

武汉大学

本科毕业论文（设计）

基于大模型的多模态医疗数据智能诊
断技术研究

姓 名： 陆 卓 然
学 号： 2020302111235
专 业： 计算机科学与技术
学 院： 计 算 机 学 院
指导教师： 杜 博 教 授

二〇二四年四月

原创性声明

本人郑重声明：所呈交的论文（设计），是本人在指导教师的指导下，严格按照学校和学院有关规定完成的。除文中已经标明引用的内容外，本论文（设计）不包含任何其他个人或集体已发表及撰写的研究成果。对本论文（设计）做出贡献的个人和集体，均已在文中以明确方式标明。本人承诺在论文（设计）工作过程中没有伪造数据等行为。若在本论文（设计）中有侵犯任何方面知识产权的行为，由本人承担相应的法律责任。

作者签名：

指导教师签名：

日 期：

年 月 日

版权使用授权书

本人完全了解武汉大学有权保留并向有关部门或机构送交本论文（设计）的复印件和电子版，允许本论文（设计）被查阅和借阅。本人授权武汉大学将本论文的全部或部分内容编入有关数据进行检索和传播，可以采用影印、缩印或扫描等复制手段保存和汇编本论文（设计）。

作者签名：

指导教师签名：

日 期：

年 月 日

摘要

在人工智能领域，基于 Transformer 的模型已成为生物医学智能分析领域的基础模块，如 x 光片分析、生物医学文本生成、疾病预测和细胞类型注释。当前的生物医学基础模型主要采用预训练和微调的方式。然而，医学是一个多模态学科，临床医生需要从多种数据中获取信息。目前的 AI 系统虽然在某些任务上取得了进展，但仍然存在着数据模态单一、模型任务单一的局限。新型基础模型的出现为重新思考生物医疗人工智能系统的发展提供了新机遇。这类模型具有多场景通用、可快速迁移或直接应用于多种下游任务的能力，有望在生物医学领域取得广泛应用。

本研究选取由通用领域多模态大模型 LLaVA 训练得到的 LLaVA-Med 以及 OFA 训练得到的 BiomedGPT，在不同下游任务上进行微调，从准确率、召回率等方面评估模型的性能表现，验证模型面对未知任务的适应能力，为医学领域的自然语言处理和计算机视觉任务提供有价值的参考。

本研究的主要内容有：

- (1) 给定纯语言的 GPT-4 接口医学图片描述，结合问题进行参考预测，用预训练模型 LLaVA-Med 生成对同一问题的回复。让 GPT-3.5 从有用性、相关性、准确性和细节级别进行评分，比较 LLaVA-Med 在医学多模态任务上与 GPT-4 的差距。
- (2) 在三个生物医学数据集上比较微调的 LLaVA-Med 与 BiomedGPT 在闭集问题上的准确率，考察微调对判断任务的性能提升。
- (3) 从召回率、BLEU、F1 分数等指标比较 LLaVA-Med 与 BiomedGPT 面对下游数据集执行生成任务的性能，考察并比较二者在医学领域生成的开放式回答的质量。

实验结果表明：LLaVA-Med 在医学多模态任务上最高可达到 GPT-4 的 77% 的性能，LLaVA-Med 和 BiomedGPT 在微调后幻觉明显减少，无论是判断任务还是生成任务的回答质量都有显著提升。

关键词：多模态；LLaVA-Med；BiomedGPT；判别任务；生成任务

ABSTRACT

In the field of artificial intelligence, Transformer-based models have become important tools for addressing biomedical challenges, such as X-ray analysis, biomedical text generation, disease prediction, and cell type annotation. Current biomedical foundation models primarily employ pre-training and fine-tuning methods. However, medicine is a multimodal discipline, and clinicians need to obtain information from various data sources. Although current AI systems have made progress in certain tasks, they still exhibit limitations in terms of single modality and single task. The emergence of new foundation models provides an opportunity to rethink the development of medical AI systems. These models possess the ability to efficiently adapt to downstream tasks and are expected to be widely applied in the medical field.

Given the relative scarcity of work in the field of multimodal medical large models, this study selects the LLaVA-Med trained from the general domain multimodal large model LLaVA and BiomedGPT trained from OFA, and fine-tunes them on different downstream tasks. The performance of the models is evaluated in terms of accuracy, recall, and other aspects to verify their adaptability to unknown tasks, providing valuable reference for natural language processing and computer vision tasks in the medical field.

The main contents of this study are: (1) Given purely language-based GPT-4 interface medical image descriptions, combined with questions for reference prediction, the pre-trained model LLaVA-Med is used to generate responses to the same questions. GPT-3.5 is used to score the responses in terms of usefulness, relevance, accuracy, and level of detail, comparing the gap between LLaVA-Med and GPT-4 in multimodal medical tasks. (2) Comparing the accuracy of LLaVA-Med and BiomedGPT on closed-set questions in three biomedical datasets before and after fine-tuning, to examine the performance improvement of fine-tuning on judgment tasks. (3) Comparing the performance of LLaVA-Med and BiomedGPT in generating tasks on downstream datasets in terms of recall, BLEU, and F1 scores, examining and comparing the quality of open-ended answers generated by both in the medical field.

The experimental results show that LLaVA-Med can reach up to 77% of the perfor-

mance of GPT-4 in multimodal medical tasks. Both LLaVA-Med and BiomeGPT have significantly reduced hallucinations after fine-tuning, and the quality of answers in both judgment tasks and generation tasks has been significantly improved.

Key words: Biomedicine; Multimodal; LLaVA-Med; Biomedgpt; Fine-tuning; Judgment task; Generation task

目 录

1	绪论	1
1.1	研究的背景与意义	1
1.2	国内外研究现状和发展趋势	2
1.2.1	医疗语言大模型研究现状	2
1.2.2	医疗多模态大模型研究现状	3
1.2.3	医疗数据集研究现状	3
1.3	本文研究内容与特色	4
1.4	本文结构	5
1.5	本章小结	6
2	医疗多模态大模型与下游数据集	7
2.1	医疗多模态大模型	7
2.1.1	LLaVA-Med	7
2.1.2	BiomedGPT	8
2.2	下游任务数据集	10
2.2.1	VQA-RAD	10
2.2.2	Slake-VQA	10
2.2.3	Path-VQA	11
2.3	本章小结	11
3	模型预训练与微调	13
3.1	LLaVA-Med	13
3.1.1	数据生成	13
3.1.2	模型预训练	15
3.1.3	下游任务微调	16
3.2	BiomedGPT	17
3.2.1	预训练任务	17
3.2.2	自回归生成训练与微调	18
3.3	本章小结	19

4	实验结果与分析	20
4.1	实施细节	20
4.2	评测指标	20
4.2.1	F1 分数、准确率与召回率	20
4.2.2	精准匹配分数	21
4.2.3	BLEU 分数	21
4.3	实验结果	21
4.3.1	LLaVA-Med 与 GPT-4 性能比较	21
4.3.2	下游任务适应能力探究	22
4.3.3	零样本迁移学习	23
5	总结与展望	25
5.1	工作总结	25
5.2	未来展望	25
	参考文献	27
	致谢	29

1 绪论

1.1 研究的背景与意义

在人工智能 (AI) 的快速发展领域, 基于 Transformer^[1] 的基础模型已经成为解决各种生物学挑战的强大工具。生物学基础模型的主流范式是预训练然后进行微调。具体来说, 首先在大规模数据集上进行预训练, 然后在下游数据集上进行微调, 促进从源域到目标域的知识转移^[2]。不同于受监督的预训练模式, 自监督学习并不受带标注数据的可用性的影响, 其可以从大量的数据中学习而不需要人为标记, 因而被广泛使用。例如, 目前 BERT 驱动或是 GPT 驱动的语言模型已经在生物学自然语言处理中得到广泛研究, 并比以前的方法获得了更好的性能^[3]。

但我们应该注意到, 医学本质上是一个多模态学科, 临床医生通常在提供护理时从临床笔记、实验室测试、生命体征和观察、医学图像、基因组学等广泛的模式解释数据。目前 AI 在生物学领域虽然取得了重大的进步, 但绝大多数的模型仍处在单一模态单任务的阶段。考虑一个现有的人工智能系统来解释乳房 x 线照片^[4], 虽然该系统在乳腺癌筛查方面获得了最先进的性能 (SOTA), 但它不能合并相关信息, 比如说患者的健康记录, 以及其他多模态信息: MRI, 或已发表的医学文献等等。此外, 系统的输出被限制为预先指定的一组可能的分类, 它不能解释其预测或参与协作对话以从医生的反馈中学习。这些单一模态, 单任务的专用人工智能系统在实际生活中的表现将被大大限制。基础模型^[5] 的出现为重新思考医疗人工智能系统的发展提供了机会。这些模型能够高效的适应下游任务, 产生诸如上下文学习或是少样本微调等涌现能力^[6]。随着生物学数据生成和创新的速度的增加, 此类模型将具有广泛的下游应用程序。

