Visual Question Answering For Disease Detection Using Stacked LSTM and CNN Architecture

Haider Ali
National University Of Science And Technology (SEECS)
Islamabad, Pakistan
hali.msai22seecs@seecs.edu.pk

***Abstract*—Visual question answering (VQA) has been extensively studied, but its application in medical imaging for disease detection is relatively limited. This research addresses this gap by proposing a novel approach called stacked LSTM-CNN networks (SLCNs) for VQA in medical imaging. The SLCN framework integrates a pre-trained VGG-16 network for image feature extraction, LSTM networks for question understanding, and dense layers with dropout for robust prediction. The proposed methodology is evaluated on the VQA-RAD dataset, achieving promising results. The model demonstrates an accuracy of 85% on the training set, highlighting its ability to capture disease-related features. However, the accuracy on the validation set is 52%, indicating the need for further refinement. This study contributes to the development of effective VQA systems for disease detection in medical imaging.**

***Index Terms*—VQA (visual Question answering), LSTM, CNN**

I. Introduction

Visual question answering (VQA) has gained significant attention in the research field, with numerous studies dedicated to exploring this area [1][3][4]. However, there has been relatively limited research on the application of visual question answering for disease detection. As natural language processing and computer vision continue to advance, there is an opportunity to explore the use of VQA in the context of medical imaging. The integration of VQA in medical imaging research can offer valuable insights and assist doctors in efficiently treating a larger number of patients, potentially saving lives. Despite the potential benefits, there is a notable scarcity of research on utilizing VQA for disease detection.

To address this research gap and evaluate the effectiveness of VQA models in medical imaging, we propose a novel approach called stacked LSTM-CNN networks (SLCNs). This architecture enables multi-step reasoning for image question-answering tasks. In this paper, we present the overall architecture of SLCNs, as illustrated in Figure 1. The SLCN framework comprises three essential components: the 1) Image mode Our image model utilizes a pre-trained VGG-16 network to extract spatial feature maps from MRI images. These maps capture important visual information. Each image region is represented by a feature vector, facilitating accurate disease detection.2) Question model: The question model converts answers into class labels for multi-class classification. Questions are tokenized and transformed into word embeddings using word2vec. LSTM networks capture contextual dependencies within the questions, enhancing understanding. Combined feature vectors from questions and answers enable accurate disease predictions and 3) Concatenation step: Concatenated feature vectors from the image and question models undergo further processing through dense layers with dropout. This ensures robustness and prevents overfitting. A softmax function generates a probability distribution across disease classes, providing reliable disease predictions. Our approach integrates multi-modal information for effective VQA in medical imaging.

II. Structure of the Paper

This research paper follows a structured approach to present the proposed methodology for visual question answering (VQA) in disease detection using stacked LSTM and CNN architecture. The paper is organized as follows:

Related Work: Reviews existing literature on VQA, medical imaging, and disease detection.

Research Gap: Identifies the need for VQA systems tailored for disease detection in medical imaging.

Proposed Methodology: Details the methodology, including data description, preprocessing steps, and the architecture combining LSTM and CNN for disease detection.

RESULTS AND ANALYSIS: Presents evaluation metrics, results, and discussions based on the VQA-RAD dataset.

Conclusion: Summarizes the key findings, contributions, and future directions for research.

References: Lists the cited sources.

III. Related Work

The applications of VQA algorithms are widespread, ranging from aiding visually impaired users to intelligence analysts, automating customer service, building surveillance video automated query systems, and providing medical assistance. Below are some of the literature reviews that have been conducted for this research:

Visual Question Answering (VQA) has been a highly researched area of visual reasoning for the past seven years and has become one of the most popular topics in this field [7]. Over the years, several datasets [6] and methods have been proposed for VQA research. Typically, a VQA dataset includes an image, a corresponding question, and an answer. DAQUAR

[4] was the first dataset to use real photos for visual question answering, while CLEVER [8] presented spatial and relational reasoning problems based on visual features. The VQA dataset [1], which is one of the largest datasets, combines open-ended questions about images with answers, and it utilizes images from the MS-COCO [5] collection.

