

# Hierarchical Deep Multi-modal Network for Medical Visual Question Answering

Deepak Gupta<sup>1,\*</sup>, Swati Suman<sup>1</sup>, Asif Ekbal

*Department of Computer Science and Engineering  
Indian Institute of Technology Patna, India*

---

## Abstract

Visual Question Answering in Medical domain (VQA-Med) plays an important role in providing medical assistance to the end-users. These users are expected to raise either a straightforward question with a *Yes/No* answer or a challenging question that requires a detailed and descriptive answer. The existing techniques in VQA-Med fail to distinguish between the different question types sometimes complicates the simpler problems, or over-simplifies the complicated ones. It is certainly true that for different question types, several distinct systems can lead to confusion and discomfort for the end-users. To address this issue, we propose a hierarchical deep multi-modal network that analyzes and classifies end-user questions/queries and then incorporates a query-specific approach for answer prediction. We refer our proposed approach as **Hierarchical Question Segregation based Visual Question Answering**, in short HQS-VQA. Our contributions are three-fold, *viz.* firstly, we propose a question segregation (QS) technique for VQA-Med; secondly, we integrate the QS model to the hierarchical deep multi-modal neural network to generate proper answers to the queries related to medical images; and thirdly, we study the impact of QS in Medical-VQA by comparing the performance of the proposed model with QS and a model without QS. We evaluate the performance of our proposed model on two benchmark datasets, *viz.*

---

\*Corresponding author

Email addresses: [deepak.pcs16@iitp.ac.in](mailto:deepak.pcs16@iitp.ac.in) (Deepak Gupta), [swati17293@gmail.com](mailto:swati17293@gmail.com) (Swati Suman), [asif@iitp.ac.in](mailto:asif@iitp.ac.in) (Asif Ekbal)

<sup>1</sup>Both authors contributed equally to this work.

RAD and CLEF18. Experimental results show that our proposed HQS-VQA technique outperforms the baseline models with significant margins. We also conduct a detailed quantitative and qualitative analysis of the obtained results and discover potential causes of errors and their solutions.

*Keywords:* Visual Question Answering, Neural Networks, Medical Domain, Support Vector Machine, Gated Recurrent Units

---

## 1. Introduction

The advancement in the field of Computer Vision (CV) (Arai & Kapoor, 2019; Guo et al., 2020; Li et al., 2020) and Natural Language Processing (NLP) (Ruder, 2019; Ruder et al., 2019; Fu, 2019; Dong et al., 2019) over the last decade, has introduced several interesting machine learning techniques. The problems such as object detection (Liu et al., 2020), segmentation (Liu et al., 2019), and image classification (Sun et al., 2020a,b) in CV, and machine translation (Yang et al., 2020), question answering (Gupta et al., 2018b, 2019; Chen et al., 2016; Gupta et al., 2018a,c), biomedical and clinical text mining (Ningthoujam et al., 2019; Yadav et al., 2018, 2019, 2020; Chen et al., 2019), speech recognition (Magnuson et al., 2020) in NLP, are being solved much more efficiently than ever before. This has facilitated the researchers to indulge into solving interdisciplinary problems that demand knowledge of both the fields.

Visual Question Answering (VQA) (Gao et al., 2019; Kafle et al., 2020) has emerged as one such problem. In VQA, the task is poised as questions being asked with respect to an image, where the machine needs to learn and generate answers of such questions based on the learned features of the input image. In contrast to the typical CV tasks which largely focus on solving problems such as *action identification*, *image classification*, VQA tasks are relatively complex. It demands more intelligence like object recognition, semantic feature extraction, external knowledge, and common sense knowledge. Many domain-specific VQA tasks have surfaced in the last few years, and VQA in the medical domain is

one such that plays a significant role in providing medical assistance. Since this task is related to the medical domain, end-users can be categorized based on the type of queries raised by them. Patients, medical students, and related users are anticipated to ask elementary questions mostly having *Yes/No* as the answer. On the other hand, clinicians and medical experts are expected to raise a more problem-specific query demanding a detailed and descriptive answer. For this reason, different portals must be created to satisfy the query-specific need, but that would lead to confusion and discomfort to end-users. To tackle this problem, in this paper, we propose a question segregation technique to segregate the user queries. For this module, we use a simple statistical machine learning model, based on simple hand-engineered, and word frequency-based features.

Towards the solution of the problem of generic VQA-Med, we propose a hierarchical deep multi-modal network that analyzes and classifies the queries at the root of the hierarchy and then incorporates the query-specific approach at leaf nodes. To predict the answers in this model, we generate the question and image representations using Bidirectional Long Short Term Memory (Bi-LSTM) (Graves & Schmidhuber, 2005) and Inception-Resnet-v2 (Szegedy et al., 2017), respectively. We fuse the representations together and pass it to the specific answer prediction model at the leaf node. For the task of question classification in the root node, we propose a question segregation technique. We use Support Vector Machine (SVM) (Cristianini et al., 2000) as the classifier with hand-engineered and word frequency-based features for QS. We use the machine learning technique for QS, as the rule-based strategy suffers from the problem of defining too many rules that may not extend to other datasets (Clark et al., 2018). The following examples from RAdiology Data (RAD) (Lau et al., 2018) show the difficulty of the rule-based approach in the medical domain.

- **Question:** Evidence of hemorrhage in the kidneys?

**Answer:** No

**Question-type:** ‘*Yes/No*’

- **Question:** Is the spleen present?

**Answer:** on the patient’s left

**Question-type:** ‘*Others*’

Careful analysis of the question reveals that the first example expects a descriptive type answer that is to list out the facts that indicate kidney hemorrhage (Question-type: ‘*Others*’), while the second example expects to confirm the presence/absence of spleen (Question-type: ‘*Yes/No*’). The presence of such anomalies in the question acts as a hindrance in the formation of robust rules for the classification of questions into their correct type.

We perform all our experiments in the RAD and ImageCLEF2018 VQA-Med 2018 (CLEF18) datasets, as they perfectly capture the problem statement that we intend to solve. Detailed discussion on the dataset can be found in Section 4. Experimental evaluation demonstrates promising results, showing the effectiveness of our proposed approach. Additionally, error analysis of the system’s outputs shows the future direction in this research area by addressing different kinds of errors.

The organization of this paper is as follows. We first discuss the related work in VQA. Then we present the details of the methodologies that we implemented to solve our specific problem. In particular, we explain our proposed HQS-VQA models in detail. Basically, we discussed the technique used for the question segregation module and the VQA components used to generate the query-specific answers. Details of experiments along with the evaluation results and necessary analysis are reported. Finally, we conclude and provide the future directions of our work.

### 1.1. Motivation

The motivation behind our work are stemmed from the following facts:

- Nowadays, medical and physiological images (“ct-scan”, “x-ray”, etc.) and reports for the patients are easily accessible with the increase in the use of medical portals. But the patient still needs to visit a medical expert to fully understand those reports and get answers to their queries. This process is both time-consuming and cost-sensitive. On the other hand, clinicians also find an efficient VQA system very useful to understand the results of a complex medical image. They may use such a system as a second opinion just to boost their self-confidence in the understanding of some specific aspects of such medical images. Although it is possible to search queries on search engines, the search results may be inaccurate, spurious, vague, and, or enormous. In the case of medical reports such inaccurate, spurious, or vague results could lead to serious after-effects. In this context, the VQA in the medical domain is getting attention as an important research problem trying to provide the answer to end-user queries related to medical images.
- VQA-Med intends to assist patients and clinicians in general, but can also be useful in medical education. A clinical apprentice or medical students who just started learning the basics of handling images of different modalities, may learn by asking queries and getting answers. Thus, developing an efficient and automated VQA system for the medical domain comes out as an essential task. Even though many medical datasets are published publicly, most of them deal with some specific disease in a particular body part with a fixed image modality. ImageCLEFtuberculosis task (Cid et al., 2018) 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, it is a more challenging problem to have a generic strategy that deals with VQA queries regarding multiple image modalities linked to multiple diseases that may appear in any part of the body. The solution to such a generic problem will be considerably less confusing to the patients.

- The difference in the type of end-users results in different query types. The queries from patients most of the time are expected to be generic in nature, having the need to produce ‘Yes’ or ‘No’ as the answer. Whereas, the queries from clinicians and medical experts are expected to be more problem-specific, which requires elaborate answers. Again, a skilled trainee is expected to ask more specific and sophisticated questions, 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 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 effective treatment.
- This difference in query type thus needs different problem-specific attention, which needs to be dealt with isolation. Again, multiple end systems for multiple types of queries may create confusion, and discomfort to the 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 to 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 7th rib in the right lung
	What are the densities in both mid-lung fields?	pleural plaques

**Table 1.** Sample, question-answer pairs formulated from a single image. More than one clinically relevant questions can be asked from a given image.

- We identify this need, and propose a SVM-based question segregation

technique to segregate the questions. We then use this information to propose a hierarchical deep multi-modal network to generate the answers.

### *1.2. Contributions*

- We propose a SVM based Question Segregation technique for the task of question classification for VQA in the medical domain.
- We propose a hierarchical deep multi-modal neural model and integrate it with the proposed question segregation module to generate proper answers to the queries related to medical images.
- We study the impact of QS in Medical-VQA by comparing the performance of the proposed model with question segregation and a model without such segregation. We also compare it with the baseline models to study its effectiveness.
- We evaluate our model on two different datasets, which demonstrate the fact that our proposed method is generic in nature.

## **2. Related Work**

The major challenges of the VQA-Med are closely related to the general VQA and QA in the medical domain and we see a lot of interesting solutions evolving over time. We present the survey with respect to the related datasets and methods in the following subsections.

