iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** V **Month of publication:** May 2024

DOI: https://doi.org/10.22214/ijraset.2024.62677

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

Visual Question Answering based Educational Tool for Medical Students using Cross-ViT

Ms. Soudamini Somvanshi[1], Dhanashree Patil[2], Pranjal Wagh[3], Atharva Sambhus[4], Prashant Singh[5], Utkarsh More[6]
Department of Computer Engineering, D. Y. Patil College Of Engineering, Akurdi, Pune-44, Maharashtra, India

Abstract: ***This paper introduces an advanced approach to medical visual question answering (VQA) using the Cross-ViT architecture. The model employs a dual-branch method to extract multi-scale feature representations from images, utilizing cross-attention mechanisms to enhance visual features. By integrating Stacked Attention Networks (SAN) and leveraging semantic extraction from LSTM for textual data, the model shows significant performance improvements. Experiments on various biomedical VQA tasks demonstrate notable improvements in retrieval accuracy and image-text correlation. The study highlights the potential of medical VQA systems to transform healthcare delivery, improve diagnostic accuracy, and facilitate patient engagement and education, with promising future applications in telemedicine, surgery assistance, and integration with electronic health records.***

Keywords: VQA, ViT, medical, SAM, SAN, ImageCLEF, VQA-RAD

I. INTRODUCTION

Medical students often encounter challenges in integrating their theoretical knowledge with practical application when interpreting medical images. While traditional learning methods focus primarily on didactic instruction, the incorporation of visual question answering (VQA) systems holds promise in augmenting medical education by fostering active engagement and enhancing diagnostic proficiency.

Our proposed Medical VQA (Medi-Vision) model endeavours to surmount these challenges by leveraging a Vision Transformer (ViT) architecture augmented with multiscale input capabilities and cross-attention mechanisms. By integrating ViT, which excels in capturing intricate visual features, with cross-attention mechanisms, which facilitate the fusion of visual and textual information, our model aims to enhance its capacity to address diverse medical queries comprehensively.

To optimize the performance of the proposed Medi-Vision model, extensive efforts have been devoted to curating and preprocessing a substantial corpus of medical VQA data. Additionally, various training techniques, including adversarial training and fine-tuning, have been employed to refine the model's understanding of the intricate relationships between visual and textual data, as well as medical knowledge.

By empowering medical students with a robust VQA model capable of effectively interpreting medical images and answering related queries, our methodology seeks to bridge the gap between theoretical knowledge and practical application, thereby fostering a more comprehensive and effective medical education paradigm.

II. ALGORITHM

A Training samples D, perturbation bound, learning rate τ, ascent steps K, ascent step size α.

A. Cross-ViT Architecture

Algorithm:

Input: An image x and a natural language question q.

Process:

- Tokenize x and q into image patch tokens and word tokens, respectively.
- Pass these tokens through two branches of the Cross-ViT architecture.
- The first branch processes small patch tokens with lower computational complexity.
- The second branch processes large patch tokens with higher computational complexity.
- Outputs of these two branches are fused together using an efficient cross-attention module.
- Let D1 and D2 be the outputs of the first and second branches, respectively.

- D = f(D1, D2) represents the fused output.

The cross-attention module f is where the multi-layer perceptron (MLP) is applied to the fused output H to forecast the response.
Application: Produces stronger visual features for VQA tasks.

B. Pseudo Code

```
Function Cross-ViT(x, q):
    image_patch_tokens = TokenizeImage(x)
    word_tokens = TokenizeQuestion(q)
  D1 = ProcessSmallPatchTokens(image_patch_tokens)
    D2 = ProcessLargePatchTokens(image_patch_tokens)
  D = CrossAttentionModule(D1, D2)
    response = MLP(D)
      return response
```

III. EXPERIMENTAL/METHODOLOGY

Accurately identifying and delineating key elements in medical images, such as tumors or lesions, presents a significant hurdle in medical VQA. SAM, however, offers a promising solution with its ability to provide precise object segmentation, a vital input for VQA models.

