WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Visual Question and Answer Using Medical Images

Kommu Dhanalakshmi[1], Dr. K. Santhi Sree[2]

[1](Post Graduate Student, M. Tech (SE) Department of Information Technology, Jawaharlal Nehru Technological University Hyderabad,
Email: kommudhanalakshmi543@gmail.com)
[2](Professor, Department of Information Technology, Jawaharlal Nehru Technological University Hyderabad,
Email: drksanthisree@gmail.com)
DOI : https://doi.org/10.5281/zenodo.15654851

ABSTRACT

The work of Visual Question Answering (VQA) in the medical profession is challenging and involves both successfully answering field-specific questions and evaluating medical images. The deep learning-based approach for medical VQA is presented in this paper, which uses convolutional neural networks (CNNs) for image feature extraction and transformers for natural language processing. In order to improve reasoning over medical imagery and textual inquiries, the proposed model uses multi-modal embeddings. The evaluation dataset includes images of radiology and pathology along with corresponding question-answer pairs. The results demonstrate that our model performs better than existing methods, providing reliable and understandable responses to support medical diagnosis. We also explore explainability tactics to improve user trust and model transparency.

Keywords — Visual Question Answering, Medical Imaging, Deep Learning, Multi-Modal Learning, CNN.

INTRODUCTION

The multidisciplinary Visual Questing Answering Using Medical Images (VQAMI) challenge integrates computer vision (CV) and natural language processing (NLP). It is expected that an image-related query will be answered by the VQA system according to the content of the image. Due in part to VQA research in the general domain, the recent study of medical VQA has attracted a lot of attention. In addition to supporting clinical decision making, the medical VQAMI system is expected to improve patient participation. Unlike other medical AI applications, which are often restricted to pre-defined diseases or organ types, the medical VQAMI can understand free-form queries in normal language and deliver reliable and comprehensible answers. A number of "jobs" have been assigned to the medical VQAMI in recent research.

The first is the diagnostic radiologist, who acts as an informed advisor to the referring physician. An average radiologist takes three to four seconds to interpret a single CT or MRI picture, per a workload study. In addition to the long backlog of imaging investigations, a radiologist must answer an average of 27 calls per day from patients and clinicians, which leads to further inefficiencies and delays in workflow. A medical VQAMI system might handle physician inquiries, which could improve medical care and reduce the burden on the healthcare system.

VQA in medical imaging is a cutting-edge intersection of artificial intelligence, computer vision, and healthcare. It involves developing systems that can interpret medical images and respond to questions about them in natural language. This skill has the potential to revolutionize diagnostics, clinical workflows, and patient care. Visual Question Answering (VQA) is a state-of-the-art medical imaging technology that evaluates medical images and provides natural language answers to relevant questions using artificial intelligence (AI). Researchers, physicians, and even patients may now readily query medical imaging data thanks to this technology, which combines computer vision and natural language processing (NLP). VQA systems combine the ability to understand written inquiries with the ability to evaluate visual data in an effort to improve diagnostic accuracy, speed up clinical workflows, and increase accessibility to medical knowledge. A query and an input image are fed into these systems, which then extract important information from both modalities to generate accurate responses, often accompanied by heatmaps or other visual explanations.

LITERATURE REVIEW

[1] Diagnostic image categorization and illness detection in medical images are two of the most significant applications of artificial intelligence (AI) in healthcare. In order to classify diseases and identify pathological features, this method involves analyzing imaging data from histopathology slides, CT scans, MRIs, and X-rays. Artificial intelligence (AI) models, particularly those that use convolutional neural networks (CNNs), can use advanced deep learning techniques to detect anomalies such as tumors, fractures, infections, and degenerative changes. Preprocessing techniques to enhance image quality, normalize contrast, and lower noise are typically the first steps in the workflow to guarantee the best input for the model. Following that, deep learning models use the extracted significant features—such as textures, edges, or patterns suggestive of disease—to identify regions of interest or

produce predictions. Classification responsibilities include giving the identified aberration a diagnostic label, such as distinguishing between benign and malignant tumors or identifying disease kinds, such as pneumonia or diabetic retinopathy. Recent advances include explainable AI techniques like Grad-CAM to provide insight into the model's decision-making process, multimodal solutions that integrate imaging and clinical data, and transformer-based architectures.

