Developing a Radiology Chatbot for Medical Scan Interpretation

Veda Sahaja Bandi

December 12th, 2024

1 Abstract

Medical imaging interpretation is a critical yet time-intensive process in healthcare, often leading to diagnostic delays as radiology professionals struggle with the exponential growth of imaging data. To address these challenges, this research introduces a Visual Question Answering System (VQA) that leverages advanced vision-language models and deep learning techniques to help radiologists efficiently analyze medical images. By integrating computer vision with natural language processing, we developed a VQA chatbot capable of interpreting medical scans and answering radiology-specific questions. The system is fine-tuned using the LLaVA-Med model and datasets such as VQA-RAD, SLAKE, and MED-VQA, employing Low-Rank Adaptation (LoRA) for optimal performance. The results demonstrate the potential of this approach to streamline diagnostic workflows, improve diagnostic accuracy, and reduce clinical decision-making delays, offering a transformative solution to the growing demands of modern radiology.

2 Introduction

The exponential growth of medical imaging data presents significant challenges in modern healthcare. Radiologists are now confronted with increasingly complex tasks that demand high precision and efficiency. They must navigate through progressively complex medical scans, manage enormous volumes of imaging data, and consistently provide timely and accurate diagnoses.

Traditional diagnostic workflows are fundamentally constrained by human limitations. The manual process of image interpretation is inherently time-intensive, and the potential for human error increases with the complexity of medical scans. These constraints create significant bottlenecks in healthcare delivery, potentially impacting patient outcomes and medical efficiency.

Visual Question Answering (VQA) in radiology emerges as a promising technological intervention to address these challenges. By combining computer vision and natural language processing, we can develop intelligent systems that transcend traditional diagnostic boundaries. Such systems have the potential to assist radiologists in image interpretation, provide rapid and contextually rich analysis of medical images, and substantially reduce the cognitive load on medical professionals.

Our project aims to develop a VQA chatbot tailored to radiology, enhancing diagnostic workflows and reducing human error. By integrating advanced AI models, this research seeks to address the challenges posed by the increasing demand for precise and efficient diagnostic tools.

3 Related Work

The landscape of medical VQA has seen several notable approaches with distinct characteristics. MedCLIP, despite achieving comparable zero-shot prediction accuracy, faces significant practical constraints[3]. The model's performance is fundamentally limited by insufficient pretraining data, rendering it challenging to deploy in real-world medical scenarios.

BiomedCLIP suffers from critical inconsistencies, with its zero-shot capabilities undermined by randomness stemming from statistical prompt relationships[4]. This inherent variability compromises its reliability in precision-demanding medical diagnostic scenarios[5].

Previous approaches like PubMedCLIP treated Medical VQA as a restrictive classification problem, artificially constraining vision-language models' capabilities and introducing significant evaluation inaccuracies[2]. A more advanced model, LLaVA-Med, represents a significant leap forward. This vision-language model integrates large language models and is specifically designed for medical visual understanding, providing more comprehensive interpretation of medical scenarios[1].

Existing approaches in the field have predominantly concentrated on general biomedical image understanding, text-image embedding techniques, and have shown limited interactive question-answering capabilities. Our research distinguishes itself by targeting specialized radiology VQA, implementing parameter-efficient fine-tuning, and developing an interactive diagnostic support system.

4 Problem Definition

The exponential growth in medical imaging data poses significant challenges to timely and accurate diagnosis. Manual interpretation remains the primary method, which is not only time-consuming but also susceptible to human error.

The objective of this project is to create an AI-driven chatbot that interprets radiology scans and answers domain-specific questions. By combining computer vision and natural language processing, this system aims to:

- Address the growing volume of imaging data.
- Reduce the time-intensive nature of manual interpretation.
- Provide reliable, real-time decision support for radiologists and clinicians.

