



华南理工大学

South China University of Technology

硕士学位论文

基于多编码器混合自注意网络的
医学视觉问答及不确定性研究

作者姓名	韦政松
学科专业	电子信息
指导教师	顾正晖 教授
所在学院	自动化科学与工程学院
论文提交日期	2023 年 6 月

LaTeX template instructions

Medical Visual Question Answering and Uncertainty Analysis Based on
Multi-Encoder Mixture Self-Attention Network.

Candidate: Wei Zhengsong

Supervisor: Prof.Gu Zhenghui

South China University of Technology

Guangzhou, China

分类号：TP273

学校代号：10561

学 号：202021017421

华南理工大学硕士学位论文

基于多编码器混合自注意网络的
医学视觉问答及不确定性研究

作者姓名：韦政松

指导教师姓名、职称：顾正晖 教授

申请学位级别：工学硕士

学科专业名称：电子信息

研究方向：医学视觉问答

论文提交日期： 年 月 日

论文答辩日期： 年 月 日

学位授予单位：华南理工大学

学位授予日期： 年 月 日

答辩委员会成员：

主席：_____

委员：_____

华南理工大学

学位论文原创性声明

本人郑重声明：所呈交的论文是本人在导师的指导下独立进行研究所取得的研究成果。除了文中特别加以标注引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写的成果作品。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律后果由本人承担。

作者签名：

日期： 年 月 日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属华南理工大学。学校有权保存并向国家有关部门或机构送交论文的复印件和电子版，允许学位论文被查阅（除在保密期内的保密论文外）；学校可以公布学位论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存、汇编学位论文。本人电子文档的内容和纸质论文的内容相一致。

本学位论文属于：

☐ 保密（校保密委员会审定涉密学位论文时间：____年__月__日），于____年__月__日解密后适用本授权书。

☐ 不保密，同意在校园网上发布，供校内师生和与学校有共享协议的单位浏览；同意将本人学位论文编入有关数据库进行检索，传播学位论文的全部或部分内容。

（请在以上相应方框内打“√”）

作者签名：

日期：

指导教师签名：

日期：

作者联系电话：

电子邮箱：

联系地址(含邮编)：广东省广州市天河区华南理工大学（五山校区）3号楼

摘 要

视觉问答 (Visual Question Answering, VQA) 指的是用户给定一张图片, 计算机理解并回答与该图片相关的自然语言问题这样一种人机交互方式, 它结合了计算机视觉和自然语言处理等领域的研究, 是一个经典的多模态问题。同时 VQA 依据使用场景可以划分成众多门类的应用, 其中医学视觉问答技术 (Medical Visual Question Answering, Med-VQA) 旨在让计算机从医学影像中获取信息并回答使用者提出的医学问题, 从而为医护人员和患者提供优质便捷高效的医疗咨询服务。然而由于医学数据样本的匮乏、医学图像和文本之间存在巨大的语义鸿沟以及医疗卫生领域的科技应用往往伴随着十分高的风险, Med-VQA 在现实世界中的应用面临着诸多挑战。鉴于现有方案都无法很好地解决这些问题, 对此本文进行了如下研究:

1. 相比于丰富的自然图像问答数据集, Med-VQA 目前存在着重大的数据缺陷, 如无法通过机器生成、具有高昂的人工标注成本、医学图像噪声大和有用信息占比低等问题。同时与 VQA 模型相比, 现有的 Med-VQA 模型在医学图像特征提取以及关联语义建模上仍有较大差距。现有 Med-VQA 方法提取到的图像特征十分有限, 同时也难以和更细粒度的语义特征相关联。针对这些问题, 本文基于混合视觉增强技术提出了一种多编码器混合自注意网络 (MEMSA) 用于 Med-VQA 研究。该网络首先引入了多个编码器来提取不同方面的图像特征, 以实现图像降噪以及提高小样本下的特征丰富度。同时设计了一种跨模态自注意力机制用于特征融合和上下文关联性语义建模, 以获得更细粒、更准确和更全面的语义特征表示。
2. 现有的 Med-VQA 模型采用的是点估计的形式来进行答案预测, 不具备输出答案不确定性的能力。但“过于自信”的回答在医学领域是一个具有高风险性的行为, 为避免医疗事故的发生, 需要采取与之相应规避方法和措施。为此, 本文在 MEMSA 和贝叶斯神经网络的基础上, 提出了一种贝叶斯分类器 (BMLP) 用于输出 Med-VQA 模型在预测时的不确定性。BMLP 使得 MEMSA 网络能够准确地回答医学图像相关问题的同时, 还能输出其答案的不确定性。同时, 本文也对 Med-VQA 中的不确定性估计开展了采样分析实验和拒绝分类实验, 论述了不确定性的由来以及阐释了拒绝分类方法如何用于减小和防范医学问答预测时所出现的不确定性。
3. 在上述模型的基础上, 针对目前 Med-VQA 落地难这一问题, 结合模态自适应、云

计算等技术，设计并实现了一个面向 Web 的在线云服务系统。该系统能够根据用户提供的医学影像和问题提供带不确定性估计的回答，随时随地为用户提供安全，可靠，尽可能准确的 Med-VQA 服务。该系统还具有对用户复杂输入模态的自适应能力，有着良好的鲁棒性和实用性的同时还可以给用户提供更丰富的人机交互体验。可应用于互联网生态以及结合各大医疗系统提供在线医疗服务。

关键词： L^AT_EX；医学视觉问答；多编码器；注意力机制；深度学习；贝叶斯神经网络

Abstract

Visual Question Answering (VQA) refers to the human-computer interaction in which a computer understands and answers natural language questions related to a picture given by the user. At the same time, VQA can be divided into many categories of applications according to the usage scenarios, among which Medical Visual Question Answering (Med-VQA) aims to provide high-quality, convenient and efficient medical consultation services for healthcare professionals and patients by allowing computers to obtain information from medical images and answer medical questions raised by users. However, the real-world application of Med-VQA faces many challenges due to the lack of medical data samples, the huge semantic gap between medical images and text, and the high risks associated with the application of technology in health care. In view of the fact that none of the existing schemes can solve these problems well, the following research is conducted in this paper:

1. Compared with rich natural image question-answering datasets, Med-VQA has significant data defects, such as high annotation costs, large noise in medical images, and low useful information proportion. Existing Med-VQA models still have significant shortcomings in medical image feature extraction and modeling methods for related semantics. The extracted image features are limited and difficult to associate with more fine-grained semantic features. To address these problem, this paper designs a Multi-Encoder Mixed Self-Attention Network (MEMSA) for medical visual question answering research based on mixed visual enhancement technology. This method first introduces multiple encoders to extract different aspects of image features, achieve image denoising, and improve the generalization representation ability under small samples. It also designs a cross-modal self-attention mechanism for feature fusion and context-related semantic modeling to obtain more fine-grained, accurate, and comprehensive semantic feature representations.
2. Existing Med-VQA models use point estimation to make answer predictions and lack the ability to output answer uncertainty. Overconfident answers are a high-risk behavior in the medical field, requiring corresponding avoidance methods and measures. Therefore, based on MEMSA, this paper designs a Bayesian MLP classifier (BMLP) for outputting the prediction uncertainty of MEMSA based on Bayesian neural networks, allow-

ing MEMSA networks to accurately answer medical image-related questions and output answer uncertainty. At the same time, sampling experiments and rejection classification experiments are carried out for uncertainty estimation in medical visual question answering, explaining the source of uncertainty and how rejection classification methods can be used to reduce and prevent uncertainty in medical question answering predictions.

3. Based on the above models, and in response to the current difficulty of Med-VQA landing, this paper designs and implements an online cloud service system that can provide answers with uncertainty estimates based on user-provided medical images and questions anytime, anywhere. This system is designed to be safe, reliable, and provide as accurate medical visual question answering services as possible. The system is also capable of adapting to complex input modes, improving system robustness and practicality, and providing users with a good human-machine interaction experience, enriching human-machine interaction methods. It can be applied to the internet ecology and provide online medical services combined with major medical systems.

Keywords: Med-VQA; Multi-Encoder; Attention Mechanism; Deep Learning; Bayesian Neural Network

目 录

摘 要	I
Abstract	III
插图目录	VIII
表格目录	X
第一章 绪论	1
1.1 研究背景和意义	1
1.2 国内外研究发展以及现状	2
1.2.1 通用领域视觉问答模型及方法	4
1.2.2 医疗领域视觉问答模型及方法	5
1.3 文章主要研究内容及章节安排	6
1.3.1 主要研究内容	6
1.3.2 章节安排	9
第二章 相关工作以及技术	11
2.1 视觉问答技术	11
2.2 医学视觉问答系统	12
2.2.1 医学视觉问答系统的组成	12
2.2.2 医学视觉问答系统的分类	13
2.3 Med-VQA 数据集	14
2.4 VQA 中的编码技术原理	16
2.4.1 图像编码技术	16
2.4.2 文本编码技术	18
2.5 跨模态自注意力机制	19
2.5.1 注意力机制	19
2.5.2 自注意力	19
2.5.3 跨模态注意力	20
2.6 贝叶斯不确定性估计理论	20
2.6.1 贝叶斯定理	20
2.6.2 贝叶斯网络概述	21

2.6.3	变分推断和蒙特卡洛方法	22
第三章	基于多编码器混合自注意网络的医学视觉问答研究	23
3.1	多编码器混合自注意网络	23
3.1.1	MEMSA 网络架构	23
3.1.2	各编码器原理	25
3.1.3	跨模态自注意力机制	28
3.2	模型预测与网络训练	30
3.2.1	模型预测	30
3.2.2	网络训练	31
3.3	实验结果	31
3.3.1	编码器模型实验对比分析	31
3.3.2	注意力模型实验对比分析	33
3.3.3	实验总结	34
3.4	讨论	34
3.4.1	按问题类型划分的准确率比较	35
3.4.2	MEMSA 模型综合评估	35
3.4.3	MEMSA 实例问答效果	38
3.5	本章小结	38
第四章	基于局部贝叶斯神经网络的医学视觉问答及其不确定性研究	39
4.1	贝叶斯神经网络	39
4.1.1	贝叶斯神经网络简介	39
4.1.2	局部 BNN 与全局 BNN	40
4.2	用于视觉问答不确定估计的 BNNs	40
4.2.1	网络搭建	40
4.2.2	先验分布和后验分布选择	41
4.3	网络训练	42
4.3.1	贝叶斯反向传播算法	42
4.3.2	网络预测	45
4.4	实验结果与分析	46
4.4.1	模型性能实验	46

4.4.2	采样-不确定性实验	46
4.4.3	拒绝分类实验	48
4.4.4	带不确定性估计的问答样例	49
4.5	本章小结	51
第五章	模态自适应医学视觉问答系统实现与在线部署	52
5.1	模态自适应系统	52
5.1.1	技术路线	52
5.1.2	设计原理	53
5.1.3	算法评估	54
5.1.4	分析	55
5.2	云端在线系统设计	55
5.2.1	功能性需求分析	55
5.2.2	非功能性需求分析	56
5.2.3	总体架构	58
5.2.4	工作数据流	59
5.2.5	后端 API 服务接口设计	59
5.3	详细设计与实现	60
5.3.1	系统架构	60
5.3.2	系统管理	61
5.3.3	信息处理	61
5.3.4	视觉问答	61
5.3.5	前端界面	62
5.4	系统接口测试	62
5.5	本章小结	63
	总结与展望	65
	参考文献	68

插图目录

图 1-1	Ai 智慧医疗体系	2
图 2-1	视觉问答样例	11
图 2-2	视觉问答系统	12
图 2-3	自注意力机制	20
图 2-4	跨模态注意力机制	21
图 3-1	总体框架	23
图 3-2	多编码融合自注意力网络	24
图 3-3	Glove 词嵌入的降维表示	28
图 3-4	LSTM 网络结构	29
图 3-5	词汇级细粒度融合特征关注图	30
图 3-6	封闭式问答混淆矩阵	36
图 3-7	开放式问答混淆矩阵	36
图 3-8	封闭式问答模型 ROC 曲线	37
图 3-9	MEMSA 与 MEVFBA (MAML+AE-BAN) 的回复比较	38
图 4-1	贝叶斯神经网络	39
图 4-2	局部不确定估计的 BNN 模型结构	41
图 4-3	MEMSA-BMLP 模型	41
图 4-4	Med-RAD 数据集性能	46
图 4-5	SLAKE 数据集性能	46
图 4-6	sample = 10	47
图 4-7	sample = 50	47
图 4-8	sample = 100	47
图 4-9	sample = 500	47
图 4-10	BMPL 采样 10 次时带不确定性问答样例	50
图 5-1	模态自适应 Med-VQA 系统实现框图	52
图 5-2	自适应交互逻辑	53
图 5-3	头部 MRI 放射学影像	54
图 5-4	胸部 X-ray 放射学影像	54

图 5-5 胸部 CT 放射学影像	54
图 5-6 病理学影像	54
图 5-7 需求分析	56
图 5-8 用户用例	57
图 5-9 模型用例	57
图 5-10 总体框架	58
图 5-11 工作数据流图示	59
图 5-12 Demo 输入栏	63
图 5-13 Demo 输出栏	63
图 5-14 Demo 反馈栏	63
图 5-15 Demo 样例栏	63
图 5-16 向 API 发送请求	64
图 5-17 获得 API 返回数据	64

表格目录

表 2-1	现有 Med-VQA 数据集	15
表 3-1	模型超参数设置	32
表 3-2	深度学习环境配置	32
表 3-3	MEMSA 与主流方法等对比	33
表 3-4	不同注意力下的模型性能对比	33
表 3-5	CLOSE-Question Results	35
表 3-6	OPEN-Question Results	36
表 3-7	不同方法的综合性能对比	37
表 4-1	BNN 与全局 BNN 的区别	40
表 4-2	不同采样率下问答结果的不确定性	47
表 4-3	拒绝分类实验	50
表 5-1	不同机器学习算法的性能评估对比	55
表 5-2	后端 API 接口通信格式	60
表 5-3	主要使用的服务框架	60
表 5-4	接口信息处理	61
表 5-5	视觉问答接口	62

第一章 绪论

1.1 研究背景和意义

打造一个既能看得见，又能听会说的 Ai 模型一直是人工智能领域的追求。这样的模型可以与人类进行更自然和高效的交互，提供近乎真实的人机交互体验。在过去几十年的发展中，Ai 技术在视觉、语音、自然语言处理等方面取得了巨大进展，其中视觉问答（VQA）技术的发展尤为迅速。通过 VQA 技术^[1]，计算机可以像人类一样理解场景和图像内容并回答相关问题^[2]，这为人机交互提供了全新的思路 and 方式。

随着人工智能技术和视觉问答技术的不断发展，医学视觉问答（Med-VQA）作为一项可以对医学图像进行分析并回答与之相关的问题的技术也在逐渐进步。近年来，由于 VQA 系统在医疗诊断、治疗和教育上的潜力，Med-VQA 还吸引了大量研究学者的关注^{daibu}。目前 Med-VQA 的一个关键性问题是建立起复杂医学图像和问题之间的联系，这通常需要研究者有着丰富的医学知识和以及对医疗影像数据有着深入的理解。同时 Med-VQA 还需要用到不同的模态信息，如视觉、语音和文本等等。为了解决这些问题，该领域的研究人员运用了各种人工智能技术，比如卷积神经网络（CNN）、递归神经网络（RNN、LSTM）和注意力机制等方法^[3]。同时，学者们还探索了不同模态信息之间的融合方法^[4]，例如图像-文本之间的跨模态语义融合以及从外部引入知识图谱，增加模态信息间的推理关系等。除此之外，如图1-1Med-VQA 有许多潜在的应用，比如为社会构筑完整的智慧医疗体系、协助放射科医生进行诊断、改善医学教育和协助医学研究等。随着深度学习技术在该领域的不断发展，相信在不远的未来，Med-VQA 将会对整个医疗卫生行业产生深远且重要的影响。

最早的视觉问答研究是针对通用场景提出的，但人们逐渐发现，相比于让 Ai 模型去学习通用的“常识”，学习一些专业性更强，更复杂的知识和技能显然更有实用价值和意义。故以 VQA 技术作为基础，各行业衍生了一系列的 VQA 相关应用。根据 QJ WU 等人^[2]的研究，VQA 领域的应用大致可以分成 5 大类：1. 基于自然场景的；2. 基于医学图像的；3. 基于人机交互的；4. 基于图像理解的；5. 基于其他领域的视觉问答应用。如游戏、社交媒体和虚拟现实等。并且，在这些应用场景中，基于医学图像的视觉问答是最具有落地潜力和现实意义方向之一，研究人员正在努力开发一套具有准确，可靠，丰富等特性的 Med-VQA 系统。例如，在这个系统中，用户可以通过在线的方式实现全自动就医，自助挂号，自助诊查，自助问询，Ai 辅助医生完成线上会诊等。这套系统可

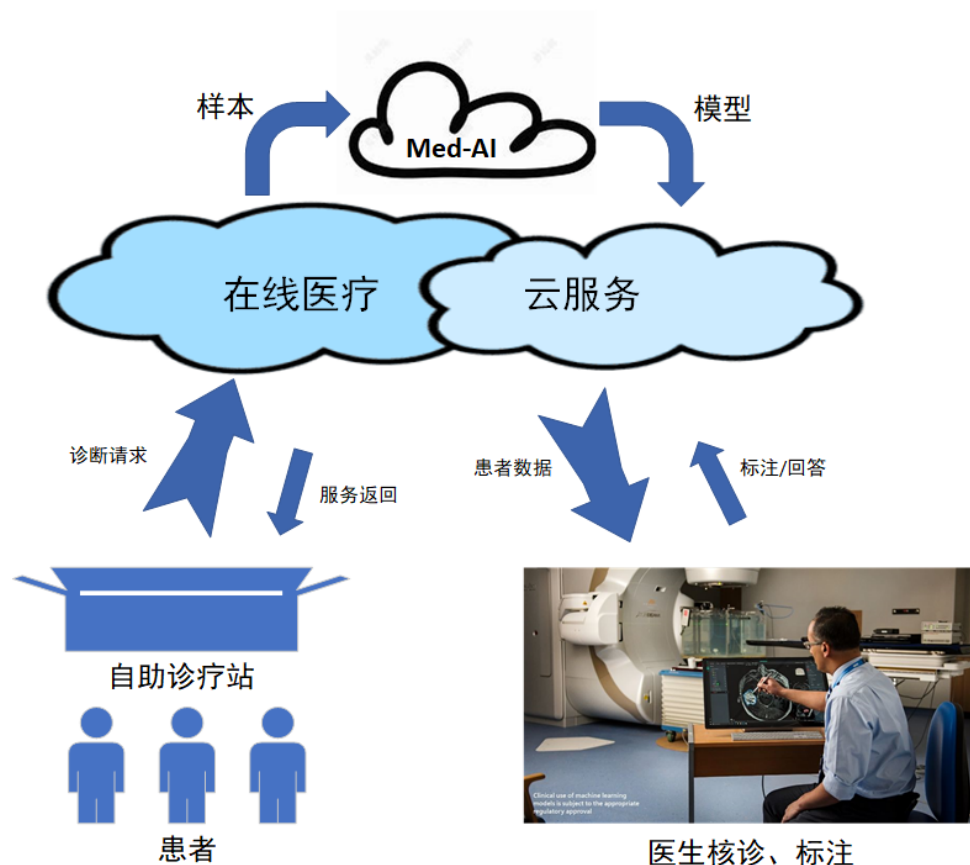


图 1-1 Ai 智慧医疗体系

以极大减轻医生的工作负担，降低误诊率，缩减患者挂号排队问询的流程，提高医院的运行效率。

目前，Med-VQA 技术处在一个蓬勃发展的时期，它涉及计算机科学，生物学，医学，认知科学，心理学等众多领域间的跨学科研究。从问诊到给患者提供更多的就医服务以及创造新应用，这个技术的出现和成熟会给现有医疗体系带来深刻变革。因此可以说，Med-VQA 技术具有非常重要的研究意义和应用价值。

1.2 国内外研究发展以及现状

视觉问答的概念最早出现在 2015 年，由斯坦福大学的 Licheng Yu 等人^[1]提出并构建了一个基于填空的数据集进行实验，成功让模型依据图像信息完成了正确的信息填写。自此以后，视觉问答逐渐成了计算机视觉、自然语言理解以及多模态领域的一个重要研究方向，同时也为传统智能对话和智能问答系统的改进提供了参考和借鉴。在视觉问答研究的早期，受模型和算力的限制，主流研究还是使用的一些非深度学习的传统方法，如 Malinowski and Fritz 等人提出了多元问答方法^[5]，Kafle and Kanan 等人^[6]在 2016

年提出了用于 VQA 的经典贝叶斯框架，通过贝叶斯算法对目标的空间关系建模，计算每个答案概率，也导致该方法比较依赖语义分割结果。

自引入深度学习后，Zhou 等人^[7]在 2015 年后提出了 ibowing 模型，们使用预训练的 GoogLeNet 图像分类模型的层输出来提取图像特征，对答案分类使用 softmax 回归，相比与传统模型有着更好的性能。接着，Ren 等人^[8]提出了一种端到端 QA 模型 Vis+LSTM，模型利用 visual semantic embedding 作为融合机制链接作用与不同模态的 CNN 与 RNN。至此，基本的”双流式”视觉问答网络被确定了下来。2016 年，Shih 等人^[9]提出基于局部注意力的 VQA 模型 WTL。WTL 具体做法是对图像经过边缘检测获得 100 个分区，采用 CNN 对这些分区进行特征提取，将各个分区和问答对的特征向量做内积运算得到 attention 权重系数，最后和文本特征并置加权求和得到加权特征。该特征进一步经过全连接映射到标签结果，通过与标签值的交叉熵损失训练模型。后来，Zhu 等人^[10]更进一步提出循环空间注意力 R-SA，引入 LSTM 网络对问题进行解码并提出循环空间注意力融合不同的模态。同时 Yang 等人^[11]参考人们对事物通常采用分层关注并逐层推理的思想提出堆叠注意力模型 SAN。SAN 在问题特征提取和图像特征提取采用 LSTM，CNN 网络来提取特征，然后用问题特征去给图像添加注意力，用注意力的结果结合问题向量再次去 attention 图像，最后产生预测。这种堆叠注意力分层抽象的思想十分适用于进行语义融合，在 Med-VQA 任务提出后就作为首个基线模型使用^[12]。后来，Kim 等人在 2018 年借鉴这个思想并参考双线性池化机制设计了双线性池化注意力网络 BAN^[13]，BAN 也迅速成为 VQA 领域内最主流的注意力方法之一。

除了融合机制的更新迭代，编码方法以及编码器性能的提升也极大地促进了 Med-VQA 的发展，Binh 等人^[14]为了克服医学图像样本的种种限制，基于模型不可知元学习方法提出 MAML 模型和混合视觉增强方法 MVEF，通过元 MAML 模型和卷积降噪自编码器^[15]的融合编码，提升了图像特征的表征能力，同时也提高了 Med-VQA 模型的问答准确率。Li-Ming Zhan 等人^[16]又针对开放式和封闭式这一问题类型不同，提出了可以区分问题类型 QCR 和 TCR 模块，该模块简化了模型设计，显著改善了模型针对不同问题的回答能力。Haifan Gong 等人^[17]又提出了集成化预训练编码器的思想以及引入跨模态注意力 CMSA，Tuong Do 等人^[18]在元学习 MAML 的基础上，又提出了基于问题增强的医学视觉元学习 MMQ 模型。随着大规模预训练模型的兴起，Sedigheh 等人^[19]将原生 CLIP 在与 PubMedshu 数据库中预训练以获得良好的医学图像文字跨模态表征。

随着近些年 Transformer 模型^[20]的广泛应用，传统 VQA 在向着可进行深度推理，

具有高可解释性的 VQA 模型逐步靠近, 知识图谱和关系推理机制也在逐渐被引入到 Med-VQA 领域^[21], 新的技术和思想正源源不断推动着 VQA 从一个传统的范式向着多元化发展^[4], 医学视觉问答任务也和很多传统一样拥有着顶级科研赛事, 国际顶级医学图像处理和人工智能会议 MICCAI 每年都会举办一场 Med-VQA 科研竞赛^[22], 旨在推动这个领域的发展。由于视觉问答本身也是一个高度复杂的多模态问题, 所以其进展也与多模态领域的发展有着紧密的联系, 特别是多模态信息融合以及相关方法的研究^[23]:

1.2.1 通用领域视觉问答模型及方法

在通用自然图像领域, 由于没有数据限制, VQA 的发展十分迅速, 从最早的非深度学习方法 Multi-World QA^[5]、ATP^[6], 到使用深度学习后但无注意力机制的 Full-CNN^[24]、AYN^[25], 进而到使用注意力机制的 WTL 模型^[26]、SAN、DAN^[27]、BAN, 最后随着 Transformer 架构的提出, NLP 领域有了大一统的模型 Bert^[28], VQA 也使用上了 ViLBERT^[29]、VisualBERT^[30]、ViLT^[31]等先进模型。