Remi and colleagues [1] introduced MuRel, a multimodal relational network designed for end-to-end reasoning over actual images. To enhance the visualization features of the network, dense vectors called MuRel cells were used to represent interactions between question and image areas. In contrast, Li et al. [2] used graphs to model both implicit and explicit relationships between items in a picture. They accomplished this using graph attention networks, which encode visual associations based on the query semantics.

Gao and co-authors [3] proposed the QLOB (Question-Led Object Attention) paradigm to improve the performance of visual question answering (VQA). The recommended model followed a three-phase approach, involving the use of an object detection network to extract improved visual features, a question model to obtain sentential semantics, and the QLOB method to select relevant object portions based on the given query. A softmax classifier was then used to predict the final answer, and the proposed model was evaluated on datasets from VQA 1.0, VAQ 2.0, and Toronto COCO-QA.

Sharma and co-author[4] presented a VQA model that incorporates contextual attention and a graph neural network (GNN) to encode visual connections between objects and produce responses. The GNN represents implicit visual associations between objects or areas within an image, while a context-aware attention model is used to highlight the most important visual connection representations. Additionally, Xi et al. employed a multi-objective relation identification approach based on word vector similarity and appearance-based criteria to obtain answers.

The emergence of large-scale datasets has led to a surge of interest in VQA research over the past decade, with the problem remaining an ongoing area of study. Existing work in this field has provided inspiration for our own research. The annual EvalAI VQA competition has showcased some of the most effective VQA models to date, with the 2019 winners [5] utilizing a complex self- and directed attention strategy that incorporated multiple layered modules for co-attention. This study highlights the importance of exploring various attentional mechanisms and developing model structures that incorporate multiple attentional processes. The most commonly used approach for Visual Question Answering (VQA) is the fusion-based method, which combines the image and question into a single representation using global features to predict the correct answer [9]. In order to extract visual information, pre-trained CNN models such as [1], [3], [10], and more recently, pre-trained Faster RCNN [2], are used to extract image features. To retrieve the questions from textual data, LSTM [1]-[3], [5], and GRU are used. Recently, attention-based techniques [2], [5], and have been employed to analyze the semantics of questions based on visual cues, leading to impressive results.

Attention mechanisms, inspired by the human visual system, have been a powerful addition to multi-modal learning tasks, including VQA. Attention helps models give varying degrees of importance to features, making neural network learning more flexible. Visual and textual attention mechanisms can be combined to filter out irrelevant information, with co-attention mechanisms being the most advanced method. These mechanisms can be used to focus on specific words or local image regions, and to gather essential information from the features of different modalities. Several approaches have been proposed, such as dual attention networks, stackable architectures, and modular co-attention networks. These have achieved state-of-the-art results in VQA by modeling intramodal and intermodal interactions. Explicit selection methods have also been used to eliminate noise information from irrelevant question words. Attention mechanisms not only improve the accuracy of VQA models but also enhance their interpretability. [11] [12]

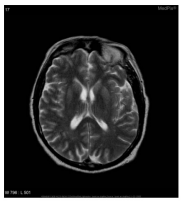
IV. Research Gap

While visual question answering (VQA) has garnered significant attention in various domains, such as scene-related VQA and object detection, there remains a substantial research gap when it comes to applying VQA techniques to medical disease detection. Unlike general VQA tasks, medical VQA involves the computation of specific regions within an image and the recognition of patterns within those regions to accurately answer questions pertaining to medical diagnoses. To the best of my knowledge, there is limited existing research on VQA for medical imaging, specifically focusing on disease detection. This gap in the literature highlights the need for further exploration and development of VQA techniques tailored specifically for medical applications. In this regard, this research aims to fill this gap by proposing a novel approach that combines LSTM (Long Short-Term Memory) and VGG16, to enhance the accuracy and effectiveness of VQA in medical imaging.