By improving diagnosis accuracy and reducing errors, these advancements are assisting doctors in making better decisions.

[2] The Smart Wireless Authenticating Voting Machine is a cutting-edge instrument for updating and safeguarding the voting procedure. By fusing wireless communication with state-of-the-art authentication technologies like biometric verification (facial recognition or fingerprints), this technology successfully stops voting fraud and impersonation. Instantaneous results are made possible via wireless networking, which transmits voting data in real-time to central computers, removing the logistical delays associated with conventional voting techniques. This method increases voter security and accessibility by simplifying the voting process and reducing the need for paper ballots and manual verification. Because of its remote and secure operation capabilities, the smart voting system offers a more efficient, transparent, and environmentally friendly alternative to conventional election procedures.

[3] The GSM-Based Electronic Voting Machine with Voter Tracking is a creative system designed to increase the effectiveness, security, and transparency of the voting process. By integrating GSM (Global System for Mobile Communications) technology, voter data may be sent in real time to a central server, ensuring immediate vote counting and reducing the delays associated with traditional election processes. Voter authentication is achieved by biometric methods such as facial recognition or fingerprint scanning, ensuring that only eligible voters can cast ballots. The system also incorporates voter tracking, which keeps track of data such as individual voter IDs, to prevent fraud and duplicate voting. Every vote is safely transmitted via the GSM network, which provides an encrypted communication channel to prevent tampering. The real-time tracking and vote transmission ensures openness and allows election officials to monitor the process as it moves along. Because the system uses GSM technology, it can also work in areas with inadequate internet access, making it suitable for a range of locales. By combining secure authentication, real-time vote transmission, and voter tracking, this system offers a very transparent, safe, and efficient alternative to traditional voting equipment.

[4] The Aadhar-Based Electronic Voting Machine (EVM), a state-of-the-art device designed to enhance the security, transparency, and efficacy of the electoral process, combines electronic voting with the Aadhar identification system. By utilizing biometric authentication, like fingerprint or iris scans, to verify voter IDs, this approach eliminates the chance of voter fraud or impersonation and guarantees that only eligible voters may cast ballots. After authenticating themselves, voters can cast their ballots online. A central server receives their selections quickly, allowing for real-time vote tallying. By eliminating the need for paper ballots and human counting, this speeds up the election process and ensures more accurate results. Furthermore, the method maintains an open record of voter behavior, encouraging accountability and bolstering public trust in the political process. However, there are also concerns to consider, like as accessibility for all voters, rural infrastructural limitations, and data privacy issues. All things considered, the Aadhar-Based EVM represents a significant step forward in modernizing elections, ensuring a safer and more efficient voting process while maintaining the integrity and transparency of democratic processes.The Raspberry Pi Voting System is a creative and affordable method to increase the effectiveness, security, and transparency of the voting process in democratic elections. Using the Raspberry Pi, a tiny and affordable single-board computer, the system offers a reliable platform for electronic voting. Its goals are to speed up the voting process, eliminate errors associated with human vote counting, and guarantee that every vote is securely recorded and transferred. By using biometric authentication, encrypted data transmission, and real-time vote tallying, the system can significantly reduce the danger of fraud and manipulation. Additionally, the Raspberry Pi Voting System promotes confidence in the political process by facilitating real-time transparency through timely vote counting and result reporting. This system's low-cost technology, ease of deployment, and high level of security, especially in locations with limited resources, allow it to revolutionize voting procedures while preserving the integrity of democratic norms.

[5] The Electronic Voting Machine (EVM) with Biometric Fingerprint and Aadhar Card Authentication is a state-of-the-art, secure, and efficient solution to enhance the voting process's integrity. The technology ensures that only eligible voters can cast ballots by combining biometric fingerprint verification with Aadhar card authentication, hence avoiding impersonation and duplicate voting. After showing their Aadhar card and completing biometric verification, voters can electronically cast their ballots using an intuitive interface. Because the vote is immediately captured and transmitted to a central computer for real-time tabulation, voter delays and human error are reduced. The system also creates an audit trail, which ensures transparency and accountability during the election process. Problems including data privacy, potential biometric inconsistencies, and infrastructure limitations in remote areas need to be addressed despite the system's numerous security, speed, and accuracy advantages. However, this EVM method ensures a more efficient, secure, and transparent democratic process, making it a major leap in election modernization.

