Pattern Recognition Letters 150 (2021) 57–75

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Visual question answering in the medical domain based on deep learning approaches: A comprehensive study

Aisha Al-Sadi[a], Mahmoud Al-Ayyoub[a,*], Yaser Jararweh[b], Fumie Costen[c]

[a] Jordan University of Science and Technology, Irbid, Jordan
[b] Duquesne University, Pittsburgh, PA, USA
[c] University of Manchester, Manchester, UK

ARTICLE INFO

Article history:
Received 11 July 2020
Revised 21 January 2021
Accepted 2 July 2021
Available online 20 July 2021

Edited by : Maria De Marsico

Keywords:
Medical visual question answering
Planes questions
Organ systems questions
Modality questions
Abnormality questions
Transfer learning
Data augmentation
Multi-Task learning
Global average pooling
Ensemble

ABSTRACT

Visual Question Answering (VQA) in the medical domain has attracted more attention from research communities in the last few years due to its various applications. This paper investigates several deep learning approaches in building a medical VQA system based on ImageCLEF's VQA-Med dataset, which consists of about 4K images with about 15K question-answer pairs. Due to the wide variety of the images and questions included in this dataset, the proposed model is a hierarchical one consisting of many sub-models, each tailored to handle certain questions. For that, a special model is built to classify the questions into four categories, where each category is handled by a separate sub-model. At their core, all of these models consist of pre-trained Convolution Neural Networks (CNN). In order to get the best results, extensive experiments are performed and various techniques are employed including Data Augmentation (DA), Multi-Task Learning (MTL), Global Average Pooling (GAP), Ensembling, and Sequence to Sequence (Seq2Seq) models. Overall, the final model achieves 60.8 accuracy and 63.4 BLEU score, which are competitive with the state-of-the-art results despite using less demanding and simpler sub-models.

1. Introduction

With the advances in the computer vision (CV) and natural language processing (NLP) fields, a new challenging task is proposed, which is Visual Question Answering (VQA), grabbing the attention of both research communities. VQA is about answering a specific question about a given image. Thus, there is a need to combine CV techniques that provide an understanding of the image's content with NLP techniques that provide an understanding of the question and the ability to produce the answer. The difficulty level of the problem depends on the expected answer types whether they are yes/no, multiple choice or open-ended.

Recently, VQA has been applied to different specific domains such as the medical domain. VQA in the medical domain has many applications such as helping clinicians in decision making in order to enhance their confidence, image interpretation for medical students, automated system for disease diagnosing, and to answer patients' questions that do not require a special visit to a doctor.

Medical VQA poses its own set of issues/challenges that are different from the ones faced in general-domain VQA. Some of these challenges are related to the processing of medical images and the difficulties in handling all kinds of images for different body parts and extracting regions of interest that vary greatly for the different medical cases and ailments. The other set of challenges are related to the understanding of the questions and the ability to process very technical medical terms as well as non-medical terms used by common users. The resources required to address all these challenges can be massive and there are many restrictions related to using them and integrating them into a single model.

The year 2018 witnessed the inauguration of a special challenge for VQA in the medical domain under the name: the VQA-Med challenge, which was organized by the reputable ImageCLEF conference [1]. The best Bilingual Evaluation Understudy (BLEU) [2] score achieved by the five participating teams was 16.2 [3], which is a very low score. However, this is expected due to the task difficulty. Moreover, the work on this problem is still in its early stages, and, with time, it is expected to improve by creat-

* Corresponding author.
E-mail addresses: asalsadi16@cit.just.edu.jo (A. Al-Sadi), maalshbool@just.edu.jo (M. Al-Ayyoub), jararwehy@duq.edu (Y. Jararweh), fumie.costen@manchester.ac.uk (F. Costen).

https://doi.org/10.1016/j.patrec.2021.07.002
0167-8655/

ing more reliable datasets and models. The problem with the 2018 dataset is that it encompasses several medical concepts within a rather small dataset. In 2019, the second instalment of the VQA-Med challenge [4] was launched with a dataset that is even more comprehensive and diverse. It consists of four question/data categories, aiming to provide a complete medical VQA system. This dataset is the one we consider.

