Beyond Imagery: AI-Enhanced Diagnostic Assistantfor Cancer and Tumor Diagnosis using Radiology Imaging

[1]Dr. Nandha Gopal S M, [2]Rahul Sharma, [3]Nithin M, [4]Prajwal B R, [5]Prashanth Kalgonda
[1]Professor, [2,3,4,5]UG Student, Department of Computer Science & Engineering, HKBK College of Engineering, Bengaluru, India

Abstract—Visual Question Answering (VQA), in the field of imaging has the potential to greatly improve decision making and patient outcomes. However, the intricate and unique nature of data presents challenges. In this research we introduce a method that combines Retrieval Augmented Generation (RAG) with cutting edge computer vision and natural language processing techniquesto tackle these challenges effectively. Our model utilizes a collection of literature and case studies to fetch relevantinformation enhancing the generation of precise, thorough and clinically significant responses to inquiries about medical images such as CT scans and MRIs. Moreover, the model can deliver personalized guidance by offering insights and recommendations based on provided images and queries. We assess our approach onselected datasets like SLAKE, Path VQA and VQA Med showcasing enhancements over models. We also examine how retrieval strategies and domain specific knowledge impact the models credibility and interpretability. The findings indicate thatincorporating retrieval based methods can significantly elevate the usefulness of VQA for medical imaging applications paving theway for intelligent and reliable clinical decision support systems applicable, across fields.

Index Terms—Natural Language Processing, Computer Vision, Retrieval-Augmented Generation, Visual Question Answering

I. INTRODUCTION

Over the past several decades, the healthcare industry has experienced a surge of advancements and achievements in technology, with those in medical imaging being especially prominent. The invention of Computed Tomography scans and Magnetic Resonance Imaging has allowed human beings to witness the inside of their bodies on an unprecedented level of detail. It is now possible for a clinician to identify relatively unknown pathologies and monitor the progression of nearly all diseases while devising personalized and targeted treatment schemes for individual patients. The high field images provide clinicians with vast knowledge about the intricate structures and functions of the human body in health and disease and haveshifted the paradigm of medical diagnosis and management.

The incredible advantages offered by these imaging methodsalso pose a challenge. The complex data they produce. Witha number of images being created every day in healthcare facilities globally along, with their intricate nature it becomes amajor hurdle for clinicians to extract valuable information and gain practical insights. Deciphering and studying images calls for knowledge as well as significant time and dedication often resulting in diagnostic delays less, than ideal treatment choicesand higher healthcare expenses.

Visual Question Answering (VQA) is seen as a way to tacklethis issue serving as a connection, between the detailed visuals found in medical images and the descriptive power of everyday language. By allowing users to ask questions about the content of images using language and getting useful data backed answers VQA systems could make medical imaging knowledge more accessible, to everyone giving healthcare providers, scientists and even patients a better grasp of their health concerns.

Nevertheless unlocking the capabilities of Visual Question Answering (VQA), in the field presents distinct and challenging obstacles. Standard VQA models, primarily trained on image datasets frequently encounter difficulties in adjusting to the intricacies and technical jargon found in medical imaging. The specialized characteristics of images require a depth of comprehension and domain knowledge that traditional VQA models might not possess, resulting in inaccuracies, uncertainties and restricted practicality within actual clinical environments.

In order to tackle these problems and unleash the ground- breaking potential of Visual Question Answering (VQA), in

imaging we suggest an approach that combines the collaborative strengths of Large Language Models (LLMs) and the Retrieval Augmented Generation (RAG) framework, supportedby creative AI methods specifically designed for healthcare applications. At the core of our strategy is the fusion of computer vision techniques and natural language processing (NLP) models, in a system known as the Large Language and Vision Architecture (LLaVA).

By using the knowledge stored in language models and combining it with relevant information, from a wide range of medical texts and real life cases our system strives to offer detailed and practical responses to various queries related to medical images such as CT scans and MRIs. Additionally, by applying AI methods our system can provide customized support, in the field of biomedicine that caters to the requirements and choices of users offering personalized treatment suggestions and educational materials tailored to their health conditions.

In the following parts of this paper, we explore the structure and execution specifics of our LLaVA powered VQA model along, with its AI driven biomedical assistance functions. Moreover, we offer an assessment of our models efficiency on selected datasets like SLAKE, Path VQA and VQA Med showcasing its ability to improve the functionality and relia- bility of VQA, for medical imaging purposes.

II. RELATED WORK

The development of advanced language models and their application to the biomedical domain has been an active area ofresearch, laying the foundation for the work presented in this paper.