### *2.1. Datasets*

A number of research projects have been initiated for the development of benchmark datasets to promote the works in the medical domain. The Genomic corpus released as part of the TREC (Hersh & Bhupatiraju, 2003) task is one

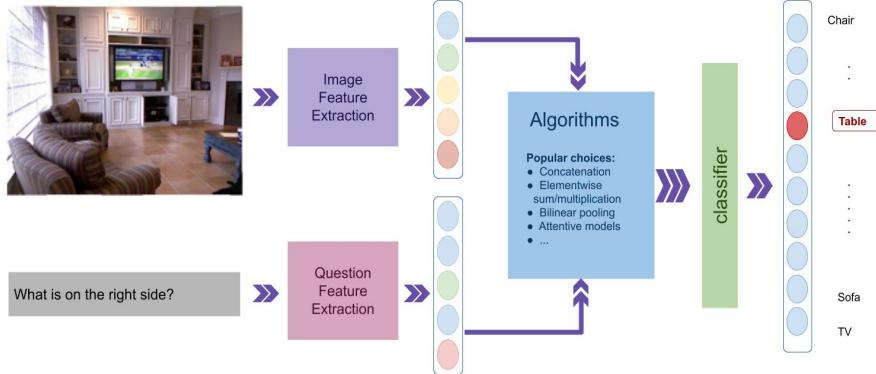
of the benchmark datasets for the medical QA task. It focuses exclusively on scientific papers. However, a small number of questions in the dataset are not sufficient to evaluate the efficiency of the large-scale QA systems. This constraint led to the release of several other datasets, such as Question Answering for Machine Reading Evaluation (QA4MRE) (Morante et al., 2013) and Biomedical Semantic Indexing and Question Answering (BioASQ) (Tsatsaronis et al., 2015). The QA4MRE consists of the Biomedical Text on Alzheimer’s Disease, while BioASQ gathers information from various heterogeneous sources to address real-life questions from the biomedical experts. A number of datasets, such as MRI-DIR (Ger et al., 2018), fastMRI (Zbontar et al., 2018), and a few more (Bradley et al., 2017; Vallieres et al., 2017; Shaimaa et al., 2017) focused on different medical tasks, are also available. However, the images in the VQA-Med dataset have different modalities and contain radiological markings such as short information, tags, etc. It may also contain a stack of sub-images that is not the case with the existing medical datasets. In addition, general VQA datasets (Lin et al., 2014; Mukoze et al., 2018; Gebhardt & Wolf, 2018; Antol et al., 2015) are task-specific, unlike VQA-Med, where a question can be asked about any disease from any part of the body.

In this work, we use the RAD (Lau et al., 2018) and CLEF18 (Ionescu et al., 2018) medical VQA datasets, which are different from the existing VQA datasets. The obvious reason is their focus on the medical domain, which offers distinguishing 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 the VQA datasets.

## 2.2. Methods

VQA tasks are primarily based 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 1 illustrates this framework where the key components can take a wide variety of forms.



**Fig 1.** Framework for VQA where, question and image are taken as input to generate or predict the answer.

VQA systems differ from each other in the way they fuse multi-modal information. Although most open-ended VQA algorithms used the classification mechanism, this strategy can only produce answers seen during training. The multi-word response is generated one word at a time using LSTM (Gao et al., 2015; Malinowski et al., 2015). The response generated, however, is still restricted to words seen in the course of training.

For question encoding, most methods for VQA uses a variant of Recurrent Neural Network (RNN) (Mikolov et al., 2010). RNNs are capable of handling sequence problems, but when RNN processes lengthy sequences, context data is easily ignored. LSTM's (Greff et al., 2016) 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 Bi-LSTM (Graves & Schmidhuber, 2005) model combines the above two points and improves the results. The Gated Recurrent Unit (GRU) (Cho et al., 2014) is also notable, and widely used, simplification of the LSTM. As for the image feature extraction, Convolutional Neural Network (CNN) (Traore et al., 2018) are used where VGG-

net (Simonyan & Zisserman, 2015) and deep residual networks (ResNet) (He et al., 2016) are the most popular choice.

The application of attention on the image can help to improve the performance of the model by discarding the irrelevant parts of the image. So, attention mechanisms (Xiong et al., 2016; Yang et al., 2016) 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 (Lu et al., 2016). 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) (Yang et al., 2016) and the Dynamic Memory Network (DMN) (Xiong et al., 2016) used image features from a CNN feature map's spatial grid. In (Yang et al., 2016) 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.

VQA depends on the image and question being processed 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. Later, multi-modal bi-linear pooling was proposed 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 (Lin et al., 2017). Multimodal Compact Bilinear (MCB) (Fukui et al., 2016) is the most significant VQA technique used in bilinear pooling. It 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. The major change was the use of MCB for feature fusion instead of element-wise multiplication.

Methods for Medical-VQA must be different from general VQA as the size of the datasets is incomparable. The other challenge with Medical-VQA 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 (Kawahara & Hamarneh, 2016; van Tulder & de Bruijne, 2016; Tarando et al., 2016). Most of them, however, deal with single disease problems and focused primarily on easily identifiable areas such as the lungs and skin. In contrast to these systems, Medical-VQA deals with multiple diseases at the same time apart from handling multiple body parts which are difficult for machines to learn.

Recently, the ImageCLEF introduced the challenge of Medical Domain Visual Question Answering, VQA-Med 2018<sup>2</sup> (Ionescu et al., 2018). The system submitted by Peng et al. (2018) achieved the best performance (in terms of BLEU score) in VQA-Med 2018 for medical visual question answering. They built their best performing systems using ResNet-152 for image feature extraction and Multimodal Factorized High-order (MFH) Yu et al. (2018) for language-vision fusion. Zhou et al. (2018) utilized the Inception-Resnet-v2 and Bi-LSTM for image and question representation, respectively. They used the inter-attention mechanism to fuse the language and vision features. Their best performing system stood second among all the submitted systems in the challenge. The third best system submitted by Abacha et al. (2018) uses the pre-trained VGG-16

---

<sup>2</sup><https://www.imageclef.org/2018/VQA-Med>

model for image feature extraction and LSTM for question representation. They utilized the stack attention network to fuse the question and image features. In the second edition<sup>3</sup> of VQA-Med, Yan et al. (2019) submitted the best system for medical visual question answering. The proposed approach utilized the BERT (Devlin et al., 2018) for question representation and pre-trained VGG-16 model for image representation. They fused the question and image features using MFB mechanism.

Inherently, questions follow a temporal sequence and naturally cluster into different types. This question-type information is very important to predict the response regardless of the image. Authors in (Kafle & Kanan, 2016) use a 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 a feature. This type of information can also affect 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 a question segregation module which segregates the learning path based on the question types (*Yes/No* and *Others*) to reap the benefits of question-type dedicated models. We use Inception-Resnet to encode image feature and Bi-LSTM for question feature creation.

### 3. Materials and Methods

#### 3.1. Problem Modeling

Given a pair  $(Q, I)$ , where  $Q$  is the textual question accompanied by any medically relevant image  $I$ , the Medical-VQA task is aimed to generate the

---

<sup>3</sup><https://www.imageclef.org/2019/medical/vqa/>

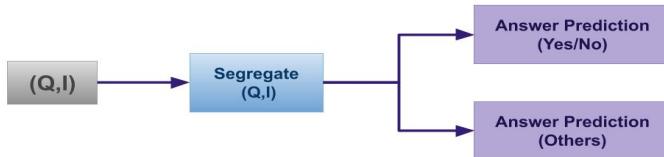
appropriate answer  $A$ . Mathematically, it can be formulated as,

$$A = f(Q, I, \alpha) \quad (1)$$

where  $f$  is the answer prediction function and  $\alpha$  denotes the model parameters. Questions in  $Q$  can be categorized into two question types ( $q\_type$ ). For questions with  $q\_type = Yes/No$ , the input  $Q$  have a straightforward binary response. While for questions with  $q\_type = Others$ , it can have a well-thought-out variable length response generated from the answer dictionary words. The problem is to develop a hierarchical model with a question segregation module to differentiate the learning path for the two  $q\_type$  for solving the generic Medical-VQA task.

### 3.2. Methodology

Our proposed approach towards the solution of the problem is to form a two-level hierarchy, where the top-level task is question segregation and the next-level task is answer prediction. The proposed hierarchical model is depicted in Fig 2. The subsequent sections describe the components of the proposed hierarchy.

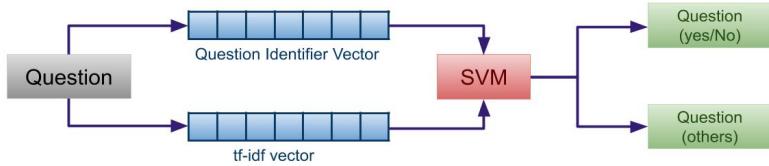


**Fig 2.** Abstract representation of the proposed hierarchical HQS-VQA model. The first level segregates the questions while the second level generates the answer using the leaf node. Answer prediction strategy is decided based on the question type.

#### 3.2.1. Question Segregation

Question Segregation, in general, segregates the questions from a question list  $Q = [q_1, q_2 \dots q_n]$  based on the  $q\_type$ , where  $n$  denotes the total number of questions. In this list,  $q_i = "w_1 w_2 \dots w_m"$  denotes the  $i^{th}$  question in

the list containing a sequence of  $m$  words. The task of question segregation is relatively an easier task compared to answer prediction. Thus, we find a simple statistical machine 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 hierarchical setup. We employ an SVM classifier for this purpose. The SVM is relatively less complex compared to the deep neural network. Moreover, the datasets, which we use here are relatively smaller, and hence deep learning models tend to perform on the lower-side compared to the classical supervised SVM based model. Fig 3 illustrates the entire question segregation process.



**Fig 3.** Proposed question segregation module with linear SVM learner as base classifier. The extracted feature vectors are fed to the SVM for question segregation.

The classifier, and input to QS module i.e. question feature vectors generated from the questions in the dataset are explained as follows:

**Question Feature Vectors:** From each question, we extract the following two vectors:

- **Question Identifier Vector:** We form a set of  $r$  question identifier words, where each word tries to represent the question motive. We then convert every question into a question vector  $V = [v_1, v_2 \dots v_r]$ , where  $v_i \in \{1, 0\}$  indicates the presence or absence of the  $i^{th}$  question identifier word in the question.
- **Tf-idf Vector:** We use tf-idf (term frequency - inverse document frequency) to assess how significant  $w_j$  is to a  $q_i$  in  $Q$ . It can be calculated

as:

$$tf\text{-}idf(w_j, q_i) = \frac{f(w_j, q_i)}{\sum_{j=1}^m f(w_j, q_i)} * \log_e \frac{n}{\sum_{i=1}^n f(w_j, q_i)} \quad (2)$$

where  $f(w_j, q_i)$  is the frequency of  $w_j$  in  $q_i$ ,  $n$  is the number of words in question  $q_i$ . From the entire vocabulary, we consider only top  $m'$  words with the highest tf-idf values. We then convert every  $q_i$  in the list Q into tf-idf vector such that position of every  $w_j$  in  $q_i$  is represented by  $tf\text{-}idf(w_j, q_i)$ .

We concatenate both the feature vectors to represent a question.