(1) 传统非深度学习的 VQA 方法

在 VQA 概念提出的同一年, Malinowski and Fritz 等人提出了多元世界问答方法 Multi-World QA, 他们将基于问题和图像的答案概率进行建模 $P(A = a|Q, W) = \sum_T (P(A = a|T, W)P(T|Q))$, 这里 T 为隐变量, 它对应于从问题中得到语义树。W 是世界, 代表图像。它可以是原始图像或从分割块获得的附加特征。使用确定性评价 (deterministic evaluation) 函数来评估 $P(A|T, W)$ 。使用简单的对数线性模型得到 $P(T|Q)$ 。这个模型也被称为 SWQA。之后, Kafle and Kanan 等人基于贝叶斯理论提出了 ATP, 其主要方法是根据图像 v 和问题 q 计算出答案 a 和答案类型 t 的概率, 再使用语义分隔来识别图像中的对象及其位置, 接着用 ResNet 对这些对象进行处理, 并使用跳级思考向量 (skip-thought vectors) 来处理文本, 然后利用贝叶斯算法对目标的空间关系进行建模, 计算每个答案的概率。ATP 也是较早期的 VQA 解决方案, 它的有效性比不上某些简单的基线模型, 其主要原因可能是 ATP 十分依赖于语义分割的结果。虽然这些非深度学习的模型或方法没有深度学习模型那么强大, 但在一些特定场景下仍然具有一定的优势和应用价值。

(2) 基于深度学习的 VQA 方法

基于深度学习的方法有注意力和非注意力两大类, 如本节引言提到的 ibowing、Vis+LSTM 模型是无注意力模型, WTL、R-SA、SAN、BAN 等都是基于注意力

的模型。与传统非深度学习方法不同的是，使用深度学习的 VQA 模型通常会使用卷积神经网络（CNN）提取图像特征和循环神经网络（RNN）或者变种（如 LSTM、GRU 等）提取文本特征。而非深度学习方法可能需要手工设计特征提取器。深度学习方法虽然提高了模型的复杂度，需要更多的训练数据和计算资源，但它的优点是可以更好地利用图像和文本的丰富特征，具有更高的准确率和泛化能力。以 VGG+LSTM 为例，该模型使用 VGG 卷积神经网络提取图像特征，使用 LSTM 循环神经网络对问题和图像特征进行编码，能够在一定程度上捕捉问题和图像之间的关系，学习到更复杂的问题和图像特征之间的关联。

1.2.2 医疗领域视觉问答模型及方法

除了本节引言所提到的主流基于注意力的 Med-VQA 方法，还有基于非注意力的 MCB-MIL、HMN-MedVQA 等方法。

(1) 基于非注意力的 Med-VQA 方法

这类方法多是简单的特征提取 + 机器学习模型，例如 MCB-MIL 使用多模态卷积块来融合图像和文本信息，然后采用多示例学习 (MIL)^[32] 框架进行分类，该方法的主要优点是能够处理多种医学图像数据，并且在准确性和效率方面都表现出色。HMN-MedVQA 方法使用一种称为医学知识网络（HMN）的结构来融合问题和答案的信息，并使用双向长短时记忆网络（Bi-LSTM）进行特征提取。然后，将这些特征传递给一个分类器进行分类。该方法的主要优点是可以利用医学知识库来提高准确性。相比于注意力的方法，这些放在模型结构上更为简洁，训练更为简单，同时具有较强的可解释性。

(2) 基于注意力的 Med-VQA 方法

基于注意力的方法通过计算关注度权重可以对图像和问题文本之间的语义特征联系建立起深度的关联。且由于医学图像中 useful 信息的占比较少，采用注意力机制的效果尤为明显。如堆叠注意力机制 SAN 可以层层递进地抽取图像-文本之间的关系语义，双线性池化注意力机制 BAN 通过线性池化操作来计算图像特征和文本特征之间的交互^[13]。同时使用多个并行的注意力分支，每个分支对不同类型的特征进行关注，以充分利用多种特征之间的交互。这使得 BAN 能够更好地理解问题和图像之间的关系，从而提高了模型问答的性能。

1.3 文章主要研究内容及章节安排

1.3.1 主要研究内容

Med-VQA 作为 VQA 中一个新兴的研究方向,近年来吸引了越来越多的研究者的关注。目前,Med-VQA 领域的研究主要集中在两个方面:首先是基于现有的 VQA 方法,对这些方法进行改进和调整,使其合适用于医学图像特征。其次是针对 Med-VQA 图像和文本之间的特殊性,提出新的算法和模型。研究者们在这些方面都做出了许多贡献,极大推动了这一领域的发展,例如采用多模态信息融合方法提升模型性能、使用图像标注和文本标注相结合的方式增强数据集的丰富程度等等。

然而,Med-VQA 领域仍然存在许多待解决的问题,一是在视觉问答领域,不同模态的数据往往是不平衡的,图像数据的量级往往要远低于文本的数据,这就可能导致某些模态对于融合的结果影响较少,进而无法发挥出多模态的作用,尤其是在医学图像中,有用的信息往往较少且相对集中,关键病变往往只出现在某些区域且不易被察觉,如何从数据中挖掘特征,如何有效地提取和利用医学图像的特征信息往往决定着一个模型的质量;同时,不同模态间还极容易出现模态缺失以及数据存在缺失值的现象,如图像中的遮挡或文本中的缺失信息。这可能会导致模型无法利用完整的多模态数据进行训练和预测,从而影响融合效果。二是多模态任务中,不同模态的信息和特征差距巨大,不同的模态的数据可能具有不同的分布和特征,这种差异性可能导致在融合过程中出现不匹配的情况,很多时候融合这些信息后并不能给模型性能带来提升而是下降,需要合理且具有一定可解释性的融合方法^[33];同时,多模态数据融合会导致数据维度的增加,这可能会带来维度灾难的问题。维度灾难会导致模型的计算复杂度增加、训练时间增加以及出现过拟合等问题。另外,多模态模型往往具有着极高的复杂度,通常需要大量的时间和资源对其进行优化三是实际的医学视觉问答系统往往都是在具有极高信任风险的医疗环境进行的,虽然 VQA 系统一般只是扮演着辅助的角色,但如果出现十分明显的错误以及给出不可靠的回答,都会影响医生和患者对系统的信任。所以如何在保证模型提高准确率的同时,降低模型在泛化时的错误率,规避具有不确定性的回答以及可能产生的误诊风险显得尤为重要。也因此,对于用于医学领域的视觉问答模型,不应该仅仅关心预测结果的精确度,更需要关注模型对其预测有多少把握。

混合视觉特征增强(Mixture of Enhanced Visual Features,MEVF)是一种用于图像特征提取的多模态特征融合的方法,通过引入多个编码器对某个模态进行特征提取,并在

特征级别上进行融合，从而提高图像特征的表示质量进而提高下游任务或模型的性能。具体来说，就是将图像以及其特征输入到不同的编码器中进行提取，然后将编码器的输出特征进行融合，最后再将融合后的特征传入后续的模型进行分类或者回归等任务。对于医学视觉问答来说，MVEF 极大增强了医学图像的代表能力，减少对同类样本的需求，并且在保留鸽子模态特征的同时，进一步挖掘其中的关联和交互信息，提高了模型的性能和稳定性。此外，MEVF 还可以提高模型的鲁棒性，减少因某个模态数据缺失而导致的性能下降。

贝叶斯神经网络 (Bayesian neural network, BNN) 是一种基于贝叶斯推断的神经网络模型。传统神经网络采用的是点估计的形式，也就是网络中的权重是一个固定具体的值，同时采用梯度下降等优化算法来调整网络的参数，目的是最小化损失函数。而贝叶斯神经网络则将参数看成是随机变量并在这些参数上设置概率分布，这些分布可以捕获网络的参数不确定性。通过对他们的集成以及贝叶斯推断来计算他们的后验分布，这样可以获得关于模型预测的不确定性。并且 BNN 的预测相对传统神经网络来说也更加鲁棒，因为它对所有可能的权值进行平均而非选择单点估计，避免了模型对于自己预测的“过度自信”。另外，BNN 通过在权值上设置概率分布 (先验) 可以将先验信息包含到权值中，从而实现对权值的正则化。因此，BNN 对小样本数据训练场景下的过拟合问题具有鲁棒性。这个特点对于十分缺乏训练数据的医学视觉问答领域来说尤为有益。

模态自适应交互系统 (Modality-adaptive interactive system) 是一种能够感知用户当前交互行为，自适应地调整交互方式和策略的系统。相比于传统只局限于使用单一模态进行交互的系统，该系统可以在多种交互模态之间进行切换以满足用户不同的交互需求，例如语音、手势、触摸、眼动等。此外，这种系统还可以根据用户的交互历史和上下文信息，预测用户可能的下一步交互行为，从而提供更高效、更便捷的交互方式。模态自适应交互系统已经广泛应用于人机交互、虚拟现实、增强现实等领域，为用户带来了更加智能、自然、舒适的交互体验。在医学视觉问答系统中加入模态自适应交互设计可以将针对不同问题训练的问答模型融合成一个具有综合能力的医学问答大模型，从而提高了模型的适用性，灵活性和可靠性；提高了交互的自然性和流畅性，提高了用户的人机交互体验。

针对以上特点，本文提出多编码器融合自注意网络用于实现高性能的医学视觉问答系统，并引入贝叶斯神经网络对模型预测及结果进行不确定性估计，最后根据模型设计和搭建了一个在线模态自适应交互系统。本文的研究内容主要分为以下三个方面：

- (1) 通过将多个编码器混合使用的方式，同时基于自注意力机制提出一种用于医学视觉问答的特征提取和融合网络（**Multi-Encoder Mixture Self-Attention Network, MEMSA**）。在特征提取阶段，相较于传统的单一编码网络，MEMSA 采用多种编码器进行特征提取，不同编码器的提取结果可以视作不同的模态表征。实验结果表明，采用同种注意力机制融合的情况下，多编码表征网络也要比单一编码表征网络具有更强的表示能力，从而提升模型的整体性能。在特征融合阶段，多编码混合编码器加跨模态自注意力机制的方法对模型的性能提升起到了较为明显的效果。证明自注意力机制可以有效筛选多编码网络这一具有复杂多模态信息的特征，并且对其中的语义关联进行更细粒地建模，从而提升开放式问答地准确率。同时，在更细致的依据问答内容进行划分的效果上看，MEMSA 在绝大多数类别都有着比传统网络更优异的性能，模型的随机问答也更准确。
- (2) 通过在多层感知机（**MLP**）^[34]的权值中引入不确定性，提出了一种基于贝叶斯神经网络的分类器（**BMLP**），用于输出模型的不确定性估计。在 BMLP 中，权值由传统的点估计形式替换成分布形式，以实现权值不确定性的建模。在预测时，使用蒙特卡洛采样方法在权值分布上进行多次采样，以生成多个子网络，这些子网络的预测差异可以解释为预测的不确定性。同时，通过集成这些子网络，可以提高网络的泛化能力。实验结果表明，在经过适当的拒绝分类后，BMLP 可以获得比传统 MLP 更高的问答准确率。不确定性有许多实际意义和价值，它可以减少网络错误分类的情况出现，提高系统的可靠性和安全性，并增加模型的鲁棒性和防止过拟合能力。我们还进行了采样-不确定性实验和拒绝分类实验，采样-不确定性实验验证了随着采样次数的增加，模型的不确定性或预测分布估计会趋近于一个真实分布，说明这是一个蒙特卡罗近似的结果。而拒绝分类实验则验证了高不确定性预测的数据样本往往是模型在点估计时容易出现错误分类的样本，不确定性与模型性能之间存在直接联系。
- (3) 为了解决在面对不同且未知的输入的复杂场景下，系统无法自发选择有效的响应形式的问题，提出了一种可以进行模态自适应的交互系统设计方法。基于闭环反馈的系统设计思想，在视觉问答交互系统中增加多交互模型控制模块，该模块由经典的机器学习分类算法进行设计实现，由用户输入作为输入，以系统反馈作为标签，两者组合成为训练样本训练该分类网络，从而不断地调整出更准确，不确定性更小的交互方案来适应不同的模态输出。同时，为了便于用户

体验和进行数据采集，还基于云服务、云计算技术为模型设计搭建了一个在线系统，该系统可以实现让用户随时随地访问医学视觉问答服务，从而减轻目前医疗卫生系统的运行负担，减轻了医疗患者的就诊压力和医生的工作压力，具有一定的实际意义。

1.3.2 章节安排

对于上述研究内容，本文的章节安排如下：

第一章，绪论。首先介绍了 Med-VQA 技术的研究背景和意义，其次概述了医学视觉问答系统的组成和分类以及介绍了目前的研究现状，最后介绍本文的大概研究内容及章节安排。

第二章，主要介绍了使用到的相关理论与技术。主要介绍了本文研究中涉及到的关键理论与技术，首先从总体上简单概述了视觉问答技术以及其技术构成；接着介绍了具体的医学视觉问答系统的组成、种类和分类；接着介绍了用于实现这个系统的核心部分编码器、各种编码器的实现原理和方法，然后介绍了用于搭建跨模态自注意力的基本注意力形式，自注意力机制和跨模态自注意力机制，最后介绍了贝叶斯定理以及一部分贝叶斯不确定性估计理论的内容。

第三章，主要介绍用于医学视觉问答的多编码器混合自注意力模型。首先介绍了问答网络和问答模型的总体架构以及如何整合各个子系统实现总体的视觉问答功能；接着详细介绍了用于特征提取的多编码器混合网络；然后也同样介绍用于特征融合的跨模态自注意力机制和用于答案预测的多层感知机模型；最后经过训练在两个数据集上验证了该网络的问答性能。

第四章，主要介绍了基于贝叶斯神经网络和贝叶斯不确定性估计的医学视觉问答研究。首先介绍了贝叶斯神经网络的结构组成以及和点估计网络之间的差别和练习，以及介绍了局部使用贝叶斯神经网络和全局使用贝叶斯神经网络的差别；接着介绍了用于预测视觉问答不确定的贝叶斯神经网络，并涉及到了先验后验分布的选择和网络训练；然后通过实验验证了模型的问答性能；最后通过采样实验和拒绝分类实验验证了不确定性估计对传统问答模型的作用和实际意义。

第五章，主要介绍了一个具备模态自适应能力的医学视觉问答系统的实现方法以及这个系统的在线部署过程。首先介绍了模态自适应系统的技术路线以及设计方法和设计原理；接着介绍了在线云服务系统的设计思路以及详细介绍了系统的实现细节和交互方法；最后对该系统的所有接口进行封装和测试，并举例展示系统的问答界面以及

问答效果。

总结与展望：对本文研究的内容进行总结，并展望未来该技术的研究方向。

第二章 相关工作以及技术

本章介绍本文研究中涉及到的数据集和关键理论与技术。数据集主要包括 Med-RAD 和 SLAKE 两大 Med-VQA 数据集；关键理论与技术主要包括视觉问答技术、视觉问答系统、VQA 中的编码技术原理，跨模态自注意力机制以及贝叶斯不确定性估计理论。

2.1 视觉问答技术

视觉问答^[35]（Visual Question Answering, VQA）是一种结合了计算机视觉和自然语言处理技术的交叉领域任务，其目标是让机器理解人类提出的关于图像内容的自然语言问题，并在语言层面上回答这些问题。如2-1所示，与传统的图像识别任务相比，VQA 需要综合考虑图像和问题两方面的信息，同时进行视觉推理和自然语言理解。



Q: What color are her eyes?
A: Brown
Q: What is the mustache made of
A: Banana



Q: How many buildings are there ?
A: 2
Q: Which side of the building is the road on?
A: Right

图 2-1 视觉问答样例

VQA 技术的基本流程包括图像特征提取、问题特征提取、多模态融合和答案预测等步骤。其中，图像特征提取可以使用传统的卷积神经网络（CNN）模型或预训练的图像识别模型（如 ResNet、Inception 等）进行。问题特征提取可以使用自然语言处理技术，例如词嵌入（word embedding）和循环神经网络（RNN）等。多模态融合可以使用传统的融合方法，例如特征串联、特征相加等，也可以使用基于注意力机制的融合方法，例如多头注意力、视觉注意力、文本注意力等。答案预测可以使用分类模型获得较为准确的结果，也可以使用生成模型获得更为灵活的回答。

2.2 医学视觉问答系统

2.2.1 医学视觉问答系统的组成

如图5-7一个完整的视觉问答系统通常由七个主要组件组成：数据集、图像处理模块、自然语言处理模块、多模态融合模块、答案生成模块、前端应用以及后端服务器。数据集是系统的基础，其质量决定了模型性能的好坏；为了更好的迭代模型以及提高系统的交互体验，一般还会给系统加入信息反馈机制，问答者可以对系统的回答做出评价以及调整意见，模型在迭代时会根据反馈对回答机制和内容进行适当的调整，从而提高系统的整体性能。

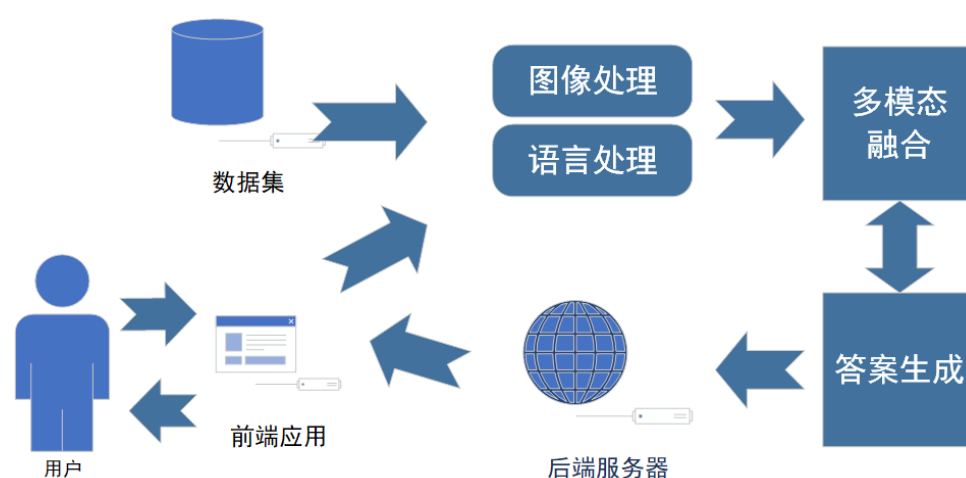


图 2-2 视觉问答系统

图像处理模块负责从输入的图像中提取有用的视觉特征；自然语言处理模块负责处理用户的自然语言输入，如问题或指令；多模态融合模块将图像特征向量和问题向量融合起来，以生成一个综合的向量表述；答案生成模块对融合后的向量进行预测，并生成一个符合语义和语法的答案。前端应用为用户提供了一个界面来输入问题，并展示系统回答的答案。通常使用 Web 或移动应用程序来实现前端应用程序。后端服务器：后端服务器是视觉问答系统的核心部分，它处理来自前端应用程序的请求，并将其传递给模型进行预测。后端服务接口通常使用 API 来实现。此外，一些系统还包括预处理模块，如图像分割和对象检测，这些组件的结合形成了一个完整的视觉问答服务系统，可以用于各种应用，如智能客服、医疗诊断和自动驾驶。

2.2.2 医学视觉问答系统的分类

- (1) 根据应用领域的不同，由于医学图像种类、数量复杂繁多，可以按其应用分为多个不同的领域，如放射学、病理学、眼科等不同领域的系统。在放射学领域，医学视觉问答系统可以用于解读医学影像，如 CT、MRI 等，帮助医生诊断疾病。在病理学领域，医学视觉问答系统可以用于分析组织切片图像，帮助医生诊断肿瘤和其他疾病。在眼科领域，医学视觉问答系统可以用于识别和诊断眼部疾病，如白内障、青光眼等。不同领域的医学视觉问答系统所需的数据、特征提取方法、模型架构等都可能存在差异，因此需要专门针对不同领域进行研究和开发。
- (2) 根据问答类型的不同，可以将医学视觉问答系统分为封闭式问答和开放式问答系统，封闭式问答是指问题的答案是事先给定的，系统只需要从给定的答案集合中选择正确的答案进行回答。这种问答方式的优点是简单高效，但是答案集合通常需要经过手动标注，且无法回答没有出现在答案集合中的问题。开放式问答则是指问题的答案不是事先给定的，系统需要从自然语言文本中寻找答案。这种问答方式的优点是可以回答更广泛的问题，但是也更加复杂和困难。开放式问答通常需要使用自然语言处理技术来理解问题和文本，以及对答案进行生成或者抽取。
- (3) 根据系统实现的不同，可以划分为基于传统机器学习方法的系统和基于深度学习方法的系统。传统机器学习方法的系统通常包括以下步骤：首先，对图像和问题进行特征提取；然后，将两者的特征向量融合在一起；最后，使用机器学习模型（如支持向量机、随机森林等）进行分类或回归预测。这种方法的优点是较为简单，容易理解和解释，但其效果受限于手动设计的特征和模型。基于深度学习方法的系统则利用深度神经网络自动地学习图像和问题的特征表示，并将二者进行融合。常用的深度学习模型包括卷积神经网络（Convolutional Neural Networks, CNN）、循环神经网络（Recurrent Neural Networks, RNN）和注意力机制（Attention Mechanism）等。深度学习方法可以更好地利用数据的内在规律，提高模型的预测性能，但需要更多的计算资源和数据量，并且模型的可解释性较差。
- (4) 根据系统目的的不同，可以划分为辅助医生诊断的系统、辅助患者自我诊断诊断的系统和用于医学教育等不同目的的系统。辅助医生的系统主要是为了协助

医生对医学图像进行分析和诊断。这类系统需要针对医生所需要的信息提供准确的答案，帮助医生做出正确的诊断和治疗决策。例如，一些放射学视觉问答系统可以帮助放射科医生快速准确地分析和诊断 CT、MRI 等医学图像。帮助病人的系统主要是为了帮助病人在家中自我进行初步的诊断和监测。这类系统需要具有简单易用、普及率高的特点，让病人能够方便地获取自己的健康信息。例如，一些眼科视觉问答系统可以帮助病人自行检查眼部问题并获取相应的诊断和治疗建议。用于医学教育的系统主要目的是帮助医学教育工作者、学生和其他研究人员更好地学习和理解医学知识。这类系统需要提供丰富、准确的医学知识和信息，并且要具有易用性和可扩展性，以满足不同的学习需求。例如，一些病理学视觉问答系统可以帮助医学学生更好地理解组织病理学的知识，同时也可以为病理医生提供更快速、准确的诊断和治疗决策支持。

2.3 Med-VQA 数据集

医学视觉问答是一个新兴领域，且由于标准成本较高，与通用领域相比，现拥有的公开数据集十分稀少。其中的主要代表是 VQA-Med、VQA-RAD、PATH、SLAKE 这四个数据集。VQA-Med-2018^[36]是领域内第一个公开可用的数据集，也是 Med-VQA 领域标准竞赛数据集。由于仅仅用于竞赛的原因，目前 VQA-Med 的缺陷是 QA 对是通过半自动化的方式生成的：由基于规则的问题生成系统（QG）通过句子简化、答案短语识别、问题生成和候选问题排序自动生成可能的问答对。然后由两名专家进行人工注释。

同年，完全由专业临床医生注释和标注的 VQA-RAD^[37]数据集被提出，是一个仅包含放射医学影像和问答对构成的数据集，同时也是一个平衡的数据集，包含 MedPix 中头部、胸部和腹部的样本。为了在现实场景中调查问题，作者将图像呈现给临床医生收集非引导问题。临床医生被要求在自由结构和模板结构中提出问题。随后，对 QA 进行人工验证和分类，以分析临床重点。答案类型要么是封闭的，要么是开放的。尽管没有大量的数据，但 VQA-RAD 数据集已经获得了医疗 VQA 系统作为 AI 放射科医生应该能够回答的基本信息。SLAKE^[38]是参考 Med-RAD 所做的数据集，该数据集是第一个具有语义标签和结构化医学知识库的数据集，同时也是目前医学视觉问答中严谨由医生做出标注的唯二放射图像数据集。

第一个病理学视觉知识问答数据集 PathVQA^[39]提出，制作初衷是探索是否有可能培养出通过医学委员会认证考试的 AI 病理学模型，配图文字的图片是从数字资源 (电

子教科书和在线图书馆)中提取的。作者开发了一个半自动的管道,将字幕转换为 QA 对,并手动检查和修改生成的 QA 对。问题可以分为七类:what, where, when, whose, how, how much/how many, and yes/no。开放性问题占有所有问题的 50.2%。对于“是/否”问题,答案是 8145 个“是”和 8189 个“否”。这些问题是根据美国病理学委员会 (ABP) 的病理学家认证考试设计的。因此,对决策支持中的“AI 病理学家”进行验证是一门考试。PathVQA 数据集表明,医疗 VQA 可以应用于各种场景。

各数据集的数据量如表2-1所示。同时为了保证模型具备实用性,本文主要采用 Med-RAD、Slake 两个经过临床医生严谨标注的数据集对模型进行训练和测试。

表 2-1 现有 Med-VQA 数据集

Dataset	Images	QA pairs	Q type	Field
VQA2.0	204K	614 K	close&open	通用领域 (对比)
VQA-Med-2018	2,866	6,413	close&open	竞赛数据集 v0
VQA-RAD	315	3,515	close&open	放射学领域
VQA-Med-2019	4,200	15,292	close&open	竞赛数据集 v1
RadVisDial	91,060	455,300	close	放射学领域
PathVQA	4,998	32,799	close&open	病理学数据集
VQA-Med-2020	5,000	5,000	close&open	竞赛数据集 v3
SLAKE	642	14 K	close&open	放射学领域
VQA-Med-2021	5,000	5,000	close&open	竞赛数据集 v4

