International Journal of Research Publication and Reviews

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

A Comprehensive Study of Multimodal Approaches for Medical Visual Question Answering

Chinni Amrutha Vamsi[1]*, Ms. U. Archana*[2]

[1](21341A1231), [2]Assistant Professor
Bachelor of Technology in Information Technology
Department of Information Technology GMR Institute of Technology
An Autonomous institute, affiliated to JNTU-GV, Vizianagaram
G.M.R. Nagar, Rajam-532127, A.P

ABSTRACT

Medical Visual Question Answering is a combination of medical artificial intelligence and popular visual question answering challenges. Medical Visual Question Answering is a multimodal task which uses deep learning to answer clinical questions about medical images. The study aims to answer medical questions based on visual content of radiology images using deep learning. Deep learning is a method in artificial intelligence (AI) that teaches computers to process data in a way that is inspired by the human brain. The purpose of Medical Visual Question Answering is to develop Artificial intelligence system that can effectively bridge the gap between medical images and questions, it is used for efficient information retrieval. First, this study employed Image-to-Question (I2Q) consideration to model the relationships between the question and both the visual and linguistic contents. In particular, we add the search query using a textual encoded and extract image features utilizing the multi modal. We concatenate the resulting visual and textual representations and feed them into a multi-modal for generating the answer. Medical Visual Question Answering represents a promising convergence of deep learning, medical imaging, and natural language processing that holds significant potential to revolutionize various aspects of healthcare, research, and education.

Keywords: Medical Visual Question Answering; Deep Learning; Multimodal; transformer model.

1. INTRODUCTION

Medical Visual Question Answering (Med-VQA) is a field that focuses on answering natural language questions about medical images. It aims to provide accurate and convincing answers to clinically relevant questions. VQA is an improved version of Natural Language Question Answering(NLQA), where we are giving a natural language question to the machine and based upon some knowledge base provided, the machine generates a natural language answer. In this paper, VQA system discuss about medical images which answers questions based on the modality of the image. It can identify the image modalities such as Xray, Computed Tomography, ultra sound, magnetic resonance imaging, mammograph, angiogram, gastro intestine imaging and positron emission tomography. The process of the VQA task can be divided into three parts: extracting image features, extracting question features, and integrating features. For image featurization, commonly used Convolutional Neural Networks (CNNs) pre-trained on ImageNet include VGGNet , ResNet, and GoogLeNet . Question featurization techniques explored include bag-of-words, LSTM encoders. For doctors, Med-VQA systems can be used to assist diagnosis by providing them a second medical opinion. The systems can also be used in clinical education to train medical professionals.

2. LITERATURE SURVEY

Following research papers are studied in details to understands the proposed recommendation technique and experimental result for predicting the output

2.1 Lubna, A., Kalady, S., & Lijiya, A. (2019). Mobvqa: A modality based medical image visual question answering system, TENCON 2019-2019.

The paper discusses the development of a modality-based medical image visual question answering (VQA) system on the ImageCLEF 2019 medical VQA dataset.It can also be used by the patients for getting a basic information about the image without consulting the doctor. We have considered the problem of answering modality based questions for medical images like X-ray etc. The approach used here is to use a Convolutional Neural Network(CNN).The proposed model shows a testing accuracy of 83.8% which is comparable with state of the art.

The approach used in the paper involves the use of a Convolutional Neural Network (CNN) to classify the input image into its modality class and generate the answer based on the CNN output. The input question is first processed using natural language processing (NLP) techniques.The paper reports a testing accuracy of 83.8% for the proposed model, which is comparable to the state of the art.

2.2 Hasan, S. A., Ling, Y., Farri, O., Liu, J., Müller, H., & Lungren, M. (2018). Overview of imageclef 2018 medical domain visual question answering task.

The paper presents an overview of the Medical Visual Question Answering task (VQA-Med) at ImageCLEF 2019, which focuses on answering medical questions based on the visual content of radiology images.The authors created a new dataset of 4,200 radiology images and 15,292 question-answer pairs for the task, covering four categories of clinical questions: Modality, Plane, Organ System, and Abnormality at ImageCLEF 2019.

Depending on the different output layers, the model can be constructed as a classifier or generator for downstream tasks. The classification mode uses images and questions (the model only contains the left part of the dotted line). The class is predicted by the first item of output hc. The generative mode includes images, questions, and masked answers. The next word piece is predicted by the output of the first mask label haj.

On the Image CLEF 2019 VQA-Med data set, this mode is used to answer the questions of yes-no, modality, plane and organ system. the generative mode is used to give the answers of abnormality questions, since there is no candidate answer in this category. The best-performing team achieved a BLEU score of 64.4% and an accuracy of 62.4%.

2.3 Liu, B., Zhan, L. M., Xu, L., Ma, L., Yang, Y., & Wu, X. M. (2021, April). Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) (pp. 1650-1654). IEEE.

The paper presents the creation of SLAKE, a large-scale, semantically annotated, and knowledge-enhanced dataset for training and testing Med-VQA systems.The paper presents the creation of the SLAKE dataset. The dataset includes comprehensive semantic labels annotated by experienced physicians and covers various modalities (e.g., CT, MRI, and X-Ray), body parts (e.g., head, neck, and chest), and question types.