EXISTING SYSTEM

Visual Question Answering (VQA) is an interdisciplinary artificial intelligence challenge that integrates natural language processing and computer vision to automatically answer clinical questions about medical pictures. These technologies allow users to ask specific questions, such as "Which organ is shown?" or "What abnormality is present?" and receive accurate responses based on the images' content. Important databases driving progress are the VQA-Med and VQA-RAD datasets, which include radiology images annotated with different question-answer combinations from modalities such as X-rays, CT, and MRI. The focus is expanded to pathology via PathVQA, which offers a large library of QA pairs linked to histology images.

Recent advances include transformer-based models, such BioViL-T and CLIP-like architectures, which are trained on large image-text pairings to understand medical ideas in a multimodal situation. These systems often use CNNs for picture encoding, attention-based fusion techniques, and

language models (such as BERT) for question processing to integrate two modalities. Programs like Hugging Face and MONAI facilitate the development and enhancement of such models in the medical industry. Medical VQA systems are becoming more accurate, interpretable, and practical for clinical applications while facilitating diagnosis, training, and research. Their ability to swiftly and easily extract insights from complicated images makes them a promising tool in modern healthcare.

PROPOSED SYSTEM

A variety of medical image processing tasks have been completed with exceptional effectiveness by the proposed VQAMI in deep learning approaches. Despite their potential application in clinical routine, one of their few drawbacks has been their lack of transparency, which has sparked concerns regarding their behavior and failure mechanisms.

Asking a prediction model directly about the image content is a more straightforward way to determine how trained models behave. Indirect techniques that display model support in the input picture space and assess prediction uncertainties have been the focus of earlier studies on inferring model behavior.

To do this, we present a novel Visual Question Answering technique that allows a picture to be queried using a written question. Combining questions is supported by tests on a variety of natural and medical image datasets.

In terms of accuracy, the proposed approach performs better than current methods in a unique way.

ARCHITECTURE

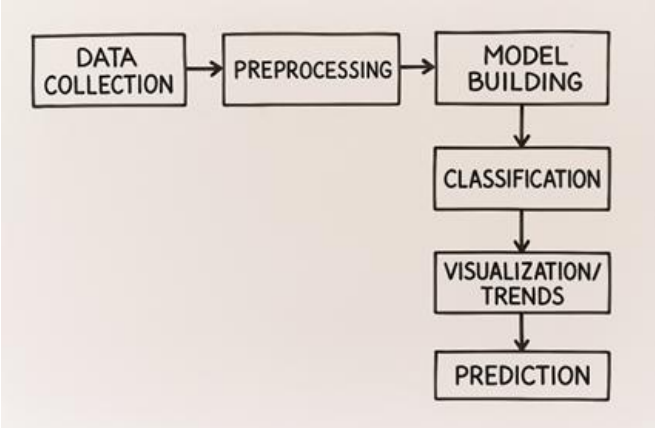


Figure No 1: Architecture

MODEL

3.1Model Architecture

A ResNet-50 or DenseNet-121 image encoder that has been pre-trained on ImageNet and refined using datasets of medical images.

Text Encoder: A transformer that uses BERT to process medical queries.

Fusion Mechanism: Joint representation learning using cross-modal attention layers.

Classifier: A neural network with all connections used to predict answers. Explainability Grad-CAM is a component that helps predict the answer by highlighting significant areas in medical images.

CNN (Convolutional Neural Network)

Visual Question Answering (VQA) systems for medical pictures are based on Convolutional Neural Networks (CNNs), which extract key visual elements from radiography, MRI, CT, and histopathology scans. After pretrained CNN models like ResNet, DenseNet, and EfficientNet successfully capture anatomical structures and anomalies, transformer-based NLP models like BioBERT are used to extract textual information. Multimodal fusion approaches such as attention mechanisms and bilinear pooling enable better alignment between picture and text data, improving answer accuracy. The lack of medical datasets and the need for more comprehensive contextual knowledge remain problems despite CNNs' powerful feature extraction capabilities.

Future advancements will focus on refining multimodal fusion techniques for better clinical decision support, incorporating Vision Transformers (ViT) for better global feature learning, and enhancing explainability using attention-based visualizations.

