Visual Question Answering for the Medical Domain

Dhruv Sharma

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Science and Applications

Chandan K. Reddy, Chair
Bimal Viswanath
Jiepu Jiang

June 18, 2020
Blacksburg, Virginia

Keywords: Visual Question Answering, deep learning, medical images

Visual Question Answering for the Medical Domain

Dhruv Sharma

(ABSTRACT)

Medical images are extremely complicated to comprehend for a person without expertise. The limited number of practitioners across the globe often face the issue of fatigue due to the high number of cases. This fatigue, physical and mental, can induce human-errors during the diagnosis. In such scenarios, having an additional opinion can be helpful in boosting the confidence of the decision-maker. Thus, it becomes crucial to have a reliable Visual Question Answering (VQA) system which can provide a "second opinion" on medical cases. However, most of the VQA systems that work today cater to real-world problems and are not specifically tailored for handling medical images. Moreover, the VQA system for medical images needs to consider a limited amount of training data available in this domain. In this thesis, we develop a deep learning-based model for VQA on medical images taking the associated challenges into account. Our *MedFuseNet* system aims at maximizing the learning with minimal complexity by breaking the problem statement into simpler tasks and weaving everything together to predict the answer. We tackle two types of answer prediction - categorization and generation. We conduct an extensive set of both quantitative and qualitative analyses to evaluate the performance of *MedFuseNet*. Our results conclude that *MedFuseNet* outperforms other state-of-the-art methods available in the literature for these tasks.

Visual Question Answering for the Medical Domain

Dhruv Sharma

(GENERAL AUDIENCE ABSTRACT)

Medical radiology scans are extremely complicated to examine for a person without expertise. The limited number of practitioners across the globe often face the issue of fatigue due to the high number of cases they examine. This fatigue, physical and mental, can induce human-errors during the diagnosis. In such scenarios, having an additional opinion can be helpful in boosting the confidence of the decision-maker. Thus, it becomes crucial to have a reliable Visual Question Answering (VQA) system which can provide a "second opinion" on medical cases. A VQA system is used to generate an answer for a question associated with an input image. However, most of the VQA systems that work today cater to real-world problems and are not specifically tailored for handling medical images. In this thesis, we propose a VQA system, *MedFuseNet*, for answering the input query associated with a medical image. We conduct an extensive analysis to evaluate the performance of our system, *MedFuseNet*. We conclude that our system outperforms the existing VQA techniques for the medical domain.

Dedication

Dedicated to my parents, my elder sister and my teachers. Also to the essential workers providing services relentlessly during this pandemic.

Acknowledgments

I would like to thank my advisor Dr. Chandan Reddy for his guidance and endless support during each phase of my masters degree. His confidence in me encouraged me to push my limits and motivated me into learning new things. I would also like to thank Dr. Bimal Viswanath and Dr. Jiepu Jiang for their support and suggestions which were invaluable for the completion of this study. I am indebted to the Computer Science Department for the opportunity to pursue and fund my Master's degree at Virginia Tech, and Sharon for helping out with all administrative issues and queries.

I would like to express my gratitude to Dr. Sanjay Purushotham for his ideas and feedback that were extremely important in helping me complete this project. I am really grateful to my flatmates Sahil, Soumil, and Utkarsh for always being there. I am equally thankful to Kedar, Sandeep, Abhishek, Vanditi, Anurag, Abhinav and all my friends here and back in India.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The advancements in the field of deep learning have demonstrated tremendous success in achieving state-of-the-art results in various problems in the fields of computer vision, natural language processing, information retrieval, to name a few. This was primarily due to the recent enhancements in the computational power of the machines, and development of new learning and optimization methods for neural networks. Several application domains have also benefitted enormously due to these recent advances. In particular, the medical domain has seen a major boost in the use of deep learning techniques for gathering more meaningful insights about various complex data sources ranging from radiology scans to medical records. Significant improvements in the performance metrics have been recorded for tasks related to image understanding such as segmentation of tumors present in brain [16], skin [9], and others [49]. There has also been a lot of compelling research done in Natural Language Processing Tasks (NLP) and medical records such as the predictive analysis using clinical records of patients [40, 50]. A more interesting and involved problem statement is the one that has both vision and NLP components in it - Visual Question Answering (VQA). The aim of VQA is to answer a natural language question associated with an image. In the medical domain, the image is replaced by a radiology scan of a patient accompanied by a clinically relevant question-answer pair. Moreover, the answer might belong to a pre-defined limited set or be a sequence of words.

As per the data provided by the World Health Organization (WHO) [52], over 45% of the

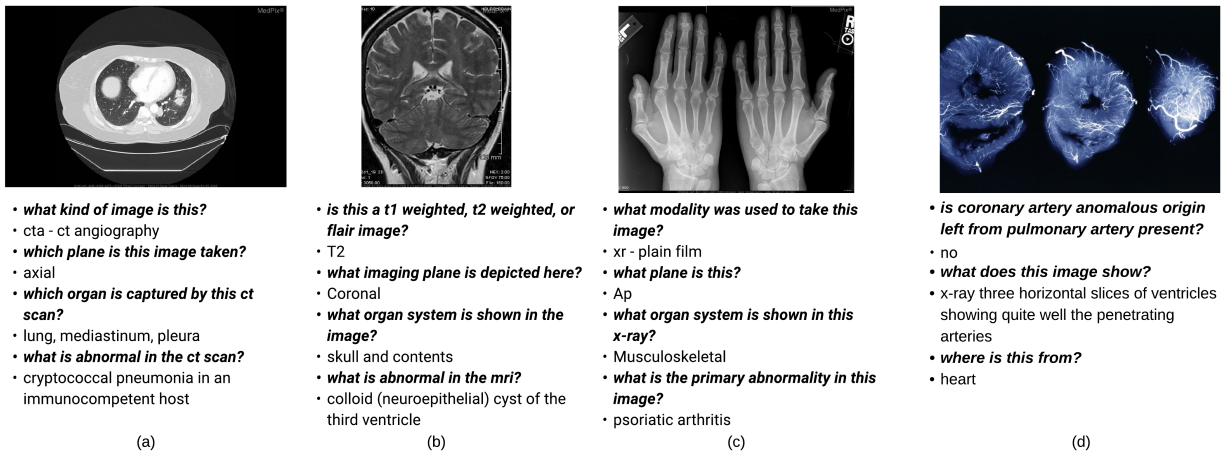


Figure 1.1: Sample radiology scans and the corresponding question-answer pairs from the MED-VQA and PathVQA dataset. The first three (a,b,c) belong to the MED-VQA dataset and the last one (d) belongs to the PathVQA dataset.

countries across the globe have less than 1 physician available per 1000 population. This burdens each medical practitioner with a large number of medical reports to examine and increases the likelihood of human error due to fatigue [5]. This gives rise to the need for capable Computer-Aided Diagnosis (CAD) systems [33]. Such systems can prove to be very useful by providing a second-opinion to the doctor and reduce the chance of human error. CAD systems can also be helpful in providing deeper insights into the case which might not be comprehensible to naked eyes. Moreover, with the medical reports being made digitally available on portals to the patients, having certain platforms that can clarify their trivial queries can help in diverging a lot of traffic from hospitals and reduce the stress over doctors. These reliable and well-evaluated portals can also reduce the risk of imparting incorrect knowledge to the patients as compared to the vast amount of misleading information available online. Thus, it becomes exceedingly important to have a visual question answering system for the medical domain to tackle all these challenges.

Apart from being a problem that has both, Computer Vision and NLP, aspects to it, VQA for the medical domain has its own new challenges. The main challenge for any supervised

machine learning problem in the medical domain is the availability of labeled data. This can be attributed majorly to the limited availability of data due to the privacy concerns of the patients. Moreover, labeling of medical data is itself a challenge due to the limited number of practitioners. When we compare the VQA datasets for real-world and medical domains, there are various datasets present for VQA related to real-world images. The medical VQA datasets have data points in the order of thousands, whereas the VQA datasets pertaining to real-world having a hundreds of times more data points. This poses a challenge in using deep learning techniques for VQA in the medical domain due to the inherent requirements of voluminous data for training deep models. Another important consideration in VQA is as follows: since VQA has a natural language question and an associated image as inputs, it is necessary that the multi-modal information is processed appropriately to maximize the information from the two modalities. Adding to this, the medical data is implicitly complicated due to the high amount of information packed in a single clinical report or a radiology scan. There can be more complications in the data due to the noise induced during scanning. Moreover, an ideal VQA system for medical images should cater to all types of queries related to any organ system instead of having a separate system for each class of organs. Another challenging aspect of this problem is the generation of the answer. This requires the model to output a meaningful sequence of words, in case of answer generation. Thus, coming up with an optimal deep learning architecture that judiciously uses the medical data to minimize the answer prediction error is of prime importance.