This work aims to gauge the effectiveness of different deep learning techniques in solving the task at hand. Therefore, an accurate and efficient medical VQA system is built consisting of several sub-models, where each sub-model is specialized in answering a specific category of questions. These sub-models vary in the used deep learning techniques and each is a result of extensive experimentation in order to reach the best performance for its respective category. These techniques include using pre-trained CNN models (with and without auxiliary information), Data Augmentation (DA), Global Average Pooling (GAP), and ensembling. All of these techniques improve the effectiveness of the overall model in solving the VQA-Med 2019 challenge. On the other hand, we also perform experiments using techniques that seem promising, but result in no improvement, such as Sequence to Sequence (Seq2Seq) models (with the encoder-decoder architecture, image captioning models, and attention mechanisms) in addition to some advanced techniques such as Multi-Task Learning (MTL) and Generative Adversarial Networks (GAN).[1] The contributions of this work are as follows.

- We present a medical VQA model based on the VQA-Med 2019 challenge. This model is simpler than the state-of-the-art (SOTA) models while achieving competitive performance.
- While building this model, we explore the use of many cutting-edge techniques, some of which are shown to be useful while the other are not. This will help guide future research in this area.
- We perform an extensive set of experiments on the presented model and its components and analyze their results. We investigate the model's weaknesses and present justifications for some of the failed cases.

The rest of this paper is organized as follows: Section 2 presents the most relevant work. Section 3 presents a detailed analysis of the dataset, which we find useful in building our models. In Section 4, we present the proposed models that achieve the best performance for answering questions of each category. These models include basic image classification models, image classification models with auxiliary information, DA models, GAP models, and ensemble models. Section 5 lists and analyzes the experiment results for the proposed work in Section 4. Section 6 summarizes the achieved work and lists the possible work that can be done in the future. Finally, the appendices contain more details about the dataset and other proposed techniques that do not yield good performance including MTL models, Seq2Seq models with their variations, and the GANs models.

2. Related works

The general VQA challenge[2] which is held every year starting from 2016, is based on a large dataset of real-world images with different question types such as yes/no questions, questions about quantities, etc. Different approaches were applied for the task and most solutions rely on deep learning techniques. These techniques combine the use of word embedding with different recurrent neural networks (RNNs) for text embedding and features extraction, and convolution neural networks (CNNs) for visual features extraction supplemented with advanced techniques such as attention mechanisms.

Several approaches have been proposed for the general domain VQA problem [5,6]. One of the machine learning based models for VQA is the one by Kafle and Kanan [7]. In this model, the image features and the questions features are fed into a Bayesian classifier, logistic regression classifier, or simple neural network. On the other hand, most of the work on VQA use CNNs for image representation and word embeddings for text (i.e., question and answer) representation, while exploiting the benefits of pre-trained models such as VGG16 [8] and ResNet [9] for extracting low and high level features of images, and word2vec [10] and GloVe [11] for text embedding in order to capture the relationships between words. Then, they usually combine both representations to answer the question. Among such approaches are [12,13]. These approaches have a limitation which is the use of all the image features in producing the answer. In most cases, only certain parts of the images are necessary for answering while the remaining parts can actually present noisy information and reduce the quality of the answer. To overcome this issue, attention mechanisms are used [14,15]. There are more advanced attention techniques such as Stacked Attention Networks (SANs) [16] and Hierarchical Co-attention [17].

2.1. Medical VQA

For the medical domain, the task is different as the nature of medical images requires knowledge in the medical domain in order to understand them. So, a special shared task or competition is dedicated to it under the name VQA-Med. The first version of this competition was held in 2018 [3] and the second version was held in 2019 [4]. In the following subsections, we discuss each of these competitions and the approaches used in their participants.

2.1.1. VQA-Med 2018

In the first version, the dataset consisted of 2866 medical images and 6413 questions-answers pairs divided into training, validation, and testing sets. Two medical experts manually checked the automatically generated questions and answers for each image. The question types are mixed including questions about a "region" within the image, what the image shows, abnormalities shown in the image, image type, etc.