在医学领域, 大数据技术结合大语言模型的应用预计将具有巨大的影响: 首先是辅助医生诊断和鉴别诊断方面, 其次是在循证医学领域, 此外医疗大数据结合大语言模型也可以应用于辅助医生进行临床和医学研究方面。通过将医疗大数据与人工智能大语言模型相结合, 可以实现更加精准、高效、智能化的医疗诊断和治疗, 并将为人类的健康领域做出更大的贡献^[7]。

本课题拟选取 LLaVA-Med^[8] 以及 BiomedGPT^[9] 两个开源多模态医疗大模型在下游任务上进行微调训练与测评。其中 LLaVA-Med 借助 GPT-4 模型进行自我指

导的开放性研究问题回答，使得模型能够学习对齐生物医学词汇并掌握开放式会话语义，并用 GPT-4 评估其在完成概念对齐与指令微调后的性能。在下游任务上，选取 Slake-VQA^[10]、VQA-RAD^[11]、Path-VQA^[12] 三个数据集，分别测试 LLaVA-Med 与 BiomedGPT 在这些数据集上执行判断任务与生成任务的能力。总体而言，本课题旨在通过研究和应用多模态医疗大模型，推动生物医学人工智能的发展，提高医疗诊断和治疗的效率和准确性。

1.2 国内外研究现状和发展趋势

多模态医疗大模型是一个相对较新的研究领域，它结合了自然语言处理、计算机视觉、和其他模态的处理能力，以应对医疗领域的复杂和多样化的数据类型。随着深度学习和大数据技术的发展，基于 Transformer 的模型如 BERT、GPT 等在自然语言处理领域取得了巨大成功。这些技术也被扩展到医疗领域，形成了一些专门针对医疗数据的模型，如 BioBERT、ClinicalBERT 等。近年来，出现了一些尝试将这些模型扩展到多模态场景的研究，如 MedBERT、CLIP 等，它们能够处理文本和图像等不同类型的医疗数据。多模态医疗大模型是一个充满活力且不断发展的研究领域，随着技术的进步和数据资源的丰富，预计将在医疗领域发挥越来越重要的作用。

1.2.1 医疗语言大模型研究现状

在近期的研究中，语言大模型（LLMs）在通用领域取得了显著成功，其中最引人注目的是 ChatGPT/GPT-4^[13]。这些模型激发了开源社区开发生物医学领域的多个 LLM 聊天机器人的兴趣。这些聊天机器人包括：

- (1) ChatDoctor：通过在生物医学领域的指令跟踪数据集上进行微调，使用 LLM 模型的 ChatDoctor 成为医疗领域的有力助手^[14]
- (2) Med-Alpaca：Med-Alpaca 是唯一接受图像输入的多模态生物医学聊天机器人，其在定制的生物医学指令跟踪数据集上进行微调^[15]
- (3) PMC-LLaMA、Clinical Camel、DoctorGLM 和 Huatuo：这些聊天机器人都基于开源 LLMs 初始化，通过微调在医学领域的指令跟踪数据集上，为了更好地了解患者需求并提供明智的建议^[16]
- (4) 科大讯飞近日发布基于讯飞星火 3.0 的医疗大模型，并推出了专门面向 C 端服务的讯飞晓医 APP，可实现病前预问诊、体检报告分析等功能。据不完全

统计,过去几个月,一些科技公司接连发布了多款医疗大模型。比如百度推出灵医大模型,商汤发布升级版医疗健康大语言模型大医,京东健康推出京医千询等^[17]。

1.2.2 医疗多模态大模型研究现状

随着 ChatGPT 的发布,各行各业开启探索、应用人工智能大模型的新时代。在海量数据、大规模算力和强大的算法优化能力等条件的支撑下,基于深度学习等技术的庞大神经网络模型问世,使自然语言处理、语音识别、计算机视觉等任务的处理得到了极大的改善。然而,多模态医疗大模型领域的工作仍相对空缺。对于这一领域的研究有望扩大大模型的应用范围。在这方面的的工作尚处于起步阶段。

- (1) 通用医疗人工智能: Moor 等人讨论了通用医疗人工智能的概念,尽管没有实现或实证结果。^[18]
- (2) Geneformer 和 Med-PaLM M: Theodoris 等人引入了 Geneformer^[19],这是一种基于 Transformer 的模型,在约 3000 万个单细胞转录组的语料库上进行了预训练,以实现低数据网络生物学应用中特定于上下文的预测。Google 的 Med-PaLM M 专门为生物医学领域设计,通过微调和对齐 PaLM-E 通用模型来构建,涵盖了多种任务,包括问答、报告生成、视觉问答、医学图像分类和基因组变异调用^[20]
- (3) 中国首个整合文本和视觉数据分析的大型视觉语言模型 Qilin-Med-VL 1^[21]结合了预先训练的视觉变压器 (ViT) 和基础 LLM。它经历了一个彻底的两个阶段的课程培训过程,包括特征对齐和指令调整。该方法增强了模型生成医学说明和回答复杂医学查询的能力。
- (4) 本课题研究用到的 BiomedGPT,是一种多任务生物医学基础模型,通过结合语言模型 (LM) 和掩码图像填充目标,在各种医学图像、医学文献和临床笔记上进行了预训练。LLaVA-Med 使用 PubMed 和 GPT-4^[22] 来管理数据集之后的多模态指令,并使用它微调 LLaVA 模型。

1.2.3 医疗数据集研究现状

多模态医疗大模型的研究依赖于大量标注良好的多模态医疗数据集。目前,一些数据集如 MIMIC-III、CheXpert 等已经被广泛用于训练和评估模型。同时,研究者们也在不断探索新的任务来验证模型的能力,如医学图像分析、疾病预测、药物发现等。以下是一些在该领域中常用的数据集以及它们的研究现状:

MIMIC-III (Medical Information Mart for Intensive Care III): 这是一个广泛使用的公开医疗数据库, 包含了来自约 4 万名重症监护病人的多种类型数据, 如生命体征、医学笔记、实验室测试结果等。**MIMIC-III** 被用于各种医疗 AI 研究, 包括自然语言处理、生理信号分析和临床事件预测等。

CheXpert: 这是一个大型公共数据集, 包含来自约 22 万张胸部 X 光片的标签。它被用于训练和评估自动化胸部 X 光片解读算法, 包括多模态医疗大模型。

eICU Collaborative Research Database: 类似于 **MIMIC-III**, **eICU** 数据库包含了来自美国多个重症监护室的病人数据, 用于支持重症医学研究。

BraTS (Brain Tumor Segmentation Challenge): 这是一个关于脑肿瘤分割的数据集, 包含了多个模态的 **MRI** 扫描图像, 用于评估和比较不同的自动分割算法。

IU X-Ray Dataset: 这个数据集包含了从印第安纳大学医院收集的 7,470 张胸部 X 光片及其相关的医学报告, 常用于医学图像报告生成的研究。

这些数据集的研究现状表明, 随着医疗大数据的日益丰富, 研究者们正在不断探索和开发新的算法和模型来更好地利用这些数据, 以提高医疗诊断和治疗的准确性和效率。然而, 由于医疗数据的敏感性和隐私问题, 数据的获取和使用仍然面临着一定的挑战。此外, 为了使模型更具泛化能力, 还需要更多多模态和多任务的医疗数据集。随着技术的进步和数据共享机制的完善, 预计未来会有更多高质量的医疗数据集出现, 以支持多模态医疗大模型的研究和应用。

1.3 本文研究内容与特色

本课题通过选取两个开源的多模态医疗大模型, 即 **LLaVA-Med** 和 **BiomedGPT**, 进行微调训练与测评。通过对 **LLaVA-Med** 和 **BiomedGPT** 模型在不同下游任务上的微调和测评, 可以验证这些模型的适应性和通用性, 即它们能否有效地适应不同类型的医疗数据和任务。这一比较可以帮助研究者了解不同模型在特定下游任务上的性能表现, 验证这些模型的适应性和通用性, 为医学领域的自然语言处理和计算机视觉任务提供有价值的参考, 促进医疗大数据的利用, 推动医疗人工智能技术的发展, 并为未来的研究提供基础。本研究的特色在于, 通过对这些模型在不同下游任务上的微调和测评, 本课题实现了医学领域知识与人工智能技术的跨领域融合, 促进了医疗人工智能技术的发展和医学领域知识的深入应用, 促进了跨领域知识融合。同时, 本课题的研究方法和成果为未来在医疗人工智能领域的研究提供了新的思路 and 基础, 促进了该领域的创新发展。