The SLAKE dataset was split into training, validation, and test sets at the image level, with a 75:15:15 ratio for each of the 8 categories: "head CT," "head MRI," "neck CT," "chest X-Ray," "chest CT," "pelvic cavity CT“..

2.4 Zhu, C., Zhao, Y., Huang, S., Tu, K., & Ma, Y. (2017). Structured attentions for visual question answering. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1291-1300).

The paper proposes a structured attention mechanism for visual question answering tasks, surpassing baseline models on the CLEVR and VQA datasets. The paper mentions three datasets that were used for evaluation: SHAPES dataset, CLEVR dataset, and VQA real-image dataset .

The dataset can be categorized into mainly 8 classes based on modality: angiogram, CT, MRI, X-ray, GI, mammography, PET and ultra sound. It covers ten organ systems such as vascular and lymphatic, gastrointestinal, breast, musculoskeletal, spine, skull, Heart, Lungs, Face and genitourinary. Convolutional Neural Network(CNN) to classify the input image to its modality class and thus generate the answer according to the CNN output.

The model outperformed the best baseline model on the CLEVR dataset by 9.5% and the best published model on the VQA dataset by 1.25%.

2.5 Ren, F., & Zhou, Y. (2020). Cgmvqa: A new classification and generative model for medical visual question answering. IEEE Access, 8, 50626-50636

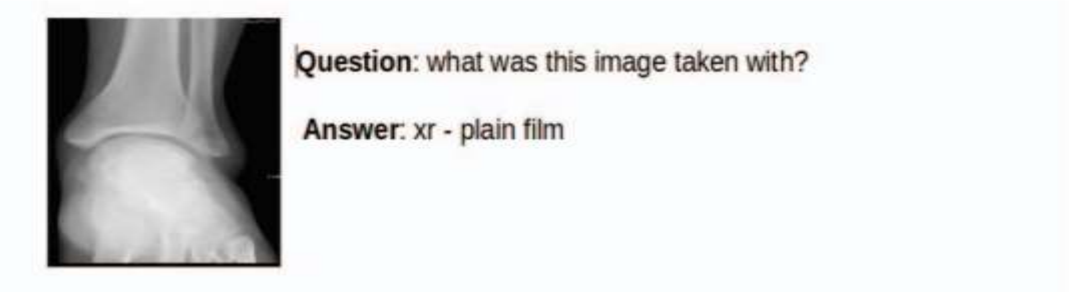
The paper proposes a model called CGMVQA that combines classification and answer generation capabilities to address the complex problem of medical visual question answering. Data augmentation is applied to images and tokenization is used for text processing .

The paper utilizes the ImageCLEF 2019 VQA-Med dataset for training, validation, and testing purposes. The dataset consists of 3200 medical images for training, 500 images for validation, and 500 images for testing. The model outperformed the best baseline model on the VQA-med dataset by 65.3% and the best published model on the VAQ dataset by 2.21%

3. METHODOLOGY

3.1 A modality based medical image visual question answering system :

The medical VQA task proposed by the ImageCLEF2019 challenge[12] consists of a dataset with 3200 training images, 500 validation images with related question-answer pairs and 500 test images with only questions. This dataset consists of questions in four categories: Modality based, Plane based, based on organ system and based on abnormality.



Question: what was this image taken with?

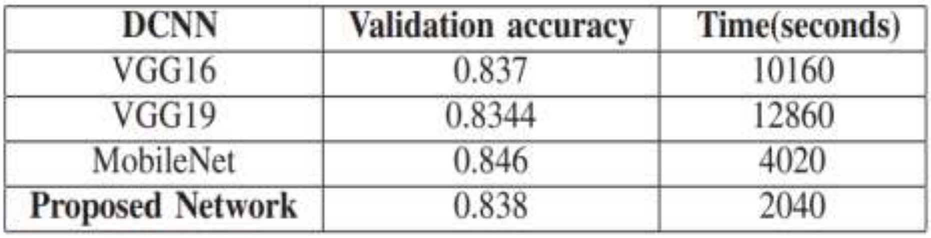
Answer: xr - plain film

1) The dataset can be categorized into mainly 8 classes based on modality: angiogram, CT, MRI, X-ray, GI, mammography, PET and ultra sound.

2) It covers ten organ systems such as vascular and lymphatic, gastrointestinal, breast, musculoskeletal, spine, skull, Heart, Lungs, Face and genitourinary.

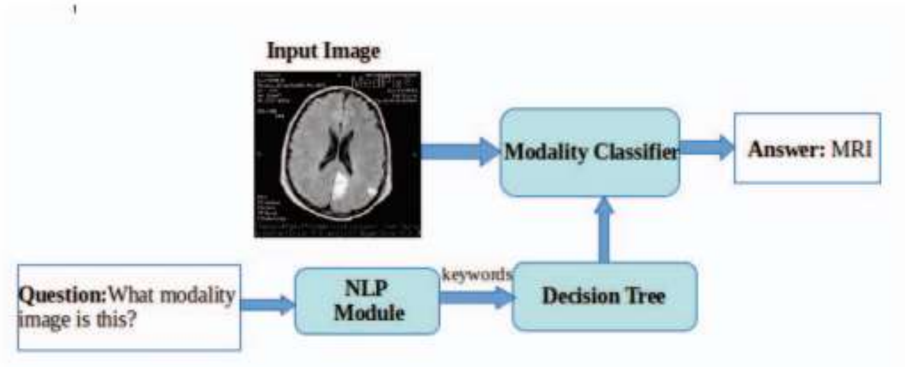
3) Convolutional Neural Network (CNN) to classify the input image to its modality class and thus generate the answer according to the CNN output.

