubMedCLIP†[5]	-	-	-	-	60.1	80	72.1
MMBERT†[9]	-	-	-	-	63.1	77.9	72.0
Ours		**54.85**	**88.85**	**74.61**	**79.19**	**81.2**	**80.48**

表 2. 消融研究。BAN、SAN 和 CMAN 分别代表双线性注意力网络 [10], 堆叠注意力网络 [16] 和我们的方法跨模态注意力网络；解码器指我们的答案查询解码器。

#	BAN	SAN	CMAN	解码器	VQA-RAD			PathVQA		
					打开	关闭	总体	自由形式	是/否	总体
1	✓				43.62	75.56	64.1	15.03	78.24	51.69
2	✓			✓	54.36	80.07	70.84	44.78	88.29	70.09
3		✓			61.07	77.07	71.33	44.58	86.29	68.88
4		✓		✓	73.83	80.08	77.83	52.88	88.44	73.51
5			✓		69.13	76.32	73.73	47.53	86.73	70.31
6			✓	✓	79.19	81.2	80.48	54.85	88.85	74.61

3.3 消融实验

为了研究我们提出的特征融合模块 CMAN 和用于答案查询的解码器所发挥的作用，我们进行了广泛的消融研究，以比较模型的不同配置，如表 3.2 所示。其中，BAN、SAN 和 CMAN 是用于融合图像和问题特征的三种注意力网络，分别代表双线性注意力网络 [10], 、堆叠注意力网络 [16] 和我们的方法跨模态注意力网络；Decoder 表示我们的答案查询解码器。符号 ✓表示包含相应组件。表 3.2 中的所有实验均基于相同的图像和问题编码器进行。

CMAN 的影响。表 3.2 中，从 #1 到 #5 或从 #3 到 #5 的提升充分体现了使用 CMAN 的益处，表明与 BAN 和 SAN 相比，我们提出的 CMAN 在图像与问题特征融合方面更为有效。这是因为，与通过相乘图像和问题特征进行融合的 BAN，或通过直接矩阵求和进行融合的 SAN 相比，我们的 CMAN 直接将两个通道的特征拼接在一起，然后通过计算注意力进行融合。通过这种方式，我们的 CMAN 缓解了 BAN 或 SAN 在特征融合过程中因相乘或求和操作而造成的信息损失。

解码器的贡献。如图所示，纳入我们的答案查询解码器可以提升模型性能。为了验证我们解码器的稳健性，我们将其与 #2、#4 和 #6 中所示的三种不同注意力模块结合。通过比较 #2 与 #1、#4 与 #3，或 #6 与 #5，可以观察到，我们的答案查询解码器与这三种注意力机制结合时都能带来显著的性能提升。特别是，将我们的 CMAN 与解码器结合后，我们可以取得新的最先进水平结果。

答案嵌入维度的影响。图 3 中的实验结果表明，随着答案嵌入维度的增加，模型性能有所提升；当嵌入维度约为 2048 时，效果最佳。然而，增大嵌入维度也会增加计算

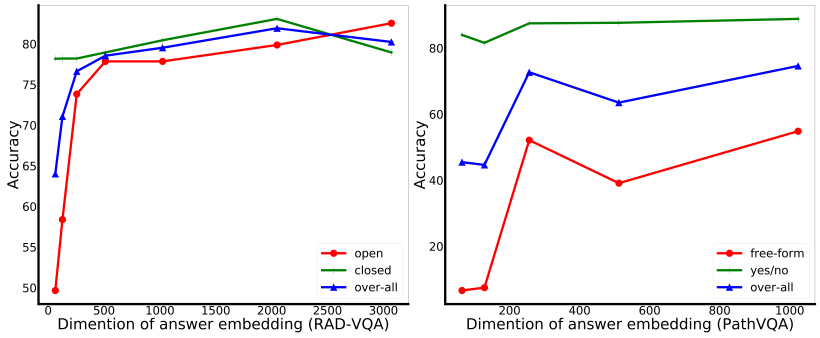


图 3. 关于答案嵌入不同维度的消融实验。

成本，而性能提升则会趋于饱和。作为折中，我们的模型采用维度为 1024 的答案嵌入。

3.4 定性结果

PathVQA 和 VQA-RAD 数据集的示例结果分别见图 4 和图 5。如图所示，对于这些使用 BAN 进行特征融合的 MMQ 失败的示例，我们的 Q2ATransformer（不含解码器）能够通过所提出的 CMAN 融合模块纠正其中的大多数。通过答案查询解码器学习候选答案嵌入与融合后的图像-问题特征之间的交互，还可以进一步提升性能。

3.5 局限性与讨论

如第 2 节所述，我们将每个答案视为可学习的嵌入，并使用所有嵌入作为查询，在我们的解码器中计算注意力图。由于我们计算全局自注意力，当答案类别数量非常庞大时，这可能会增加计算开销。在处理长序列时，自然语言处理领域也遇到过这一问题。已有一些解决方案被提出，例如动态计算稀疏注意力，这可以显著降低计算开销，也是我们未来将探索的方向。

4 结论

本文提出了一种医学视觉问答半开放式框架。该框架通过我们设计的机制，将答案语义信息成功融入答案类别预测过程，并将答案嵌入与融合后的图像-问题特征相关联，从而显著提高了准确率。该框架丰富了现有的医学视觉问答封闭式和开放式框架，并刷新了两个基准测试上的最先进水平，尤其是在开放式问题上。