**Question Classifier:** The SVM (Cortes & Vapnik, 1995; Cristianini et al., 2000) is a statistical classification technique and inspired by its performance in (Wang et al., 2018; Zhi et al., 2018). we use SVM learner as the base classifier. It takes the question feature vectors as input during the training stage to segregate the questions according to its type. It is a linear function which can be represented as,

$$f(v_i) = \langle v_i, w^T \rangle + b, \quad \text{where } \langle v_i, w^T \rangle = \|v_i\| \|w^T\| \cos(\theta) \quad (3)$$

where  $v_i$ ,  $w^T$ , and  $b$  are feature vector of  $i^{th}$  question, weight, and bias respectively. So,  $\forall i; i \in [1, n]$ , either of the following equations can be true based on  $q\_type$ .

$$\langle v_i, w^T \rangle + b \geq 1, \quad \text{if}(q\_type == Yes/No) \quad (4)$$

or

$$\langle v_i, w^T \rangle + b \leq -1, \quad \text{if}(q\_type == Others) \quad (5)$$

We use the SVM with linear kernel, and use ‘*hinge*’ as the loss function. The hinge loss  $\ell(y)$  of the prediction  $y = f(v_i)$  for a true  $q\_type$   $t$  and a classifier score  $y$  is defined as,

$$\ell(y) = \max(0, 1 - t * f(v_i)) \quad (6)$$

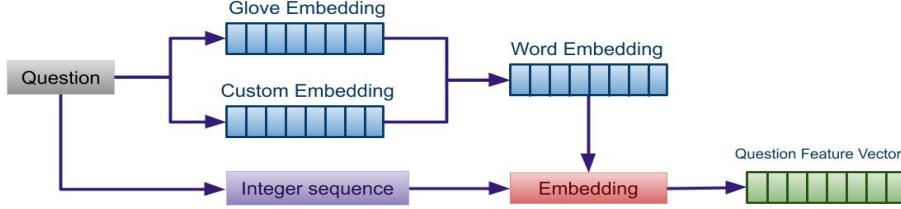
### 3.2.2. Answer Prediction

Answer prediction component is at the second level of our hierarchy, where two separate models  $M = \{m_1, m_2\}$  deal with the problem of answer generation at the lowest level nodes, i.e. leaf nodes. While  $m_1 = Yes/No$  deals with the problem of producing simple *Yes/No* answers from simple incoming queries at the first leaf node,  $m_2 = Others$ , on the other hand, deals with complex queries to produce other expertise answers at the second leaf node. We extract the question and image feature vectors which are then fused together. Based on  $q\_type$  this model passes the fused vector through several layers to finally generate the answer. We outline these tasks with more details in the following subsections.

**Question Feature Extraction:** We first pre-process the questions by converting the words into lowercase and then lemmatizing them to reduce the ambiguity among their different forms. Next we remove 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 the complexity of information in questions. We then generate the integer sequence from the pre-processed questions that are finally fed to the embedding layer together with the word embedding for the extraction of question feature ( $F^Q$ ) of dimension  $m \times d$  where  $m$  is the total number of words in the question and  $d$  denotes the dimension of the word embedding vector. Fig 4 illustrates the entire process.

The components of the question feature extraction process are described below:

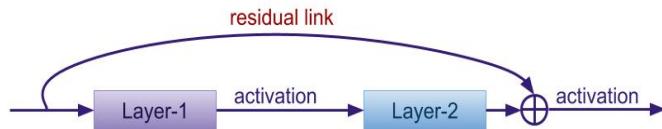
- **Word embedding:** We use word embedding to vectorize words to capture their meaning. For this we first generate  $d_1$  dimensional vectors  $G = [g_1, g_2 \dots g_{d_1}]$  using the GloVe (Pennington et al., 2014) vector. We also introduce the sub-word embedding to capture the embedding of unknown word in medical terminology. For sub-word embedding, we follow the



**Fig 4.** Flowchart of generating question embedding. The word embedding is the concatenation of GloVe and Custom (sub-word) embedding, which are used along with the integer sequence representation of the questions by the embedding layer.

(Bojanowski et al., 2017) work on FastText vector and generate the subword embedding vector of dimension  $d_2$  as  $C = [c_1, c_2 \dots c_{d_2}]$ . We next concatenate the embeddings to create the final  $d$ -dimensional word embedding vector  $E = [g_1, g_2 \dots g_{d_1}, c_1, c_2 \dots c_{d_2}]$ .

**Image Feature Extraction:** For image feature extraction, we use the Inception-Resnet-v2 model. It is a type of advanced CNN that integrates the inception module (Szegedy et al., 2014) with ResNet. Inception enables one to accomplish a very good performance at comparatively low computational costs, while residual connections considerably speed up network training by enabling connection shortcuts. Together they allow the development of deeper and wider networks in the inception-resnet-v2 model. Basically, the network utilizes residual links (He et al., 2016) (Fig 5) to combine filters of varying dimensions, which not only prevent the issue of degradation caused by deep structures but also decreases the training cost.

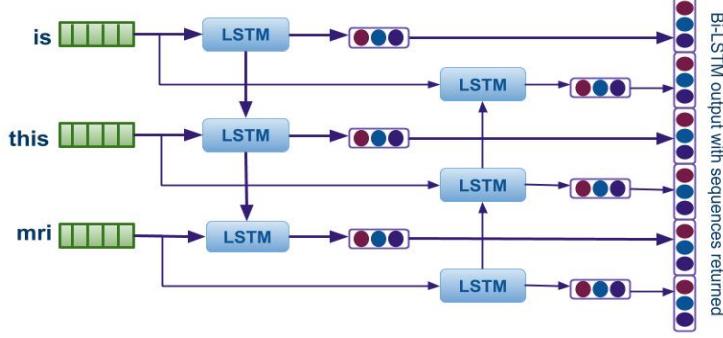


**Fig 5.** Canonical form of a 2-layer ResNet block. Layer-2 is skipped over activation from layer-1 using residual link.

Since the model expects an input of dimension  $299 \times 299$ , we re-size the input images to suit that dimension. We then initialize the model with weights pre-trained on ImageNet (Russakovsky et al., 2015). Such initialization facilitates transfer learning (Long et al., 2017), 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 a comparatively smaller dataset. 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 (Tajbakhsh et al., 2016). Activations are obtained from the model’s last fully connected layer as it represents the detected features of the medical image ( $F^I$ ). The features generated are of size 1000.

**Question and Image Feature Fusion:** Unidirectional LSTM layer helps to capture the sequence information in the question, but it can only retain prior information as it has only seen the past inputs. In bidirectional layer inputs run bidirectionally in two ways, one from the past to the future and vice-versa. Therefore, before bi-modal feature fusion, we feed the extracted question feature vector  $F^Q$  to the bidirectional layer with LSTM as input for the recurrent instance. The process to generate the representation of question by Bi-LSTM is depicted in Fig 6.

It helps to preserve the information from both past and future. We use it 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, such that  $\vec{H} = [\vec{h}_1, \vec{h}_2 \dots \vec{h}_m]$ , and  $\overleftarrow{H} = [\overleftarrow{h}_m, \overleftarrow{h}_{m-1} \dots \overleftarrow{h}_1]$ . Here, for forward and backward directions,  $\vec{H}$ , and  $\overleftarrow{H}$  are the sequence of hidden state outputs, while  $\vec{h}_i$ , and  $\overleftarrow{h}_i$  are the hidden state outputs at  $i^{th}$  time-step. The final Bi-LSTM output is then  $\leftrightarrow{H} = [\leftrightarrow{h}_1, \leftrightarrow{h}_2 \dots \leftrightarrow{h}_m]$ , with  $\leftrightarrow{h}_i = \vec{h}_i \odot \overleftarrow{h}_{m-i+1}; \forall i \in [1, m]$ , where  $\odot$  denotes the concatenate operator. To minimize the problem of overfitting due to small amount of training data we also used dropout value of 50% in the this layer.

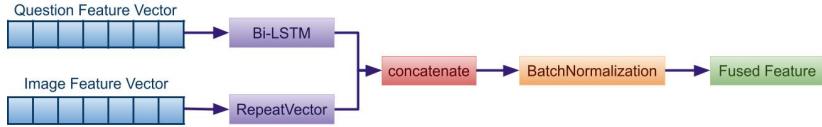


**Fig 6.** The processing of a question by Bi-LSTM to get the question representation. The word embedding of each word in the question is fed to a Bi-LSTM network, and the forward and backward hidden state outputs are concatenated at each time-step to get the final representation of the question.

We feed  $F^I$  to the RepeatVector<sup>4</sup> layer to make it's dimension same as  $F^Q$  for the modeling convenience. We then concatenate the repeated  $F^I$ , and  $\overset{\leftrightarrow}{H}$  for fusion. We finally feed the output to Batch-Normalization(Ioffe & Szegedy, 2015) layer for regularization and to increase the stability of the network. Thus, the normalized fused feature ( $F$ ) is:

$$F = BN((\text{Bidirectional}(LSTM(F^Q))) \oplus (\text{RepeatVector}(m)(F^I))) \quad (7)$$

where,  $BN$  and  $\oplus$  represents Batch-Normalization and concatenation respectively. The process of feature fusion is illustrated in Fig 7.



**Fig 7.** The architecture represents the fusion of the two modalities' feature vectors using concatenate layer, the output of which is normalized to generate the fused feature vector.

**Answer Prediction - Yes/No:** We treat model  $m_1$  (c.f. Section 3.2.2) as a two-class classification model. To generate the answer, we flatten the normalized

<sup>4</sup>RepeatVector(n) replicates the input feature vector n times.

fused feature ( $F$ ) to generate a single long fused feature ( $F'$ ) which we pass through the fully-connected layer with two output neurons. We formulate the prediction procedure to predict the answer using Softmax as the activation function in the fully-connected layer as,

$$\begin{aligned}\hat{a} &= \mathcal{P}(a_i|F', \mathbf{W}, b) = \text{softmax}(F' \mathbf{W}_i + b_i) \\ &= \frac{e^{F' \mathbf{W}_i + b_i}}{e^{F' \mathbf{W}_{Yes} + b_{Yes}} + e^{F' \mathbf{W}_{No} + b_{No}}}, \quad i \in \{Yes, No\}\end{aligned}\tag{8}$$

where,  $\hat{a}$  is the prediction probability of selecting the  $i^{th}$  answer word ( $a_i$ ) given  $F'$  bias ( $b_i$ ), and weight matrix  $\mathbf{W}_i$  ( $i \in \{Yes, No\}$ ). We use categorical cross entropy as loss function having the following formula.

$$\mathcal{L}(a, \hat{a}) = -(a_{Yes} * \log(\hat{a}_{Yes}) + a_{No} * \log(\hat{a}_{No}))\tag{9}$$

where,  $a_i$  and  $\hat{a}_i$  denote the actual and predicted probability, respectively, of selecting ‘Yes/No’ as answer.

**Answer Prediction - Others:** We treat model  $m_2$  (c.f. Section 3.2.2) as a multi-label classification model for which we create a separate word-index dictionary for answers  $D_a = \{w_1 : 1, w_2 : 2 \dots w_z : z\}$ , where  $z$  is the total number of unique words in the answer list. We also transform the  $x^{th}$  answer list  $A(x)$  of answer length  $r$  to  $A'(x) = [a'_1, a'_2 \dots a'_r]$ , where  $a'_i$  is the  $i^{th}$  answer which is encoded in the form of one-hot vector (a binary vector with values 0 and 1). We pass the normalized fused feature ( $F'$ ) to the fully-connected layer with  $t$  output neurons to generate  $F''$ . We formulate the recursive prediction procedure to predict the answer words using TimeDistributed<sup>5</sup> layer having Softmax activation as,

$$\hat{a} = \mathcal{P}(a_i|F'', \hat{a}_{i-1}, b) = \text{softmax}(F'' \mathbf{W}_i + b_i) = \frac{e^{F'' \mathbf{W}_i + b_i}}{\sum_{j=1}^z e^{F'' \mathbf{W}_j + b_j}}\tag{10}$$

---

<sup>5</sup>[https://keras.io/api/layers/recurrent\\_layers/time\\_distributed/](https://keras.io/api/layers/recurrent_layers/time_distributed/)

where,  $\hat{a}$  is the probability of selecting the  $i^{th}$  answer word ( $a_i$ ) given  $F''$ , bias ( $b$ ), and the set of probabilities of previously predicted answer words ( $\hat{a}_{i-1}$ ), and  $\mathbf{W}_i$  is the weight matrix.  $z$  is the number of the words in vocabulary. We use categorical cross entropy as the loss function as follows:

$$\mathcal{L}(a, \hat{a}) = - \sum_{i=1}^z \sum_{j=1}^r (a_{ij} * \log(\hat{a}_{ij})) \quad (11)$$

In Eq (11), for  $i^{th}$  word in the answer sequence of length  $r$ ,  $a_{ij}$  and  $\hat{a}_{ij}$  denotes the actual and predicted probability of selecting the  $j^{th}$  word of the answer dictionary having  $z$  words.

### 3.2.3. Hyper-parameters

In the QS module to create the tf-idf vector, we consider top 500 words with the highest tf-idf values from the vocabulary in the training set of 2000 words. After studying the training set, we select 10 words ('is', 'was', 'are', 'how', 'can', 'does', 'which', 'what', 'type', and 'there') to form the set of question identifier words. We use the default values of the rest of the parameters (e.g., the c in the SVM). For question embedding, we create 600-dimensional word embedding by concatenating the two 300-dimensional GloVe and FastText embeddings, following an approach similar to the work proposed in (Ghannay et al., 2016). We create a question dictionary of size 1050 to capture the most frequent words in the questions. For answers, we create a separate answer dictionary of size equal to the count of unique word in the answer list. As a negligible number of questions is of length greater than 21, we fix the maximum question length as 21. 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 is 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. For the hidden layer of Bi-LSTM, we fix 128 neurons in each

direction. We use the Categorical Accuracy<sup>6</sup> as the metrics to calculate the mean accuracy rate for multiclass classification problems across all the predictions. We consider a batch of size 256 for training. We set the number of epochs to 251 and to optimize the weights during training we use Adam optimizer (Kingma & Ba, 2014). We obtained the optimal hyper-parameters value based on the model performance on validation dataset.

### 3.2.4. Evaluation Metric

In our work, the datasets we use for evaluating our proposed model has only one reference answer. Thus, reporting high accuracy means generating a high number of the answers with exact same words as in the reference answer. This may not be necessary at all times and is a very complex task even in the medical domain. However, more than one answer may be correct. This is explained with an example given in Table 2.

<b>Question</b>	<b>Multiple correct answers</b>
Where is the lung lesion located?	<ul style="list-style-type: none"> <li>• Right lobe</li> <li>• Lower lobe</li> <li>• Right lower lobe</li> </ul>

**Table 2.** Example where the absence of the degree of specification makes all the answers to the question are correct.

In this regard, we follow the standard evaluation schemes from the ImageCLEF2018 VQA-Med 2018 challenge. The evaluation metrics are as follows:

- **BiLingual Evaluation Understudy (BLEU) (Papineni et al., 2002):**  
It stands for and it is a popular evaluation metric in machine translation, which compares the generated answer with the reference answer based on the number of n-grams of the generated answers that match with the

---

<sup>6</sup><https://github.com/keras-team/keras/blob/master/keras/metrics.py>

reference answer, along with the brevity penalty for shorter output. BLEU is computed using the multiple modified n-gram provisions<sup>7</sup>

$$\text{BLEU} = \text{BP} \cdot \exp \left( \sum_{n=1}^N w_n \log_e p_n \right) \quad (12)$$

where  $p_n$  is the modified n-gram precision, BP is the brevity penalty to penalize short answer,  $w_n$  is weight between 0 and 1 for  $\log_e p_n$  and  $\sum_{n=1}^N w_n = 1$ ,  $N$  is the maximum length of n-gram. BP can be computed as follows:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ \exp \left( 1 - \frac{r}{c} \right) & \text{if } c \leq r \end{cases} \quad (13)$$

where  $c$  is the number of unigrams in all the candidate answers  $r$  is the best match lengths for each candidate answer in the dataset.

BLEU score serves as a better evaluation metric in this work, but it is not that effective when more than one medical term indicates the same part or symptom (e.g. the words ‘*Lung*’, and ‘*Lobe*’ refers to the same organ). As shown in Table 3, all the answers are correct for the question, but the BLEU score decreases as it fails to consider synonyms during evaluation.

Question	Multiple correct answers
Where is the lesion located?	<ul style="list-style-type: none"> <li>– Right lobe</li> <li>– Right lung</li> <li>– Right lobe of the lung</li> </ul>

**Table 3.** Example of the semantically similar answers. Although not all the answers to the question are the same, but they are semantically correct.

- Word-based Semantic Similarity (WBSS) (Hasan et al., 2018): It is another evaluation metrics used to assess the performance of the systems

---

<sup>7</sup>To compute modified n-gram precision, all candidate n-gram counts and their corresponding maximum reference counts are collected. The candidate counts are clipped by their corresponding reference maximum value, summed, and divided by the total number of candidate n-grams. Specifically,

submitted in the VQA-Med 2018 challenge. WBSS metric based on Wu-Palmer Similarity<sup>8</sup> Wu & Palmer (1994) with WordNet ontology in the backend. It computes a similarity score between the ground truth answer and system-generated answer by considering the word-level semantic similarity.

We follow the evaluation setup discussed in ImageCLEF2018 VQA-Med 2018 challenge overview paper (Hasan et al., 2018) to evaluate the performance of the system. Towards this, we first pre-process the predicted and ground-truth answers and then calculate the scores. For pre-processing, we convert the answers to lower-case, remove the punctuations, and apply tokenization<sup>9</sup> to the individual words in the answer. We also remove the stopwords in the answer from the NLTK’s<sup>10</sup> English stopword list.

We also use the following metrics to evaluate the effectiveness of our Answer Prediction module for  $q\_type = \text{‘Yes/No’}$  (c.f. Section 3.2.2):

- **Precision (P):** It reflects the fraction of correctly predicted instances of a class (say  $c$ ) from the total number of predicted instances as  $c$ .

$$P = \frac{|\{\text{instances of } c\} \cap \{\text{predicted instances as } c\}|}{|\{\text{predicted instances as } c\}|} \quad (14)$$

We report the macro-averaged precision  $P_m$ , where  $C$  is a set of all the possible classes,

$$P_m = (\sum_{c \in C} P_c) / \|C\| \quad (15)$$

- **Recall (R):** It reflects the fraction of correctly predicted instances from the total number of actual instances belonging to  $c$ .

$$R = \frac{|\{\text{instances of } c\} \cap \{\text{predicted instances as } c\}|}{|\{\text{instances of } c\}|} \quad (16)$$

---

<sup>8</sup><https://datasets.d2.mpi-inf.mpg.de/mateusz14visualturing/calculate>

<sup>9</sup><http://www.nltk.org/modules/nltk/tokenize/punkt.html#PunktLanguageVars>.

word tokenize

<sup>10</sup><http://nltk.org/>

We report the macro-averaged recall ( $R_m$ ) for the set of all classes C similar to Eq (15).

- **F1-score (F1):** It is a function of P and R.

$$F1 = 2 * ((P * R) / (P + R)) \quad (17)$$

We report the macro-averaged F1-score similar to Eq (15).

- **Accuracy (A):** It reflects the fraction of correctly predicted instances from the total number of instances.

$$A = \frac{|\{\text{correctly predicted answers}\}|}{|\{\text{answers}\}|} \quad (18)$$

#### 4. Data Description

Datasets for Medical-VQA 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 the 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 the variety in orientation and modalities, images can be of any body part or organ such as *heart, lung, skull*, etc. (Fig 8).



(a) Brain (b)Breasts (c) Chest (d) CNS (e) AP (f) Axial (g)Coronal(h) PA

**Fig 8.** Sample images in the Medical-VQA dataset. The images in this dataset can be of different organs (a to d) and/or modalities (e to h).

#### 4.1. RAD Dataset

RAD dataset is a recently released dataset for VQA in the medical domain. Statistics of the dataset are as follows:

- The training set consists of 3,064 question-answer pairs.
- The test set consists of 451 question-answer pairs.

Some of the images in the dataset are blurred while others contain markings such as short information, tags, etc. But none of the images in the dataset contains a stack of sub-images. The questions are primarily categorized into 11 categories viz. abnormality, attribute, color, counting, modality, organ system, other, plane, positional reasoning, and size. The average question length is 5 to 7 words which is greater than the answer length. 53% of the answers are of *Yes/No* type while rest of them are either one word or short phrase answers.

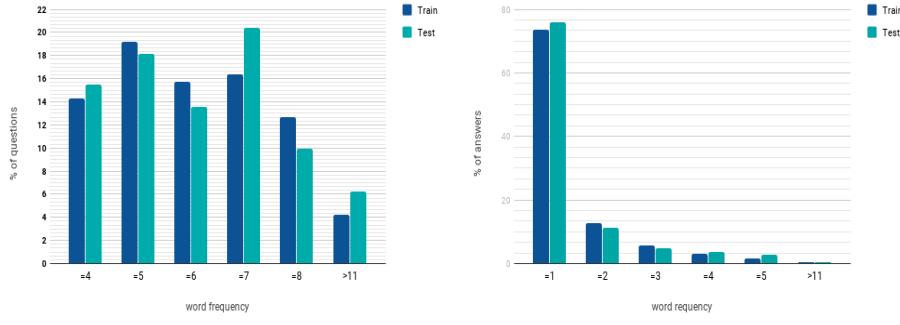
The maximum question length in the dataset is about 21 words, with an average of 7 words. It should also be noted that many questions are being rephrased, which are semantically similar. For example,

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

From the 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 of the answers are of *Yes/No* type. This is visualized by the peak in the Fig 9b for one-word answers.

#### 4.2. CLEF18 Dataset

CLEF18 task is similar to RAD task where a semi-automatic approach is used to generate the questions and answers from the captions. It uses the



(a) Questions

(b) Answers

**Fig 9.** Word-Frequency distribution in the RAD dataset. The graph demonstrates that for both the questions and the answers, this distribution is almost similar for train and test data splits.

radiology images and their respective captions extracted from the PubMed Central articles<sup>11</sup> (essentially a subset of the ImageCLEF2017 caption prediction task (Eickhoff et al., 2017)). Due to the way the question-answer (QA) pairs are generated, they are diverse and descriptive. The dataset also contains a lot of artificial questions that are semantically invalid. Table 4 demonstrates the complexity of a sample question-answer pair from the training data.