COMPARISON WITH PRE-TRAINED MODELS



DCNN	Validation accuracy	Time(seconds)
VGG16	0.837	10160
VGG19	0.8344	12860
MobileNet	0.846	4020
Proposed Network	0.838	2040

The pre-trained network are trained on the image classification task of the dataset ImageNet which consists of natural images in 1000 categories. The pre-trained network have very complex structures with several layers and is computationally expensive. A training accuracy of 96.4% and testing accuracy of 83.8% was obtained for the given model.



1) VQA has applications like providing a second opinion to radiologists about their analysis of the image.

2) It can also be used by the patients for getting a basic information about the image without consulting the doctor.

3) We have considered the problem of answering modality based questions for medical images like X-ray etc.

4) The approach used here is to use a Convolutional Neural Network(CNN).

5) The proposed model shows a testing accuracy of 83.8% which is comparable with state of the art.

3.2 A semantically-labeled knowledge-enhanced dataset for MVQA :

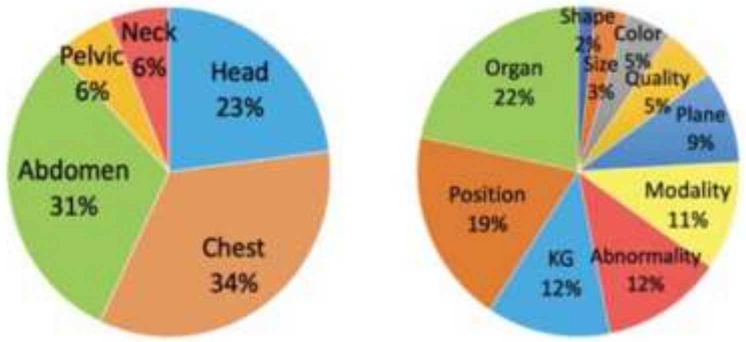


Figure 5 : System Architecture

1) The images include 140 head CTs or MRI. The distribution is shown in Figure 2

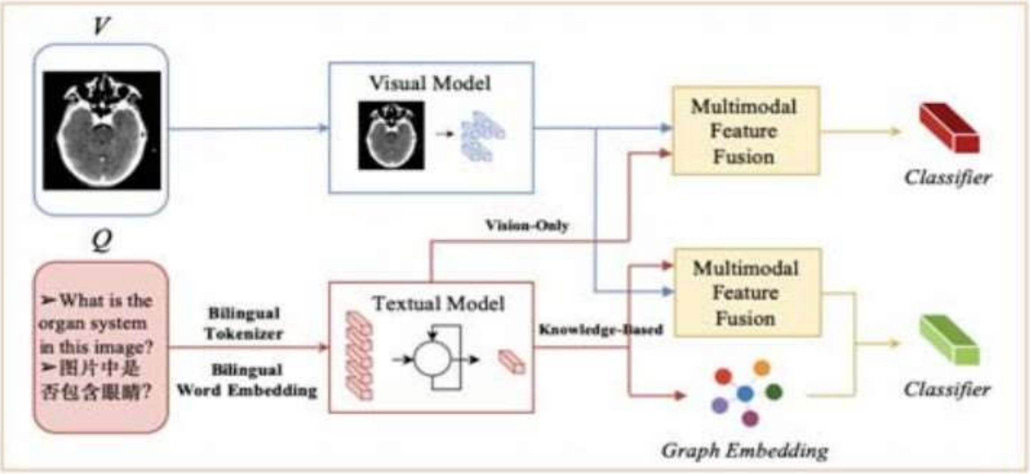
2) The number of images for each body part is set based on the complexity of the body part.

3) For example, the number of diseases and organs in abdomen is much more than that in neck, so there are more images of abdomen than neck in the dataset.

Dataset:

1) We elaborate on the construction of our SLAKE dataset. In general, we ensure the diversity of the dataset in terms of modalities (e.g., CT, MRI, and X-Ray)

2) It also covered body parts (e.g., head, neck, and chest), and question types. (e.g., vision-only, knowledge based)



Algorithm:

1.Naive bayes:

Naive Bayes (NB) classifiers belong to the family of probabilistic models. Naive Bayes is a probabilistic classifier commonly used in machine learning. It's based on Bayes' theorem and assumes independence between features, which simplifies calculations. Naive Bayes is often used in text classification.

2. Support Vector Machine (SVM):

A Support Vector Machine (SVM) is an intelligent way to draw lines, or hyperplanes, in high-dimensional spaces to separate things into categories, like cats and dogs. It looks for the best line that maximizes the gap between the categories, making it useful for tasks like image recognition and email sorting.

