Group- 8 :Visual Instruction Tuning for Medical Image Question Answering

Aditya Dhavala **Vaishnavi Gadhikar** **Vineeth Veligeti** **Divyagna Bavikadi**

Abstract

Medical Image Question Answering (MIQA) has emerged as a critical area within healthcare informatics, aiming to leverage the vast volumes of medical imaging and text data to enhance clinical decision-making and patient care. This literature survey explores the recent advancements, methodologies, and challenges in MIQA systems. With the proliferation of deep learning techniques, particularly convolutional neural networks (CNNs) and natural language processing (NLP) models, MIQA frameworks have witnessed significant progress in understanding and interpreting medical images based on textual queries. Through an extensive review of the existing literature, this survey delves into the key components of MIQA systems, including image feature extraction, question formulation, and answer generation. Furthermore, it investigates various datasets and evaluation metrics commonly used in assessing the performance of MIQA models. Despite notable achievements, several challenges persist, such as domain-specific language understanding, model interpretability, and generalization across different medical specialties and modalities.

1 Introduction

The development of Medical Image Question Answering (MIQA) systems is motivated by the pressing need to address the escalating complexity and volume of medical imaging data within healthcare. Traditional manual interpretation methods are time-consuming and prone to variability, leading to potential delays in diagnosis and treatment initiation. MIQA systems offer a solution by automating image analysis and interpretation, thus expediting diagnostic processes and treatment planning. Moreover, MIQA facilitates personalized medicine by integrating patient-specific data to tailor treatment strategies according to individual needs and preferences. This personalized approach not only optimizes treatment outcomes but also enhances patient satisfaction and quality of life. Additionally, MIQA research fosters the integration of artificial intelligence (AI) into clinical practice, enabling interdisciplinary collaboration between computer scientists, healthcare professionals, and biomedical researchers. By driving innovation and advancing healthcare informatics, MIQA contributes to the broader goal of leveraging technology to improve human health and well-being. In addition to expediting diagnosis and treatment planning, MIQA systems hold the potential to revolutionize healthcare delivery by facilitating more efficient and accurate clinical workflows. By automating the interpretation of medical images, MIQA systems reduce the burden on healthcare professionals, allowing them to allocate more time to patient care and complex medical decision-making. Furthermore, MIQA enables the extraction of nuanced insights from medical imaging data that may not be readily apparent through traditional manual methods. This deeper understanding of imaging data can lead to earlier detection of pathologies, more precise treatment recommendations, and improved prognostic assessments, ultimately enhancing patient outcomes and reducing healthcare costs.

Moreover, MIQA research addresses critical challenges in healthcare informatics, such as the integration of heterogeneous data sources and the interoperability of healthcare systems. By developing

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

robust MIQA frameworks that can seamlessly integrate with existing electronic health record (EHR) systems and medical imaging repositories, researchers are laying the foundation for a more interconnected and data-driven healthcare ecosystem. This integration not only improves the accessibility and usability of medical imaging data but also enhances collaboration among healthcare providers and researchers, driving advancements in clinical practice and medical research.

Furthermore, MIQA systems offer insights into the underlying biological mechanisms of diseases and the effectiveness of various treatment modalities. By analyzing patterns and correlations within large-scale medical imaging datasets, MIQA can identify novel biomarkers, uncover disease subtypes, and predict treatment responses, thus informing the development of personalized therapeutic interventions. Additionally, MIQA research contributes to the refinement and validation of AI algorithms for medical applications, addressing concerns related to reliability, interpretability, and ethical considerations. By rigorously evaluating MIQA models using diverse datasets and standardized evaluation metrics, researchers can ensure the robustness and generalizability of these systems across different clinical settings and patient populations.

In summary, the development of MIQA systems represents a pivotal advancement in healthcare informatics, with far-reaching implications for clinical practice, medical research, and healthcare delivery. By leveraging AI technologies to bridge the gap between medical imaging data and clinical decision-making, MIQA holds the promise of enhancing diagnostic accuracy, treatment efficacy, and patient outcomes. Furthermore, MIQA research drives innovation in healthcare informatics by addressing challenges related to data integration, interoperability, and algorithm validation, ultimately paving the way for a more efficient, effective, and patient-centered healthcare system.

2 Data Set

2.1 PMC : Pub Med Central

The PMC (PubMed Central) dataset is a collection of full-text biomedical and life sciences research articles made freely available by the National Institutes of Health (NIH) through the PubMed Central repository. PubMed Central is a digital archive of peer-reviewed scientific literature in the fields of biomedicine and life sciences[7].
Here are some key points about the PMC dataset:

Content
The PMC dataset includes a wide range of biomedical and life sciences research articles spanning various disciplines, including but not limited to medicine, biology, genetics, neuroscience, public health, and bioinformatics.

Structured Metadata
In addition to the full-text articles, the PMC dataset also includes structured metadata associated with each article. This metadata may include information such as article title, authors, abstract, publication date, journal information, keywords, and MeSH (Medical Subject Headings) terms.

Data Availability
The PMC dataset is available for download in various formats, including XML, plain text, and other structured formats. Researchers and developers can access the dataset directly from the PubMed Central website or through specialized repositories and APIs.

Research and Analysis
The PMC dataset is a valuable resource for researchers, data scientists, and developers working in biomedical informatics, natural language processing (NLP), text mining, machine learning, and related fields. It enables various types of research and analysis, including but not limited to literature reviews, meta-analyses, text classification, information extraction, and knowledge discovery.

Usage and Citation
While the PMC dataset is freely available for non-commercial use, users are often required to adhere to certain terms and conditions specified by PubMed Central. Additionally, proper citation of the

original articles and appropriate acknowledgment of PMC as the data source are encouraged.

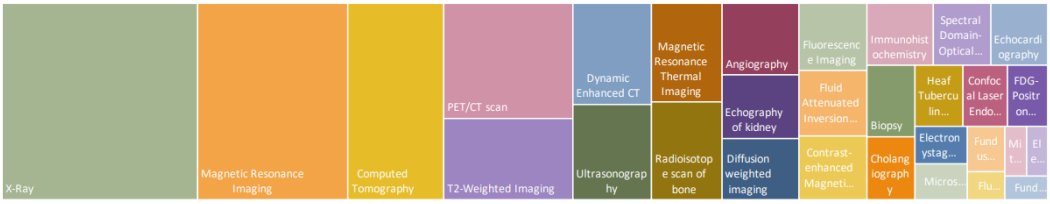

Figure 1: The top 20 figure types in PMC-VQA, cover a wide range of diagnostic procedures.

2.2 VQA-RAD

The VQA-RAD dataset addresses the need for automated systems in radiology by providing a comprehensive resource for visual question answering (VQA) within the domain of medical imaging. It aims to assist radiologists in interpreting diagnostic images efficiently and accurately.

Dataset Composition: The dataset comprises a diverse collection of medical images, predominantly X-rays, covering various anatomical regions and clinical conditions. Each image is accompanied by a set of questions designed to test different aspects of medical diagnosis and interpretation.

Question Types: The questions in VQA-RAD cover a wide range of topics related to medical imaging, including identifying abnormalities, assessing disease severity, and making differential diagnoses based on visual cues present in the images. These questions are formulated to mimic real-world scenarios encountered by radiologists.

Annotations: In addition to questions and images, the dataset provides annotations for referring expressions and attributes. Referring expressions annotate words or phrases in the questions that refer to specific regions or features within the images, while attributes describe characteristics or properties of the objects depicted in the images.

Purpose and Applications: VQA-RAD serves as a valuable resource for training and evaluating VQA models tailored to the domain of radiology. It enables the development of algorithms that can automatically answer questions about medical images, leading to improved computer-aided diagnosis systems and decision support tools for radiologists.

Research Opportunities: Researchers can leverage VQA-RAD to study various aspects of medical image understanding, including the interpretation of natural language questions in the context of radiology. The dataset facilitates investigations into the interaction between textual and visual information in medical image analysis, advancing the field of healthcare AI.

3 Comparative study

3.1 Initial Approach

We initially planned to use the VQA RAD[1] dataset to create a Neuro Symbolic Visual Question Answering (NSVQA) model. The problem statement for the VQA model was to answer questions about a given image. The goal of Neuro Symbolic VQA[5] is to develop a model that can autonomously acquire visual concepts, language comprehension, and sentence semantic parsing without explicit supervision. Rather, the model picks up knowledge by looking at pictures and analyzing questions and responses in pairs. The suggested model generates executable, symbolic programs from natural language queries and builds an object-based scene representation. In order to enable communication between the language and visual domains, these programs are implemented on the latent scene representation using a neuro-symbolic reasoning module. The acquired visual notions also help in learning new words and comprehending new sentences[3]. This approach is specifically applied to VQA tasks within radiology data, leveraging the unique characteristics of the VQA RAD dataset. Due to the challenges we faced with this approach we decided to study other state-of-the-art approaches for medical VQA.

3.1.1 Challenges Encountered

Lack of Sufficient Data
The Neuro Symbolic VQA model was initially trained on the CLEVR dataset which consists of synthetic images of 3D-rendered objects placed on a blank background in a closed-world setting, unlike our application. We decided to take a subset of the medical data and acquire external knowledge to convert them to a symbolic domain, but this would require expert curation. We suggest providing some feature values to enclose a knowledge base to make it usable in a more logical domain.

Synthetic Data from CLEVR Dataset
The objects in the dataset come in various shapes, colors, sizes, and materials, and they are arranged in different spatial configurations. Each image in the dataset is accompanied by a set of questions that require complex reasoning about the scene depicted in the image. The questions in the CLEVR dataset are designed to assess the model's ability to understand the relationships between objects, their properties, and their spatial arrangements.

Real Medical Images in RAD Dataset
The RAD dataset is a collection of medical images, particularly focusing on radiology images, used for training and testing algorithms in medical image analysis and computer-aided diagnosis (CAD) systems and does not contain 3D rendered images to pinpoint spatial data. Also, for answering the spatial questions we explored a method where we relate the location of the pixel to sub-labels of the image but realized this would not be a generalized solution.

Scale of the CLEVR Dataset
The original CLEVR dataset consists of 700,000 images, each accompanied by questions and answers. We explored data augmentation of the medical images but learned that this is creating a ias and learning spurious correlations in the machine learning algorithms.

Limitations of Radiology Datasets
The size of radiology datasets is often limited by factors such as data collection and annotation costs, as well as privacy concerns regarding patient data. This makes it more important for the processed data passed to the machine learning algorithm to contain factors that influence and correlate to the corresponding answers.

3.2 Other approaches

The main dataset we found is PMC (PubMed Central). the state of the art model we found on this dataset is PMC-VQA on Visual question answering, PMC-Llama for textual question answering.

3.2.1 PMC Llama

PMC-LLaMA (PubMed Central Language Model for Medical Applications) is a specialized language model tailored specifically for medical applications. It represents a significant advancement in adapting general-purpose foundation language models to the medical domain, which demands precision, accuracy, and domain-specific knowledge. The development of PMC-LLaMA involved systematically integrating biomedical academic papers and medical textbooks into the training process, ensuring that the model captures and synthesizes domain-specific knowledge effectively.

Development and Training:
The development of PMC-LLaMA primarily involved fine-tuning an existing language model called LLaMA on a large corpus of medical literature. This corpus consisted of approximately 4.8 million biomedical academic papers sourced from PubMed Central (PMC) and 30,000 medical textbooks. The fine-tuning process aimed to align the model with domain-specific nuances and instructions, enhancing its ability to understand and generate medical text. Additionally, a comprehensive dataset was utilized for instruction tuning, encompassing various medical tasks such as question-answering (QA), rationale for reasoning, and conversational dialogues, totaling approximately 202 million tokens.

Model Variants and Performance:
PMC-LLaMA is available in several versions, with PMC-LLaMA-13B being the latest iteration

fine-tuned on an instructions-following dataset. This version has demonstrated improved performance in following user instructions compared to its predecessor, MedLLaMA-13B. The PMC-LLaMA-13B model is easily accessible through the transformers library in Python and has been observed to converge more rapidly and achieve lower loss compared to the original LLaMA model, particularly when trained on larger datasets with more trainable parameters.

Significance and Applications:
PMC-LLaMA holds significant promise for various medical applications, including clinical decision support, medical information retrieval, automated medical report generation, and medical education. By leveraging the vast amount of medical literature available in PubMed Central and medical textbooks, PMC-LLaMA offers a robust foundation for understanding and generating medical text with a high degree of accuracy and relevance. Its ability to follow user instructions effectively makes it a valuable tool for assisting healthcare professionals in tasks such as medical documentation, literature review, and evidence-based decision-making.

In summary, PMC-LLaMA represents a pivotal advancement in adapting language models to the specialized domain of medicine, offering a powerful resource for advancing medical research, practice, and education.

3.2.2 PMC VQA

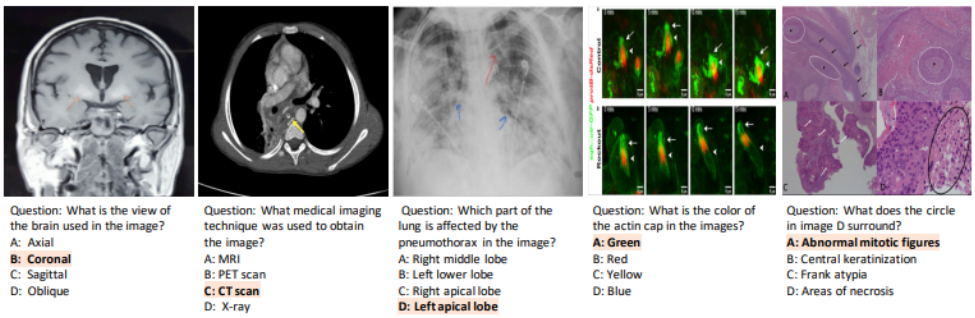

Figure 3: Several examples of challenging questions and answers along with their respective images. To answer questions related to these images, the network must acquire sufficient medical knowledge, for example, for the first two images, it is essential to recognize the anatomy structure and modalities; for the third image, recognizing the X-ray image pattern of pathologies is necessary; for the final two images, apart from the basic biomedical knowledge, the model is also required to discern colors, differentiate subfigures, and perform Optical Character Recognition (OCR).

The PMC-VQA (PubMed Central Visual Question Answering) model is a generative model designed to predict answers to questions based on both an image and a text caption. It utilizes a modified version of the T5 model, a text-to-text transformer architecture developed by Google. The PMC-VQA model leverages pre-trained visual and language models and is fine-tuned on the PMC-VQA dataset.

Here's a breakdown of the PMC-VQA model architecture and its components:

Visual Feature Extractor: Responsible for extracting visual features from the input image. Utilizes the CLIP (Contrastive Language-Image Pre-training) model, which is trained to associate images and text captions. CLIP extracts visual features from the image, which are then used as input to the text-to-text transformer.

Text-to-Text Transformer: Based on the T5 model, a transformer-based architecture trained on various text-based tasks. Fine-tuned on the PMC-VQA dataset, enabling it to generate answers to questions given both image and text inputs. Takes as input the question and text caption concatenated with a special token, along with the extracted visual features represented as embeddings. Generates the answer to the question based on the combined input.

Training Data:
Question-answer pairs used for training the PMC-VQA model are generated by a pre-trained teacher model called MedVint-TE. MedVint-TE is a variant of the T5 model architecture, specifically fine-

tuned on the PMC-VQA dataset. Question-answer pairs generated by MedVint-TE serve as training data for the PMC-VQA model, enabling it to learn to predict answers to questions.

The PMC-VQA model is designed to handle visual question answering tasks, where understanding both visual content (image) and textual context (caption and question) is essential. By leveraging pre-trained models and fine-tuning on the PMC-VQA dataset, the model can effectively generate accurate answers to questions posed about the content of biomedical and life sciences research articles available in the PubMed Central repository.

3.2.3 BioMedGpt

Introduction

The paper introduces BiomedGPT, a unified biomedical generative pre-trained transformer that can be fine-tuned for various biomedical tasks, including vision, language, and multimodal tasks. The authors aim to address the challenges of limited data and domain-specific models in biomedical research by developing a single, versatile model that can be adapted to different tasks and datasets.

Background

The authors provide an overview of the current state of biomedical research, highlighting the importance of integrating multimodal data and the limitations of traditional approaches. They also discuss the recent advancements in transformer-based models and their applications in natural language processing and computer vision.

Methodology

The BiomedGPT model is based on the transformer architecture, which consists of an encoder and a decoder. The encoder is responsible for processing input sequences, while the decoder generates the output sequence. The authors use a combination of vision and language tasks to pre-train the model, including:

1. Image captioning: generating captions for biomedical images 2. Medical text classification: classifying medical text into different categories 3. Multimodal sentiment analysis: analyzing sentiment in biomedical text and images

The pre-training process involves optimizing the model's parameters using a combination of cross-entropy loss and masked language modeling. The authors also introduce a new task, "image-text matching," which involves matching biomedical images with corresponding text descriptions.

Experiments

The authors conduct extensive experiments to evaluate the performance of BiomedGPT on various biomedical tasks, including:

1. Image captioning: BiomedGPT outperforms state-of-the-art models on the Medical Image Captioning (MIC) dataset 2. Medical text classification: BiomedGPT achieves competitive results on the 20 Newsgroups dataset 3. Multimodal sentiment analysis: BiomedGPT outperforms state-of-the-art models on the BioSentiment dataset 4. Image-text matching: BiomedGPT achieves state-of-the-art results on the Image-Text Matching dataset

Results

The authors present a comprehensive analysis of the results, highlighting the strengths and limitations of BiomedGPT. They also discuss the potential applications of the model in biomedical research, including:

1. Automated image annotation 2. Biomedical text analysis 3. Multimodal data integration

3.3 Evaluation

Upon examination of medical Visual Question Answering (VQA) datasets and methodologies, we have identified several prevailing issues. In contrast to conventional VQA in general domains, the specialized needs and real-world implementation scenarios within the medical context pose distinct and novel challenges.

PMC-Llama reported an average Accuracy of 64.43 among three MedQA datasets. MedVInT-TD (7B parameters) reported an average Accuracy of 86.6 for VQA-RAD dataset, while BioMedGPT-B gave an accuracy of 81.3 for the same dataset with just 182 M parameter sized model. Compared to VQA in a generic sense, medical VQA needs to be evaluated considering the class imbalance and that the samples are more skewed to certain classes, so it will be interesting to see other metrics like MaP, Auc etc.

For VQA-RAD, the common step among the top three approaches is to have extra pre-training on their image encoders. it shows that enhancement of the image encoder is the essential ingredient of achieving state-of-the-art performance in most medical VQA datasets.

Another important finding is that for most datasets, on average the attention-based fusion algorithms outperform the ones without attention. it's been observed in VQA-Med datasets and many others. It suggests that attention-based fusion algorithms are suitable for medical VQA datasets.

3.3.1 Question Diversity

Question diversity is one of the most significant challenges of medical VQA. The VQA-RAD[8] investigates the natural questions in clinical conversation. The questions can be categorized into modality, plane, organ system, ab- normality, object/condition presence, positional reasoning, color, size, attribute other, counting, and other. In other datasets such as VQA-Med-2019 [9] and PathVQA [10], the question categories tend to be less diverse than the VQA- RAD dataset. In the RadVisDial [11], VQA-Med-2020 [12], and VQA-Med-2021 [13], the question category is reduced to the abnormality presence only. However, most questions about an abnormality in existing datasets are about presence without further inquiry, like the location of tumors or tumor size. Therefore, questions remain to be added to diversify the medical VQA dataset. Future research should concentrate on defining relevant question categories that are in line with practical needs in order to improve the integration of medical Visual Question Answering (VQA) into clinical processes. For example, information about the imaging modality and particular organs being studied might not be required as these are usually recorded in study records, which could avoid using the VQA system at all. Increasing the number of data sources is crucial. Currently, the majority of questions in medical VQA datasets are synthesized from image captions and medical reports, which restricts the range of topics that can be covered. Textbooks and other unofficial sources could be added to the textual corpus to enhance its quality. Moreover, collecting real-world clinical conversations especially ones that involve patient interactions would yield insightful information about realistic needs. It's critical to loosen the requirement that questions only relate to visual material. In real-world clinical settings, conversations on prognosis and illness progression frequently take place in addition to the visible visual content. Therefore, a key component of dataset design is question diversity to provide medical VQA systems with thorough coverage, practical applicability, and improved user engagement.

3.3.2 Integrating additional information

Including the new data into the inference process presents another difficulty for medical VQA. One example is the finding of Kovaleva et al. [11] that answering questions more effectively is achieved when the patient's medical history is addressed.

3.3.3 Interpretability and Reliance

Deep learning has long struggled with interpretability. Interpretability is a problem for medical VQA because deep learning techniques are frequently the foundation of these models. Interpretability is more significant to the medical VQA system than it is to the broader domain since a poor decision could have disastrous results. It establishes the reliability of the anticipated answer. This issue has been addressed by general-domain VQA researchers, who have looked into a number of approaches to assess a model's capacity for inference.

3.3.4 Generalizability

In medical AI research, generalizability is a common topic and a necessary concern for systems operating in real-world settings. The practical input may be outside of the training data's distribution (OOD), which is the root cause of the generalizability problem. The patient's race and the imaging equipment are just two examples of the many variables. Domain shift is the term used to describe

the difference in data distributions. The generalizability challenge is more complex and dual for downstream activities such as VQA. The VQA models typically comprise multiple sub-models, each of which may possess a pre-established weight. A domain shift between the pre-train and current training data is introduced at this point. Second, there will be a domain shift between training and real-world data following the development and implementation of a VQA model. This domain shift is often assessed using cross-dataset validation. Several approaches have examined the domain shift in sub-model pre-training and the acquisition of medical pre-trained models as image encoders among the studied methods [14]. Nevertheless, no research has looked into the possibility of a domain shift in language encoders. Furthermore, little thought or research has been done on the domain shift among medical VQA datasets. There has been enough material for transferring learning experiments like domain adaptation and domain generalization thanks to the increasing amount of medical VQA datasets. Enhancing and quantifying the generalizability of models will be a viable and significant area of study.

3.3.5 Integration with medical workflow

The attempt to include AI decision-supporting technologies into the clinical flow is not new to the community. Effective communication and assistance will be provided by integrating the medical VQA system into clinical practice. Hekler et al. discovered that when it comes to skin cancer classification, the combined use of artificial and human intelligence can produce better outcomes than either system working alone [15].

Tschandl et al. also discovered that doctors receiving AI-based assistance performed better than doctors or AI alone. In the meanwhile, AI-based support benefits clinicians with the least amount of expertise the most [16]. Tschandl et al. also investigated how human-computer collaboration could increase skin cancer recognition accuracy. They discovered that multiclass probabilities performed better in mobile technology than either malignancy probability or content-based image retrival (CBIR) [16]. These results suggest that a variety of aspects, including the cognitive style, cognitive error, personality, experience, and acceptance of AI by doctors, need to be taken into account in order to generate successful outcomes for medical AI-supporting system integration. Furthermore, the majority of AI decision support tools utilized in clinical settings are only effective at responding to pre-formulated queries. On the other hand, medical VQA has the benefit of real-time communication and understanding free-form queries. This benefit can be addressed by streamlining and eliminating unnecessary querying from the procedure. For instance, prior to the QA session, it is possible to always inquire about the existence of abnormalities. Therefore, maintaining an operational question database may be aided by carefully choosing relevant question categories and continuing online education. A medical VQA system's ultimate objective is to provide answers to open-ended inquiries (what, which, e.t.c.). According to the "human-in-the-loop" theory, achieving this goal will add to the complexity and uncontrollable aspects. When the medical VQA system's assessment differs from a clinician's, for instance, a disagreement may need to be settled. As a result, the training corpus might help with negotiation preparation. It will take more work and research in this area to successfully integrate medical VQA into the clinical pipeline. It's also necessary to look into a number of factors, including user satisfaction, collaborative answer accuracy, time spent, and work-flow efficiency.