Question	Answer
what reveals prominent bilateral enhancing parietal occipital lesions on flair and t2 sequences and small areas of hyperintensity in the left periventricular white matter on diffusion weighted images?	mri of the brain
what does mri in sagital plane show?	the collection was superficial to the muscles of the back and the gluteal region but deep to the posterior layer of the thoraco lumbar fascia

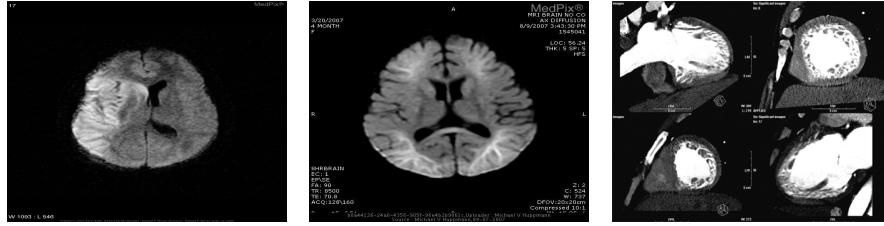
**Table 4.** Sample examples from the CLEF18 training data. Most of the questions and answers are descriptive and complicated in this dataset as they are generated semi-automatically.

<sup>11</sup><https://www.ncbi.nlm.nih.gov/pmc/>

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 10a) and most of the images contain radiology markings (Fig 10b) such as short information, tags, arrows, etc. A few of them even consists of a stack of sub-images (Fig 10c).



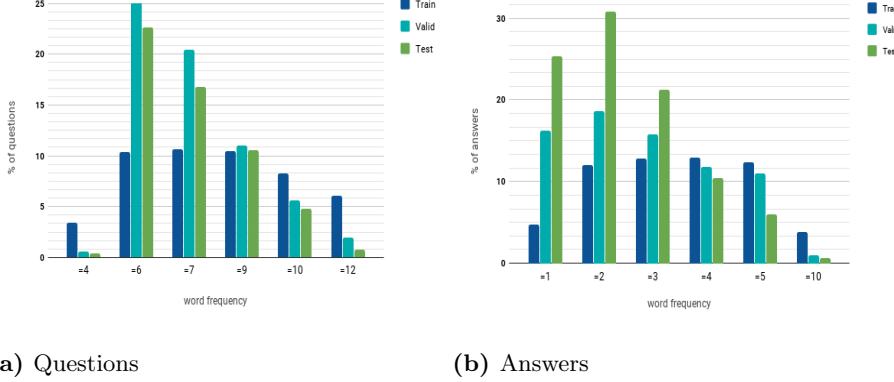
**Fig 10.** Sample images in the CLEF18 dataset. Some of the images in this dataset are blurred (hazy/not clear), and/or contains short information in the form of radiology markings, and/or contains stack of sub-images.

Question categorization is not present in the dataset and only 0.6% of the answers in training, 6% in validation, and 10% in test data are of *Yes/No* type. From Fig 11, we can see that the average question length is more than answers, and the word frequency distribution is not even in the data splits.

## 5. Results and Analysis

### 5.1. Baselines

We compare our proposed methodology with the existing works in VQA-Med.



**Fig 11.** Word-Frequency distribution in the CLEF18 dataset. The graph shows that for the data splits, this distribution is not comparable for both the questions and the answers.

Towards this, we use the following baseline models.

1. ResNet152 + LSTM + MFH (Peng et al., 2018): We compare our approach with the best system reported in the ImageCLEF2018 challenge of VQA-Med 2018. They used LSTM to extract the question features, whereas the image features were extracted from the ResNet152 model pre-trained on the Imagenet dataset. For question and image feature fusion, they employed the co-attention mechanism with MFH to generate the question-image representation. They predicted the probable words to form the answers using the multi-label classification and then generated the answers using sampling.
2. Inception-Resnet-v2 + Bi-LSTM + Attention (Zhou et al., 2018): Our second baseline model corresponds to the second best system participated in the ImageCLEF2018 challenge of VQA-Med 2018. Before applying the question and image to their proposed model, they performed the pre-processing steps on image and question both. They employed Inception-Resnet-v2 model to extract image features, and Bi-LSTM model to encode the questions. They utilized the attention mechanism to fuse the image and question features.
3. VGG-16 + LSTM + SAN (Abacha et al., 2018): This is the third best

system participated in the ImageCLEF2018 challenge of VQA-Med 2018. They used LSTM to extract the question features, whereas the image features were extracted from the last pooling layer of VGG-16 pre-trained on the Imagenet dataset. The stacked attention network (Yang et al., 2016) is used to fuse the question and image features to obtain a single feature representation. These features are used to predict the answers from the given answer list, which is compiled from the training dataset.

4. VGG16 + BERT + MFB (Yan et al., 2019): We compare the performance of our proposed model with the best performing system participated in the ImageCLEF2019 challenge of VQA-Med 2019. They utilized the BERT model to extract the question features, whereas the image features were extracted from the multiple pooling layers of VGG-16 pre-trained on Imagenet dataset. Multi-Modal Factorized Bilinear (MFB) (Yu et al., 2017) pooling were used to fuse the question and image features.

RAD			CLEF18			CLEF18+RAD			
Yes/No	Others	Overall	Yes/No	Others	Overall	Yes/No	Others	Overall	
<b>Precision</b>	0.98	0.99	0.99	1.00	0.93	0.93	0.99	1.00	0.99
<b>Recall</b>	0.99	0.98	0.99	0.28	1.00	0.93	0.99	1.00	0.99
<b>F<sub>1</sub>-score</b>	0.99	0.98	0.99	0.44	0.96	0.91	0.99	1.00	0.99

**Table 5.** Performance of Question-Segregation model in terms of Precision, Recall, and F<sub>1</sub>-score.

### 5.2. Result

Table 5 shows the performance of our QS model on RAD, CLEF18, and CLEF18+RAD dataset in terms of Precision, Recall, and F<sub>1</sub>-score. QS using SVM shows impressive results on the stated datasets. For CLEF18 dataset Recall and F<sub>1</sub>-score for ‘Yes/No’ type question is little less due to the less number of such questions in training example.

Table 6 shows the comparison of our proposed approach with the baseline models on the datasets in terms of BLEU and WBBS scores. Our proposed

Model	RAD		CLEF18		CLEF18+RAD	
	BLEU	WBSS	BLEU	WBSS	BLEU	WBSS
Peng et al. (2018)	–	–	0.161	0.184	–	–
Peng et al. (2018)*	0.023	0.104	0.023	0.072	0.027	0.081
Zhou et al. (2018)	–	–	0.134	0.173	–	–
Zhou et al. (2018) <sup>\$</sup>	0.522	0.532	0.072	0.112	0.277	0.299
Abacha et al. (2018)	–	–	0.121	0.174	–	–
Abacha et al. (2018)*	0.035	0.213	0.051	0.170	0.036	0.173
Yan et al. (2019)*	0.002	0.011	0.005	0.069	0.002	0.023
Ours	0.411	0.437	0.132	0.162	0.257	0.288

**Table 6.** Comparison between the baseline models and our model in terms of BLEU and WBBS scores. Star (\*) denote the re-implementation of the proposed work with the authors reported experimental setups. Dollar (\$) denote the official implementation of the approach proposed by the author.

approach for medical visual question answering achieves the BLEU score of 0.132 and WBBS score of 0.162 on the CLEF dataset. Peng et al. (2018) reports the BLEU score of 0.161 on the CLEF dataset. Since there is no official open-source implementation available for their system, we re-implemented the approach with the official experimental setup discussed in the paper, but only achieves the BLEU score of 0.023 and WBBS score of 0.072. We also fine-tune the hyper-parameters of their network with the available validation dataset, but we were not able to improve the performance further. Peng et al. (2018) used the sampling approach to generate the answer in contrast with the other systems that participated in the ImageCLEF2018 VQA-Med 2018 challenge. The comparatively less (only 5413 examples) amount of data to train the system raises the question of the effectiveness of the sampling approach to generate the correct answer. We also extend the experiments for the RAD and CLEF+RAD dataset on the re-implementation of the approach of Peng et al. (2018). For both the datasets, we randomly select 10% of the data from the training set to fine-tune the network hyper-parameters. We obtain the BLEU score of 0.023 and WBBS score of 0.104 on the RAD dataset. The improvement (in terms

of WBBS) in the RAD dataset can be understood by the fact that the RAD test set contains questions, which can be answered in a single-word. A similar observation is made on the results of the CLEF+RAD dataset.

Zhou et al. (2018) reported the performance in terms of BLEU (0.134) and WBBS (0.173) scores on the CLEF dataset. We evaluated the performance of their implementations<sup>12</sup> on the CLEF dataset, and recorded the BLEU and WBBS score of 0.072 and 0.112, respectively. We achieve the BLEU score of 0.522 and WBBS score of 0.532 on the RAD dataset with the re-implementation of their approaches. The reason for significant improvement for the RAD dataset is that Zhou et al. (2018) performed considerable pre-processing on image, question, and answer and also post-processed the generated answers. In their pre-processing steps, they adopted image enhancement and reconstructed the images with exceedingly small random rotations, offsets, scaling, clipping, and increase to 20 images per image. For questions, they utilized the methods like stemming and lemmatization to alter verbs, nouns, and other words into their original forms. Furthermore, they replaced all the medical terms with their abbreviations, a combination of letters and numbers are replaced with ‘pos’ token and the pure numbers are mapped to an ‘num’ token. For answers, they used lemmatization and removed all the stop words. They also replaced all the words associated with any number in answer. These words are generally the measures, for e.g. ‘cm’ in ‘5 cm’. As a post-processing step, they added several simple rules to the generated answers to make these more reasonable. In addition, they also deleted extra prepositions and additional words from the answers of yes/no questions. These additional pre-processing and rule-based post-processing steps make the system highly focused on medical VQA and not adaptable to the other domains of VQA. Additionally, the rules favor the short answers, which are only useful in the case of answer prediction and may suffer for answer generation. In contrast, our proposed approach achieves better performance for the CLEF dataset and

---

<sup>12</sup><https://github.com/youngzhou97qz/CLEF2018-VQA-Med>

comparable performance on the CLEF+RAD dataset without any additional processing on questions, images, or answers. Our system is generic, and it can be adaptable to any domain of VQA.