1.4 本文结构

第 1 章介绍了本论文的研究背景，并从生物医学语言模型、多模态大模型以及生物医学数据三个方面简要说明了当前多模态医疗大模型的研究现状，同时介绍了本研究的主要创新点。多模态医疗大模型是一个相对较新的研究领域，研究现状多样且充满活力，各个方面都值得深入挖掘研究。本研究通过借助 GPT-4 为模型响应打分，在闭集问题上测量准确率，在开放式生成上评估回答质量，旨在深入了解当前多模态大模型医疗领域的最新进展，研究大模型在面对未知任务的适应能力，为未来该领域的研究提供有价值的参考。

第 2 章对本研究需要用到的预训练大模型 LLaVA-Med、BiomedGPT 以及下游任务需要用到的数据集 VQA-RAD,SLAKE 以及 PATH-VQA 进行简要说明。LLaVA-Med 是一个基于大型语言模型的医学领域预训练模型，专为处理医学文本和图像数据而设计，其通过 LLaVA 在 PMC-15M 数据集上进行了概念对齐与指令微调得到，而 BiomedGPT 基于 OFA，使用了多个多模式和多任务生物医学数据集进行预训练，能够理解专业的医学术语和概念。

第 3 章详细介绍 LLaVA-Med 与 BiomedGPT 预训练与微调中的具体流程与方法。不论是 LLaVA-Med 还是 BiomedGPT，作为一种专门为生物医学领域设计的基于 Transformer 架构的模型，都是建立在现有通用数据统一模型的成功之上。预训练过程中他们都遵循统一模型的基本原理：模态不可知，任务不可知，以及模态和任务综合性。通过编码器-解码器架构将具有任务相关指令的多模态数据映射到公共表示空间，以消除不同数据模态之间的差异。而在微调阶段，实验中使用了诸如端到端学习、数据增强等方法，不仅优化了模型的任务特定表现，还增强了模型对新场景和数据类型的适应能力。

第 4 章说明本实验的实验结果，包括对利用 GPT-4 为问题提供参考预测并为模型响应打分，在三个生物医学数据集 VQA-RAD、SLAKE 和 PATH-VQA 上进行了微调和评估。本文的研究发现，经过微调后，这两个模型在闭集问题的准确率、生成任务的召回率、BLEU 和 F1 分数等方面都表现出了显著的提升。这些结果表明，LLaVA-Med 和 BiomedGPT 在医学多模态任务上具有强大的潜力。

第 5 章对本论文的工作进行了总结，并对未来的探索进行了展望。本课题探讨了两个开源的多模态医疗大模型——LLaVA-Med 和 BiomedGPT——在不同医学多模态任务上的应用和性能。这些大模型在医学多模态任务上具有强大的潜力，并能为医学领域的自然语言处理和计算机视觉任务提供有价值的参考。未来可以

考虑引入更多种类的医学数据集，以进一步验证模型的泛化能力和适应性，同时在此基础上研究和开发更先进的多模态医疗大模型，以提高处理复杂医学数据的能力。未来的工作将继续探索更多的可能性，以推动医疗人工智能技术的发展和医疗健康领域的进步。

1.5 本章小结

本章主要介绍了多模态医疗大模型在人工智能领域的研究背景和意义，以及国内外研究现状和发展趋势。首先，介绍了基于 Transformer 的模型在生物医学领域的应用，特别是在自监督学习和多模态数据处理方面的重要性。然后，针对目前医疗人工智能系统存在的单一模态、单任务限制，提出了多模态医疗大模型的发展机遇。接着，介绍了国内外在生物医学领域语言模型、多模态医疗大模型和医疗数据集方面的研究现状，以及结合两领域的研究趋势。最后，提出了本课题的特色与创新点，包括多模态医疗大模型的应用、跨领域知识融合、促进医疗大数据利用和为未来研究提供新思路等方面。

2 医疗多模态大模型与下游数据集

2.1 医疗多模态大模型

2.1.1 LLaVA-Med

在通用领域中，存在大量可用的并行图文数据，例如网络图像及其相关图像描述。生成式预训练已被证明可以有效地利用这种并行数据进行自监督视觉语言建模，如 OpenAI 的多模态大模型 GPT-4 和微软开源的多模态大模型 LLaVA。通过指令调整模型根据多模态输入与人类概念对齐，由此产生的多模态大模型 (LMMs) 在各种面向用户的视觉语言任务（例如图像理解和推理）上表现出强大的零样本任务完成性能。作为第一个尝试将多模态指令调整扩展到生物医学领域的工作，LLaVA-Med 做出了以下贡献：

- (1) 生物医学多模态指令跟踪数据: 提出了一种新的数据生成方式来创建不同的（图像、指令、输出）实例。通过从 PMC-15M（一个覆盖广泛的数据集，包含从 PubMed Central1 中提取的 1500 万对生物医学图像-文本对）中采样生物医学图像-文本对并使用 GPT-4 从图像文本描述中创建问答对（预期输出）。这一过程不需要人工标注，利用 PMC-15 创建了一个极其多样化的视觉指令跟踪数据集。
- (2) 新的模型训练方法: 提出了一种新的模型训练方法，使用我们的自生成的生物医学多模态指令跟踪数据集将 LLaVA 适应生物医学领域。具体来说，首先通过图像文本对来对齐生物医学词汇。（即文本简单描述医学图像）。在概念对齐完成后，利用上文提到的 GPT-4 生成的视觉指令跟踪数据集，让模型充分学习开放式对话语义。具体的流程如图2.1所示：

LLaVA-Med 类似于语言模型 (LMs) 的前缀微调 (prefix-tuning)。通过新的可训



图 2.1 LLaVA-Med 使用通用域 LLaVA 初始化，然后以课程学习方式不断训练（第一个生物医学概念对齐，然后是成熟的指令调整）。

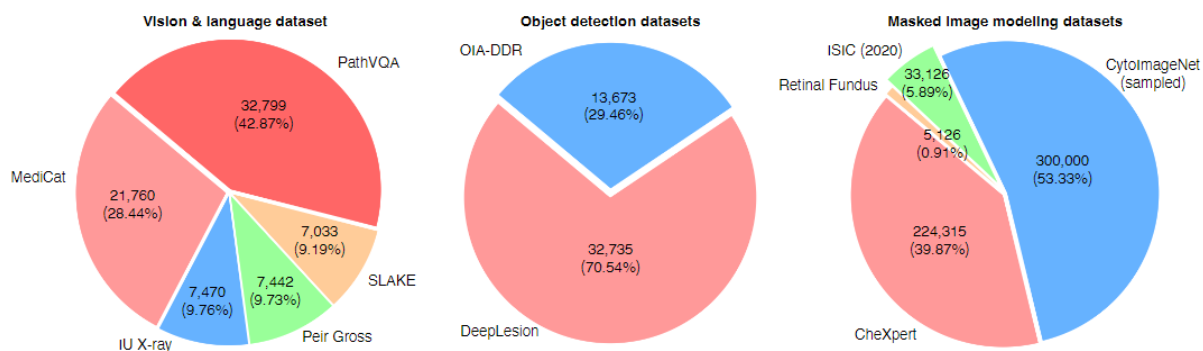


图 2.2 视觉和语言数据集、对象检测数据集和仅图像数据集的预训练数据集的分布

训练模块连接图像编码器和语言模型 LM。在训练过程中冻结图像编码器，对语言模型与投影层进行微调。

2.1.2 BiomedGPT

BiomedGPT 实现为序列到序列模型，架构设计原则旨在编码器有效地将各种模态映射到统一的语义表示中。作为一个轻量级模型，他的参数仅 3300 万到 1.82 亿个参数，而 LLaVA-Med 达到了 70 亿到 130 亿。尽管如此，通过在一系列数据集 2.2 上进行预训练，BiomedGPT 仍在多个任务上取得了 SOTA 结果。

BiomedGPT 以 BART 为骨干，在损坏的文本和 GPT 风格的从左到右自回归解码器上使用 BERT 风格的编码器实现为序列到序列模型。所有这些模型都依赖于广泛使用的具有多头注意力机制的 Transformer，这允许模型共同关注来自不同表示子空间的信息。BiomedGPT 在此之上做了部分改进。如图 2.2 所示，首先，为了提高预训练的收敛速度和稳定性，每一层都添加了三个归一化操作：自注意力前归一化 (LN)、后第一个 FFN 前归一化 (Post-attention LN) 和自注意力头缩放。同时，为了对位置信息进行编码，为文本和图像合并了两组绝对位置嵌入。(图 2.4) BiomedGPT 做出了以下贡献：

- (1) 作为一个完全开源与轻量级的通用生物学模型，虽然只有 3300 万到 1.82 亿个参数，但在一系列数据集上提供了 SOTA 结果，并且在五种不同的任务中表现出色。通过完全开源，详细的数据预处理工作流程和代码，都很容易被公众访问。这种透明度不仅有助于复制和验证其工作，而且还可作为推进生物学 AI 领域研究的宝贵资源。
- (2) 在现实世界的医疗保健环境中成功部署和评估。BiomedGPT 成功部署麻省

