

VISUAL INQUIRY ANSWER ON MEDICAL SPACE

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE AWARD OF DEGREE

OF

MASTER OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted By:

VARUN SINGH

2K17/CSE/18

Under the supervision of

MANOJ KUMAR

(Head, Computer Centre and Associate Professor (CSE))



DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College Of engineering)

Bawana Road, Delhi-110042

JUNE, 2019

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College Of engineering)

Bawana Road, Delhi-110042

DECLARATION

I, Varun Singh, Roll No. 2K17/CSE/18 student of M.Tech (Compter Science & Engineering), hereby declare that the Project Dissertation titled “**Visual Inquiry Answer On Medical Space**” which is submitted by me to the Department of Computer Science & Engineering , Delhi Technological University, Delhi Report of the Major II which is being submitted to Delhi Technological University, Delhi, in partial fulfillment for the requirement of the award of degree of Master of Technology for the requirements of the award of degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associate ship, Fellowship or other similar title or recognition.

Place: DTU, Delhi

Date: 22-06-2019

Varun Singh

(2K18/CSE/11)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College Of engineering)

Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled “**Visual Inquiry Answer On Medical Space**” which is submitted by Varun Singh, Roll No. 2K17/CSE/18, Department of computer Science & Engineering, Delhi Technological University, Delhi in partial fulfilment for the requirement of the award of degree of Master of Technology (Computer Science and Engineering) is a record of a project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Date:

(Manoj Kumar)

Head, Computer Centre

and Associate Professor

Department of Computer Engineering

Delhi Technological University

ABSTRACT

This proposal thinks about strategies to tackle Visual Inquiry Answer in Medical Space (VQA-Medical) assignments with a Deep Learning system.

As a primer advance, we investigate Long Short Memory (LSTM) networks utilized in Natural Language Processing (NLP) to handle Question-Replying (content based).

We at that point adjust the past model to acknowledge a picture as a contribution to expansion to the inquiry. For this reason, we investigate the Origin ResNet v2 systems to extricate visual highlights from the picture. These are converged with the word inserting of the inquiry to foresee the appropriate response.

This work was a piece of the Visual Inquiry Noting Challenge CLEF2018 and data set which was released by Nature Recently.

The created programming has pursued the best programming practices and Python code style, giving a predictable gauge in Keras for various designs.

ACKNOWLEDGEMENT

I am most thankful to my family for constantly encouraging me and giving me unconditional support while pursuing this research.

I am extremely grateful to **Manoj Kumar** Head, Computer Centre and Associate Professor, Department of Computer Science Engineering, Delhi Technological University, Delhi for providing invaluable guidance and being a constant source of inspiration throughout my research. I will always be indebted to her for the extensive support and encouragement she provided.

I also convey my heartfelt gratitude to all the research scholars of the web Research Group at Delhi Technological University, for their valuable suggestions and helpful discussions throughout the course of this research work.

Varun Singh

Roll No. 2K17/CSE/18

TABLE OF CONTENTS

Candidate’s Declaration.....	(i)
Certificate.....	(ii)
Abstract.....	(iii)
Acknowledgement.....	(iv)
Table of Contents.....	(v)
List of Figures.....	(vii)
List of Tables.....	(viii)
Chapter 1: Introduction and Outline.....	1
1.1. Overview.....	1
1.2. Applications.....	3
1.3. Specification.....	3
1.4. Organization of Thesis.....	4
Chapter 2: State of Art.....	5
2.1. Related Work.....	5
2.2. Preprocessing.....	7
2.2.1. Image.....	6
2.2.2. Text.....	8
2.3. Visual Question Answering.....	10
2.4. Motivation.....	13
2.5. Contribution.....	15
Chapter 3: Proposed Method.....	16
3.1. Data Set.....	16

3.2. Method.....	16
3.3. Material and Method.....	23
3.3.1 Problem Modeling.....	23
3.3.2 Data Description	24
3.3.2.1 RAD DataSet.....	24
3.3.2.2 CLEF18 DataSet.....	25
3.2. Image Analysis.....	20
3.2.1. Pre-Processing.....	21
3.2.2. Feature Extraction.....	21
3.2.3. Visual vocabulary-(BoVW) model.....	22
3.2.4. Machine Learning Classification.....	23
Chapter 4: Final Model.....	28
4.1.Methodology.....	28
4.2 Hyper-Parameter.....	32
4.3. Evaluation Metric.....	32
Chapter 5: Result & Error Analysis.....	34
5.1. Baseline.....	34
5.2. Evaluation Result.....	34
5.3. Quantitative Analysis.....	35
5.4. Qualitative Analysis.....	35
5.5. Conclusion.....	37
References.....	38

List of figures

Fig 1.1	Example of Model.....	1
Fig 2.1	Real example of the Visual Question-Answering in Medical dataset.....	3
Fig 2.2	An Example of original (Left) and enhanced (Middle and Right) image.....	6
Fig 2.3	LeNet, an example.....	7
Fig 2.4	Word embedding representation in 2-d.....	9
Fig 2.5	Question and got answer from image.....	12
Fig 2.6	Neural Network.....	13
Fig 3.1	Sample question and answer from DAQUAR and COCO-QA dataset.....	17
Fig 3.2	Key components of the VQA framework.....	17
Fig 3.3	Hierarchical Problem model with Question Segregation module.....	23
Fig 3.4	Images in the RAD dataset.....	24
Fig 3.5	Word Frequency distribution in the RAD dataset.....	25
Fig 3.6	Stacked Image.....	27
Fig 3.7	Word Frequency distribution in the CLEF18 dataset.....	27
Fig 4.1	Proposed model architecture.....	28
Fig 4.2	Yn and Oth model architecture.....	30
Fig 4.3	Working of Bi-LSTM for a sample question.....	30
Fig 4.4	Block diagram of Inception-Resnet-v2.....	31
Fig 5.1	Comparison of BLEU scores of the model with/without QS on different datasets.....	35

List of Tables

Table 1	Question-answer pairs about the content of the image	15
Table 2	Sample example from the CLEF18 training data.	26
Table 3	Multiple correct answers	33
Table 4	Semantically similar answers	33
Table 5	BLEU score of yes/no, others and overall question	34
Table 6	Confusion BLEU score of the model with (w.) and without (w/o.)	34
Table 7	Answer prediction of question with type yes/no by model with (w.)/without (w/o.)	36
Table 8	Sample question-answer pair in RAD and CLEF18 dataset	36

there. The reason is that Deep Learning has appeared extraordinary execution taking care of a ton of issues that were recently handled by progressively great AI calculations and it has likewise opened the way to increasingly complex assignments that we couldn't comprehend previously. We people are always posing inquiries and noting them. This is the manner in which we learn, we move information and finally we speak with one another. This essential structure of correspondence has propelled different methods for interchanges, for example, the HTTP Protocol which is fundamentally a mix of a solicitation (question) and a reaction (answer). Frequently Asked Question (FAQ) additionally utilizes this arrangement. Yet, shouldn't something be said about machines? Artificial Intelligence is a tremendous sci-fi subject and it is re-as of now everywhere throughout the news and media, however the fact of the matter isn't that a long way from that point. Deep Neural Systems are these days utilized in our regular day to day existence when we surf the net, when we use proposal frameworks or programmed interpretation frameworks. This has likewise been reached out to handle Question-Answer errands from the Natural Language Processing point of view (for example Facebook artificial intelligence Exploration exhibited a lot of assignments, called bAbI [23], to assess man-made intelligence models' content understanding and thinking).

Visual Inquiry Answer Medical space has developed as an advancement of these content based QA frameworks. These models expect to have the option to respond to a given common inquiry identified with a given picture. One of the interests in such models is that so as to prevail in these visual QA undertakings (or even just content based QA), they need an a lot further dimension of thinking and comprehension than other comparative models, for instance picture inscribing models. A case of a VQA undertaking is appeared in Figure 1.2 This proposition thinks about new models to handle VQA issues. The regular purpose of every one of these models is that they use Convolutional Neural Network(CNN) to process the picture and concentrate visual highlights, which are an abridged portrayal of the picture, and Long Short-Term Memory systems (LSTM), a kind of Recurrent Neural System(RNN), to process the inquiry arrangement.

In light of the given setting, the fundamental goals of this task are:

- Explore the techniques used for text-based Question-Answering
- Build a model able to perform visual question-answering in Medical Domain
- Try different architectures and parameters to increase model's accuracy

- Develop a reusable software project using programming good practices



	<p>Which organ is captured by ct scan?</p>	<p>Lung, pleura, mediastinum.</p>
	<p>What is primary abnormality of the image?</p>	<p>Triplanar fracture of distal tibia.</p>

Fig. 1.2 Real example of the Visual Question-Answering in Medical dataset

1.2. APPLICATIONS

(a)Addresses add-on patient queries

It allows the patient to obtain health information directly, thereby improving the efficiency of diagnosis and treatment.

(b)Used for second opinion

It provides a reference of diagnosis to the doctor.

1.3. SPECIFICATION AND REQUIRMENTS

The specifications are the accompanying:

- Python is used as a programming language
- Manufacture the task utilizing a deep learning framework. Keras has been picked as the structure and it can keep running upon Theano or TensorFlow backends.

As to this, the necessities of this requirements are the accompanying:

- Build up a product that can be utilized later on to continue doing research in this field, having a skeleton/base venture to begin with
- Fabricate a Deep Neural Network that utilizations NLP and CV strategies to process the inquiry & the images individually
- Attempt diverse model arrangements to build the exactness of the first model

1.4. ORGANIZATION OF THESIS

The project report has been divided into five chapters. Each chapter deals with one component related to this thesis. Chapter 1 being introduction to this thesis, gives us the brief introduction about the project, thereafter chapter 2 tells about the literature survey which further includes related work section. Following up is chapter 3 which tells about the proposed work. Chapter 4 provides us with the experiments and results followed by final chapter, chapter 5, that is the conclusion of the thesis.

CHAPTER 2

STATE OF ART

In previous ages, multidisciplinary issues of vision, language and Reasoning had developed as a pattern Artificial Intelligence (AI) investigation. This undertakings joint Knowledge Representation (KR), Natural Language Processing (NLP), Computer Vision (CV), and to have the option to manufacture models that can collaborate with together picture and language input/yield. Be that as it may, this models still come up short accomplishing exactness's near human dimension.

Visual Inquiry Answering has showed up as an issue where models should most likely perform distinctive sub-issues of the over three fields so as to succeed. To take care of this issues the models need an a lot further understanding and perception of the scene in the picture, what the question is referring to and how the things are connected.

2.1. Related Work

A very close study to the VQA-Med task is the VQA challenge. The VQA challenge has been held every year since 2016. The dataset is based on open domain and includes more than 260 thousand images and 5.4 questions on average per image. Kafle K et al. [8] and other researchers summarized quite a few methods for VQA. The majority of them used recurrent neural networks such as LSTM to encode questions, and used deep convolutional neural networks such as VGG16 to focus on image recognition in advance. On the basis of these, there were variant models such as attention mechanisms [17], neural modules [1], dynamic memory [10], and even the addition of external knowledge bases [16], to improve the accuracy of the answers. Deep convolutional neural networks [9] (CNN) can be used to extract the features of an image and identify the objects in it. The Inception-Resnet-v2 model [15] is one kind of advanced convolutional neural network that combines the inception module with ResNet. The remaining connections allow shortcuts in the model to make the network more efficient. Elman J L [3] first used a recurrent neural network (RNN) to handle

sequences problems. Nevertheless, context information is easily ignored when RNN processes long sequences. The proposal of LSTM [6] alleviated the problem of long-distance dependence. Furthermore, the researchers also found that if the input sequence is reversed, the corresponding path from the decoder to the encoder will be shortened, contributing to network memory. The Bi-LSTM model[4] combines the two points above, and makes the result better. On the other hand, there have been many Computer-aided diagnosis systems in medical imaging [2]. However, the majority of them are dealing with single-disease problems, and mainly concentrated on easily-determined regions such as the lungs and skins. The progress of the complex parts is slow. Compared with detection technology, the global lesions and structural lesions are still intensely difficult for the machines to learn. The VQA-Med task differs from the VQA challenge in that it requires the understanding of different kinds of medical images with different body parts.