4 Conclusion

The development of Medical Image Question Answering (MIQA) systems represents a significant advancement in healthcare informatics, with the potential to revolutionize clinical practice and medical research. By leveraging deep learning techniques, particularly convolutional neural networks (CNNs) and natural language processing (NLP) models, MIQA frameworks have made substantial progress in understanding and interpreting medical images based on textual queries. The VQA-RAD dataset serves as a valuable resource for training and evaluating MIQA models tailored to the domain of radiology. It enables the development of algorithms that can automatically answer questions about medical images, leading to improved computer-aided diagnosis systems and decision support tools for radiologists. However, several challenges persist in the field of MIQA, such as domain-specific language understanding, model interpretability, and generalization across different medical specialties and modalities. To further advance the field, future research should focus on increasing the diversity and realism of medical VQA datasets, integrating additional patient-specific information into the inference process, and improving the interpretability and generalizability of MIQA models. By

addressing these challenges, MIQA systems can become more robust, reliable, and clinically relevant, ultimately enhancing patient outcomes and transforming healthcare delivery.

References

[1] Lau, J. J., Gayen, S., Ben Abacha, A., & Demner-Fushman, D. (2018). A dataset of clinically generated visual questions and answers about radiology images. Scientific data, 5(1), 1-10.

[2] Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., & Girshick, R. (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2901-2910).

[3] Hasan, S.A., Ling, Y., Farri, O., Liu, J., Müller, H., Lungren, M.P., 2018. Overview of ImageCLEF 2018 medical domain visual question answering task., in: CLEF (Working Notes).

[4]Sharma, D., Purushotham, S., Reddy, C.K., 2021. Medfusenet: An attention-based multimodal deep learning model for visual question answering in the medical domain. Scientific Reports 11, 1–18.

[5]Yi, Kexin, et al. "Neural-symbolic vqa: Disentangling reasoning from vision and language understanding." Advances in neural information processing systems 31 (2018).

[6]Zhang, X. et al. (2023) PMC-VQA: Visual instruction tuning for medical visual question answering, arXiv.org. Available at: https://arxiv.org/abs/2305.10415 (Accessed: 23 March 2024).

[7]PMC Open Access Subset [Internet]. Bethesda (MD): National Library of Medicine. 2003 - 2024. Available from https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/.

[8]Lau, J.J., Gayen, S., Abacha, A.B., Demner-Fushman, D., 2018. A dataset of clinically generated visual questions and answers about radiology images. Scientific Data 5, 1–10.

[9]Ben Abacha, A., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Müller, H., 2019. VQA-Med: Overview of the medical visual question answering task at imageclef 2019, in: CLEF2019 Working Notes, CEUR-WS.org, Lugano, Switzerland.

[10] He, X., Zhang, Y., Mou, L., Xing, E., Xie, P., 2020. PathVQA: 30000+ questions for medical visual question answering. preprint arXiv:2003.10286

[11]Kovaleva, O., Shivade, C., Kashyap, S., Kanjaria, K., Wu, J., Ballah, D., Coy, A., Karargyris, A., Guo, Y., Beymer, D.B., et al., 2020. Towards visual dialog for radiology, in: Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing, pp. 60–69.

[12]Ben Abacha, A., Datla, V.V., Hasan, S.A., Demner-Fushman, D., Müller, H., 2020. Overview of the VQA-Med task at ImageCLEF 2020: Visual question answering and generation in the medical do- main, in: CLEF 2020 Working Notes, CEUR-WS.org, Thessaloniki, Greece.

[13]Ben Abacha, A., Sarrouti, M., Demner-Fushman, D., Hasan, S.A., Müller, H., 2021. Overview of the VQA-Med task at ImageCLEF 2021: Visual question answering and generation in the medical domain, in: CLEF 2021 Working Notes, CEUR-WS.org, Bucharest, Romania.

[14]Liu, B., Zhan, L.M., Wu, X.M., 2021a. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images, in: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer International Publishing, Cham. pp. 210–220.

[15]Hekler, A., Utikal, J.S., Enk, A.H., Hauschild, A., Weichenthal, M., Maron, R.C., Berking, C., Haferkamp, S., Klode, J., Schadendorf, D., et al., 2019. Superior skin cancer classification by the combination of human and artificial intelligence. European Journal of Cancer 120, 114–121.

[16]Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J., et al., 2020. Human-computer collaboration for skin cancer recognition. Nature Medicine 26, 1229–1234.