

基于 ChatGLM 微调的医疗问答系统

姚杰恺

广东实验中学，广东 广州 510380

摘要：为提升大模型在医疗领域的专业性，本文采用 Lora 微调技术，利用 huatuo26M 数据集，对 ChatGLM3-6B 大模型进行微调，构建医疗问答系统。研究结果显示，该方法显著提高了医疗问答的专业度、准确性及对话流畅性。该系统在医疗咨询与健康指导中展现应用价值，并具备推广至其他专业领域大模型微调的潜力。

关键词：医疗问答系统；ChatGLM 模型；大模型；Lora 微调

A Medical Q&A System Fine Tuned Based on ChatGLM

Yao Jiekai

Guangdong Experimental High School, Guangzhou, Guangdong 510380

Abstract : To enhance the professionalism of the ChatGLM3-6B big model in the medical field, this article adopts Lora fine-tuning technology and utilizes the Huatuo26M dataset to fine tune the ChatGLM3-6B big model and construct a medical question answering system. The research results show that this method significantly improves the professionalism, accuracy, and fluency of medical Q&A. This system demonstrates application value in medical consultation and health guidance, and has the potential to be promoted to other professional fields for large-scale model fine-tuning.

Keywords : medical question answering system; ChatGLM model; large model; Lora fine-tuning

一、引言

随着信息技术的飞速发展和互联网医疗的普及，医疗问答系统作为连接患者与医疗专业知识的桥梁，其重要性日益凸显。传统的医疗咨询方式受限于时间、地点和专家资源，难以满足广大患者的即时需求。而医疗问答系统通过自然语言处理技术，能够24小时不间断地为用户提供准确、可靠的医疗信息，有效缓解了医疗资源紧张的问题。

ChatGLM3-6B^[1]作为新一代的大型语言模型，具备强大的语言理解和生成能力，为构建高性能的医疗问答系统提供了坚实的基础。然而，直接将通用领域的 ChatGLM3-6B 模型应用于医疗领域，可能会因领域差异而导致回答不够准确或专业。因此，通过微调技术将 ChatGLM3-6B 模型适应于医疗领域，成为提升医疗问答系统性能的关键^[2]。

近年来，基于深度学习的自然语言处理技术在医疗领域取得了显著进展^[2-5]。许多研究致力于将大型语言模型应用于医疗问答系统，以提高系统的专业性和准确性。然而，目前的研究多集中在特定疾病或症状的问答上，缺乏全面的医疗知识覆盖和高效的模型微调方法。此外，如何在保证模型性能的同时，减少计算资源消耗和避免知识遗忘，也是当前研究面临的重要挑战。

本文的研究贡献主要有：采用低秩适应（Low-Rank Adaptation, Lora）微调技术^[6]，利用 huatuo26M^[7]数据作为训练集，对 ChatGLM3-6B 模型进行医疗领域的适应性训练，提高模型在医疗问答任务中的专业性和准确性，以及对话流畅性。

二、ChatGLM3-6B 模型与微调技术

ChatGLM3-6B 是基于 Transformer 结构开发的一种大型语言模型，拥有 60 亿参数，具备强大的文本生成和理解能力。该模型在多个自然语言处理任务中表现出色，如文本分类、情感分析、问答系统等。

微调是一种将预训练的大型语言模型适应于特定任务或领域的技术。通过在预训练模型的基础上，使用特定领域的数据进行训练，可以使得模型更好地适应新任务或领域。微调技术可以显著提高模型在特定任务上的性能，同时减少对新领域数据的依赖。

Lora 微调是一种轻量级的微调方法，它通过向预训练模型中添加低秩矩阵来实现参数的更新。这种方法可以在不改变模型主体结构的情况下，实现模型的快速适应，同时减少计算资源的消耗。另一种常用的微调技术是 Freeze 微调，在训练过程中，它固定模型的大部分参数不变，只对部分参数进行微调，保持模型的稳定性。Lora 微调相比 Freeze 微调，以低秩分解减少参数，提升训练效率与灵活性，同时保持模型性能，减少显存占用，不增加推理延迟。因此，选择用 Lora 微调展开实验。