We will modify a portion of the writing associated with the way toward structure a VQA model, from image and text processing, to the best in class approaches for VQA assignments.

2.2. Pre-processing

2.2.1. Image

For images, we use Inception-Resnet-v2 models to generate their features. In order to reduce the over fitting case, we adopt some image enhancement methods. Considering there are position judgments in the task, we reconstruct the picture with exceedingly small random rotations, offsets, scaling, clipping, and increase to 20 images per image.

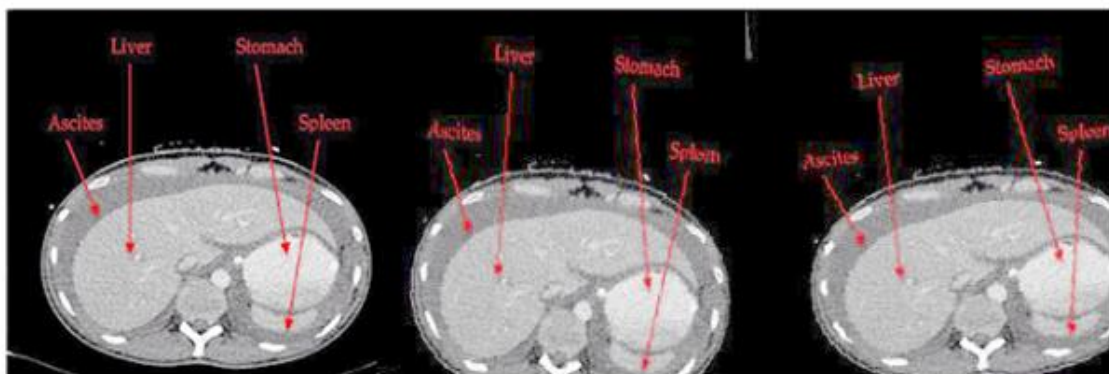


Fig 2.1 An Example of original (Left) and enhanced (Middle and Right) image

Deep Convolutional Neural Networks (CNN) have been demonstrated to accomplish cutting edge results in normal Computer Vision undertakings, for example, image recovery, object discovery and item acknowledgment.

A typical methodology when managing images is to utilize an off-the-rack model (VGG [21], AlexNet [10], GoogLeNet [22], and so on.) pre-prepared to do such assignments with some huge image dataset, for example, ImageNet 1 [4] and utilize a portion of the internal layer's yields as a portrayal of the visual features of the image.

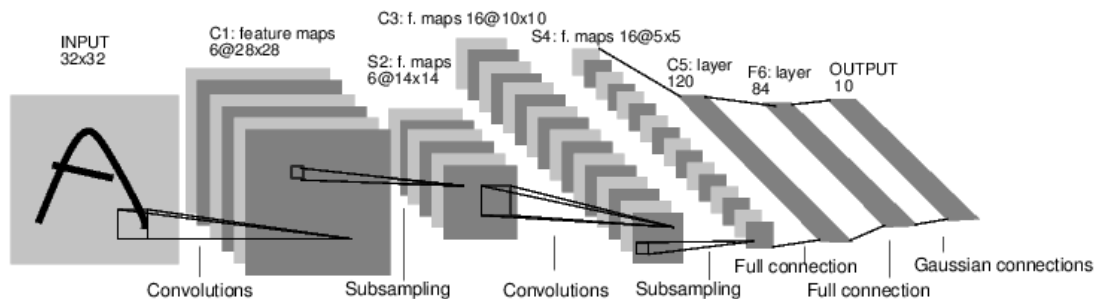


Fig 2.2: LeNet, an example of Convolutional Neural Network

Normally these models have various kinds of layers, among the most widely recognized convolutional layers (that give the name to the CNNs) and completely associated layers. The convolutional layers utilized in image preparing perform 2D convolutions of the past layer yield (which can be an image) where the loads indicate the convolution channel. Conversely, completely associated layers take each yield from the past layer and interface them to the majority of its neurons, losing the spatial information so they can be viewed as one dimensional. A standout among the most widely recognized completely associated layers is the purported softmax layer, which is a standard completely associated with the softmax as actuation work. Its yield pursues a conveyance like shape, taking qualities from 0 to 1 and being the expansion of every one of them equivalent to 1.

2.2.2. Text Processing

For questions, we adopt some methods like stemming and lemmatization to alter verbs, nouns, and other words into original forms, to prevent overfitting. Furthermore, there is a situation that both full name and abbreviation coexist, like “inferior vena cava” and “IVC”. We have changed all these medical terms into abbreviation. There are also a lot of pure numbers and combinations of numbers and letters. Therefore, the combinations of letters and numbers used to represent positions are mapped to an “pos” token, and the pure numbers are mapped to an “num” token, so as to reduce information complexity. In addition, we try to remove useless information such as stop words. According to the word frequency distribution in data analysis, we remove the low-frequency words to ensure training efficiency. In the meanwhile, we establish the dictionaries separately and make sure that the sizes of the dictionaries are both within one thousand. There are some high-frequency verbs like “show” that emerge in almost every question. Several less useful adjectives like “large” also appear in questions from time to time. To cooperate with image enhancement methods, these verbs and adjectives are removed in the questions each time, so that each question is enhanced to 20 questions, and the answer remains unchanged at the same time.

Word embedding techniques (e.g., Word2Vec, GloVe) have been as of late utilized for assortment of utilizations with very great rate of accomplishment. They permit to catch word semantics and syntactics with diminished dimensionality dependent on the idea of distributional vector representations. Vector representations can be then utilized for likeness correlation. In any case, in the event that we treat the word embedding as a sort of encryption process, it is hard to decode their importance. This makes it dangerous to legitimize why specific terms ought to be viewed as comparative just as to demonstrate that the general nature of the prepared vector space is high. Assessing the exactness of the closeness calculation between any two given terms is troublesome because of the absence of solid confirmations to clarify and bolster the comparability. In this paper, we propose a novel method to naturally concentrate confirmations spoke to as term sets to clarify the likeness of self-assertive terms. Our methodology is unsupervised and can be connected to either homogeneous or heterogeneous vector spaces *“Word Embedding is a representation of text where words that have the same meaning have a similar representation. In other words it represents words in*

a coordinate system where related words, based on a corpus of relationships, are placed closer together". In the deep learning frameworks such as TensorFlow, Keras, this part is usually handled by an **embedding layer** which stores a lookup table to map the words represented by numeric indexes to their dense vector representations.

	MAN (5391)	WOMEN (9853)	KING (4914)	QUEEN (7157)	APPLE (456)	ORANGE (6257)
GENDER	-1	1	-0.95	0.97	0.00	0.01
ROYAL	0.01	0.02	0.93	0.97	-0.01	0.00
AGE	0.03	0.02	0.70	0.69	0.03	-0.02
FOOD	0.09	0.01	0.02	0.01	0.95	0.97

Fig 2.3 Word embedding analogies

Deep network takes the sequence of embedding vectors as input and converts them to a compressed representation. The compressed representation effectively captures all the information in the sequence of words in the text. The deep network part is usually an RNN or some forms of it like LSTM/GRU. The dropout is added to overcome the tendency to overfit, a very common problem with RNN based networks.

The **fully connected layer** takes the deep representation from the RNN/LSTM/GRU and transforms it into the final output classes or class scores. This component is comprised of fully connected layers along with batch normalization and optionally dropout layers for regularization.

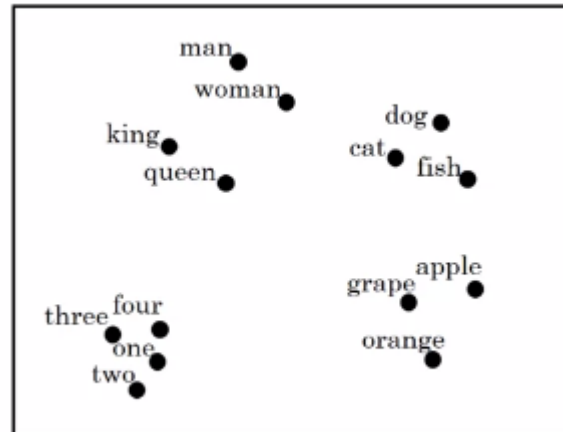


Fig 2.4 Word embedding representation in 2-d

Based on the problem at hand, this layer can have either **Sigmoid** for binary classification or **Softmax** for both binary or multi classification output.

2.3. Visual Question Answering

Computer can solve our problem or can give answer from understanding image is a very novel problem for the CV and NL network, but recently lot of work had been done in this sector all credit goes to VAQ challenge for releasing dataset and the metric. The big firm like Apple, Google, Microsoft had contributed in very large amount to happen this.

The main approach that was used by everyone is by mining the visual feature of the picture and using them on the pretrained network and question are processed in the form of sentence or the word embeddings [1][13][14][6][8]

Baseline model that uses VGG-16 was used by Antol et al[1] the creators of the VAQ dataset and the organizer of the challenge. The model which which is used as baseline was used to extract the visual features of the image. The features are normalized and given to fully connected layer to convert the received vector into common space with the question representation. In the baseline model 2-layer LSTM model is used which take input as word embedding of each question as token and when whole question is introduced it output the status of the question embedded to it. This vector with the dimension of 2048 is given a fully connected layer as we have done same with the image all the feature are combined with

the help of fully connected layer and with help of softmax classifier we can predict the class answer. There are 1000 most predictable answer the class chosen as most predictable class. Zhou et.al on his paper presented a model on VAQ the model uses GoogLeNet for processing of the Image and a word embedded model both then are concatenated this was basic approximation to VAQ.

Zhou et.al on his paper presented a model on VAQ the model uses GoogkeLeNet for processing of the Image and a word embedded model both then are concatenated this was basic approximation to VAQ.

Noh et.al propose the different method that was previously presented,he proposed dyanamic parameter prediction(DPPnet).In this model to solve problem it was stated that different network to be used to solve different problem, depending upon the question. To accomplish they use the question to predict the weight of the network and changing he test time of each sample. VGG-16 and the ImageNet model was taken as pretrained in starting point. They have removed softmax and added fully connected layer after that softmax was added then this network was know as modified VGG-16 network. The new fully connected layer was having dynamic parameter layer this mean at training and testing time parameter where changing new model was VGGNet.

GRU(Gated Recurrent Unit) Network help us to predict the weight of the model know as parameter prediction model connected to fully connected model.GRU is the modified version of the RNN and LSTM embedding each layer as input and when whole question has been passed in network then the last state predict the correct weight of the network. Hashfuction is used to reduce the number of parameter to predict the weight of the model.

Some author introduce attention layer model. This model was introduce to improve the performance of the model. What exactly attention model do?

Attention layer make it work better it is very influencing idea in the AI. First idea that we were memorizing the whole sentence for example we wan to translate the france sentence to English sentence. What previous model was doing they were memorizing the whole sentence then converting to English language sentence. Hence BLEU Score for the very long sentence is reducing. What in real life human do it will take 5-6 word in starting then will translate. Hence the same concept will be applied in the attention model.

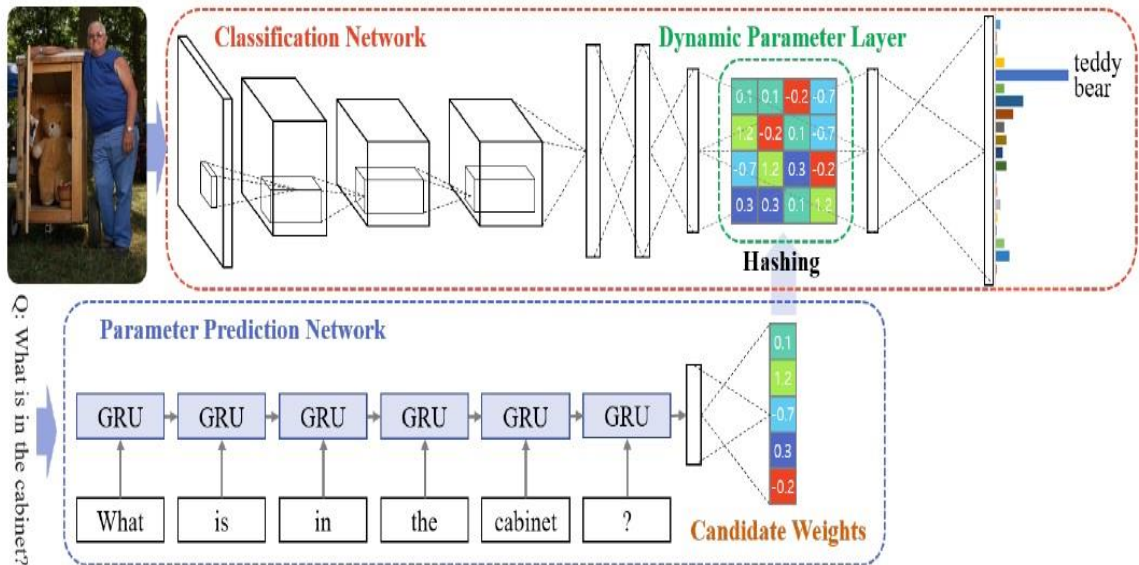


Fig 2.5 question and got answer from image

The intuition of the attention layer model is suppose we use bidirection RNN suppose output output is $y^{<1>}y^{<2>}...y^{<n>}$. We will use another RNN let name this second RNN. First RNN is connected to the second RNN. First one is at the top and the second one at bottom the input to the second layer is alpha weighted from the first RNN model. So this will tell us what should be pair on RNN unit. The RNN on first and second layer are bidirectional.

Let's talk about the basic building block of this thesis. Why we need RNN model? Why simply we can't work with neural network.

Problem with the neural network

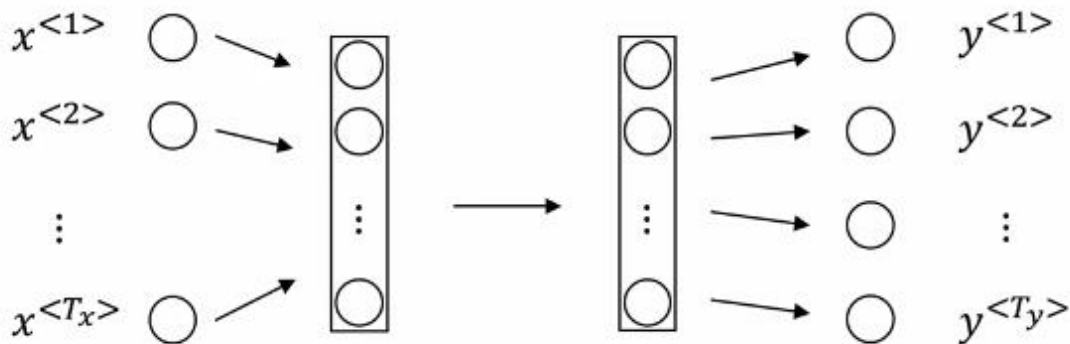


Fig 2.6 Neural Network

- Input output can be of different length but in neural network it should be of same length.
- Doesn't share feature learned across different position of the text.

Problem can be explained with the help of the example suppose if Harry is the name of a person and it appears on the first position of the sentence then it should remain a name only if it occurs at the fourth, fifth, or last word of the sentence but a neural network does not treat it like this. RNN (Recurrent Neural Network) doesn't have a problem of this kind.

In an RNN model we will give one input vector $x_{<1>}$ at the first layer then we will wait for the output vector $y_{<1>}$ and input $x_{<2>}$ is concatenated with the output of the first layer. To predict the value of $y_{<3>}$ not only $x_{<3>}$ is responsible, $x_{<1>}$ and $x_{<2>}$ are also responsible.

One weakness of this kind of model is that it only uses the previous value of $y_{<3>}$; it does not use $x_{<4>}$, $x_{<5>}$, ..., $x_{<n>}$. So this is the problem. For example, he said, "Teddy Roosevelt was a great prime minister". So the model may think it is the name of a thing like a teddy bear. So, by giving the first few words of the sentence you can't figure out what it is, so something else. To overcome this issue we will talk about LSTM, bidirectional LSTM, GRU; these are all upgraded versions of the RNN.

2.4 Motivation

The development and advancement of VQA is seen as a solution to various real-world challenging problems. The field of medicine is one such area. Medical and physiological images (CT-scan, X-ray, etc.), and reports for the patients are nowadays easily accessible with the increase in use of medical portals. But to understand the reports, and get answers to the related queries, a patient still needs to visit a medical expert. This involves payment of considerable fees, even to get the answer to simple queries. Also, investment of time plays a crucial role, which at times creates a vacuum in the timely delivery of an answer. Clinicians on the other hand can find it useful for reporting findings of a complicated medical image or can use it as a second opinion just to boost their self-confidence in understanding a particular aspect of such reports. Although it is possible to search the queries on search engines, but the results over there may be inaccurate, spurious, vague, and/or enormous. In this context

the VQA in medical domain (VQA-Med) is getting attention as an important problem domain, trying to solve queries related to medical images.

VQA-Med intends to assist mainly patients and clinicians but can also be useful in Medical education, as clinical apprentice or medical students who just started learning the basics of handling images of various medical modality (e.g radiology images) may learn by asking queries. Thus, developing an efficient and automated VQA system for medical domain comes out as an essential task. Even though many medical datasets are published publicly, most of them deals with some specific disease in a particular body part with a fixed image modality. ImageCLEFtuberculosis task [11] is one such example which was published to build models for detection, classification, and severity measurement of TB from the provided chest-CT). On the contrary a generic approach, which deals with VQA queries with respect to various image modalities, related to diverse diseases appearing in any part of the body is a much more challenging task. Moreover, a generic solution will be considerably less confusing to the patients, as a single portal may provide answers to all their queries regardless of modality, body part and disease.

The difference in type of end users results in different query type. The queries from patients most of the times are expected to be generic in nature, having the need to produce 'Yes' or 'No' as answer. Whereas, the queries from clinicians and medical experts are expected to be more problem specific, that requires elaborate answers. Again, a skilled trainee is expected to ask more specific and sophisticated question, while queries from beginners are likely to be simple and straightforward. For example, a naive trainee may inquire about the presence of any abnormality in the image whereas, a senior trainee may already identify the abnormality of 'intraventricular hemorrhage' from the image, and want to understand more about the grade and effect of the hemorrhage. They can then draw inferences from the acquired data for treatment.

This difference in query type thus needs different problem specific attention, which needs to be dealt separately in isolation. Again, multiple end systems for multiple types of queries may create confusion, and discomfort to end user. There should be a single end user module to solve both the complex, and simple queries. Table 1 demonstrates one such system where

any clinically relevant question can be asked about the image. Here, image plays an important role as the answer of the questions may vary based on the provided image.


Image	Question	Answer
	Is this a cyst in the left lung?	No
	Has the left lung collapsed?	Yes
	Where is the nodule?	Below the 7 th Rib
	What are the densities in both mid-lung fields?	Pleural plaques

Table 1. Question-answer pairs about the content of the image

We identify this need, and propose a hierarchical multimodal deep neural model, which at first deals with the problem of question type identification, then generate answers based on the question type.

2.5. Contributions

In this work, we try to approach a solution for the generic problem in VQA-Med. Particularly our work focuses on proposing a multimodal machine learning model, which generates proper answers to queries related to medical images of various modality. In addition we also deal with segregation of input features based on query/question type. We propose a hierarchical deep neural model, with a segregation node at the root, and separate deep neural models at different leaf branches dealing with different types of queries. The problem intended to be solved by the dataset published by Nature Publishing Group in the journal Scientific data [12], perfectly captures our problem statement that we aim to solve in this work. We perform all our experiments in the said dataset and ImageCLEFVQA-Med2018(CLEF18) [13] dataset. More detailed discussion on the dataset can be found in the section Data Description. Experimental evaluation shows promising results, depicting effectiveness of our approach. Further our detailed analysis of errors of the system output, highlights the room for improvement in the task.

In brief our contribution is:

- A hierarchical deep neural multimodal model, which generates proper answers to queries related to medical images.
- A question segregation module to differentiate the learning path for questions with yes/no answers from other types of questions, for solving the generic VQA-Med problem.
- quantitative and qualitative analysis, to demonstrate that the proposed method enhances the performance of a base VQA-Med model by a significant margin

CHAPTER 3

PROPOSED METHOD

3.1 Data Set

At the beginning of 2014, five major VQA datasets were released. These datasets facilitate the training and evaluation of VQA systems. The Dataset for Question Answering on Real-world Images (DAQUAR) [14] is the first benchmark dataset among them. It contains images from the dataset of NYUDepth V2 [15] containing images along with their semantic segmentations. All the images are of indoor scenes where each pixel of the image is marked with an object class out of 894 possible classes or as no-object class. The dataset is made up of 6794 training, and 5674 test QA pairs which are of two types: synthetic and human. There is also a reduced form of the same dataset with only 37 object classes. The limitations of DAQUAR is the restriction of answers to a predefined set and strong bias in human annotations. This led to several other datasets being released, most prominently COCO-QA [16]. It is based on images from MS-COCO [17] dataset, where significant efforts have been made to increase the amount of training data. Using the image captions from MS-COCO, both questions and answers were produced automatically. There are a total of 123,287 images in the dataset with one question per image with one-word answers.



(a) DAQUAR

Que: What is the object on the chair?

Ans: pillow



(b) COCO-QA

Que: What is the color of the coat?

Ans: yellow

Fig 3.1. Sample question and answer from DAQUAR and COCO-QA dataset. Taken from [16]

Other datasets were also created from the same MS-COCO dataset such as, Visual7W [18] with 47,300 images and 327,939 questions and Visual Madlibs [19] with 10,738 images and 360,001 questions. Visual7W is for questions of multiple choices only, and this dataset does not contain binary questions to make the QA pairs more diverse. While Visual Madlibs comprises of both fill-in-the-blanks and multiple choice questions. The descriptions in this dataset go beyond the objects in the image and are more detailed than a generic description of the image as a whole

Visual Question Answering (VQA 1.0) [9], is another dataset created from MS-COCO dataset. It is the most significant and commonly used dataset for the VQA task which was published as a part of the VQA challenge. It is divided into two components: one includes MS-COCO real-world pictures, and the other includes abstract clip scenes generated from human and animal models. It has 204,721 images with 614,163 questions and 50,000 abstract scenes with 150,000 questions respectively. About 40% of the questions have a yes or no as answer. For each image, there are three free-form natural-language questions with ten concise open-ended responses each with two task formats: open-ended and multiple-choice. The VQA 1.0 dataset's primary issue was its inherent bias and language priors too had a major impact on the responses to the questions that inspired the design of the second version of this dataset. VQA 2.0 [10] is a larger dataset with 265,016 images and abstract scenes in total and an average of 5.4 questions per image. Specifically, it is a balanced version of the popular VQA 1.0 dataset. It has supplementary pictures such that each question in the balanced dataset is connected not only with a single image, but with a couple of comparable pictures resulting in two distinct responses to the issue.