Med Palm 2 is a large language model developed by Google for the medical domain, which has shown exceptional learning abilities and impressive performance on medical exams, including the USMLE examination, scoring 82%.It is designed to assist and potentially automate medical decision-making, and its performance suggests that it could revolutionize health- care by providing accurate and relevant information to medical professionals.

PMC-LLaMA is a large language model fine-tuned on med- ical papers, specifically on the PubMedQA and MedMCQA training set, which has demonstrated superior performance on various medical QA datasets compared to the original LLaMA model. The model has been fine-tuned on 4.8 million biomedical academic papers and has shown better performance in medical knowledge and context understanding.

MedLLaMA is a pretrained medical language model that hasbeen shown to have a better ability to follow user instructionsthan other models, such as MedLLaMA-13B. It has been pretrained on medical corpus and can be easily loaded using theHugging Face Transformers library.

LLaVA-Med is a large language and vision model trained using a curriculum learning method for adapting LLaVA to the biomedicine space. It is an open-source release intended for research use only and can be helpful for exploring various biomedical vision-language processing tasks. The model is built upon LLaVA and Vicuna and can be used for visual question answering in the biomedicine domain.

III. METHODOLOGY

A VQA model used in imaging, like LLaVA Med typically comprises the following elements;

- Image Encoder; Processes visual input, such as images. Transforms it into a feature vector representation. This is often done using a network (CNN) or a pre trained modellike ResNet.
- Question Encoder; Handles textual input, like natural language questions and converts it into a feature vector representation. Typically achieved with a network (RNN) such as LSTM or a transformer-based model.
- Multimodal Fusion; Integrates the image and question embeddings into a representation for tasks like classification or sequence to sequence models. Techniques used include concatenation, element wise multiplication and attention mechanisms.
- Classifier or Sequence to Sequence Model; Generates answers to questions by processing the multimodal fusion output to provide the answer. For multiple choice questions the classifier outputs the probability distribution, over answer choices.

LLaVA Med is built on the PMC 15M dataset, a collection of image text data designed for vision language tasks. The model undergoes training using the log likelihood loss function. Is fine-tuned on studies extracted from the MIMIC CXR dataset. It attains a BLEU 4 score of 0.264 and a precision scoreof 0.311.

A. *Vision Encoder*

The vision encoder for Visual Question Answering (VQA) in medical imaging typically utilizes a convolutional neural network (CNN) or a pre-trained model like ResNet-18 to process visual input, such as medical images, and convert it into a feature vector representation. This architecture comprises several convolutional layers followed by pooling layers toreduce spatial dimensions, where filters extract features and pooling layers combine them. For ResNet-18, which includes 18 convolutional layers, each layer possesses varying numbersof filters and kernel sizes, progressively reducing spatial dimensions to a 1x1 feature map. The vision encoder undergoes training on medical imaging datasets such as VQA- RAD or PathVQA, optimizing parameters to minimize loss functions like cross-entropy or mean squared error.

Once trained, the vision encoder collaborates with a language encoder and multimodal fusion component to answer questions regarding medical images. After processing the inputimage, the vision encoder generates a feature vector, subsequently combined with a language encoder's feature vector via the multimodal fusion component. This combined feature vector then enters a classifier or sequence-to-sequence model to produce the final answer.

In essence, the vision encoder's role in VQA for medical imaging involves processing visual input using CNNs or pre-trained models to create a feature vector representation. Through training on medical imaging datasets, it learns relevant features for the VQA task and collaborates with other components to answer questions about medical images.

B. Question Encoder

In a Visual Question Answering (VQA) system the question encoder processes text input, like a natural language query. Converts it into a feature vector representation. Typically, it uses a network (RNN) such as long short-term memory (LSTM) or a transformer-based model to extract relevant features for theVQA task.

In imaging the question encoder examines questions to extract details. For example, it can determine if the question relates to a part of an image like an organ or anomaly. It can also identify whether the question is about a type of scan like Xray or CT.

After processing the text input, the question encoder creates a feature vector. This vector is combined with the image encoders feature vector using fusion. The resulting combined vector is then inputted into a classifier or sequence to sequencemodel to generate the answer.

In transformer-based models the question encoder may use self-attention mechanisms to analyze questions. For instance, it could employ attention to capture features, from image regionsfor answering that particular question.

The question encoder, for VQA, in imaging essentially uses RNNs or transformer-based models to transform text input intoa feature vector representation. By being trained on imaging datasets it understands features and works together with the vision encoder and multimodal fusion component to respond to queries regarding medical images.