We also perform an additional experiment to further examine the performance improvement by Zhou et al. (2018) on RAD and CLEF+RAD datasets. Toward this, we introduce an attention mechanism similar to the Zhou et al. (2018) in our proposed model. The introduction of basic attention leads to significant performance improvement on the RAD dataset. With the new model (our proposed + attention), we achieve 0.542 and 0.553 BLEU and WBBS scores, respectively, on the RAD dataset. The new model achieves the 0.110 (0.313) and 0.051 (0.142) BLEU (WBBS) scores on CLEF and CLEF+RAD datasets, respectively. The experiment with the new model shows that the attention favors the RAD dataset well, and we achieve much better BLEU, and WBBS scores compared to Zhou et al. (2018). However, it goes against the CLEF dataset, and we report the performance degradation on CLEF and CLEF+RAD datasets.

Abacha et al. (2018) reported the BLEU score of 0.121 on the CLEF dataset. The official implementation of their system is not available. Therefore, we re-implemented the approach with the official experimental setup discussed in the paper, but only achieved the BLEU score of 0.051 and WBBS score of 0.170. Similar to the re-implementation of Peng et al. (2018), we fine-tune the hyper-parameters of their network with the available validation dataset. In our re-implementation, we use the fine-tuned hyper-parameters to train the network on the training and validation dataset of CLEF (as done by Abacha et al. (2018)) to evaluate the performance on the test dataset of CLEF. We perform experiments with the RAD and CLEF+RAD datasets on the re-implemented approach of Abacha et al. (2018). For both the RAD datasets, we randomly select 10% of the data from the training set to fine-tune the network hyper-parameters. For CLEF+RAD dataset, we fine-tune the hyper-parameters on the combination of validation data of CLEF and 10% of the RAD training set. We

obtain the BLEU score of 0.0351 and WBBS score of 0.213 on the RAD dataset. For CLEF+RAD dataset, we report the BLEU and WBBS scores of 0.0365 and 0.173, respectively.

We also compare the performance of our system with Yan et al. (2019). Similar to the other works, we re-implement all the existing approaches and report the results on all the datasets. By observing performance on all the experiments, we conclude that the number of training samples is not sufficient enough to use for the sophisticated language-vision fusion mechanism such as SAN, MFB, or MFH. The model quickly gets over-fitted, which leads to lower performance on the test dataset. However, much simpler models (our proposed and Zhou et al. (2018)) without any sophisticated language-vision fusion mechanism performs comparatively better on the task.

We follow the evaluation setup discussed in ImageCLEF2018 VQA-Med 2018 challenge overview paper (Hasan et al., 2018) to evaluate the performance of the system. But we cannot directly compare the results of our implementation of the ImageCLEF2018 VQA-Med 2018 participating systems (Peng et al., 2018; Zhou et al., 2018; Abacha et al., 2018) and our proposed approach, as the participants did not use their own evaluation setup/script. The ImageCLEF2018 VQA-Med 2018 challenge organizers release the performance scores in terms of BLEU and WBBS scores. Further, the official evaluation scores are not readily available for use. To make fair comparison and reproducibility of our works, we make our source codes available<sup>13</sup> along with the evaluation script of all the re-implementations of the existing works and our proposed approach.

We also analyze that, in general, a model performs better if more training samples are present, thus the models must have better scores when trained and tested on the combined dataset. But further analysis of the results reveals that the models perform better on the individual datasets than on the combined one.

---

<sup>13</sup><https://bit.ly/3aH6EFm>

This is due to the difference in the size of the datasets, and the way the questions are framed and answers are generated. While the CLEF18 dataset is created by a semi-automatic approach, the RAD dataset is manually created. The difference in quality and complexity of the generated sequence in the two datasets is clearly visible in Table 7, where both the questions require liver identification, but there is a considerable difference in complexity of the question as well as the answer. Due to the complications and the limited number of examples in the datasets, the model fails to learn efficiently.

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

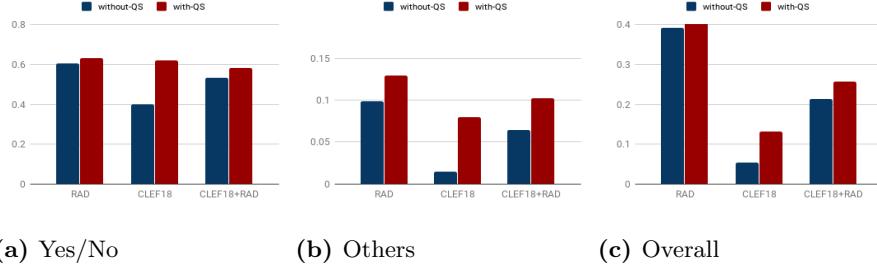
**Table 7.** Example of question-answer pair, which shows that question and answer structure in CLEF18 is more complicated than in RAD.

### 5.2.1. Impact of QS module

Table 8 demonstrates the impact of QS (c.f. Section 3.2.1) module. It also reveals that QS improves the model’s performance by a significant margin regardless of the question type. The performance difference is clearly visible in Fig 12.

	<b>RAD</b>		<b>CLEF18</b>		<b>CLEF18+RAD</b>	
	without	with	without	with	without	with
<b>Yes/No</b>	0.606	<b>0.634</b>	0.400	<b>0.620</b>	0.534	<b>0.581</b>
<b>Others</b>	0.099	<b>0.129</b>	0.015	<b>0.080</b>	0.064	<b>0.102</b>
<b>Overall</b>	0.392	<b>0.411</b>	0.053	<b>0.132</b>	0.213	<b>0.257</b>

**Table 8.** Result of our model with and without question type segregation, for RAD, CLEF18, and CLEF18+RAD dataset.



**Fig 12.** Impact of QS on the model performance. it shows that with QS the model performs better regardless of the type of question or dataset.

**Impact of QS on questions with type ‘Yes/No’:** For this type of question, the main advantage of QS is that it prevents the model from predicting any answer phrases other than a straightforward ‘Yes’ or ‘No’. While a model without QS can predict any answer word or sequence of answer words that turns out to be irrelevant. Table 9 shows the model’s predicted responses with and without the QS module.

<b>Question :</b>	Is the GI tract is highlighted by contrast?	Is the surrounding phlegmon normal?	Were both sides affected?	Does the PET scan show abnormal tracer accumulation?
<b>Ans (w/o.) :</b>	bilateral	bronchiectasis	axial	internal whorled
<b>Ans (w.) :</b>	Yes	Yes	Yes	No
<b>Ans (GT) :</b>	Yes	No	Yes	No

**Table 9.** Comparison of Ground Truth (GT) answer with the predicted answer having question-type Yes/No by the model with (w.) and without (w/o.) QS. Without QS, our model predicts answer words that are not expected for question in this category.

With QS, the search space is reduced to only two words that is ‘Yes’ and ‘No’

while a model without QS will have to unnecessarily predict the answer words from the entire answer dictionary (a dictionary containing all the possible answer words). This reduction in search space leads to a better chance of predicting the right answer. Table 10 demonstrates the effectiveness of QS in our model for ‘Yes/No’ type questions. The scores show that, for RAD, CLEF18 and CLEF18+RAD datasets, QS improves precision (Eq. 14), recall (Eq. 16) and f1-score (Eq. 17) by 0.3, 0.1, 0.2 and 0.3, 0.5, 0.2 points respectively.

	RAD		CLEF18		CLEF18+RAD	
	w/o.	w.	w/o.	w.	w/o.	w.
<b>precision</b>	0.38	<b>0.63</b>	0.60	<b>0.72</b>	0.37	<b>0.58</b>
<b>Recall</b>	0.37	<b>0.64</b>	0.12	<b>0.63</b>	0.35	<b>0.58</b>
<b>F1-score</b>	0.37	<b>0.63</b>	0.17	<b>0.62</b>	0.34	<b>0.58</b>

**Table 10.** Performance of our model with (w.) and without (w/o.) QS module for Y/N type questions.

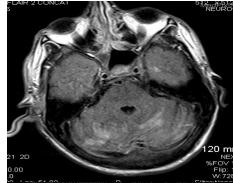
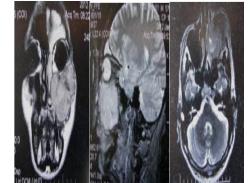
sample-imgs/chest

**Impact of QS on questions with type ‘Others’:** A model without QS can predict ‘Yes’ or ‘No’ as an answer which is completely irrelevant for the questions with type ‘Others’. But the quality of prediction increases when QS is integrated with the same model. Table 11 includes several such instances where the model with QS fails to predict the answer correctly but produces an answer that is applicable to the question-image pair and is more acceptable than straightforward ‘Yes’ or ‘No’.

### 5.3. Error Analysis

On the outputs generated by our model, we conduct thorough error analysis and classify the main sources of errors in the following types:

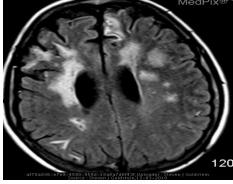
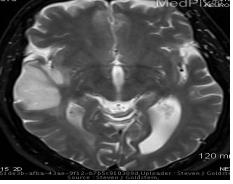
1. **Semantic Error:** This type of error occurs when the system predicts the answer words that are semantically comparable but fails to predict the exact same words as the ground-truth answer.

			
<b>Question :</b>	(1) The image is taken in what plane?	(2) Where are the infarcts?	(3) Is this patient male or female?
<b>Ans (w/o.) :</b>	Yes	No	No in
<b>Ans (w.) :</b>	pa	in right	chest
<b>Ans (GT) :</b>	axial	basal ganglia	female
			tumor

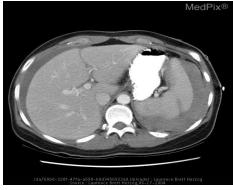
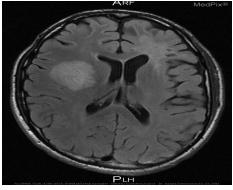
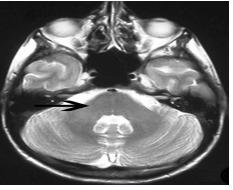
**Table 11.** Comparison of Ground Truth (GT) answer (Ans) with the predicted answer having question-type Yes/No by the model with (w.) and without (w/o.) QS. Without QS, our model predicts Yes or No along with other answer words that are not expected for question in this category.

2. **Modality/Plane Confusion:** This type of error specifically occurs for questions that require identification of the modality/plane. For such questions, the system fails to identify whether only the plane, modality, or a prediction of modality subtype is sufficient, or the question requires a possible combination of these.
3. **Specification Error:** When the question itself fails to specify how much information is desired in the answer, this type of error occurs where more than one correct answer is possible.
4. **Boundary Loss:** This type of error occurs when the system predicts the correct answer but does not predict the unimportant ground truth answer words that the question itself can determine.
5. **Miscellaneous Error:** This type of error occurs when the system predicts the information needed to answer the question, but fails to analyze the information collected to answer the question.