As the field began to mature, researchers found the significance of biases, which resulted in the release of more balanced datasets such as CLEVR [20]. It also led to the assessment of the significance of the common-sense knowledge needed to solve a specific problem, which in turn led to attempts to deliver ground-based reality with a wealthy description of the scene, facilitating both training and assessment (e.g. Visual Genome [21]).

The VQA-Med [12] [13] dataset is very different from the VQA datasets discussed in this section. The obvious reason is its focus on the medical domain, which gives this dataset its unique set of challenges. The images, questions and answers must be clinically relevant in order to be a part of this dataset which is not a constraint in VQA datasets. The building of phrases for sentences is another distinction. Most of the VQA-Med sentences are complex with lots of medical terms, while they are simple and straightforward in the stated datasets. Another difference is the incomparable size of the dataset, medical domain data resources are limited compared to the general domain data resources which are usually huge (e.g., thousands vs. millions of samples). The images in this dataset may have different modalities, may also contain radiology markings such as short information, tags, etc., and may also contain a stack of sub-images which is not the case with the existing VQA datasets. The number of reference answers, which is just one, is another drawback with this dataset.

3.2 Method

VQA tasks are based primarily on three key components: generating representations of images and questions; passing these inputs through a neural network to produce a co-dependent embedding; and then generating the correct response. Fig. illustrates this framework where the key components can take wide variety of forms.

VQA systems differ from each other in the way they fuse multimodal information. A few such examples which are relevant to our work are:

- Multimodal feature fusion using simple mechanisms, e.g., concatenation, element-wise addition/multiplication, followed by feeding them to a linear-classifier or a neural network [22].
- Multimodal bilinear pooling or related feature fusion strategy [23] [24]

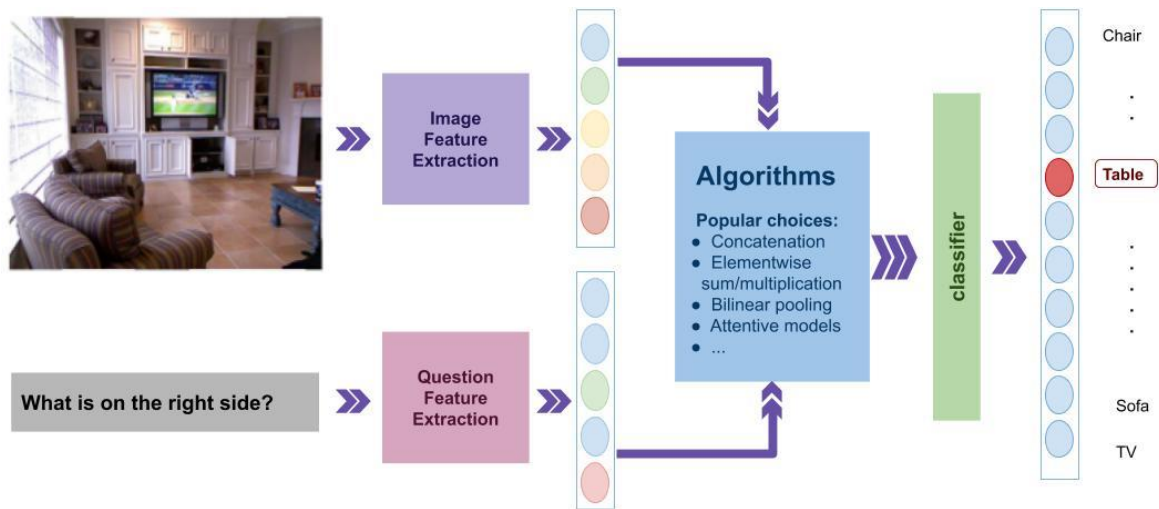


Fig 3.2. Key components of the VQA framework.

- A classifier that utilizes the question features to calculate maps of spatial attention for the visual features or that adaptively scales local characteristics according to their relative significance [25] [26] [27] [28].

Although most open-ended VQA algorithms use the classification mechanism, this strategy can only produce answers seen during training. Multi-word response is generated one word at a time using an LSTM [29] [10]. The response generated, however, is still restricted to words seen in the course of training.

For question encoding, most methods for VQA uses variant of recurrent neural network (RNN) [30]. RNNs are capable of handling sequence problems, but when RNN processes lengthy sequences, context data is easily ignored. LSTM's (long Short-Term Memory) [31] proposal mitigated the long-distance dependency issue. In addition, the researchers also discovered that the respective route from the decoder to the encoder will be reduced if the input sequence is inverted, contributing to network memory. The Bidirectional LSTM (BiLSTM) [32] model combines the above two points and improves the results. The Gated Recurrent Unit (GRU) [33] is also a notable, and widely used, simplification of the LSTM. As for the image feature extraction, Convolutional Neural Network (CNN) [34] are used where VGG-net [35] and deep residual networks (ResNet) [36] are most popular. Since these features are from different feature spaces there needs to be some mechanism to fuse them together in order to preserve the relationship between them. Numerous multimodal fusion

methods ranging from simple concatenation or element-wise summation to advance methods are available these days.

LSTM was used to extract question features, operating on a one-hot sentence encoding, and GoogLeNet was used to featurize the image in [9]. Length of the feature vectors of both image and question was mapped to the same dimensional space and then the two vectors' Hadamard product was fused together. As an input to an MLP, the fused vector was used to generate the answer. In [10], an LSTM model was supplied a sequential embedding of each term with fused CNN features. This persisted until the question ended. A list of responses was produced using the subsequent time steps. A similar method was used in [37], where CNN features were supplied to an LSTM in the first and last time steps, with word features between them. The image features served as the sentence's first and last words. To predict the response a softmax classifier followed the LSTM network. A comparable strategy was used in [38], but the CNN features were only fed into the LSTM at the end of the question and a separate LSTM was used to produce the response one word at a time instead of a classifier.

Application of attention on image can help to improve the performance of the model by discarding the irrelevant parts of the image. So, attention mechanisms [39] [26] are usually incorporated in the models so that they may learn to 'attend' to the important regions of the input image. However, attending image is not enough but question attention is important too as most of the words in the question may be irrelevant so simultaneous integration of both question and image attention is advised [40]. The fundamental concept behind all these attentive models is that for answering a specific question, certain visual areas in an image and certain words in a question provides more information than others. The Stacked Attention Network (SAN) [26] and the Dynamic Memory Network (DMN) [39] used image features from a CNN feature map's spatial grid. Both the approach resized the image to 448 * 448 to generate a feature map of dimension 512 at each grid location using the last VGG-19 convolution layer. In [26] an attention layer is specified by a single layer of weights using the question and image feature defined to calculate attention distribution across image locations. Using a weighted sum, this allocation is then applied to the CNN feature map to pool across spatial feature locations. It creates a global representation of the image that highlights certain spatial regions. Then they are fused to generate the answer. This approach

was generalized to manage multiple (stacked) layers of attention. Similar to this, the Spatial Memory Network [25] uses spatial attention which is generated by predicting the correlation of image segments with individual words in the question. [39] proposed another related strategy that uses CNN feature maps to incorporate attention. Hierarchical Co-Attention [40] gives attention to both modalities to reason jointly on the two different streams of information. The visual attention in this model is analogous to the technique used in the Spatial Memory Network [25]. [41] also explored the joint image and question attention.

VQA depends on the image and question being processed together. Features of these two modalities are required to be integrated together. This was achieved earlier by using simplified methods such as concatenation or element-wise product, but these methods fail to capture the complex interactions between these two modalities. But to capture these complicated interactions, an outer product can be used. Later, multimodal bilinear pooling was proposed as a novel method for joint feature representation where the idea was to approximate the outer product between the two features, enabling a much deeper interaction between them. Similar concepts have been shown to work well to improve the fine grained image recognition [1]. Multimodal Compact Bilinear (MCB) [23], Multi-modal Low-rank Bi-linear pooling (MLB) [42] are the two most significant VQA techniques used in bilinear pooling. MCB calculates the outer product in a reduced dimensional space instead of explicit calculation to minimize the number of parameters to be learned. Then this is used to predict the relevant spatial features according to the question. A slight variation of this model that uses soft attention was used in [26]. The major change was the use of MCB for feature fusion instead of element-wise multiplication. Later, to further reduce the number of parameters that are to be learned, MLB was introduced that uses Hadamard product and a linear mapping. Afterwards [43] developed a refined version of MLB i.e., Multimodal Factorized Bilinear pooling (MFB) where the multimodal-features were expanded to a high dimensional space and then squeezed to create the fused feature. A more generalized version of MFB, multimodal factorized high-order(MFH) [44] pooling containing N-MFB modules was then proposed.

Methods for VQA-Med must be different from general VQA as the size of the datasets are incomparable. The other challenge with VQA-Med is to balance the number of image features (usually thousands) with the number of clinical features (usually just a few) in the

deep learning network to avoid drowning out of the clinical features. Attention based on bounding box too cannot be applied directly as medical images lack the bounding box information. For medical imaging there are many computer-aided diagnostic systems [45] [46] [47]. Most of them, however, deal with single disease problems, and focused primarily on easily identifiable areas such as the lungs and skins. In contrast to these systems VQA-Med deals with multiple diseases at the same time apart from handling multiple body parts which is difficult for machines to learn.

Inherently, question pursue a worldly succession and normally group into various sorts. This inquire type information is important to predict the response regardless of the image. [48] use similar approach where they first identify the question type and use this information for answer generation. Our work however isolates the learning path based on question-type rather than using this knowledge as feature. This type information can also affect the model performance as some of the VQA models perform better than others for certain types of questions. Therefore, these models can be intelligently combined to leverage their varied strengths. We propose a simple model with question segregation module which segregates the learning path based on the question type(yes/no and others) to reap the benefits of question-type dedicated models. We use Inception Resnet to encode image feature and BiLSTM for question feature creation. Significant improvement is clearly visible when we compare our model models with the baseline modal.

3.3 Materials and methods

3.3.1 Problem modeling

Given a doublet (Q, I), where Q is the question accompanied with any clinically relevant image I, the VQA-Med task is to generate the appropriate answer (A) which can either be yes/no or open-ended (answers other than yes/no). Mathematically formula ,

$$A = f(Q, I; \alpha)$$

where, f is the answer prediction function and α denotes the model parameters.

Our approach towards solution of the problem statement is to view the task as an hierarchy of two different task. At the top level of the hierarchy we view our task as identifying type of question (yes/no or others), to differentiate the learning path for questions with yes/no answers from other types of questions. We propose a question segregation module to handle this task. At the next level of the hierarchy the task of answer generation gets divide into two leaf branches. One branch deals with the problem of producing simple (Yes/No) answers from simple incoming queries with respect to given images, while the other branch deals with complex queries to produce open ended expertise answers. The problem model is depicted in the fig. 3.3.

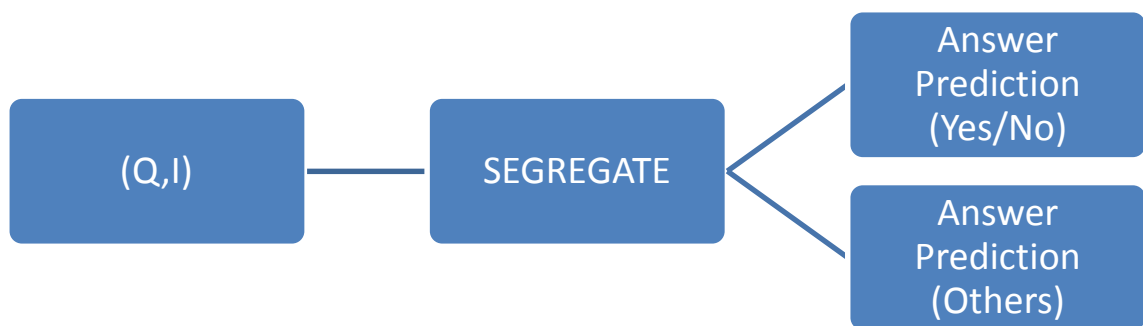


Fig 3.3 Hierarchical problem model with question segregation module.

3.3.2 Data Description

Datasets for VQA-Med consists of Natural Language Questions about the content of radiography images, and the task is to generate the appropriate answer. The questions are framed on different modalities of medical image like angiogram, magnetic resonance imaging, computed tomography, ultrasound, etc that describes how the image is taken. These images can have different orientations e.g. sagittal, axial, longitudinal, coronal, etc. Along with variety in orientation and modalities images can be of any body part or organ such as heart, lung, skull, etc.

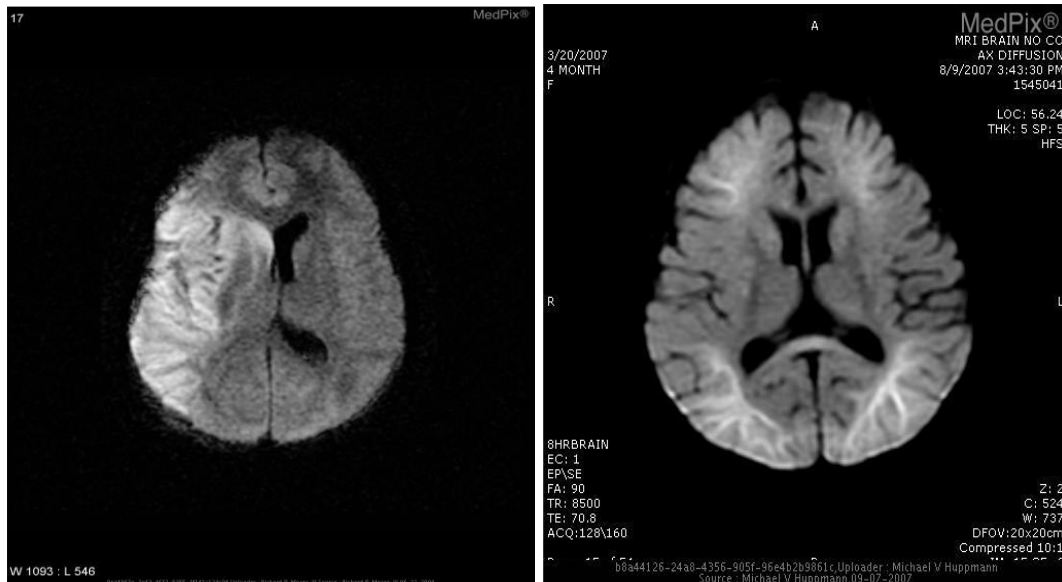
3.3.2.1 RAD Data Set

RAD [12] dataset is recently launched dataset for VQA in medical domain. Statistics of the dataset are as follows:

- The training set consists of 1,797 question-answer pair.

- The test set consists of 451 question-answer pair.

Some of the images in the dataset are blurred as depicted in fig. 3.4a. While others contain markings such as short information, tags, etc (fig. 3.4b), but none of the images in the dataset contains stack of sub-images.



(a) Blurred image.

(b) Image with Radiology markings.

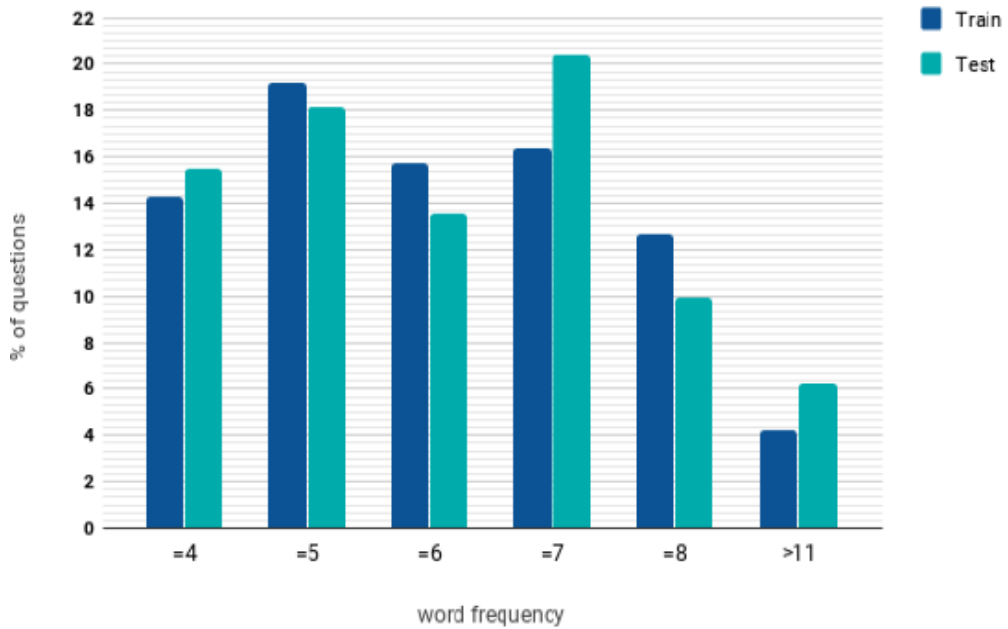
Fig 3.4. Images in the RAD dataset.

The questions are primarily categorized into the 11 categories viz. abnormality, attribute, color, counting, modality, organ system, other, plane, positional reasoning, and size. The average length of question is 5 to 7 words which is greater than that of the answers having 53% percent of answers being Yes/No. Apart from Yes/No, most of the answer are either of one or two words, or short phrases. The maximum length of questions in the dataset is around 21 words, with average being of 7 words. Also it needs to be noted that a lot of questions are being rephrased, which are similar semantically.

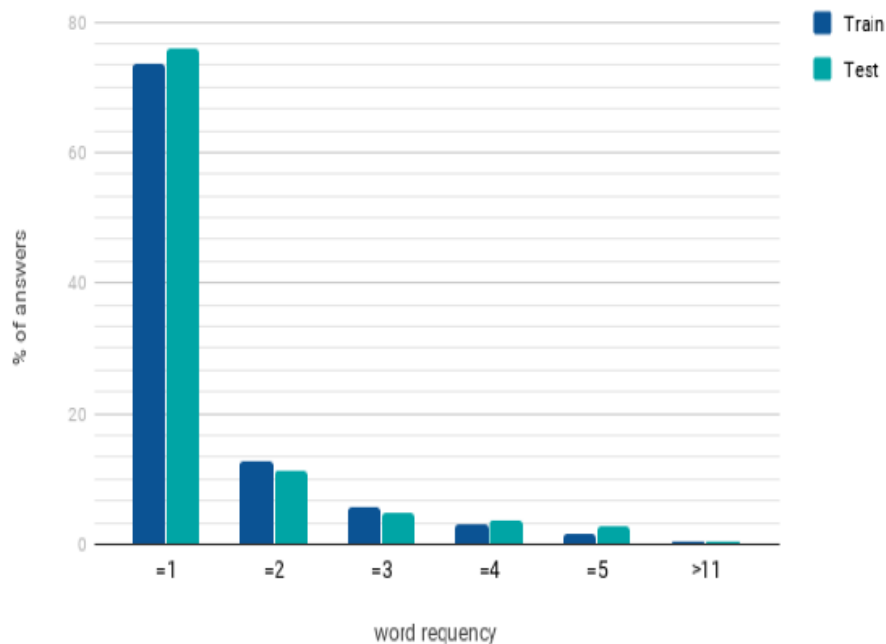
For example,

- What is the size and density of the lesion?
- Describe the size and density of this lesion?

From statistical study of the dataset, we find that only 87% of the free-form, and 93% of the rephrased questions are unique, while only 32% answers are unique. More than half answers are of yes/no type. This constitutes to the fact which shows the peak in the graph 5b for answers of length 1.



(a) Questions



(b) Answers

Fig 3.5. Word Frequency distribution in the RAD dataset.

3.3.2.2 CLEF18 Dataset

ImageCLEFVQA-Med2018(CLEF18) [13] task is similar to RAD task. PubMed Central articles¹ (essentially a subset of the ImageCLEF2017 caption prediction task [49]) is used to extract the radiology images and their respective captions. A semi-automatic approach is then used to generate the questions and answers from the extracted information. Due the way the QA pairs are generated they are diverse and descriptive. Dataset also contains a lot of artificial questions that are semantically invalid. Table 2 demonstrates a few question-answer pair from the training data.

Question	Answer
What uncovers noticeable reciprocal improving parietal occipital sores on style and t2 arrangements and little regions of hyper force in the left periventricular white issue on dissemination weighted pictures?	MRI of the brain
What does MRI in sagittal plane show?	the accumulation was shallow to the muscles of the back and the gluteal district however profound to the back layer of the thoraco lumbar belt

Table 2. Sample example from the CLEF18 training data.

Statistics of the provided dataset are as follows:

- The training set consists of 5413 questions along with their respective answers about 2,278 images.
- The validation set consists of 500 questions along with their respective answers about 324 images.
- The test set consists of 500 questions about 264 images.

Some of the images in the dataset are blurred fig. 3.4a and most of the images contains radiology markings fig. 3.4b such as short information, tags, arrows, etc. A few of them even consists of stack of sub-images(Fig.3.6).

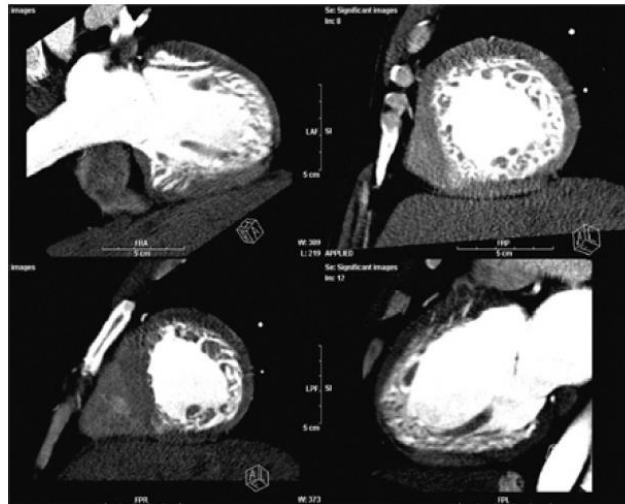
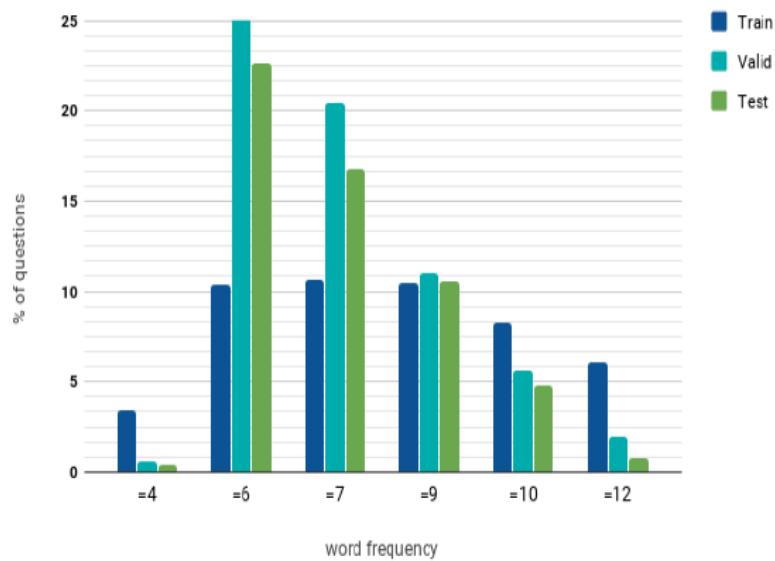
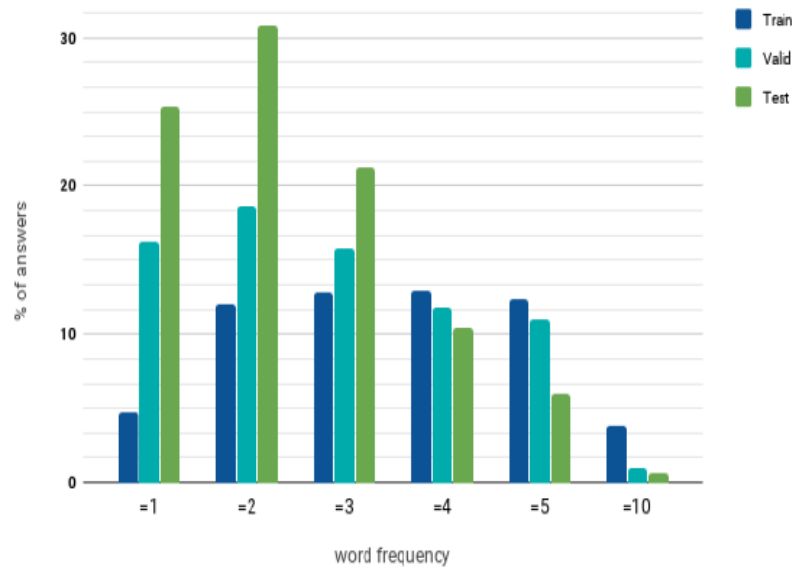


Fig3.6 Stacked Image

Question categorization is not present in dataset. Only 0.6% of answers in train, 6% in valid, and 10% data are Yes/No. From table 2 The average length of questions are more than answers and distribution of word frequency is not similar in the train, test, and valid data.



(a)Question



(b) Answer

Fig 3.7. Word Frequency distribution in the CLEF18 dataset.

First Step

Questions and Answers. At first we convert words of the question and answer into lowercase, and then lemmatize them to reduce ambiguity among different forms. Next we remove some words like 'the', 'and', 'with' etc. to discard useless information. We then map pure numbers to `num` token and alphanumeric words to `pos` token to minimize complexity of information in questions. For answers, we replace low-frequency words by `abnormality`. We create separate vocabulary dictionaries for questions, and answers. As negligible number of questions are of length greater than 21 so, thus we fix the maximum question length as 21. Similarly, for answers having type others, we prune the maximum length to 11. However, for yes/no type answers, the length is 1, as only yes or no are the probable answer. Consistency is maintained in the input length by appending "blank" at the end for the shorter sequences, and curbing longer sequences up to the required length.

CHAPTER 4

FINAL MODEL

4.1 Methodology

Segregation. Our proposed system is a hierarchical Multi-modal deep neural model, which deals with the problem of question segregation in order to generate answer based on the question type as described in fig. 4.1.

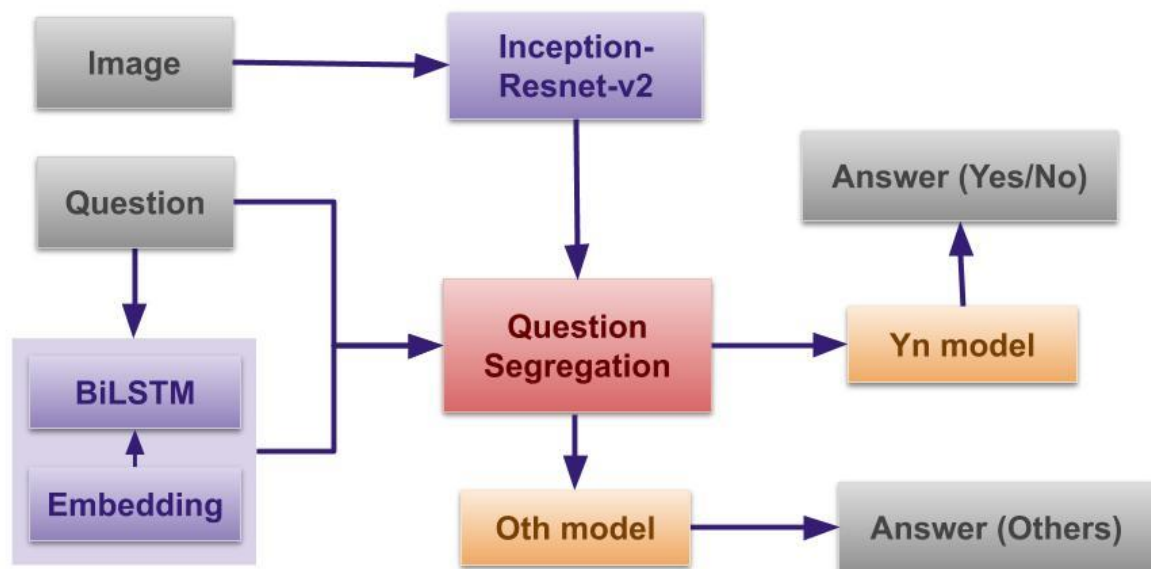


Fig 4.1. Proposed model architecture.

The question-type knowledge may not be readily available. But the question type can still be predicted from the text. As, the task of question segregation is relatively an easier task compared to answer prediction. Thus, we find a simple statistical learning model, based on simple hand engineered, and word frequency based features to effectively solve the problem poised in the top tier (segregation) of our proposed hierarchy model (fig. 4.1). We use Linear SVM learner as our base classifier with the following two feature set:

- **TF-IDF vector:** The questions are converted into one hot vectors, with each word position represented by the corresponding tf-idf value of the word. The entire vocabulary in the training set have nearly 2000 words. Out of that we consider only top 500 words with the highest tf-idf values.

- **Question Identifier Vector:** After studying the training set, we form a set of identifier words, where each word tries to represent a question motive. The identifier word set consists of 10 words ['is','was','are','how','can','does','which', 'what', 'type', 'there']. Each question is converted into a vector of length 10, where each position marks the presence or absence of the corresponding word in the question with 1 or 0 respectively.

Yn and Oth model. The model (for both yes/no, and other types) Fig.4.2 mainly consists of Feature extraction and Feature Fusion.

For the extraction of question features, we first generate a list by gathering the questions in the data set. Next we create a word-index dictionary of the top 1050 frequent words in the list. We then transform the generated list in each split to a sequence of integers. Basically we take every word in the text and replace it with its respective index value from the dictionary. To maintain consistency in the length of the sequence, for the shorter sequences, we attach "blank" at the end and cut the longer sequences to the required length. We then generate 300-dimensional vectors for question embedding using GloVe [50]. But for questions other than yes/no we create 600-dimensional vectors by appending the custom word embedding vectors pre-trained

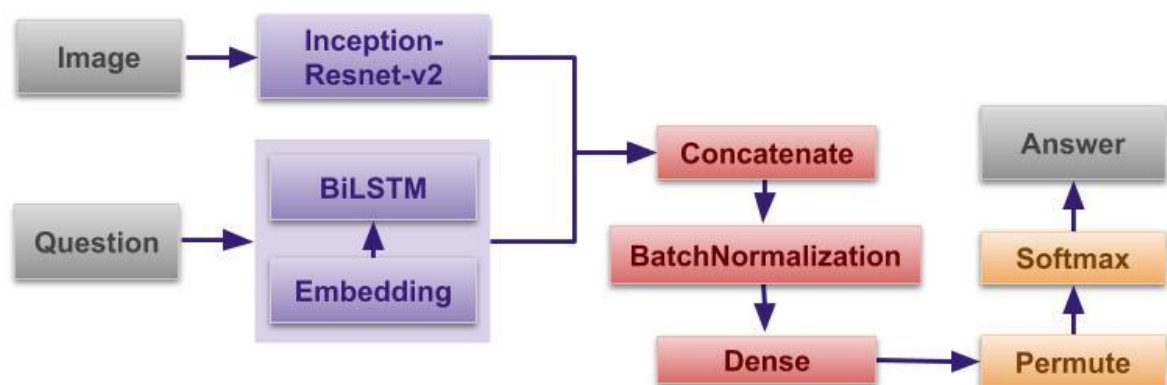


Fig 4.2. Yn and Oth model architecture.

On GloVe to the previously generated vectors. For custom word embedding we use all the questions and answers from the dataset. Questions with type yes/no are simple and straightforward. Thus, we don't need embedding combination. But, for the rest of the questions we follow the approach similar to the work proposed by Ghannay et al. [51], and use combined embedding. According to the question-type we feed the respective embedding vectors to the Bidirectional LSTM (Bi-LSTM) layer to capture the sequence information in the question. Unidirectional LSTM retains prior information as it has only seen past inputs and in bidirectional layer inputs will be run bidirectionally in two ways, one from the past to the future and vice-versa. Therefore, we use a bidirectional layer with LSTM as input for the recurrent instance to preserve information from both past and future. Working of Bi-LSTM for a sample question is depicted in fig. 4.3

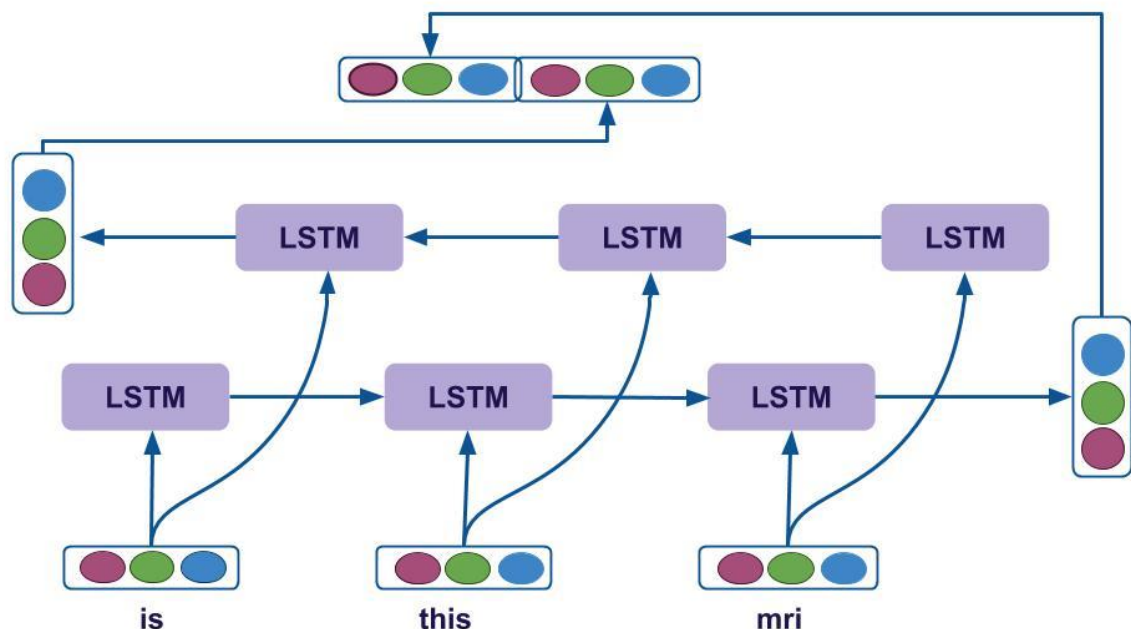


Fig 4.3. Working of Bi-LSTM for a sample question.

We use this Bi-LSTM layer with sequences returned so that LSTM hidden layer returns a sequence of values one per time-step instead of returning a single value for the entire sequence. To minimize the problem of overfitting due to small amount of training data we also use dropout value of 50% in the BiLSTM layer. The last time step output of the Bi-LSTM layer is then used as the textual features of the questions.