三、医疗问答系统的构建

(一) 数据来源

医疗问答系统需要大量的数据来训练模型，使其能够理解用户提出的问题，并给出准确的答案。通过收集医疗问答数据，可以为

作者简介：姚杰恺，男，主要研究领域为人工智能。

用户提供更准确、更及时的医疗健康信息。这有助于改善用户的体验，提高用户对问答系统的满意度，提升系统效果和性能。为此，本次实验所用的数据集是在 Hugging Face 网站的 huatuo26M 数据集。Huatuo-26M 是一个大规模的医疗问答数据集，以中国古代伟大的医生华佗命名，含有超过 2600 万个问答对，是目前最大的中文医疗问答数据集。其中，训练集和测试集的比例为 9:1。数据来源包括医疗问答网站、健康论坛、医学资讯网站等多种渠道。这些数据来源覆盖了各种医疗健康问题，涵盖了疾病、症状、诊断、治疗方法、用药指导等多个方面的话题。

训练前先要对数据进行预处理，包括数据清洗、分词与去除停用词、数据转换等操作，保证数据集转换成训练所需的格式。

(二) ChatGLM3-6B 模型微调策略

ChatGLM3-6B 模型的微调策略主要涉及针对特定任务或数据集对预训练模型进行调整，以便更好地适应新的应用场景。微调过程通常包括准备相关数据集、选择合适的优化器和学习率，以及进行多轮训练等步骤。在微调过程中，需要密切关注模型的性能变化，及时调整参数以获得最佳效果。Lora 微调的实施细节如下：

(1) 确定微调目标：首先明确微调的具体目标，例如提高模型的准确性、降低计算成本或优化特定任务的性能。

(2) 数据准备：收集并整理用于微调的数据集。这些数据应该与微调目标密切相关，并具备足够的多样性和代表性。

(3) 模型加载：加载预训练的 ChatGLM3-6B 模型作为微调的基础。

(4) Lora 层引入：在模型中引入额外的 Lora 层。这些层通常具有较少的参数，但能够有效地捕获任务相关的特征。

(5) 模型训练：使用准备好的数据集对模型进行训练，并通过调整学习率、优化器等参数来优化训练过程。模型的损失函数选择交叉熵损失函数。具体公式见式 1。

$$L = -\frac{1}{N} \sum_i -[y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)] \quad (1)$$

其中：表示样本 i 的标签，正类为 1，负类为 0；表示样本 i 预测为正的概率。

四、实验

在本次实验的微调训练工作中，软件和硬件需求不算高。关键的配件 GPU，选用了型号为 RTX 4090(24G)，算是轻量级配置。具体配置信息如表 1 所示。

表 1 模型微调服务器规格

软硬件	配置信息
软件	PyTorch 2.0.0; Python 3.8(ubuntu20.04); Cuda 11.8
GPU	RTX 4090(24GB)
CPU	12 vCPU Intel(R) Xeon(R) Platinum 8352v CPU @ 2.10GHz
内存	90GB
硬盘	系统盘: 30GB; 数据盘: 50GB

实验采用 LLaMA Factory 作为 LLM 微调工具，该工具支持多种 LLM 模型与微调策略，集成前沿方法，并面向非专业人员提

供 WebUI 操作界面。实验前，下载并配置 LLaMA Factory，随后导入数据集，并获取 ChatGLM3 模型及其依赖。利用 AutoDL 平台的网页微调功能，根据平台要求调整了 LLaMA Factory 的端口设置，以访问微调界面。在界面中，选择了 Lora 微调策略，设置训练参数，启动训练流程。训练完成后，通过加载模型，并在系统界面进行测试提问，验证模型效果。最后，成功导出微调模型至本地，以供后续评估与应用。

针对此次模型微调训练的结果，提出“饮酒以后可以服用西地那非吗”的问题，根据 ChatGLM3-6B 微调前后的回复进行对比，如表 2 所示，二者的回复的篇幅差不多，原始的 ChatGLM 的回复还有明显的错误，西地那非的英文名是 Sildenafil，而不是 Cialis (Cialis 是另一种壮阳药他达那非的英文名称)。而微调之后的 ChatGLM 没有这个幻觉回复。ChatGLM 微调之后的回复更接近人类医生角色的回复，而微调前的回复更像知识的摘抄和罗列。