A typical deep learning approach for VQA has four major components - image feature extraction, question feature extraction, the combination of the two feature vectors, and the answer prediction module. The high-level illustration of this approach can be seen in Figure 1.2. Another important aspect is the attention modules which help the model in confining the focus to the most relevant part of the inputs. The answer prediction module further has

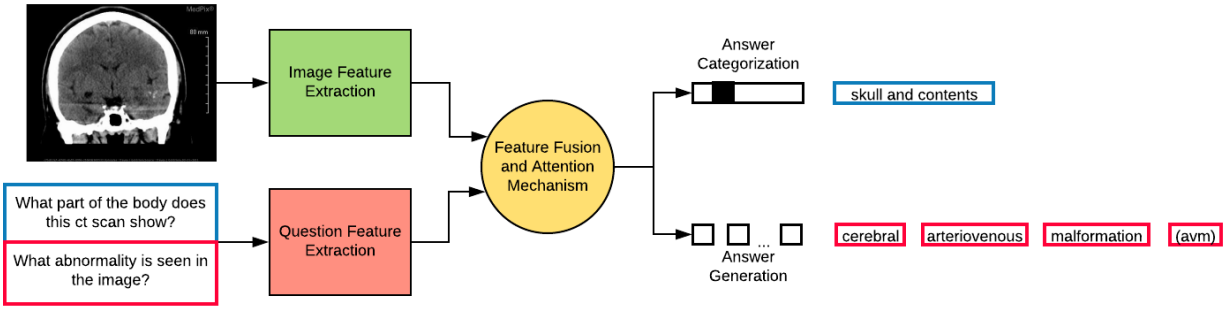


Figure 1.2: A high-level model design for the task of VQA. The model has four major components - image feature extraction, question feature extraction, feature fusion amalgamated with the attention mechanism, followed by answer categorization or generation depending on the task.

two subparts - answer categorization and answer generation. In answer categorization, the model selects an answer from the set of possible answers while answer generation requires the model to produce a meaningful sequence of words that answers the input question. This means that we need a full-fledged generative decoder for the task of answer generation, which adds up to the challenges of the problem. To tackle the aforementioned challenges, we propose a similar deep learning network that tackles the VQA task as a collation of various components. For the purpose of our experiments, we use the MED-VQA 2019 dataset and PathVQA dataset. Sample question-answer pairs from the dataset are shown in Figure 1.1. We aim at maximizing the learning of the model with the minimal model complexity. The major contributions of this thesis are as follows:

- Propose a model for medical domain VQA which focuses on 1. the accurate representation learning of the multi-modal inputs and 2. the optimal fusion of the feature vectors to maximize the information gain from the combination of the input feature vectors.
- Introduce an LSTM-based decoder for the task of answer generation and implement heuristics that help in improving the model performance.
- Conduct an exhaustive ablation study to observe the importance of each component in

the model and find a combination of the underlying model components that generate the best performance metrics for the ImageCLEF MED-VQA dataset.

- Exploring the interpretability of the model by visualizing various attention mechanisms used in the model. This provides a deeper insight into understanding the answering capability.

The rest of the thesis is organized as follows. Chapter 2 explores the existing methods for learning feature from the multi-modal inputs, their fusion, and the existing models for VQA pertaining to real-world and medical VQA. In Chapter 3, we discuss the insights about the ImageCLEF MED-VQA dataset and present the entire framework used to tackle the VQA problem for the medical domain. This is followed by the comprehensive discussions of the experiments and the results in Chapter 4. Chapter 5 puts forward the conclusions and the scope of future work.

Chapter 2

Related Work

2.1 Visual Question Answering

VQA for real-world domains has been a well-explored problem with various datasets such as DAQUAR [31], VQA [4], VQA 2.0 [15], CLEVR [21], to name a few. For VQA related to natural images, the work in [57] is motivated by the use of attention modules to focus on the important parts of the image relevant to the question. The works in [10, 13, 23, 30] also make use of attention schema to cater to the image and question while answering. The set of experiments in all these works affirm the idea of using attention modules in VQA models to confine the focus of the model to the appropriate portions of the input and has motivated us to pursue that in our model as well. However, not every work uses an attention mechanism. for example, the model presented in [47, 51]. Some of the existing works encourage the idea of using a simple concatenation of better features of the image and the question since a more robust representation of these inputs helps the model to converge faster. For VQA in the medical domain, the majority of work has been regarding dealing with the task as a classification problem [2, 3, 55]. There has been a very limited amount of work performed in answer generation for VQA. Work in [41] presents a method to tackle the VQA in the medical domain for answer generation as well as classification. It uses the transformer model to build the architecture for generating a sequence of words. Work in [48] presents a different perspective on solving VQA for the medical domain by presenting a model that is more aware

of the input question. However, all these papers do not present an exhaustive comparison of the models and an interpretation of the results. Another concern in the medical domain is that there are very limited VQA datasets like RAD-VQA [27], Indian Diabetic Retinopathy Image Dataset (IDRiD) [38], and ImageCLEF MED-VQA 2019 [1].

For our experiments, we explore two medical VQA datasets - MED-VQA [1] and PathVQA [18]. MED-VQA dataset was released as a part of the ImageClef Challenge 2019. PathVQA dataset is a huge pathology base VQA dataset. Next we discuss research done for each component as shown in Figure 1.2.

2.2 Deep Learning for VQA

2.2.1 Image Representation Learning

The superior performance of the Convolutional Neural Networks (CNN) in computer vision tasks has established CNN models as a reliable tool for robust feature representation of the spatially correlated data. Generally speaking, the intermediate layer just before the output layer is used as the feature vector. For the models like VGGNet [45], AlexNet [26], DenseNet [20], ResNet [17], the models are usually trained on large-scale image datasets such as ImageNet [11]. The image features, thus manufactured, using the intermediate layers of these pre-trained networks provide a rich feature representation of the input image.

2.2.2 Textual Representation Learning

For textual data, there have been various strategies to represent the features. Word2Vec [32], GloVe [37], FastText [8] are some of the word embedding algorithms that have been successful in a robust representation of the text at a word level. Sequential networks such as Recurrent Neural Networks (RNNs) ([44]) or Long-Short Term Memory (LSTM) networks [19] are used to make more sense out of these embeddings. BERT [12] and XLNet [56] are the state-of-the-art models for the NLP tasks and, hence, are a good candidate for feature extraction tasks as well due to the generalization they provide.

2.2.3 Feature Fusion Techniques

The most intuitive way of combining the feature vectors is through the element-wise multiplication of vectors. However, due to the limited interaction of the elements of the two participating vectors, the outer product or the bilinear product of the two vectors is a better strategy to capture a robust and complete interaction of all the elements. Various fusion techniques relevant to VQA have been devised over time to maximize vector interaction and reduce computational cost. These include Multimodal Compact Bilinear Pooling (MCB) [13], Multimodal Low-rank Bilinear Pooling (MLB) [22], Multimodal Tucker Fusion (MUTAN) [6], Multimodal Factorized Bilinear Pooling (MFB) [58]. They all share the same idea of making the bilinear pooling of two vectors computationally feasible.

In this project, we also use the bilinear product to fuse the two extracted feature vectors. We work on each component independently and explore the possible options. Finally, we present a combination of all these components that provide the best results. We also present both quantitative and qualitative reasoning for the performance of the model.

Chapter 3

Proposed Model

In this chapter, we will first define the problem statement and then discuss each component of the model and provide the details of the final deep learning based architectures used for solving the VQA problem in the medical domain.

3.1 Problem Definition

Definition 3.1. Answer Classification. Given a medical image v, an associated natural language question q, the aim is to produce the answer $\tilde{a}$ from a possible set of answers $\mathcal{A}$, where the ground truth answer is represented by a. This can be formulated as follows

$$\tilde{a} = \underset{a \in \mathcal{A}}{\operatorname{argmax}} P(a|v, q; \Theta) \tag{3.1}$$

where Θ is the set of model parameters, v is the input radiology scan, and q is the natural language question associated with the image in Equation (3.1).

Definition 3.2. Answer Generation. Given a medical image v, a natural language question associated with the image q, the aim is to generate a sequence of words $\tilde{a} = [\tilde{a}_1, \ldots, \tilde{a}_i]$, where the ground truth answer is represented by $a = [a_1, \ldots, a_j]$, where $\tilde{a}_1, \ldots, \tilde{a}_i$ and

$a_1, \ldots, a_j$ belong to the answer word vocabulary $W_{\mathcal{A}}$. This can be represented as

$$[\tilde{a}_1, \ldots, \tilde{a}_i] = \underset{a_1,\ldots,a_j \in W_{\mathcal{A}}}{\operatorname{argmax}} P(a_1, \ldots, a_j | v, q; \Theta) \tag{3.2}$$

where Θ is the set of model parameters, v is the input radiology scan, and q is the natural language question associated with the image. We tackle the problem of VQA also as a answer generation problem wherein we generate a sequence of words from the answer word vocabulary $W_{\mathcal{A}}$ as shown in Equation (3.2).

We use the Softmax Cross-Entropy loss function to find the error in the answer prediction of the model. This can be formulated as follows:

$$\mathcal{L}(a, \tilde{a}) = \sum_i -p(a_i) \log(p(\tilde{a}_i)) \tag{3.3}$$

where $p(\tilde{a}_i)$ is the probability of $\tilde{a}_i$ being the answer, and $p(a_i)$ is the probability of a_i being the ground-truth answer, for the task of answer classification. In the case of answer generation, the cross-entropy loss, as defined in Equation (3.3), refers to the error in predicting each word of the generated answer from the word vocabulary $W_{\mathcal{A}}$.

In our approach, we aim at decomposing the problem into multiple components - Image feature generation, question feature generation, feature fusion, and answer prediction. The image feature generation component will have the image v as input and will output the image feature vector $\hat{v}$. Similarly, the question feature generation component will generate the feature vector $\hat{q}$ for the input question q. The feature vectors would then be combined to form z. We also use attention mechanisms to enhance the performance of the model. The combined vector z would further be processed to predict the answer depending on the task - categorization or generation. The exhaustive list of notations used in this thesis is provided

in Table 3.1.

Table 3.1: Notations used in this thesis.

Notation	Description
v	input image
$\hat{v}$	image feature vector
$\hat{v}_e$	attended image feature vector
q	input question
$\hat{q}$	question feature vector
$\hat{q}_e$	attended question feature vector
z	combined feature vector
d_i	attention output for the i^{th} step of the decoder
h_i	LSTM output for the i^{th} step of the decoder
g	number of attention glimpses
a	actual answer
$\tilde{a}$	predicted answer
$[a_1, \ldots, a_j]$	actual answer sequence
$[\tilde{a}_1, \ldots, \tilde{a}_i]$	predicted answer sequence
Θ	model parameters
$\mathcal{L}$	Loss function
$\mathcal{A}$	possible set of answers
$W_{\mathcal{A}}$	vocabulary of words in answers
$\circ$	inner product operation
N_b	batch size
$\mathcal{E}_v$	Image Attention Mechanism
$\mathcal{E}_q$	Question Attention Mechanism
$\mathcal{E}_d$	Decoder Attention Mechanism

3.2 Components of the Model

3.2.1 Image Feature Generation

The feature learning from images has been an active research area for decades. An intermediate layer of a CNN captures the features of the image at varying levels of abstraction. While the shallow layers represent a more elementary level of features, the deeper layers encapsu-

late a more generalized version of the features. Exploiting this interpretation, generally, the intermediate layer just before the output layer is used as the feature vector. We use the following models for image feature extraction. For our experiments, we have used VGGNet-16, DenseNet-121, and ResNet-152 models. Since the medical domains have inherent complexity as compared to the real-world images, models like DesneNet and ResNet which have skip connections provide more robust features by the virtue of deeper convolutional layers. These models are pre-trained on the ImageNet dataset. The intermediate output from the last convolutional block of each model was used as the feature representation of the medical image. Due to the superior performance of ResNet-152 over the other two, we propose it as the best option for image features.

3.2.2 Question Feature Generation

As discussed earlier in Chapter 2.2.2, word embeddings form the primary method for expressing the underlying context of natural language. However, they are insufficient and lack in capturing the context properly. While modeling the feature representation of the natural text, it is necessary that we appropriately capture the positional semantics of each word and not just the word-level semantics. Thus, we experiment with the state-of-the-art NLP architectures to model the features of the input question - BERT and XLNet. The primary idea behind these models is to learn an exhaustive textual representation of the question. We use the pre-trained versions of both the models for the feature extraction of the question. The set of experiments with these two methods also reveal that using BERT over XLNet leads to better results in most of the cases.

3.2.3 Feature Fusion Techniques

The most intuitive way of combining the feature vectors is through the inner product or the element-wise multiplication of vectors. However, due to the limited interaction of the elements of the two vectors in the inner product, it is a very primitive strategy for feature fusion. The outer product or the bilinear product of the two vectors is a better strategy to capture a robust and complete interaction of all the elements. A simple bilinear model for two vectors $v \in \mathbb{R}^m$ and $q \in \mathbb{R}^n$ is shown in Equation (3.4).

$$z_i = v^T W_i q \tag{3.4}$$

where $W_i \in \mathbb{R}^{m \times n}$ and $z_i \in \mathbb{R}^o$. Thus, the model needs to learn the parameter matrix $W = [W_1, \ldots, W_o] \in \mathbb{R}^{m \times n \times o}$. However, for values of $m = 1024, n = 1024, o = 512$, the number of parameters in the projection matrix W will be ~ 530 million parameters, which makes it computationally expensive and infeasible. There have been various ways that have been proposed to solve this problem. For our experiments, we have used Multimodal Compact Bilinear (MCB) Pooling [13], Multimodal Tucker Decomposition (MUTAN) [39], and Multimodal Factorized Bilinear Pooling (MFB) [58]. Each of these techniques simplify the process of Bilinear Pooling by presenting a way to decompose the outer product projection matrix W. Due to the simplicity of the MFB algorithm, we prefer using it over the other two methods which is also supported by the results of the ablation study presented in the latter part of the thesis.

The ease in implementation and the high convergence rate make this fusion strategy very impactful. In addition to the normal process, to avoid the model from converging to local minima, the output of the MFB module is normalized using power normalization and L-2 normalization. A detailed explanation of the process is available in Yu et al. [58].

3.2.4 Attention Mechanisms and Feature Fusion

A primitive model for any task with multimodal data, like VQA, is to first extract the feature vectors, combine the vectors using any one of the above-stated techniques, and then predicting the answer from the combined vector. However, questions that are very specific to the input image require a more specific context of the image. This is where attention mechanisms prove to be useful by focusing on the most relevant parts of the input. We explore two types of attention mechanisms in this thesis - Image Attention and Co-Attention scheme.

Image Attention: The image attention mechanism aims at spanning the attention of the model to the most relevant part of the image on the basis of the input question. This establishes a correlation between the multimodal input and helps the model converge fast. The image attention mechanism amalgamates the feature fusion technique with the attention maps to come up with the attended image feature vector as explained in lines 20-30 of Algorithm 1. Firstly, the image features $\hat{v}$ and question features $\hat{q}$ are combined using the fusion technique (line 21). The attention maps are then computed from this combined feature vector (lines 22-23). The input image features $\hat{v}$ are then overlaid with the attention glimpses (lines 24-28) to get the attended image feature vector $\hat{v}_e$. The pictorial representation of the algorithm is shown in Figure 3.1.

Image-Question Co-Attention: The image attention mechanism focuses on the significant parts of the image, however, it takes the entire question into consideration. A co-attention mechanism exploits the intuition that the key parts of the question can be solely computed for the question which can further be used to enhance the image attention. So, the model first computes the attended question feature vector $\hat{q}_e$ using the Question attention mechanism $\mathcal{E}_q$ as shown in Figure 3.1. It then uses this attended vector as an input to the

image attention mechanism as described in Algorithm 1 from lines 8-18, instead of question feature vector $\hat{q}$.

3.3 MedFuseNet

3.3.1 Answer Categorization

After discussing the various components of the model, we present the model that aims at maximizing the performance for answer prediction and minimizing the model complexity - *MedFuseNet*. The three major components of the model are as follows - 1. **image feature -** pre-trained ResNet-152, 2. **question feature -** pre-trained BERT, 3. **feature fusion -** MFB. Moreover, *MedFuseNet* uses Image-Question Co-Attention technique so that the model focuses only on the most relevant parts of the image and the question while predicting the answer. The pictorial representation of the model is shown in Figure 3.1.

The model *MedFuseNet* tackles all the challenges stated in Chapter 1. The following characteristics help in boosting the performance of the model:

- Since the ResNet and BERT models are pretrained on very large datasets, they provide a much better generalization for the features by the virtue of transfer learning.
- Due to the simplistic implementation of MFB, it reduces the complexity of calculating the outer product to a large extent, while conserving the information from the fusion of the two modalities. This reduces the model parameters and works well with the limited MED-VQA dataset.
- The co-attention mechanism helps in reducing the attention span of the model to the significant parts of the input, thus, reducing the search space for the model.

Algorithm 1: *MedFuseNet* Training Algorithm

Input: Image v, Question q, Answer a, Batch size N_b
Output: Trained model parameters Θ

Extract the image features $(\hat{v})$, from image (v)
Extract the question features $(\hat{q})$ from question (q)
for *a few iterations* **do**
  for *batch of size* N_b *in* $\{\hat{v},\ \hat{q},\ a\}$ **do**
    Perform Question Attention $\mathcal{E}_q(q)$ on $\hat{q}$ to get attended question features $(\hat{q}_e)$
    Perform Image Attention $\mathcal{E}_v(\hat{v}, \hat{q}_e, MFB, 2)$ on $\hat{v}$ to get attended image features $(\hat{v}_e)$
    Combine $\hat{q}_e$ and $\hat{v}_e$ using MFB($\hat{q}_e$, $\hat{v}_e$, 5000, 3) to get intermediate vector (z)
    Find the predicted answer $(\tilde{a})$ depending on the task as defined in Eq. (3.1) and Eq. (3.2)
    Calculate the loss $\mathcal{L}$ for a and $\hat{a}$ using Eq. (3.3)
    Update the model parameters Θ with the loss $\mathcal{L}$
  end
end
return trained model parameters Θ
Procedure `MFB`$(\hat{v}, \hat{q}, d_o, k)$
  $v' = Fully - Connected(\hat{v}, m, d_o)$
  $q' = Fully - Connected(\hat{q}, n, d_o)$
  Compute and store inner product $(\circ)$ of vector v' and vector q' in vector z
  Perform SumPooling of vector z with a window size of k
  Normalize vector z using L2-normalization
  return z
Procedure `Image Attention`$(\hat{v}, \hat{q}, \mathcal{F}, g)$
  Combine $\hat{v}$ and $\hat{q}$ using $\mathcal{F}(\hat{q}_e, \hat{v}_e)$ to get intermediate vector f
  $f_{conv} = \text{ReLU}(\text{Conv2d}(f, d_o, 512))$
  $f_{AttMaps} = \text{Softmax}(\text{Conv2d}(f_{conv}, 512, g))$
  Initialize v_e as an empty list to store the attention glimpses
  for $i \leftarrow 1$ **to** g **do**
    Find the attended image feature e_i for i_{th} glimpse as follows:
    $e_i = f_{AttMaps}[i] \circ \hat{v}$
    Add e_i to the list v_e
  end
  Sum over all the attention glimpses in v_e to get attended image feature vector $(\hat{v}_e)$
  return $\hat{v}_e$

As shown in Algorithm 1 (lines 1-12), the model first extracts the feature vectors $\hat{v}$ and $\hat{q}$ correspondingly for input image v and question q. This is followed by the computation of the attended question features $\hat{q}_e$ using question attention mechanism $\mathcal{E}_q(q)$. Then we use the Image Attention mechanism $\mathcal{E}_v$ as explained in Algorithm 1 (lines 20-30) to get the attended image features $\hat{v}_e$. $\hat{v}_e$ and $\hat{q}_e$ are then combined using MFB (lines 13-19) to get vector z. A classification model is then built over z to find the loss and update the model parameters Θ.

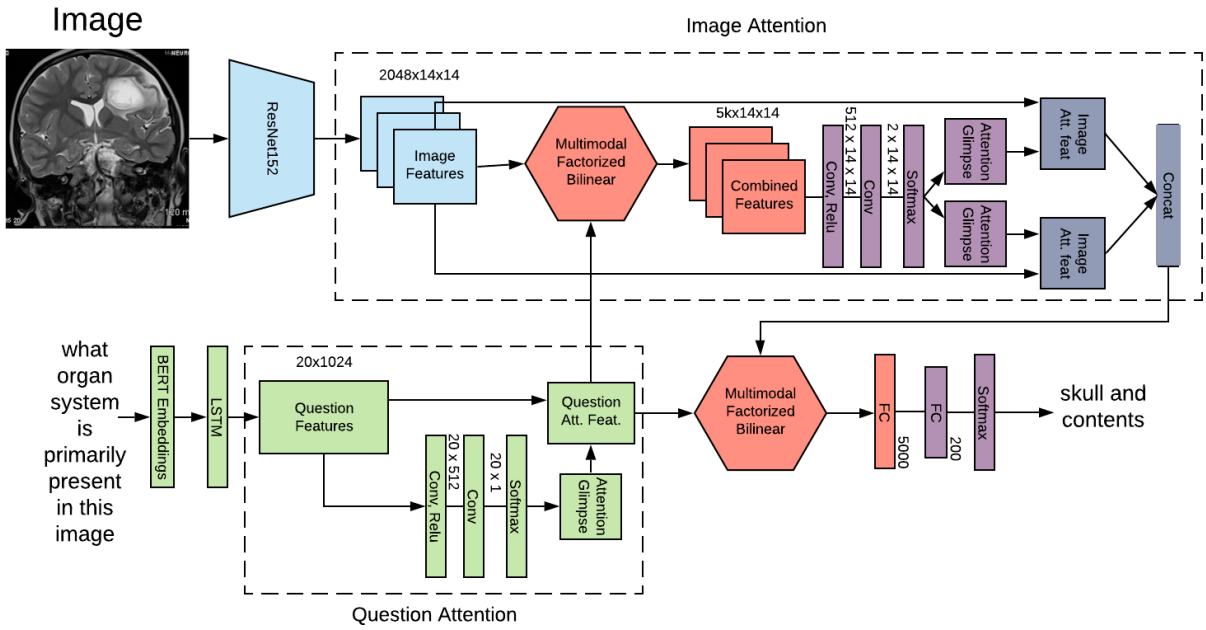


Figure 3.1: Our end-to-end framework for Medical Visual Question Answering for answer categorization. It takes the medical image and the associated question as the inputs, followed by the feature extraction. The question features are further processed using the question attention mechanism. The attended question features and the image features are then passed through the image attention mechanism to get the attended image features. These attended vectors are finally combined using MFB to build the answer classification module.

3.3.2 Answer Generation

As described in definition 3.2, the problem of answer generation is not a straightforward task as we need to generate a meaningful sequence of words from the answer word vocabulary $W_{\mathcal{A}}$ to predict the answer. Hence, we need a more sophisticated model for the fourth component, i.e., the answer prediction. We present an LSTM-based decoder model that builds over the fused feature vector. Our decoder model is inspired by the work presented in [54]. The pictorial representation of the decoder is shown in Figure 3.2. The key characteristics of our decoder are as follows:

- Due to the inherent complexity of the task of sequence generation, the model is susceptible to a slower convergence rate. Moreover, the limited amount of data in the medical domain may cause more hindrance to the model convergence rate. Thus, to increase the learning rate of the model, we use Teacher Forcing [7]. As a part of it, we pass the ground-truth word for the i^{th} time-step as well to the LSTM.

- To make each LSTM step prediction more accurate, we also incorporate the attention mechanism in the decoder. We use the output of the $i-1^{th}$ time-step to span the focus of the model on those parts of the image feature vector $\hat{v}_e$ that have already been answered. This helps the model to guide its search for the i^{th} word in the generated answer more precisely.

- During inference, we use Beam Search heuristic [53] to avoid the model from greedily generating the answer by choosing the best word at each decoding step.

Before generating the answer sequence using the decoder, we fuse the input image v and and question q to get the attended image features $\hat{v}_e$ as described in the procedure Image Attention of the Algorithm 1. This obtained vector $\hat{v}_e$ is passed to the decoder to generate

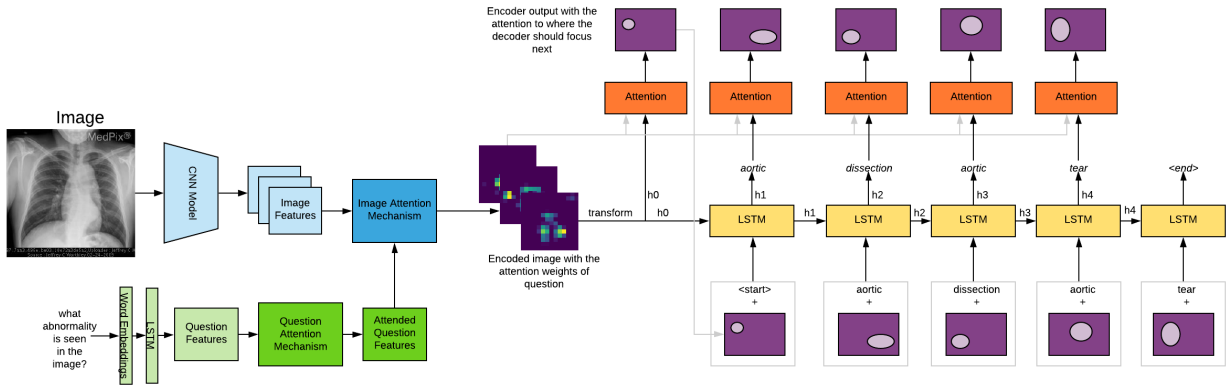


Figure 3.2: The architecture used for the answer generation task. This module takes the image and the question as the input. It generates the feature vectors for both and produces the combined vector after fusing them using MFB as a part of the image-question co-attention mechanism. This is followed by an LSTM-based decoder to generate the answer. The two major characteristics of this decoder are - the attention mechanism and teacher forcing. The attention mechanism helps the model in focusing on various parts of the image while generating a word, and the teacher enforcing helps the model converge faster.

the answer. As shown in Algorithm 2, $\hat{v}_e$ is first used to initialize the states of the LSTM (line 1). Following this, for the i^{th} step of the decoder, we concatenate the output d_{i-1} of the attention mechanism $\mathcal{E}_d$ for $(i-1)^{th}$ step with the i^{th} word in the ground truth answer, that is a_i, as shown in line 3 in Algorithm 2. This concatenated vector is then fed to the LSTM cell to get h_i which is also $\tilde{a}_i$, the i^{th} word in the predicted answer (lines 4-5 in Algorithm 2). The vectors h_i and $\hat{v}_e$ are then fed to the attention mechanism (lines 6-7 in Algorithm 2). The pictorial representation of the end-to-end model for answer generation is shown in Figure 3.2.

Algorithm 2: Decoder Algorithm for Answer Generation

Input: Attended Image Features $\hat{v}_e$, Answer $a_1, \ldots, a_n$

Output: Generated Answer $\tilde{a} = [\tilde{a}_1, \ldots, \tilde{a}_n]$

Initialize the decoder LSTM states using image features ($\hat{v}_e$)
Initialize generated answer $\tilde{a}$ as an empty list
Initialize d_0 as image features ($\hat{v}_e$)
for *each step i in* $[a_1, \ldots, a_n]$ **do**
    Concatenate a_i and d_{i-1}, the output of Decoder Attention $\mathcal{E}_d$ for $(i-1)^{th}$ step
    Feed this concatenated vector to the i^{th} decoder step
    Add h_i, which is also $\tilde{a}_i$, to list $\tilde{a}$
    Feed $\hat{v}_e$ and h_i to decoder attention $\mathcal{E}_d$ to get d_i
end
return Generated Answer $\tilde{a}$

Chapter 4

Experiments

We conduct several experiments on two real-world medical VQA datasets to compare the performance of our proposed model with the state-of-the-art VQA models. Our experiments help us to answer the following key questions: 1) How does the proposed (*MedFuseNet*) model perform w.r.t. the state-of-the-art models for different types of question categories 2) Can we visualize and explain the results of our proposed model? 3) What is the impact of different attention mechanisms on model performance? 4) How good are the answers generated by the proposed model in terms of BLUE scores?

4.1 Datasets for Answer Classification

4.1.1 MED-VQA

We use the ImageCLEF 2019 MED-VQA challenge dataset [1] to conduct our experiments. The dataset is well-structured= with 4200 images and medical questions associated with each image as shown in Figure 1.1. Each question belongs to one of the three categories - Modality, Plane, and Organ. In total, there are 3,825 image-question-answer triplets for each category combining all the splits. The detailed distribution of the data can be seen in Table 4.1. A detailed explanation of the three categories is as follows:

1. **Modality:** This category pertains to the modality of the input medical image. In

Table 4.1: Data distribution of the training, validation, and the testing split for the yes-no type question-answer pairs in MED-VQA dataset.

Split	Modality	Plane	Organ
Train	3200	3200	3200
Validation	500	500	500
Test	125	125	125

total, the training data set has 3200 images and question-answer pairs with modalities varying from 36 types.

2. **Plane:** This category pertains to the plane in which the medical image has been taken. In total, the training data set has 3200 images and question-answer pairs with planes varying from 16 types.

3. **Organ System:** This category describes the organ system captured by the image. There are 3,200 question-answer pairs in this category as well, and a total of 10 unique organ systems.

The maximum question length for the three categories combined is 13 words and the average question length is 8 words. The combined vocabulary of the questions contains about 100 words.

4.1.2 PathVQA

Another dataset that we use in our experiments is PathVQA dataset [18]. This is the VQA dataset on pathology images prepared using a novel pipeline from the captions of the images in textbooks. The dataset has 9,000+ medical images and 47,000+ QA pairs. Out of the entire dataset, we use only the yes/no type question-answer pairs for answer categorization experiments. The PathVQA dataset has three splits in it - train, validation, and test split.

All the three splits have fairly well distributed yes-no types question answers with almost a 1:1 proportion. The details of the dataset are presented in Table 4.2.

Table 4.2: Data distribution of the training, validation, and the testing split for the yes-no type question-answer pairs in PathVQA dataset.

Split	Medical Images	Yes type QA Pairs	No type QA Pairs
Train	4271	9305	9163
Validation	1176	2359	2335
Test	942	1874	1853

4.2 Datasets for Answer Generation

4.2.1 MED-VQA

Other than the three categories, as mentioned in Chapter 4.1.1, there is one more class of answers - abnormality prediction. The answers in this category are open-ended and cater to the type of abnormality that is being filmed using the radiology scan. Answering these types of questions are more useful to the healthcare providers as it can help them in getting a second opinion on the critical cases. The questions category 4 of MED-VQA are more inclined towards this motivation as they have open-ended questions with answers which do not belong to a limited set of answers. In total, we have 3,817 question-answer pairs with a wide variety of possible answers. The combined word vocabulary of the answers is of size 2109 words, out of which 756 words have an occurrence of one in the dataset. This poses a greater challenge to the model the answer generation for this skewed dataset. The average length of an answer is 2.63 words and the average length of a question is ~ 7 words.

4.2.2 PathVQA

As discussed in Chapter 4.1.2, PathVQA is a dataset about the question-answers related to pathology images. Apart from the yes-no type question-answers, it also has a great proportion of open-ended answer type data. For the set of experiments related to the task of answer generation, we subsample a dataset from this portion of the PathVQA dataset. To assure that the data is not skewed enough, we sample only those answers which have a frequency of at least 5 in the entire dataset. This gives us a total of 6,770 question-answer pairs as a part of 4,192 unique cases. The vocabulary size of the answers is about 480 words. The average number of words in an answer is 2.76 words. The average question length is ~ 6 words.

For all the datasets described above, the medical images were resized to be of the same dimension of $224 \times 224 \times 3$. This was done as most of the well-accepted pre-trained models take the input in this dimension. For each question, we first tokenized using the NLTK library in python [28]. Then, the question vocabulary was prepared and the tokens in the vocabulary were enumerated, which was used to convert the question to a list of numbers. The questions were also padded to make them all of the same lengths.

4.3 Evaluation Metrics

For evaluating the performance of the model in all the datasets discussed in Chapters 4.1 and 4.2, we use stratified 5-fold cross-validation after combining the training, the validation, and the testing pool. This helps in understanding the generalization capability of the model in a better manner. The combined data was split into 5 folds such that each fold had an appropriate representation of each class.

4.3.1 Answer Classification

We use three metrics to evaluate the performance of the model - *Accuracy, Area Under Curve - Receiver Operator Characteristics(AUC-ROC)*, and *Area Under Curve - Precision-Recall Curve (AUC-PRC)* for the task of answer classification. Accuracy is the primary metric used for any classification task and it quantifies the performance of the model in distinguishing between various classes. However, accuracy scores can be misleading for the data with imbalanced classes, as in the case of the MED-VQA dataset. So, we also calculate the AUC-ROC and AUC-PRC. AUC-ROC is defined by the area under the Receiver Operating Characteristics (ROC) Curve. A ROC curve describes the ability of the model to separate between various classes by plotting False Positive Rate (FPR) on X-axis and True Positive Rate (TPR) on the y-axis. Higher the area under the curve, better is the performance of the model. Similarly, AUC-PRC is the area under the curve with Precision on Y-axis and Recall on X-axis. Higher the value AUC-PRC, better is the performance. These metrics help us gauge the performance of the model with respect to the answer prediction considering the class imbalance as well. For the PathVQA dataset, we use only the accuracy as a metric to evaluate the performance of the models as the classes are fairly balanced with an equal proportion of yes and no type answers.

4.3.2 Answer Generation

To evaluate the answer generation capability of our model, we use generated sequence evaluation metrics such as Bilingual Evaluation Understudy (BLEU) score [34]. BLEU score calculates the similarity of the reference (ground truth answer) and the hypothesis (predicted answer) at an n-gram level. Thus, it is a very useful metric for comparing two sequential entries. Specifically, we use BLEU-1, BLEU-2, and BLEU-3 scores to compare the sequences

at 1-gram, 2-gram, and 3-gram levels, respectively. Apart from the BLEU score, we also compute the F-1 score of the generated answer. In terms of sequence generation, the F-1 score gives an idea about the performance of the model in generating the correct words. We use the NLTK library in Python for calculating these metric scores.

4.4 Baseline Models

We establish the superior performance of *MedFuseNet* by comparing it with the five baselines for the task of answer categorization. Three of the baselines are attention-based VQA models, while the other two are popular VQA models.

- **VIS + LSTM (V+L) [42, 43] -** This is a relatively simpler model that uses vanilla LSTM for question features and a CNN model for image feature extraction. The LSTM of the question feature was initialized using the image features. The last output of LSTM was used to predict the answer by having a dense-layer.

- **Deeper LSTM + Norm. CNN (d-L+n-I) [29] -** This model again uses a VGG16 for image feature extraction and a 2-layer LSTM model for question features. The two feature vectors are then combined using a simple element-wise multiplication to get the output vector.

- **Stacked Attention Networks (SAN) [57] -** SAN is an attention-based model and it uses multiple attention layers to refine the search space of the two feature vectors more thoroughly. It uses VGG16 based image features and CNN to extract the features of the question text. It then stacks attention layers over image vector and then applies an array of attention vectors on the question to obtain the final combined feature vector.

- **Hierarchical Co-attention (HiCAt) [30] -** This is again an attention-based model. The image features are CNN-based while the question features here use 1-D convolution over the word-embedding to get a hierarchy of the text. Two attention schemes are used in this work: parallel attention and alternating co-attention. In parallel attention, the model captures the attention of both vectors simultaneously while the latter one alternates the attention between the feature vectors of the two inputs.

- **Bilinear Attention Networks (BAN) [23] -** BAN is a novel baseline method that was proposed in 2018. It presents an architecture that tries to maximize the model performance by weaving the attention mechanism with the feature fusion technique. It uses a modified version of MFB model for feature fusion where in the attention mechanisms come into action during combination. It uses FasterRCNN features with the aim of using localized feature fusion instead of a global feature vector.

For the task of answer generation, there are no appropriate baselines that are suitable for comparison. Hence, we use BAN as one of the baseline and plug-in a decoder into the model architecture to make it compatible for answer generation. This decoder is a simple LSTM-based model. We also incorporate teacher forcing method in this decoder to help the model converge faster.

4.5 Implementation Details

We have implemented all the components of *MedFuseNet* using Pytorch [35]. The image feature extraction was developed using pre-trained models available in Keras [14]. Embedding-as-a-Service [46] was used for extracting the features for question from the pre-trained BERT and XLNet models. The size of each question was made uniform with 20 tokens. The size

of the combined feature vector is set to be 16,000 for MCB 5000 for MFB and MUTAN. These figures were derived empirically as stated by the authors of the respective works. The number of LSTM steps used were 1024. For attention modules, 2 attention glimpses were used. We used the ADAM optimizer [24] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ with a learning rate of 0.001. Cross-Entropy Loss was used to calculate the error between the predicted and the actual answer. The model was trained for 100 epochs with a batch-size of 32. We used the Scikit-Learn package [36] to calculate the performance metrics. he codes for implementing different fusion techniques are available at MCB[1], VQA PyTorch[2], OpenVQA[3].

The implementation of the Decoder part of *MedFuseNet* is also in Pytorch. The code for the same is adapted from Image-Captioning-Pytorch[4]. We used the ADAM optimizer with a learning rate of $10e^{-4}$ and Cross-Entropy Loss Function to calculate the sequence generation loss. The model was trained for 30 epochs with a batch-size of 32. The BLEU-scores were evaluated using the NLTK Module[5].

For the first three baselines, the code was adapted from SAN-VQA[6]. For HiCAt, the code was adapted from HiCAt[7]. The code for BAN was adapted from ban-vqa[8]. The FasterRCNN features for BAN were extracted using the code available in FasterRCNN-Visual Genome[9]. All the implementation codes are in PyTorch and will be made publicly available.

[1] https://github.com/gdlg/pytorch_compact_bilinear_pooling
[2] https://github.com/Cadene/vqa.pytorch
[3] https://github.com/MILVLG/openvqa
[4] https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning
[5] https://www.nltk.org/_modules/nltk/translate/bleu_score.html
[6] https://github.com/Shivanshu-Gupta/Visual-Question-Answering
[7] https://github.com/karunraju/VQA
[8] https://github.com/jnhwkim/ban-vqa
[9] http://github.com/shilrley6/Faster-R-CNN-with-model-pretrained-on-Visual-Genome

4.6 Experimental Results

4.6.1 Comparison with the Baselines

We will now quantitatively evaluate the performance of *MedFuseNet* and compare it with the baseline models described in Chapter 4.4 for the tasks of answer categorization and answer generation.

The performance values of each model for answer categorization task with the MED-VQA dataset are summarized in Table 4.3. Comparing the accuracy scores for all three question categories, we can clearly see that *MedFuseNet* outperforms the BAN model. *MedFuseNet* achieves accuracy scores of 0.840 for category 1, 0.780 for category 2, and 0.746 for category 3. While the BAN model is more competitive to *MedFuseNet* model for category 3, it lags our model for category 1 by 2 percent and category 2 by 1.4 percent. In terms of AUC-ROC, BAN model outperforms *MedFuseNet* with a scores of 0.961 for category 1, 0.929 for category 2, while *MedFuseNet* leads with a score of 0.800 for category 3. For AUC-PRC scores, *MedFuseNet* outperforms all the baselines. This superior performance of *MedFuseNet* reckons that very basic models (like VIS + LSTM and Deeper LSTM + normalized CNN) may be insufficient to model the underlying patterns. On the other hand, more complicated attention mechanisms as presented in SAN and Hierarchical Co-Attention model might make the architecture more complex which requires more data to learn the parameters. This idea is again reinforced by the AUC-PRC scores as shown in Table 4.3. It clearly shows that simpler models like VIS + LSTM outperform the attention-based models. Alhough, BAN proves out to be a strong contender, *MedFuseNet* quantitatively outperforms the baselines due to the simpler model complexity and better captures the intricacies in the presence of limited amount of data in the medical domain. Another observation worth noting is the huge difference in the AUC-ROC and AUC-PRC scores of the model as shown in Table 4.3. This

shows that our model is comparably better in detecting true negatives, due to comparably high AUC-ROC score, than detecting true positives, because of the low AUC-PRC score. This can be attributed to the high class-imbalance.

Table 4.3: Comparison of *MedFuseNet* with the baseline models on MED-VQA answer classification dataset.

Methods	Accuracy			AUC-ROC			AUC-PRC		
	Modal	Plane	Organ	Modal	Plane	Organ	Modal	Plane	Organ
V+L[43]	0.704	0.701	0.652	0.899	0.851	0.775	0.478	0.453	0.456
d-L+n-I [29]	0.723	0.719	0.672	0.909	0.862	0.777	0.474	0.459	0.450
SAN [57]	0.669	0.729	0.669	0.926	0.870	0.783	0.459	0.415	0.406
HiCAt [30]	0.760	0.740	0.668	0.929	0.869	0.797	0.468	0.431	0.430
BAN [23]	0.820	0.766	**0.750**	**0.961**	**0.929**	0.800	0.600	0.521	0.456
MedFuseNet	**0.840**	**0.780**	0.746	0.942	0.901	**0.800**	**0.618**	**0.526**	**0.510**

For the PathVQA dataset with yes-no type answers, the accuracy scores of the baselines and *MedFuseNet* are presented in Table 4.4. Since the PathVQA data is very much balanced for yes and no type answers, and hence accuracy scores are a good metric to evaluate the performance of the models. There is no need for evaluating based on other metrics. Even with this dataset, *MedFuseNet* outperforms the baselines with an accuracy score of 0.636. Amongst other baseline methods, we can observe that the performance of SAN [57] and Hierarchical Co-Attention Networks [30] is competitive, while that of BAN [23] is relatively inferior. This could be attributed to the fact that the answer categorization task for PathVQA might not be inherently complex to justify the need for more complex models. Moreover, the performance of the BAN is highly dependant on the bounding boxes extracted from the pre-trained FasterRCNN model. These bounding boxes might not always be informative since the FasterRCNN model is pre-trained using real-world images dataset like Visual Genome [25]. Using such a model for pathological images might provide misleading results.

The second set of experiments are related to answer generation task using the MED-VQA abnormality category dataset and the open-ended answer types questions in PathVQA dataset.

Table 4.4: Comparison of *MedFuseNet* with the baseline models on PathVQA yes-no answer type dataset.

Methods	Accuracy
VIS + LSTM [43]	0.603
d-LSTM + n-CNN [29]	0.607
SAN [57]	0.627
HiCAt [30]	0.629
BAN [23]	0.604
MedFuseNet	**0.636**

The metric scores of these experiments are summarized in Table 4.5. To start with the MED-VQA dataset, we observe that *MedFuseNet* with the decoder performs better than the BAN model (with Decoder) for the metrics of BLEU-1 and BLEU-3 scores, while BAN (with Decoder) is better in terms of BLEU-2 and F-1 scores. This gives us an idea about the almost comparable performance of the two models on this dataset. Since there are 2-3 words on an average in the answer of the MED-VQA dataset, we do not have a clear winner since *MedFuseNet* is marginally better at a 3-gram level while BAN (with Decoder) leads at answer generation evaluation at the 2-gram level. For open-ended question-answer pairs of the PathVQA dataset, *MedFuseNet* with the decoder significantly outperforms the BAN model with decoder. It outperforms the baseline method in all the four metric scores with a BLEU-1 score of 0.605, BLEU-2 score of 0.303, BLEU-3 score of 0.073, and an F-1 score of 0.381. This establishes the supremacy of *MedFuseNet* for the answer generation part as well. In addition, it should be noted that our main contribution here is the integration of decoder to the basic VQA model and the decoder is flexible to be Incorporated into any other encoding structure beyond the one build in this work.

Table 4.5: Comparison of *MedFuseNet* with the baseline models on answer generation dataset.

Dataset	Method	BELU-1	BLEU-2	BLEU-3	F-1
MED-VQA	BAN + Decoder	0.266	**0.083**	0.013	**0.274**
	MedFuseNet + Decoder	**0.276**	0.076	**0.016**	0.229
PathVQA	BAN + Decoder	0.542	0.216	0.054	0.378
	MedFuseNet + Decoder	**0.605**	**0.303**	**0.073**	**0.381**

4.6.2 Ablation Study

To justify the importance of each component in *MedFuseNet*, we conduct an ablation study where we compare the performance of *MedFuseNet* against the various possible combinations of Image Feature - Question Feature - Fusion Technique for all the three categories of questions. We use 3 types of image features - VGG16, DenseNet121, and ResNet152, 2 types of question features - BERT and XLNet, and 3 types of fusion techniques - MCB, MUTAN, and MFB, along with the attention mechanisms. In total, there are 18 types of possible combinations that needed to be tested. The evaluation metric scores generated for each of the possible combination and question categories are summarized in Table 4.6. In terms of accuracy, *MedFuseNet* (BERT + ResNet + MFB) performs the best for question category 1 (Modality) with an accuracy of 0.840 and category 2 (Plane) with an accuracy of 0.780. Another close model for these two categories is BERT + DenseNet + MFB with 0.813 accuracy score for Modality and 0.757 for Plane. These scores suggest that image features are more generic for models with skip connections. Moreover, this asserts the power of MFB as a fusion model. For category 3 (Organs), the XLNet + ResNet + MFB combination achieves the best accuracy score of 0.844.

In terms of AUC-ROC scores, BERT + VGG16 + MFB performs the best with a score of 0.954 and is marginally ahead of *MedFuseNet* with a score of 0.942 for Modality. For category 2 (Plane), *MedFuseNet* again has the highest AUC-ROC score of 0.921. *MedFuseNet* also

performs well on category 3 questions with an AUC-ROC score of 0.800. The highest AUC-ROC score for category 3 is from BERT + ResNet + MUTAN with a value of 0.854. These figures demonstrate that *MedFuseNet* performs well with the inherent class imbalance in the data.

The trend for accuracy scores continues for AUC-PRC scores as well. *MedFuseNet* has the highest AUC-PRC for category 1 and category 2 with values of 0.618 and 0.526 respectively. In category 3, the highest AUC-PRC is for BERT + XLNet + MFB with 0.578 followed by *MedFuseNet* with a score of 0.510. This quantitative analysis establishes *MedFuseNet* as superior compared to all the other combinations with consistently performing and achieving the maximum scores for the majority of the metrics.

For a similar ablation study using the PathVQA yes-no type dataset, we observe that the combination of BERT + VGG16 + MFB performs best with an accuracy score of 0.645. This is followed by BERT + VGG16 + MUTAN and BERT + DenseNet121 + MFB with accuracy scores of 0.637 and 0.636, respectively. The combination of BERT + ResNet152 + MFB has an accuracy score of 0.621. This ablation study again strengthens the claim that the PathVQA dataset for yes-no type answers is not very complex, which is also supported by the results of the baseline methods. Thus, simpler models like VGG16 and BERT tend to perform better for the answer classification task with the PathVQA dataset.

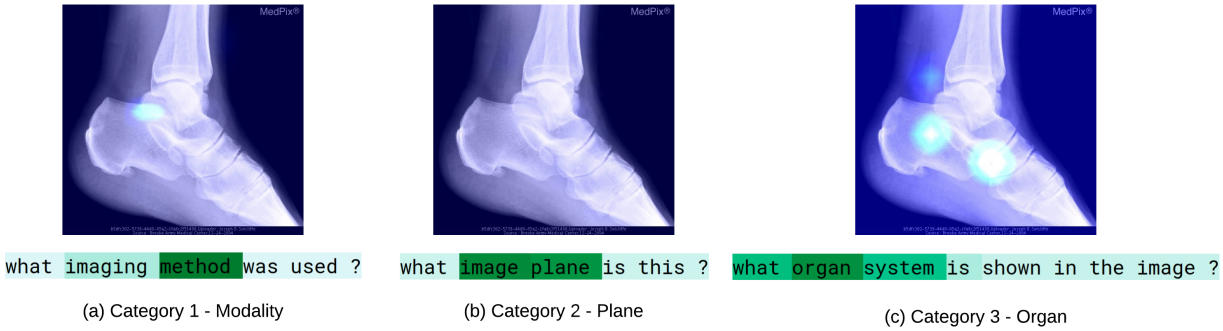
4.7 Attention Visualization

In this subsection, we perform the qualitative analysis of *MedFuseNet* and compare the results to the ones from SAN, and Hierarchical Co-Attention models. Since VIS+LSTM and Deeper-LSTM + Norm. CNN do not have any attention modules, hence we do not perform a qualitative analysis for the same. We visualized the image attention maps for each model

in various cases to understand the performance of the model. These interpretable results are summarized in Table 4.8. We have considered four cases, where each image belongs to a different organ system. This helps us interpret how well the model is learning the underlying nuances of the medical images. As mentioned in Chapter 4.5, we use two attention glimpses. For the first scan of the ankle, SAN can be seen to have a distributed attention span with a certain focus on the upper part of the ankle, while Hierarchical Co-Attention focuses on two different parts of the ankle. *MedFuseNet* has its attention maps spanned over the ankle joints and the lower bone. In the knee scan, SAN again fails to focus on the appropriate location in the image and has distributed attention. Hierarchical Co-Attention spans its attention to the posterior ligament. The *MedFuseNet* has a distributed attention span over the cartilage and the lower shin bone, also known as the tibia. This again supports the fact that *MedFuseNet* is able to attend to the crucial discriminatory parts of the organ. The third case is a radiology scan of the skull. *MedFuseNet* again has attention maps catered to both halves of the skull. The fourth case is a CT scan of the spine and contents. As we can see that from the attention maps that *MedFuseNet* is again able to focus on different parts of the scan, thus justifying the prediction. Thus, observing the visualization of the attention maps gives us interesting insights on where the models are focusing while trying to answer the questions related to the medical scans. *MedFuseNet* is able to learn the major distinguishing parts of the medical image which qualitatively justifies its metric scores.

In Figure 4.1, we analyze the co-attention schema of the *MedFuseNet* model by laying the image and question attention maps for a particular case over the input image and question. For the first category, we can see that model spans its attention over keywords like “method” in the question which shows that the model is learning to be aware of the modality. Similarly, Figure 4.1(b) shows how the model focuses on the keyword ”plane” in the category 2 question. Through the image attention maps, we can infer that model has an evenly

distributed attention to find the plane for the image. For category 3, again the question attention highlights the words like ”organ” and ”system”, thus, supporting the fact that the model knows where to span the textual attention. The image attention for category 3 also has a distributed attention span over multiple regions of the image.



(a) Category 1 - Modality (b) Category 2 - Plane (c) Category 3 - Organ

Figure 4.1: Co-Attention Maps for a sample case to display the attention span of *MedFuseNet* with the input image and the corresponding question attention. Figure (a) displays the image attention map and the corresponding question attention map for category 1 - modality, figure (b) for category 2 - plane, and figure (c) for category 3 - organ.

In Figure 4.2, we visualize the attention maps obtained from *MedFuseNet* while generating each word in the answer. As described in Chapter 3.3.2, for each time step t_i, the attention maps of the previous time step t_{i-1} are also fed into the LSTM. Figure 4.2 demonstrates the attention map that fed with each word to the model for three cases. The first case (a) is of sarcoidosis in the genitourinary organ system. Our model generates an extra word ”medullary” which is related to the medulla oblongata, located in the stem of the spinal cord near the skull. For the other two cases, the model predicts the answer correctly along with the punctuation of comma (,). The second case (b) is of a brain injury. In this case, we can observe how the model is attending different parts of the brain to discover the cause of injury. The third case (c) is of salter and harris fracture, a fracture specifically caused at the joint of two bones. As we can see in the attention maps that the model is specifically attending at the joint portion of the scan multiple times while generating the words ”salter-harris” and

"salter". This shows that the model is slowly and steadily learning to identify this special type of fracture and also localize it in the medical image.

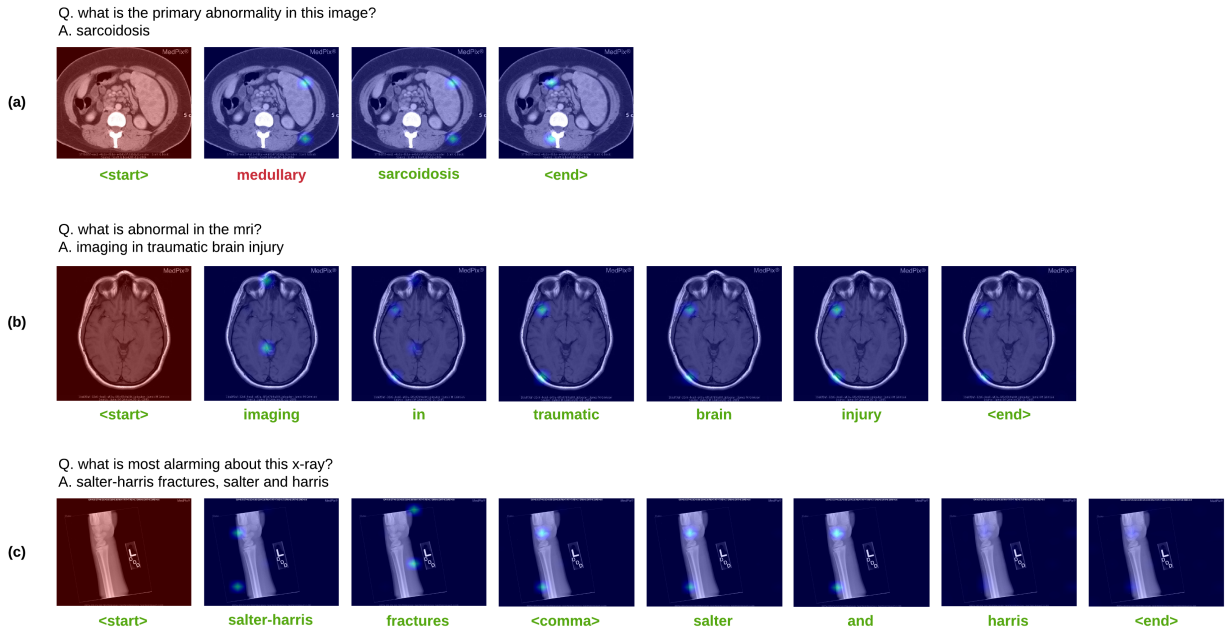


Figure 4.2: The attention maps produced by *MedFuseNet* while generating the words in the answer. There are three cases (a) sarcoidosis in the genitourinary system, (b) anoxic brain injury, and (c) salter-harris fracture in the bone.

Table 4.6: Performance metric scores for the ablation study experiments on MED-VQA dataset.

Accuracy

Question Category	Image Feature	MCB		MUTAN		MFB	
		BERT	XLNet	BERT	XLNet	BERT	XLNet
Category 1 Modality	VGG16	0.718	0.697	0.751	0.686	0.805	0.680
	DenseNet121	0.704	0.675	0.768	0.688	0.813	0.675
	ResNet152	0.731	0.663	0.783	0.716	**0.840**	0.701
Category 2 Plane	VGG16	0.706	0.697	0.750	0.605	0.749	0.629
	DenseNet121	0.719	0.643	0.754	0.643	0.757	0.655
	ResNet152	0.712	0.659	0.763	0.693	**0.780**	0.735
Category 3 Organ System	VGG16	0.718	0.625	0.785	0.683	**0.798**	0.692
	DenseNet121	0.753	0.630	0.774	0.696	0.774	0.720
	ResNet152	0.669	0.672	0.705	0.649	0.746	0.682

AUC-ROC

Question Category	Image Feature	MCB		MUTAN		MFB	
		BERT	XLNet	BERT	XLNet	BERT	XLNet
Category 1 Modality	VGG16	0.845	0.697	0.896	0.710	**0.954**	0.738
	DenseNet121	0.854	0.675	0.898	0.659	0.934	0.703
	ResNet152	0.861	0.703	0.906	0.740	0.942	0.700
Category 2 Plane	VGG16	0.833	0.697	0.866	0.718	0.899	0.729
	DenseNet121	0.832	0.743	0.867	0.801	0.894	0.839
	ResNet152	0.840	0.685	0.881	0.849	**0.921**	0.891
Category 3 Organ System	VGG16	0.655	0.619	0.689	0.622	0.691	0.730
	DenseNet121	0.667	0.700	0.691	0.626	0.690	0.650
	ResNet152	0.803	0.674	**0.854**	0.795	0.800	0.790

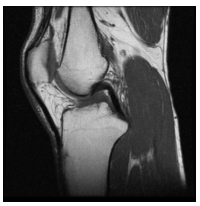
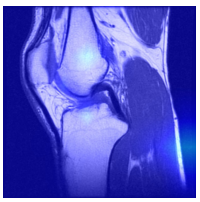
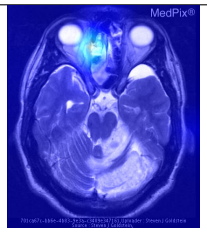
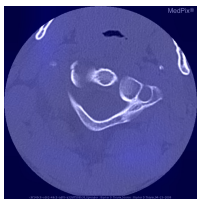
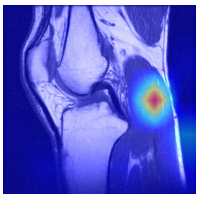
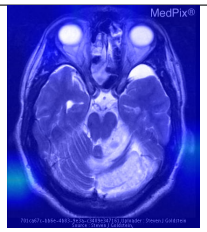
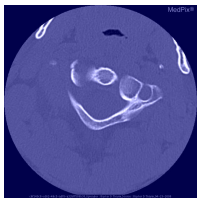
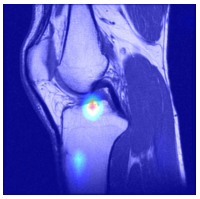
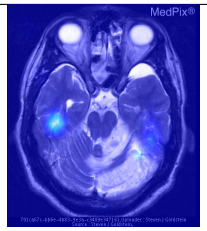
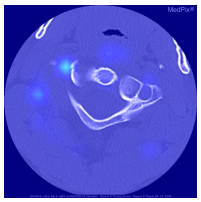
AUC-PRC

Question Category	Image Feature	MCB		MUTAN		MFB	
		BERT	XLNet	BERT	XLNet	BERT	XLNet
Category 1 Modality	VGG16	0.322	0.312	0.379	0.373	0.590	0.352
	DenseNet121	0.287	0.310	0.407	0.390	0.572	0.219
	ResNet152	0.361	0.208	0.469	0.343	**0.618**	0.224
Category 2 Plane	VGG16	0.252	0.368	0.331	0.370	0.439	0.288
	DenseNet121	0.269	0.279	0.347	0.335	0.437	0.351
	ResNet152	0.248	0.293	0.365	0.321	**0.526**	0.435
Category 3 Organ System	VGG16	0.341	0.348	0.393	0.289	0.443	0.351
	DenseNet121	0.364	0.420	0.377	0.289	0.433	0.330
	ResNet152	0.428	0.322	0.473	0.396	**0.510**	0.352

Table 4.7: Accuracy scores for the ablation study experiments of PathVQA yes-no answer type dataset.

Image Feature	MCB		MUTAN		MFB	
	BERT	XLNet	BERT	XLNet	BERT	XLNet
VGG16	0.614	0.502	0.637	0.513	**0.645**	0.507
DenseNet121	0.609	0.503	0.624	0.514	0.636	0.507
ResNet152	0.611	0.505	0.620	0.505	0.621	0.503

Table 4.8: Image Attention visualization for SAN, Hie. Co-Att, and *MedFuseNet*.

Method	musculoskeletal - ankle	knee	skull and contents	spine and contents
Original				
SAN [57]				
HiCAt [30]				
MedFuseNet				

Chapter 5

Conclusions

Efficient and effective Visual Questions Answering systems for medical images can be extremely helpful in providing the doctors with a second-opinion. In this thesis, we discussed the challenges of having a deep learning-based VQA system. After understanding each component of the problem, we presented the *MedFuseNet* which aims at maximizing the learning and minimizing the model complexity. This was verified by a rigorous quantitative and qualitative analysis of the model's performance using the MED-VQA dataset and the PathVQA dataset. The model is able to learn the essential components of a medical image and effectively answer questions related to it. We also explored the possible models for answer generation which has been a very sparsely researched topic in the VQA domain. In the medical domain, the challenges are even more severe due to the innate variance and vast vocabulary of the dataset. In many cases, there are not sufficient examples available in the dataset which makes the model training process more complicated. For instance, open-ended types of question-answer pairs in PathVQA and MED-VQA category-4 dataset have a huge disparity in the data present in the training and testing split. This causes a hindrance in the performance of the models over the testing split. Moreover, the vocabulary of the medical dataset varies vastly from that of the real-world dataset. Thus, VQA for the medical domain is very challenging and has immense scope for various future research directions to improve the model performance. The performance of the model for answer generation of the model can be improved by enhancing the decoder part of the model that caters to the challenges of

the problem statement. Another interesting aspect would be to generate and annotate more VQA datasets in the medical domain. This would be really beneficial to tackle the enormous variance, in terms of types of scans, organs, diseases, present in medical datasets.

Bibliography

[1] Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *CLEF2019 Working Notes. CEUR Workshop Proceedings*, pages 09–12, 2019.

[2] Aisha Al-Sadi, Bashar Talafha, Mahmoud Al-Ayyoub, Yaser Jararweh, and Fumie Costen. Just at imageclef 2019 visual question answering in the medical domain. *Working Notes of CLEF*, 2019.

[3] Imane Allaouzi and Mohamed Ben Ahmed. Deep neural networks and decision tree classifier for visual question answering in the medical domain. In *CLEF (Working Notes)*, 2018.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[5] David W Bates and Atul A Gawande. Error in medicine: what have we learned? *Annals of internal medicine*, 132(9):763–767, 2000.

[6] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620, 2017.

[7] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum

learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

[8] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[9] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

[10] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017.

[11] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018.

[13] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Conference on Empirical Methods in Natural Language Processing*, pages 457–468. ACL, 2016.

[14] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2019.

[15] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017.

[16] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv*, pages arXiv–2003, 2020.

[19] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

[20] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269. IEEE, 2017.

[21] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional

language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017.

[22] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *arXiv*, pages arXiv–1610, 2016.

[23] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1564–1574, 2018.

[24] DP Kingma and LJ Ba. Adam: A method for stochastic optimization. 2015.

[25] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.

[27] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

[28] Edward Loper and Steven Bird. Nltk: The natural language toolkit. *arXiv*, pages cs–0205028, 2002.

[29] Jiasen Lu, Xiao Lin, Dhruv Batra, and Devi Parikh. Deeper lstm and normalized cnn visual question answering model. *GitHub repository*, 6, 2015.

[30] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances in neural information processing systems*, pages 289–297, 2016.

[31] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in neural information processing systems*, pages 1682–1690, 2014.

[32] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. 2013.

[33] Nabil W Moukheibir. Universal computer assisted diagnosis, February 1 2000. US Patent 6,021,404.

[34] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[35] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[36] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[37] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods*

in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://www.aclweb.org/anthology/D14-1162.

[38] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid): A database for diabetic retinopathy screening research. *Data*, 3(3):25, 2018.

[39] Stephan Rabanser, Oleksandr Shchur, and Stephan Günnemann. Introduction to tensor decompositions and their applications in machine learning. *arXiv*, pages arXiv–1711, 2017.

[40] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1):18, 2018.

[41] F. Ren and Y. Zhou. Cgmvqa: A new classification and generative model for medical visual question answering. *IEEE Access*, 8:50626–50636, 2020.

[42] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In *Advances in neural information processing systems*, pages 2953–2961, 2015.

[43] Mengye Ren, Ryan Kiros, and Richard Zemel. Image question answering: A visual semantic embedding model and a new dataset. *Proc. Advances in Neural Inf. Process. Syst*, 1(2):5, 2015.

[44] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*, pages arXiv–1409, 2014.

[46] Aman Srivastava. embedding-as-service. https://github.com/amansrivastava17/embedding-as-service, 2019.

[47] Damien Teney and Anton van den Hengel. Zero-shot visual question answering. *arXiv*, pages arXiv–1611, 2016.

[48] Minh H Vu, Tommy Löfstedt, Tufve Nyholm, and Raphael Sznitman. A question-centric model for visual question answering in medical imaging. *IEEE Transactions on Medical Imaging*, 2020.

[49] Guotai Wang, Wenqi Li, Maria A Zuluaga, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE transactions on medical imaging*, 37(7):1562–1573, 2018.

[50] Ping Wang, Tian Shi, and Chandan K Reddy. Text-to-sql generation for question answering on electronic medical records. In *Proceedings of The Web Conference 2020*, pages 350–361, 2020.

[51] Zhengyang Wang and Shuiwang Ji. Learning convolutional text representations for visual question answering. *Proceedings of the 2018 SIAM International Conference on Data Mining*, page 594–602, May 2018. doi: 10.1137/1.9781611975321.67. URL http://dx.doi.org/10.1137/1.9781611975321.67.

[52] who.int. Stats and analysis, March 2019. URL https://www.who.int/gho/health_workforce/physicians_density/en/.

[53] Sam Wiseman and Alexander M Rush. Sequence-to-sequence learning as beam-search optimization. In *EMNLP*, 2016.

[54] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[55] Xin Yan, Lin Li, Chulin Xie, Jun Xiao, and Lin Gu. Zhejiang university at imageclef 2019 visual question answering in the medical domain. *Working Notes of CLEF*, 2019.

[56] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763, 2019.

[57] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016.

[58] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 1821–1830, 2017.