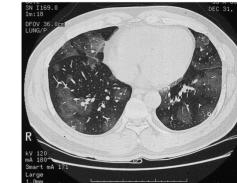
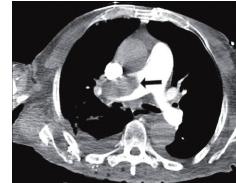
Table 12, Table 13, and Table 14 include a variety of examples along with

<b>Semantic Error:</b>				
<b>Question :</b>	Where do you see acute infarcts?	How is the patient oriented?	How was this image taken?	where is the lesion located?
<b>Ans(GT) :</b>	R frontal lobe	Posterior-Anterior	T2-MRI	Lower lobe of the right lung
<b>Ans(pred.) :</b>	right frontal lobe	pa	mri t2 weighted	right lower lobe
<b>Comments :</b>	R refers to Right.	pa is acronym for Posterior-Anterior.	T2-MRI denotes mri (t2-weighted).	Semantically same answer with different sentence structure.
<b>Modality/ Plane Confusion:</b>				
<b>Question :</b>	What imaging modality is this?	What kind of image is this?	What type of image is this?	What type of image is this?
<b>Ans(GT) :</b>	Sagittal view of t2 weighted mri	X-ray	Plain film x-ray	CT with contrast
<b>Ans(pred.) :</b>	mri t2 weighted	axial x ray	x ray	ct
<b>Comments :</b>	Identification of plane is not required as per the question.	Predicted answer is more precise.	GT answer is more precise.	GT answer consists of modality and it's subtype information.

**Table 12.** Error analysis (Semantic Error and Modality/Plane confusion): Comparison of Ground Truth answer with the predicted answer for understanding the error types.

Specification Error:				
<b>Question :</b> What does the ct scan of the chest show?		What lobe of the brain is the lesion located in?	What is the location of the lesion?	Where are the lesions formed?
<b>Ans(GT) :</b> A large mass		Right frontal lobe	Right lower lateral lung field	Mediastinum and Hilum of the right lung
<b>Ans(pred.) :</b> mass	right lobe	lower lung	in right lung	
<b>Comments :</b> does not specify the amount of mass.	Predicted answer lacks image-plan specification.	GT answer is more specific in terms of lesion location.	GT answer specifically defines the lesion formation.	
Boundary Loss:				
<b>Question :</b> In what plane was this image taken?		What lobe of the brain is the lesion located in?	Is the spleen present?	What kind of image is this?
<b>Ans(GT) :</b> Axial plane	Right frontal lobe	On the patient's left	T2 weighted mri	
<b>Ans(pred.) :</b> axial	right frontal	left	t2 weighted	
<b>Comments :</b> 'plane' is uninformative word in the GT answer.	Uninformative word 'lobe' in the GT answer.	GT answer consists additional and uninformative words.	Predicted answer lacks modality information.	

**Table 13.** Error analysis (Specification Error and Boundary Loss): Comparison of Ground Truth answer with the predicted answer for understanding the error types.

<b>Miscellaneous</b>				
<b>Error:</b>				
<b>Question :</b>	Is this an MRI or a CT scan?	Is this patient male or female?	Where are the lesions found?	What does the ct pulmonary angiogram show?
<b>Ans(GT) :</b>	MRI	Female	In both lungs	Massive filling defect
<b>Ans(pred.) :</b>	brain	chest	in right	large defect
<b>Comments :</b>	Question demands modality identification but organ is identified and predicted.	Question demands gender identification for which chest analysis is required	Partial prediction.	Type of defect is not identified.

**Table 14.** Error analysis (Miscellaneous Error): Comparison of Ground Truth answer with the predicted answer for understanding the error types.

the justifications to better understand each of these error types. To quantify each error types, we randomly choose 100 incorrectly generated samples from the CLEF+RAD dataset and categorize them into the five error types. We quantitatively analyze the errors and found that 20.54% errors belong to *Modality/Plane Confusion* and 14.16% errors belong to *Semantic Error*. Similarly, we found 16.48% errors fall into the *Specification Error* type. *Boundary Loss* type error contribute to the 20.49% of the total errors. The remaining errors (28.33%) associated with the *Miscellaneous Error* type.

## 6. Conclusion

In this paper, we propose a hierarchical multi-modal approach to tackle the VQA problem in the medical domain. In particular, we use a question segregation 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. Our proposed approach can be applied to any related problem where such segregation is possible but it does require non-trivial changes in the architecture. To evaluate the usefulness of our proposed model, we conduct experiments on two different datasets, RAD and CLEF18. We also perform experiments on the combined data 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 stated baseline models. It suggests that questions with different types learn better in isolation having their individual learning paths. Experimental results indicate the effectiveness of our work, depicting its value for the VQA in the medical domain. We also find out that even simple versions of our model are competitive.

Further analysis of the obtained results reveals that the evaluation metric needs improvement to evaluate VQA in the medical domain. For future work, we plan to investigate a better evaluation strategy for evaluating the task apart from devising a detailed scheme for QS. We also plan to introduce better individual models to handle each of the leaf node tasks.

## Acknowledgment

Asif Ekbal gratefully acknowledges the Young Faculty Research Fellowship (YFRF) Award supported by the Visvesvaraya PhD scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government

of India, and implemented by Digital India Corporation (formerly Media Lab Asia).

## References

- Abacha, A. B., Gayen, S., Lau, J. J., Rajaraman, S., & Demner-Fushman, D. (2018). Nlm at imageclef 2018 visual question answering in the medical domain. In *CLEF (Working Notes)*.
- Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., & Parikh, D. (2015). VQA: Visual Question Answering. In *Proceedings of the IEEE international conference on computer vision* (pp. 2425–2433).
- Arai, K., & Kapoor, S. (2019). *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC)* volume 2. Springer.
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5, 135–146. URL: <https://www.aclweb.org/anthology/Q17-1010>. doi:10.1162/tacl\_a\_00051.
- Bradley, E., Zeynettin, A., Jiri, S., & Panagiotis, K. (2017). Data From LGG-1p19qDeletion. URL: DOI:<https://doi.org/10.7937/K9/TCIA.2017.dwehtz9v>.
- Chen, D., Bolton, J., & Manning, C. D. (2016). A thorough examination of the cnn/daily mail reading comprehension task. *arXiv preprint arXiv:1606.02858*,
- .
- Chen, D., Mao, Y., & Zhou, J. (2019). Constructing medical image domain ontology with anatomical knowledge. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1750–1757). IEEE.
- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). 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)* (pp. 1724–1734). Doha, Qatar: Association for Computational Linguistics. URL: <https://www.aclweb.org/anthology/D14-1179>. doi:10.3115/v1/D14-1179.

Cid, Y. D., Liauchuk, V., Kovalev, V., & Müller, H. (2018). Overview of ImageCLEFtuberculosis 2018-Detecting Multi-drug Resistance, Classifying Tuberculosis Type, and Assessing Severity Score. In *CLEF2018 Working Notes, CEUR Workshop Proceedings, Avignon, France*.

Clark, A. T., Megerian, M. G., Petri, J. E., & Stevens, R. J. (2018). Question Classification and Feature Mapping in a Deep Question Answering System. US Patent 9,911,082.

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine learning*, 20, 273–297.

Cristianini, N., Shawe-Taylor, J. et al. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge university press.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, .

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., & Hon, H.-W. (2019). Unified Language Model Pre-training for Natural Language Understanding and Generation. In *Advances in Neural Information Processing Systems* (pp. 13042–13054).

Eickhoff, C., Schwall, I., de Herrera, A. G. S., & Müller, H. (2017). Overview of ImageCLEFcaption 2017-Image Caption Prediction and Concept Detection for Biomedical Images. In *CLEF (Working Notes)*.

- Fu, Z. (2019). An Introduction of Deep Learning Based Word Representation Applied to Natural Language Processing. In *2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)* (pp. 92–104). IEEE.
- Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., & Rohrbach, M. (2016). Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 457–468). Austin, Texas: Association for Computational Linguistics. URL: <https://www.aclweb.org/anthology/D16-1044>. doi:10.18653/v1/D16-1044.
- Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., & Xu, W. (2015). 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* (pp. 2296–2304). Cambridge, MA, USA: MIT Press. URL: <http://dl.acm.org/citation.cfm?id=2969442.2969496>.
- Gao, P., Jiang, Z., You, H., Lu, P., Hoi, S. C., Wang, X., & Li, H. (2019). Dynamic Fusion with Intra-and Inter-modality Attention Flow for Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6639–6648).
- Gebhardt, E., & Wolf, M. (2018). Camel Dataset for Visual and Thermal Infrared Multiple Object Detection and Tracking. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 1–6). IEEE.
- Ger, R. B., Yang, J., Ding, Y., Jacobsen, M. C., Cardenas, C. E., Fuller, C. D., & Howell, R. M. (2018). Data from Synthetic and Phantom MR Images for Determining Deformable Image Registration Accuracy (MRI-DIR). URL: <TheCancerImagingrcive.DOI:10.7937/K9/TCIA.2018.3f08iejt>.

- Ghannay, S., Favre, B., Esteve, Y., & Camelin, N. (2016). Word Embedding Evaluation and Combination. In *LREC* (pp. 300–305).
- Graves, A., & Schmidhuber, J. (2005). Framewise Phoneme Classification with Bidirectional LSTM and other Neural Network Architectures. *Neural networks : the official journal of the International Neural Network Society*, 18, 602–10. doi:10.1016/j.neunet.2005.06.042.
- Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). LSTM: A Search Space Odyssey. *IEEE transactions on neural networks and learning systems*, 28, 2222–2232.
- Guo, J., He, H., He, T., Lausen, L., Li, M., Lin, H., Shi, X., Wang, C., Xie, J., Zha, S. et al. (2020). Gluoncv and Gluonnlp: Deep Learning in Computer Vision and Natural Language Processing. *Journal of Machine Learning Research*, 21, 1–7.
- Gupta, D., Ekbal, A., & Bhattacharyya, P. (2019). A deep neural network framework for english hindi question answering. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19, 1–22.
- Gupta, D., Kumari, S., Ekbal, A., & Bhattacharyya, P. (2018a). MMQA: A Multi-domain Multi-lingual Question-Answering Framework for English and Hindi. In N. C. C. chair), K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, & T. Tokunaga (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Paris, France: European Language Resources Association (ELRA).
- Gupta, D., Lenka, P., Ekbal, A., & Bhattacharyya, P. (2018b). Uncovering code-mixed challenges: A framework for linguistically driven question generation and neural based question answering. In *Proceedings of the 22nd Conference on Computational Natural Language Learning* (pp. 119–130).