Beyond refining VQA precision, SAM offers invaluable insights into critical visual features crucial for different medical image analyses. Through scrutinizing SAM's object proposals and their corresponding VQA outcomes, researchers can glean deeper understandings of the visual cues essential to diverse medical conditions.

By integrating SAM into medical VQA systems, there's the potential to significantly enhance the accuracy and efficiency of medical image interpretation and diagnosis. By facilitating precise object segmentation and enabling nuanced object-focused reasoning, SAM equips healthcare professionals with the tools to swiftly and accurately interpret medical images, leading to more informed decisions and ultimately improving patient outcomes.

Hence, SAM is integrated in the system in order to efficiently segment medical images and to increase efficiency of the model.

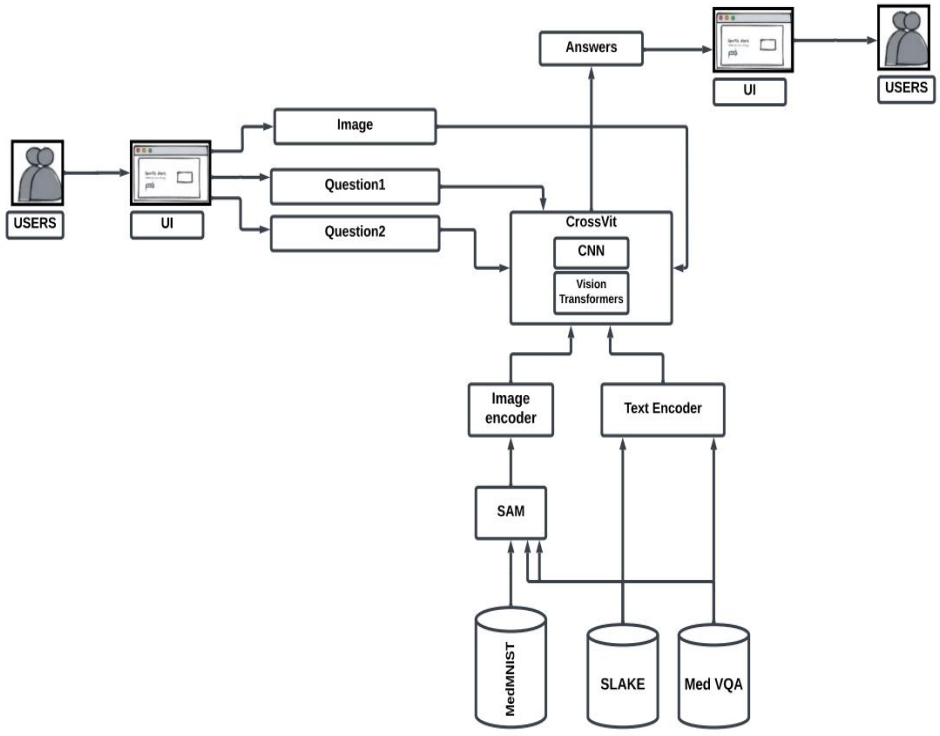


Fig. 1 System Architecture

A. Dataset

1) The VQA-RAD: VQA-RAD dataset includes 3,515 question-and-answer pairs and 315 images of medical objects.

2) ImageCLEF: ImageCLEF dataset evaluates technologies for visual data annotation, indexing, and retrieval across lifelogging, medicine, nature, and security domains

3) Most of the time, patient images are acquired from various imaging techniques such as CT scans and MRIs. These images can be either two-dimensional (2D) or three-dimensional (3D) and often come labeled or annotated with different types of metadata, including patient demographics, imaging parameters, and diagnostic or pathological information.