5 Methodology

Our approach integrates existing frameworks with task-specific optimizations to build an efficient VQA system for radiology. The methodology consists of the following steps:

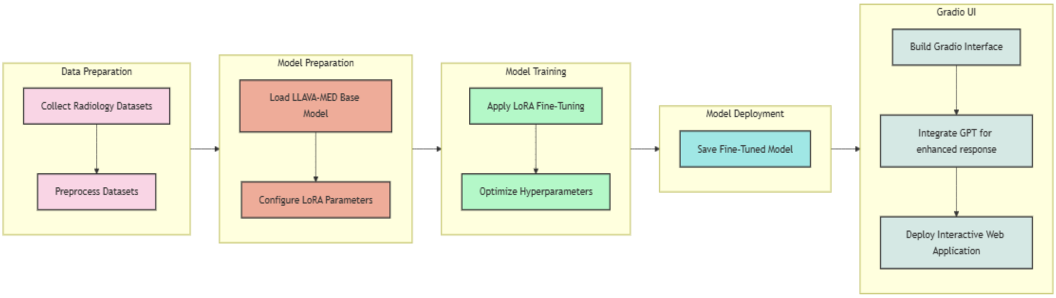


Fig 1: Methodology Pipeline

5.1 Data Preparation

We utilized three key datasets: VQA-RAD[10], SLAKE[12], and MED-VQA[11]. These datasets were preprocessed to ensure compatibility with the fine-tuning process. Preprocessing steps included resizing images, normalizing pixel values, and organizing question-answer pairs to align with the model's input format.

5.2 Model Preparation

The LLaVA-Med model, with 7 billion parameters[7], was selected as the base model for this project due to its advanced ability to integrate visual and textual contexts effectively. This model, which leverages state-of-the-art transformer architectures, is pre-trained on medical imaging and language tasks, making it highly suitable for handling multimodal data. The choice of LLaVA-Med is motivated by its proven capability in processing and understanding complex radiological and clinical information.

However, fine-tuning all layers of such a large model is a computationally expensive task. Given the scale of LLaVA-Med, directly fine-tuning all 7 billion parameters would require significant resources and time. To overcome this challenge, we implemented Low-Rank Adaptation (LoRA)[6], a method designed to optimize the fine-tuning process while minimizing computational costs. LoRA works by introducing small, trainable matrices into specific layers of the transformer architecture. By targeting critical layers such as the query, key, and value projections within the transformer blocks, LoRA enables efficient fine-tuning of the model, retaining the overall architecture while adapting it for radiology-specific tasks. This method not only improves efficiency but also enhances scalability, allowing the model to be adapted to diverse hardware configurations.

This approach is especially beneficial in contexts like medical VQA, where the need for fine-tuning with domain-specific data is crucial, yet computational resources are often limited. By implementing LoRA, we ensure that the model remains both powerful and computationally feasible for practical applications in radiology.

5.3 Model Training

Fine-tuning the LLaVA-Med model involved training on the prepared datasets to tailor the model for radiology-specific tasks. A carefully optimized learning rate over five epochs ensured stable and effective training, allowing the model to adapt efficiently. Mixed precision training with bfloat16 was employed to reduce memory usage and accelerate computations, making the process feasible on GPUs with limited resources while maintaining the precision required for multimodal learning.

The training process utilized the Paged AdamW optimizer, renowned for its ability to handle large-scale models efficiently. By leveraging paged memory techniques, this optimizer facilitated stable convergence without overwhelming GPU memory. These choices ensured the fine-tuning process was computationally efficient while enabling the model to handle complex medical question-answering tasks with precision and scalability.

5.4 Model Evaluation

The fine-tuned model's performance was evaluated using a multifaceted approach to ensure robustness and reliability. The primary evaluation metric was training loss, monitored across five epochs to observe convergence and stability. Additionally, exact match accuracy was employed to measure the model's ability to generate correct answers to specific queries.

5.5 Model Deployment

The fine-tuned model was deployed on Hugging Face Spaces at [8] to ensure scalability and accessibility for real-world applications. This platform provides a robust infrastructure for hosting machine learning models, making them publicly accessible and easy to integrate into various workflows. Deployment involved saving the optimized model and preparing it for seamless interaction, enabling both technical and non-technical users to utilize its capabilities effectively.

5.6 User Interface Using Gradio

A Gradio-based user interface (UI) was developed to simplify interaction with the model. The UI allows users to upload radiology images and pose questions through a conversational format. Key features include real-time image processing, dynamic question-answer generation, and GPT integration for providing more detailed and contextually rich answers. This integration enhances the model's ability to deliver nuanced responses, improving its practical utility in radiology-specific applications. The user-friendly design ensures accessibility for a broad audience, bridging the gap between advanced AI functionality and healthcare needs. The UI has been hosted on Hugging Face at [9].