Five teams submitted their work, most of which were based on deep learning techniques. They used pre-trained CNN models (such as VGG16 and ResNet) to extract image features. Some of them used the encoder-decoder architecture with different components such as Long Short-Term Memory (LSTM) or Bidirectional LSTM (Bi-LSTM), with or without attention. In addition, some teams used more advanced techniques such as the SANs and Multimodal Compact Bilinear (MCB) pooling. The following paragraphs discuss the work of each of these teams.

The top team in 2018, Team UMMS [18], used ResNet-152 to extract image features, and a pre-trained word embedding on PubMed articles, Pittsburgh clinical notes, and Wikipedia pages for text features. They created multiple attention maps using a co-attention mechanism between image features and text features. Then, they generated answers using a sampling method treating the task at hand as a classification task. Their best run yielded a BLEU score of 16.2.

The second team in the 2018 competition, Team TU [19], provided two models. In the first model, which is the same architecture as [20], they used the pre-trained Inception-ResNet-v2 model to extract image features and Bi-LSTM instead of LSTM as [20]. In their second model, they computed the attention between the image features and the question features and concatenated it with

[1] Code for the final models will be made available through a public GitHub repository.

[2] https://visualqa.org/index.html.

the question features before feeding it to a Softmax layer for prediction. Their best run yielded a BLEU score of 13.5.

Similar to the second team, the third team, Team NLM [21], also created two models. For the first model, they used a SAN with VGG16 for image features and LSTM for question features. As for the second model, they used MCB pooling with ResNet-50 and ResNet-152 for image features and 2-layer LSTM question features. In the SAN model, they computed the attention over the image, then combined the image features and question features for the second attention layer. Then they passed it to a Softmax layer as a classification problem. For the MCB model, they fine-tuned ResNet-50 and ResNet-152 on external medical images, then they combined the image features and question features to create a multimodal representation to predict the answer. Their best run yielded a BLEU score of 12.1.

The fourth team, Team JUST [20], used VGG16 for image features extraction. They used an LSTM-based encoder-decoder model where they fed the question to the encoder and then concatenated the hidden state of the encoder with the image features to feed them to the decoder as the initial hidden states. Their best run yielded a BLEU score of 6.1.

The final team, Team FSST [22], treated the task at hand as a multi-label classification problem. They extracted image features using VGG16. They computed the question's word embeddings and fed them to a Bi-LSTM network in order to extract question features. Then, they concatenated question features and image features and fed them to a decision tree classifier. Their best run yielded a BLEU score of 5.4.

2.1.2. VQA-Med 2019

For the second version of this competition, which took place in 2019, the dataset was larger, more diverse and better organized. Moreover, unlike the 2018 dataset, only the test set was manually checked by medical experts. The organizers used exact match accuracy in addition to BLEU scores to rank the competitors. Since this dataset is what we use in this work, more details about it can be found later in this paper (Section 3).

Overall, 16 teams participated and their best result reached 62.4 accuracy and 64.4 BLEU score. All teams used deep learning techniques, where the use of pre-trained CNNs models (such as VGG16, ResNet-50, ResNet-152, and Inception ResNet) for image features extraction is the main part of all models. Similar to the 2018 teams, for the question features, most teams used word embeddings such as word2vec pre-trained on medical data. However, the best model's approach in question features extraction used Bidirectional Encoder Representations from Transformers (BERT) [23], which is based on the Transformer model [24]. To concatenate image features and question features, pooling approaches, such as Multimodal Factorized Bilinear (MFB) pooling and Multimodal Factorized High-order (MFH) pooling, are used by the teams with the best results, while traditional concatenation techniques using the encoder-decoder model are used by the other teams. Moreover, most teams supplemented their approaches with some attention mechanism. Descriptions of the work of the top five teams are as follows. In the following paragraphs, we discuss the most interesting approaches and results on the VQA-Med 2019 dataset.