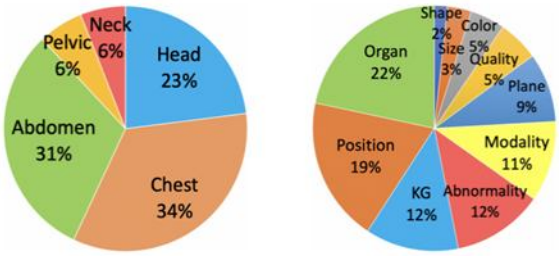


Fig.2 Image distribution based on body parts and the distribution of question types

TABLE I
VQA-RAD AND IMAGECLEF ARE CONTRASTED

Dataset	Images	QA Pairs	Question Type	Language	Knowledge Graph
VQA-RAD	315	3500	Vision	English	NO
ImageCLEF	3200	12792	Knowledge based & Vision	English	Yes

B. Pre-processing.

1) *Segmentation*: Segmentation stands as an important initial phase in Med-VQA, encompassing the isolation of targeted regions or entities within an image for subsequent examination. Within our study, we harnessed the Segment Anything Model (SAM), developed by Meta, for segmentation tasks using the Slake dataset. SAM emerges as a robust and groundbreaking AI model, uniquely adept at precisely delineating objects in images or videos with notable efficiency. What sets SAM apart are its remarkable attributes, including its adaptability to prompts, diverse data handling capabilities, capacity for generalization, and remarkable zero-shot performance.

2) *Multi Modal Collator*: The integration of data from various modalities into a cohesive input suitable for a neural network's visual question answering (VQA) process is facilitated by a multimodal collator. This component, typically implemented as a function or class in PyTorch, plays a vital role in VQA systems by enabling the utilization of both images and queries to generate answers. A fundamental reason for employing a multimodal collator is the disparate representations of images and questions. Images may exist as 2D or 3D arrays of pixel values, whereas questions are often represented as sequences of text tokens. Consequently, a multimodal collator serves to harmonize these diverse formats into a unified structure that can be effectively processed by the VQA model. Essentially, the multimodal collator acts as a bridge between the visual and textual components of the VQA system, ensuring seamless integration of image and question data for accurate answer generation. Its ability to translate different data types into a common format enhances the overall functionality and performance of the VQA model.

Pre-Processing steps

a) *Image Pre-processing*: Prior to utilization in a VQA model, images from the ImageCLEF dataset may require resizing, cropping, or normalization. Typical techniques involve adjusting images to a standardized size, focusing on pertinent areas via cropping, and standardizing pixel values.
b) Text Pre-processing: Textual data from the VQA-RAD dataset might necessitate pre-processing due to medical terminology and technical terms. This involves specialized tokenization methods, potentially leveraging medical ontologies or knowledge bases. Removal of stop words, punctuation, and stemming or lemmatization of words may also be required.
c) Knowledge Graph Integration: Integration of knowledge graph data from the VQA-RAD dataset could enhance VQA model performance. Pre-processing of this data ensures alignment with dataset questions and images, optimizing its structure and relevance.
d) Data Augmentation: Augmentation techniques can enhance the generalizability of VQA models trained on the VQA-RAD dataset. Common methods include random cropping, flipping, and introducing noise or perturbations to images, enriching the dataset and improving model robustness.
e) Encoding: Following pre-processing, encoding techniques such as bag-of-words, TF-IDF, or word embeddings can be applied to the VQA-RAD dataset. The choice of encoding depends on task specifics and dataset characteristics, facilitating effective data representation for VQA tasks.

C. Model

The Cross-ViT architecture employs a dual-branch strategy to derive multi-scale feature representations from images. It begins by tokenizing the input image x and natural language question q into image patch tokens and word tokens, respectively. These tokens are then processed separately by two branches within the Cross-ViT architecture.

The first branch handles small patch tokens, prioritizing computational efficiency, while the second branch tackles larger patch tokens, albeit with a higher computational cost. Subsequently, the outputs of these two branches, denoted as D1 and D2 respectively, are fused together employing an efficient cross-attention module.

This fusion mechanism enables the branches to complement each other, yielding more robust visual features tailored for Visual Question Answering (VQA) tasks. The fused output

$D=f(D1, D2)$ encapsulates the integrated representation.

The cross-attention module f leverages a multi-layer perceptron (MLP) applied to the fused output H to forecast the response, thereby facilitating seamless integration of visual and textual information for accurate predictions.