6 Experimental Setting

The experimental setup in our study utilized three key datasets—VQA-RAD, SLAKE, and MED-VQA—each chosen for their relevance to radiology and clinical applications. These datasets were instrumental in training and evaluating the model on specialized medical question-answering tasks.

- VQA-RAD[10] consists of 1.7k training images and 451 testing images, focusing on clinical images with associated radiology-specific questions and answers, ideal for assessing model performance in real-world medical settings.
- SLAKE[12] includes 4.92k training images and 1.06k testing images, providing a larger corpus that covers a wide range of radiological questions, ensuring a robust evaluation of the model's generalization capabilities across diverse medical scenarios.
- MED-VQA[11] contains 635 training images and 159 testing images, offering a smaller but still relevant dataset for fine-tuning, particularly beneficial for focusing on the intricacies of medical image interpretation.

The model was fine-tuned using NVIDIA RTX A100 GPUs and later deployed on NVIDIA A10G GPUs, allowing for efficient training and inference. Hyperparameters were carefully selected, including an optimized learning rate over five epochs, mixed precision training with bfloat16 to accelerate computations, and the Paged AdamW optimizer, which provided efficient training on constrained hardware.

For performance evaluation, metrics such as training loss, exact match accuracy were employed to quantify the model's ability to understand and respond to medical queries accurately. These metrics are critical for measuring the model's success in providing clinically relevant and precise answers based on radiological images.

This setup underscores the significance of domain-specific datasets in refining and validating advanced VQA systems for medical imaging tasks.

7 Experimental Results and Analysis

7.1 Model Performance

The training process of the model was monitored over five epochs, with training loss recorded after each epoch to observe the convergence trend. As shown in the Table 1 and Figure 2, training loss steadily decreased across epochs, demonstrating effective learning and stability. This consistent decline indicated that the model adapted well to the provided datasets, optimizing its parameters for radiology-specific question answering.