Language Model

The Visual Question Answering (VQA) language model, in imaging powered by the LLaVA Med framework utilizes opensource models like LLaMA 2 and GPT 3.5 to enhance its versatility and effectiveness. These models play a role in improving the VQA system by leveraging their capabilities in understanding and generating natural language.

LLaMA 2, a variation of the Large Language Model and Vision Architecture (LLaVA) enhances the VQA models ability to analyze and interpret inputs related to images. With its features and expertise in the biomedical domain LLaMA 2 enriches the question answering process by offering precise andcontextually relevant answers.

Moreover integrating GPT 3.5, a AI model further boosts the conversational skills of the VQA system. GPT 3.5 is adept at producing responses tailored to queries facilitating seamless interactions between users and the VQA system.

By incorporating these open source models the LLaVA Med framework taps into intelligence and advancements in natural language processing and generative AI. This collaborative approach enhances the systems capability to understand and address inquiries, about images accurately with adaptability. In the end the collaboration, among LLaMA 2, GPT 3.5 and various open source elements in the LLaVA Med platform provides professionals with an flexible tool to analyze com- municate about and utilize biomedical images, in both clinical settings and research endeavors.

Multimodal Fusion

In the context of LLaVA-Med, multimodal fusion is the process of amalgamating information from both the imageencoder and the question encoder to form a unified representation. This combined representation is pivotal for generating answers to medical questions pertinent to the provided image. LLaVA-Med employs a method known as" tuning multi-modal MLP adapter" for multimodal fusion. This method entails integrating a small adapter network into the pre-trained vision tower, like OpenAI CLIP-ViT-L/14, and fine-tuning it to learn the correlation between image and question embeddings. The adapter network, comprising a multi-layer perceptron (MLP) with a smaller hidden layer size than the input embeddings, synthesizes concatenated image and question embeddings to yield a unified representation capturing relevant information from both modalities.

Furthermore, LLaVA-Med adopts a technique termed" feature selection" to discern the most pertinent features

International Journal On Engineering Technology and Sciences – IJETS
ISSN (P):2349-3968, ISSN (O): 2349-3976 Volume XI - Issue III, March – 2024

from image and question embeddings. This process involves selecting features with the highest cosine similarity to the query vector, ensuring that the fused representation encapsulates the salient aspects of both the image and the ac- companying question. By leveraging these multimodal fusion methods, LLaVA-Med enhances the synergy between visual and textual inputs, facilitating accurate and contextually rich responses to medical inquiries related to provided images.

C. *Answer Decoder*

The answer decoder, in the LLaVA Med model plays a role in crafting responses to inquiries concerning images. This key element takes input from the multimodal fusion module, which blends textual data from the image and question encoders to generate a natural language response. Utilizing a transformer architecture the answer decoder excels at handling data, text. Featuring layers with self-attention mechanisms and feed forward neural networks this architecture allows for processing of various parts of the input sequence aiding in managing lengthy sequences and identifying relationships within the data. Additionally, the feed forward neural networks enable the model to capture linear connections between input and output sequences.

During training the answer decoder refines its parameters using a entropy loss function that gauges the difference between predicted answers and ground truth responses. By leveraging a dataset of images along with associated questions the model grasps complex relationships between visual and textual information and their corresponding answers. Once trained it can adeptly generate responses to queries, about images. Users provide an image and a question prompting the model to generate a natural language response based on

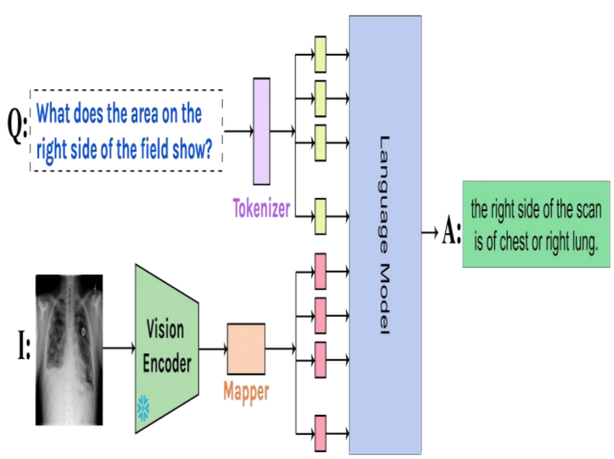

Fig. 1. Model architecture of our proposed open-ended generative VQA method textual details.