The top team in the 2019 competition (which represented the SOTA at the time this paper was written) was a team of researchers from Zhejiang University, China, and National Institute of Informatics (NII), Japan. The team used two names: Hanlin and yan. This team used a modified version of VGG16 model with GAP strategy [25] for image features extraction and, for the question features, they used the BERT model. Then, they concatenated both features using MFB pooling with a co-attention mechanism. Their best run yielded 62.4 accuracy and 64.4 BLEU score [26].

The second team, Team Minhvu [27], processed the data first by removing texts and backgrounds from images and applied DA techniques such as random rotation and scaling. They used ResNet-152 for the image features and BERT for question features. Then, they applied the attention mechanism and passed the output to a bilinear transformation unit to output the answer. Their best run yielded 61.6 accuracy and 63.4 BLEU score.

The third team in the 2019 competition, Team TUA1 [28], augmented data randomly with rotate, shear, and shift operations. They built two models, the first was the classifier model for the plane, organ, and modality questions, while the other was dedicated to the abnormality questions called the generator model. In both models, the Inception-Resnet-v2 model was used for image features extraction, and the BERT model for question features extraction. For the classifier model, they concatenated the features using Multi-Layer Perceptron followed by a Softmax layer to predict the answer. On the other hand, for the generator model, they concatenated the features and passed them to an LSTM to produce the answer word-by-word and used beam search to get the final answer. Their best run yielded 60.6 accuracy and 63.3 BLEU score. It should be noted that, in parallel to our work, this team has been working on improving their results. They recently published a paper [29] with a new model called CGMVQA. This model uses several techniques, such as multi-head self-attention mechanism, data augmentation and tokenization on texts. Moreover, the model employ pre-trained ResNet-152 to extract image features in addition to three kinds of embeddings together to deal with texts. Their new results represent a new SOTA results with 64 accuracy and 65.9 BLEU score.

The top team from the 2018 competition, Team UMMS [30], participated in the 2019 and it was ranked in the fourth place. They used an approach similar to the one that won them the first place in the 2018 competition. For question features, they used a pre-trained word-embedding on PubMed articles and clinical notes from MIMIC-III database. As for the image features, they used the ResNet-152 model. Two additional inputs were created which are the question category (out of the four categories under consideration) and the question topic (out of ten topics they used) as determined by the Entity Topic Modeling (ETM) technique of [31]. Then, they calculated the attention between all the features. After that, they used MFH pooling for features concatenation to predict the answer. The first three categories (plane, organ, and modality) were considered as a single classification task, whereas they dealt with the abnormality questions as a multi-label classification task. Their best run yielded 56.6 accuracy and 59.3 BLEU score.

Our team's participation, Team JUST [32], was ranked in the fourth place and the results we obtained then were not far from the best known results at the time with 57 accuracy and 59.1 BLEU score. In our participation, we used the basic image classification model for the four categories (planes, organs, modalities, and abnormalities) in addition to the basic encoder-decoder model for the abnormalities category, and images similarity model for the same category. In this paper, we investigate different approaches to address the problem at hand starting from the basic image classification used in [32]. We build new models based on it by leveraging extra information from other categories, augmenting the data, using other ideas from the literature, and employing ensembling techniques. At the same time, we experiment with more models based on the Seq2Seq architecture for the abnormality category, by adding the attention mechanism to our old model in [32]. We also build an image captioning model with and without using the attention mechanism. Other advanced techniques are used in our models which are MTL and GANs. We use two MTL models for the planes and organs categories, by sharing all layers or sharing some layers. While we started experimenting with generating data with basic GANs and Deep Convolutional GAN (DCGAN) [33] for or-

gans images, however, due to resource limitations, these efforts are postponed as future work.

Another notable effort was by the IBM Research AI team [34], which merged the original training and validation data, shuffled them and re-sampled a new training and validation sets with 95%-5% split. They proposed a model called Supporting Facts Network (SFN), which takes an additional input parameter of the image size. The image features were extracted using the VGG16 model, while the question features were extracted using GloVe embedding followed by LSTM. The attention between the features were calculated using the multimodal pooling techniques. Then, they classified the answer directly for the first three question categories of plane, organ, and modality, where each category had a dedicated model. As for the abnormality category, its model exploited the information from the other categories as supporting facts. Their best run yielded 55.8 accuracy and 58.2 BLEU score.