Epoch	**Training Loss**
1	1.436800
2	0.751200
3	0.594200
4	0.474500
5	0.370700

Table 1. Training Loss at Different Steps

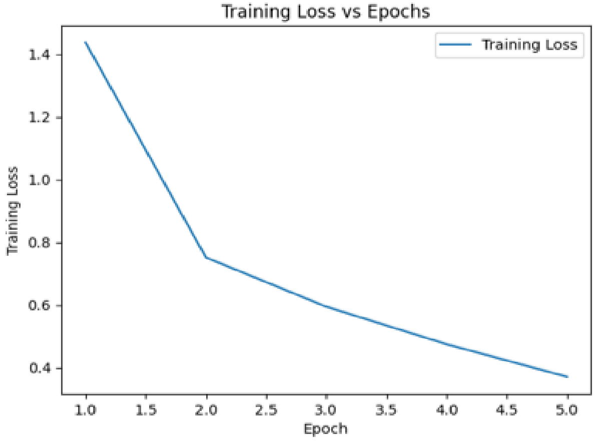


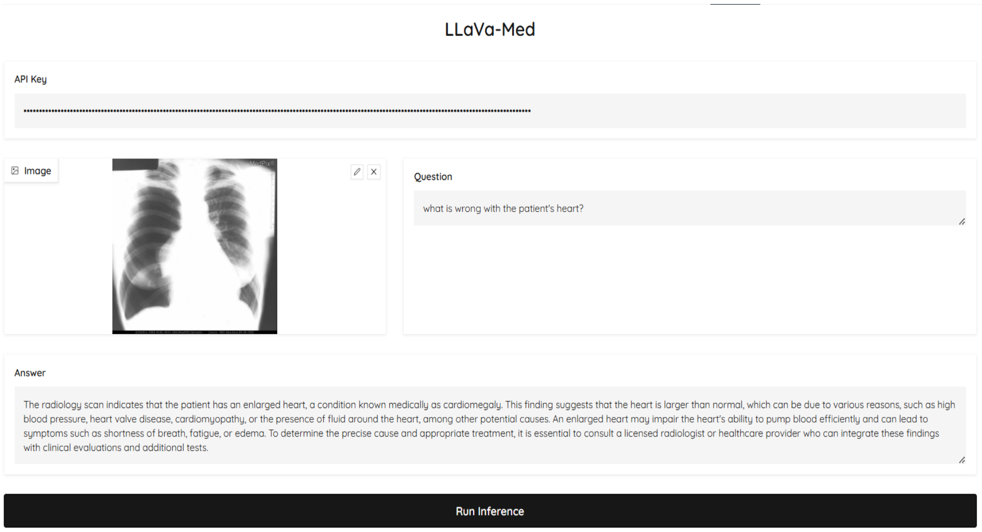
Fig 2: Epoch Vs Training Loss

The testing evaluation was conducted on the VQA-RAD test dataset, where the model achieved an exact match accuracy of 49.45 %. This metric reflects the model's ability to generate precise answers for radiology-specific questions, demonstrating the effectiveness of fine-tuning for this domain. However, exact match accuracy can be particularly challenging, especially for complex and nuanced medical queries.

To obtain a more comprehensive assessment, additional metrics like the BLEU score or GPT-based evaluations would provide better insights into the model's performance. These metrics are more flexible and can capture variations in phrasing and context, offering a broader view of the model's ability to generate accurate and relevant answers in real-world medical settings.

7.2 UI Demonstration

The model was deployed with a user-friendly interface. Users can input an OpenAI API key, upload a medical image, and ask a query related to the radiology domain. The UI will then generate an answer based on the model's interpretation of the image, providing insights tailored to medical imaging. Figure 3 and 4 showcases two examples of the UI demo in action, where users can interact with the system by uploading radiology images and submitting queries. These examples illustrate how the model processes the images in real-time and generates relevant answers to radiology-specific questions.



LLaVa-Med

API Key

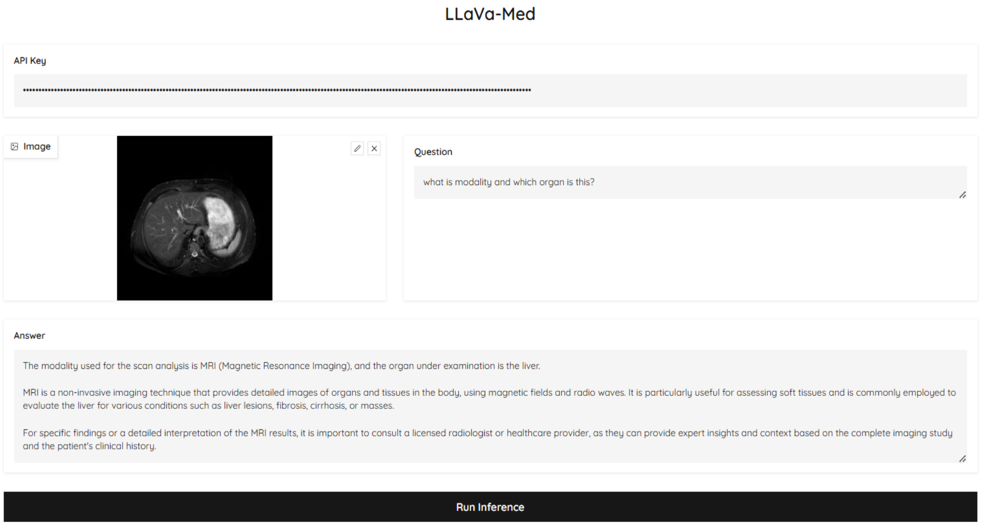
Question

what is wrong with the patient's heart?

Answer

The radiology scan indicates that the patient has an enlarged heart, a condition known medically as cardiomegaly. This finding suggests that the heart is larger than normal, which can be due to various reasons, such as high blood pressure, heart valve disease, cardiomyopathy, or the presence of fluid around the heart, among other potential causes. An enlarged heart may impair the heart's ability to pump blood efficiently and can lead to symptoms such as shortness of breath, fatigue, or edema. To determine the precise cause and appropriate treatment, it is essential to consult a licensed radiologist or healthcare provider who can integrate these findings with clinical evaluations and additional tests.

Fig 3: UI Demonstration - Example 1



LLaVa-Med

API Key

Question

what is modality and which organ is this?

Answer

The modality used for the scan analysis is MRI (Magnetic Resonance Imaging), and the organ under examination is the liver.

MRI is a non-invasive imaging technique that provides detailed images of organs and tissues in the body, using magnetic fields and radio waves. It is particularly useful for assessing soft tissues and is commonly employed to evaluate the liver for various conditions such as liver lesions, fibrosis, cirrhosis, or masses.