表 2 针对问题“饮酒以后可以服用西地那非吗” ChatGLM3-6B 微调前后的回复

模型	输出
ChatGLM	<p>西地那非 (Cialis, 亦称伐地那非, Viagra) 是一种用于治疗勃起功能障碍 (ED) 的药物。癌症治疗和西地那非的使用需要谨慎评估，因为癌症和其治疗可能对性功能产生影响，而西地那非可能与其他癌症治疗药物有相互作用，或者在某些癌症类型中可能不适宜使用。</p> <p>在考虑是否使用西地那非时，应咨询医生或医疗专业人员，他们会根据患者的具体情况来决定是否适合使用西地那非：</p> <p>癌症类型和治疗：某些类型的癌症和治疗（如化疗、放疗）可能会影响性功能。医生会评估癌症的类型和治疗方式，以确定是否适合使用西地那非。</p> <p>药物相互作用：西地那非与其他药物（如某些用于治疗心脏病的药物、高血压药物、某些抗生素和抗真菌药）可能有相互作用，影响药物的效果或增加副作用的风险。医生需要评估患者的整体用药情况。</p> <p>副作用和安全性：西地那非可能会引起一些副作用，如头痛、面部潮红、消化不良等。在癌症背景下，可能需要更加小心地评估这些副作用。</p> <p>患者的整体状况：患者的整体健康状况、年龄、其他并发症和偏好也是考虑的因素。</p> <p>总之，癌症患者是否可以使用西地那非应由医生进行全面评估后决定。如果医生认为使用西地那非是安全的并且有益的，他们可能会开具处方。重要的是，患者不应自行服用西地那非或其他任何药物，而应始终在医生的指导下进行治疗。</p>
微调之后的 ChatGLM	<p>如果您被诊断为癌症患者，应避免使用西地那非 (Sildenafil)。西地那非是一种 PDE5 抑制剂，可能会与某些癌症治疗方法（如化疗、放疗或激素治疗）相互作用，从而降低癌症治疗的疗效或增加副作用的风险。此外，一些研究表明 PDE5 抑制剂可能促进某些类型癌症（如肺癌、乳腺癌和前列腺癌）的生长和进展。西地那非还可能增加出血风险，这对于已经因癌症或其治疗而存在出血风险的患者来说可能是个问题。关于西地那非在癌症患者中的安全性和功效的研究有限，长期在此人群中使用该药物的效果尚不清楚。</p> <p>因此，最重要的是与您的医疗服务提供者和肿瘤科医生讨论您的治疗选择，以确定在您的情况下最佳的行动方案。他们可以帮助您权衡不同治疗的潜在益处和风险，包括西地那非，并推荐对您安全有效的替代药物或治疗方法。请记住，始终将您的健康放在首位，并遵循医疗专业人员的指导。</p>

五、总结

本文采用 Lora 微调技术，利用 huatuo26M 数据集进行训练，构建高效、准确的医疗问答系统。经过微调后的医疗问答系统在准确性方面表现出色，充分证明了微调策略的有效性，为医疗行业信息服务提供了强有力的技术支持。这不仅有助于改善患者的就医体验，还能提高医疗资源的利用效率，对推动医疗服务的智能化发展具有重要意义。本文不仅验证了大型语言模型在

特定领域微调后的优越性能，还为医疗问答技术的发展提供了有力支持。进一步地，本文的微调策略和方法论为跨领域智能问答系统的构建提供了一定的参考，可以推广到更多行业，如教育、金融等，推进各领域智能问答系统的升级。

本文用于微调训练的数据有限，今后研究可以加大训练数据，也可以增加如药物数据等，提高问答系统的能力。也可以尝试更多的微调方法，如监督微调（Supervised Fine-Tuning, SFT）、前缀微调（Prefix Tuning）等，以得到更好的效果。

参考文献：

-
- [1] Du Z, Qian Y, Liu X, et al. GLM: General Language Model Pretraining with Autoregressive Blank Infilling [J]. 2021. DOI: 10.18653/v1/2022.acl-long.26.
 - [2] 张伟, 李明. 基于深度学习的医疗问答系统研究综述. 计算机工程与应用, 2022. 58(12), 1-10.
 - [3] 王晓霞, 刘强. 面向医疗领域的自然语言处理技术研究进展. 智能科学与技术学报, (2021). 3(2), 152-160.
 - [4] 赵瑞, 陈志军. 大型语言模型在医疗领域的应用与挑战. 计算机科学, 2020. 47(6), 1-8.
 - [5] 李华, 张敏. 自然语言处理技术在医疗问答系统中的应用. 计算机应用研究, 2019. 36(10), 2925-2930.
 - [6] Hu E J, Shen Y, Wallis P, et al. LoRA: Low-Rank Adaptation of Large Language Models [J]. 2021. DOI: 10.48550/arXiv.2106.09685.
 - [7] Li Jianquan, Wang Xidong, Wu Xiangbo, et al. Huatuo-26M: A large-scale Chinese medical question and answer dataset [J], 2023. arXiv. <https://arxiv.org/abs/2305.01526>.