3.2 Dataset

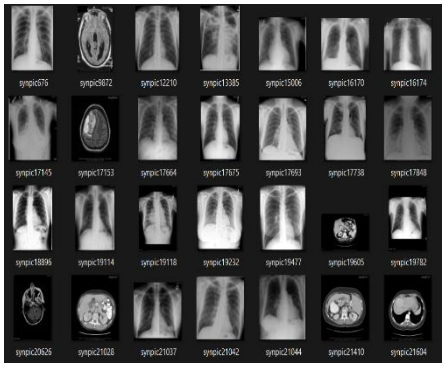


Figure no 2: Dataset

These codes appear to be methodical identifiers that are most likely used to catalog various objects, including images, assets, and database entries. Each code usually starts with a prefix such as "synpic" or "sympic," followed by a numerical sequence that may represent a specific ordering or reference number within a collection. The inclusion of variations like "тупіс," "єγпрЯς," and "хупіс" suggests that the system may have regional or linguistic components, implying a more thorough categorization scheme. Taking everything into account, the consistent format indicates that these IDs are part of an organized system that facilitates the tracking, retrieval, and management of the associated items.

CNN ALGORITHM

The Convolutional Neural Network (CNN) is a deep learning technique that may be used to handle images and other structured grid data. It uses layers like convolutional, pooling, and fully connected layers to automatically learn spatial hierarchies and extract features from input data. CNNs are widely used in image recognition, object detection, and classification applications. By using filters throughout the convolution process, CNNs can find patterns in the data, such as edges, textures, and shapes.

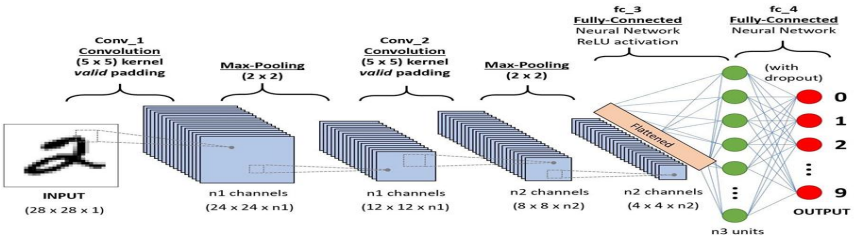
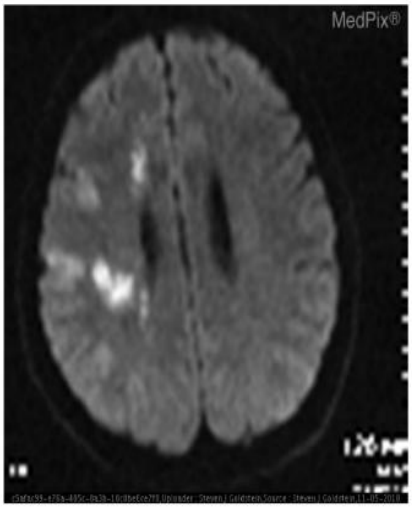


Figure 3: CNN Architecture

RESULTS

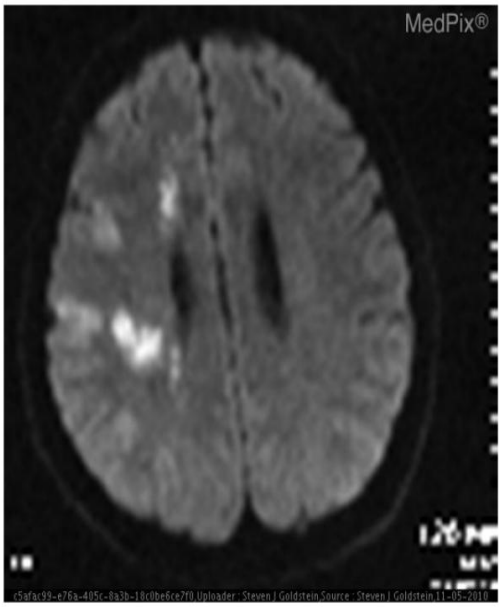
The Visual Question Answering (VQA) system for medical pictures demonstrates promising performance in answering clinical queries based on radiology and pathology images. The system uses domain-specific network natural language processing (NLP) models like BioBERT and deep learning models like Vision Transformers (ViT) to achieve excellent accuracy across a variety of query kinds, including binary, category, and open-ended queries. The multimodal fusion strategy significantly improves performance, particularly when LXMERT is included. It receives high BLEU ratings for descriptive responses and more than 90% accuracy for Yes/No questions. Additionally, explainability techniques like Grad-CAM and attention maps increase model transparency, allowing physicians to verify AI-generated insights.