For specific findings or a detailed interpretation of the MRI results, it is important to consult a licensed radiologist or healthcare provider, as they can provide expert insights and context based on the complete imaging study and the patient's clinical history.

Fig 4: UI Demonstration - Example 2

7.3 Error Analysis

Error analysis revealed that the model performed well in identifying abnormalities in chest-related scans, but struggled with brain scans and multi-abnormality detection. These difficulties were attributed to the under representation of such images in the training data, which hindered the model's ability to handle these cases accurately.

Moreover, the model showed challenges in addressing ambiguous queries and complex multi-abnormality scenarios. This suggests that additional training data, especially for overlapping and nuanced medical conditions, would improve the model's robustness and help it generalize better across diverse diagnostic situations.

8 Future Work

In future, we will focus on expanding the training dataset to include a wider range of modalities such as brain scans and PET scans. These additions will address current limitations where certain imaging types are underrepresented, such as in Neuroimaging. This diversification will improve the model's ability to interpret a broader spectrum of medical images. Additionally, incorporating ambiguous cases and multi-answer scenarios will allow the model to handle more complex diagnostic situations, enhancing its robustness in real-world clinical settings.

Another important direction for future improvement is evaluating the model's generalization capabilities on unseen datasets. This step is essential to assess whether the model can effectively handle previously unseen cases, which is critical for real-world applications in healthcare. Moreover, interactive features such as chatbot history will be explored, allowing for more dynamic and context-aware interactions. This would enable the model to remember prior questions and answers, improving its contextual understanding and creating a more engaging and intuitive user experience.

9 Conclusion

This project successfully demonstrates the feasibility of fine-tuning the LLaVA-Med model for radiology-specific Visual Question Answering (VQA) tasks. By implementing Low-Rank Adaptation (LoRA), we were able to achieve resource-efficient fine-tuning, making the model more scalable and adaptable for real-world healthcare applications. The fine-tuning process led to significant improvements in the model's ability to interpret and respond to medical images, showcasing the potential of AI in healthcare.

The model's success in interpreting radiology images and answering domain-specific questions highlights the transformative potential of AI in streamlining diagnostic workflows and supporting clinical decision-making. By reducing the cognitive load on medical professionals, such systems can enhance the efficiency and accuracy of medical diagnoses, ultimately improving patient outcomes. Future advancements, such as incorporating more diverse datasets and enhancing interactivity, will continue to push the boundaries of AI's role in healthcare, making it an invaluable tool in medical practice.

References

1. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. Microsoft (2023)
2. Ha, C. N., Asaadi, S., Karn, S. K., Farri, O., Heimann, T., Runkler, T.: Fusion of Domain-Adapted Vision and Language Models for Medical Visual Question Answering. (2024)
3. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. 2022
4. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Tupini, A., Wang, Y., Mazzola, M., Shukla, S., Liden, L., Gao, J., Lungren, M. P., Naumann, T., Wang, S., Poon, H.: BiomedCLIP: A Multimodal Biomedical Foundation Model Pretrained from Fifteen Million Scientific Image-Text Pairs. 2023
5. Van, M.-H., Verma, P., Wu, X.: On Large Visual Language Models for Medical Imaging Analysis: An Empirical Study. 2024
6. LLaVA-Med, https://github.com/onyekaokonji/LLaVA-Med. Last accessed 8 Dec 2024
7. Microsoft LLaVA-Med , https://huggingface.co/microsoft/llava-med-v1.5-mistral-7b. Last accessed 8 Dec 2024
8. Finetuned LLaVA-Med (Deployed on Hugging Face), https://huggingface.co/Veda0718/llava-med-v1.5-mistral-7b-finetuned. Last accessed 9 Dec 2024
9. LLaVA-Med UI (Deployed on Hugging Face) , https://huggingface.co/spaces/Veda0718/Llava-Med. Last accessed 9 Dec 2024
10. Flavia Giammarino, F.: VQA-RAD https://huggingface.co/datasets/flaviagiammarino/vqa-rad
11. Arjun Gupte: MedVQA https://huggingface.co/datasets/agupte/MedVQA
12. I Made Wiratathya Putramas: SLAKE-VQA-English https://huggingface.co/datasets/mdwiratathya/SLAKE-vqa-english