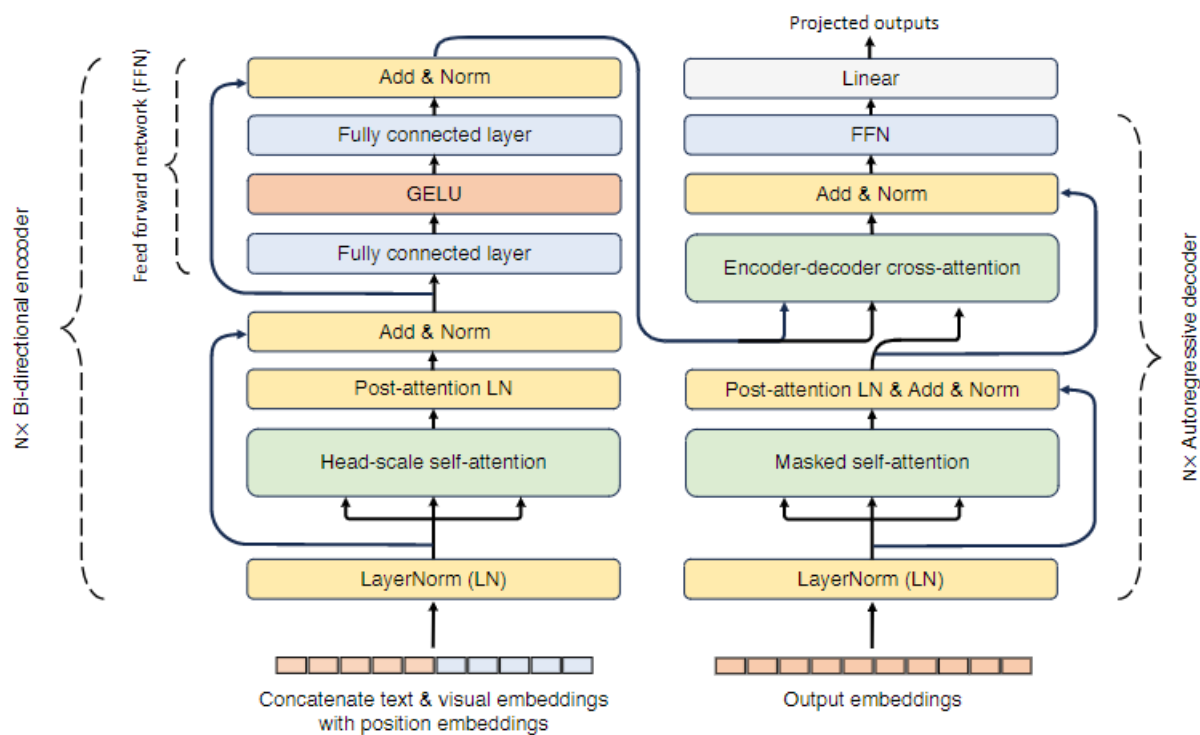


图 2.3 BiomedGPT 在多头注意力 Transformer 上做出改进

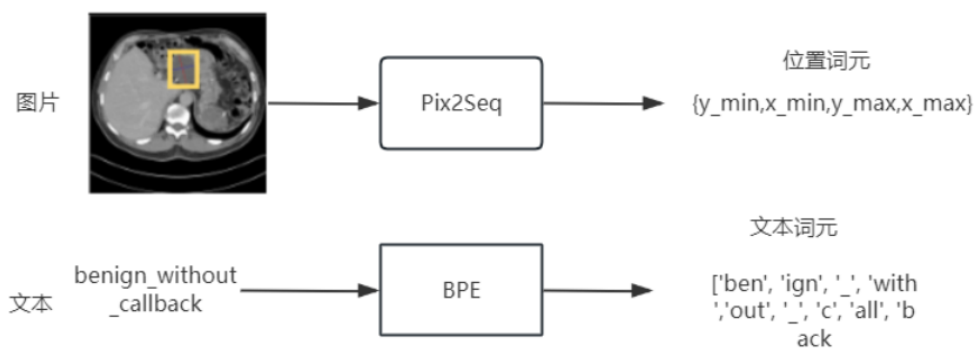
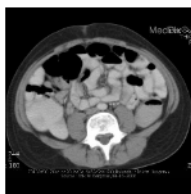


图 2.4 分别通过 Pix2Seq 和 BPE 对边界框和文本进行标记



VQA-RAD

Axial.

Instructions: You are a helpful medical assistant. The following are questions about medical knowledge. Solve them in a step-by-step fashion, referring to authoritative sources as needed.

Given . Q: Can you diagnose a pericardial effusion from this image? (closed domain)

A: No.

Given . Q: What cut of the body is this image? (open domain)

A:

图 2.5 VQA-RAD 数据示例

理工学院附属医院，并在与胸部 X 射线图像相关的问题的 AI 生成答案的准确性和可靠性上大大超过了 GPT-4(参考结果)。

2.2 下游任务数据集

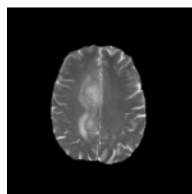
下游数据集在机器学习和人工智能领域中至关重要，尤其是在处理多模态模型和应用中这些数据集通常具有真实世界数据的特征，可以帮助研究者和开发者了解模型在实际应用中的效果。例如，在医学图像处理领域，模型必须能够准确识别和解释各种医学影像，下游数据集则提供了这些场景的代表性样本。下面我们将介绍用到的三个数据集：VQA-RAD、SLAKE 和 PATH-VQA。

2.2.1 VQA-RAD

VQA-RAD 是一个放射学视觉问答 (VQA) 数据集，由 315 张放射图像和 3,515 个由临床医生创建和验证的问答对组成。从三种成像方式 (CT、MRI 和 X 射线) 和三个解剖区域 (头部、腹部、胸部) 中选择放射图像。125 问题类型分为 11 类，包括模态、平面、器官系统、异常、大小、平面、位置推理、颜色、计数、属性等。58% 的问答对是封闭式的 (是/否或有限的选择, 即判断任务), 其余 42% 是开放式的 (生成任务)。我们采用官方的训练/测试拆分, 其中训练集包含 1,797 个 QA 对 (仅包含自由形式和释义问题), 测试集包含 451 个 QA 对 (未过滤)。

2.2.2 Slake-VQA

Slake-VQA 是放射学图像上的语义注释和知识增强双语 (英文和中文) VQA 数据集。它包含 642 张带注释的图像, 其中包含 14,028 个问答对, 涵盖 12 种疾病、39 个器官系统和 3 种成像方式 (CT、MRI 和胸部 X 射线)。问题要么是开放式的 (自由形式的, 生成任务), 要么是封闭式的 (是/否, 判断任务), 与图像内容的各个方面有关, 包括平面、质量、位置、器官、异常、大小、颜色、形状、知识图等。



Slake-VQA

Instructions: You are a helpful medical assistant. The following are questions about medical knowledge. Solve them in a step-by-step fashion, referring to authoritative sources as needed.

Given ``. Q: Is the lung healthy?

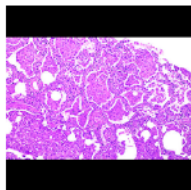
A: No.

Given ``. Q: Which part of the body does this image belong to?

A:

Brain.

图 2.6 Slake-VQA 数据示例



Path-VQA

Instructions: You are a helpful medical assistant. The following are questions about medical knowledge. Solve them in a step-by-step fashion, referring to authoritative sources as needed.

Given ``. Q: What is present ? (other)

A: Abdomen.

Given ``. Q: What is there of large numbers of macrophages within the alveolar spaces with only slight fibrous thickening of the alveolar walls? (other)

A:

accumulation of large numbers of macrophage.