Finally, a team of researchers from NITC, India, considered a subset of the VQA-Med 2019 dataset and presented their Modality based Medical Image Visual Question Answering (MoBVQA) system [35]. The MoBVQA system is a simple CNN that consists of three convolution layers, each followed by a max pooling layer. After that, a dropout layer is used followed by two fully connected layers then a final dense layer that uses Softmax to give the final output. They worked only on the modality questions and limited their attention to the main eight categories. As can be seen from Table A.16, the dataset contained 36 subcategories and the all papers mentioned so far (including our work) are evaluated based on these subcategories. This makes MoBVQA incomparable to the other papers. The only work we can compare MoBVQA with is ours. While the reported validation accuracy for MoBVQA is 83.3%, our model's is 90%.

3. Dataset

The dataset used in VQA-Med 2019 was generated from the MedPix[3] database. It consists of: 3,200 medical images with 12,792 Question-Answer (QA) pairs as training data, 500 medical images with 2000 QA pairs as validation data, and 500 medical images with 500 QA pairs as test data. The data is equally distributed over the four categories of plane, organ, modality, and abnormality categories. Each image has a question in each one of these four categories. The images sizes vary a lot with the smallest image size being 106×109 and the largest one being 2268×2040. However, most images sizes are within 1000×1500.

After careful inspection of the dataset, the first thing we observed is that the question category can be determined directly from the question words. I.e., if the word 'plane' appears in the question, then, this is a plane question. On the other hand, words like 'organ' or 'part' indicate that the question is from the organ type. Similarly, abnormality questions have words like 'normal', 'abnormal', 'alarm' and 'wrong'. If none of these conditions applies, then it is a modality question. This proves to be very useful for test data questions since the category of the question is not explicitly given as is done in the training and validation questions.

3.1. Plane category

Questions on planes come in one of the following forms: "in which plane", "which plane", "what plane", "in what plane", "what is the plane", "what imaging plane is", and "what image plane". There are 16 planes under consideration. The list of planes along with sample images are included in Appendix A. Table 1 shows the main planes and their distributions in training and validation data.

[3] https://medpix.nlm.nih.gov.

Table 1
Planes distribution in training & validation data.

Plane Type	Training Data	Validation Data
Axial	49%	43%
Sagittal	15%	16%
Coronal	12%	13%
Anteroposterior (AP)	6%	7%
Lateral	5%	6%
Frontal	4%	4%
Other	9%	11%

Table 2
Organ systems distribution in training & validation data.

Organ Type	Training Data	Validation Data
Skull and Contents	38%	35%
Musculoskeletal	13%	15%
Gastrointestinal	11%	12%
Lung, mediastinum, pleura	8%	7%
Spine and contents	7%	10%
Genitourinary	7%	6%
Face, sinuses, and neck	6%	6%
Vascular and lymphatic	4%	4%
Heart and great vessels	4%	3%
Breast	2%	2%

As evident in this table, the data is imbalanced, with some planes being more frequent than the others. Moreover, the table shows that the training data and the validation data have slightly different distributions. In fact, these observations are noticeable across all categories.

3.2. Organ systems category

Questions on organ systems come in one of the following forms: "what part of the body is", "the CT/MRI/Ultrasound/X-Ray scan shows what organ system", "which organ system is", "what organ system is", etc. There are ten organ systems; organ systems list with sample images are in Appendix A. Table 2 shows all organ systems and their distribution in training and validation data. Similar to the plane category, data in this category is imbalanced with different training and validation distributions.

3.3. Modality category

There are eight main modality categories:

- XR: X-ray.
- CT: Computer tomography.
- MR: Magnetic resonance imaging
- US: Ultrasound.
- MA: Magnetic resonance angiography.
- GI: Gastrointestinal.
- AG: Angiography.
- PT: Positron tomography.