D. Experiment Setup

Experiments are conducted on the VQA-RAD dataset using Stacked Attention Networks (SAN), a widely utilized model in the Med-VQA framework. Visual features are extracted from electronic images and segmented using the Segment Anything Model (SAM). For processing questions, tokens are created to represent insertions in English. A 1024-dimensional Long Short-Term Memory (LSTM) network is employed to extract the semantics and classify the query.

The model operates through two pipelines. For non-visual processing, SAN's multi-modal fusion module generates unified features by combining visual and textual content for classification. For cognitive tasks, query-related links are extracted from knowledge maps and combined with various types of data to enhance classification accuracy.

IV. RESULTS

The image given below is uploaded to our model along with question. The following table shows questions(prompt) and answers generated by model based on given image.

Fig 3 Image given to model along with prompt

TABLE 2
PROMPT AND RESPECTIVE ANSWERS GENERATED BY

Sr. No	Question	Answer
1	What is in this image?	The image features a close-up view of a human head, focusing on the frontal part of the brain.
2	what type of scan this is?	This is a medical imaging scan, specifically an MRI (Magnetic Resonance Imaging) of the human head.
3	Is there any sign of tumour or damage?	The scan displays the brain's anatomy, including the cerebral cortex, cerebellum, and other structures. However, there are no visible signs of tumours or damage in the scan
4	List parts of brain visible in this image along with their position	In the image, the visible parts of the brain include the frontal lobe, which is the area in front of the forehead, and the occipital lobe, which is the area at the back of the head.