- (1) 如上表所示, Med-RAD 数据集有 315 张图片, 3515 对问答对, 是一个包含 X-ray、CT 和 MRI 三类图片以及 14 类不同问题的放射医疗影像数据集, 主要数据来源于美国国家在线医疗图像数据库 MedPix。Med-RAD 同时也是 Med-VQA 领域中最轻量化的数据集, 十分适合开展小样本数据下的训练和研究。本文选择 Med-RAD 作为主要的训练、测试和验证数据集, 不但可以检验模型的性能, 还可以测试 MEMSA 模型在小样本下的拟合能力、健壮性以及泛化能力。
- (2) SLAKE 数据集是近两年最新推出的新数据集, 其制作方式参考了 Med-RAD, 不同的是, SLAKE 的数据的样本量上有了明显的提升。SLAKE 有 642 张医学影像图片, 14K 问答对, 样本总量约为 Med-RAD 的 8 倍, 同样是一个放射学影像数据集, 相比于 Med-RAD, SLAKE 是一个具有语义标签和结构化医学知识

库的综合数据集。图像的语义标签为可视对象提供掩码（分割）和边界框（检测），并以知识图谱的形式提供了医学知识库。本文选择具有同样图像样本以及问答样本类型的 SLAKE 数据集作为 Med-RAD 的对比测试数据集，测试模型在数据量增大近一个量级的数据集上的综合表现。

2.4 VQA 中的编码技术原理

2.4.1 图像编码技术

在计算机视觉等相关领域，图片作为最重要的信息来源，决定了提取图像特征是一个十分重要且关键的环节，高效的特征提取方法可以充分挖掘图像潜在的信息，显著地加强图像的表征能力，有利于后续的模型进行训练以及推理，同时也可以防止特征冗余并且提高整个特征处理过程的效率。

现有方法中，主要的图像特征提取方法大致分为两类，一类是基于卷积神经网络（CNN）或者注意力机制 CBAM、ViT（Vision Transformer）等深度神经网络模型去处理图像，通常是采用预训练好的方式去处理目标图像，并以最后一层的输出结果作为对整个图像编码输出。另一种是基于特征金字塔的特征提取方法，将图像按照不同的区域或者按不同的尺度进行处理，提取图像中的对象特征：首先同样用网络提取图片网格级或者像素级特征，然后识别出图像中的显著对象，之后将多个具有显著对象的区域特征作为提取的图像特征表示。

在医学图像问答（Med-VQA）领域，受到医学问答数据样本十分匮乏的影响，Nguyen,Binh 等人提出了一种启发式元学习（Model-Agnostic Meta-Learning, MAML）^[14]方法，用于在缺乏大量标记数据的情况下快速适应各种下游任务，通过先学习如何快速适应不同的任务，然后再利用这些知识来更好地适应新任务。该方法主要有以下三个步骤：1. 首先，通过在一个大型的数据集上进行预训练，学习如何快速适应各种任务。2. 接着，在每个新的任务上，利用少量的标记数据进行快速微调。3. 最后，将微调后的模型应用于该任务，以获得更好的性能。相比于传统的方法，模型启发元学习方法具有在相似特征的任务之间共享知识的能力，从而可以更好利用有限的数据库。

同时，为了缓解医学图片中的噪声的影响，网络中加入了卷积去噪自编码器（Convolutional Denoising Auto-Encoder, CDAE）^[15]作为一种辅助的图像编码方式，具体而言，自编码器通过将输入数据压缩成一种低维特征表示，然后将其解压缩回原始输入数据来学习特征表示。这种无监督学习方法可以帮助模型从数据中提取重要的特征，并

且对于缺少标记数据的 Med-VQA 任务特别有用。同时，自编码器还尝试用于与其他深度学习方法结合使用，例如卷积神经网络（CNN）或者循环神经网络（RNN），以实现更好的特征提取和融合效果。此外，自编码器还可以作为一种无监督的预训练方法，为其他监督学习任务提供初始化参数或特征表示。

- (1) MAML 模型：模型无关元学习是一种元学习方法，用于在训练模型时获得一个较好的可学习的初始化参数以快速适应新的任务。具体来说，MAML 首先在元学习环境中训练一个通用模型，然后在每个小任务中使用少量的训练数据来更新模型的参数，从而使得其能够在该任务上取得更好的性能。在 Med-VQA 任务中，MAML 模型的训练包含以下几个步骤：1. 初始化模型参数；2. 选择小任务；3. 内部更新；4. 外部更新；5 重复步骤 2-4，直到模型收敛或者达到迭代次数。MAML 方法的优势在于，它可以训练一个通用模型，使其能够在不同的任务中进行快速适应和学习，从而提高模型的泛化能力和鲁棒性。在 Med-VQA 中，MAML 可以用于训练一个能够快速适应不同医学图像问答任务的模型，从而提高模型的性能和效率。
- (2) CDAE 模型：降噪自编码器是一种无监督学习的神经网络模型，它的主要目标是学习数据的低维表达形式。与传统的自编码器不同，降噪自编码器在训练过程中通过对输入数据加入噪声来增强其鲁棒性。在训练阶段，DAE 首先将输入数据添加噪声例如高斯噪声、脏数据等，得到噪声数据。接着，模型将噪声数据作为输入，并讲其映射到一个低维度的隐蔽表示，这个过程称为编码。然后，模型将隐藏表示映射回重构数据，这个过程称为解码。最终，模型通过计算输入局和重构数据之间的差异来衡量其性能，并使用反向传播算法来更新模型参数，从而使其能够学习到输入数据的低维表示。
- (3) 基于对比学习预训练的医学图像编码器：随着 OpenAI 在 2021 年发布对比学习预训练 CLIP^[40]（Contrastive Language-Image Pre-Training）方法，越来越多的深度学习任务采用了预训练 + 微调这一模式，既在上游通过同类型数据对初始模型进行预训练，然后再针对下游任务特性进行微调的过程。CLIP 常用于将自然语言和视觉信息融合在一起，实现对图像和文本的联合理解，其核心思想使用对比学习来训练一个大型的神经网络，使其能够对图像和文本进行联合编码。

2.4.2 文本编码技术

词向量编码技术又称为词嵌入技术，是将自然语言中的单词映射到向量空间中的过程，目的是将单词转换为机器可读的形式，以便于在自然语言处理任务中使用。词嵌入的主要目的是将单词的语义信息编码到向量表示中，以便于在机器学习任务中进行处理。在词向量编码中，向量的每个维度通常代表着不同的语义信息，如词性、情感、主题等。常见的词嵌入方法有 Word2Vec、GloVe、BERT 等，在自然语言处理领域、文本的编码以及向语义空间的转换一直是学者们重点关注的话题。

(1) Word2Vec

由 Google 在 2013 年提出^[41-42]，主要有两种模型：连续词袋模型（Continuous Bag of Words, CBOW）和 Skip-Gram 模型。这两种模型都是基于神经网络的训练方法，通过将单词表示为向量，将单词的语义信息转换为连续的向量空间。Word2Vec 的词向量在自然语言处理领域中非常流行。

(2) GloVe:

GloVe（Global Vectors for Word Representation）是一种基于全局词汇统计信息的词嵌入方法^[43]，由斯坦福大学的研究者开发。它的目标是学习一个向量空间，使得在该空间中的向量可以很好地表示不同单词之间的语义关系。与 Word2Vec 等方法不同，GloVe 使用了全局的共现矩阵来学习词嵌入，而不是像 Word2Vec 那样基于局部上下文信息。GloVe 的训练过程通过最小化一个特定的损失函数来进行，该损失函数旨在最大化两个单词之间的共现概率和它们之间的向量差异。GloVe 的优点是可以在较小的数据集上训练得到良好的词向量表示，并且可以通过简单的线性变换进行求和、平均等操作。

(3) Bert 编码

Bert 作为一个强大的基于 Transformer 的预训练语言模型^[28]，在单独作为编码器使用时有几个独有的优势：1. 双向性，Bert 采用了双向 Transformer 来建模输入句子的上下文信息，使得模型可以从左向右和从右向左两个方向去理解句子，从而更好地捕捉句子的上下文信息。2. 预训练和微调，Bert 在大规模语料上进行预训练，学习得到了通用的语言表示。在实际的应用中，可以通过微调将 Bert 模型应用到特定的自然语言处理任务中这样可以在相对较少的标注数据上获得良好的效果。此外，Bert 还具有多层次抽象的能力以及针对不同的下游任务有着很强的适应能力。

(4) 对比学习预训练的医学文本编码器 PubMedCLIP

随着 OpenAI 在 2021 年发布对比学习预训练 CLIP 方法，越来越多的任务采用了预训练 + 微调这一模式，既在上游通过同类型数据对初始模型进行预训练，然后再针对下游任务特性进行微调的过程。CLIP 常用于将自然语言和视觉信息融合在一起，实现对图像和文本的联合理解，其核心思想使用对比学习来训练一个大型的神经网络，使其能够对图像和文本进行联合编码。

2.5 跨模态自注意力机制

2.5.1 注意力机制

注意力是深度学习中的一个强大机制，它允许模型关注输入数据的特定部分，同时过滤掉不相关的信息。注意力最早是在自然语言处理（NLP）任务的背景下引入的，如机器翻译，它被用来对齐输入和输出序列。从本质上讲，注意力使神经网络能够有选择地处理输入数据的不同部分，为不同的输入元素分配不同的权重。这些权重是在训练过程中学习的，并且是基于模型对其对手头任务的重要性的预测。最简单的注意力形式在数学上可以表达为(2-1):

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right) \mathbf{V} \quad (2-1)$$

其中， \mathbf{Q} 、 \mathbf{K} 、 \mathbf{V} 分别是注意力的查询、键和值矩阵， d_k 是键向量的维数。等式说明注意力机制计算的是 \mathbf{V} 的加权和，权重是查询和键直接的相关关系（此处为点积），以键向量的平方根为尺度，使用 softmax 函数将权重大小归一化，同时也代表键值的概率分布。

除 NLP 外，注意力机制还可以被广泛应用于各种任务，包括图像说明、视觉问题回答和语音识别等等。在这些应用中，注意力允许模型关注输入图像或音频信号的特定区域或特征，以便更好地理解底层结构并提取相关信息。总的来说，注意力机制已经成为提高各领域深度学习模型性能的重要工具，并为可解释人工智能和多模态数据融合等领域的研究开辟了新途径。

2.5.2 自注意力

自注意力机制（Self-Attention Mechanism）由 Vaswani 等人^[20]提出，是一种用于序列建模的机制。它可以在不使用递归和卷积的情况下将整个序列考虑在内，并且将序列中的每个元素与序列中的其他元素进行交互，从而计算每个元素的重要性权。换句话说来说，自注意力机制可以学习序列中每个元素之间的关系，从而更好地理解序列。

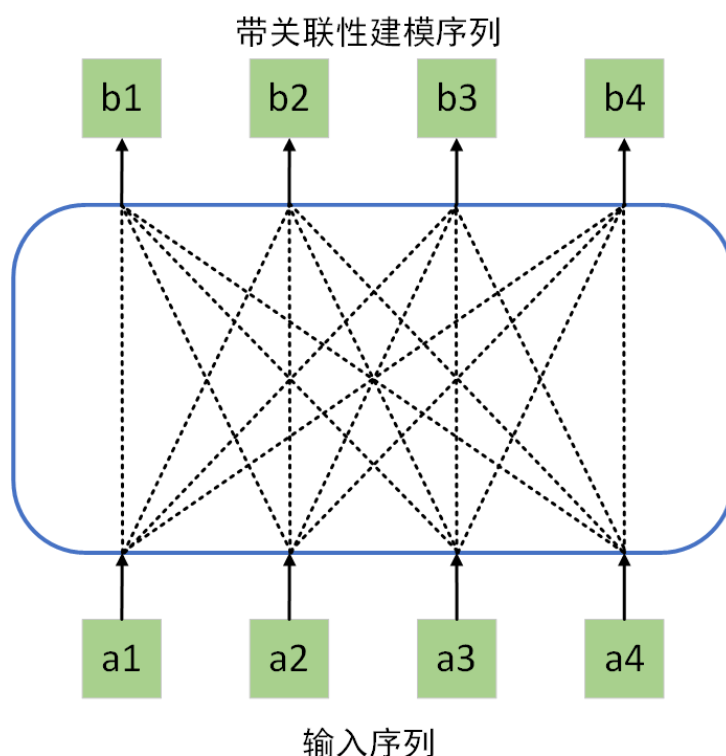


图 2-3 自注意力机制

2.5.3 跨模态注意力

跨模态注意力机制由 Xu, Kelvin 等人^[44]提出，是指利用注意力机制对来自不同模态的信息进行整合。它是深度学习中用于多模态数据融合的一种技术，数据可以来自不同的模态，如图像、文本或音频，跨模态注意力机制允许模型关注每个模态中与特定任务最相关的部分。例如，在视觉问题回答（VQA）的背景下，该模型不但可以关注图像中的特定区域，同时也考虑问题文本以生成答案。该机制通过计算不同模态之间的注意分数来工作。给定一个模态的输入，该模型计算出一组注意力权重，表明其他模态的每个元素的相关性。然后，注意力权重被用来计算其他模式元素的加权和，提供一个跨模式的表示，然后用于具体的下游任务。

2.6 贝叶斯不确定性估计理论

2.6.1 贝叶斯定理

贝叶斯定理是概率论中的一个基本公式，描述了在给定一些证据（观测值）的情况下，如何更新对一个事件发生的概率的估计。条件概率（又称后验概率）就是事件 A 在另外一个事件 B 已经发生条件下的发生概率。条件概率表示为 $P(A|B)$ ，读作“在 B 条件

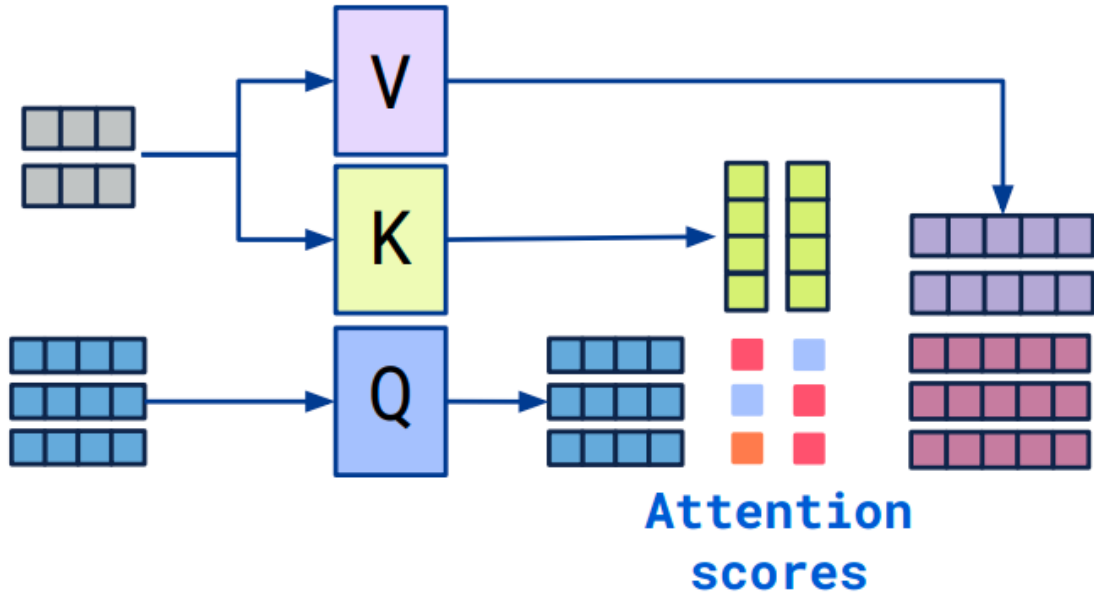


图 2-4 跨模态注意力机制

下 A 的概率”。其计算方式可以描述为，假设在同一个样本空间 Ω 中的事件或者子集 A 与 B，如果随机从 Ω 中选出的一个元素属于 B，那么这个随机选择的元素还属于 A 的概率就定义为在 B 的前提下 A 的条件概率(2-2)：

$$P(A | B) = \frac{P(A \cap B)}{P(B)} \quad (2-2)$$

$P(A)$ 和 $P(B)$ 是事件 A 和 B 发生的先验概率， $P(A \cap B)$ 是它们的联合概率。

2.6.2 贝叶斯网络概述

贝叶斯网络 (Bayesian network)，又称信念网络 (Belief Network)，或者有向无环模型，同时也是一种概率图模型，于 1985 年由 Judea Pearl^[45] 首先提出。它是一种模拟人类推理过程中因果关系的不确定性处理模型，其网络拓扑结构是一个有向无环图 (DAG)。图中的节点表示随机变量 $\{X_1, X_2, \dots, X_n\}$ ，他们可以是可观察到的变量，隐变量，或者未知参数等。具有因果关系 (或非条件独立) 的变量或者命题则用箭头来连接。若两个节点间以一个单箭头连接，表示两个节点之间是据因推果，则产生相应的一个条件概率值。

故而，把某个系统中涉及到的随机变量，根据是否条件独立绘制在一个有向图中，就形成了贝叶斯网络。该网络主要用来描述随机变量之间的条件依赖，对于任意的随机变量，其联合概率可以由各自的局部条件概率分布相乘而得出：

$$P(x_1, \dots, x_k) = P(x_k | x_1, \dots, x_{k-1}) \dots P(x_2 | x_1) P(x_1)$$

2.6.3 变分推断和蒙特卡洛方法

变分推断（Variational Inference）和蒙特卡洛方法（Monte Carlo Methods）都是概率图模型中常用的推断方法。变分推断是一种近似推断方法^[46]，通过在一族指定的分布族中寻找一个最接近真实后验分布的分布，从而近似后验分布。具体来说，变分推断将后验分布近似为一个分布族 $q(\theta)$ 中与真实后验分布 $p(\theta|D)$ 最为接近的分布 $q^*(\theta)$ ，即最小化 $q(\theta)$ 与 $p(\theta|D)$ 之间的 KL 散度：

$$\text{KL}(q(\theta)||p(\theta | D)) = \int q(\theta) \log \frac{q(\theta)}{p(\theta | D)} d\theta$$

变分推断的目标就是寻找一个最优的 $q(\theta)$ ，使得 KL 散度最小。变分推断方法的优点是得到一个解析形式的近似后验分布，计算速度较快。但是由于对分布族的限制，其对真实后验分布的逼近能力有限。

蒙特卡洛方法^[47]是一种基于采样的方法，通过从后验分布中采样得到样本，从而近似后验分布。常用的蒙特卡洛方法有马尔科夫链蒙特卡洛（MCMC）和重要性采样（Importance Sampling）等。蒙特卡洛方法的优点是能够直接从后验分布中采样，对分布形态的限制较少，能够得到更加精确的后验分布近似。但是，蒙特卡洛方法在计算量上通常较大，需要进行大量的采样。

第三章 基于多编码器混合自注意网络的医学视觉问答研究

究

本章提出一种多编码器混合自注意网络（Multi-Encoder Mixture Self-Attention Network, MEMSA）用其展开对医学视觉问答（MED-VQA）问题的研究，在传统采用单一自编码器进行图像编码的模型中分别加入元学习，对比学习等方法训练混合编码器模型，多个编码器可以从不同的“视角”和维度，充分挖掘并提取图像信息，获得更丰富的信息表征能力，是一种有效的视觉增强方法^[14]。同时采用跨模态自注意力的思想融合不同模态的编码信息，以获得模态间重要的特征关联。实验证明，多编码器混合方法以及自注意网络都有效地提高了 MED-VQA 模型在特征提取部分的融合表征能力，从而提高了视觉问答模型的问答准确率。

3.1 多编码器混合自注意网络

3.1.1 MEMSA 网络架构

在计算机科学领域，每个问题往往都有着一套通用的解决方案和基础结构。目前解决 VQA 问题的模型基本方法是 Q（问题）+I（图像）的联合嵌入方法^[48]。该方法的框架如图3-1所示,由主要由特征提取、特征融合、回答预测三部分构成，包含图像编码器、文本编码器、特征融合算法模型以及根据任务需要设计的回答预测四大组件。

MEMSA 网络架构

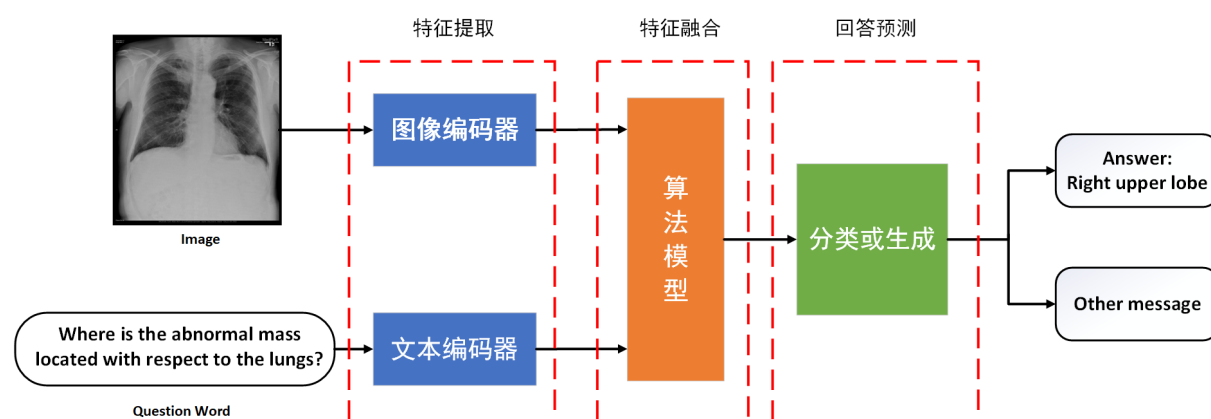


图 3-1 总体框架

如图3-1,Med-VQA 中图像编码器使用的一般是目前比较成熟的卷积神经网络，如 VGGNet、ResNet；文本编码器主要用于编码问题句子，常用的有目前主流的语言编码

模型，如 LSTM、Bert 等。编码器通常都使用预训练权重进行初始化，在训练过程中可以冻结也可以用端到端的方式进行微调。特征融合以及算法部分目前主流的方式有加权融合，拼接融合，注意力融合等方法。回答预测部分根据任务的不同往往采用不同的设计：如针对分类问题，组件通常是一个神经网络分类器；针对生成问题，组件通常是循环神经网络（RNN）语言生成器或者注意力模型。

MEMSA 模型设计

依据上述视觉问答的基本框架，并针对当前医学视觉问答（Med-VQA）问题存在的几大难点提出了如图3-2所示的多编码器混合自注意力网络模型（MEMSA）：

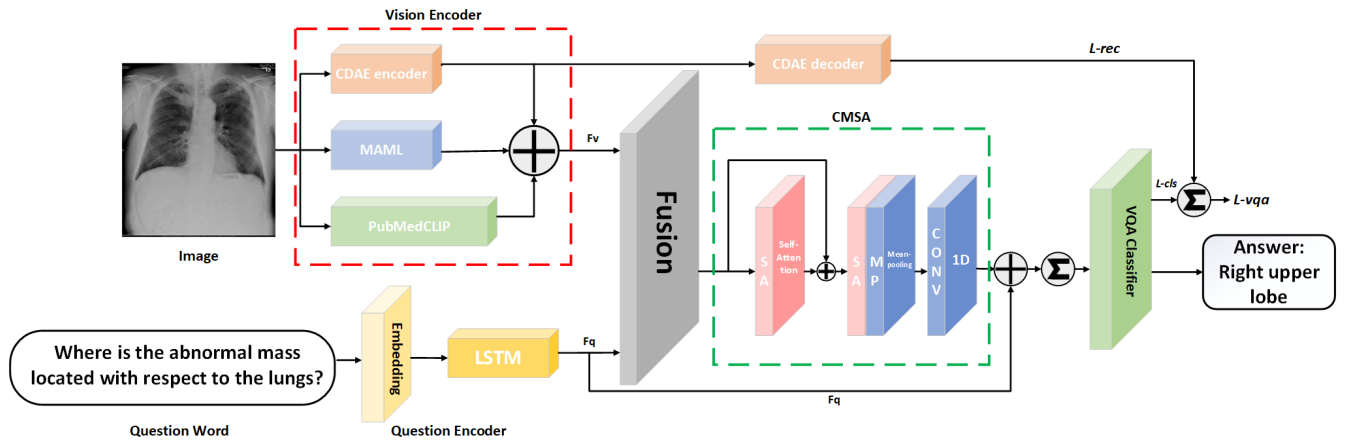


图 3-2 多编码融合自注意力网络

MEMSA 网络的输入为图像和针对图像提出的开放式或者封闭式文本问题。首先，MEMSA 在特征提取部分，使用多个编码器分别提取图像的特征表示，再使用如图 GloVe 词编码空间和 LSTM 网络提取问题的文本特征表示。

接着，在多模态融合部分，先将图像的特征表示和文本的特征表示拼接，拼接后和自身的转置组合成类似于共现矩阵的形式，再对该矩阵计算跨模态自注意力，通过注意力加权提取出其融合表征，通过残差连接将不同层提取到的信息保持，最终通过每一层特征加和的方式得到一个具有高度相关性的图像-文本联合表征。

最后，在答案推理预测部分，由于 Med-VQA 的特殊性，将答案预测视为多标签分类问题，使用多层感知机对该联合表征进行求解得到最终的预测，然后再通过最小化预测与标签之间的交叉熵损失来训练调整模型。接下来各小节将逐一介绍 MEMSA 各模块的组成和工作原理。