V. Proposed Methodology

A. Data

In this research paper, we utilized the publicly available VQA-RAD dataset [13], which was manually curated and consists of a diverse range of clinically generated medical images. The dataset includes 104 head MRI scans, 107 chest x-rays, and 104 abdominal axial CTs, totaling to 315 MRI images in JPEG format. Sample dataset is shown below:



Question: Is the cardiac silhouette enlarged?
Answer: Yes

Question: How would you describe the liver?
Answer: Shrunken and nodular

One of the notable strengths of the VQA-RAD dataset is its meticulous construction, ensuring its quality and relevance for our project. Within this dataset, we encountered a total of 3,515 visual questions, each composed of 5 to 7 words. These questions cover various aspects of the medical images and serve as valuable inputs for our research.

The question-answer pairs in the VQA-RAD dataset are categorized into two types: open-ended questions and close-ended questions. Open-ended questions account for 42% (1,476) of the total pairs, while close-ended questions, such as "Yes" or "No," make up the remaining 58% (2,038). Among the close-ended answers, the most frequently occurring answer was "No," which was observed 820 times.

To facilitate training and evaluation, the dataset is divided into a training set and a validation set, allowing us to effectively assess the performance of our proposed methods. The distribution of the dataset between the two sets is presented below.

Table 1. Dataset Distribution

	Image	Questions	Answers
Train Split	220	3,000	3,000
Validation Split	95	515	515

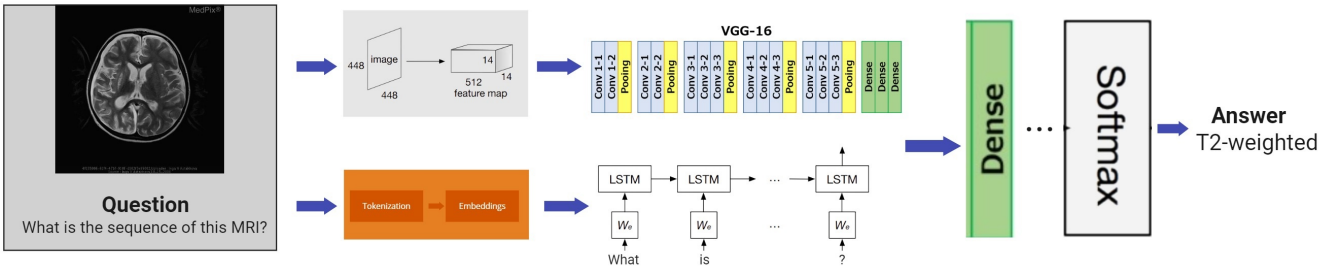
B. Method

In this section, we present the methodology employed for VQA in disease detection. Our approach encompasses several crucial steps, starting with dataset preprocessing. The dataset utilized in this study comprises images along with corresponding questions and answers. Prior to preprocessing, we carefully split the dataset into a training split consisting of 3,000 examples and a validation split containing 515 examples. Each example within the dataset constitutes a triplet comprising an image, a question, and an answer. Notably, the answers are classified into 486 distinct possible answer classes. To facilitate effective feature extraction, all images are resized to dimensions of 448x448x3. Subsequently, these images undergo processing using the VGG16 architecture, renowned for its capability to extract meaningful image features. By leveraging the power of VGG16, we aim to capture essential visual information that can aid in disease detection. Furthermore, we address the textual component of the VQA system by standardizing the length of questions. To achieve this, all questions are padded to ensure a uniform length of 21 words. This preprocessing step ensures that the input data is consistent and compatible for subsequent processing. Now, let us delve into the details of each step involved in our proposed methodology.

1) Image Model: One crucial component of our proposed model is the image model, which plays a vital role in obtaining essential features from MRI images. We leverage the power of a pre-trained VGG-16 network to extract spatial feature maps from the images. Specifically, we extract these feature maps from the last convolutional block of the VGG-16 network, capturing high-level information and intricate details. Each image is then represented by a 3D volume with dimensions of 512x14x14, signifying the presence of 512 spatial feature maps, each of size 14x14. This representation enables us to effectively capture the diverse visual information present in the MRI images, facilitating accurate disease detection. To further process and refine the image representation, we reshape the image into a dimension of 196x512. This reshaping results in each region being represented by a feature vector of dimensions 1x1x512. Notably, each image is now divided into 196 distinct regions, with each region carrying crucial information about the underlying pathology. By extracting feature vectors for the 196 regions of each image, our model becomes capable of making accurate predictions regarding the presence of diseases. This comprehensive representation of the image enables our model to effectively capture localized abnormalities and assess their significance in the overall diagnostic process. Through the utilization of the image model, we aim to enhance the accuracy and reliability of disease detection using MRI images. By capturing and leveraging the distinct features of each region, our model strives to provide accurate and interpretable predictions, contributing to the field of medical imaging analysis.

2) Question Model: Another essential component of our model is the question model, responsible for extracting feature vectors from the questions and answers. To facilitate this process, we employ several steps to effectively represent and process the textual input. Firstly, we convert the answers into class labels, thereby transforming our problem into a multi-class classification task. Each answer is assigned a unique number ranging from 0 to 485, representing the 485 distinct answer classes. This conversion enables us to frame our problem within the context of classification, facilitating accurate prediction of the answers. Moving on to the questions, we tokenize each question to break it down into individual words or tokens. Subsequently, we utilize GoogleNews vectors word2vec to convert these tokenized words into embeddings. By leveraging word embeddings, we can effectively capture

Fig. 1. The VQA Model for Medical Imaging integrates image and question inputs. Images are preprocessed into 14x14x512 feature images via the VGG-16 architecture, renowned for effective feature extraction. Questions undergo tokenization and embedding, followed by processing through an LSTM network, capturing sequential and contextual dependencies. LSTM and VGG-16 outputs are concatenated and fed into dense networks, facilitating complex representation learning and capturing image-question interactions. The model predicts answers using a softmax function, offering accurate and interpretable responses in medical-related inquiries.



the semantic meaning and contextual information embedded within the questions. Upon obtaining the question vector and answer label, we can create feature vectors for both questions and answers. These feature vectors encapsulate important information derived from the input text, enabling our model to capture the essential features necessary for accurate disease detection. To further capture the contextual dependencies within the questions, we employ Long Short-Term Memory (LSTM) networks. By passing the question through the LSTM layer, our model can effectively capture the sequential information and contextual dependencies embedded within the text. This step plays a crucial role in ensuring that our model comprehends the nuanced meaning and dependencies within the questions, enhancing the accuracy of disease prediction. By combining the extracted feature vectors from questions and answers and leveraging the power of LSTM for contextual understanding, our model becomes equipped to make accurate predictions in disease detection. This comprehensive approach, encompassing both textual and visual information, contributes to the overall effectiveness and reliability of our VQA system for disease diagnosis.

3) Concatenating Image and Question Models for Prediction: Having obtained the feature vectors from the question, answer, and image components, our next step involves integrating these vectors to make accurate disease predictions. To accomplish this, we employ a concatenation strategy, combining the VGG16 and LSTM architectures using a fully connected dense layer. The feature vectors derived from the VGG16 and LSTM networks are concatenated to create a comprehensive representation of the combined information captured from the image, question, and answer. This concatenation enables our model to effectively leverage both visual and textual cues, maximizing the richness of information for disease detection. Following the concatenation step, the concatenated feature vector is passed through additional dense layers. These dense layers further process and refine the combined information, enhancing the model's ability to capture intricate patterns and relationships. To prevent overfitting, we incorporate a dropout technique with a rate of 0.5, ensuring robustness and generalizability of the model. Finally, we utilize a softmax function for prediction. The softmax function normalizes the output of the model, generating a probability distribution across the different disease classes. This probability distribution allows for accurate and interpretable predictions, providing insights into the likelihood of specific diseases based on the input image, question, and answer.

By incorporating the concatenated feature vectors, employing dense layers with dropout, and utilizing softmax for prediction, our model aims to achieve accurate and reliable disease detection. This comprehensive approach facilitates the integration of multi-modal information and effectively addresses potential overfitting concerns, resulting in a robust VQA system for medical imaging analysis.

VI. Results and Analysis

Our VQA model for disease detection underwent training for 50 epochs using the Adam optimizer with a learning rate of 0.01. To ensure efficient training, a batch size of 32 was employed, with all these parameters being carefully selected after extensive experimentation. The loss function employed for training was categorical crossentropy. Impressively, upon completion of the training process, our model achieved an accuracy of 85% on the training set, signifying its capacity to effectively learn and capture disease-related features. The training loss graph and accuracy graph are presented below:

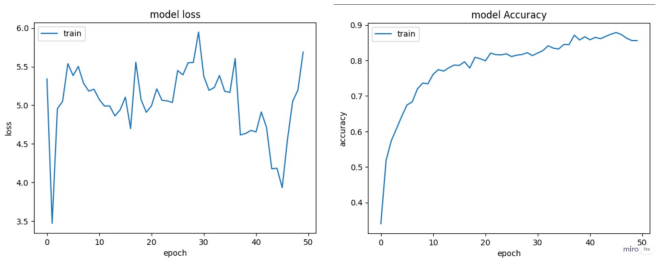
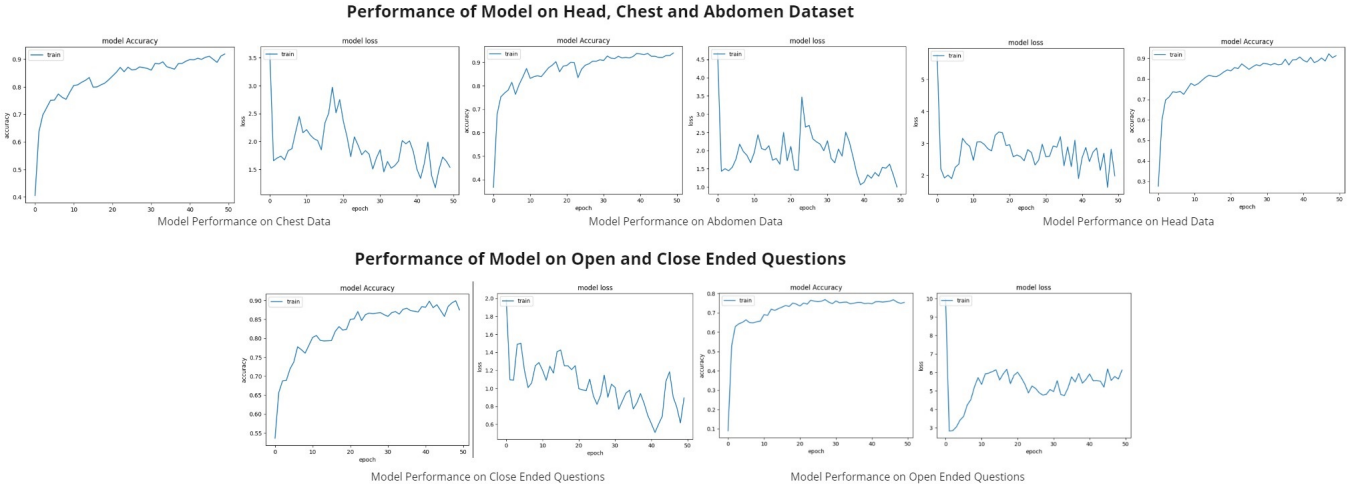


Fig 3. Model accuracy and loss graph

However, during the evaluation on the validation set, the model exhibited a lower accuracy of 52%. One of the primary factors contributing to this diminished accuracy could be the size of the dataset. A larger dataset would provide the model with a wider variety of instances to learn from,

Fig. 2. Depiction of performance across various dataset categories, including head, chest, and abdomen, as well as open-ended and closed-ended questions.



potentially leading to enhanced accuracy. Further analysis is required to identify additional reasons and potential remedies for the observed performance. To conduct a comprehensive assessment of our model, we thoroughly analyzed its performance across various data classes, including chest, head, and abdomen. Evaluating its proficiency in comprehending and answering questions, we specifically examined its performance on open-ended and closed-ended questions. The results of this analysis, including performance metrics and loss values, are summarized in Table 2. Additionally, Figure 2 illustrates the accuracy and loss graph, providing insights into the model's performance across all categories. Interestingly, when considering the performance across different classes, the abdomen class exhibited the highest accuracy, reaching a rate of 91%. Moreover, our model displayed better performance on closed-ended questions compared to open-ended questions. This observation suggests that the model might encounter challenges when confronted with the intricacies and variability inherent in open-ended questions.