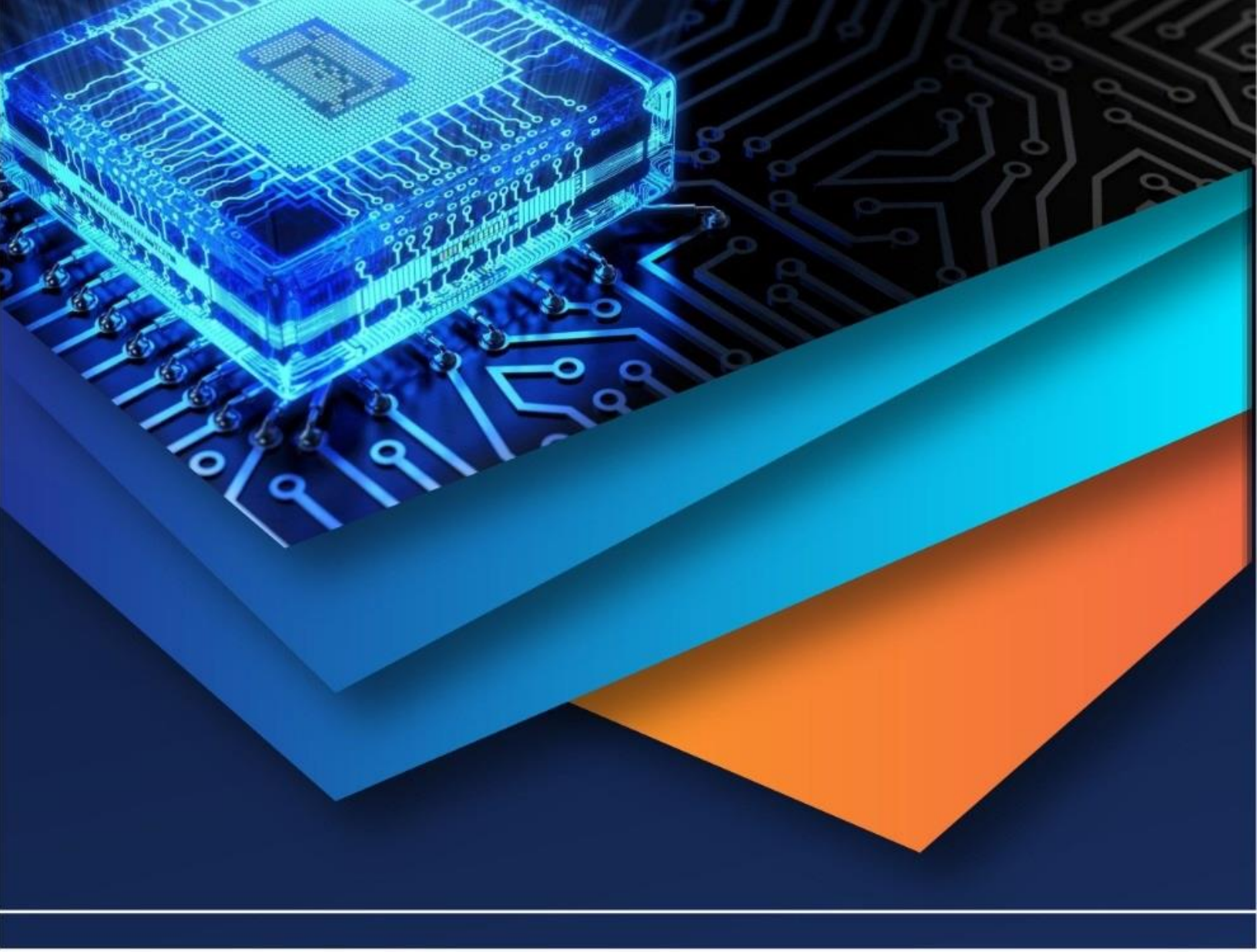
V. CONCLUSIONS

Thus, this study proposes the approach of using Cross-Vit for medical visual question and answering questions that call for lengthy chains of deduction and short-term memory, Incorporating SAM for segmentation.

The future scope of Medical VQA is broad and promising, with the potential to significantly impact the healthcare industry. It can aid in accurate diagnoses, empower patients through education and engagement, enhance telemedicine and remote care, assist in surgeries, support medical research and education, integrate with electronic health records, and expand across various medical specialties. As this technology evolves, it is poised to transform healthcare delivery, improve patient outcomes, and drive advances in medical knowledge and practice.

REFERENCES

[1] J. Ma, J. Liu, Q. Lin, B. Wu, Y. Wang and Y. You, "Multitask Learning, for Visual Question Answering," in IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 3, pp. 13801394, March 2023, doi: 10.1109/TNNLS.2021.3105284.
[2] S. Antol et al., “VQA: Visual question answering,” in Proc. IEEE Int. Conf. Computer. Vis. (ICCV), Dec. 2015, pp. 2425–2433.
[3] D. Gurari et al., “VizWiz grand challenge: Answering visual questions from blind people,” in Proc. IEEE/CVF Conf. Computer Vis. Pattern Recognition, Jun. 2018, pp. 3608–3617
[4] D. Zhang, R. Cao, and S. Wu, “Information fusion in visual question answering: A survey,” Inf. Fusion, vol. 52, pp. 268–280, Dec. 2019 .
[5] Y. Zhu, O. Groth, M. Bernstein, and L. FeiFei, “Visual7W: Grounded question answering in images,” in Proc. IEEE Conf. Computer Vis. Pattern Recognition (CVPR), June. 2016, pp. 4995–5004.
[6] Y. Goyal, T. Khot, D. SummersStay, D. Batra, and D. Parikh, “Making the v in VQA matter: Elevating the role of image understanding in visual question answering,” in Proc. IEEE Conf. Computer Vis. Pattern Recognition (CVPR), Jul. 2017, pp. 6904– 6913.
[7] H. Xu and K. Saenko, “Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering,” Computer Vision – ECCV 2016. Springer International Publishing, pp. 451–466, 2016.
[8] X. Gao, Y. Qian, and A. Gao, “COVID-VIT: Classification of COVID-19 from CT chest images based on vision transformer models.” arXiv, 2021.
[9] A. M. H. Tiong, J. Li, B. LiS. Savarese, and S. C. H. Hoi, “Plug-and-Play VQA: Zero-shot VQA by Conjoining Large Pretrained Models with Zero Training.” arXiv, 2022.
[10] Ma, Jun, and Bo Wang. "Segment anything in medical images." arXiv preprint arXiv:2304.12306 (2023)



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY