

MedVQA: A Visual Diagnosis Support System Using CNN and GloVe

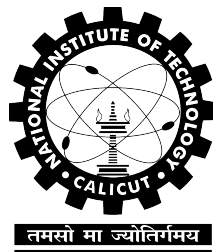
CS4099D Project
End Semester Report

Submitted by

Gagan Lal (B190480CS)
Geethika S (B190449CS)

Under the Guidance of

Dr.Saidalavi Kalady
Associate Professor

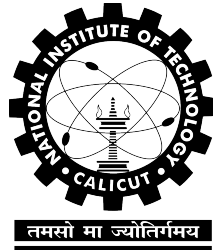


Department of Computer Science and Engineering
National Institute of Technology Calicut
Calicut, Kerala, India - 673 601

May 2023

NATIONAL INSTITUTE OF TECHNOLOGY CALICUT
KERALA, INDIA - 673 601

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING



CERTIFICATE

Certified that this is a bonafide report of the project work titled

**MEDVQA: A VISUAL DIAGNOSIS SUPPORT SYSTEM
USING CNN AND GLOVE**

done by

**Gagan Lal
Geethika S**

*of Eighth Semester B. Tech, during the Winter Semester 2022-'23, in
partial fulfillment of the requirements for the award of the degree of
Bachelor of Technology in Computer Science and Engineering of the
National Institute of Technology, Calicut.*

08-05-2023

Date

(Dr. Saidalavi Kalady)

(Associate Professor)

Project Guide

DECLARATION

I hereby declare that the project titled, **MedVQA: A Visual Diagnosis Support System Using CNN and GloVe**, is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material that has been accepted for the award of any other degree or diploma of the university or any other institute of higher learning, except where due acknowledgment and reference has been made in the text.

Place : NIT Calicut
Date : 08-05-2023

Name : Gagan Lal
Roll. No. : B190480CS
Name : Geethika S
Roll. No. : B190449CS

Abstract

A Visual Question Answering system is expected to answer an image-related question through the understanding of the provided image and question. VQA systems can be used in the medical domain. Medical Visual Question Answering (MedVQA) is an emerging field that combines medical imaging with natural language processing to provide answers to questions related to medical images. Deep learning approaches, as well as image and question classification techniques, are used in current Med-VQA frameworks. However, the volume and variety of questions and images in the medical domain make Med-VQA a difficult task. In this project, we aim to develop a MedVQA system that can understand and answer questions based on medical images, assisting healthcare professionals in decision-making and diagnosis. When given a medical image and a clinically relevant question, the system is supposed to predict a convincing answer. The project contributes to the growing field of MedVQA and demonstrates the feasibility of combining medical imaging and natural language processing for medical question answering.

ACKNOWLEDGEMENT

We would like to express our sincere and heartfelt gratitude to our guide and mentor Dr. Saidalavi Kalady and Lubna A, who have guided us throughout the course of the final year project. Without their active guidance, help, cooperation, and encouragement, we would not have made headway in the project. We would like to thank our parents and the faculty members for motivating us and being supportive throughout our work. We also take this opportunity to thank our friends who have cooperated with us throughout the course of the project.

Contents

1	Introduction	3
2	Problem Statement	6
3	Literature Survey	7
3.1	Methods for VQA	7
3.2	Datasets	8
3.3	Medical Visual Question Answering	9
4	Dataset and Design Overview	17
4.1	Dataset Used	17
4.2	Design	20
4.2.1	Question Feature Extraction	21
4.2.2	Image Feature Extraction	22
5	Implementation Of MVQA	24
5.1	Dataset Preparation	24
5.2	Classification	24
5.3	Question Feature Extraction	25
5.4	Image Feature Extraction	26
5.5	Feature Fusion	28
6	Evaluation Of MVQA	29
7	Conclusion	32
	References	34

List of Figures

1.1	An example of Medical VQA where an image and question are given as input and the system predicts the correct answer. . .	5
4.1	Proposed Design	23
5.1	Open - Close Classifier	25
6.1	Sample output	30
6.3	Few examples of manual testing	31

Chapter 1

Introduction

A VQA system is expected to answer an image-related question through the understanding of the provided image and question. For example, given an image of a bird, a person can ask "What is shown in the figure?" or "What is the color of the bird?" The VQA system is expected to provide the correct answer. VQA requires a deep comprehension of both images and textual questions. It is a purely supervised learning setting. VQA systems combine NLP, which provides an understanding of the question and the ability to produce an answer, with CV techniques, which provide an understanding of the content of the image. VQA shows great potential in interpreting automated medical imagery and machine-supported diagnoses.

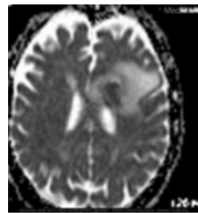
The field of medical imaging plays a crucial role in modern healthcare, enabling clinicians to visualize internal structures of the human body for diagnosis, treatment planning, and monitoring of diseases. In the field of medical imaging, vast amounts of visual data are generated daily, ranging from X-rays and CT scans to histopathological images and endoscopic videos. These images contain valuable diagnostic information that can aid clinicians in accurate disease detection, treatment planning, and monitoring. However, extracting meaningful insights from these images often requires domain

expertise and extensive manual analysis, which can be time-consuming and prone to human errors.

With the advancements in deep learning and natural language processing techniques, there has been a growing interest in developing intelligent systems that can understand and answer questions based on medical images. This emerging field, known as Medical Visual Question Answering (MedVQA), holds great potential to assist healthcare professionals in decision-making, enhance patient care, and facilitate medical education. In a Med VQA system, both the image and question given to the system are from the medical domain. The potential applications of MedVQA systems are diverse and impactful, ranging from assisting radiologists in interpreting medical images to empowering patients to gain better insights into their health conditions.

Since 2016, a general VQA challenge has been issued every year to answer questions of various types. In 2018, ImageCLEF launched the VQA-Med challenge which is specific to the medical field and is held annually. Med VQA is technically more challenging than general VQA because of the following factors: (i) creating a large-scale Med VQA dataset is challenging because expert annotation is expensive for its high requirement of professional knowledge and QA pairs cannot be synthetically generated directly from images. (ii) Answering questions according to a medical image also demands a specific design of the VQA model. (iii) A question can be very professional, which requires the model to be trained with medical knowledge base rather than a general language.

A complete Med VQA system can directly review patients' images and answer any kind of questions. It will ease the shortage of medical resources and provide convenience for patients as well as medical professionals. The most recent model proposed in 2022 [10], has an accuracy of 82.4 % for SLAKE dataset.



Q: What is abnormal in the MRI?

A: A brain abscess, central nervous system

Figure 1.1: An example of Medical VQA where an image and question are given as input and the system predicts the correct answer.

Chapter 2

Problem Statement

Medical Visual Question Answering (MVQA) addresses the challenge of effectively utilizing the vast amount of visual medical data, such as medical images and videos, to provide accurate and timely answers to specific clinical queries. The problem lies in the complex nature of medical images and the difficulty in extracting meaningful information from them without extensive manual analysis by domain experts. This manual analysis is time-consuming, subjective, and prone to human errors, hindering efficient diagnosis and treatment planning. This project aims to develop an intelligent MVQA system that can understand, interpret, and answer specific clinical questions based on medical images. The goals of the project include overcoming the limitations of manual analysis, handling a wide range of questions and imaging modalities, and integrating visual and textual information effectively.

Chapter 3

Literature Survey

3.1 Methods for VQA

Wu Q et al, 2017 [1] have focused on studying the various VQA methods which are divided into four categories based on the nature of their main contribution and datasets available for training and evaluating VQA systems. The methods mentioned in the paper are (i) Joint embedding approaches which allow one to learn representations in a common feature space. Joint embedding model was proposed by H.gao et al. for implementing VQA. It focuses on the global features of an image. Image and textual questions are taken as input and features of both are to be extracted through different deep learning and NLP techniques. After getting these features, both the feature vectors are jointly embedded into a common feature space, and then these combined feature vectors are fed into a classifier. The classifier then predicts the answer to the question. (ii) Attention mechanism focuses on question specific region of an image rather than all the global features of an image. They use local image features and allow the model to assign different importance to features from different regions. (iii) Compositional models are useful when questions require multi-step reasoning to answer properly. They

facilitate transfer learning as the same module can be used and trained within different overall architectures and tasks. (iv) Models using external knowledge bases are useful when additional background knowledge or common sense is required to answer properly.

3.2 Datasets

In Wu Q et al, 2017 [1], the authors did a detailed survey about the datasets available for training and evaluating VQA systems. The general datasets mentioned in the paper are (i) DAQUAR (Dataset for question answering real-world images) which was the first VQA dataset to be designed as a benchmark. The images in DAQUAR are split into 795 training and 654 test images. (ii) COCO-QA dataset includes 123,287 images (72,783 for training and 38,948 for testing). This dataset uses images from the Microsoft Common Objects in Context data. (iii) Freestyle multilingual image question answering dataset uses 123,287 images. (iv) VQA-real is one of the most widely used datasets. VQA-real comprises of 123,287 training and 81,434 test images. (v) Balanced dataset contains 10,295 and 5328 pairs of complementary scenes for the training and test set respectively. (vi) KB-VQA dataset contains questions requiring topic-specific knowledge that is present in DBpedia. (vii) FVQA dataset contains only questions that involve external information. It was designed to include additional annotations.

In Lin Z et al, 2021 [2], the authors conducted a detailed study of medical VQA datasets. There are 8 public-available medical VQA datasets up to date: (i) VQA-MED-2018 is a dataset that was proposed in the ImageCLEF 2018. In this dataset, the QA pairs were generated from captions by a semi-automatic approach. (ii) VQA-RAD is a radiology-specific dataset proposed in 2018. (iii) VQA-MED-2019 was published during the ImageCLEF

2019 challenge. This dataset addressed the four most frequent question categories: plane, modality, organ system, and abnormality. (iv) RadVisDial is the dataset for visual dialog in radiology. It consists of two datasets: a silver standard and a gold standard dataset. (v) PathVQA explores VQA for pathology. The questions are designed according to the pathologist certification of the American Board of Pathology. (vi) VQA-MED-2020 was published in ImageCLEF 2020 challenge. The images are selected with the limitation that the diagnosis was made according to image content. The questions are specifically addressing on abnormality. (vii) SLAKE has both semantic labels and a structural medical knowledge base. (viii) VQA-MED-2021 is published in ImageCLEF 2021 challenge. The training set is the same as those in VQA-MED-2020 but the test is new.

3.3 Medical Visual Question Answering

Teney D et al, 2018 [3] proposes using a meta-learning approach to overcome the scalability issues faced when VQA is trained by using a large training set of example questions, images, and answers. Meta-learning approach implies that the model learns to learn i.e. it learns to use a set of examples provided at test time to answer the given question. The model is initially trained on a small set of questions/answers and is provided with a support set of additional examples at test time. The model proposed in the paper by the authors is a deep neural network that takes advantage of the meta-learning scenario. The approach mentioned is to provide the model with supervised data at test time. The conclusion drawn from this experiment was that even though the baseline is most effective with frequent answers, the proposed model fares better in the long tail of rare answers. The proposed model had better sample efficiency and a unique capability to learn to produce a novel answer. This model improved the practicality and scalability of the system.

Zhang A et al, 2022 [4] mainly focuses on type-aware medical visual question answering. Since medical images may restrict to a specific part of the human body, identifying the type of the image helps to successfully exploit the characteristics of the image. The authors proposed an image feature extraction module that extracts type points. The textual features are joined with type point embeddings. This improved the ability of semantic alignment between modalities. Further, it enhanced the applicability of the fusion method for Med VQA. The model achieves state-of-art with the VQA-RAD dataset.

Gasmi K et al, 2022 [5] proposes an optimal deep neural network-based model for answering visual medical questions. The medical questions were classified based on a BERT model. The authors used EfficientNet as a deep learning model to extract visual features. These features were then combined using an attention model. Adaptive generic algorithm model was used to determine the optimal learning parameters. This model performed better than the runs of ImageCLEF 2019 participants.

Gasmi K, 2022 [6] proposed a hybrid deep learning model for answering visual medical questions. The medical questions were classified based on the BERT model. The features of the medical image were extracted by a hybrid deep-learning model of VGG and ResNet. The text features were extracted using a Bi-LSTM mode. The authors combined these features extracted on a classifier based on the softmax layer and got the most accurate answer. The dataset used for this model was ImageCLEF2019.

Liu S et al, 2022 [7] proposed a bi-branch model for medical VQA. The first branch transformer structure is the main framework of parallel structure mode. A parallel network was adopted to extract the image features. An

improved CNN model was used to extract spatial features of the medical 4 images. Then RNN model was used to extract the sequence features of the medical images. The text features are also extracted. The image features are then embedded into the front part of the question (text) features. These two features are integrated into a feature matrix and then input into the stacked four-layer transformer structure. As a result, the model can learn about the dependency between image features and question features and capture the internal structure of the input vector. In the second branch, the answers of the training set are used as labels of the corresponding images. This model achieves state-of-art-result with 3 datasets – ImageCLEF2018, ImageCLEF2019, VQARAD. The main metric score exceeds the best results so far by 0.2%, 1.4%, and 1.

Eslami S et al, 2021 [8] assess the efficiency of Contrastive Language Image Pre-training (CLIP) for the MedVQA task. CLIP is a neural network trained on several image and text pairs. Without specifically optimizing for the task, it can be told in natural language to guess the most pertinent text excerpt when given an image. To predict the right pairings of a batch of (image, text) training examples, CLIP jointly trains an image encoder and a text encoder. In the paper, to fine-tune CLIP for the medical domain, the training and validation data splits from the original paper were employed. The refined model was referred to be PubMedCLIP. The authors examine the impact of employing PubMedCLIP as a pre-trained visual encoder in MedVQA models. They took into account MEVF and QCR, two well-known MedVQA techniques that use MEVF as their visual encoders, for the investigation. The MEVF was modified by substituting its pre-trained MAML module with PubMedCLIP. The experiments were conducted on two benchmark MedVQA datasets namely VQA-RAD and SLAKE. The results show that PubMedCLIP performs up to 3% better than the previously employed pre-trained visual encoders in MedVQA. The accuracy scores for VQA-RAD

and SLAKE datasets were 72.1% and 80.1% respectively.

Zhu H et al, 2022 [9] focuses on improving the interactions of particular characteristics from relevant radiological images and questions. The authors explore the use of the Corresponding Feature Fusion (CFF) approach, utilizing a CNN-based type classifier to classify multimodal inputs and subsequently perform feature fusion for the corresponding image-question pairs. They designed a semantic attention (SA) module for the extraction of textual features. As a result, the model is better prepared to consciously concentrate on the important words in different questions while paying less attention to irrelevant data. Since the radiology images mainly focus on three categories of human body regions: abdomen, brain, and chest, the authors utilize a type classifier to classify each pair of multimodal inputs (radiology images and clinical questions) into given categories. Word Embedding is used in conjunction with LSTM to extract question features and then passed through the SA module to determine the attention weight. Both features are then combined and sent through the VQA classifier. The VQA-RAD and filtered SLAKE were used to validate the model. The accuracy scores for VQA-RAD and SLAKE datasets were 75.4% and 82.4%.

Haridas H.T et al, 2022 [10] focuses on MED-GPVS, a tailored deep learning-based GPVS (General Purpose Vision System) on biomedical images that can perform a variety of vision tasks on medical images, such as object detection and visual question answering to support precision medicine/e-health services. When given an image and some natural language text as inputs, the MED-GPVS will produce bounding boxes, confidence scores, and a caption. The authors develop a customized deep learning-based model that accepts an image as input and generates a list of pre-defined output classes. In hybrid customization of the DETR and ViLBERT architectures, the visual and text encoders are combined. In order to detect objects, the vision en-

coder uses transformer encoder-decoder architecture with a CNN backbone for feature extraction. The BERT model and the co-attention module of ViLBERT are combined to provide the outputs of the text encoder model. A ResNet-50 framework is used to extract image features from the input image before passing it on to the DETR transformer encoder architecture. A decoder architecture with a predetermined set of object queries receives the output from this. The input text is tokenized and converted to a sub-token token representation using a BERT model, which is then transferred to the ViLBERT co-attention module for cross-contextualization. The model reached an accuracy of 82.40%.

Table 1 : Tabular Comparison Of Literature Survey

Paper	Method Used	Pros	Cons
Teney D et al, 2018	A deep neural network that takes advantage of the meta-learning scenario	The model had better sample efficiency and a unique capability to learn to produce a novel answer	Handling the memory of dynamic weights was not accurate
Zhang A et al, 2022	TI, TQ (Type Image, Type Question) modules exploit characteristics of input data	Achieves state-of-art with VQA RAD, Very high accuracy	Restricted to a specific class, not a complete solution for VQA
Gasmi K et al, 2022	BERT model on Question and feature extraction using a Bi-LSTM mode Efficient-Net as a deep learning model to extract visual features	Performed better than runs of ImageCLEF 2019 participants, Very High accuracy rate.	Information provided in the questions and the corresponding images are not always sufficient to predict the right answer, and answering the questions often requires external knowledge resource
Gasmi K, 2022	The hybrid deep learning model of VGG and ResNet on ImageBERT model on Question and feature extraction using a Bi-LSTM mode	On using various optimization algorithms on the Image-CLEF2019 dataset – Adam and SGD performed better	A better question classification system is needed. Abnormality question answering is poo

Table 1 : Tabular Comparison Of Literature Survey

Paper	Method Used	Pros	Cons
Eslami S et al, 2021	PubMedCLIP as a pre-trained visual encoder for medical images	PubMedCLIP outperforms previously used pre-trained visual encoders in MedVQA by 3%	The model fails to yield the correct answer to some of the questions when images from the VQA-RAD dataset are provided. The model has trouble understanding the semantics of the question.
Zhu H et al, 2022	Corresponding Feature Fusion (CFF) approach, utilizing a CNN-based type classifier to classify multimodal inputs and subsequently perform feature fusion for the corresponding image-question pairs.	Model is better prepared to consciously concentrate on the important words in different questions while paying less attention to irrelevant data	The questions that can be answered by the model are only intuitive questions raised according to the content of clinical images.
Haridas H.T et al, 2022	A customized deep learning model - MED-GPVS was demonstrated to execute two tasks simultaneously: object detection and visual question answering	Performs reasonably well on the object identification task and substantially well on the visual question-answering task.	The model's ability to forecast a wide range of diseases is comparatively low. The model has complexity issues.

The literature survey for MVQA revealed that the use of deep learning techniques, such as CNNs, and LSTMs has significantly improved the performance of MVQA models. Dataset preparation is also an important aspect of MVQA. Some of the studies use less complex techniques to simulate VQA frameworks. Most of the MVQA model's average accuracy ranges from 0.6-0.7.

The motivation of developing an MVQA system is to use machine learning and artificial intelligence to help clinical decision-making and solve problems related to the increasingly complex nature of medical data. With the increasing availability of medical images, clinicians now have an overwhelming amount of data to analyze and evaluate. By giving doctors an effective tool to analyze and interpret medical pictures, MVQA systems can help bridge this gap. The lack of medical specialists in particular areas or specializations can also be addressed with the use of MVQA systems. An efficient MVQA system can provide non-expert clinicians and even patients with access to insightful medical suggestions based on their medical images and associated inquiries.

Chapter 4

Dataset and Design Overview

4.1 Dataset Used

We use Semantically-Labelled Knowledge Enhanced (SLAKE) Dataset [11] for training our VQA models. The dataset consists of 642 images in total comprising 12 diseases and 39 organs. It has 2,094 testing and 9,835 training examples, all of which are question-answer pairs. SLAKE dataset is a relevant repository with extensive labels, manually annotated by medical experts. Total number of unique questions in training set is 1189 and total number of unique questions in the test set is 605. There were a total of 484 unique training set answers and 270 unique test set answers. The dataset includes questions in both English and Chinese languages and is divided into open-ended and closed-ended categories. We are only taking into consideration questions in English for our project. Questions were of two base types - vqa and kvqa.

The starting words of each question present in the dataset are also analyzed. Questions are either of the form:

- 1) Starting with what/where/how/why/when/which
- 2) Starting with is/does/are/in/do/can

Table 2 : Question Types		
Question Type	Training	Test
Open-ended	2976	645
Close-ended	1943	416

Table 3 : Question Base Types		
Base Type	Training	Test
VQA	4337	913
KVQA	582	148

Questions are divided into 10 types based on their content. They are- Color, Abnormality, Modality, KG (Knowledge Base), Size, Organ, Plane, Shape, Position, and Quantity. Each type represents a specific aspect of the question that pertains to different attributes or characteristics of medical images. By categorizing questions into these 10 types, the SLAKE dataset provides a comprehensive set of query variations that cover different aspects of medical image interpretation. They cover a wide range of content aspects in medical image interpretation, including visual features, anatomical structures, abnormalities, imaging modalities, size, shape, position, and quantitative information.

Table 4 : Question Content Types		
Content Type	Training	Test
Organ	1,269	253
Position	847	186
KG	582	148
Abnormality	716	150
Modality	534	108
Plane	249	58
Quantity	220	52
Color	137	34
Size	323	65
Shape	42	7

The SLAKE dataset consists of medical images from three different modalities: X-Ray, CT (Computed Tomography), and MRI (Magnetic Resonance Imaging). Each modality provides unique information and captures different aspects of the human body.

Table 5 : Image Modality Types		
Modality Type	Training	Test
X-Ray	1,423	361
MRI	1,129	228
CT	2,367	472

In the SLAKE dataset, the medical images are categorized into 10 specific areas or regions. These regions represent different anatomical locations or parts of the human body. They are - Chest heart, Abdomen, Chest lung,

Lung, Neck, Chest mediastinal, Pelvic Cavity, Brain Tissue, Brain Face, and Brain.

Table 6 : Image Locations		
Location	Training	Test
Lung	1,710	828
Abdomen	1,506	607
Brain Tissue	691	166
Brain	272	42
Pelvic Cavity	222	47
Chest Lung	151	21
Neck	129	35
Brain Face	140	29
Chest Heart	86	0
Chest mediastinal	12	10

4.2 Design

The proposed design for the Medical Visual Question Answering (MVQA) system incorporates feature fusion techniques to enhance the performance and accuracy of the question-answering process. Feature fusion aims to combine and integrate different types of features extracted from medical images and textual questions to generate more informative and comprehensive answers.

The questions in the SLAKE dataset are divided into ten types based on their content. They are - Color, Abnormality, Modality, KG, Size, Organ, Plane, Shape, Position, and Quantity. These questions are either close-ended

or open-ended. Initially, the questions are classified into 10 categories by using a classifier. An open-close classifier is also used for categorising questions based on their type (open-ended or close-ended).

4.2.1 Question Feature Extraction

Question feature extraction plays a crucial role in understanding the textual input and generating informative answers. Two key components used in this process are GloVe (Global Vectors for Word Representation) and LSTM (Long Short-Term Memory).

GloVe is a word embedding technique that represents words as dense vectors in a continuous vector space. It captures the semantic relationships between words based on their co-occurrence statistics in a large corpus of text. GloVe provides a pre-trained word embedding model where each word is represented by a fixed-length vector. The vectors capture the meaning and context of words, allowing for meaningful comparisons and semantic similarity calculations. By utilizing GloVe embeddings, the textual questions in the MVQA system can be represented as numerical vectors, facilitating further processing and analysis.

LSTM is a type of recurrent neural network (RNN) that is effective in capturing long-range dependencies and sequential patterns in textual data. It addresses the vanishing gradient problem of traditional RNNs by introducing specialized memory cells. LSTM cells can selectively remember and forget information over time, enabling the model to retain relevant context and ignore irrelevant information. In the question feature extraction process, LSTM is employed to process the sequence of word embeddings obtained from GloVe. LSTM models can effectively capture the dependencies and relationships between words in the question, enabling a better understanding of the question's semantics.

4.2.2 Image Feature Extraction

A Convolutional Neural Network (CNN) is used to extract features from medical images to answer corresponding questions. A CNN is an algorithm that can be used to extract specific features from an image and learn the necessary weights to help distinguish images. It can be used to capture the relevant features in an image, potentially providing much better accuracy in answering the VQA model.

The CNN architecture for image feature extraction typically consists of several convolutional layers, each followed by a pooling layer to reduce the dimensionality of the feature maps. The input image is passed through a CNN architecture which consists of three convolutional layers. Each convolutional layer is followed by a max pooling layer to reduce the spatial dimensions of the previous layer's output. The pooling operation helps to extract the most important features while reducing the size of the output, thereby reducing the computational cost.

After the three convolutional and pooling layers, the output is flattened to a one-dimensional vector and passed through a fully connected dense layer with 256 neurons. This layer acts as a bottleneck between the convolutional layers and the output layer.

The output of the dense layer is then used to compute the final output of the CNN, which can be used as an input feature vector for MVQA.

The image features and question embeddings obtained are then concatenated into a single vector representation. The fused feature representation is passed through a fully connected layer followed by an activation function to generate the answer.

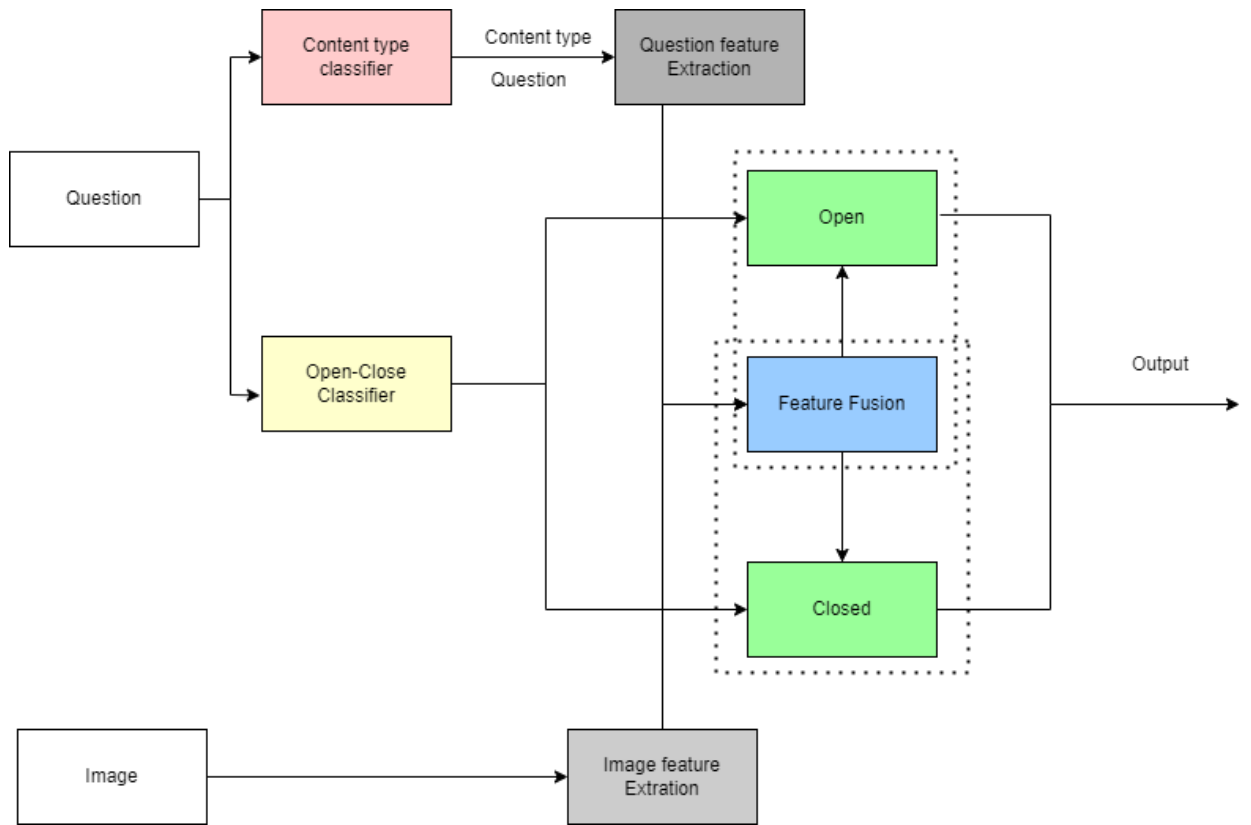


Figure 4.1: Proposed Design

Chapter 5

Implementation Of MVQA

5.1 Dataset Preparation

The SLAKE dataset is obtained, which consists of images and corresponding questions related to medical visual understanding. The dataset is then split into training, validation, and testing sets. The training set is used to train the MVQA model, the validation set is used for hyperparameter tuning and model selection, and the testing set is used for evaluating the final model's performance. The images are preprocessed to normalize their pixel values and to resize them to a uniform size that is appropriate for the MVQA model. This ensures that all images have the same dimensions, which is necessary for efficient processing.

5.2 Classification

An open-close classifier using BERT (Bidirectional Encoder Representations from Transformers) is used for categorizing questions based on their type (open-ended or close-ended). This classifier is used to classify incoming questions and guide subsequent processing steps based on the question type. It has an accuracy of 100%.

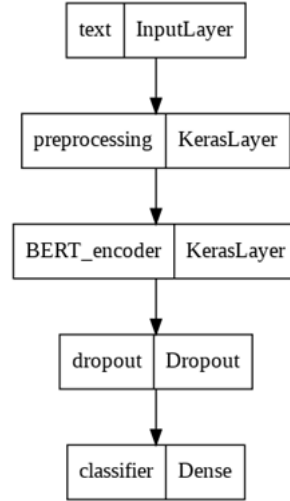


Figure 5.1: Open - Close Classifier

A content type classifier for question classification based on the ten different content types namely Color, Abnormality, Modality, KG, Size, Organ, Plane, Shape, Position, and Quantity is also tested. The classifier was 99.8% accurate.

5.3 Question Feature Extraction

Question features are extracted using GloVe and LSTM. This process involves mathematical operations to transform the textual input into numerical representations.

GloVe embeddings are obtained using co-occurrence statistics in a large corpus of text. Given a vocabulary of V words, each word is assigned a vector

representation of dimension d . The GloVe embedding matrix E contains the vector representations for each word. The GloVe embedding for a word w_i is denoted as e_i .

Given a textual question, it is tokenized into a sequence of words or subword units. This step ensures that each word is treated as a separate entity for further analysis. Each token q_i in the question is mapped to its corresponding GloVe embedding vector e_i . The GloVe embeddings form a matrix E which is fed into the LSTM model sequentially. The LSTM has weight matrices and bias vectors to learn the transformation functions. At each time step t , the LSTM processes the current input x_t (corresponding to the GloVe embedding e_i) and updates its hidden state h_t .

Once the LSTM processes the entire sequence of word embeddings, the final hidden state of the LSTM is extracted. This hidden state represents a condensed representation of the question, capturing its important features and semantic context.

By incorporating GloVe for word embeddings and LSTM for sequence modeling, the question feature extraction stage enhances the system's ability to understand the textual input, capture contextual information, and generate more accurate and contextually relevant answers in the MVQA system.

5.4 Image Feature Extraction

The image features are extracted using a CNN model which consists of three convolutional layers, each followed by a pooling layer and finally a flattening operation. The convolutional layers are the core building blocks of the CNN architecture. Each convolutional layer consists of multiple learnable filters or kernels that scan the input image to extract local features. The filters slide across the image, performing element-wise multiplications and

summations to produce feature maps.

Mathematically, the convolution operation can be represented as follows:

Output feature map (O) = Convolution(Input image (I), Kernel (K)) + Bias (B)

where each element in the output feature map is computed as the sum of element-wise products between the kernel and the corresponding patch of the input image, plus a bias term.

Each convolutional layer is followed by a pooling layer which is used to downsample the feature maps. It helps reduce the spatial dimensions of the feature maps while retaining the most important information. Max pooling is used, which selects the maximum value within a spatial window.

Mathematically, the max pooling operation can be represented as follows:

Output feature map (O) = MaxPooling(Input feature map (I))

where the maximum value within each pooling window is selected to form the output feature map.

After the last pooling layer, the feature maps are flattened into a 1D vector. Flattening converts the spatially arranged feature maps into a sequential representation that can be input to the fully connected layers.

The flattened feature vector is then passed through a fully connected dense layer. The dense layer has 256 units, which means each unit is connected to every element in the flattened feature vector. The dense layer performs a linear transformation followed by a non-linear activation function, allowing the network to learn complex relationships between the features.

Mathematically, the dense layer operation can be represented as follows:

Output (O) = Activation(W * Input (I) + Bias (B))

where the weights (W) and biases (B) are learnable parameters, and the activation function applies a non-linear transformation to introduce non-linearity into the network.

5.5 Feature Fusion

Once the question features and image features are extracted, they are combined using concatenation. Mathematically, the fusion of question and image features using concatenation can be represented as follows:

$$\text{Fused Features} = \text{Concatenate}(\text{Question Features}, \text{Image Features})$$

The fused features capture both textual and visual information, allowing the model to reason about the relationship between the question and the image to generate the final answer.

Chapter 6

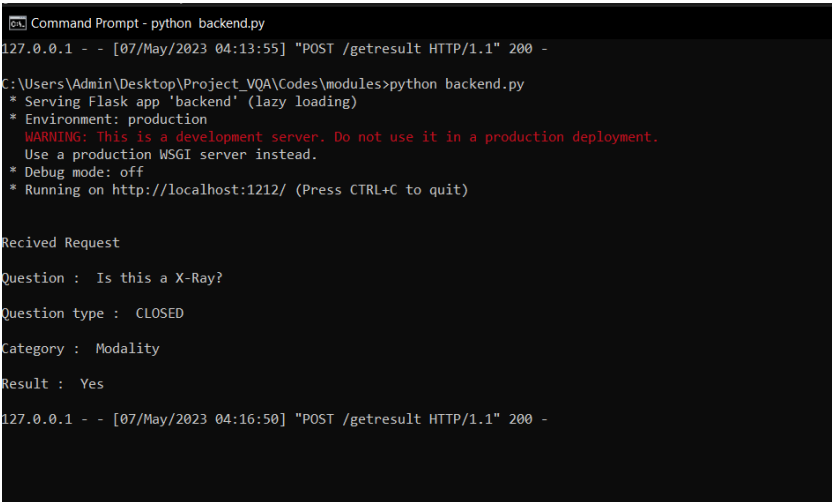
Evaluation Of MVQA

We evaluate the MVQA model using accuracy metric. The accuracy metric measures the percentage of correct answers predicted by the model.

The model produced good results during training, testing, and validation. The open-close classifier classified the questions into open and close-ended questions with an accuracy of 100%. The content type classifier used for question classification based on the ten different content types namely Color, Abnormality, Modality, KG, Size, Organ, Plane, Shape, Position, and Quantity was 99.8% accurate.

Table 7 : Classifiers	
Type	Accuracy
Open-Close	100%
Content	99.8%

We evaluated the model's performance on open-ended and close-ended questions separately. The model achieved an accuracy of 70.9% on open-ended questions and 57.2% on close-ended questions.



```

Command Prompt - python backend.py
127.0.0.1 - - [07/May/2023 04:13:55] "POST /getresult HTTP/1.1" 200 -

C:\Users\Admin\Desktop\Project_VQA\Codes\modules>python backend.py
* Serving Flask app 'backend' (lazy loading)
* Environment: production
  WARNING: This is a development server. Do not use it in a production deployment.
  Use a production WSGI server instead.
* Debug mode: off
* Running on http://localhost:1212/ (Press CTRL+C to quit)

Received Request
Question : Is this a X-Ray?
Question type : CLOSED
Category : Modality
Result : Yes

127.0.0.1 - - [07/May/2023 04:16:50] "POST /getresult HTTP/1.1" 200 -

```

Figure 6.1: Sample output

Table 8 : Accuracies	
Model	Accuracy
Open-ended	70.9%
Close-ended	57.2%

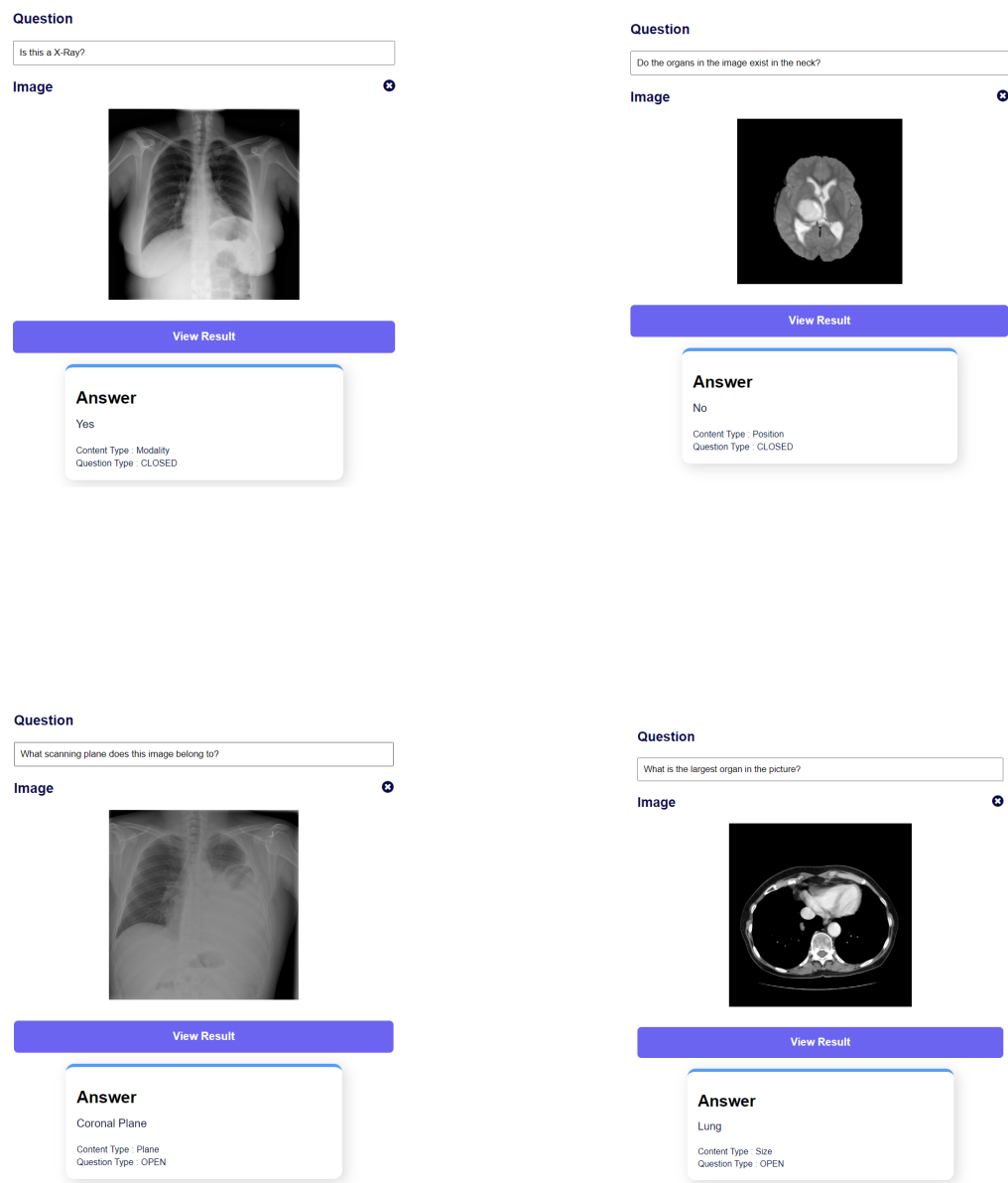


Figure 6.3: Few examples of manual testing

Chapter 7

Conclusion

Med VQA can boost trust in disease diagnosis and aid in patients' understanding of their medical conditions. The usage of Medical VQA frameworks can facilitate the diagnosis process. This project aims to develop a system that can effectively answer questions related to medical images. It uses the SLAKE dataset, which contains a diverse set of questions categorized into different types based on their content.

Throughout the project, we have implemented various techniques and architectures to address the challenges of MVQA. We have explored question feature extraction using GloVe embeddings and LSTM networks to capture the semantic information and contextual understanding of the questions. Additionally, image feature extraction has been performed using a CNN architecture to extract visual patterns and semantics from medical images. To combine the question and image features, we have employed techniques like concatenation, allowing us to create a joint representation that combines both textual and visual information. This fusion of features enables the model to reason and generate accurate answers based on the given question and medical image.

Although our model showed good accuracy in the training, testing, and validation datasets, the accuracy can be increased by using more robust feature extraction models like BERT and VGG in place of GloVe. Better feature fusion techniques can also be utilized in order to improve the accuracy of the model.

UNet architecture has the potential to improve the accuracy and effectiveness of the model. It can potentially improve the accuracy of image feature extraction, particularly in tasks that involve image segmentation or localization of specific regions of interest. Future work could focus on developing a model which uses UNet architecture.

In conclusion, while the task of creating highly accurate and effective medical VQA models is challenging, research efforts are driving significant progress in this field. With advancements in deep learning techniques, the availability of large-scale medical datasets, and the incorporation of domain-specific knowledge, the future of medical VQA looks promising. Continued research and collaboration among experts in the medical and AI communities will lead to the development of highly accurate and reliable models that can truly benefit the medical domain.

References

- [1] Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A. and Van Den Hengel, A., 2017. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163, pp.21-40.
- [2] Lin, Z., Zhang, D., Tac, Q., Shi, D., Haffari, G., Wu, Q., He, M. and Ge, Z., 2021. Medical visual question answering: A survey. *arXiv preprint arXiv:2111.10056*. Inchur, Vilas and L S, Praveen and Shankpal, Preetham. (2020).
- [3] Teney, D. and van den Hengel, A., 2018. Visual question answering as a meta learning task. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 219-235)
- [4] Zhang, A., Tao, W., Li, Z., Wang, H. and Zhang, W., 2022, May. Type-Aware Medical Visual Question Answering. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4838-4842). IEEE.
- [5] Gasmi, K., Ltaifa, I.B., Lejeune, G., Alshammari, H., Ammar, L.B. and Mahmood, M.A., 2022. Optimal deep neural network-based model for answering visual medical question. *Cybernetics and Systems*, 53(5), pp.403-424.
- [6] Gasmi, K., 2022. Hybrid deep learning model for answering visual medical questions. *The Journal of Supercomputing*, pp.1-18.

- [7] Liu, S., Zhang, X., Zhou, X. and Yang, J., 2022. BPI-MVQA: a bi-branch model for medical visual question answering. *BMC Medical Imaging*, 22(1), pp.1- 1
- [8] Eslami, S., de Melo, G. and Meinel, C., 2021. Does CLIP Benefit Visual Question Answering in the Medical Domain as Much as it Does in the General Domain?. *arXiv preprint arXiv:2112.13906*.
- [9] Zhu, H., He, X., Wang, M., Zhang, M. and Qing, L., 2022. Medical visual question answering via corresponding feature fusion combined with semantic attention. *Mathematical Biosciences and Engineering*, 19(10), pp.10192-10212.
- [10] Haridas, H.T., Fouda, M.M., Fadlullah, Z.M., Mahmoud, M., El-Halawany, B.M. and Guizani, M., 2022, May. MED-GPVS: A Deep Learning-Based Joint Biomedical Image Classification and Visual Question Answering System for Precision eHealth. In *ICC 2022-IEEE International Conference on Communications* (pp. 3838-3843). IEEE.
- [11] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, Xiao-Ming Wu. "SLAKE: A Semantically-Labelled Knowledge-Enhanced Dataset For Medical Visual Question Answering". In:arXiv:2102.09542v1(18 Feb 2021).
- [12] 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)*.
- [13] 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 domain, in: *CLEF 2020 Working Notes*, CEUR-WS.org, Thessaloniki, Greece.

- [14] Gupta, A.K., 2017. Survey of visual question answering: Datasets and techniques. arXiv preprint arXiv:1705.03865
- [15] Y. Khare, V. Bagal, M. Mathew, A. Devi, U. D. Priyakumar and C. Jawahar, "MMBERT: Multimodal BERT Pretraining for Improved Medical VQA," 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), 2021, pp. 1033-1036, doi: 10.1109/ISBI48211.2021.9434063
- [16] Kougia V, Pavlopoulos J, Androutsopoulos I. Aueb nlp group at image-clefmed caption 2019. In CLEF (Working Notes), 2019.