For image feature extraction, we use the Inception-Resnet-v2 [52] model, after resizing the input image to a dimension of 224 X 224. Block diagram of the compressed version of the model is depicted in fig. 4.4 which is taken from googleblog.

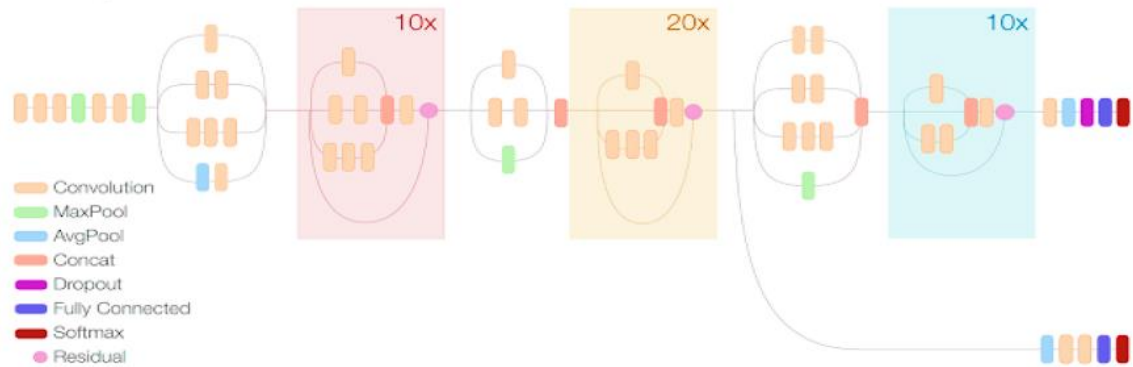


Fig 4.4. Block diagram of Inception-Resnet-v2 (Compressed view).

It is a type of advanced CNN that integrates the inception module with ResNet where connections allow shortcuts in the model to improve the efficiency of the network. Basically, it utilizes residual links to combine filters of varying dimensions, which not only prevents the issue of degradation caused by deep structures but also decreases the training cost. We initialize the model with weights pre-trained on imagenet [53] based on the Apache License3. Such initialization facilitates transfer learning [54] which is incorporated to enable a model to learn from another model pre-trained on a bigger dataset. It helps to train our deep neural network with comparatively small data. Though the type of images in the medical domain is very different from those in the general domain still, transferring learned knowledge is more promising than training straight from scratch [55]. We extracted the last layer of this model, as global features of the image. For multi-modal feature fusion, we pass the extracted features from both the image and question through the concatenation layer. We then pass the output of this layer to BatchNormalization layer for regularization and to increase the stability of the network. We pass the normalized fused feature to the dense layer where the number of neurons depends on the answer length of respective question-type. The fusion information is then finally used to predict the answer using multi-class classification method. For this, we use TimeDistributed layer with dense layer having softmax activation. The number of neurons in the dense layer depends on size of the answer dictionaries that we create

separately for the two question types. We use categorical cross entropy as the loss function having the following formula.

$$H(T; q) = \sum_{i=1}^n \frac{1}{N} \log_2 q(x_i)$$

where, $q(x)$ is the probability of event x calculated from the training set, and N is the size of validation set.

4.2 Hyper-parameters

According to the model performance, the hyper-parameters are set as follows. Dictionary size is 1050 for questions to capture the most frequent words. Answer dictionary is of size equals to the count of unique word in the ans-list. Question length is 21. Answer length for yes/no and others type question is 1 and 11 respectively. The hidden layer of Bi-LSTM has 128 neurons in each direction. Categorical accuracy is the metrics method. Adam is the optimizer. For training, batch size is 256. Monitor is on validation loss. Epoc is set to 51.

4.3 Evaluation Metric

The BLEU score metric proposed by [56], is implemented using NLTK tool4. It is a popular evaluation metric in machine translation, which compares the generated answer with the reference answer based on number of n-grams of generated answer that matches with the reference answer, along with brevity penalty for shorter output.

In our work, the dataset we use for evaluating our proposed hypothesis has only a single reference answer. Thus, reporting high accuracy means generating high number of answers with exact same words as in the reference answer. This may not always be a necessary, and is a very complex task even in the medical domain. However, more than one answer may be correct e.g. in the absence of the degree of specification. This is explained elaborately with an example given in table 3.

Question	Answer
Where is the lung lesion located?	<ul style="list-style-type: none"> • Right lobe • Lower lobe • Right Lower lobe

Table 3. Multiple correct answers

Thus, achieving high accuracy is desirable but not a very good metric to evaluate the model. BLUE score on the other hand serves as a better evaluation in this work. But still it may not give best evaluation in every case, e.g in the data instance given in table 4, there may be more than one medical term indicating the same part or symptom e.g. the words 'Lung', and 'Lobe' refers to the same organ, but BLEU score decreases due to unavailability of a second reference answer.

Question	Answer
Where is the lesion located?	<ul style="list-style-type: none"> • Right lobe • Right lungs • Right lobe of the lungs

Table 4. Semantically similar answers

Before calculating the scores each predicted and ground-truth answers are converted to lower-case and then the punctuations are removed. Apart from that answer tokenization⁵ is also applied to individual words before removing the stopwords from NLTK's⁶ English stopword list. The score over the whole dataset is the average of the score per answer for all the samples.

CHAPTER 5

RESULT AND ERROR ANALYSIS

5.1 Baseline

Following the approach from [12], for effective comparison with our proposed approach, we use MCB [23] and SAN [26] trained on RAD, CLEF18, and the combined version of the previous two dataset to create the baseline models. These models were pre-trained on ResNet18 and VGGNet19 respectively for image feature extraction. Question were passed through Bi-LSTM layer to generate the textual features.

5.2 Evaluation Result

Comparison between the baseline models and our model is depicted by Table 5 for RAD, CLEF18, and CLEF18+RAD datasets respectively. Table 6 represents the results of the model without Question Segregation and the same model with Question Segregation for the stated datasets.

	RAD			CLEF18			RAD+CLEF18		
	MCB	SAN	OUR	MCB	SAN	OUR	MCB	SAN	OUR
Yes/No	0.168	0.622	0.598	0.003	0.020	0.440	0.167	0.167	0.555
Others	0.031	0.058	0.115	0.006	0.011	0.031	0.003	0.034	0.057
Overall	0.106	0.372	0.383	0.007	0.012	0.072	0.095	0.109	0.207

Table 5. BLEU score of yes/no, others and overall question.

	RAD		CLEF18		RAD+CLEF18	
	w.	w/o.	w.	w/o.	w.	w.o.
Yes/No	0.494	0.598	0.300	0.440	0.518	0.555
Others	0.098	0.115	0.018	0.031	0.056	0.057
Overall	0.318	0.383	0.460	0.072	0.202	0.207

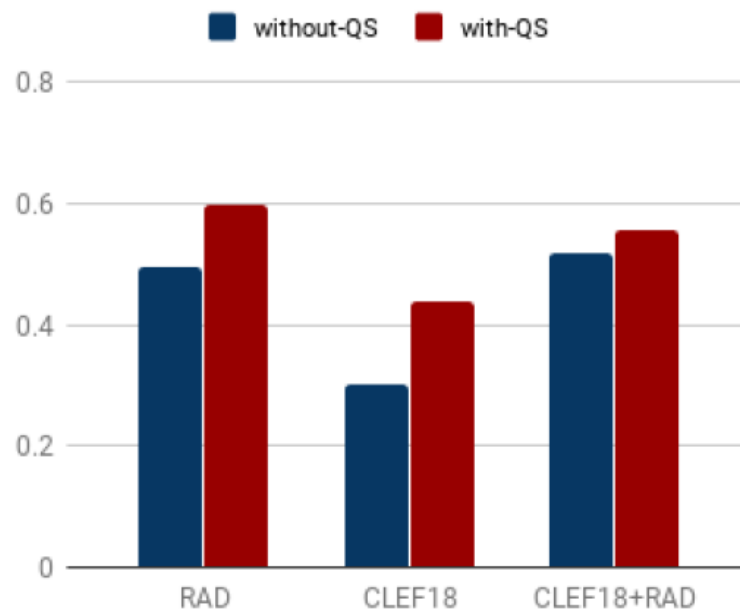
Table 6. BLEU score of the model with (w.) and without (w/o.)

Question-Segregation on different datasets.

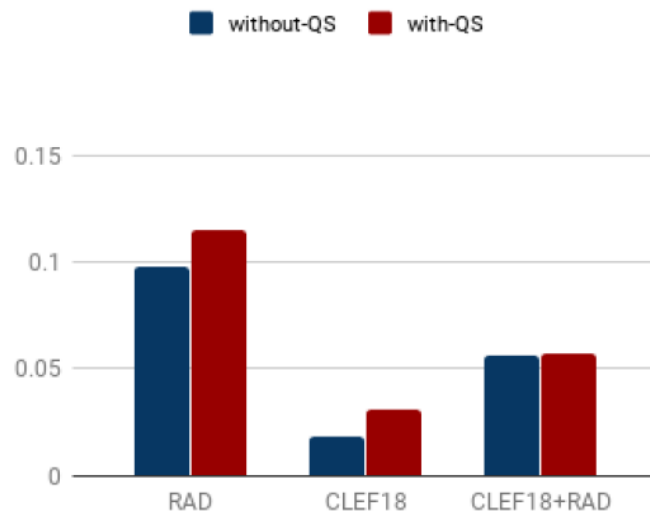
5.3 Quantitative Analysis

Table 5 shows that for answers having type others, our model outperformed the other two models. It has a difference of at least 0.06, 0.02, and 0.02 in BLEU score for RAD, CLEF18 and CLEF18+RAD dataset respectively. For yes/no type questions, our model is comparable with other models with slightest of difference i.e; maximum of 3.8% in the BLEU score when exclusively trained and tested on RAD dataset. For CLEF18 and CLEF18+RAD dataset, it beats other models having a difference not less than 0.42 and 0.39 respectively. For all the questions combined our model is superior with impressive scores.

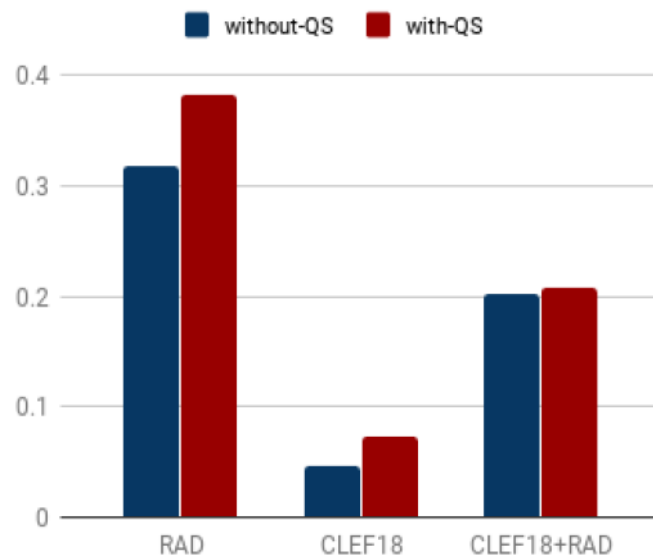
Table 6 depicts that our proposed model improves the performance of the same model without QS by significant margins regardless of the question type. The stated difference is clearly visible in table 8. It shows that other similar tasks where categorization is present can reap the benefits from our hierarchical model with QS to enhance their performance.



(a) Yes/No



(b)Others



(c)Overall

Fig 5.1. Comparison of BLEU scores of the model with/without QS on different datasets.