3. LSTM Networks:

LSTM stands for long short-term memory networks, used in the field of Deep Learning. It is a variety of recurrent neural networks (RNNs) that are capable of learning long-term dependencies, especially in sequence prediction problems. It aimed to deal with the vanishing gradient problem present in traditional RNNs. LSTMs have been widely adopted in deep learning applications for their ability to model and understand sequential patterns in data.

3.3 Structured attentions for visual question answering:

Dataset:

1) Since we have found MF-SIG and LBP-SIG are the best on CLEVR, in this part, we mainly compare the two models with different.

2) The optimal choice in these experiments is MF-SIG-T3, which is 0.92% higher in overall accuracy than the previous best method

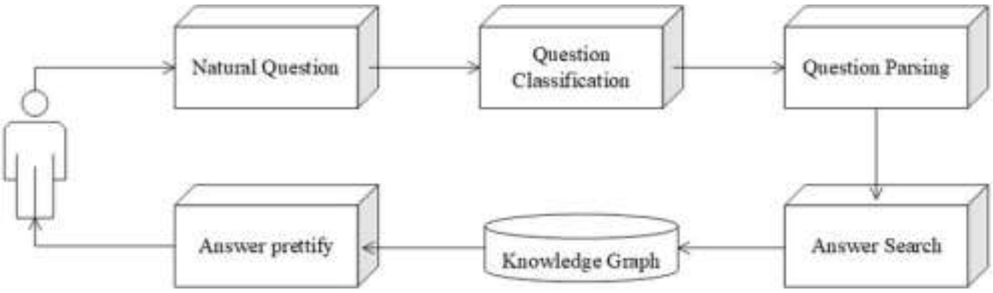
3) Outperforms previous methods on all 3 general categories of questions.

4) We then use external data from Visual Genome to train MF-SIG-T3 and MF-T3.

5) The maximum margin of MF/LBP vs. SIG on overall accuracy is 2.62% and 1.33% w

6) The vocabulary sizes for the questions and answers are 82 and 28 respectively.

7) best model surpasses the best baseline model by more than 9.5% on the test set.



The SHAPES dataset:

1) It is a synthetic dataset consisting of images containing 3 basic shapes in 3 different colors with a resolution of 30×30, and queries about the arrangements of the basic shapes, as shown in

2)The answer is "yes" when the image satisfies the query, and "no" otherwise. There are 3 different lengths of queries.

3) The original dataset [1] has 14592 and 1024 image/question pairs for the training and test sets. All the queries in the test set do not appear in the training set

4) The parser-based method may not perform well in more general tasks such as the VQA

3.4 A new classification and generative model for MVQA:

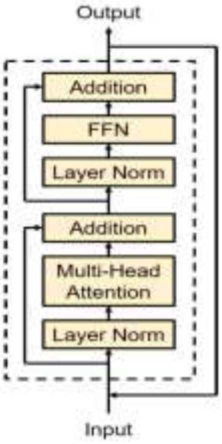


FIGURE 3. Architecture of pre-layer-normalization multi-head self-attention and feed-forward network transformer.

1) There is no candidate answer to the abnormality question, which is different from that of other categories.

2) We employ the generative mode to obtain the answer. Unlike the classification mode, we add the masked answer in generative mode training.

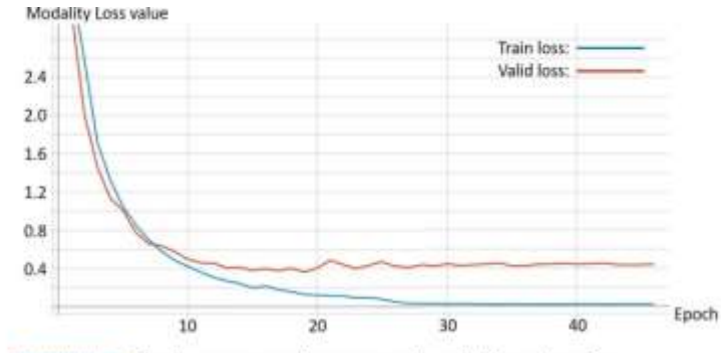


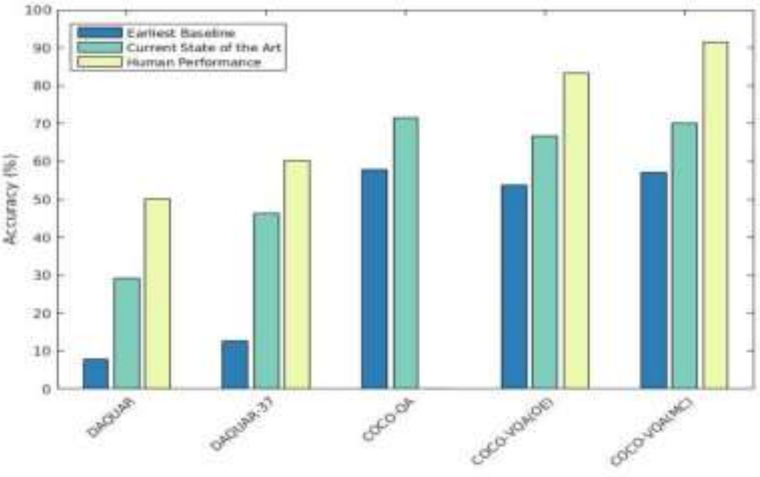
FIGURE 10. **The loss curves of yes-no and modality categories.**

1) In the modality category with the largest number of classes, our model has an improvement rate of 15.2% in accuracy compared to the baseline.

2) The data enhancement we used is an transformation based on the existing image.

2) High frequency predictions account for 22.6%.

3.5 Medical domain visual question answering task:



1) The predictive power of language over images have been corroborated by ablation studies.

2) They found that the image-only model’s predictions differed from the combined model 40% more often than the question only model.

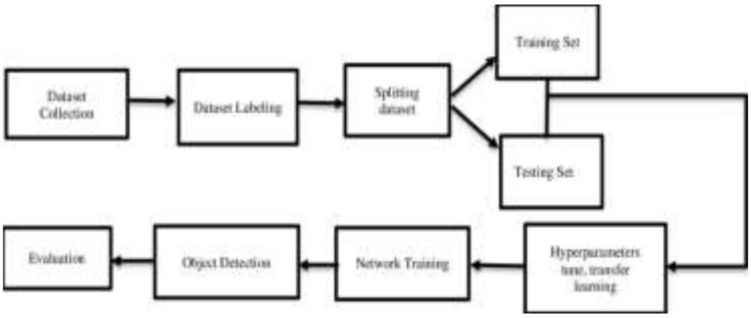
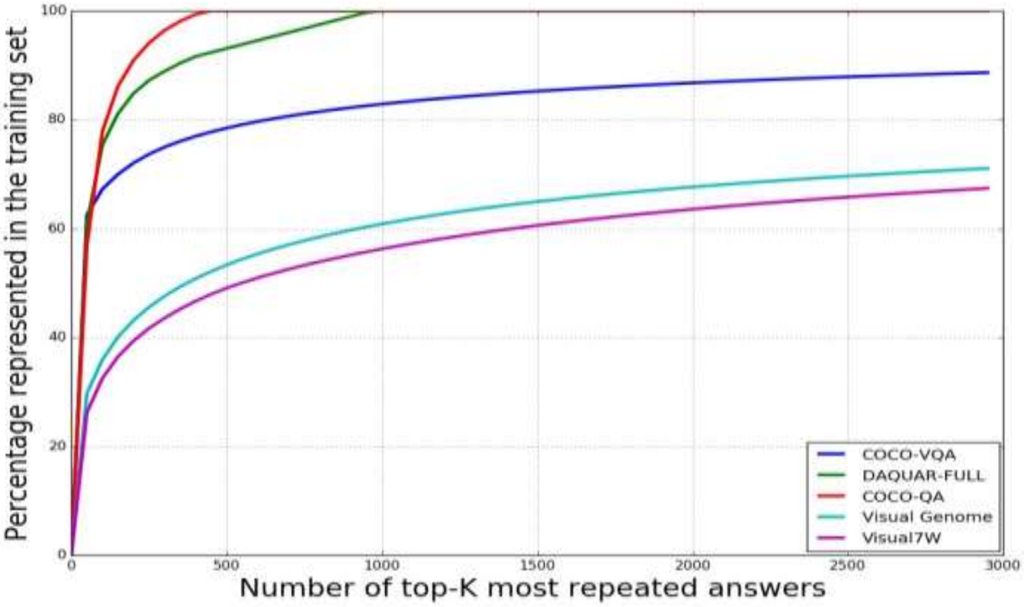


Figure 10 : Flow Diagram for MVQA

1) QA pairs are created for images using an Natural Language Processing (NLP) algorithm that derives them from the COCO image captions.

2) Most questions ask about the object in the image (69.84%), with the other questions being about color (16.59%), counting (7.47%) and location (6.10%)



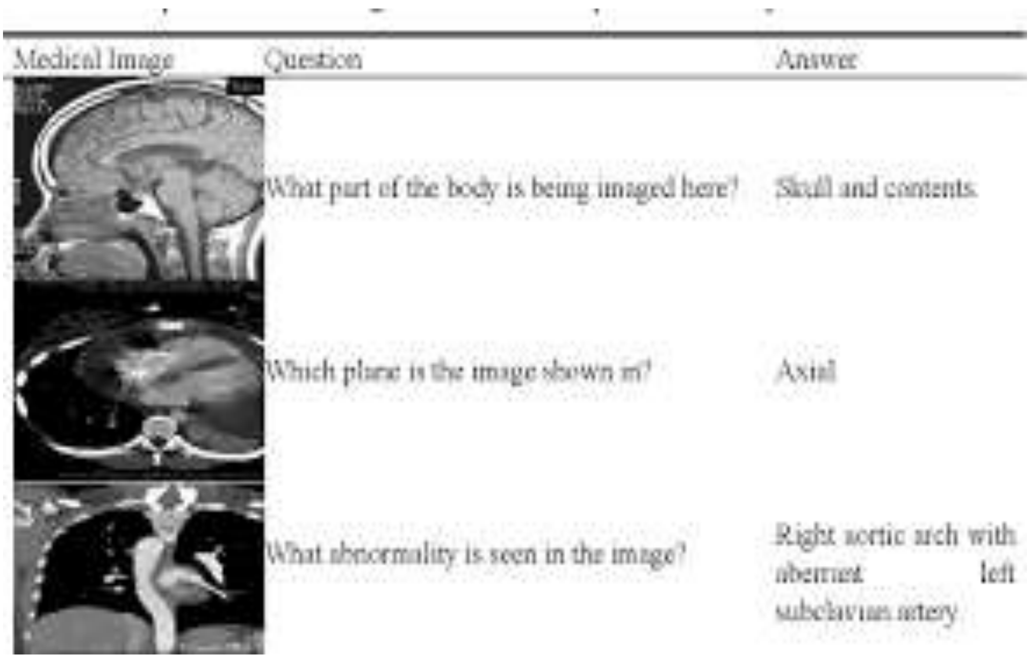
1) This graph shows the long-tailed nature of answer distributions in newer VQA datasets.