图 2.7 Path-VQA 数据示例

训练集、验证集和测试集分别包含 9,849,2,109 和 2,070 个样本。

2.2.3 Path-VQA

Path-VQA 是一个病理学 VQA 数据集，总共包含 4,998 张病理图像，包含 32,799 个问题问答对。病理图像是从医学教科书和在线数字图书馆中提取的。每个图像都与多个 QA 对相关，这些对与病理学的不同方面有关，包括颜色、位置、外观、形状等。开放式问题占有所有问题的 50.2%，分为 7 个类别：什么、在哪里、何时、何地以及多少/多少，占有所有问题的 50.2%。其余的是带有简单“是/否”答案的封闭式问题。采用官方数据划分，其中训练集、验证集和测试集分别包含 19,755、6279 和 6,761 个 QA 对。

2.3 本章小结

在本章，本课题对研究需要用到的预训练大模型 LLaVA-Med、BiomedGPT 以及下游任务需要用到的数据集 VQA-RAD,SLAKE 以及 PATH-VQA 进行简要说明。LLaVA-Med 利用从大规模生物医学图像-文本对数据库 PMC-15M 中抽取的数据，自动生成问答对，从而创造了一个针对生物医学的多模态指令跟踪数据集。这种方法减少了对昂贵人工标注的依赖，同时增加了数据的多样性和覆盖范围。而作为一个轻量级的序列到序列模型，BiomedGPT 通过结构上的创新和优化提供了与大型模型相媲美的表现，尽管其参数数量显著少于 LLaVA-Med。BiomedGPT 特别

强调预训练的效率和稳定性,通过加入多个归一化步骤并改进位置编码机制,增强了模型对各种医学图像和文本数据的处理能力。作为一个轻量级的序列到序列模型, BiomedGPT 通过结构上的创新和优化提供了与大型模型相媲美的表现,尽管其参数数量显著少于 LLaVA-Med。BiomedGPT 特别强调预训练的效率和稳定性,通过加入多个归一化步骤并改进位置编码机制,增强了模型对各种医学图像和文本数据的处理能力。

3 模型预训练与微调

模型预训练与微调是深度学习领域两个极其关键的过程，它们通常在大规模数据集上训练和优化复杂模型（如神经网络）时使用。这两个步骤在许多应用中都有着至关重要的作用，特别是在自然语言处理（NLP）、计算机视觉和其他需要处理大量数据和特征的领域。预训练指的是在一个通常较大且多样化的数据集上训练模型的过程，目的是使模型学习到广泛的特征和模式。这一步骤通常不专注于特定任务，而是旨在让模型掌握尽可能多的通用知识，从而能够更好地理解或处理后续任务中的数据。预训练模型可以被多次重用，适用于多种下游任务，减少了从头开始训练模型的需要。而微调是在预训练的基础上，对模型进行进一步训练以适应特定任务的过程。在这个阶段，模型利用较小的、特定任务的数据集进行训练，调整其权重以更好地完成这些特定任务。本章将介绍 LLaVA-Med 和 BiomedGPT 预训练与微调的具体流程与方法。

3.1 LLaVA-Med

3.1.1 数据生成

目前的多模态医疗大模型领域，缺乏多模态生物医学数据集来训练指令跟踪助手。为了填补这一空白，LLaVA-Med 通过借助纯语言的 GPT-4 接口，从现有广泛的生物医学图像-文本对上创建医学对话数据集。该数据集包含两个部分，分别用于概念对齐与指令微调。

对于概念对齐的数据集部分，对于生物医学图像 X_v 与其对应描述 X_c ，对问题 X_q 进行回答采样，直接让图片描述作为回答，创建指令跟踪示例：

$$Human : X_q X_v < STOP > \backslash n Assistant : X_c < STOP > \quad (3.1)$$

而对于指令微调，使用医学图像对应描述提示 GPT-4 来生成医学视觉指令跟踪数据，如3.1所示。注意，医学图像对于 GPT-4 而言是不可见的。通过提示仅语言的 GPT-4 来呈现和管理与有关提供的生物医学图像的多轮对话的各种指令跟踪数据。

```
1 def few_shot_messages_gen(query_context, use_inline_mentions=True):
2     messages = [
```

```

3     {"role": "system", "content": ""You are an AI assistant specialized in
      biomedical topics.
4
5     You are provided with a text description (Figure Caption) of a figure image
      from a biomedical research paper. In some cases, you may have
      additional text (Figure Context) that mentions the image. Unfortunately
      , you don't have access to the actual image.
6
7     Your task is to generate a conversation between a person (User) inquiring
      about the image and you (Assistant) responding to their questions. The
      conversation should proceed as though both the User and Assistant are
      viewing the image, while not referring to the text information (Figure
      Caption and Figure Context).
8
9     Below are requirements for generating the questions and answers in the
      conversation:
10    - Avoid quoting or referring to specific facts, terms, abbreviations, dates
      , numbers, or names, as these may reveal the conversation is based on
      the text information, rather than the image itself. Focus on the visual
      aspects of the image that can be inferred without the text information
      .
11    - Do not use phrases like "mentioned", "caption", "context" in the
      conversation. Instead, refer to the information as being "in the image
      ."
12    - Ensure that questions are diverse and cover a range of visual aspects of
      the image.
13    - The conversation should include at least 2-3 turns of questions and
      answers about the visual aspects of the image.
14    - Answer responsibly, avoiding overconfidence, and do not provide medical
      advice or diagnostic information. Encourage the user to consult a
      healthcare professional for advice.
15    ""},
16    ]
17    for ex in instruct_few_shot_examples.fs:
18        messages += [
19            {"role": "user", "content": PromptGenerator.context_gen(ex,
20                use_inline_mentions)},
21            {"role": "assistant", "content": conv_to_str(ex["conversations"])},
22        ]
23    messages.append({"role": "user", "content": query_context})
    return messages

```

Listing 3.1 使用医学图像对应描述提示 GPT-4 来生成医学视觉指令跟踪数据 (医学相关问答), 图像对于 GPT-4 并不可见。

3.1.2 模型预训练

与通用域多模态大模型 LLaVA 的架构相同, 生物医学领域的 LLaVA-Med 通过线性投影层连接视觉编码器与语言模型。预训练过程分为以下两个阶段:

- (1) 概念对齐。为了在概念覆盖率和训练效率之间取得平衡, LLaVA-Med 从 PMC-15M 中过滤了 600k 图像文本对用于概念微调。这些图像文本通过上一小节提到的朴素拓展方法 3.1 转换为指令跟踪数据: 指令简单地呈现描述图像的任务。对于每个样本, 给定语言指令和图像输入, 我们要求模型预测原本图片描述。在训练期间, 我们保持视觉编码器和 LM 权重冻结, 并且只更新投影矩阵。这样, 大量新型生物医学的图像特征可以与预训练 LM 中的文本词嵌入对齐。这一阶段可以理解为将对齐的图文标记的词汇扩展到生物医学领域。
- (2) 端到端指令微调。为了收集图像标题及其上下文, LLaVA-Med 过滤 PMC-15M 以保留仅包含单个图的图像, 从五种最常见的成像方式中采样 60K 图像-文本对: CXR (chest X 射线)、CT (计算机断层扫描)、MRI (磁共振成像)、组织病理学和重度 (即宏观) 病理。然后, 从原始 PubMed 论文中提取提及图像的句子作为标题的附加上下文。为了生成高质量的回答, 还维 GPT-4 提供少样本样例 3.1。在这一阶段, 只冻结视觉编码器的权重, 利用 3.1 方法得到的生物医学语言图像指令跟踪数据更新投影层与语言模型的权重。

图 3.1 少样本样例

图片描述:

Chest X-ray. Cardiomegaly with diffuse bilateral interstitial infiltrates and a right-sided pleural effusion.

图片上下文:

The patient's respiratory status continued to worsen and on hospital day number 14, he required intubation and mechanical ventilation for hypoxemic respiratory failure .

```

1 User: Which side of the lung has a pleural effusion?
2 Assistant: The right side of the lung has a pleural effusion.
3
4 User: According to the CT image, does the patient has cardiomegaly?
5 Assistant: The image appears to be a chest x-ray. However, it
   indicates that the patient has cardiomegaly, which means an
   enlarged heart. Additionally, there are diffuse bilateral
   interstitial infiltrates and a right-sided pleural effusion
   present in the image.
6
7 User: Additional Info: The patient's respiratory status continued
   to worsen and on hospital day number 14, requiring intubation
   and mechanical ventilation for hypoxemic respiratory failure.
8 Assistant: Thank you for providing additional information about the
   patient's condition. Based on the information provided, it
   appears that the patient's respiratory status continued to
   worsen and on hospital day number 14, they required intubation
   and mechanical ventilation for hypoxemic respiratory failure.
   This suggests that the patient's lung function has
   significantly declined, and they may be experiencing respiratory
   distress. It is important for the medical team to closely
   monitor the patient's condition and provide appropriate
   treatment to manage their symptoms and improve their lung
   function.