5.4 Qualitative Analysis

Questions having answer of type yes/no are much easy to predict than that of type others. But a model without QS will have to predict the words of the answer from the entire answer dictionary which is unnecessary. With QS the search space is reduced to only two words that is "yes" and "no" which results in better probability of correct answer prediction which is depicted by table 8

Question	GT ans.	Ans. W.	Ans w/o.
Is the GI tract is highlighted contrast?	Yes	Yes	Bilateral
Is the surrounding normal?	No	Yes	bronchiectasis

Table 7. Answer prediction of question with type yes/no by model with (w.)/without (w/o.) QS. (GT ans is the Ground Truth ans)

For other questions it is difficult to predict the answer due to the considerable size of answer vocabulary. An answer can be written in different ways with varying details or synonyms of the words. For Example, "Posterior-Anterior" , "Posteroanterior", and "PA" are correct answer for the question "how is the patient oriented?". In the absence of multiple reference answers like in general VQA, predicting the exact terms as in gold answer is challenging. Additional challenge is sequence detection, that is to predict the order in which related words follow each other. These differences for predicting the answers from particular question type is clearly visible in Table 5.

For RAD+CLEF18 dataset the model must have better score than the scores on RAD and CLEF18 dataset individually. But this is not the case due to the difference in size of the dataset and way the questions are framed and answers are generated. CLEF18 dataset is created by a semi-automatic approach while RAD dataset is manually created. The difference in quality and complexity of the generated sequence in the two datasets is clearly visible in Table 8.

	Question	Answer
RAD	What solid organ is seen on the right side of this image?	The liver
CLEF18	What shows the dilated common bile duct with a filling defect within it indicating the tumor extending?	Magnetic resonance imaging image of the Liver

Table 8. Sample question-answer pair in RAD and CLEF18 dataset

Both the questions requires liver identification but, there is a huge difference in complexity of the question as well as the answer. Due to the complications and the limited number of examples in the datasets, model fails to learn efficiently.

5.5 Conclusion

In our work we propose a hierarchical multimodal approach to tackle the VQA problem in medical domain. In particular we use a Question Segregation (QS) module at the top level of our hierarchy to divide the input questions into two different types (Yes/No, and others), followed by individual and independent models at the leaf level, each dedicated to the type of question segregated at the previous level. Proposed approach can be applied to any related problem where such segregation is possible but, it does require non-trivial changes in the architecture. We use SVM for QS but based on the requirements more rigorous QS techniques can be implemented. To evaluate the usefulness of our proposed hierarchy we conduct our experiments on two different datasets, RAD [12] and CLEF18 [13]. We also perform experiments on the combined 533data of the above two datasets to show the generalisability of our approach. Models when trained with the proposed hierarchy with QS scored better, outperforming all the baseline models. It suggests that questions with different types learn better in isolation having their individual learning paths. Experimental results indicates the effectiveness of our work, depicting it's value for the VQA in medical domain.

Our analysis of results showcase that the evaluation metric needs improvement while evaluating VQA in the medical domain. Particularly, BLUE score fails to consider semantic similarity of medical terms, answers of varying length, which are important while correctly evaluating any machine learning task in medical domain. For future work we plan to introduce better individual models for handling each of the leaf level problems. We also plan to use a better evaluation strategy for evaluating the task apart from devising a detailed schemes for QS.

REFERENCES

- [1] Lin TY, RoyChowdhury A, Maji S. Bilinear convolutional neural networks for fine-grained visual recognition. *IEEE transactions on pattern analysis and machine intelligence*. 2017;40(6):1309{1322.
- [2] Lee JY, Deroncourt F. Sequential short-text classification with recurrent and convolutional neural networks. *arXiv preprint arXiv:160303827*. 2016;.
- [3] Yin W, Ebert S, Schutze H. Attention-based convolutional neural network for machine comprehension. *arXiv preprint arXiv:160204341*. 2016;.
- [4] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*; 2016. p. 779{788.
- [5] Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. In: *Advances in neural information processing systems*; 2015. p. 91{99.
- [6] Girshick R, Donahue J, Darrell T, Malik J. Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence*. 2015;38(1):142{158.
- [7] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*; 2012. p. 1097{1105.
- [8] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:14091556*. 2014;.
- [9] Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Lawrence Zitnick C, et al. Vqa: Visual question answering. In: *Proceedings of the IEEE international conference on computer vision*; 2015. p. 2425{2433.
- [10] Malinowski M, Rohrbach M, Fritz M. Ask your neurons: A neural-based approach to answering questions about images. In: *Proceedings of the IEEE international conference on computer vision*; 2015. p. 1{9.
- [11] Cid YD, Liauchuk V, Kovalev V, Muller H. Overview of ImageCLEFtuberculosis 2018-Detecting multi-drug resistance, classifying tuberculosis type, and assessing severity score. In: *CLEF2018 Working Notes, CEUR Workshop Avignon, France*; 2018.

- [12] Lau JJ, Gayen S, Abacha AB, Demner-Fushman D. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*. 2018;5:180251.
- [13] Ionescu B, Müller H, Villegas M, de Herrera AGS, Eickhoff C, Andrearczyk V, et al. Overview of ImageCLEF 2018: Challenges, datasets and evaluation. In: *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer; 2018. p. 309{334.
- [14] Malinowski M, Fritz M. A multi-world approach to question answering about real-world scenes based on uncertain input. In: *Advances in neural information processing systems*; 2014. p. 1682{1690.
- [15] Silberman N, Hoiem D, Kohli P, Fergus R. Indoor segmentation and support inference from rgb-d images. In: *European Conference on Computer Vision*. Springer; 2012. p. 746{760.
- [16] Ren M, Kiros R, Zemel R. Image question answering: A visual semantic embedding model and a new dataset. *Proc Advances in Neural Inf Process Syst*. 2015;1(2):5.
- [17] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft coco: Common objects in context. In: *European conference on computer vision*. Springer; 2014. p. 740{755.
- [18] Zhu Y, Groth O, Bernstein M, Fei-Fei L. Visual7w: Grounded question answering in images. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*; 2016. p. 4995{5004.
- [19] Yu L, Park E, Berg AC, Berg TL. Visual madlibs: Fill in the blank description generation and question answering. In: *Proceedings of the IEEE international conference on computer vision*; 2015. p. 2461{2469.
- [20] Johnson J, Hariharan B, van der Maaten L, Fei-Fei L, Lawrence Zitnick C, Girshick R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2017. p. 2901{2910.
- [21] Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*. 2017;123(1):32{73.
- [22] Zhou B, Tian Y, Sukhbaatar S, Szlam A, Fergus R. Simple baseline for visual question answering. *arXiv preprint arXiv:151202167*. 2015;.
- [23] Fukui A, Park DH, Yang D, Rohrbach A, Darrell T, Rohrbach M. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:160601847*. 2016;.

- [24] Kim JH, Lee SW, Kwak D, Heo MO, Kim J, Ha JW, et al. Multimodal residual learning for visual qa. In: Advances in neural information processing systems; 2016. p. 361{369.
- [25] Xu H, Saenko K. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In: European Conference on Computer Vision. Springer; 2016. p. 451{466.
- [26] Yang Z, He X, Gao J, Deng L, Smola A. Stacked attention networks for image question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 21{29.
- [27] Ilievski I, Yan S, Feng J. A Focused Dynamic Attention Model for Visual Question Answering. CoRR. 2016;abs/1604.01485.
- [28] Shih KJ, Singh S, Hoiem D. Where to Look: Focus Regions for Visual Question Answering. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. p. 4613{4621.
- [29] Gao H, Mao J, Zhou J, Huang Z, Wang L, Xu W. Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2. NIPS'15. Cambridge, MA, USA: MIT Press; 2015. p. 2296{2304. Available from: <http://dl.acm.org/citation.cfm?id=2969442.2969496>.
- [30] Elman JL. Finding structure in time. Cognitive science. 1990;14(2):179{211.
- [31] Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Comput. 1997;9(8):1735{1780. doi:10.1162/neco.1997.9.8.1735.
- [32] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural networks : the official journal of the International Neural Network Society. 2005;18:602{10. doi:10.1016/j.neunet.2005.06.042.
- [33] Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning Phrase Representations using RNN Encoder{Decoder for Statistical Machine Translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics; 2014. p. 1724{1734. Available from: <https://www.aclweb.org/anthology/D14-1179>.
- [34] LeCun Y, Bengio Y. Convolutional networks for images, speech, and time series; 1998.
- [35] Wang L, Guo S, Huang W, Qiao Y. Places205-vggnet models for scene recognition. arXiv preprint arXiv:150801667. 2015;

- [36] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770{778.
- [37] Ren M, Kiros R, Zemel R. Exploring models and data for image question answering. In: Advances in neural information processing systems; 2015. p. 2953{2961.
- [38] Gao H, Mao J, Zhou J, Huang Z, Wang L, Xu W. Are you talking to a machine? dataset and methods for multilingual image question. In: Advances in neural information processing systems; 2015. p. 2296{2304.
- [39] Xiong C, Merity S, Socher R. Dynamic Memory Networks for Visual Question Answering. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48. ICML'16. JMLR.org; 2016. p. 2397{2406. Available from: <http://dl.acm.org/citation.cfm?id=3045390.3045643>.
- [40] Lu J, Yang J, Batra D, Parikh D. Hierarchical question-image co-attention for visual question answering. In: Advances In Neural Information Processing Systems; 2016. p. 289{297.
- [41] Nam H, Ha J, Kim J. Dual Attention Networks for Multimodal Reasoning and Matching. CoRR. 2016;abs/1611.00471.
- [42] Kim J, On KW, Lim W, Kim J, Ha J, Zhang B. Hadamard Product for Low-rank Bilinear Pooling. CoRR. 2016;abs/1610.04325.
- [43] Yu Z, Yu J, Fan J, Tao D. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 1821{1830.
- [44] Yu Z, Yu J, Xiang C, Fan J, Tao D. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. IEEE transactions on neural networks and learning systems. 2018;(99):1{13.
- [45] Kawahara J, Hamarneh G. Multi-resolution-Tract CNN with Hybrid Pretrained and Skin-Lesion Trained Layers. In: MLMI@MICCAI; 2016.
- [46] van Tulder G, de Bruijne M. Combining Generative and Discriminative Representation Learning for Lung CT Analysis With Convolutional Restricted Boltzmann Machines. IEEE Transactions on Medical Imaging. 2016;35(5):1262{1272. doi:10.1109/TMI.2016.2526687.
- [47] Tarando SR, Fetita C, Faccinnetto A, Brillet PY. Increasing CAD system efficacy for lung texture analysis using a convolutional network. In: Medical Imaging 2016: Computer-Aided Diagnosis. vol. 9785. International Society for Optics and Photonics; 2016. p. 97850Q.

- [48] Kafle K, Kanan C. Answer-Type Prediction for Visual Question Answering. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. p. 4976{4984.
- [49] Eickhoff C, Schwall I, de Herrera AGS, Muller H. Overview of ImageCLEFcaption 2017-Image Caption Prediction and Concept Detection for Biomedical Images. In: CLEF (Working Notes); 2017.
- [50] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. In: 2014 conference on empirical methods in natural language processing (EMNLP); 2014. p. 1532{1543.
- [51] Ghannay S, Favre B, Esteve Y, Camelin N. Word embedding evaluation and combination. In: LREC; 2016. p. 300{305.
- [52] Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-First AAAI Conference on Artificial Intelligence; 2017.
- [53] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Ieee; 2009. p. 248{255.
- [54] Torrey L, Shavlik J. Transfer learning. In: Handbook of research on machine learning applications and trends: algorithms, methods, and techniques. IGI Global; 2010. p. 242{264.
- [55] Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, et al. Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE transactions on medical imaging. 2016;35(5):1299{1312.
- [56] Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: A Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. ACL '02. Stroudsburg, PA, USA: Association for Computational Linguistics; 2002. p. 311{318. Available from: <https://doi.org/10.3115/1073083.1073135>.