3.1.2 各编码器原理

多编码器的混合的提出最早是为了解决现有医学图像样本的几大缺陷^[4]：1. 样本数量少；2. 噪声大，有用信息占比少；3. 专业性强，成本高 4. 同类图像整体相似度高，难以区分；5. 多为灰度图，有用的特征较少。MEMSA 网络分别采用降噪自编码器、元学习模型、对比学习预训练模型作为图像编码器。降噪自编码器可以降低图像噪声，提取有用信息。元学习模型通过在元学习环境中训练的一个通用模型，从而达到在小样本条件下快速适应新任务的效果，十分适用于医学视觉问答这一领域。对比学习预训练模型由于在大样本预训练时充分学习了图像和语义的对齐关系，所以在编码上有着更丰富的表示，有利于完成开放式问题的回答。

图像编码器

用于 MEMSA 的图像编码器，由卷积降噪自编码器 CDAE、模型不可知元学习模型 MAML 以及医学对比学习预训练模型 PubMedCLIP 三个部分混合构成：

(1) 卷积降噪自编码器 CDAE

卷积降噪自编码器 (Convolutional Denoising Auto-Encoder, CDAE) 有编码器和解码器两部分，编码器通过卷积和池化操作提取特征并输出给具有较低维度的隐藏层学习其表示，该表示包含了图像的主要特征而过滤掉了噪声（采用均方误差 MSE），再由解码器重新进行上采样重建后获得了更纯净的图像表示，DAE 首先将原始图像输入 x (带噪声) 映射到保留有用信息的潜在表示 z ，然后再通过解码器将 z 转换为输出 y 。自编码器的训练目标是 최소화输出 y 与原始图像 x 之间的重建损失：

$$L_{rec} = \|x - y\|_2^2 \quad (3-1)$$

遵循上述原理设计降噪自编码器。为了得到较好的隐变量表示，降噪自编码器由 5 层卷积层堆栈而成。解码器是一个反卷积层和卷积层的堆栈。带噪声的 x' 是通过在原始图像 x 上添加高斯噪声来实现的。训练结束后使用编码器和解码器的权值在模型中进行微调。

(2) 模型不可知元学习编码器 MAML

基于元学习方法搭建的启发式元学习 (Model-Agnostic Meta-Learning, MAML) 编码器，其原理为抽取一定数量的样本，在元环境下进行训练得到一个初始化

的权重，该权重可以在小样本下就能促进模型达到一个比较好的效果。MAML 由带初始化元参数 θ 的参数化函数 f_θ 表示。当需要学习一个新任务 T_i 时，设模型参数为 θ'_i ，设训练所用的数据集为 \mathcal{D} 样本大小为 N ，在使用元学习方法进行少样本学习时，任务往往可以定义为是一个“k-shot n-way”分类问题。对每个元任务所用的训练集 \mathcal{D}' 来源于数据集 \mathcal{D} 中的 n 个不同的类，训练时，将 \mathcal{D}' 平均分成训练集 \mathcal{D}^{tr} 和验证集 \mathcal{D}^{val} ，每个类包含 k 个训练图像。迭代 h 次，生成用于元学习训练批处理的 m 个任务，对于每个任务 \mathcal{T}_i ，计算和更新初始化权重的方法如下：

$$\theta'_i = \theta - \alpha \nabla_{\theta} L_{\mathcal{T}_i} (f_{\theta} (\mathcal{D}_i^{\text{tr}})) \quad (3-2)$$

其中 $L_{\mathcal{T}_i}$ 为任务 i 的分类损失，在计算完 m 个任务的参数后，通过随机梯度下降（SGD）法更新元模型的参数 θ ，如下：

$$\theta \leftarrow \theta - \beta \nabla_{\theta} \sum_{\mathcal{T}_i} L_{\mathcal{T}_i} (f_{\theta'_i} (\mathcal{D}_i^{\text{val}})) \quad (3-3)$$

遵循上述规则设计元学习编码器。由 4 层 3×3 的卷积层，每层卷积后跟着一层归一化层，步幅为 2，每个卷积层后接大小为 64 的 BN 层进行批次归一化处理，以 Relu 作为激活函数，解决深度网络中数值不稳定的问题，缓解梯度的消失，提高训练效率。

(3) 对比学习预训练编码模型 PubMedCLIP

PubMedCLIP 是基于 PubMed 文章数据库对经典 CLIP 模型进行预训练以及微调得到一个具有医学图像-文本表示能力的编码模型，同时也是一个跨模态模型。PubMedCLIP 由图像编码器和文本编码器构成，图像编码器可选用基于 CNN 的模型，也可以选用基于 Transformer 的模型；文本编码器可选中基于 RNN 的模型，同样也可以选用基于 Transformer 的模型。以本文使用到的编码器为例，受制于医学图像的样本数量，选用了样本需求较少的 Resnet50 模型，然后基于 Pelka 等人提出的 ROCO 数据集上对模型进行预训练。对比学习预训练过程是通过计算图像文本对 $I - T$ 之间的余弦相似度得到。余弦相似度越大，表明图像 I 和文本 T 的对应关系就越强，反之越弱。所以只需要最大化正样本的余弦

相似度，最小化负样本的余弦相似度即可完成优化目标：

$$L_{clp} = \min \left(\sum_{i=1}^N \sum_{j=1}^N (I_i \cdot T_j)_{(i \neq j)} - \sum_{i=1}^N (I_i \cdot T_i) \right) \quad (3-4)$$

其中，N 为图像文本样本对 $I - T$ 的个数，训练完成后得到 PubMedCLIP 的图像编码模型在编码图像时就有了和对应文本的相似关系，从而使得该向量包含了相关含义的语义表示，便于进行下游图像-文本任务以及迁移学习。

文本编码器

用于 MEMSA 的问题文本编码器，主要分为 Word embedding 和 Question embedding 两部分，一个输入文本或者词汇先由 Word embedding 按照某一词向量空间表示方法将其映射到一个向量表示空间中，成为一个词向量，然后这些词向量再通过 Question embedding 进行不同向量间的语义关系建模和学习。

(1) Word embedding

词编码器（Word embedding）是一个词嵌入算法模型，通常由两层全连接神经网络构成，用于学习文本的词向量表示。MEMSA 使用 GloVe6b 作为词向量编码模型，假设有一个大小为 V 的词汇表，其中包含单词 i。GloVe 模型中的目标是学习每个单词的向量表示，使得这些向量可以在预测上下文单词和中心单词时发挥良好的作用。具体地，GloVe 使用一个共现矩阵 X 来表示单词之间的共现信息，其中 X_{ij} 表示单词 i 和 j 在同一个上下文中出现的次数。然后，GloVe 定义了每个单词的两个向量表示：中心单词的向量表示 w_i 和上下文单词的向量表示 w_j 。GloVe 模型的学习目标是 minimize 损失函数 J (3-5)。

$$J = \frac{1}{2} \sum_{i,j=1}^{|V|} f(X_{ij}) (\mathbf{v}_i^T \mathbf{u}_j + b_i + b_j - \log X_{ij})^2 \quad (3-5)$$

其中， $f(X_{ij})$ 是一个权重函数，用于对共现矩阵进行加权。 b_i 和 \tilde{b}_j 是偏置项，用于偏移单词向量表示的值。通过最小化上述损失函数，可以学习到每个单词的向量表示。这些向量表示可以作为文本特征进行使用。经医学图像文本预训练的 glove 词嵌入表示经过 t-SNE 方法降维后如图3-3，可知经过一万次的迭代降维后，Glove 词向量之间已经在二维平面上初步呈现出各类词向量编码之间的区别，表示数字的语义和具有相似意义的词向量已经被模型归类出来划分到了一个具有近似意义的空间中。

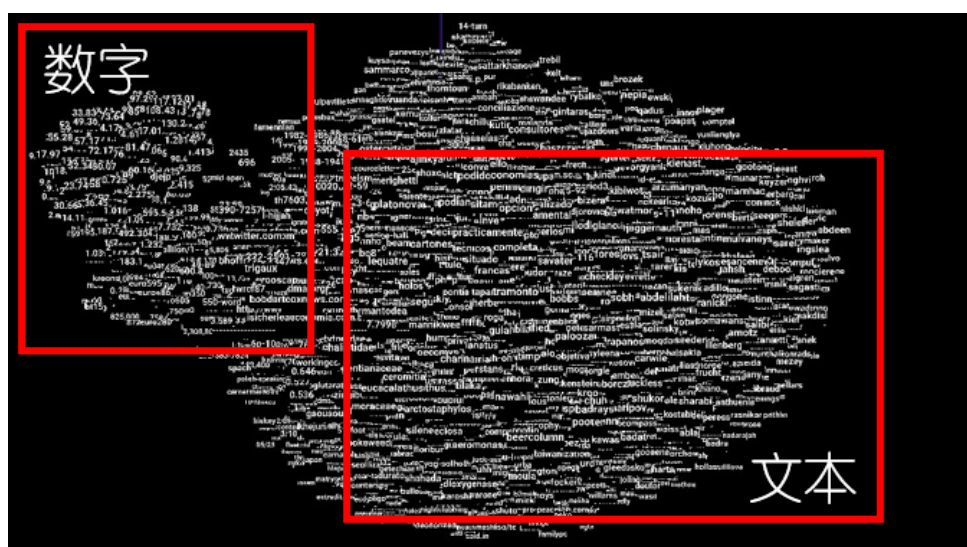


图 3-3 GloVe 词嵌入的降维表示

(2) Question embedding

句编码器（Question embedding）是一个用于学习词向量间的语义关系的编码模型，通常可以是循环神经网络（RNN）或者是纯注意力结构。由于医学文本的数据限制，且为了更好地捕获问句序列上下文间的语义联系，MEMSA 采用长短时记忆网络 Long short-term memory（LSTM）作为 Question embedding。如图3-4, LSTM 的基本单元包括三个门（输入门、遗忘门、输出门）和一个记忆细胞，其中输入门控制输入信息的重要性，遗忘门控制上一时刻记忆细胞的重要性，输出门控制输出信息的重要性。记忆细胞可以存储和读取信息，从而实现长期依赖关系的建立和维护。例如在抽取一个句子 X 的语义特征时，可以使用 LSTM 来处理这个句子。假设这个句子 X 包含 n 个词，每个词用一个 d 维的词向量表示，那么可以将这个句子的词向量表示为一个 $n \times d$ 的矩阵 $X=[x_1, x_2, \dots, x_n]$ ，其中 x_i 表示第 i 个词的 d 维词向量。将这个矩阵输入到 LSTM 中，经过多个时间步的计算，LSTM 会输出一个长度为 h 的向量表示句子的语义特征，其中 h 是 LSTM 的隐藏单元数。这个向量可以作为句子的表示，用于后续的任务，比如分类、聚类等。

3.1.3 跨模态自注意力机制

跨模态自注意力机制由最早 P Gao 等人^[49]提出，是一种基于自注意力机制实现的跨模态序列融合方法，受 Haifan Gong 等人^[17]将其作用在医学图像和文本上的启发。由于自注意力机制可以通过对序列中不同位置的元素之间的相互作用进行建模来捕捉序列

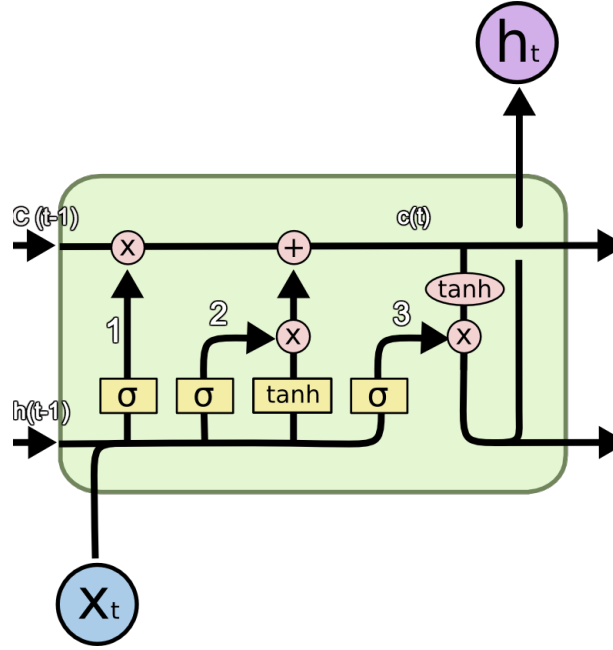


图 3-4 LSTM 网络结构

中的长程依赖关系。

如图不同于传统计算图像文本相似度关系的跨模态注意力机制，MEMSA 采用的是从图形文本拼接特征中用自注意力机制提取融合特征；不在各模态中单独计算注意力权重矩阵，而是将不同模态特征统一到一个特征空间里进行表示，通过上建立自我内部序列之间的关联来对各模态之间的特征相关关系进行建模，从而捕获更好的图像-语义关联。本文使用跨模态自注意力机制进行特征融合时，通过先将问题句子按词汇进行词向量拆分，每个词向量再和图像特征进行拼接，最后重新组合成句子。通过这种方式来捕获语义的基本单元-词汇级的细粒度，不但可以把握词汇之间的上下文联系，同时也能把握图像和词汇、文本之间的语义对齐关系。

根据自注意力机制和数学原理，在模态融合前，首先需要将编码得到的视觉特征 v 和问题特征 q 相连接，为了提高表征的细粒度，先将问题特征 q 按词重新拆分，都将其与视觉特征 v 相拼接得到局部特征图 f ，然后重新合并组合成融合的多模态特征图 F 。然后对该特征图建立注意力矩阵 Q, K, V ，使用 3D 卷积分别进一步提取 Q, K, V 特征并重塑成二维的特征表示。用该二维特征与其自身的转置相乘得到特征之间的关联表示，再经过 softmax 函数缩后放后就得到了对应的注意力图：

$$A = \text{softmax}(QK^T) \quad (3-6)$$

这样，注意力图 A 中就蕴含了融合特征 F 内各部分特征之间的相关关系，相比于一维

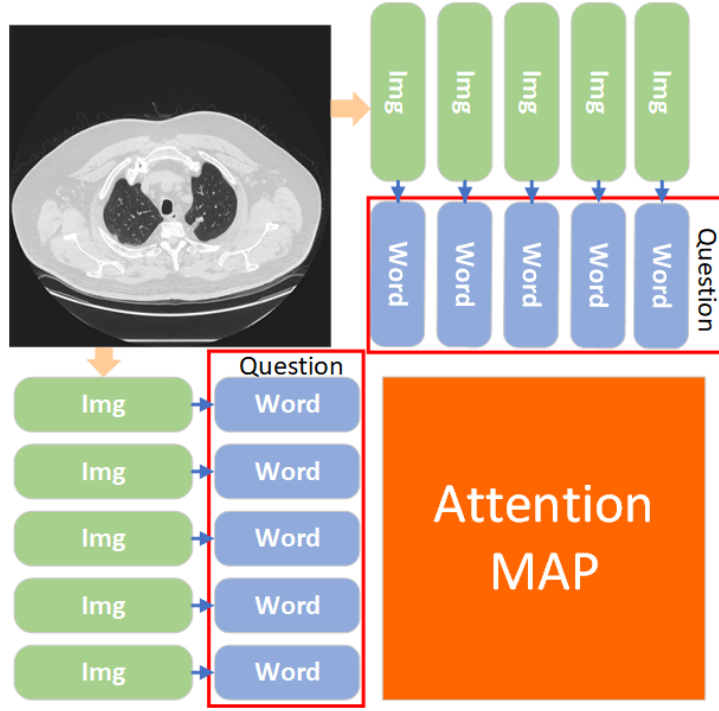


图 3-5 词汇级细粒度融合特征关注图

序列上的自注意力，这是一种在二维特征图上的自注意力机制。接着，特征图 V 经过注意力 A 加权后得到注意力增强的多模态表示 F' ：

$$F' = AV \quad (3-7)$$

经过注意力增强后 F' 再次经过转置，重塑，反卷积操作还原成最初多模态特征图 F 的形状，此外，受到 Resnet 网络结构的启发，多次重复了注意力模块并使用残差连接保留所有的局部特征。经过均值和池化处理后，最终的多模态表示为：

$$\hat{F}_i = \frac{\sum_{j=1}^m \sum_{k=1}^n (F'_{ijk} + F_{ijk})}{m \times n} \quad (3-8)$$

其中 i, j, k 为特征图 F 构建时所包含的词汇数位以及特征图高度、宽度， m, n 为该特征图总尺寸，最终 \hat{F}_i 经过线性层映射到和文本特征 q 相同的维度。

3.2 模型预测与网络训练

3.2.1 模型预测

感知机（Multilayer Perceptron, MLP）是一种简单、常用的分类网络，有输入层，隐藏层，输出层组成，并且各层皆为全连接网络。输入层接收数据，隐藏层对数据进行非线性变换和特征提取，提高数据的表达能力，层数越多，感知机对特征的抽象能力往往

越强，内部非线性由激活函数激活，最后计算输出损失，根据反向传播算法优化网络的权重和偏置，从而达到学习的效果。医学视觉问答由于分为开放式问答和封闭式问答，对于检索式问答系统，可将这两种问答分别看成是多标签分类问题和二分类问题。为了提高模型的简洁度，MEMSA 模型同样将 yes, no 同样视为两个独立的标签参与多分类。虽然给分类增加了难度，但也打开了问题形式的局限。多标签分类问题往往通过计算多标签损失（MSE）来作为梯度信息进行反向传播。多模态表示 \hat{F} 经过线性映射后维度和问题 q 相同，与问题 q 句子中的所有词上进行求和，最后将其输入到一个两层 MLP 中进行答案预测，获得答案的预测分数 s ，答案的预测分数计算如表达式(3-9):

$$s = MLP \left(\sum_{i=1}^N (\hat{F}_i + q_i) \right) \quad (3-9)$$

变量 N 为控制问题句子长度的超参数，同时也指示问题句子所包含的词汇个数。

3.2.2 网络训练

MEMSA 模型由多个部分组成，包括图像编码，问题编码，跨模态注意力融合以及答案预测。故采用多任务损失并以端到端的方式训练模型：

$$L = L_{vqa} + \alpha L_{rec} \quad (3-10)$$

L_{vqa} 是基于分类的答案预测和标签之间的交叉熵损失； L_{rec} 如上述是自编码器输出 y 与原始图像 x 之间的重建交叉熵损失。 α 是平衡两个损失项的超参数，往往设置为 0.5。

实验超参数设置

模型的超参数设定如表3-1:

实验条件与环境

实验环境设置如表3-2:

3.3 实验结果

3.3.1 编码器模型实验对比分析

目前各混合视觉增强方法在 Med-RAD、SLAKE 数据集上的开放式问答，封闭式问答，以及总准确率如表??。其中，因为 SLAKE 数据集缺少 MTPT 方法所需要的图像来源标签，所以此处不予计算和讨论。

从表3-3可以预见，在使用经典的双线性池化注意力机制（BAN）以及跨模态自注

表 3-1 模型超参数设置

参数名	参数值	参数含义
Word token	12	输入问句的划分长度
Word embedding dim	600	词向量维度 (GloVe)
Visual feature dim	1152	图像特征维度
LSTM dim	1024	文本特征维度
Joint feature dim1	2184	多编码特征混合后的的维度
Joint feature dim2	1024	特征经注意力融合后的的维度
Batch Size	32	训练批量大小
Learn rate	(0,0.01)	动态学习率区间

表 3-2 深度学习环境配置

环境名称	版本
处理器	Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz
内存	64 GB
操作系统	Ubuntu 18.04
开发语言	Python 3.6
GPU	NVIDIA TITAN Xp 12GB
CUDA	CUDA Toolkit 11.2
深度学习开发库	Pytorch 1.7.1 for GPU

注意力机制（CMSA）作为融合网络时，PubMedCLIP 的使用和在多编码器混合作用的效果下，模型在小样本数据集 Mad-RAD 下的开放式问答表现有着明显的提高，采用多编码器混合加双线性池化注意力机制 MEMBA 这一方法，相比 18 年 Binh 等人提出的基线模型 MVEF-BAN^[14],MEMBA 模型针对开放式回答的效果提升了 8.1%, 封闭式回答提升了 2.9%, 总体性能提升了 5.6%。采用 MEMSA 模型，综合性能分别提升了 21.5%、2.5% 和 10%。在 SLAKE 数据集上，MEMBA 综合提升了 1% 2%，MEMSA 提升了 2% 3%，性能提升不明显的主要原因是 SLAKE 数据量级较大，模型抽取的特征相对完备，单一或简单编码器组合也能取得良好的性能，就会导致针对小样本情况下设计的 MEMSA 带来的性能提升不明显。综上，通过采用具有不同功能的多编码器混合以及跨模态自注意力方法，MEMSA 可以通过提升特征的多样化表示，在开放式问题下获得更好的性能

表 3-3 MEMSA 与主流方法等对比

Dataset	Med-RAD			SLAKE		
Methods	Open	Closed	ALL	Open	Closed	ALL
RAD-SAN ^[37]	24.2%	57.2%	44.0%	X	X	X
RAD-MCB ^[37]	25.4%	60.6%	46.5%	X	X	X
MVEF-BAN ^[14]	43.9%	75.1%	62.7%	75.0%	76.4%	75.6%
MTPT-BAN	47.2%	77.8%	65.6%	X	X	X
CPAE-BAN ^[19]	48.6%	78.1%	66.5%	76.2%	79.9%	77.6%
MEMBA	52.0%	78.0%	68.3%	76.3%	78.6%	77.2%
MVEF-CMSA	43.9%	75.1%	62.6%	75.8%	81.5%	78.0%
MTPT-CMSA ^[17]	56.1%	77.3%	68.8%	X	X	X
CPAE-CMSA	63.7%	76.1%	71.2%	77.2%	81.5%	78.9%
MEMSA	65.4%	77.6%	72.7%	76.0%	81.7%	78.2%

效果，从而提升模型整体的问答性能。

3.3.2 注意力模型实验对比分析

表 3-4 不同注意力下的模型性能对比

Dataset	Med-RAD			SLAKE		
Methods	Open	Closed	ALL	Open	Closed	ALL
MVEF-SAN ^[14]	40.7%	74.1%	60.8%	72.9%	77.6%	74.7%
MVEF-BAN ^[14]	43.9%	75.1%	62.6%	75.0%	76.4%	75.6%
MVEF-CMSA	43.9%	75.1%	62.6%	75.8%	81.5%	78.0%
MTPT-SAN	43.0%	72.8%	61.0%	X	X	X
MTPT-BAN ^[17]	56.1%	75.7%	67.9%	X	X	X
MTPT-CMSA ^[17]	56.1%	77.3%	68.8%	X	X	X
MEM-SAN	53.6%	75.7%	67.0%	74.9%	82.0%	77.7%
MEM-BAN	52.0%	78.0%	68.3%	76.3%	78.6%	77.2%
MEMSA	65.4%	77.6%	72.7%	76.0%	81.7%	78.2%

使用不同注意力机制时，模型性能如表3-4。固定同一编码方式为 MVEF、MTPT、

MEM 的情况下,采用几个常用的基于注意力的特征融合方法:堆叠注意力机制(SAN),双线性池化注意力机制(BAN)和跨模态自注意力(CMSA)进行对比分析:在 Med-RAD 数据集上,每种注意力之间呈现逐渐递增的态势。并且发现编码器多样性越高,跨模态自注意力机制的效果就越明显,这充分说明了自注意力在跨模态或者多模态信息的建模上具有独特的优势。

并且通过三类编码方案综合对比,多编码器混合方式对于各种注意力融合机制具有最佳的适配性,在 SAN、BAN 等传统方法上都可以大幅度提升性能。在数据量较大,基准准确率较高的 SLAKE 数据集上,模型的综合性能也得到了小幅度提升。

3.3.3 实验总结

综合上述实验,在使用不同的编码方式时,具有不同作用的多编码器可以提高模型的代表能力,提高模型的分类准确率;

我们知道,向量之间的主要运算有 cat, sum, mul 等形式。受空间约束和相关性的影响,属于不同模态,代表不同意义的向量之间无法直接产生有效的运算形式。如果不经坐标变换,映射,归一化等操作再运算的话则会导致严重的信息丢失。可以预见,在神经网络中,直接拼接组合是相比于运算更为有利的信息融合方式。因为维度代表一种信息的尺度,在编码时,每增加一个维度就会带来这个信息新的“描述”,从而提升了这个信息的表达能力和空间上的可区分度。值得注意的是,不同的编码器在编码同一信息时难免会产生冗余,比如描述的向量空间中的某些分量是呈现线性相关的。这时候就会引入重复的信息和增大特征维度给模型设计和计算资源带来难题。

而且相比于传统的特征处理方式,基于注意力是一种更为灵活,高效,而且可以由神经网络自发的学习和选择不同模态中的有效特征的一种多模态融合方式。为了解决不同模态编码信息维度之间有效匹配的问题本文采取了词汇级细粒度表征的跨模态自注意力的方法,跨模态自注意力通过计算不同模态信息之间的匹配程度得到这些信息维度的重要性权值用于筛选出真正有效(线性无关)的特征信息,从而实现更高效地捕获不同模态间的联系以及和自身的上下文关联。