```

3.1.3 下游任务微调

对于某些特定的生物医学场景，需要开发高度准确和特定于数据集的模型来提高助手的服务质量。本实验在三个生物医学 VQA 数据集上微调 LLaVA-Med，涵盖不同的数据集大小和不同的生物医学主题。给定一个生物医学图像作为上下文，提供了 5 个多个自然语言问题，助手针对闭集和开集问题以自由形式的文本响应，每个闭集问题的提示中构建的候选答案列表问题。

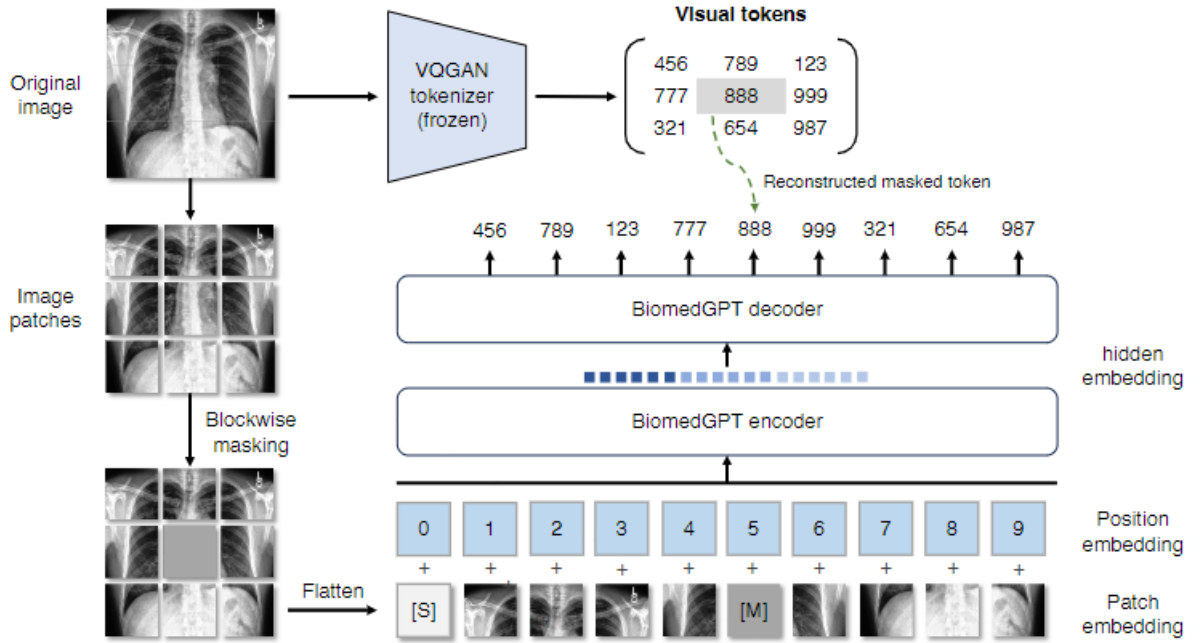


图 3.2 预训练中掩模图像建模的图形说明，旨在通过重构掩模补丁来学习表示

3.2 BiomedGPT

BiomedGPT 是一种专门为生物医学领域设计的基于转换器的架构，建立在现有通用数据统一模型的成功之上。BiomedGPT 设计统一模型的基本原理：1) 模态不可知，2) 任务不可知，3) 模态和任务综合性。与 LLaVA 的相似，BiomedGPT 使用视觉编码器结合语言模型的思想来实现输入/输出统一。其编码器-解码器架构将具有任务相关指令的多模态数据映射到公共表示空间，这有助于解决生物医学模态之间的差异。

3.2.1 预训练任务

BiomedGPT 在预训练中仅考虑两个视觉任务：图像填充（image infilling）与目标检测。对于图像填充，BiomedGPT 借用了块掩码的思想，将输入图像分割成多个小块（或称为“补丁”），然后随机选择其中的一些区域施加掩码。在 BiomedGPT 的预训练阶段，特别是处理这些掩码图像时，模型不仅要识别未被遮盖的部分，还要生成缺失部分的内容。相应的指令是“中间部分的图像是什么？”。如3.2所示。

对于目标检测，训练 BiomedGPT 生成图像中各对象的边界框“图像中的对象是什么？”这一指令引导模型不仅识别图像中的各种对象，还要精确地定位它们的位置。在训练过程中，模型需要分析图像的视觉内容，并预测每个对象的边界框，这包括对象的类型、位置和大小。通过这种方式，模型学习如何在接收到类似查

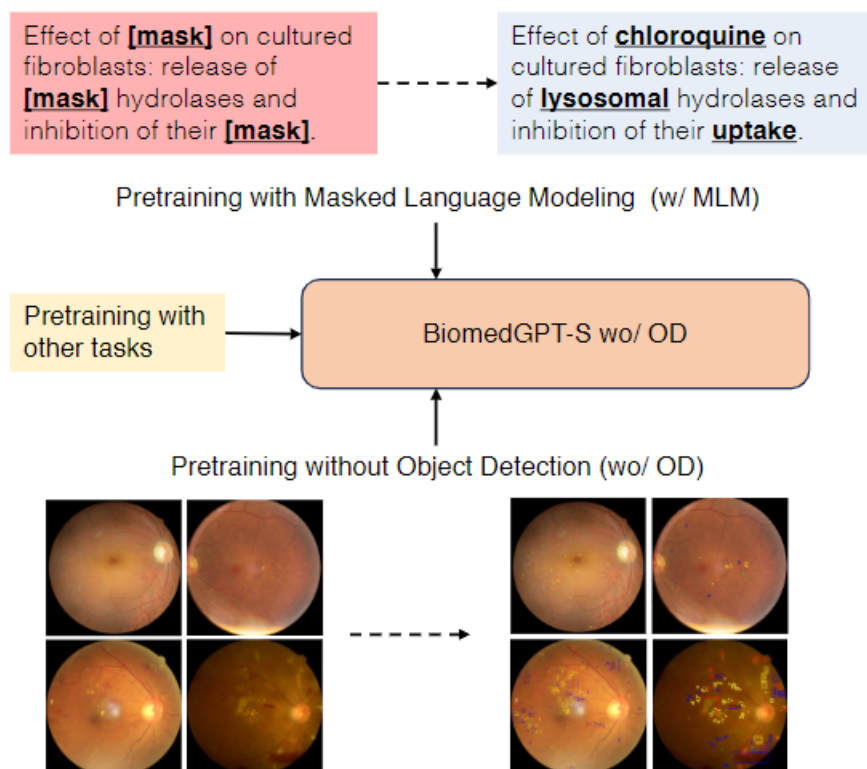


图 3.3 目标检测与掩码语言模型

询时，解析图像并生成准确的对象定位信息。

而对于纯文本任务，与目标检测的预训练策略相类似，BiomedGPT 对文本内容采用了掩码语言建模方法。这一策略中，部分文本内容被掩码（即隐藏），模型的任务是预测这些被掩码的文本。具体的训练指令为：“Text 的完整文本是什么？”这要求模型基于上下文的线索，填补缺失的文本信息，如3.3所示。

3.2.2 自回归生成训练与微调

自回归：这一术语源自统计学，描述的是模型在预测未来值时使用其先前值的方法。在序列到序列学习的上下文中，自回归模型在生成每个新元素时，都会将之前生成的所有元素作为上下文信息。这种依赖之前输出的方式使得生成的序列在语法和语义上更加连贯和准确。

BiomedGPT 采用自回归序列到序列 (seq2seq) 学习，这是大型语言模型训练的常用方法。形式上，假设我们得到一个标记序列 $x_{i,b}$ 作为输入，索引 $i = 1 \dots I$ 标记数据样本， $b = 1 \dots B$ 标记训练批次，模型由 θ 参数化，通过最小化自回归训练

模型:

$$L\theta(x_{1,1}, \dots, x_{i,b}) = - \sum_{b=1}^B \log \left(\sum_{i=1}^I p_{\theta}(x_{i,b} | x_{1,b}, \dots, x_{i-1,b}) \right) = - \sum_{b=1}^B \sum_{i=1}^I p_{\theta}(x_{i,b} | x_{<i,b}) \quad (3.2)$$

在 BiomedGPT 的上下文中, x 可以参考预训练任务中的语言和视觉标记, 包括子词、图像编码和位置标记。子词由 BPE 分词器提取, 位置标记则使用 Pix2Seq 生成如图2.4。在经过数据预处理后, 使用 VQ-GAN(图像编码器) 量化生物学医学图像。

下面是 BiomedGPT 预训练的参数设置:

- (1) 参数 $\beta_1 = 0.9$ 、 $\beta_2 = 0.999$ 和 $\epsilon = 1e-8$ 的 AdamW 优化器
- (2) 峰值学习率设置为 $1e-4$
- (3) 对于正则化, 我们将 dropout 设置为 0.1, 并使用 0.01 的权重衰减
- (4) 使用随机深度, 其速率为 0.1
- (5) 多模态、纯文本、仅视觉和对象检测样本。应用的比例为 8:2:1:1, 强调学习和增强视觉和语言之间的交互。

对于下游任务微调, 采用同样的自回归序列到序列学习。

3.3 本章小结

本章展示了 LLaVA-Med 和 BiomedGPT 两种多模态医疗大模型在预训练与微调阶段的复杂流程和技术细节, 突出了这些技术在提升模型性能中的作用。通过深入的预训练和针对性的微调, 这些模型不仅在理解和生成医学相关内容方面表现出色, 也展现了处理复杂医学数据的潜力。LLaVA-Med 通过与 GPT-4 接口合作, 从生物学医学图像-文本对生成医学对话数据集。这一过程包括概念对齐与指令微调两部分, 旨在通过精细化的数据处理提高模型的语义理解能力。模型在预训练阶段采用了两步法策略, 首先进行概念对齐, 然后进行端到端指令微调, 通过这一策略有效提高了模型对生物学医学图像描述的理解和生成能力。BiomedGPT 则采用了自回归序列到序列学习方法, 处理图像填充与目标检测任务。这种方法强调了视觉与文本信息的融合, 以增强模型的多模态处理能力。

4 实验结果与分析

4.1 实验细节

本实验在 4 块 NVIDIA GeForce RTX 3090 上进行, 测试代码采用 Python 语言与 PyTorch 框架。考虑到调用 GPT-4 接口的费用过高, 在 GPT-4 对 LLaVA-Med 评分阶段替换为 GPT-3.5-Turbo。同时考虑到预训练多模态医疗大模型开销过大, 本地测评采用了开源的模型权重以节省成本。

为了探究当前多模态医疗大模型在生物医学领域做出的进步, 本研究选取了 VQA-RAD、PATH-VQA 和 SLake 三个判别式下游任务与开放式问答生成任务上测试模型推理的性能指标。此外, 借助 gpt-3.5-turbo 将预训练得到的 LLaVA-Med 与 GPT-4 生成的参考答案针对五个模态 (CXR(胸部 X 光)、CT(计算机断层扫描)、MRI(磁共振成像)、组织病理学和粗 (即宏观) 病理) 从相关性、帮助程度、准确度、细节水平上进行对比打分。

4.2 评测指标

在下游任务上, 对于判断任务, 仅需计算模型是否判断正确, 给出其回答准确度即可。对于开放式问答任务, 选取了 F1 分数、准确率、召回率、BLEU 分数与精确匹配得分作为性能的评测指标。

4.2.1 F1 分数、准确率与召回率

F1 分数是一种常用于统计分类任务性能的度量, 特别是在数据集类别分布不均时。它是精确率 (Precision) 和召回率 (Recall) 的调和平均, 用于评估模型对正类预测的准确性和完整性。

- (1) 精确率 (Precision): 也称为正预测值, 是模型预测为正类中实际为正类的比例。公式为:

$$Precision = \frac{TruePositives(TP)}{TruePositives(TP) + FalsePositives(FP)}$$

精确率衡量的是模型的预测结果中有多少是正确的。

- (2) 召回率 (Recall): 也称为真正率或敏感性, 是实际为正类中被模型正确预测

为正类的比例。公式为：

$$Recall = \frac{TruePositives(TP)}{TruePositives(TP) + FalseNegatives(FN)}$$

召回率衡量的是模型找出所有正类案例的能力。

(3) F1 分数是精确率和召回率的调和平均，给出了这两个指标的单一度量。

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

F1 分数通过综合精确率和召回率，提供了一个全面的性能评价，帮助开发者平衡预测结果中的假正例与假反例影响，从而更合理地评估和改进模型的性能。

4.2.2 精准匹配分数

Exact Match Score（精确匹配得分）是一种在自然语言处理（NLP），特别是在机器理解和问答系统评估中常用的指标。它用于衡量系统生成的答案与标准答案之间的精确一致性。如果生成的答案与参考答案完全一致，即完全匹配，那么得分为 1（或 100%），否则得分为 0。通过计算响应中出现的参考答案包含的词汇占比来获取 Exact Match Score。

4.2.3 BLEU 分数

BLEU（Bilingual Evaluation Understudy）分数是一种广泛用于评估机器翻译质量的自动度量方法，也逐渐被应用于其他自然语言生成任务如文本摘要。由 IBM 的研究人员在 2002 年提出，BLEU 分数通过比较机器生成的翻译文本与一个或多个人工翻译的参考文本之间的相似度来工作。其核心思想是：机器翻译得到的结果越接近专业人士的翻译，其质量就越高。BLEU 分数主要关注字面上的 n-gram（n 元语法）重叠。本研究选取 1-gram、2-gram、3-gram 作为参考。

4.3 实验结果

4.3.1 LLaVA-Med 与 GPT-4 性能比较

为了评估 LLaVA-Med 在生物学多模态对话上的性能，我们构建了一个包含 193 个新问题的评估数据集。对于这个测试数据集，我们从 PMC-15M 中随机选择 50 个未参与预训练的图像和字幕对，并生成两种类型的问题：对话和详细描述。其中对话数据由 3.1 提到的方法获得。采用 GPT-3.5 对 LLaVA-Med 生成的回答从相关性、帮助程度、准确度、细节水平上与参考回答进行对比打分（0-10 分，以

GPT-4 的回答作为 10 分)。得到如下结果4.1:

表 4.1 多模态聊天指令跟踪能力的性能比较，通过 gpt-3.5 评估的相对分数来衡量

模型	对话 (143)	描述 (50)	胸部 X 光 (37)	磁共振成像 (38)	组织学 (44)	粗略解剖学 (34)	CT(40)	总计
LLaVA	62.94455	44.14705	66.4575	52.4933	65.4650	46.670168	57.1875	58.0747
LLaVA-Med	81.24375	65.87002	86.80180	67.83521	79.63855	77.36344	74.6875	77.26092

无论是从响应类型（对话/描述），还是不同的模态，具有完整两阶段训练的 LLaVA-Med 始终优于通用领域 LLaVA。对于最好的 LLaVA-Med，从 GPT-3.5 的视角来看，它匹配了 GPT-4 的 77.26% 性能。

4.3.2 下游任务适应能力探究

VQA-RAD、PATH-VQA 和 SLAKE 三个数据集的组成如下表所示4.2:

表 4.2 下游数据集官方给出的训练，测试，验证样本划分

数据集	VQA-RAD		SLAKE			PathVQA		
	训练	测试	训练	验证	测试	训练	验证	测试
# 图像	313	203	450	96	96	2599	858	858
# 问答对	1797	451	4919	1053	1061	19755	6279	6761
# 开放式	770	179	2976	631	645	9949	3144	3370
# 闭集	1027	272	1943	422	416	9806	3155	3391

比较 LLaVA-Med 与 BiomedGPT 在下游数据集上的表现，考察模型对下游任务的适应能力。

VQA-RAD: 表中各项数据的单位均为%。

模型	F1 分数	精确率	召回率	精确匹配得分	一元语法 bleu	二元语法 bleu	三元语法 bleu	闭集判断准确率
LLaVA-Med (微调前)	4.72	2.85	19.9957	2.3788	2.37878	0.366806	0.0130429	50.3827
LLaVA-Med-VQA	91.99	92.67	91.83	92.6	91.35	45.05	25.66	82.44
BiomedGPT-VQA	90.26	91.37	89.16	90.77	90.36	36.23	16.65	81.3

表 4.3 大模型在 VQA-RAD 上的表现

从表4.3中可以看出，LLaVA-Med 和 BiomedGPT 在该数据集上的回答都取得了非常优异的表现, 在开放式问答上几乎能复现原回答，在闭集任务也就是判断任务上，其成功率也达到了 80% 以上。二者性能相差不打，LLaVA-Med 略优于 BiomedGPT。

实验还单独测量了微调前的 LLaVA-Med 各项指标，与微调后的结果差距较大。这是因为在微调前模型无法完成下游任务上的指令遵循，在开放式问答上，无法按照 VQA-RAD 指定的回答方式生成响应，故而即使有部分医疗领域的专业知识，依然无法取得高质量回复。

SLAKE:

由于大模型的预训练阶段采用的均为英文文本数据，从该双语数据测试集中过滤得到纯英文样例，下表4.4为 LLaVA-Med 和 BiomedGPT 在 SLAKE 上微调后的测试结果：

模型	F1 分数	精确率	召回率	精确匹配得分	一元语法 bleu	二元语法 bleu	三元语法 bleu	闭集判断准确率
LLaVA-Med-SLAKE	83.4233	83.5599	83.99	82.84	882.26	22.91	6.63	72.35
BiomedGPT-SLAKE	88.36	87.64	89.09	90.00	91.62	46.31	25.66	87.3

表 4.4 大模型在 SLAKE 上的表现

可以看到，两模型在 SLAKE 上取得优异表现，在该任务上 BiomedGPT 优于 LLaVA-Med。

PATH-VQA: LLaVA-Med 和 BiomedGPT 在 PATH-VQA 上的表现如4.5所示。相对于前两个下游任务，在该数据集上的开放式问答上，LLaVA-Med 和 BiomedGPT 所面临的挑战显著增加。由于开放式问题的答案可能具有更高的变异性，并且模型生成的回答需要更准确地反映出医学图像的细节，两个医疗大模型没能给出较为优质的答案。而在闭集任务上，两个模型表现出的高准确率可能得益于对训练数据的良好拟合，这意味着模型能够有效地识别和回忆在训练期间学到的模式和关联。此外，闭集任务的固定答案范围限制了模型需要考虑的可能性，从而简化了推理过程。