Moreover, the model can be fine- tuned using datasets related to tasks or areas to improve its performance. When incorporating RAG (Retrieval Augmented Generation) features into a model designed to address queries and offer advice through Langchain the model would utilize RAG capabilities to boost its effectiveness, in generating responses. RAG, a model that merges retrieval and generation methods aims to elevate the quality and relevance of generated answers

by integrating information from retrieved documents.

By integrating RAG features into the model, it can efficiently retrieve data from biomedical literature or medical databases to provide precise and contextually suitable responses to medical inquiries. The Langchain element would aid in formulating responses based on the retrieved information ensuring that the answers are coherent, informative and tailored to the query.

This integration would empower the model to tap into the wealth of knowledge found in texts and databases to deliver personalized advice and responses across various medical topics. Leveraging RAG capabilities would enhance the models capacity to furnish current information by retrieving and consolidating content from diverse sources ultimately enhancing the systems overall performance, in addressing biomedical queries and offering medical guidance.

IV. RESULTS AND DISCUSSIONS

This section provides a comprehensive analysis and interpretation of the performance of the proposed LLaVA- Med model on three biomedical VQA datasets: VQA-RAD, SLAKE, and PathVQA. This section elucidates the model's efficacy in addressing both closed-set and open-set questions across these datasets, shedding light on its ability to interpret and generate responses to a diverse range of medical inquiries.

- VQA-RAD: This dataset comprises 3515 QA pairs generated by clinicians, distributed over 315 radiology im- ages covering head, chest, and abdomen regions. Questions encompass 11 categories, including abnormalities, attributes, modalities, and organ systems. Half of the answers are closed-ended, while the remainder are open- ended with one-word or short phrase responses.
- SLAKE: A Semantically-Labeled Knowledge-Enhanced dataset containing 642 radiology images and over 7000 QA pairs. Questions may involve external medical knowledge and are annotated by experienced physicians. Im- ages feature rich visual annotations such as semantic segmentation masks and object detection

bounding boxes. The dataset encompasses various human body parts and is bilingual with English and Chinese subsets, though only the English subset is considered in comparisons.

- PathVQA: This dataset consists of 4998 pathology im- ages with 32,799 QA pairs. Questions cover aspects like location, shape, color, and appearance, categorized into open-ended and closed-ended varieties.

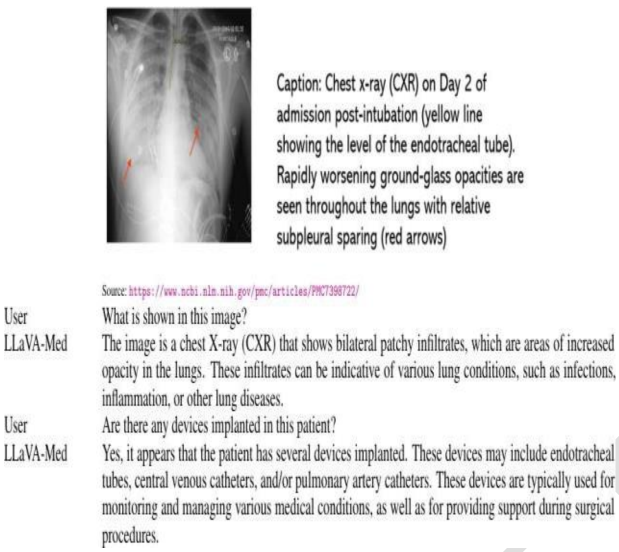

Fig. 2. Example of the our VQA model

a) Evaluation Metrics: Accuracy is reported for closed-set questions, while recall is used for open-set questions, measuring the presence of ground-truth tokens in generated sequences.

b) Comparisons with State-of-the-Art (SoTA): LLaVA- Med variants surpass LLaVA, with slight performance differences based on language model and vision encoder initialization. LLaVA-Med outperforms supervised SoTA for closed- setquestions on VQA-RAD and Path VQA, highlighting its efficacy in following instructions for biomedical tasks. How- ever, performance varies for open-set questions across datasets,with LLaVA-Med achieving SoTA on SLAKE but facing limitations elsewhere, possibly due to the ambiguity of open- set biomedical questions.

In summary, LLaVA-Med demonstrates strong performancein closed-set question answering, particularly when provided clear instructions. However, its effectiveness in open-set ques-tion scenarios varies across datasets, suggesting challenges in handling ambiguous biomedical queries without constrained answer options.