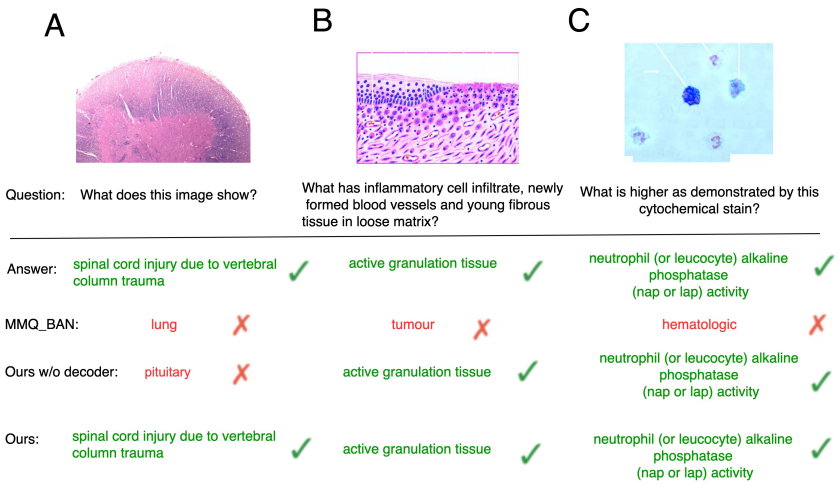
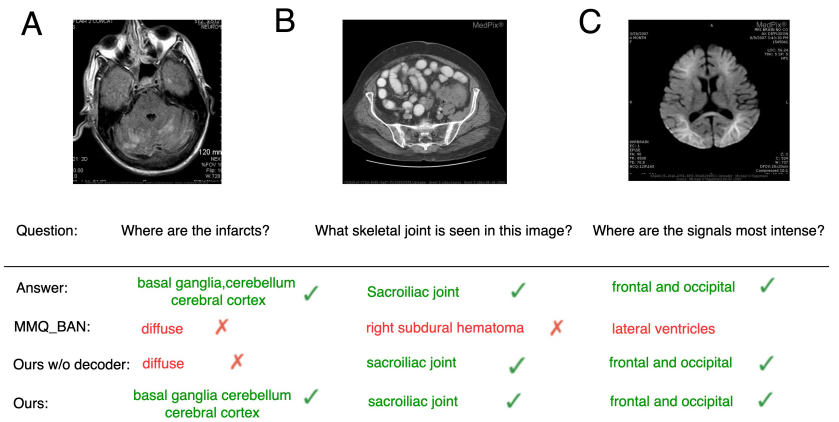


图 4。 PathVQA 数据集中的示例结果。



F图 5. VQA-RAD 数据集的示例结果 set.

参考文献

1 2 2. PEIR 数字图书馆。http://peir.path.uab.edu/library/index.php?/category/。Ambati, R., Dudyala, C.R.：用于 ImageCLEF 2018 医学领域视觉问答的序列到序列模型方法。载于：2018年第15届IEEE 印度理事会国际会议（ INDICON ）。第 1–6 页。IEEE （ 2018 ） 3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.：BERT：用于语言理解的深度双向 Transformer 预训练。arXiv 预印本 arXiv:1810.04805（2018）4. Do, T., Nguyen, B.X., Tjiputra, E., Tran, M., Tran, Q.D., Nguyen, A.：用于医学视觉问答的多元元模型量化。载于：医学图像计算与计算机辅助干预国际会议。第 64–74 页。Springer（ 2021）5. Eslami, S., de Melo, G., Meinel, C.：CLIP 在医学领域的视觉问答中获得的益处是否与其在通用领域中一样多？arXiv 预印本 arXiv:2112.13906（2021）6. Finn, C., Abbeel, P., Levine, S.：用于深度网络快速适应的模型无关元学习。载于：国际机器学习会议。第 1126–1135 页。PMLR（2017）7. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.：PathVQA：用于医学视觉问答的 $30000+$ 问题。arXiv 预印本 arXiv:2003.10286（2020）8. Jiang, H., Misra, I., Rohrbach, M., Learned-Miller, E., Chen, X.：为视觉问答中的网格特征辩护。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第 10267–10276 页（2020）9. Khare, Y., Bagal, V., Mathew, M., Devi, A., Priyakumar, U.D., Jawahar, C.：Mm-bert：用于改进医学视觉问答的多模态BERT预训练。载于：2021年 IEEE 第18届生物医学成像国际研讨会（ISBI）。第 1033–1036 页。IEEE（2021）10. Kim, J.H., Jun, J., Zhang, B.T.：双线性注意力网络。神经信息处理系统进展 **31** （2018）11. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.：关于放射影像的临床生成的视觉问题与答案数据集。Scientific Data**5**(1)，1–10（2018）12. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.：Swin Transformer：使用移位窗口的分层视觉 Transformer。载于：IEEE/CVF国际计算机视觉会议论文集。第 10012–10022 页（2021）13. Nguyen, B.D., Do, T.T., Nguyen, B.X., Do, T., Tjiputra, E., Tran, Q.D.：克服医学视觉问答中的数据限制。载于：医学图像计算与计算机辅助干预国际会议。第 522–530 页。Springer（2019）14. Ren, F., Zhou, Y.：CGMVQA：一种用于医学视觉问答的新型分类生成模型。IEEE Access **8**，50626–50636（2020）15. Wu, C., Liu, J., Wang, X., Li, R.：用于视觉问答的差分网络。载于：AAAI人工智能会议论文集。第 33 卷，第 8997–9004 页（2019）16. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.：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）