3.4 讨论

为了更加深入地了解多模态混合自注意力网络的机制,理解其提升医学视觉问答效果的途径和原理,本小节将继续讨论 MEMSA 网络具体回答 Med-RAD 数据集中某一类问题的能力,同时对 MEMSA 网络的分类预测能力进行综合评估,最后通过问答样例对

比直观地感受 MEMSA 网络在回答医学视觉问题时的实际效果。

3.4.1 按问题类型划分的准确率比较

除上述表??中针对回答方式（OPEN，CLOSE）的划分外，还可以针对不同的提问内容（如提问计数、颜色、尺寸、模态等）进行问答划分，分析 Med-VQA 模型在不同内容提问下的准确率。用 Med-RAD 数据集进行模型性能测试，其在开放式问答和封闭式问答中针对具体的回答内容的分类性能分别如表3-5和表3-6，

表 3-5 CLOSE-Question Results

Model	MEMSA					MVEF-BAN (base)				
QA-type	count	real	TRUE	real %	score %	count	real	TRUE	real %	score %
COUNT	4	4	3	100 %	75 %	4	4	3	100 %	75 %
COLOR	4	4	4	100 %	100 %	4	4	4	100 %	100 %
ORGAN	2	2	2	100 %	100 %	2	2	2	100 %	100 %
PRES	124	124	106	100 %	85.5%	124	124	104	100 %	83.9 %
PLANE	12	11	7	91.7 %	58.3%	12	11	4	91.7 %	33.3 %
MODALITY	17	17	12	100 %	70.6%	17	17	8	100 %	47.1 %
POS	8	8	5	100 %	62.5%	8	8	4	100 %	50 %
ABN	38	35	28	92.1 %	73.7 %	38	35	28	92.1 %	73.7 %
SIZE	41	40	35	97.6 %	85.4 %	41	40	35	97.6 %	85.4 %
OTHER	11	10	4	90.9 %	36.4 %	11	10	8	90.9 %	72.7 %
ATTRIB	16	16	14	100 %	87.5 %	16	16	15	100 %	93.8 %

可见，MEMSA 模型的性能提升主要来自于其能够更好回答开放式问答和封闭式问答中有关 PRES、PLANE、MODALITY、POS 这几类数据集占比较大的问题。

3.4.2 MEMSA 模型综合评估

建立模型预测的混淆矩阵是综合评价一个模型分类性能的有效方法，对于一个分类模型，可以建立一个类别数为 N 的混淆矩阵，行表示预测标签，列表示真实标签。矩阵中的每个元素表示真实标签和预测标签的组合，即在真实标签为该行所表示的标签时，模型预测为该列所表示的标签的样本数量。在非生成式视觉问答场景中，封闭式问答为归为一个二分类问题，开放式问答可以归为一个多分类问题，Li-Ming Zhan 等人给模型增加了一个即插即用的 QCR 和 TCR 模块^[16]用于问题分类，并解释了将问题按照类型

表 3-6 OPEN-Question Results

Model	MEMSA					MVEF-BAN (base)				
QA-type	count	real	TRUE	real %	score %	count	real	TRUE	real %	score %
COUNT	2	2	2	100 %	100 %	2	2	2	100 %	100%
COLOR	0	0	0	0 %	0 %	0	0	0	0 %	0%
ORGAN	8	6	3	75 %	37.5 %	8	6	3	75 %	37.5%
PRE	47	36	31	76.6 %	66%	47	36	21	76.6 %	44.7%
PLANE	14	13	11	92.9 %	78.6 %	14	13	11	92.9 %	78.6%
MODALITY	14	10	3	62.5 %	18.8 %	16	10	5	62.5 %	31.2%
POS	53	47	42	88.7 %	79.2%	53	47	40	88.7 %	75.5%
ABN	18	18	17	100 %	94.4%	18	18	11	100 %	61.1%
SIZE	5	5	3	100 %	60 %	5	5	4	100 %	80%
OTHER	15	3	0	20 %	0 %	15	3	0	20 %	0%
ATTRIB	4	4	3	100 %	75 %	4	4	3	100 %	75 %

进行事先划分可以更好地提高模型问答的准确率。通过观察混淆矩阵3-6和3-7，可以直观地了解模型的对各问题的分类情况，例如哪些类别容易被混淆、哪些类别分类效果较好等，进而对模型进行优化。

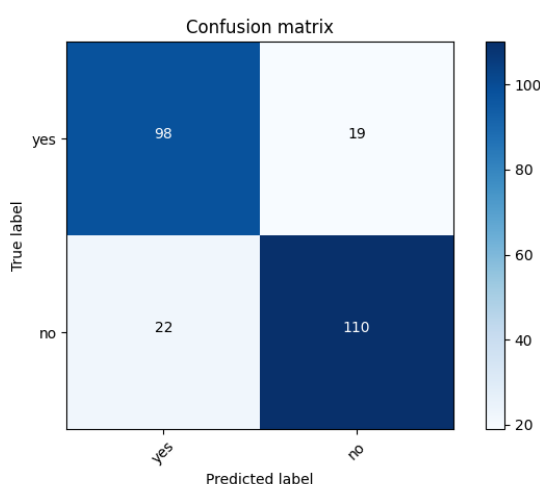


图 3-6 封闭式问答混淆矩阵

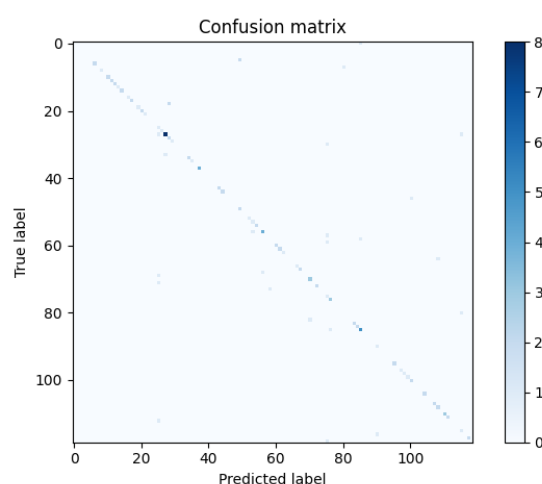


图 3-7 开放式问答混淆矩阵

对于分类检索式问答来说，准确率如表??代表了模型分类到正确的答案占总样本数的比例，它反映了模型整体的分类性能也是最重要的指标；关于表3-7，精确率表示模型在预测为正例的样本中真正为正例的样本数占预测为正例的样本数的比例，它反

表 3-7 不同方法的综合性能对比

方法	TP (sum)	TN (sum)	FP (sum)	Precision (avg)	Recall (avg)	F1 (avg)
MVEF-BAN	290	73	73	0.658	0.643	0.644
CPAE-BAN	282	72	72	0.671	0.625	0.637
MTPT-CMSA	296	89	89	0.648	0.656	0.645
MEMSA	309	73	73	0.685	0.685	0.678

映了模型在预测为正例的样本中的准确程度，从表中可得知，MEMSA 模型具有最大的 Precision 值，相较于主流模型具有比较优秀的精度，可以保证回答的质量；召回率是指模型在所有真正的正例样本中预测为正例的样本数占有所有真正的正例样本数的比例，它反映了模型在发现真正的正例样本方面的能力。同样，MEMSA 也具有最大的 Recall 值，在问答时尽可能地预测出所有正确的回答。F1 分数是精确率和召回率的调和平均数，它综合了精确率和召回率的优缺点，用于综合评价模型的性能，综合来看 MEMSA 具有最高的 F1 分数，说明 MEMSA 相较其他主流模型存在优势。

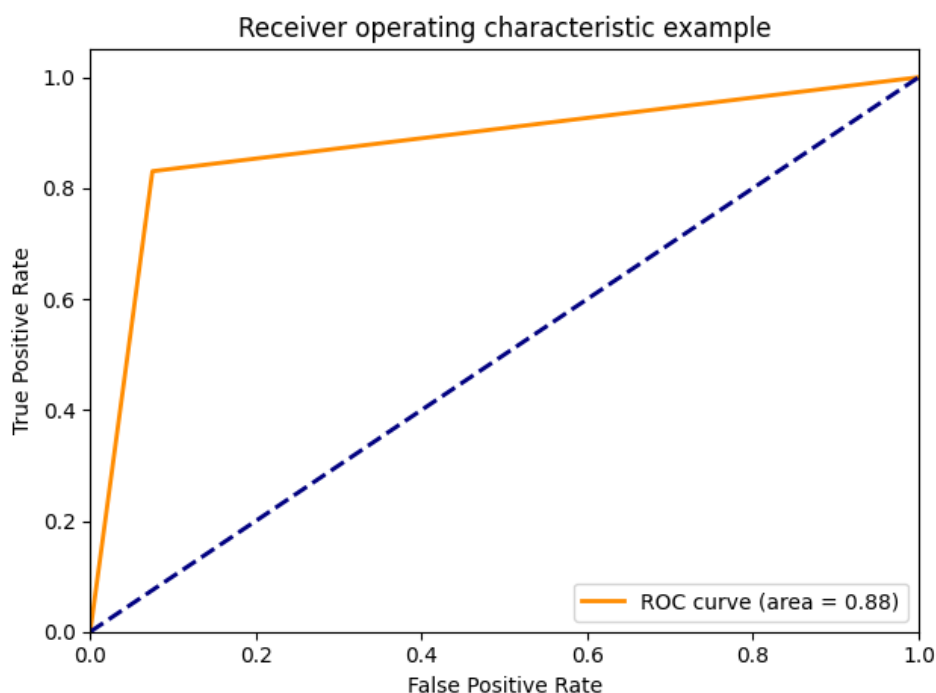


图 3-8 封闭式问答模型 ROC 曲线

从图3-8可以看出，对于回答封闭式这一二元分类问题，在模型默认的绝对分类阈值以及分类方式下，模型 AUC 值达到 0.88，说明 MEMSA 模型在回答二分类问题时具有较强的泛化能力。由于开放式问答（多标签分类问题）的 ROC 曲线不具有显著的实

际意义，故本文在此不做详细讨论。

3.4.3 MEMSA 实例问答效果

为了直观地展现 MEMSA 在医学视觉问答上的实际效果，具体问答样例如图3-9，选用 MEMSA 和基线模型 MEVF-BAN 作问答对比，模型在做开放式问答时，MEMSA 准确预测了问题出现在肺叶的右上叶（right upper lobe），相比于采用双线性池化的 MVEF-BAN 模型预测的不具备直接语义的回答“right upper base”MEMSA 具有更好的细粒度语义解析能力和对语句序列的上下文进行语义建模的能力。同时，在具有多个正确答案的开放式问答图中具有什么器官中，主流模型往往倾向于回答脊髓（spinal cord）MEM



	
Close_Q: Is the heart abnormally large? MEVFBA_A: No(✓) MEMSA_A: No(✓) Right answer:No	Close_Q: Does the picture contain lung? MEVFBA_A: no(✗) MEMSA_A: yes(✓) Right answer:Yes
Open_Q: Where is the abnormal mass located with respect to the lungs? MEVFBA_A: right lung base(✗) Right answer: MEMSA_A: right upper lobe(✓) Right upper lobe	Open_Q: What is the main organ in the image? MEVFBA_A: spinal cord(✓) MEMSA_A: lung(✓) Right answer: Lung/Spinal Cord

图 3-9 MEMSA 与 MEVFBA（MAML+AE-BAN）的回复比较

3.5 本章小结

本章首先介绍了主流视觉问答网络的实现原理和总体架构，包括特征提取、特征融合和答案预测等组件。接着，根据目前主流模型的一些缺陷和医学视觉问答的特点，提出了多编码器混合自注意力模型 MEMSA。然后在具体的实验部分介绍了使用的数据集和模型细节，实验所处的环境和实验条件。本章在实验中从整体性能，问答类型性能等多个角度综合对比了多编码器混合自注意力模型相比于主流模型的性能差异并对原因进行了分析，接着做了分解对照实验，固定编码器模型，重新分析了不同注意力模型对问答性能的影响，证明了多编码器模型在非跨模态自注意力模型的情况下也能提升部分模型性能，说明了正确使用多模态编码模型时的一个优越性。

第四章 基于局部贝叶斯神经网络的医学视觉问答及其不确定性研究

贝叶斯神经网络的应用领域十分广泛，不但可以应用于传统的分类和回归任务的学习以及预测，还可以通过学习给定状态下行动的概率分布来学习强化学习问题中的最优策略，也可以通过将从一个任务中学到的后验分布作为另一个相关任务的先验分布从而将知识从一个任务转移到另一个任务实现迁移学习。最为特殊的是，贝叶斯神经网络可以用来估计模型预测的不确定性，这在一些具有极高不确定性量化需求的任务中具有十分重要的用途。

4.1 贝叶斯神经网络

4.1.1 贝叶斯神经网络简介

传统深度神经网络使用点估计作为权值，忽略了权值中的不确定性，往往会导致网络对自身决策“过度自信”^[50]。为了解决这个问题，Blundell 等人^[51]受贝叶斯网络的启发提出了一种贝叶斯神经网络（BNN，Bayes neural network）模型。与传统神经网络不同的是，该 BNN 的权值是一个概率分布，从而在权值上引入了不确定性。通过权值不确定性来估计预测的不确定性。这些权值分布的后验参数通过贝叶斯反向传播算法（BBB）学习得到。如4-1展示了 BNN 和传统神经网络的区别。

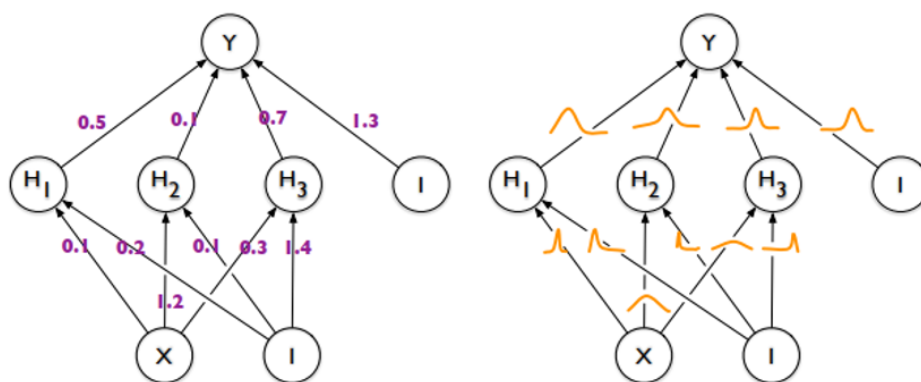


图 4-1 贝叶斯神经网络

相比传统神经网络，BNN 具有明显优势。一方面，通过引入权值的概率分布，BNN 可以看作是无数个神经网络的集成，这些集成模型的预测差异可以视为预测的不确定性。BNN 不确定性估计原理也可以用 Dropout 方法模拟近似^[52]，通过集成各个网络可以得到更可靠的预测。另一方面，研究证明，贝叶斯方法比最大似然估计更适合于小样

本数据建模。在贝叶斯方法中，参数的后验分布是通过先验和似然的乘积获得的，因此贝叶斯方法可以通过先验分布将先验信息包含到模型中。在样本数据较少的情况下，先验分布可以在计算后验分布时发挥着重要作用，使得贝叶斯方法在小样本情况下仍然能够很好地收敛^[53]。因此，通过在权重上设置先验分布，可以实现对权重的正则化，从而降低网络在小样本训练下的过拟合风险^[54]。

4.1.2 局部 BNN 与全局 BNN

局部使用贝叶斯神经网络（L-BNNs）和全局都使用贝叶斯神经网络（G-BNNs）是贝叶斯神经网络的两种不同的应用形式。局部使用贝叶斯神经网络通常是指仅在神经网络的某些层中使用贝叶斯神经网络，而其他层则采用传统的深度神经网络。这种方法不但可以预测该局部网络输出的不确定性，还可以帮助解决神经网络的过拟合问题，并提高神经网络的泛化能力。全局使用贝叶斯神经网络可以更好地建模模型的不确定性，并且在训练数据不足是仍然具有较好的预测性能。但是伴随着多次采样，全局 BNN 计算成本往往十分高昂，需要更长的训练时间和更多的算力资源，效率显著降低。并且在遇到具有较多层数的深度网络或者模块众多的复杂网络时，全局采用贝叶斯神经网络结构不但会增加训练难度，网络的可解释性也会进一步下降。由此，全局的贝叶斯神经网络

表 4-1 BNN 与全局 BNN 的区别

网络结构	权重参数形式	不确定性估计范围	计算效率	计算成本	实用性
G-BNNs	概率分布	全面	低	高	较差
L-BNNs	确定值 & 概率分布	局部	高	低	较好

结构更适合用在一些轻量网络从而获得网络的全局不确定性估计的能力，当涉及复杂的深度神经网络模型，由于网络认知不确定性的前向传播特性^[52]，使用局部的贝叶斯神经网络进行不确定估计可以在一定程度上替代对其进行全局估计。

4.2 用于视觉问答不确定估计的 BNNs

4.2.1 网络搭建

通常参考的是点估计神经网络中的感知机结构，将采用点估计的全连接层替换为 BNN 结构，通过计算输出的后验分布来预测分类结果以及不确定性。受 Zhijie Deng 等人^[55]将局部 BNN 采样结构用于网络对抗攻击检测的启发，在 Med-VQA 问题中，由于与输入图像以及问题相关词向量在经过训练后会具有某种语义关联，低维空间上表现

为具有较短的余弦距离。其输出的不确定性往往意味着输入也带有极大的不确定性，所以在语义空间中，不确定估计往往评估的是语义特征的分度，高不置信情况下网络提取到的语义特征往往也是无空间关联的，所以采用如4-2的思想重新设计一个可以预测 MEMSA 模型预测不确定性的网络结构。

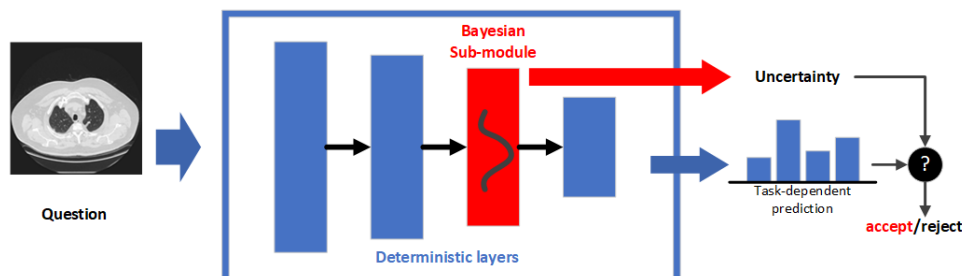


图 4-2 局部不确定估计的 BNN 模型结构

如4-3新的 MEMSA-BNNs 在原模型中最后用于预测的多层感知机网络替换成了多层全连接形式的贝叶斯神经网络结构，此处定义其为 BMLP（Bayes Multilayer Perceptron）其作用也是对 MEMSA 提取到的融合特征进行学习和分类。不同的是 BMLP 可以依据先验分布形式对融合特征进行蒙特卡洛采样，从而得到众多子网络，通过这些子网络预测出的结果以及后验分布形式估计预测的不确定性。

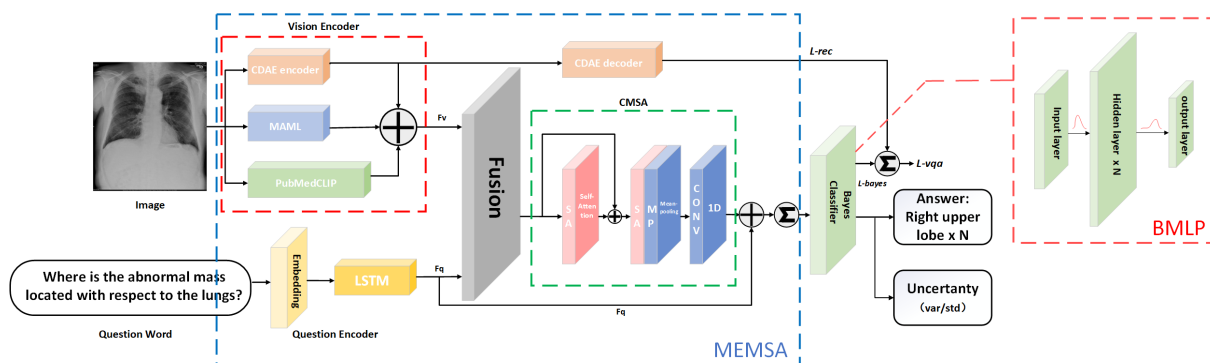


图 4-3 MEMSA-BMLP 模型

4.2.2 先验分布和后验分布选择

BNN 中先验分布和后验分布地选择往往直接影响模型的性能。先验的选择应该反映设计者对模型的认识以及对模型参数分布的事先预测，例如预测模型经过训练哪些参数更可能接近零，哪些参数可能是正数等等。先验分布应该尽量简单以避免过拟合，常用的分布有高斯分布或者拉普拉斯分布，针对实际问题，可以采用交叉验证或者网格搜索来评估不同的先验分布和后验分布，以选择最佳的组合。

在一般情况下, Blundell 等人^[51]建议将混合高斯分布 (Mixture of Gaussian Distribution) 作为权值的先验分布, 该混合高斯由两个高斯密度按照比例混合而成, 其中两个高斯密度的均值为 0, 但方差不一致。具体来说, 该混合高斯先验分布的表达式为:

$$P(w) = \prod_j \pi N(w_j | 0, \sigma_1^2) + (1 - \pi) N(w_j | 0, \sigma_2^2) \quad (4-1)$$

其中 w_j 是网络中的第 j 个权值, $N(w_j | 0, \sigma_1^2)$ 是 w_j 处均值为 μ , 标准差为 σ 的高斯密度。该种混合高斯的第一个混合成分的标准差要大于第二个成分的标准差 ($\sigma_1 > \sigma_2$), 从而提供比普通高斯先验更重的尾部。第二个混合成分的标准差要远小于 1, 从而使得许多权值落在零附近。所有的先验参数在所有权值之间是相同的, 这样可以避免要事先对这些权值进行先验优化。本文同样采用了这个混合高斯分布。其中, 第一个混合成分的标准差 σ_1 设置成了 0.1, 第二个混合成分的标准差 σ_2 设置成了 0.0001。比例权重系数 π 设置成了 0.5。本文尝试了各种先验参数的组合并通过实验效果得出了这一最优化参数组合。尝试的先验有: σ_1 (0.1、0.5), σ_2 (0.001、0.0005 和 0.0001) 以及 π (0.25、0.5 和 0.75) 虽然可以让先验分布参数跟随网络一起优化, 但一般无法提升性能, 反而容易导致训练收敛变得缓慢, 陷入局部最小值从而导致交差的性能。因此将该先验分布的参数设置为超参数的形式。

在本文中, 借鉴了 Blundell 等人^[51]的方法, 将权值的变分后验分布设置为高斯分布的对角矩阵, 其中的均值和标准差即为 BNN 需要优化的参数。相较于拥有相同网络结构的 CNN, BCNN 的需要优化的参数数量增加了一倍。

4.3 网络训练

4.3.1 贝叶斯反向传播算法

贝叶斯反向传播算法是一种使用贝叶斯推断来进行神经网络训练的方法, 与传统的反向传播算法相比, BBB 不仅可以优化神经网络中的权重还可以通过计算权重的后验分布来考虑权重的不确定性。为了得到 BNN 每个权值的分布, 需要在给定数据情况下计算权值的后验分布, 即 $P(w|D)$ 。然而, 权值的后验分布形式往往十分复杂, 其解析形式很难直接通过计算得到, 因此提出了很多近似方法来对权值的后验分布进行逼近。其中, Hinton 和 VanCamp^[56]以及 Graves^[57]提出通过变分近似来估计权值的真实后验分布。具体来说, 为神经网络的每个权值设定一个易于处理的变分后验分布 $q(w|\theta)$, 其中 θ 是变分参数。之后通过最小化 $P(w|D)$ 和 $q(w|\theta)$ 之间的 Kullback-Leibler (KL) 散度,

从而使得变分后验分布不断逼近真实后验分布。获得变分后验分布之后，BNN 的权值后验分布就可以替换为变分后验分布。 $P(w|D)$ 和 $q(w|\theta)$ 之间的 KL 散度公式为：

$$\begin{aligned}
 \text{KL}[q(w|\theta)||P(w|D)] &= \int q(w|\theta) \log \frac{q(w|\theta)}{P(w|D)} dw \\
 &= \int q(w|\theta) \log \frac{q(w|\theta)P(D)}{P(w)P(D|w)} dw \\
 &= \int q(w|\theta) \left[\log P(D) + \log \frac{q(w|\theta)}{P(w)} - \log P(D|w) \right] dw \\
 &= \log P(D) + \text{KL}[q(w|\theta)||P(w)] - \int q(w|\theta) \log P(D|w) dw
 \end{aligned} \tag{4-2}$$

因此，可以得到：

$$\begin{aligned}
 \log P(D) &= \text{KL}[q(w|\theta)||P(w|D)] - \text{KL}[q(w|\theta)||P(w)] \\
 &\quad + \int q(w|\theta) \log P(D|w) dw
 \end{aligned} \tag{4-3}$$

因为 $\log P(D)$ 是一个常数，为了最小化 $\text{KL}[q(w|\theta)||P(w|D)]$ ，需要最小化以下公式：

$$\mathcal{F}(D, \theta) = \text{KL}[q(w|\theta)||P(w)] - \int q(w|\theta) \log P(D|w) dw \tag{4-4}$$