模型	F1 分数	精确率	召回率	精确匹配得分	一元语法 bleu	二元语法 bleu	三元语法 bleu	闭集判断准确率
LLaVA-Med-PVQA	37.65	38.63	37.82	38.59	36.502	9.28	3.402	91.0824
BiomedGPT-PVQA	28.95	27.33	30.74	32.56	29.31	10.71	1.66	87.55

表 4.5 大模型在 PATH-VQA 上的表现

4.3.3 零样本迁移学习

虽然 LLaVA-Med 的训练阶段不包括中文指令跟踪数据，但在 SLAKE 提供的中文测试集上，LLaVA-Med 能够理解中文问题并生成正确的英文响应。4.6

这可能源于最初的语言模型 llama 学习到的多语言知识。

提示	模型回答
这张图片的成像方式是什么?	MRI
这张 MRI 图片用的什么加权方式?	T2 weighted
肝脏在哪个位置?	Right lobe
异常病变在哪个位置?	Right lung

表 4.6 中文问题零样本迁移

5 总结与展望

5.1 工作总结

本研究探讨了基于 Transformer 的模型 LLaVA-Med 和 Biomedgpt 在生物医学领域的应用，特别是它们在多模态任务中的性能表现。通过对这些模型在不同生物医学数据集上进行预训练和微调，我们评估了它们在准确率、召回率、BLEU 分数和 F1 分数等多个方面的表现。研究显示，这些基于 Transformer 的模型不仅能有效适应多种医学任务，还能处理复杂的多模态数据输入，如结合医学图像描述与自然语言处理。

研究结果表明，经过微调的 LLaVA-Med 模型在医学多模态任务上能够达到 GPT-4 性能的 77%，证明了其在处理医学领域特有问题上的强大潜力。此外，LLaVA-Med 和 Biomedgpt 在经过针对性微调后，表现出在减少幻觉和提高回答质量方面的显著进步，这对于提高医学诊断和研究的准确性和可靠性具有重要意义。

5.2 未来展望

多模态医疗大模型作为医学人工智能领域的前沿技术，整合了来自多个数据源的信息，如图像、文本和声音等，以解决复杂的医学问题。该领域虽然已经取得了显著的研究进展，但仍然存在诸多挑战和发展潜力：

数据集和标准化问题：为了充分发挥多模态医疗大模型的潜力，需要构建更为全面和标准化的医学数据集。这不仅需要医学专家对数据进行精确的分类和标注，还需要确保数据在收集和处理过程中的高质量和一致性。

集成更多数据类型：随着医疗科技的发展，可用于诊断和治疗的数据类型也在不断增加。未来的多模态医疗模型将可能集成更广泛的数据类型，如基因组数据、生化指标、电子病历、实时生理信号等。这些数据的集成能够帮助模型从多个维度全面理解患者的健康状态，从而提供更精准的诊断和个性化的治疗方案。

模型泛化能力的进一步提升：尽管现有的多模态医疗模型在特定任务上已显示出卓越的性能，但它们在面对新的、未见过的任务时往往难以保持同样的效果。因此，未来的研究需要探索如何通过更深层的语义理解和上下文分析，进一步提高模型的泛化能力。

这些挑战的解决将极大推动医疗人工智能领域的发展，使多模态医疗大模型不仅能在实验室环境中表现出色，也能在真实世界的医疗实践中发挥关键作用，最终实现改善患者健康和优化医疗资源的目标。

参考文献

- [1] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J/OL]. CoRR, 2017, abs/1706.03762. <http://arxiv.org/abs/1706.03762>.
- [2] NIU S, LIU Y, WANG J, et al. A decade survey of transfer learning (2010–2020) [J/OL]. IEEE Transactions on Artificial Intelligence, 2020, 1: 151-166. <https://api.semanticscholar.org/CorpusID:232151586>.
- [3] KRALJEVIC Z, SHEK A, BEAN D M, et al. Medgpt: Medical concept prediction from clinical narratives[J/OL]. ArXiv, 2021, abs/2107.03134. <https://api.semanticscholar.org/CorpusID:235755486>.
- [4] MCKINNEY SM G V G J A N A H B T C M C G D A E M G V F G F H B M H D J S K A K C K D L J M Sieniek M. Addendum: International evaluation of an ai system for breast cancer screening[J]. Nature., 2020 Oct;586(7829):E19.
- [5] BOMMASANI R, HUDSON D A, ADELI E, et al. On the opportunities and risks of foundation models[J/OL]. CoRR, 2021, abs/2108.07258. <https://arxiv.org/abs/2108.07258>.
- [6] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners [J/OL]. CoRR, 2020, abs/2005.14165. <https://arxiv.org/abs/2005.14165>.
- [7] 陈润生. 医疗大数据结合大语言模型的应用展望[J/OL]. 四川大学学报(医学版), 2023, 54(5): 855-856. <https://ykxb.scu.edu.cn/cn/article/doi/10.12182/20230960301>.
- [8] LI C, WONG C, ZHANG S, et al. Llava-med: Training a large language-and-vision assistant for biomedicine in one day[M]. 2023.
- [9] ZHANG K, YU J, YAN Z, et al. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks[M]. 2023.
- [10] LIU B, ZHAN L, XU L, et al. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering[J/OL]. CoRR, 2021, abs/2102.09542. <https://arxiv.org/abs/2102.09542>.
- [11] LAU J J, GAYEN S, BEN ABACHA A, et al. A dataset of clinically generated

- visual questions and answers about radiology images[J/OL]. Scientific Data, 2018. <http://dx.doi.org/10.1038/sdata.2018.251>.
- [12] HE X, ZHANG Y, MOU L, et al. Pathvqa: 30000+ questions for medical visual question answering[J/OL]. CoRR, 2020, abs/2003.10286. <https://arxiv.org/abs/2003.10286>.
- [13] OPENAI, ACHIAM J, ADLER S, et al. Gpt-4 technical report[M]. 2024.
- [14] LI Y, LI Z, ZHANG K, et al. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge[M]. 2023.
- [15] HAN T, ADAMS L C, PAPAIOANNOU J M, et al. Medalpaca – an open-source collection of medical conversational ai models and training data[M]. 2023.
- [16] WU C, ZHANG X, ZHANG Y, et al. Pmc-llama: Further finetuning llama on medical papers[J/OL]. ArXiv, 2023, abs/2304.14454. <https://api.semanticscholar.org/CorpusID:263888272>.
- [17] 童同, 李雪, 朱永北, 等. 基于多模态大模型的动作识别关键技术研究与应用[J]. 通信世界, 2023(22): 48-49.
- [18] MOOR M, BANERJEE O, SHAKERI Z, et al. Foundation models for generalist medical artificial intelligence[J].
- [19] THEODORIS C V, XIAO L, CHOPRA A, et al. Transfer learning enables predictions in network biology[J/OL]. Nature, 2023, 618(7965): 616—624. <https://europepmc.org/articles/PMC10949956>.
- [20] TU T, AZIZI S, DRIESS D, et al. Towards generalist biomedical ai[J]. 2023.
- [21] LIU J, WANG Z, YE Q, et al. Qilin-med-vl: Towards chinese large vision-language model for general healthcare[M]. 2023.
- [22] BUBECK S, CHANDRASEKARAN V, ELDAN R, et al. Sparks of artificial general intelligence: Early experiments with gpt-4[M]. 2023.

致谢

随着本篇毕业论文的完成，我不禁回想起从确定选题到设计实验、从应对挑战到寻找解决方案、再到论文最终成稿的整个过程。在此期间，我有幸得到了杜博老师与刘菊华老师的精心指导与无私帮助。他们不仅在学术上给予我宝贵的指点，更在精神上给予我莫大的鼓励与支持。对此，我表示最深切的感激和最高的敬意。

我还要感谢实验过程中给予我无私帮助的所有师兄和同学。你们的耐心答疑和宝贵经验分享，对我完成论文具有不可估量的价值。在此向你们致以衷心的感谢。

这四个多月的毕业设计，不仅是对一个课题的探索，更是我四年大学生活的缩影和升华。我感谢武汉大学为我提供的宽广舞台和卓越资源，使我能在珞珈山的怀抱中汲取知识、丰富人生。我感激计算机学院的每一位老师，你们的教诲如灯塔，照亮了我前行的道路。我亦要感谢一路上给予我支持和帮助的所有师长与同侪，正是你们共同编织了我大学生生活中最灿烂的篇章。

最后，我必须要感谢我的父母，你们默默的付出和坚强的背后支持，是我能够在学业上不断前行的强大动力。在未来的旅途中，我将带着感恩的心，珍惜这段难忘的经历，并以此为起点，继续追寻学术与人生的更高峰。感谢你们，这份谢意，将随我一生。