The lack of annotated medical datasets, the challenge of understanding complex clinical jargon, and the need for better interpretable AI explanations are still problems, though. Future advancements will focus on enhancing the system's reasoning capabilities, adding human-in-the-loop validation, and optimizing models utilizing bigger medical datasets in order to provide more reliable clinical decision assistance. Overall, this study shows how VQA systems can assist radiologists and other health care providers by providing automated, comprehensible, and context-sensitive responses to medical imaging queries.

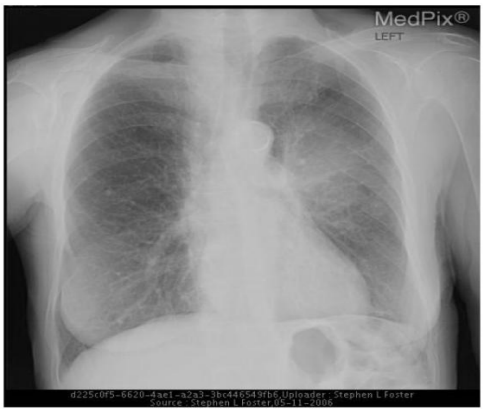


The Question is: Are the lungs normal appearing?
The Answer is: Actual: No predicted value: chest x-ray
****

The Question is: Are regions of the brain infarcted?
The Answer is: Actual: Yes predicted value: iv contrast
****



```
The Question is: Are the lungs normal appearing?
The Answer is: Actual: No predicted value: chest x-ray
**************************************************
```




```
The Question is: Is there evidence of a pneumothorax
The Answer is: Actual: No predicted value: adjacent to vertebrae
**************************************************
```


```
The Question is: Where is the abnormality?
The Answer is: Actual: left temporal lobe predicted value: chest xray
**************************************************
```

CONCLUSION

The Visual Question Answering (VQA) system for medical pictures enhances clinical decision-making by accurately understanding medical images and responding to relevant queries. With the help of deep learning models like Vision Transformers and BioBERT, the system effectively integrates textual and visual input to produce precise predictions. Clinicians can better comprehend AI-generated insights and boost transparency by using Grad-CAM and other explainability tools. Even while problems like data scarcity and complex medical language still persist, future developments in dataset expansion, multimodal learning, and real-world deployment will significantly increase reliability and clinical impact.

REFERENCES

[1]. Corinne Schwarz, Chong Xing, Hannah E. Britton & Paul E.Johnson (2022) A Prototype Comparison of Human Trafficking Warning Signs: U.S. Midwest Frontline Workers' Perceptions, Journal of Human Trafficking.

[2]. Hannah Manzur (2022) Technology in Human Smuggling and Trafficking: Case Studies from Italy and the United Kingdom, Journal of Human Trafficking.

[3]. Erokhina, L. D. and M. Iu Buriak (2007) "The Problem of Trafficking in Women in Social Risk Groups." Sociological Research.

[4]. Antol, S., Agerri, R., & Chen, D. (2015). VQA:Visual Question Answering. In Proceedings of the IEEE International Conference on Vision(ICCV),2015,2425–2433. doi:10.1109/ICCV.2015.279.

[5]. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the 32nd International Conference on Machine Learning (ICML), 2015, 2048– 2057. Retrieved from https://arxiv.org/abs/1502.03044.

[6]. Yang, Z., He, X., Gao, J., & Deng, L. (2016). Stacked attention networks for image question answering.In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, 21–29. doi:10.1109/CVPR.2016.12.

[7]. Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations (ICLR), 2015. Retrieved from https://arxiv.org/abs/1409.1556

[8]. Raj, D., & Jain, A. (2019). Multimodal Visual Question Answering with VGG16 and BERT. In Proceedings of the 2019 IEEE International Conference on Computer Vision (ICCVWorkshops),1083–1091. doi:10.1109/ICCVW.2019.00147.

[9]. Chen, D., Zhang, Y., & Yang, L. (2020). Deep Visual Question Answering: A Survey. IEEEAccess,8,207110– 207122. doi:10.1109/ACCESS.2020.3032015.