$\mathcal{F}(D, \theta)$ 即为 BNN 的损失函数，通过最小化 $\mathcal{F}(D, \theta)$ 即可获得变分后验参数。该损失函数称为变分自由能^[58-59]或期望下限^[60]。 $\mathcal{F}(D, \theta)$ 的第一项被称为复杂损失项，他依赖于先验。第二项称为似然损失项，他依赖于数据 D 。由于损失函数 $\mathcal{F}(D, \theta)$ 是针对整个数据集的，但在神经网络中往往将数据集分成小批次进行训练。假设数据集 D 被分成 M 个大小相同的批次，那么对于每个批次 D_i ，复杂性损失项需要乘上加权值 $1/M$ ，此时损失函数变为^[51]：

$$\mathcal{F}(D_i, \theta) = \frac{1}{M} \text{KL}[q(w|\theta)||P(w)] - \int q(w|\theta) \log P(D_i|w) dw \tag{4-5}$$

这样，每个批次的损失之和 $\sum_i \mathcal{F}(D_i, \theta)$ 等于 $\mathcal{F}(D, \theta)$ 。此外，Blundell 等人^[51]还提出了另一种对复杂性损失项加权的方法：

$$\mathcal{F}(D_i, \theta) = \pi_i \text{KL}[q(w|\theta)||P(w)] - \int q(w|\theta) \log P(D_i|w) dw \tag{4-6}$$

其中加权值 π_i 为：

$$\pi_i = \frac{2^M - i}{2^M - 1} \tag{4-7}$$

可见 π_i 随着训练过程而减小。这种训练模式对于网络的初始训练阶段特别有用，因为在刚开始的几个训练批次中，加权值 π_i 最大，因此复杂性损失占主要作用，此时权值的变化是比较轻微的。随着训练数据越来越多， π_i 减小，此时数据对权值的更新的作用越来越来，而先验的作用越来越小。

对于神经网络，精确的最小化 $\mathcal{F}(D, \theta)$ 是不可行的，因此，Blundell 等人^[51]提出了贝叶斯反向传播算法，通过神经网络中的梯度下降来求解变分参数 θ 。首先，通过展开 $\text{KL}[q(\mathbf{w} | \theta) \| P(\mathbf{w} | D)]$ 这一项， $\mathcal{F}(D, \theta)$ 可以化解为：

$$\mathcal{F}(D, \theta) = \int q(\mathbf{w} | \theta) [\log q(\mathbf{w} | \theta) - \log P(\mathbf{w}) - \log P(D | \mathbf{w})] d\mathbf{w} \quad (4-8)$$

在中，通过蒙特卡洛采样方法来近似以上期望，即：

$$\mathcal{F}(D, \theta) \approx \sum_{i=1}^n \log q(\mathbf{w}^{(i)} | \theta) - \log P(\mathbf{w}^{(i)}) - \log P(D | \mathbf{w}^{(i)}) \quad (4-9)$$

其中 $\mathbf{w}^{(i)}$ 代表从变分后验 $q(\mathbf{w} | \theta)$ 采样的第 i 个蒙特卡洛样本。在 Blundell 提出的方法中，变分后验分布被设置为协方差矩阵为对角矩阵的高斯分布。令 μ 代表高斯分布的均值， σ 代表高斯分布的标准差。因为 σ 总是非负的，为了保证这一条件将 σ 参数化为：

$$\sigma = \log(1 + \exp(\rho)) \quad (4-10)$$

ρ 称为标准差参数。此时变分后验参数 $\theta = (\mu, \rho)$ 。由于从分布中采样是不可导的，这会导致神经网络无法进行反向传播。这一问题可以通过采用重参化^[61]技巧解决：

$$\mathbf{w} = \mu + \log(1 + \exp(\rho)) * \epsilon, \quad \epsilon \sim N(0, I) \quad (4-11)$$

其中 w 相当于变分分布中采样得到的权值， ϵ 是噪声变量，它从标准正态分布中采样得到。由于以上采样过程只涉及线性操作，因此该采样的过程是可导的。总的来说，BBB 中的重参化技术将随机性从权重中移除，使其可以直接计算梯度。这使得 BBB 能够高效地进行反向传播，并且可以在大型的神经网络中使用。由此，可以推导出贝叶斯反向传播算法的过程^[51]：

1. 从标准正态分布中采样噪声：

$$\epsilon \sim N(0, I) \quad (4-12)$$

2. 从变分后验中采样权值:

$$\mathbf{w} = \mu + \log(1 + \exp(\rho)) * \epsilon \quad (4-13)$$

3. 计算损失:

$$f(\mathbf{w}, \theta) = \log q(\mathbf{w} | \theta) \quad (4-14)$$

4. 计算损失关于均值的梯度:

$$\Delta_{\mu} = \frac{\partial f(\mathbf{w}, \theta)}{\partial \mathbf{w}} + \frac{\partial f(\mathbf{w}, \theta)}{\partial \mu} \quad (4-15)$$

5. 计算损失关于标准差参数的梯度:

$$\Delta_{\rho} = \frac{\partial f(\mathbf{w}, \theta)}{\partial \mathbf{w}} \frac{\epsilon}{1 + \exp(-\rho)} + \frac{\partial f(\mathbf{w}, \theta)}{\partial \rho} \quad (4-16)$$

6. 更新变分参数:

$$\begin{aligned} \mu &\leftarrow \mu - \alpha \Delta_{\mu} \\ \rho &\leftarrow \rho - \alpha \Delta_{\rho} \end{aligned} \quad (4-17)$$

其中 α 是学习率。以上步骤是一次蒙特卡洛采样训练过程。但为了更精准的估计公式(4-8)的积分, 在网络训练过程中往往会进行多次的蒙特卡洛采样, 计算出多个损失后再求平均。

4.3.2 网络预测

通过贝叶斯反向传播算法可以使得权值的变分后验分布不断逼近真实后验。当网络训练完成之后, 用得到的变分后验分布代替真实后验, 之后便能够对数据进行预测。BNN 对数据 x 的预测公式为:

$$\begin{aligned} P(y | x, D) &= \int P(y | x, \mathbf{w}) P(\mathbf{w} | D) d\mathbf{w} \\ &= \int P(y | x, \mathbf{w}) q(\mathbf{w} | \theta) d\mathbf{w} \end{aligned} \quad (4-18)$$

其中 $P(y|x, D)$ 是给定权值 \mathbf{w} 时网络的预测输出。这一积分相当于是对无数个模型的预测进行平均。但是在整个 \mathbf{w} 空间上进行积分一般是难以计算的, 因此这项积分通常采用蒙特卡洛方法进行估计, 此时 BNN 的预测公式为:

$$P(y | x, D) \approx \frac{1}{T} \sum_{t=1}^T P(y | x, w_t), \quad w_t \sim q(\mathbf{w} | \theta) \quad (4-19)$$

其中 w_t 是从 $q(w) | \theta$ 中采样得到的权值，采样的次数为 T 次，之后这 T 次预测的平均作为 BNN 的输出。这 T 个模型预测的差异可以视为预测的不确定性。可见 BNN 相当于多个模型的集成，这可以提高网络的泛化能力，减少过拟合的风险。

4.4 实验结果与分析

4.4.1 模型性能实验

贝叶斯在小样本分类下有着优异的性能，可以有效防止过拟合。以下是模型在两个不同量级数据集下将局部网络结构中替换成 BNNs 所取得的性能效果：

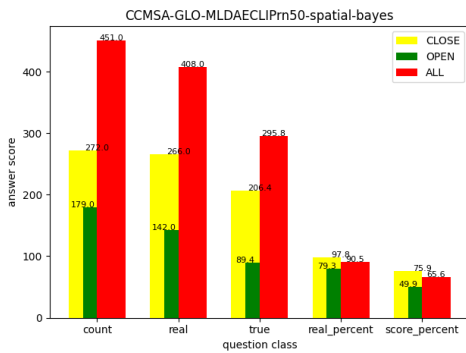


图 4-4 Med-RAD 数据集性能

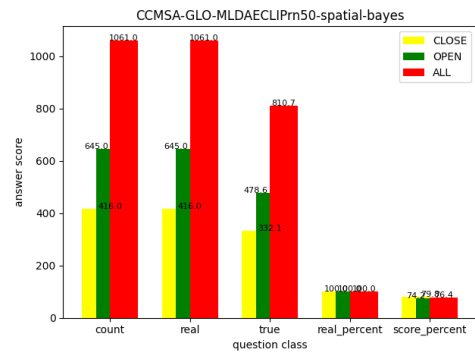


图 4-5 SLAKE 数据集性能

Dataset	Methods	Open	Closed	ALL
Med-RAD	Ours(no.bnns)	65.4%	77.6%	72.7%
	Ours(add.bnns)	56.4% ± 1.5%	73.9% ± 1.1%	66.7% ± 1.8%
SLAKE	Ours(no.bnns)	76.0%	81.7%	78.2%
	Ours(add.bnns)	73.5% ± 1.2%	81.7% ± 0.3%	76.7% ± 0.9%

通过上述图表可以发现，在更小的数据集上，加入贝叶斯分类器也仍然可以取得平均以上的性能，这说明贝叶斯分类器有防止模型过拟合的作用。

4.4.2 采样-不确定性实验

贝叶斯网络的不确定性预测其实本质也来自于多个模型集成的思想。经过对分布的采样获得多个模型预测结果的均值代表模型最终的预测，而方差则代表了其对于预测结果的不确定程度。方差越大，模型越不置信，而为了保证医学问答场景下的严谨和安全，医学问答模型能够获得对自身结果预测的不确定程度（讨论和 softmax 的区别）具有十分重要的意义。

由于贝叶斯神经网络是通过不断进行蒙特卡洛下采样来获取网络中的不确定性信息的，采样频率与不确定性估计间往往存在关联：从图??得见，随着采样频率的增

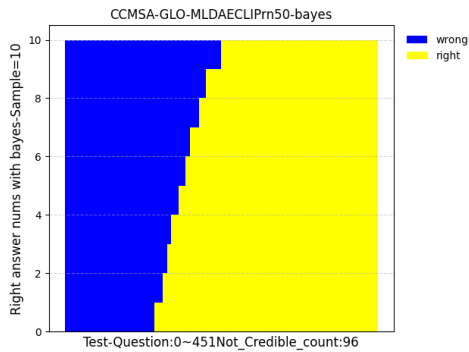


图 4-6 sample = 10

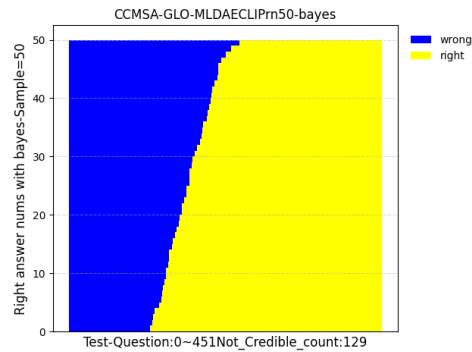


图 4-7 sample = 50

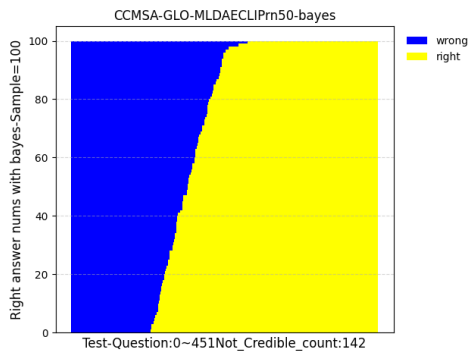


图 4-8 sample = 100

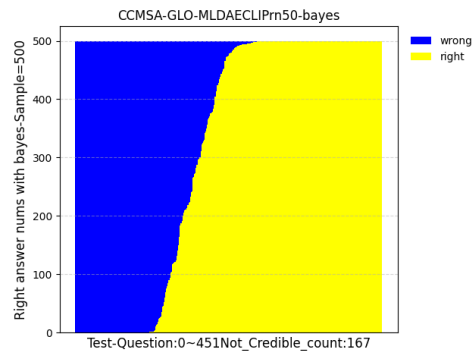


图 4-9 sample = 500

表 4-2 不同采样率下问答结果的不确定性

Dataset	采样频率	不确定性的问答数 U/总问答数 A	回答预测分布的方差
Med-RAD	10	0.21	2639.96
	50	0.25	2656.54
	100	0.30	2711.51
	500	0.31	2758.05
Slake	10	0.21	355.00
	50	0.24	396.82
	100	0.29	401.05
	500	0.32	398.77

大，模型捕获不确定性的能力增强，可以充分捕获目标样本点以及其周围邻域是否有充足的近似样本分布来支撑模型对其结果的确信程度以及消解对其预测的不确定性。在

表4-2中，可见由于 Med-RAD 数据集相比 SLAKE 数据集数据量较少，在分类空间上缺少“支撑样本”，从而导致整体预测的不确定性，也就是预测回答分布的方差较大，并随着采样频率的增大，这个数值也会增大。这一实验一定程度上解释了分类模型的不确定性和输入样本分布之间的关联，以及通过增大采样频率，可以增强模型预测不确定性的能力。但同时从图中也可以看出，由采样定理这种增强并不是线性增长，而是逐渐接近一个极限值。

4.4.3 拒绝分类实验

从理论上来说，由于预测不确定是贝叶斯神经网络最突出的优点之一，预测不确定性越高，说明网络对预测结果的把握程度越低，所以对于不确定性过高的预测，网络往往可以通过拒绝分类的形式，即不给出分类结果，从而提高模型分类的准确性和可靠程度。在实际的医学问答等具有风险性的情景中，当面对具有高不确定的诊断场合时，医生也会拒绝给出确切的答复，并通过其他手段继续收集信息以降低不确定性后再给出确诊意见，从而防止漏诊、误诊以及误治等具有极大事故风险的情况出现。

为了验证拒绝分类这一想法的可行性，本文利用用于分类预测的 BNNs 探究拒绝不确定性高（ $\geq 50\%$ ）的分类是否能够提高模型整体的准确率。在实验前，首先需要一种计算样本预测不确定性的方法。由于 BNN 的权值为变分的形式，所以其预测输出也是一种变分分布。Kwon 等人^[62]和 Shridhar 等人^[63]提出通过计算变分预测分布 $q(y^* | x^*)$ 的方差来度量预测不确定性：

$$\begin{aligned} \text{Var}_{q(y^*|x^*)}(y^*) &= \int [\text{diag}\{E_{p(y^*|x^*,w)}(y^*)\} - E_{p(y^*|x^*,w)}(y^*)^{\otimes 2}] q(w) dw \\ &+ \int \{E_{p(y^*|x^*,w)}(y^*) - E_{q(y^*|x^*)}(y^*)\}^{\otimes 2} q(w) dw \end{aligned} \quad (4-20)$$

其中 $v^{\otimes 2} = vv^T$, x^* 和 y^* 分别是测试样本以及其预测输出， $p(y^* | x^*, w)$ 表示给定 w 下网络的预测， $q(w)$ 表示权值的变分分布， $\text{diag}(v)$ 是以向量 v 为对角线元素的对角矩阵。公式通常难以精确计算，但可以通过蒙特卡洛方法进行近似：

$$\text{Var}_{q(y^*|x^*)}(y^*) = \frac{1}{T} \sum_{t=1}^T \text{diag}(\hat{p}_t) - \hat{p}_t \hat{p}_t^T + \frac{1}{T} \sum_{t=1}^T (\hat{p}_t - \bar{p})(\hat{p}_t - \bar{p})^T \quad (4-21)$$

其中 T 为蒙特卡洛采样次数， \hat{p}_t 代表第 t 次采样的网络输出的预测概率，而 \bar{p} 是所有 \hat{p}_t 的均值，即 $\bar{p} = \sum_{t=1}^T \hat{p}_t / T$ 。在 Kwon 等人^[62]的推导中，公式右边第一项为偶然不确定性，第二项为认知不确定性。偶然不确定性描述的观测数据中固有的噪声，因此也被称为数据不确定性。这是一项无法消除和避免的误差，即时收集更多的数据，也无法降低

偶然不确定性。认知不确定性描述的训练模型中的不确定性，通常也被称之为模型不确定性。认知不确定性反映了模型缺乏模型知识。例如，对于来源于数据稀疏区域或者远离训练集的测试点，模型的认知不确定性就会显著增加^[64]。通过收集更多的数据训练模型，可以减少认知不确定性。Shridhar 等人^[63]的实验表明，分类准确性和认知不确定性之间呈负相关的关系，即分类准确性随着认知不确定性降低而提高。Shridhar 等人^[63]还证明偶然不确定性取决于数据集而不是模型，同一个数据集在不同模型下具有相同的偶然不确定性。因此，本文将认知不确定性作为衡量预测不确定性高低的指标。本文设计的拒绝分类实验过程如下：测试时，首先对模型进行 10 次蒙特卡洛采样，计算所有样本的认知不确定性。之后讲所有版本按照认知不确定性的大小进行排序，并按照不同的拒绝比（例如 10%、10%、10%）将不确定性高的样本从舍弃，相当于 BNNs 拒绝对这些样本进行分类，之后重新计算拒绝分类后的准确率。

表4-3显示了模型按不同比例拒绝高认知不确定样本后的问答准确率。当拒绝比例为 $n\%$ 时，代表 $n\%$ 认知不确定性值最高的样本被拒绝分类。可以看到，当网络拒绝分类部分预测不确定性高的样本之后，在开放式问答（Open）和封闭式问答（Close）上的回答准确率均有提升；可以看到，当以百分之 50 的拒绝比例拒绝高认知不确定性样本时，网络在 Med-RAD、Slake 这两个数据集上的总问答准确率分别提高到了 69.8% 和 82.1%，接近甚至超越了在没有引入不确定估计时的模型的预测水平，在该数据集上也是一个相当高的准确率。且对于数据不确定性较高的小样本数据集 Med-RAD，拒绝比例越大，准确率的增长率越大，提升幅度也越明显。以上结果说明，被拒绝分类的样本中大部分为容易分类错误的样本，同时也解释了拒绝分类可以升 Med-VQA 模型问答性能。

综上所述，认知不确定性可以在一定程度上反映样本被错分类的可能性，认知不确定性高的样本中，错分类的样本占多数，说明模型对容易错分类的样本会具有较大的认知不确定性。因此，贝叶斯神经网络模型可以通过拒绝认知不确定性高的样本进行分类来提高模型的准确性，同时尽可能避免了错分类的产生，可以说 BNNs 的加入提高了整个医学视觉问答系统的性能和可靠性。

4.4.4 带不确定性估计的问答样例


当设定 BMLP 的采样频率（sample）取 10 的时候，模型不确定性估计样例如图4-10所示，针对 Q1 提问，当所有的采样子网络预测结果都指向一个答案标签（Right upper lobe）时，依据上文提到的贝叶斯不确定性估计相关原理^[51]，此估计方差和标准

表 4-3 拒绝分类实验

Dataset	拒绝比例	Open	Close	Overall	Growth rate	U/A
Med-RAD	0 %	49.4 %	75.1 %	64.9 %	X	0.215
	10 %	49.4 %	76.2 %	65.7 %	1.2 %	0.197
	25 %	52.4 %	76.0 %	67.0 %	1.9 %	0.185
	50 %	52.1 %	80.3 %	69.8 %	4.1 %	0.121
Slake	0 %	75.1 %	80.9 %	77.5 %	X	0.172
	10 %	78.3 %	81.7 %	79.6 %	2.7 %	0.145
	25 %	78.5 %	83.2 %	80.4 %	1.0 %	0.121
	50 %	80.2 %	85.0 %	82.1 %	2.1 %	0.081

差为 0 我们认为这个答案是具有低不确定性的，具有更好的可靠性。

Sample = 10



Q: In which lobe do you see an abnormal mass in the image?

All_answers: Right upper lobe、Right upper lobe、Right upper lobe、Right upper lobe、Right upper lobe、Right upper lobe、Right upper lobe、Right upper lobe、Right upper lobe、Right upper lobe

(Low Uncertainty)

True answer: Right upper lobe

Q: Size of the mass in the right upper quadrant?

All_answers: 3.4cm,3.4cm,3.4cm,no,emphysam,emphysam,no,no,ct-ratio,no

(High Uncertainty)

True answer: 3.4cm

图 4-10 BMPL 采样 10 次时带不确定性问答样例

对于 Q2 提问，不同子网络给出了不同的答案，而且呈现零散分布的态势，具有极高的方差和标准差，所以是具有高不确定性的答案。即时其中包含了正确的答案，但我们通过不确定性预测认为这个答案是不可靠的，因为按先验分布对样本点以及其周围邻域采样时会出现影响结果的样本点，说明该预测是一种在超平面上，也就是对边界样本的预测，模型对该类型预测往往具有极高错分类的概率。可见，基于 BNNs 搭建的 BMLP 可以知晓自己预测时的样本分布情况，从而预测出其结果的不确定性，这种预测优势是传统点估计网络所无法具备的。BMLP 也模仿了人在事前对信息（样本分布以及

先验) 掌握不充分的情况下做出预测时会给出的一种不确定, 不肯定评估, 这种不确定性评估虽然不具备信息增强的能力, 但可以帮助我们在某些场合进行预测时规避掉重大的风险, 从而步步为营, 以较高质量完成后续任务以及目标。

4.5 本章小结

本章首先介绍了贝叶斯神经网络的原理、结构以及该结构和不确定估计之间的关系, 然后对网络中局部使用贝叶斯神经网络和全局使用贝叶斯神经网络进行了简单的阐述和定性分析, 接着网络搭建部分, 先后介绍了用于医学视觉问答网络中的 BNNs 结构以及 BNN 中先验和后验的选择依据和技巧, 紧接着网络训练部分详解了用于 BNN 训练的贝叶斯反向传播算法以及网络如何进行输出和预测以及与不确定性的关系, 介绍了模型的参数构成与复杂度。最后通过模型问答性能实验、采样与不确定性的关系实验和拒绝分类实验验证了 BNNs 相比传统点估计网络所具有的防止过拟合、获得预测分布以及不确定性以及通过拒绝分类的形式可以提升模型性能以及防止错分类等等优势。

综上, 本章验证了通过将贝叶斯神经网络引入具有风险性的医学视觉问答场景不但可以提高模型性能、提高模型在缺少数据样本时的鲁棒性和防止过拟合, 还可以通过不确定估计和拒绝分类的形式, 极大地提高了模型甚至医学视觉问答系统的可靠性, 具备相当的实用价值。

第五章 模态自适应医学视觉问答系统实现与在线部署

为了提高医学视觉问答技术的实用性，本章设计、实现并成功在云端部署了一个提供医学视觉问答服务的在线系统，在第三，第四章节的模型和算法的基础上，结合机器学习、云计算技术、模型部署和服务化技术、安全和隐私保护技术、自动化运维技术以及图像上传、机器翻译等组件/API，能够根据用户拍摄或者上传的图片以及问题，系统自适应地选择与图像问题相匹配的交互模型，给出该模型对应的预测答案。该系统可以提供稳定、便捷和高效的云端视觉问答服务，并且可以实现在线更新，很好地丰富了人机交互方式。本章内容包括模态自适应系统设计，在线云服务系统设计两部分内容。

5.1 模态自适应系统

5.1.1 技术路线

模态自适应系统可以在面临不同输入的复杂场景下，自适应地选择合适的模型与用户完成交互，其主要实现原理是在网络中加入模型控制单元，通过输出反馈和学习算法对输入和系统状态关系进行建模并训练相应的控制模型，让控制模型自主学习合适的交互策略。同时也能在这一闭环内建立完整的对话样本采集学习的内环回路，在实现模型自适应交互的基础上还可以完成对话样本的高效收集。如图5-7, 在传统的对话路线之

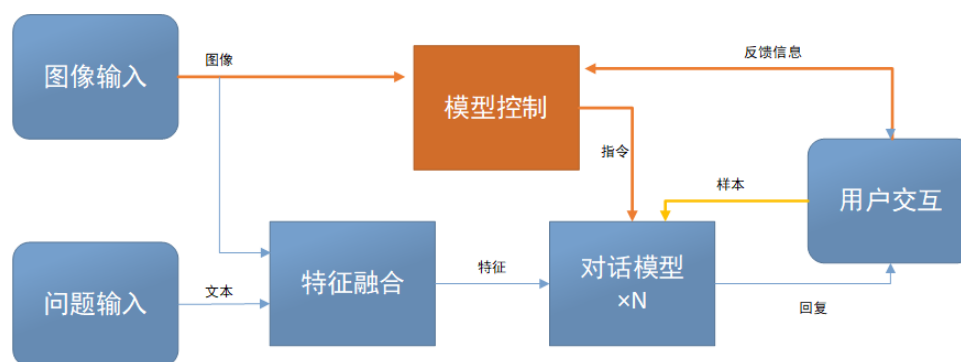


图 5-1 模态自适应 Med-VQA 系统实现框图

外，新增加了对图像输入-对话模型之间的关联控制，通过一些轻量化的机器学习算法或经典的分类网络在不同类型的图像输入和相应的对话模型之间建立关联，通过用户对话反馈以及输出的不确定性设计反馈逻辑获得可用于训练的对话决策样本，从而获得在面临不同医学图像输入（例如放射学、病理学）模型可以自主识别用户的输入类型并选择合适的对话策略。这个设计思想和“专人专攻”的思想相近似，其优势是不用重新设计和训练可以处理多种不同数据特征的多模态“大模型”，这种方式不仅代价高昂并且