2) choosing the 500 most repeated answers in the training set would cover a 100% of all possible answers in COCO-QA but less than 50% in the Visual Genome dataset.

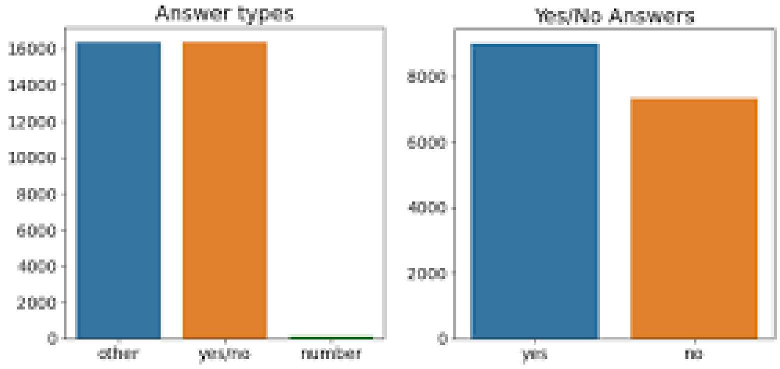
3) VQA datasets may indicate that it is only capable of analyzing images in a limited manner

4. RESULTS

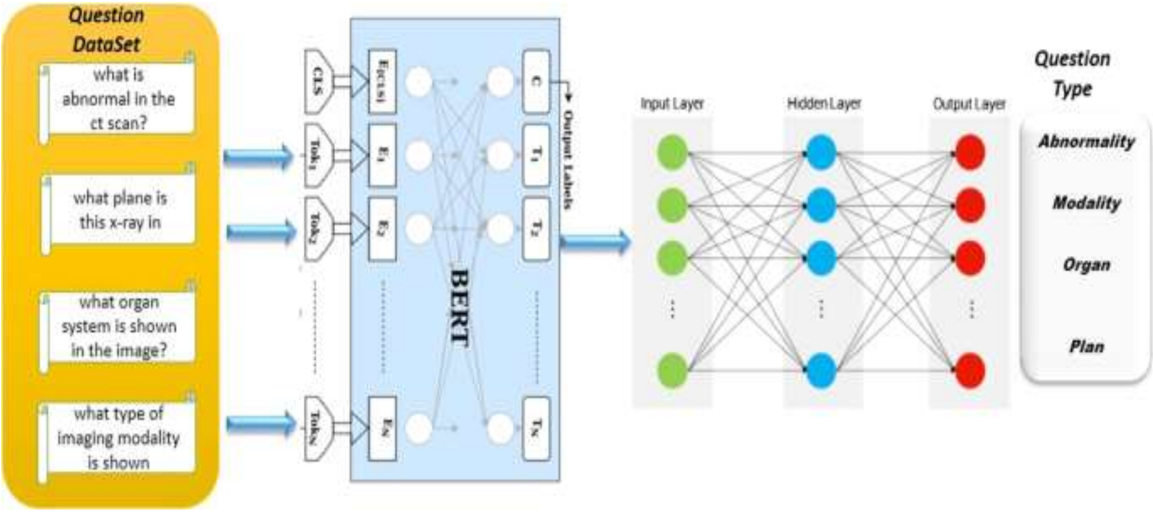
4.1 A modality based medical image visual question answering system:

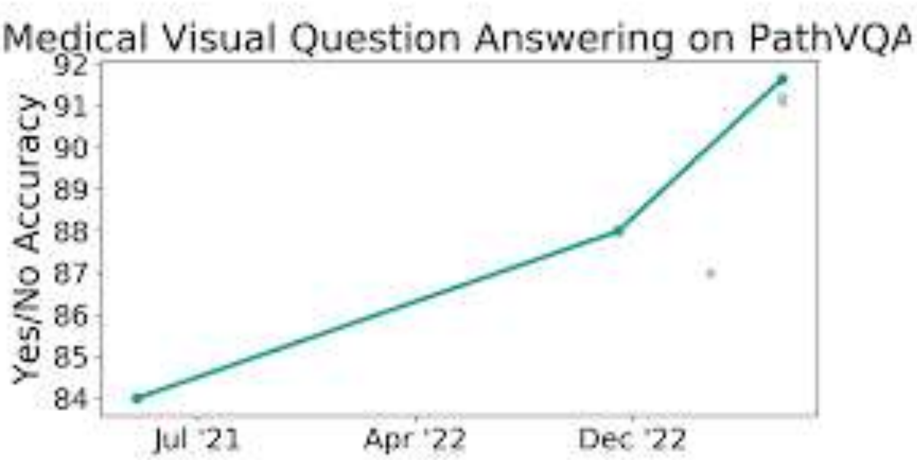


Medical Image	Question	Answer
	What part of the body is being imaged here?	Skull and contents.
	Which plane is the image shown in?	Axial
	What abnormality is seen in the image?	Right aortic arch with aberrant left subclavian artery

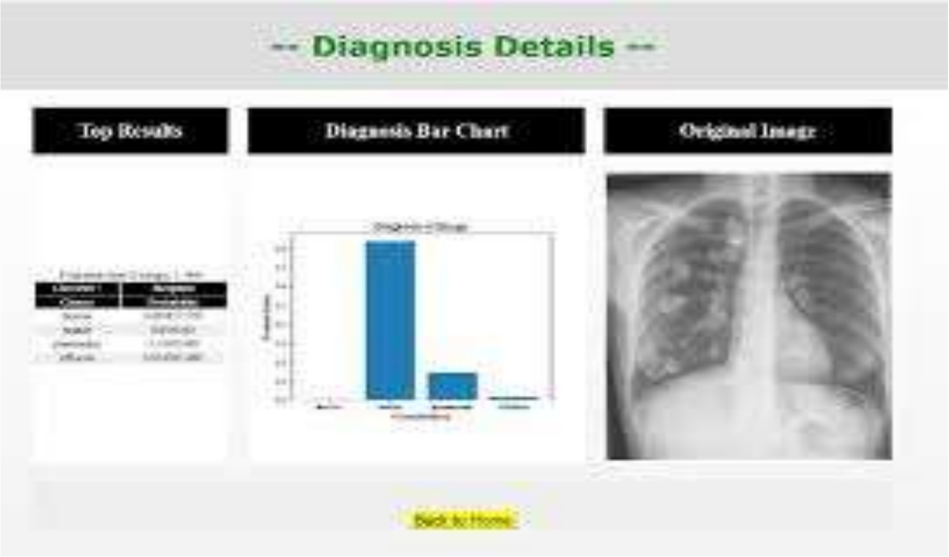


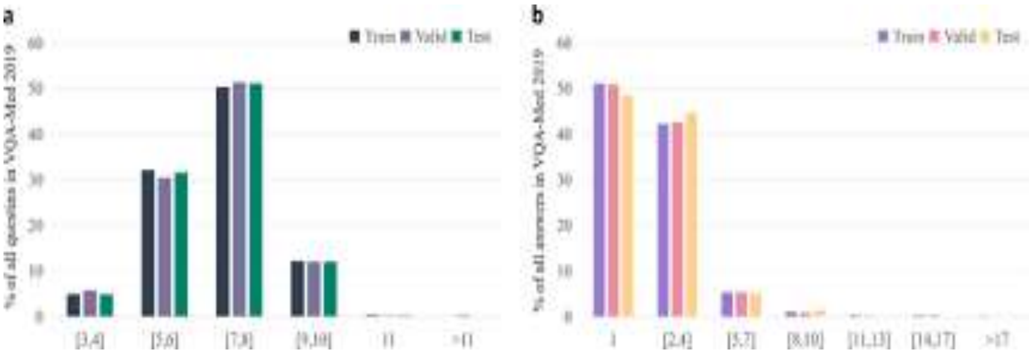
4.2 A semantically-labeled knowledge-enhanced dataset for MVQA:



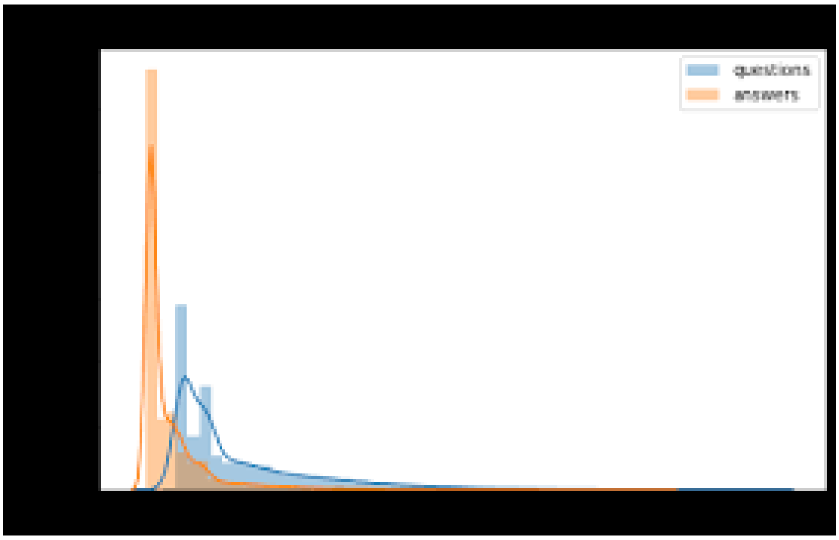


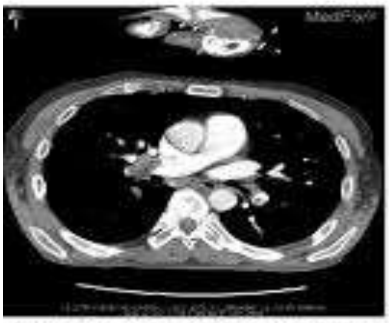
4.3 Structured attentions for visual question answering:



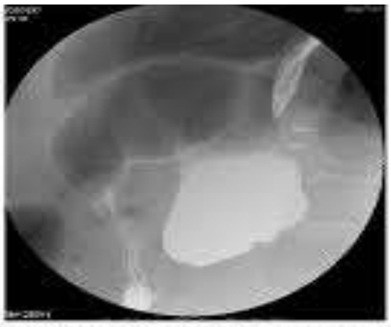


4.4 A new classification and generative model for MVQA:



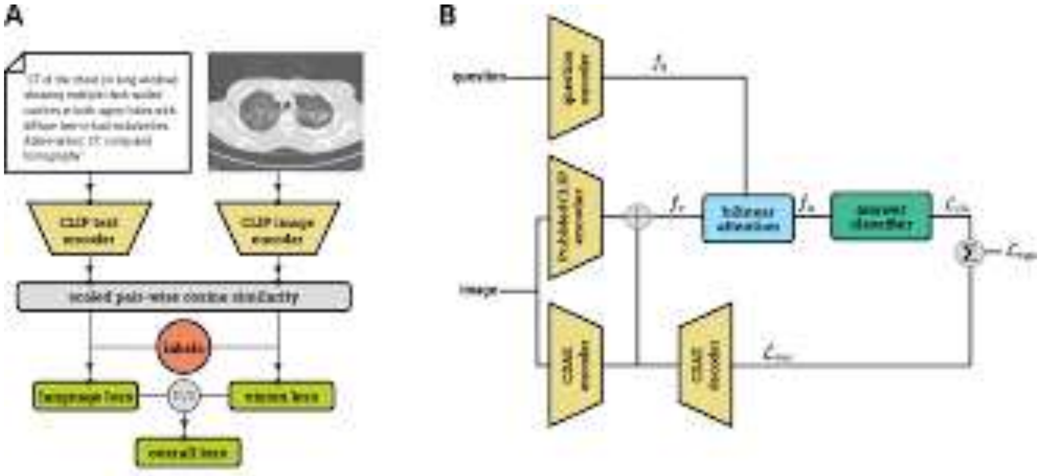


(g) **Q**: which organ system is shown in the ct scan? **A**: lung, mediastinum, pleura



(h) **Q**: what is abnormal in the gastrointestinal image? **A**: gastric volvulus (organoaxial)

4.5 Medical domain visual question answering task:



5. CONCLUSION

In conclusion, Medical Visual Question Answering (Med-VQA) stands at the forefront of innovation, merging medical artificial intelligence with visual question answering capabilities. This pioneering approach leverages deep learning to effectively respond to clinical inquiries rooted in the visual content of medical images, aiming to bridge the gap between these images and pertinent questions for efficient information retrieval.The integration of Image-to-Question considerations, multi-modal representations, and the utilization of advanced technologies such as Convolutional Neural Networks (CNNs) and linguistic encoders showcases the potential of Med-VQA to revolutionize healthcare, research, and education. It promises accurate, clinically relevant responses to natural language queries about various medical imaging modalities, from X-rays to MRIs and more.

Ultimately, Med-VQA systems hold immense promise for medical professionals, offering valuable support by providing secondary medical opinions to aid in diagnoses. Moreover, these systems have the potential to significantly contribute to the education and training of medical practitioners, thereby shaping the future landscape of medical imaging and diagnostics.

6. REFERENCES

1. M. Malinowski, M. Rohrbach, and M. Fritz, “Ask your neurons: A deep learning approach to visual question answering,” arXiv preprint arXiv:1605.02697, 2021.

2. Q. Wu, P. Wang, C. Shen, A. Dick, and A. V. D. Hengel, “Ask me anything: Free-form visual question answering based on knowledge from external sources,” in Proc. CVPR, 2020..

3. J. Lu, J. Yang, D. Batra, and D. Parikh, “Hierarchical Question-Image Co-Attention for Visual Question Answering,” arXiv preprint arXiv:1606.00061v5, 2021

4. S.A. Hasan, Y. Ling, O. Farri, J. Liu, H. Muller, and M. Lungren, “Overview of ImageCLEF 2018 Medical Domain Visual Question Answering Task,” in Proc. CEUR Worskshop, 2022

5. H. Noh, P. H. Seo, and B. Han, “Image question answering using convolutional neural network with dynamic parameter prediction,” in Proc. CVPR, 2020