	Accuracy	Loss
CHEST	0.9191	1.5360
HEAD	0.9130	1.9764
ABDOMEN	**0.9399**	**1.0040**

VQA-RAD on different ORGANS

	Accuracy	Loss
OPEN	0.7534	6.1258
CLOSED	**0.9130**	**1.9764**

VQA-RAD on CLOSED and OPEN answer

Table 2. Accuracy and Loss values

To enhance the model's performance on open-ended questions, several factors can be taken into consideration. Firstly, augmenting the training dataset by including a larger and more diverse set of open-ended questions would allow the model to gain exposure to a wider range of question types and variations. Additionally, exploring advanced natural language processing techniques, such as incorporating pre-trained language models or leveraging attention mechanisms, holds promise for improving the model's handling of open-ended questions.

Conclusion and Future Work

This paper proposes a novel approach called stacked LSTM-CNN networks (SLCNs) for visual question answering (VQA) in disease detection using medical imaging. The SLCN framework integrates a pre-trained VGG-16 network for image feature extraction, LSTM networks for question understanding, and dense layers with dropout for robust prediction. The proposed methodology is evaluated on the VQA-RAD dataset, achieving promising results with an accuracy of 85% on the training set. However, further refinement is needed as the accuracy on the validation set is 52%. The research addresses the research gap in applying VQA techniques to medical disease detection and contributes to the development of effective VQA systems in the medical imaging domain. Future research can focus on improving the model's performance on the validation set and exploring other architectures and datasets for VQA in disease detection.

In our future work, we can enhance the performance of our model significantly through the incorporation of attention mechanisms and expanding the size of our dataset. These two contributions have the potential to play a pivotal role in elevating the performance levels of our model.

References

[1] Cadene R, Ben-Younes H, Cord M and Thome N (2019) Murel: Multimodal relational reasoning for visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1989–1998

[2] Li L, Gan Z, Cheng Y, and Liu J (2019) Relation-aware graph attention network for visual question answering, In Proceedings of the IEEE International Conference on Computer Vision, pp. 10313–10322. 2019

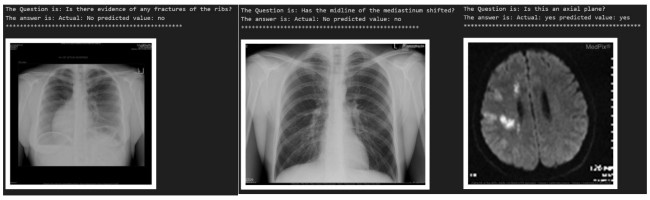
[3] Gao L, Cao L, Xu X, Shao J, Song J (2020) Question-Led object attention for visual question answering. Neurocomputing 391:227–233

[4] Sharma H, Jalal AS (2021) Visual question answering model based on graph neural network and contextual attention. Image and Vision Comput 110:104165
[5] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.
[6] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, et al., "Vqa: Visual question answering", Proceedings of the IEEE international conference on computer vision, pp. 2425-2433, 2015.
[7] R. Cadene, H. Ben-Younes, M. Cord and N. Thome, "Murel: Multimodal relational reasoning for visual question answering", Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1989-1998, 2019
[8] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning", Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2901-2910, 2017.
[9] J. Lu, J. Yang, D. Batra and D. Parikh, "Hierarchical question-image co-attention for visual question answering", Advances in neural information processing systems, vol. 29, 2016.
[10] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering", Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6904-6913, 2017.
[11] Nam H, Ha J-W, Kim J (2017) Dual attention networks for multimodal reasoning and matching. CVPR:2156–2164
[12] Nguyen D-K, Okatani T (2018) Improved fusion of visual and language representations by dense symmetric Co-Attention for visual question answering. CVPR:6087–6096
[13] Lau, J., Gayen, S., Ben Abacha, A. et al. A dataset of clinically generated visual questions and answers about radiology images. Sci Data 5, 180251 (2018).

VII. Appendix

Fig 4. Prediction on sample data

CORRECT RESULTS



INCORRECT RESULTS