Under each of these categories, there is a number of subcategories as follows. Each of XR and MA has one subcategory, while each of US, AG, and PT has two subcategories. GI has four subcategories, CT has seven subcategories, and, finally, MR has 17 subcategories. Modality main categories with their subcategories are listed in Appendix A.

The questions on the modality part are more diverse. We can classify them into four types as follows.

- Type 1: Questions whose answer is one of the main modality categories and its subcategory. Examples include "what modality was used to take this image", "how was this image taken", "what kind of image is this", etc.

Table 3
Modality questions distribution.

Type	Training Data	Validation Data
Type 1	43%	46%
Type 2	37%	36%
Type 3	14%	14%
Type 4	6%	4%

Table 4
Modality main categories distribution in training & validation data.

Main Category	Training Data	Validation Data
XR	14.3%	17.2%
MR	39.2%	37.4%
CT	27.4%	24.6%
US	6%	4%
AG	2.5%	3.6%
GI	1.7%	2.6%
MA	0.75%	1.6%
PT	0.66%	0.2%
Can't infer	7.4%	8.8%

- Type 2: Yes/no questions. Examples include "is this an mri image", "was gi contrast given to the patient", etc.
- Type 3: Questions whose answer is one of the choices explicitly mentioned in the question itself. Examples include "is this a contrast or noncontrast ct", "is this a t1 weighted, t2 weighted, or flair image", etc.
- Type 4: Questions whose answer is among one, two or three choices that are not explicitly mentioned in the question. Examples include "what type of contrast did this patient have", "what is the mr weighting in this image", etc.

Table 3 shows modality question types distribution in training data and validation data.

Table 4 shows the distribution of images of each main modality category from all question types. Note that, in some cases (238 in the train data and 44 in the validation data), we are unable to infer the modality from the question, such as the case with the question "is this an mri image" which has "no" as the answer.

Due to the variations in modality question types and the large number of subcategories for some modality categories, we dedicate considerable effort to the modality determination problem and present several ways/models to address it. We re-formulate the problem in different ways in order to focus on specific aims for each model.

3.4. Abnormality category

Questions of this category come in one of the following forms.

- Type 1: Questions asking about abnormality in the image; for example, "what is wrong/alarming/the abnormality in this image". This type represents 97% of abnormality questions and 95% of abnormality validation questions.
- Type 2: Questions with yes/no answers such as "is this image normal" or "is this image abnormal". This type represents 3% of abnormality training questions and 5% of abnormality validation questions.

For Type 1 questions, there are 1461 different abnormalities in the 3082 training images, and 407 different abnormalities in the 477 validation images, i.e., a total of about 1600 different abnormalities in the whole data. Moreover, there are labels in the validation data that are not in the training data and vice versa.

3.5. Dataset errors

It is worth mentioning that the dataset has wrong answers for some questions that might affect the accuracy of any model trained on it. This is expected since the data was generated automatically with only the test part undergoing manual checking by medical experts [4], which makes this problem a distant supervised learning problem rather than a traditional supervised learning problem [36]. Even for non-medical people like ourselves, we can detect some obvious errors such as organ systems errors and we can easily fix them. However, we are not able to fix all errors on our own. In fact, in some cases, we know that the label is wrong, but we are not sure about the correct label. Fig. 1 shows some samples of these cases. Consequently, medical experts are needed to determine all wrong answers and correct them.

Regarding the test data, the task organizers state that the answers were checked by a medical doctor and a radiologist [4]. One important issue to note about the test data is that, in some cases, multiple answers are considered correct. For example, the organ system answer for a certain image can be either 'Skull and contents' or 'Face, sinuses, and neck'; both of which are considered true. Another example is the case of 'lung, mediastinum, and pleura' and 'heart and great vessels', both can be correct for the same image. However, such cases are not present in the training and validation data, which might negatively affect the accuracy on models trained on them as such models may learn to exclude or reject correct answers. More discussion of this issue and its effect is presented in Section 5.2.

4. Methodology

Since there are different types of questions from different categories in this work, a special model for each category is created. These models are combined into one model to be used for predicting answers. In order to use them correctly to answer a given question about a given image, we need to detect the suitable model to answer the question on the image using the question words.