由于样本之间的互相干扰，模型很难学习到合适的特征对其进行划分，往往也会导致各类型任务的分类性能下降。另一方面，也避免了必须独立使用具有相似功能的模型的情况，提高了系统功能的集成化。所以模态自适应技术可以在一个独立系统内将用于处理各个模态的轻量化模型进行高效集成，从而在保证各个子系统原本性能的同时，解放了输入数据的限制，提高了模型的泛化能力和交互体验。

5.1.2 设计原理

在多模型交互的场景下，可以基于规则、检索等传统方式选择合适的交互模型，但这种模型是固定的逻辑，存在泛化能力弱不具备自主调节能力等弊端。如图5-2基于机器学习方法搭建的模型控制器可以根据用户或者系统的反馈，实现一个可自适应的对话模型控制器，可根据用户的输入和反馈将问题分类到合适的模型中进行处理。而多模型的集成（例如投票、加权）也可以提高系统的准确性和鲁棒性，以及根据用户的反馈，不断对模型进行调整和优化，以提高对话系统的性能和用户体验。

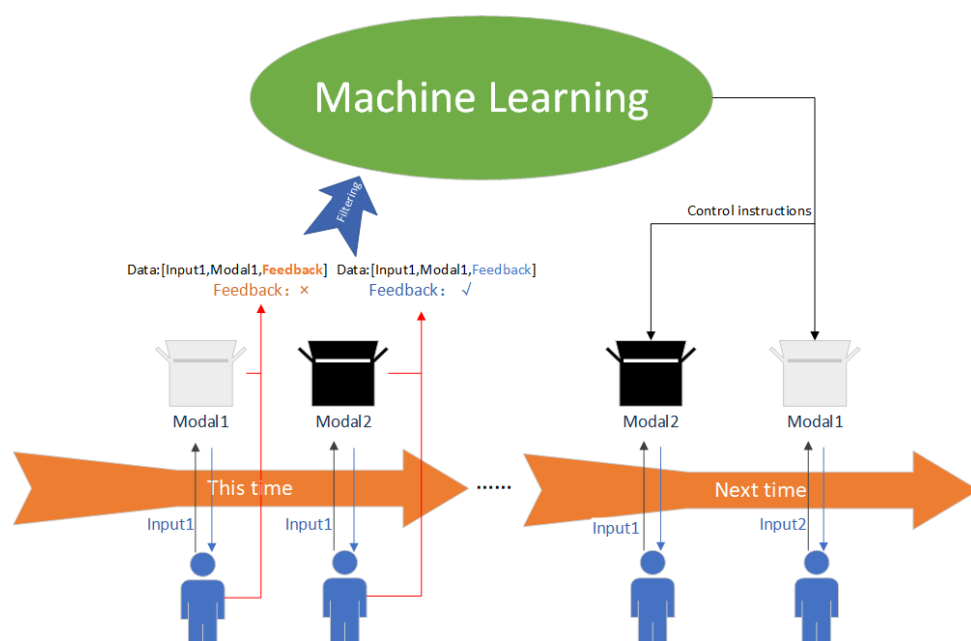


图 5-2 自适应交互逻辑

在现有的公开医学视觉问答数据集中，主要是放射学和病理学两个类别的数据，针对每个类型的数据和数据集都可以训练出一个性能较为优秀的模型，模态自适应系统通过多模型选择、用户反馈、模型策略优化、模型集成和迭代等步骤逐渐获得最优的效果，即时是面对复杂抽象的输入，只要获得正确的反馈并加以迭代，模型也能够对其进行识别和适应。

5.1.3 算法评估

基于现有的医学问答场景问答诉求，主要的图像输入如??为放射学图像以及病理学图像，故综合对比了目前常见的机器学习图像分类算法以及系统响应性能，准确率高，系统响应快，具有实时调整能力的机器学习算法模型往往是使用在即时对话系统中进行模型控制的较好选择。为了较好的初始化这个自适应模型，选用了 Med-RAD 中的所有放射学图像 315 张以及 Path 数据集中的前 315 张病理学图像，总计 630 张组成一个新的数据集并按 8:2 的比例划分训练集和测试集，其所在的数据集作为每张图片的分类标签，选用不同的分类模型进行分类。



图 5-3 头部 MRI 放射学影像



图 5-4 胸部 X-ray 放射学影像



图 5-5 胸部 CT 放射学影像

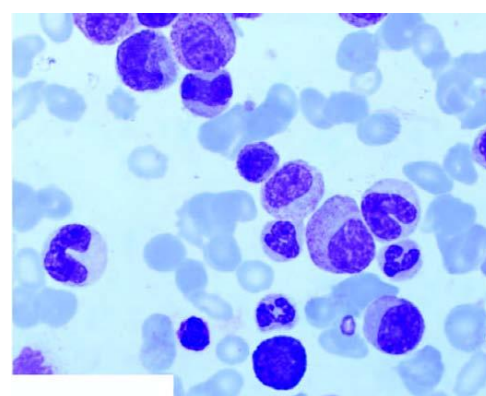


图 5-6 病理学影像

最终在测试集上的表现如表5-1。经过对比，随机森林算法具有最好的准确率、精确率和 F1 值，且在此类无正负例倾向的分类场景中，准确率和 F1 分数为最主要的参考指标，且随机森林方法在防止过拟合、鲁棒性和可解释性上相比其他算法君有一定的优

表 5-1 不同机器学习算法的性能评估对比

方法	TN	FP	FN	TP	Acc	Precision	Recall	F1
朴素贝叶斯	48	7	10	61	0.865	0.897	0.859	0.876
决策树 (avg)	54	1	4	69	0.960	0.985	0.943	0.964
感知机	51	4	0	71	0.968	0.947	1.0	0.973
逻辑回归	51	4	0	71	0.968	0.946	1.0	0.972
支持向量机	52	3	0	71	0.976	0.959	1.0	0.979
随机森林	55	0	2	69	0.984	1.0	0.972	0.986

势，故而选择随机森林算法作为模型自适应控制的学习算法。

5.1.4 分析

综上所述，该模态自适应系统设计可以自适应地不同医学影像输入间进行转换和交互。它可以使得机器学习模型能够在多种输入模态下进行训练和推断，从而提高了模型的鲁棒性和实用性。但同时，这个系统本身也存在一定的不足之处。例如，接受反馈以及模型调整存在明显的滞后效应，单个用户往往不会询问一个问题多次。其次就是反馈的数据并不能直接用于训练，还需要专业的医生进行筛选甄别，这无形中会增大其成本。并且要想达到很好的跨模态效果，还需要大量相关数据的支持，这也是医学领域所匮乏的。

5.2 云端在线系统设计

5.2.1 功能性需求分析

一个视觉问答系统的核心功能是协助用户完成相关的视觉问答任务，通常包括系统管理、信息处理、问答模型三个部分。系统管理主要负责系统日志控制，系统内外部的信息通信；信息处理负责对系统输入的图像或者文本信息以及数据流进行预处理，转换成问答模型所需要的信息格式；问答模型是本系统的核心模块，包括模型控制、模型问答两部分。系统的总体功能需求如图5-7。

视觉问答系统中的主要角色包括用户和视觉问答模型，以下通过图例来描述他们所涉及的各个功能模块。

- (1) 用户用例5-8：对于使用本系统的用户，系统主要负责处理用户的信息采集以及问答请求，其中信息采集包括用户上传的医学影像图片以及相对应的问题，其

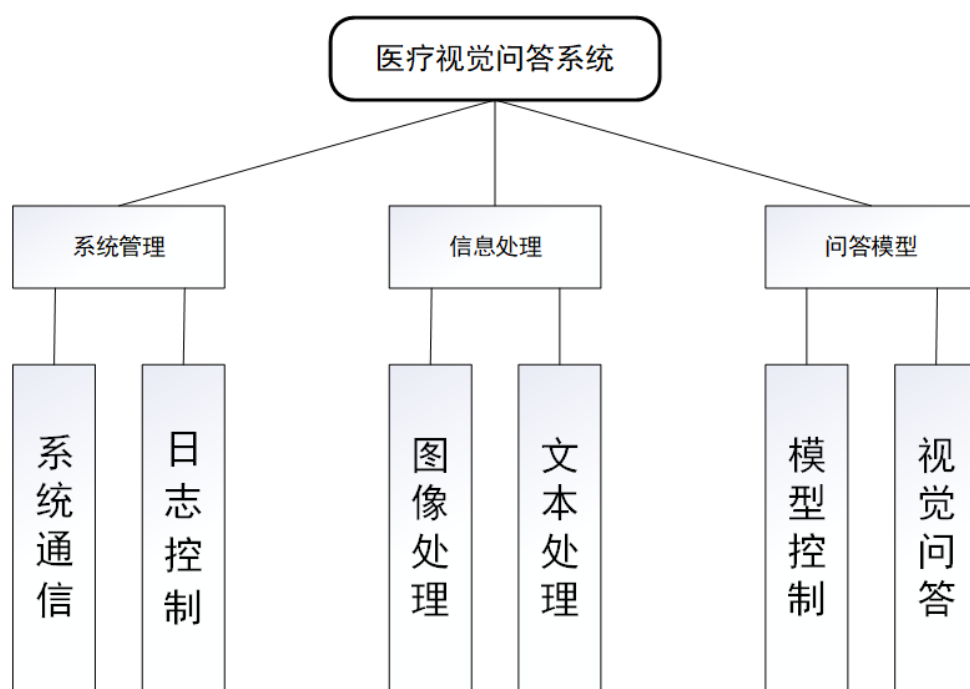


图 5-7 需求分析

中这两个输入都要做到良好的兼容性，图像采集需做到兼容常见的图像格式和资源形式，例如兼容 png、jpg 等格式的图片以及用户采用的是本地上传或者是网络资源定位符 url 的方式提供图像信息。问题文本采集应充分考虑语种，语言类型的差异，对不同语种问题的输入，接口需具备基础的翻译能力。同时在处理问答请求是要区分是控制请求还是服务请求，控制请求分为 API 调用或者前端调用两种不同的服务请求，确认后系统才能够提供准确的返回形式。服务请求则需要系统调用视觉问答模型对问题输入进行回复，为用户提供服务。

- (2) 模型用例如 5-9：对于搭载在系统上视觉问答模型，系统主要通过各种接口为其实现机器翻译，答案返回（云端下发）以及样本收集等功能。在面对非英语输入时，系统需要调用翻译组件进行翻译，最后交由模型进行问答，当模型返回答案预测时，系统会将其整理成带标签的字典，并打包成 JSON 信息下发给用户。在面对需要收集的样本时，系统先将数据按样本和标签进行分类整理，并调用本地 OS 将其保存在数据文件夹或者数据库文件中。

5.2.2 非功能性需求分析

非功能需求是所述系统在满足完成主体功能的情况下，系统还需要有其他的功能来满足系统运行时的实用性和可靠性。视觉问答系统的非功能性需求主要包括高可用性和高响应性。如果是处于生产环境部署的情况下，同时考虑到其安全性和可扩展性。

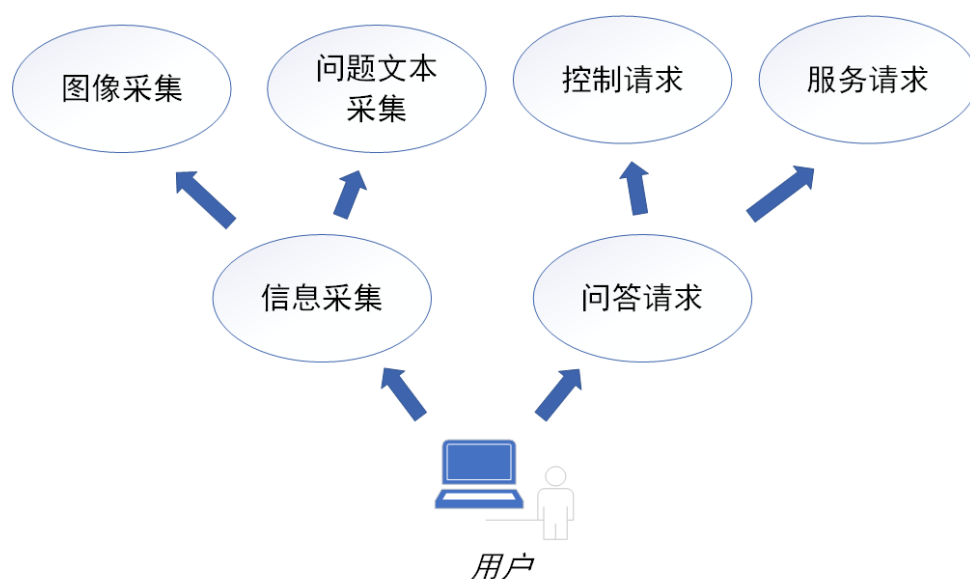


图 5-8 用户用例

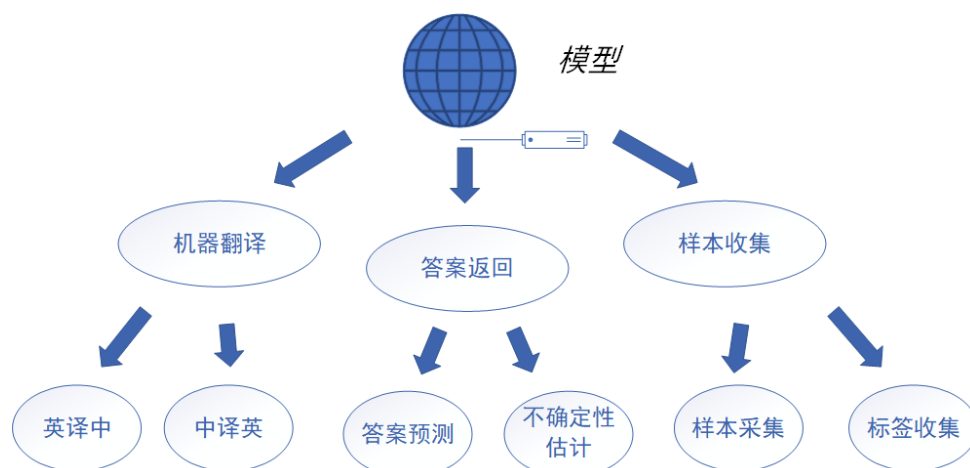


图 5-9 模型用例

高可用性又包涵高可靠性、灵活性以及可视化等几个特点。高可靠性是指模型需要在面对各种异常情况（如网络延迟、服务器故障等）时保证系统的稳定性；灵活性是指模型具备灵活部署、维护和升级的能力，以适应不同业务场景的需求；可视化则要求模型具备可视化的特点，能够为用户提供直观，清晰的结果展示，以使用户更好的使用和理解模型。由于视觉问答系统内部模型复杂，集成算法较多，出现故障的可能性极高。为此采用多机分布式部署的方案来保证系统的可靠性和高可用性。

高响应性是指系统对于用户操作或者请求的快速响应能力，包括用户界面的快速响应，数据的快速检索和处理等。一个实时性要求比较高的模型，部署时需要考虑到响应性的问题，以保证良好的用户反馈和用户体验。而由于视觉问答模型中的特征提取和答案推理过程都需要极大的算力支撑，并且往往较为费时。因此为了减少用户的整个问答

过程的响应时间，考虑使用缓存数据的方式来减少数据读取和存储的时间，同时采用消息队列技术吗，可以实现削峰、解耦以及异步处理系统其他功能（例如日志存储）防止信道阻塞，避免其增加核心问答功能的响应时间。

5.2.3 总体架构

总体设计分为服务端和客户端两大模块，服务端组织构成上包含后端服务器和代理服务器，开放 API 访问接口以及相关技术文档提供服务，由服务端主机在内网发布，再通过内网穿透 + 代理服务器开放到公网提供外部访问，客户端则是一些前端交互实现，总体架构如图所示。客户端：基于 Web 网页端实现，用户可以通过访问网页（前端界

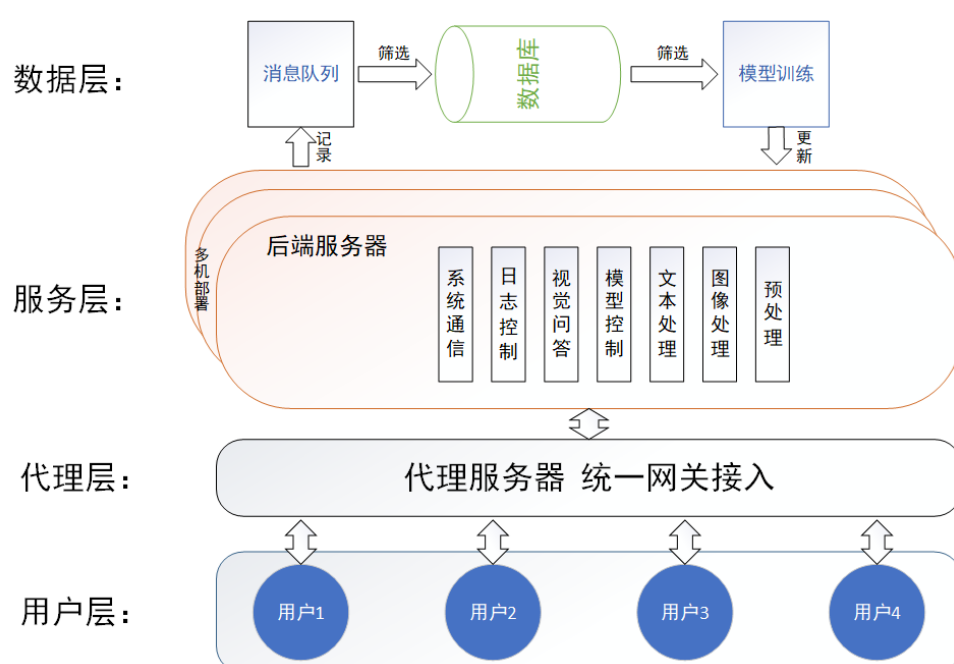


图 5-10 总体框架

面）上传图片信息和问题信息来和视觉问答系统进行交互。网页接收信息流后向后端服务 API 发起访问，并接收返回的回答内容并转换成文本或者语音的形式呈现给用户。

代理服务器：面对公网开放服务时，代理服务器是客户端到后端服务器的连接桥梁，代理服务器具有路由映射组件。通过路由映射表将来自于公网的服务请求路由到后端服务器，访问服务器上的对应接口，并在获得服务 API 的接口返回数据后再发送回请求方。使用代理服务器有许多优势，一是避免了服务器主机直接开放到公网上，防止服务被劫持。二是可以有效管理访问流量，代理服务器的负载均衡组件可以让客户端的请求流量按策略请求分发到后端服务器，一定程度上避免了流量过大时服务器会出现的宕机情况。同时，代理服务器具备的心跳检测组件可以发现故障的后端服务器，如果内网

中仍有其他服务器可以正常提供服务，代理服务器可以将后续的请求路由到正常运行的服务器上，提高了服务的稳定性。

后端服务器：后端服务可以采用单机部署，为了提高服务的效率和稳定性，往往也采用多机部署的方式。主要包括系统管理模块、信息处理模块视觉问答模块。系统管理模块提供用户注册登录等接口，信息处理模块提供视觉问答数据需要的转化接口，视觉问答模块则搭载了主要的视觉问答模型，接收数据并预测结果返回。

5.2.4 工作数据流

如图5-11所示，为了更好地理解整个架构和数据流转方式，设计的模态自适应医学视觉问答系统的数据 workflows 主要包含了用户请求、问答信息、问答记录三个部分。要实现分别记录不同用户的问答信息并保存在文件中作为数据反馈，需要一个高效的数据 workflows 处理方式。用户数据是系统接口收集到的用户信息和登录信息，问答记录数据记录

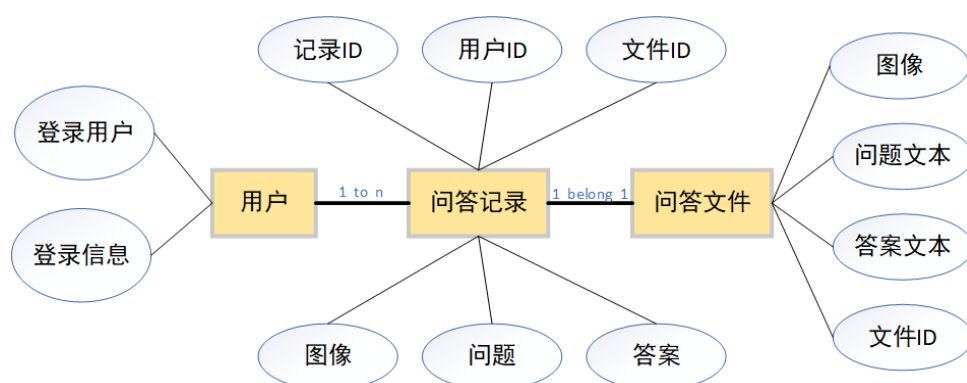


图 5-11 工作数据流图示

了和用户的交互以及反馈信息，包含三种 ID 数据和三类问答数据，一个用户包含了许多的问答记录，所以是 1 to n 的关系。问答文件是问答记录的实际载体，一个问答文件对应一次问答记录，以便于在系统中进行数据整理。

5.2.5 后端 API 服务接口设计

设计一个合理的 API 接口往往包含了开发语言和框架的选择、输入输出参数和格式的确定、编写接口功能、部署服务、测试、文档化、发布等过程。

服务器端主要通过 HTTP/HTTPS 协议为客户端提供视觉问答接口。其中信息处理模块集成在了接口上，接口的通信格式为 json，通信内容如表5-2。视觉问答接口是实现系统问答和提供服务的核心，在内部调用了各个算法服务，包括视觉问答算法服务，模型请求和信息处理中的各种算法。客户端通过 HTTP 经代理发送请求给后端服务器，

表 5-2 后端 API 接口通信格式

字段名	类型	描述
code	int	状态码
message	String	消息
data	json	数据

并携带想要提问的图像和文本问题，后端服务器将最终的回答返回给客户端。

5.3 详细设计与实现

5.3.1 系统架构

如表5-3, 要搭建实现一个在线医学视觉问答系统，并保证其服务的稳定性、可靠性、闭环性以及安全性，需要一个较好的集成化设计，框架分用户层、代理层、服务层、数据层进行了设计。

表 5-3 主要使用的服务框架

用户层	代理层	服务层	数据层
Gradio	Coplar	Docker	Flask

- (1) Gradio 是一个基于 Web 的交互式界面构建框架，可以用于构建机器学习模型的演示和应用。使用 Gradio 可以将模型部署为具有自定义用户界面的 Web 应用程序。
- (2) Coplar 框架是一个用于提供内网穿透服务的应用框架，用于解决在内网环境下的外部访问问题，例如在企业内部网络中，往往存在一些需要从外网访问的资源，例如 Web 服务、数据库等，但这些资源由于被部署在内网中，因此无法直接通过公网 IP 进行访问。Coplar 内网穿透服务可以将内网资源映射到公网上，从而实现对内网资源的远程访问。
- (3) Docker 是一种容器化平台和服务，可以帮助开发人员和系统管理员在虚拟化环境中轻松地构建、部署和运行应用程序。Docker 容器还提供了一个隔离环境，使得应用程序可以在一个独立的运行环境中运行，不受其他应用程序或系统资源的影响。
- (4) Flask 框架的核心是一个 WSGI(Web Server Gateway Interface) 应用程序，其实现

非常简单，一般只有几个 Python 文件，其中包括了应用程序、路由、视图函数、模板、静态文件等。Flask 通过装饰器机制实现路由映射，使开发人员可以通过定义路由和视图函数的方式，轻松地实现各种 HTTP 请求的响应处理。有利于收集用户数据，建立数据库和服务之间的快速通路。

5.3.2 系统管理

系统管理主要是在数据以及通信层管理包括控制模型调用、访问控制、日志记录等系统功能。并通过编程和逻辑设计解决一系列部署中会遇到的问题。例如如何调配资源和模型对处理访问申请，如何在用户访问过程中进行并行的日志记录，如何监控异常行为，返回错误信息等。

5.3.3 信息处理

信息处理单元一般都集成在接口处，处理包括读取用户请求输入的图像数据和文本数据。由于在线系统的图片输入一般会分成互联网统一资源定位符 (url) 或者本地上传形式，所以都需要先将其统一读取成二进制形式交由模型处理。文本作为一个字符串，直接由模型读取并转化成响应的 token。以下是系统中会遇到的信息以及其格式和处理方法。

表 5-4 接口信息处理

信息内容	接口信息处理操作
图像操作	按 84×84 进行裁剪 (AE)
	按 128×128 进行裁剪 (MAML)
	按 250×250 进行裁剪 (CLIP)
文本操作	语种识别 (langdetect)
	语言翻译 (API)

5.3.4 视觉问答

视觉问答又是模型服务端，也就是模型的预测函数。通常由模型结构 + 加预训练好的参数构成，在调用问答服务时，模型会处理接口端传输来的图像和文本信息，然后预测输出并返回；通常为了保证深度学习模型和其环境的稳定性，主流上都会使用 Docker 组件对模型以及所需要的环境进行封装并打包到第一个可移植的容器中。以便在任何地方都可以轻松部署。Docker 容器是镜像运行的实例。负责管理容器的创建、启动、停止

等操作。视觉问答接口请求如下表：

表 5-5 视觉问答接口

接口名	视觉问答
请求协议	HTTP\HTTPS
请求方法	Post
请求 url	\post
携带参数	图像 url、文本
返回结果	文字和不确定估计结果

5.3.5 前端界面

在前端的编写使用中最注重的就是可视化。**Gradio** 是一个轻量级的工具，主要用于快速编写可视化的 **Demo**，搭建模型的 **Demo** 并部署深度学习模型。但其本身并不具备模型管理功能，必须采用其他方式管理模型。因而后端为了与前端有更好的区分，采用了 **Flask** 框架编写。