V. CONCLUSION

In conclusion, LLaVA-Med is a large multimodal model (LMM) designed for answering medical questions related to medical images. It uses a multimodal fusion approach to com- bine information from both the image and the question, and it is a key research frontier for machine learning in the medical domain. By incorporating RAG (Retrieval-Augmented Gener- ation) techniques, LLaVA-Med can enhance its performance in generating answers to medical questions by retrieving relevant information from a large corpus of biomedical literature or medical databases.

The RAG capabilities enable the model to provide accurate and contextually appropriate answers to questions related to medicine and healthcare, and they can improve the overall performance and utility of the system in answering biomedical questions and offering medical recommendations. The integra- tion of RAG techniques with LLaVA-Med represents a promising approach to addressing the challenges of VQA in medical imaging, as it allows the model to leverage the vast amount of knowledge available in biomedical texts and databases to offer personalized medicine suggestions and responses to a wide range of medical queries.

REFERENCES

[1] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, Jianfeng Gao LLaVA- Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day https://doi.org/10.48550/arXiv.2306.00890

[2] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Ku¨ttler, Mike Lewis, Wen-tau Yih, Tim Rockta¨schel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. NeurIPS, 2020.

[3] Yakoub Bazi, Mohamad Mahmoud Al Rahhal, Laila Bashmal, and Mansour Zuair. Vision– language model for visual question answering in medical imagery. Bioengineering, 2023.

[4] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In ECCV. Springer, 2022

[5] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286, 2020.

[6] Singhal, K., Azizi, S., Tu, T. et al. Large language mod- els encode clinical knowledge. Nature 620, 172–180 (2023). https://doi.org/10.1038/s41586-023-06291-2

[7] Tom van Sonsbeek, Mohammad Mahdi Derakhshani, Ivona Naj- denkoska, Cees GM Snoek, and Marcel Worring. Open-ended medical visual question answering through prefix tuning of language models. arXiv preprint arXiv:2303.05977, 2023.

[8] Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. arXiv preprint arXiv:2303.14070, 2023.

[9] OpenAI. ChatGPT. https://openai.com/blog/chatgpt/, 2022.

[10] Haotian Liu, Kilho Son, Jianwei Yang, Ce Liu, Jianfeng Gao, Yong Jae Lee, and Chunyuan Li. Learning customized visual models with retrieval-augmented knowledge. arXiv preprint arXiv:2301.07094, 2023.

[11] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semanticallylabeled knowledge-enhanced dataset for medical visual question answering. In International Symposium on Biomedical Imaging (ISBI). IEEE, 2021.

[12] Pengfei Li, Gang Liu, Lin Tan, Jinying Liao, and Shenjun Zhong. Self-supervised visionlanguage pretraining for medical visual question answering. arXiv preprint arXiv:2211.13594, 2022.

[13] Iryna Hartsock, Ghulam Rasool Vision-Language Models for Med- ical Report Generation and Visual Question Answering: A Review https://doi.org/10.48550/arXiv.2403.02469

[14] Miao, J.; Thongprayoon, C.; Suppadungsuk, S.; Garcia Valencia, O.A.; Cheungpasitporn, W. Integrating Retrieval-Augmented Generation with Large Language Models in Nephrology: Advancing Practical Applications. Medicina 2024, 60, 445. https://doi.org/10.3390/medicina60030445

[15] Mercy Ranjit, Gopinath Ganapathy, Ranjit Manuel, Tanuja Ganu Re- trieval Augmented Chest X-Ray Report Generation using OpenAI GPT models https://doi.org/10.48550/arXiv.2305.03660

[16] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286, 2020.

[17] Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grund- mann, Tom Oberhauser, Alexander Lo¨ser, Daniel Truhn, and Keno K Bressem. Medalpaca–an open-source collection of medical conversa- tional ai models and training data. arXiv preprint arXiv:2304.08247, 2023.

[18] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. arXiv preprint arXiv:2303.00915, 2023.

[19] Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Qian Wang, and Dinggang Shen. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. arXiv preprint arXiv:2304.01097, 2023.

[20] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-llama: Further finetuning llama on medical papers. arXiv preprint arXiv:2304.14454, 2023.

[21] Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. Huatuo: Tuning llama model with chinese medical knowledge, 2023.

[22] A Venigalla, J Frankle, and M Carbin. BiomedLM: a domain-specific large language model for biomedical text. MosaicML. Accessed: Dec, 23, 2022.

[23] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie- Anne Lachaux, Timothe´e Lacroix, Baptiste Rozie`re, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023

[24] Chang Shu, Baian Chen, Fangyu Liu, Zihao Fu, Ehsan Shareghi, and Nigel Collier. Visual med-alpaca: A parameter-efficient biomedical llm with visual capabilities. 2023.