- Gupta, D., Pujari, R., Ekbal, A., Bhattacharyya, P., Maitra, A., Jain, T., & Sengupta, S. (2018c). Can Taxonomy Help? Improving Semantic Question Matching using Question Taxonomy. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 499–513). Association for Computational Linguistics. URL: <http://aclweb.org/anthology/C18-1042>.
- Hasan, S. A., Ling, Y., Farri, O., Liu, J., Lungren, M., & Müller, H. (2018). Overview of the imageclef 2018 medical domain visual question answering task. In *CLEF2018 Working Notes. CEUR Workshop Proceedings, Avignon, France (September 10-14 2018)*.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).
- Hersh, W. R., & Bhupatiraju, R. T. (2003). TREC GENOMICS Track Overview. In *TREC*.
- Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37 ICML'15* (pp. 448–456). JMLR.org. URL: <http://dl.acm.org/citation.cfm?id=3045118.3045167>.
- Ionescu, B., Müller, H., Villegas, M., de Herrera, A. G. S., Eickhoff, C., Andreadczyk, V., Cid, Y. D., Liauchuk, V., Kovalev, V., Hasan, S. A. et al. (2018). Overview of ImageCLEF 2018: Challenges, datasets and evaluation. In *International Conference of the Cross-Language Evaluation Forum for European Languages* (pp. 309–334). Springer.
- Kafle, K., & Kanan, C. (2016). Answer-Type Prediction for Visual Question Answering. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4976–4984). doi:10.1109/CVPR.2016.538.

- Kafle, K., Shrestha, R., Cohen, S., Price, B., & Kanan, C. (2020). Answering Questions about Data Visualizations using Efficient Bimodal Fusion. In *The IEEE Winter Conference on Applications of Computer Vision* (pp. 1498–1507).
- Kawahara, J., & Hamarneh, G. (2016). Multi-Resolution-Tract CNN with Hybrid Pretrained and Skin-Lesion Trained Layers. In *MLMI@MICCAI*.
- Kingma, D., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*, .
- Lau, J. J., Gayen, S., Abacha, A. B., & Demner-Fushman, D. (2018). A Dataset of Clinically Generated Visual Questions and Answers about Radiology Images. *Scientific data*, 5, 180251.
- Li, X., Grandvalet, Y., Davoine, F., Cheng, J., Cui, Y., Zhang, H., Belongie, S., Tsai, Y.-H., & Yang, M.-H. (2020). Transfer Learning in Computer Vision Tasks: Remember where you come from. *Image and Vision Computing*, 93, 103853.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740–755). Springer.
- Lin, T.-Y., RoyChowdhury, A., & Maji, S. (2017). Bilinear Convolutional Neural Networks for Fine-grained Visual Recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40, 1309–1322.
- Liu, C., Chen, L.-C., Schroff, F., Adam, H., Hua, W., Yuille, A. L., & Fei-Fei, L. (2019). Auto-deeplab: Hierarchical Neural Architecture Search for Semantic Image Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 82–92).
- Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., & Pietikäinen, M. (2020). Deep Learning for Generic Object Detection: A survey. *International journal of computer vision*, 128, 261–318.

- Long, M., Zhu, H., Wang, J., & Jordan, M. I. (2017). Deep Transfer Learning with Joint Adaptation Networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (pp. 2208–2217). JMLR. org.
- Lu, J., Yang, J., Batra, D., & Parikh, D. (2016). Hierarchical Question-image Co-attention for Visual Question Answering. In *Advances In Neural Information Processing Systems* (pp. 289–297).
- Magnuson, J. S., You, H., Luthra, S., Li, M., Nam, H., Escabi, M., Brown, K., Alloppenna, P. D., Theodore, R. M., Monto, N. et al. (2020). EARSHOT: A Minimal Neural Network Model of Incremental Human Speech Recognition. *Cognitive Science, 44*.
- Malinowski, M., Rohrbach, M., & Fritz, M. (2015). Ask Your Neurons: A Neural-based Approach to Answering Questions about Images. In *Proceedings of the IEEE international conference on computer vision* (pp. 1–9).
- Mikolov, T., Karafiat, M., Burget, L., Černocky, J., & Khudanpur, S. (2010). Recurrent Neural Network based Language Model. In *Eleventh annual conference of the international speech communication association*.
- Morante, R., Krallinger, M., Valencia, A., & Daelemans, W. (2013). Machine Reading of Biomedical Texts about Alzheimer’s Disease. *CEUR Workshop Proceedings, 1179*.
- Mukuze, N., Rohrbach, A., Demberg, V., & Schiele, B. (2018). A Vision-grounded Dataset for Predicting Typical Locations for Verbs. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Ningthoujam, D., Yadav, S., Bhattacharyya, P., & Ekbal, A. (2019). Relation extraction between the clinical entities based on the shortest dependency path based lstm. *arXiv preprint arXiv:1903.09941*, .
- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th*

*annual meeting on association for computational linguistics* (pp. 311–318). Association for Computational Linguistics.

Peng, Y., Liu, F., & Rosen, M. P. (2018). Umass at imageclef medical visual question answering (med-vqa) 2018 task. In *CLEF (Working Notes)*.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. In *2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Ruder, S. (2019). *Neural Transfer Learning for Natural Language Processing*. Ph.D. thesis NUI Galway.

Ruder, S., Peters, M. E., Swayamdipta, S., & Wolf, T. (2019). Transfer Learning in Natural Language Processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials* (pp. 15–18).

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. et al. (2015). ImageNet Large Scale Visual Recognition Challenge. *Int J Comput Vis*, 115, 211–252.

Shaimaa, B., Olivier, G., Sebastian, E., Kelsey, A., Mu, Z., Majid, S., Hong, Z., Weiruo, Z., Ann, L., Michael, K., Joseph, S., Andrew, Q., Daniel, R., Sylvia, P., & Sandy, N. (2017). Data for NSCLC Radiogenomics Collection. URL: <http://doi.org/10.7937/K9/TCIA.2017.7hs46erv>.

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.

Sun, X., Xv, H., Dong, J., Zhou, H., Chen, C., & Li, Q. (2020a). Few-shot Learning for Domain-specific Fine-grained Image Classification. *IEEE Transactions on Industrial Electronics*, .

- Sun, Y., Xue, B., Zhang, M., Yen, G. G., & Lv, J. (2020b). Automatically Designing CNN Architectures Using the Genetic Algorithm for Image Classification. *IEEE Transactions on Cybernetics*, .
- Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, Inception-Resnet and the Impact of Residual Connections on Learning. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2014). Going Deeper with Convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 1–9).
- Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? *IEEE transactions on medical imaging*, 35, 1299–1312.
- Tarando, S. R., Fetita, C., Faccinetto, A., & Brillet, P.-Y. (2016). Increasing CAD System Efficacy for Lung Texture Analysis using a Convolutional Network. In *Medical Imaging 2016: Computer-Aided Diagnosis* (p. 97850Q). International Society for Optics and Photonics volume 9785.
- Traore, B. B., Kamsu-Foguem, B., & Tangara, F. (2018). Deep Convolution Neural Network for Image Recognition. *Ecological Informatics*, 48, 257–268.
- Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M. R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., Almirantis, Y., Pavlopoulos, J., Baskiotis, N., Gallinari, P., Artieres, T., Ngonga, A., Heino, N., Gaussier, E., Barrio-Alvers, L., Schroeder, M., Androutsopoulos, I., & Paliouras, G. (2015). An Overview of the BIOASQ Large-scale Biomedical Semantic Indexing and Question Answering Competition. *BMC Bioinformatics*, 16, 138. URL: <http://www.biomedcentral.com/content/pdf/s12859-015-0564-6.pdf>. doi:10.1186/s12859-015-0564-6.

- Vallieres, M., Kay-Rivest, E., Perrin, L. J., Liem, X., Furstoss, C., Khaouam, N., Nguyen-Tan, P. F., Wang, C.-S., & Sultanem, K. (2017). Data from Head-Neck-PET-CT. URL: [TheCancerImagingArchive.doi:10.7937/K9/TCIA.2017.8oje5q00](https://TheCancerImagingArchive.doi:10.7937/K9/TCIA.2017.8oje5q00).
- van Tulder, G., & de Bruijne, M. (2016). Combining Generative and Discriminative Representation Learning for Lung CT Analysis With Convolutional Restricted Boltzmann Machines. *IEEE Transactions on Medical Imaging*, 35, 1262–1272. doi:10.1109/TMI.2016.2526687.
- Wang, H., Zheng, B., Yoon, S. W., & Ko, H. S. (2018). A Support Vector Machine-based Ensemble Algorithm for Breast Cancer Diagnosis. *European Journal of Operational Research*, 267, 687–699.
- Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics* (pp. 133–138). Las Cruces, New Mexico, USA: Association for Computational Linguistics. URL: <https://www.aclweb.org/anthology/P94-1019>. doi:10.3115/981732.981751.
- Xiong, C., Merity, S., & Socher, R. (2016). Dynamic Memory Networks for Visual and Textual Question Answering. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48* ICML’16 (pp. 2397–2406). JMLR.org. URL: <http://dl.acm.org/citation.cfm?id=3045390.3045643>.
- Yadav, S., Ekbal, A., Saha, S., & Bhattacharyya, P. (2019). A unified multi-task adversarial learning framework for pharmacovigilance mining. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 5234–5245).
- Yadav, S., Ekbal, A., Saha, S., Bhattacharyya, P., & Sheth, A. (2018). Multi-task learning framework for mining crowd intelligence towards clinical treatment. In *Proceedings of the 2018 Conference of the North American Chapter of the*

*Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (pp. 271–277).

Yadav, S., Ramteke, P., Ekbal, A., Saha, S., & Bhattacharyya, P. (2020). Exploring disorder-aware attention for clinical event extraction. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16, 1–21.

Yan, X., Li, L., Xie, C., Xiao, J., & Gu, L. (2019). Zhejiang university at imageclef 2019 visual question answering in the medical domain. In *CLEF (Working Notes)*.

Yang, M., Liu, S., Chen, K., Zhang, H., Zhao, E., & Zhao, T. (2020). A Hierarchical Clustering Approach to Fuzzy Semantic Representation of Rare Words in Neural Machine Translation. *IEEE Transactions on Fuzzy Systems*,

Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2016). Stacked Attention Networks for Image Question Answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 21–29).

Yu, Z., Yu, J., Fan, J., & Tao, D. (2017). Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Yu, Z., Yu, J., Xiang, C., Fan, J., & Tao, D. (2018). Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems*, 29, 5947–5959.

Zbontar, J., Knoll, F., Sriram, A., Muckley, M., Bruno, M., Defazio, A., Parente, M., Geras, K., Katsnelson, J., Chandarana, H., Zhang, Z., Drozdzal, M., Romero, A., Rabbat, M., Vincent, P., Pinkerton, J., Wang, D., Yakubova, N., Owens, E., & Murrell, T. (2018). fastMRI: An Open Dataset and Benchmarks for Accelerated MRI. URL: <https://fastmri.org/>.

Zhi, J., Sun, J., Wang, Z., & Ding, W. (2018). Support Vector Machine Classifier for Prediction of the Metastasis of Colorectal Cancer. *International journal of molecular medicine*, 41, 1419–1426.

Zhou, Y., Kang, X., & Ren, F. (2018). Employing inception-resnet-v2 and bi-lstm for medical domain visual question answering. In *CLEF (Working Notes)*.