如图5-12和图5-13, 前端页面设计采用了 **VQA—Demo** 最常用的双流输入-输出的方式，也就是一侧负责模型输入，一侧负责模型输出。这种设计更具有和模型对话的效果，可以提高用户的交互体验。

如5-14, 为了更好地收集用户反馈，在输入栏下方设计了反馈栏，这样便于用户同时将反馈信息输入给模型。此外，如图5-15为了更方便用户进行体验，**Demo** 还在最后设置了三类输入样例、8 组对话样例，用户在进入界面后可以直接上手使用。

5.4 系统接口测试

本小节使用最常见的接口测试工具 **Postman** 来测试问答接口，测试时请求需要携带的参数为图像信息和问题文本，请求遵守使用 **HTTP** 协议，请求方式为 **POST**，测试 **URL** 路径为 **/post**，请求格式 **JSON** 例子如图5-16。测试的返回结果为文本。经过测试返回结果如图5-17，从图中可以看出，本次测试成功从接口中获得了预期结果的返回，返回的 **json** 数据中的 **data** 字段包含了模型回答、模型不确定性预测、结果方差以及不确定标签评估。

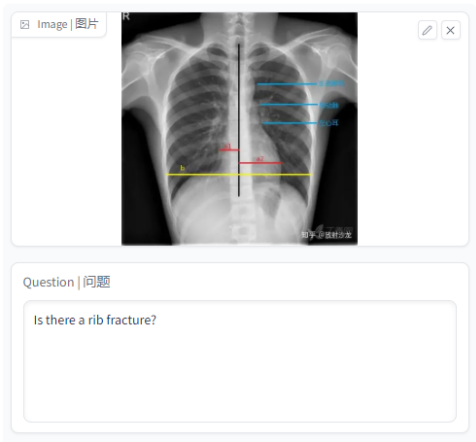


图 5-12 Demo 输入栏

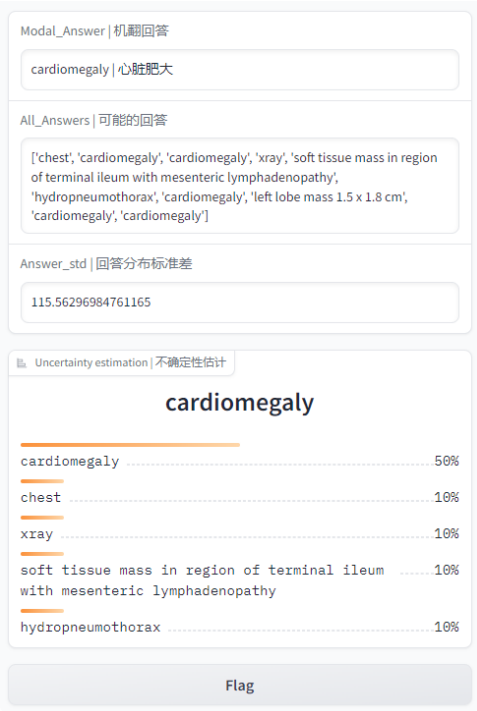


图 5-13 Demo 输出栏

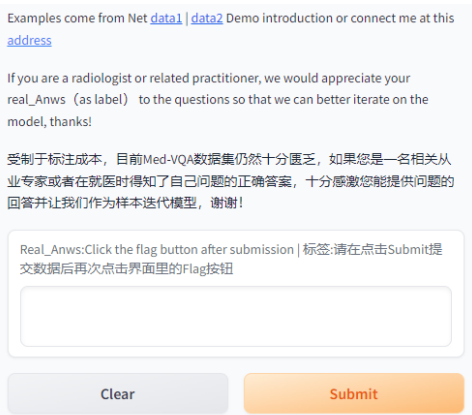


图 5-14 Demo 反馈栏

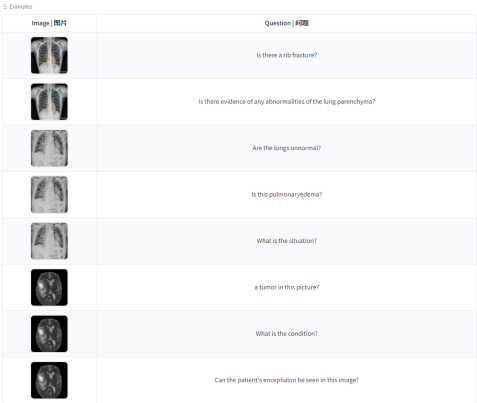


图 5-15 Demo 样例栏

5.5 本章小结

本章首先介绍了模态自适应系统的技术路线和设计原理，接着针对目前场景在评估了各个经典机器学习算法的性能，并选用最优的算法作为控制交互模型的自适应方法，紧接着又介绍了云端在线系统的设计过程，从分析用户功能需求的角度确定系统需要实现的服务目标并以此设计出在线系统框架，从而设计数据工作流和相应后端服务支持。然后分别从系统架构、系统管理、信息处理、视觉问答以及前端页面五个部分介绍了从设计到实现的过程。最后，通过网络测试工具对模型服务进行了在线测试，得到了正确

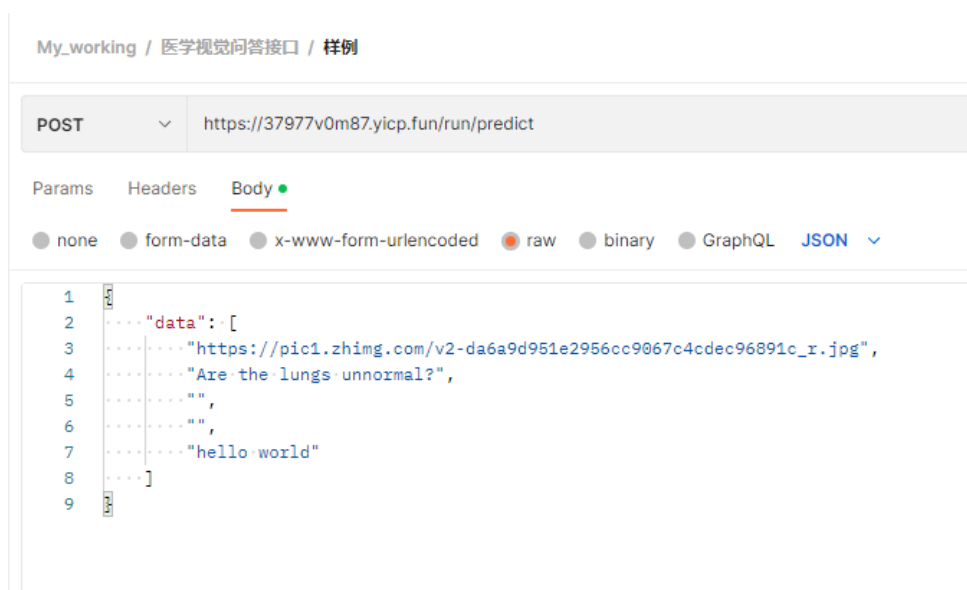


图 5-16 向 API 发送请求

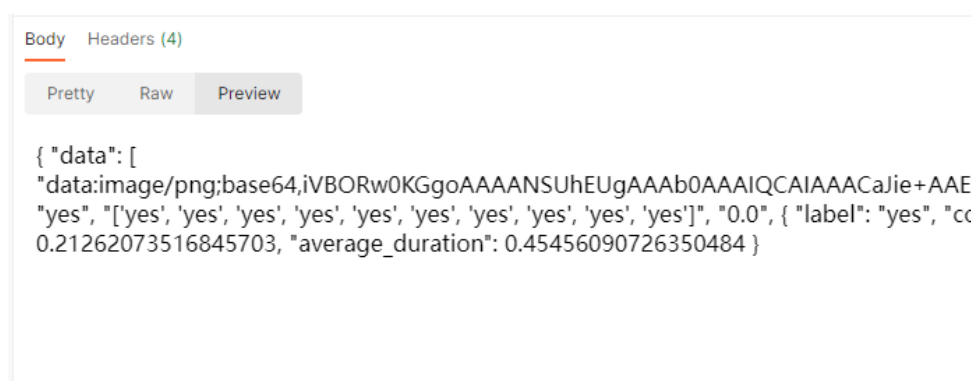


图 5-17 获得 API 返回数据

的返回。验证了这个在线系统的可靠性，为整个系统的实用化打下基础。

综上，本章通过分别设计和实现了模态自适应系统和在线系统设计部署，让一个本地模型可以通过互联网远程地为用户们提供医疗视觉问答服务，并适应用户的实时反馈和交互，是一个可以实现在线学习和更新的系统，具备一定的实际意义和创新价值。

总结与展望

总结

医学视觉问答是近几年新兴起的一项研究，相比于传统的针对自然图像开展的基于常识的视觉问答（VQA），具备专业化知识的视觉问答模型可以真正切实地解决用户的疑难，让 VQA 系统真正具备实用性和落地价值。但由于目前医学图像问答样本的稀缺以及高昂的人力标注成本，Med-VQA 模型在问答性能、泛化能力以及实用性上都还有着巨大的发展潜力。限制 Med-VQA 模型性能的不仅是数据的匮乏，还因为医学图像相比自然图像具有更大的噪声，关键信息占比也更小并且多为光谱单一的灰度图，可提取的特征十分有限。因此在不断丰富样本量的同时，也需要关注如何提高模型提取特征，并且融合特征的效果和效率，即小样本下进行特征提取和特征融合的能力；且由于问答对样本的缺乏，如何让模型在有限的文本里与图像一起建立更多、更有效的医学视觉关联，从而获得更好的开放式问答能力也十分重要。

同时，由于医学问答是一个极具风险性的信息交换场景，错误的信息相比自然图像问答更容易造成严重的后果，所以此类系统必须具备一套完善的风险评估和预测的能力，也就是 Med-VQA 模型不应该仅仅关心问答预测的准确率，更需要关注模型对其预测结果有多少把握，即不确定性量化的能力。目前 QA 领域大部分深度网络的权值都是点估计的形式，基于点估计的预测可能会提供虚假的高置信度的错误回答，且无法对预测结果中的不确定性进行有效度量。针对以上这些问题，本文基于增强视觉特征理论和贝叶斯不确定性估计原理，提出了一种多编码器混合自注意力网络 MEMSA 对医学图像-文本多模态特征进行高效的特征提取和关联性建模，同时依据贝叶斯神经网络设计了一种贝叶斯分类器，用来估计模型进行分类预测时的不确定性并探讨其与模型性能之间的关系。

最后，为了提高模型在真实场景下的可用性以及体验实际的医学视觉问答效果，本文还设计了可搭载不同医学视觉问答模型用于不同场景进行交互的模态自适应系统，并融合云服务等手段，实时在线不受时间和空间约束地为使用者提供医学视觉问答服务。论文的主要工作总结如下：

- (1) 提出一种多编码器混合自注意网络（MEMSA）用于医学图像-文本特征提取和关系建模，在经典的特征提取网络设计的基础上，本文针对医学图像以及文本数据的特点和痛点，针对性地设计了用于卷积去噪的自编码器，适用于小样本

学习的模型不可知元学习编码器以及具备图像-文本语义联系和良好跨模态能力的对比学习预训练模型。再对这些编码器进行集成，分别用于解决目前医学图像噪声大、可用信息少、小样本下多模态大模型学习不充分以及 Med-VQA 模型在面对开放式问答场景时缺乏良好的跨模态图像-文本关系特征抽取和建模能力等问题。本文还通过对比实验与目前主流的医学视觉特征提取模型以及特征融合机制进行了比较以及实例分析。通过编码器-注意力模型单一变量对照实验，实验表明 MEMSA 可以取得比其他方法更出色的特征提取性能以及跨模态特征融合能力，通过详细分析各子类问答的准确率，说明了性能提升的来源。同时通过问答样例对比凸显了 MEMSA 模型在语义关系建模上的能力。

- (2) 基于经典贝叶斯神经网络 (BNN) 提出了一种用于估计医学视觉问答模型不确定性的方法以及网络模型。在经典 Med-VQA 分类模型中将分类器的权值由传统的点估计替换为概率分布的形式。同时阐述了局部贝叶斯进行不确定性估计的原理和应用性。在深度网络中，网络模型的不确定性会通过权值分布形式向后传递，局部的贝叶斯不确定性估计具有合理性以及更好的实用性。在预测时，通过变分推断和蒙特卡洛采样方法从权值分布中进行多次采样获得多个子网络，对它们进行集成，不仅可以获得更可靠的预测，而且可以获得预测的不确定性，也就是模型对自己预测的把握程度。通过采样频率和不确定性估计实验，阐释了增大的采样频率可以增强模型不确定性预测能力的原理机制。通过拒绝分类实验，发现在医学视觉问答模型中，容易错分类的往往也是具有高不确定性的样本，从而使得拒绝对高不确定性样本进行预测这一方式可以提升模型的整体性能表现，同时，这一方式以及机制也极大地保障了医学视觉问答模型的可靠性和安全性。
- (3) 针对视觉问答具有复杂的输入模态这一特点以及不同数据集训练的模型适配的场景不同，并且它们之间往往存在难迁移、不通用等问题设计了模态自适应交互系统。在传统的视觉问答路线中增加了模型控制模块和反馈回路，这一机制使得模型可以通过收集用户反馈，通过设计合适的自学习或分类算法选择合适的模型与用户进行交互，更好、更优质地解决用户的问题。在医学视觉问答场景下，图像输入往往各种类医学影像。为了实现良好的实时交互和学习效果，选用图像分类的方式确定交互模型。为此用不同数据集的图像数据重构了一个用于分类的新数据集，通过对比各种经典机器学习方法后发现，随机森林方法

在这一类问题的表现上较为突出，具有最好的效果。同时为了方便用户体验和收集样本，改善模型效果，本文还将模型以及系统部署成云端服务，可以随时随地给用户提供医学视觉问答咨询服务。提高了医学视觉问答研究的实用性，易用性，可以让人人都拥有一个在线“医生”提供医疗诊断服务。

展望

目前基于多编码器、跨模态自注意力以及不确定性估计的医学视觉问答研究处于刚萌芽的阶段，相关方法甚少。本文提出基于多编码器混合自注意网络的医学视觉问答及其不确定性研究，并通过云服务搭建模态自适应交互系统提高了医学视觉问答模型及系统的回答交互效果并提高了其可靠性和易用性，但仍然存在一些问题，需要进一步的研究和探索：

第一，本文提出的基于多编码器混合自注意力网络 MEMSA。实际上，MEMSA 多编码特征之间是直接混合拼接的方式，相比于混合拼接，采用直接融合或者系数可学习的加权融合方式或许更能凸显图像-文本之间的多模态关联。因此，多编码器融合以及其配对的注意力研究是一个值得研究的问题。

第二，本文提出的局部贝叶斯不确定估计只针对决策层，度量的信息有限，并且没有开展信号扰动，对抗攻击等系统鲁棒性实验。贝叶斯神经网络由于采用蒙特卡洛采样方法，模型预测时容易采集到扰动信号从而获得错误的分类，抗扰动，对抗攻击性能较差。因此，通过 BNN 开展医学视觉问答不确定性研究，做抗扰动、对抗攻击等实验是一个值得探索的方向。

第三，本文提出的在线模态自适应医学视觉问答模型以及系统仍然缺乏医学模型和优质的样本。若要在实际使用中获得良好的效果，需要源源不断的相关数据进行训练。同时模型控制器还研究最了各种机器学习算法用于收集样本和用户反馈，迭代以提升模型的效果。

参考文献

- [1] Yu L, Park E, Berg A C, et al. Visual madlibs: Fill in the blank description generation and question answering[C]//Proceedings of the IEEE international conference on computer vision. 2015: 2461-2469.
- [2] Wu Q, Teney D, Wang P, et al. Visual question answering: A survey of methods and datasets[J]. Computer Vision and Image Understanding, 2017, 163: 21-40.
- [3] Teney D, Wu Q, van den Hengel A. Visual question answering: A tutorial[J]. IEEE Signal Processing Magazine, 2017, 34(6): 63-75.
- [4] Lin Z, Zhang D, Tac Q, et al. Medical visual question answering: A survey[J]. arXiv preprint arXiv:2111.10056, 2021.
- [5] Malinowski M, Fritz M. A multi-world approach to question answering about real-world scenes based on uncertain input[J]. Advances in neural information processing systems, 2014, 27.
- [6] Kafle K, Kanan C. Answer-type prediction for visual question answering[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 4976-4984.
- [7] Zhou B, Tian Y, Sukhbaatar S, et al. Simple baseline for visual question answering[J]. arXiv preprint arXiv:1512.02167, 2015.
- [8] Ren M, Kiros R, Zemel R. Exploring models and data for image question answering[J]. Advances in neural information processing systems, 2015, 28.
- [9] Shih K J, Singh S, Hoiem D. Where to look: Focus regions for visual question answering [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 4613-4621.
- [10] Zhu Y, Groth O, Bernstein M, et al. Visual7w: Grounded question answering in images [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 4995-5004.
- [11] Yang Z, He X, Gao J, et al. Stacked attention networks for image question answering[C] //Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 21-29.
- [12] Liu B, Lane I. Attention-based recurrent neural network models for joint intent detection

- and slot filling[J]. arXiv preprint arXiv:1609.01454, 2016.
- [13] Kim J H, Jun J, Zhang B T. Bilinear attention networks[J]. Advances in neural information processing systems, 2018, 31.
 - [14] Nguyen B D, Do T T, Nguyen B X, et al. Overcoming data limitation in medical visual question answering[C]//Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22. 2019: 522-530.
 - [15] Masci J, Meier U, Cireşan D, et al. Stacked convolutional auto-encoders for hierarchical feature extraction[C]//Artificial Neural Networks and Machine Learning–ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, Part I 21. 2011: 52-59.
 - [16] Zhan L M, Liu B, Fan L, et al. Medical visual question answering via conditional reasoning[C]//Proceedings of the 28th ACM International Conference on Multimedia. 2020: 2345-2354.
 - [17] Gong H, Chen G, Liu S, et al. Cross-modal self-attention with multi-task pre-training for medical visual question answering[C]//Proceedings of the 2021 international conference on multimedia retrieval. 2021: 456-460.
 - [18] Do T, Nguyen B X, Tjiputra E, et al. Multiple meta-model quantifying for medical visual question answering[C]//Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24. 2021: 64-74.
 - [19] Eslami S, de Melo G, Meinel C. Does clip benefit visual question answering in the medical domain as much as it does in the general domain?[J]. arXiv preprint arXiv:2112.13906, 2021.
 - [20] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
 - [21] Chen Z, Li G, Wan X. Align, Reason and Learn: Enhancing Medical Vision-and-Language Pre-training with Knowledge[C]//Proceedings of the 30th ACM International Conference on Multimedia. 2022: 5152-5161.
 - [22] Ben Abacha A, Sarrouiti M, Demner-Fushman D, et al. Overview of the vqa-med task at

- imageclef 2021: Visual question answering and generation in the medical domain[C]//
Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum-working
notes. 2021.
- [23] Ngiam J, Khosla A, Kim M, et al. Multimodal deep learning[C]//Proceedings of the 28th
international conference on machine learning (ICML-11). 2011: 689-696.
- [24] Ma L, Lu Z, Li H. Learning to answer questions from image using convolutional neural
network[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 30: 1.
2016.
- [25] Malinowski M, Rohrbach M, Fritz M. Ask your neurons: A deep learning approach to
visual question answering[J]. International Journal of Computer Vision, 2017, 125: 110-
135.
- [26] Zitnick C L, Dollár P. Edge boxes: Locating object proposals from edges[C]//Computer
Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12,
2014, Proceedings, Part V 13. 2014: 391-405.
- [27] Nam H, Ha J W, Kim J. Dual attention networks for multimodal reasoning and matching
[C]//Proceedings of the IEEE conference on computer vision and pattern recognition.
2017: 299-307.
- [28] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers
for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
- [29] Lu J, Batra D, Parikh D, et al. Vilbert: Pretraining task-agnostic visiolinguistic repre-
sentations for vision-and-language tasks[J]. Advances in neural information processing
systems, 2019, 32.
- [30] Li L H, Yatskar M, Yin D, et al. Visualbert: A simple and performant baseline for vision
and language[J]. arXiv preprint arXiv:1908.03557, 2019.
- [31] Kim W, Son B, Kim I. Vilt: Vision-and-language transformer without convolution or re-
gion supervision[C]//International Conference on Machine Learning. 2021: 5583-5594.
- [32] Maron O, Lozano-Pérez T. A framework for multiple-instance learning[J]. Advances in
neural information processing systems, 1997, 10.
- [33] Huang Y, Du C, Xue Z, et al. What makes multi-modal learning better than single (prov-
ably)[J]. Advances in Neural Information Processing Systems, 2021, 34: 10944-10956.

-
- [34] Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain.[J]. Psychological review, 1958, 65(6): 386.
- [35] Antol S, Agrawal A, Lu J, et al. Vqa: Visual question answering[C]//Proceedings of the IEEE international conference on computer vision. 2015: 2425-2433.
- [36] Hasan S A, Ling Y, Farri O, et al. Overview of ImageCLEF 2018 Medical Domain Visual Question Answering Task.[C]//CLEF (Working Notes). 2018.
- [37] Lau J J, Gayen S, Ben Abacha A, et al. A dataset of clinically generated visual questions and answers about radiology images[J]. Scientific data, 2018, 5(1): 1-10.
- [38] Liu B, Zhan L M, Xu L, et al. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering[C]//2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). 2021: 1650-1654.
- [39] He X, Zhang Y, Mou L, et al. Pathvqa: 30000+ questions for medical visual question answering[J]. arXiv preprint arXiv:2003.10286, 2020.
- [40] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. 2021: 8748-8763.
- [41] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.
- [42] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[J]. Advances in neural information processing systems, 2013, 26.
- [43] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.
- [44] 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. 2015: 2048-2057.
- [45] Pearl J. Bayesian networks: A model of self-activated memory for evidential reasoning [C]//Proceedings of the 7th conference of the Cognitive Science Society, University of California, Irvine, CA, USA. 1985: 15-17.
- [46] Jordan M I, Ghahramani Z, Jaakkola T S, et al. An introduction to variational methods

- p for graphical models[J]. Machine learning, 1999, 37: 183-233.
- [47] Metropolis N, Rosenbluth A W, Rosenbluth M N, et al. Equation of state calculations by fast computing machines[J]. The journal of chemical physics, 1953, 21(6): 1087-1092.
 - [48] Malinowski M, Rohrbach M, Fritz M. Ask your neurons: A neural-based approach to answering questions about images[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1-9.
 - [49] Gao P, Jiang Z, You H, et al. Dynamic fusion with intra-and inter-modality attention flow for visual question answering[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 6639-6648.
 - [50] Shridhar K, Laumann F, Liwicki M. A comprehensive guide to bayesian convolutional neural network with variational inference[J]. arXiv preprint arXiv:1901.02731, 2019.
 - [51] Blundell C, Cornebise J, Kavukcuoglu K, et al. Weight uncertainty in neural network[C]//International conference on machine learning. 2015: 1613-1622.
 - [52] Gal Y, Ghahramani Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning[C]//international conference on machine learning. 2016: 1050-1059.
 - [53] McNeish D. On using Bayesian methods to address small sample problems[J]. Structural Equation Modeling: A Multidisciplinary Journal, 2016, 23(5): 750-773.
 - [54] Welling M, Teh Y W. Bayesian learning via stochastic gradient Langevin dynamics[C]//Proceedings of the 28th international conference on machine learning (ICML-11). 2011: 681-688.
 - [55] Deng Z, Yang X, Xu S, et al. Libre: A practical bayesian approach to adversarial detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 972-982.
 - [56] Hinton G E, Van Camp D. Keeping the neural networks simple by minimizing the description length of the weights[C]//Proceedings of the sixth annual conference on Computational learning theory. 1993: 5-13.
 - [57] Graves A. Practical variational inference for neural networks[J]. Advances in neural information processing systems, 2011, 24.
 - [58] Neal R M, Hinton G E. A view of the EM algorithm that justifies incremental, sparse,

- and other variants[J]. Learning in graphical models, 1998: 355-368.
- [59] Friston K, Mattout J, Trujillo-Barreto N, et al. Variational free energy and the Laplace approximation[J]. Neuroimage, 2007, 34(1): 220-234.
- [60] Jaakkola T S, Jordan M I. Bayesian parameter estimation via variational methods[J]. Statistics and Computing, 2000, 10: 25-37.
- [61] Kingma D P, Welling M. Auto-encoding variational bayes[J]. arXiv preprint arXiv: 1312.6114, 2013.
- [62] Kwon Y, Won J H, Kim B J, et al. Uncertainty quantification using Bayesian neural networks in classification: Application to biomedical image segmentation[J]. Computational Statistics & Data Analysis, 2020, 142: 106816.
- [63] Shridhar K, Laumann F, Liwicki M. Uncertainty estimations by softplus normalization in bayesian convolutional neural networks with variational inference[J]. arXiv preprint arXiv:1806.05978, 2018.
- [64] Kendall A, Gal Y. What uncertainties do we need in bayesian deep learning for computer vision?[J]. Advances in neural information processing systems, 2017, 30.