A model is built to classify the question category. It is a rule-based model that does not require training (i.e., it is not a machine/deep learning model). The simplicity and efficiency of this part of our work is one of the advantages it has over other works that rely on heavy word embedding techniques. Fig. 2 shows how this model determines the question category. This process is rather simple for the plane category, the organ systems category, and both types of the abnormality category. However, for the modality category, there are many considerations into determining the required modality type correctly. For simplicity, these details are not shown in the figure. The accuracy of this model is 100% for the dataset at hand. I.e., It is able to classify any question to its correct category. For the selected question category, the appropriate model is then used to generate the answer. Fig. 3 shows a flowchart of the entire model.

It should be noted that our model is simpler than existing models for two reasons. The first one is the way it process each question. Instead of complying heaving word embedding techniques, our model uses a simple yet effective keywords spotting technique, which proved to be perfect for a task like the one at hand. The other reason is that the core components of our model rely on a pre-trained model (VGG16) that is considered lighter and simpler than what most other models use, which is ResNet and its derivatives.

This section explains the main proposed models to solve the answer prediction task. Some of the proposed ideas are applied to all categories, while the rest are only applied to specific questions categories. The specific details about the architecture of these mod-

Fig. 1. Organ systems errors samples. The left image is labeled as 'Breast' organ while it should be labeled as 'Skull and contents' organ. The middle image is labeled as 'Skull and contents' organ while it should labeled as 'Vascular and Lymphatic' organ. Finally, the right image is labeled as 'Vascular and Lymphatic' organ while it should be labeled as either 'Lung, Mediastinum, and Pleura' organ or 'Heart and great Vessels' organ. A medical expert is needed to determine which one of these two options is the correct one. This is an example of the errors in the dataset that cannot be corrected with a medical expert.

Algorithm 1: Prediction Steps

```
Input: Image and Question
  if Plane word in the question then
    Predict image plane using the best results Plane model
  else if Organ or Part Words in the question then
    Predict organ plane using the best results Organ model
  else if Normal, Abnormal, Alarm, or Wrong words in the questions then
    if question starts with 'is this' or 'is there' or 'does this' or 'is the' or 'are there' then
      Predict using Normal/Abnormal model and answer yes/no based on that
    else
      Predict using the best results abnormality model
    end if
  else
    if Modality type-1 question then
      Predict main modality category using MM1 model
      Predict subcategory model based on the predicted main category from models MM7-MM11
    else
      Predict Answer using models MM2-MM6 based on what the question ask about
    end if
  end if
Output: Answer
```

Fig. 2. Prediction steps for the final VQA medical system.

els are determined empirically after performing a very large number of experiments using different configurations for each model.

In this work, we present four approaches to build a medical VQA system. They are as follows. The first approach is to address this problem as an image classification task. The second approach extends the first one by feeding the image classification models with extra information. The third approach aims to improve the results by using data augmentation techniques. The fourth and final approach takes the core idea of the top team at the 2019 VQA-Med challenge and combine it with previous models in order to get better results. The other approaches that we use without success (such as Seq2Seq, MTL and GANs) are explained in Appendix B.

4.1. Basic image classification

In all four categories, we notice that the images are enough to determine the answers and the sole purpose of the questions is simply reduced to determining the suitable model for answering. Therefore, an image classification model is created for each question type.

Since the dataset used in this work is rather small, pre-trained models are suitable due to their benefits in detecting low-level features, despite the fact that it has been trained on images from a different domain. The pre-trained VGG-16 network, winner of ImageNet challenge of 2014 [37], is used in this work. The following is the description for creating an image classification model for each category.

4.1.1. Planes image classification model

We use the pre-trained model VGG16 with some modeifications. Specifically, we remove the last layer (the Softmax layer) of VGG16 and freeze all layers of VGG16 except for the last four. Freezing a layer means using its generated weights from training the original model on the ImageNet dataset. These layers are useful in detecting low-level features. We do not use the pre-trained weights for the last four layers (i.e., we do not freeze them), since

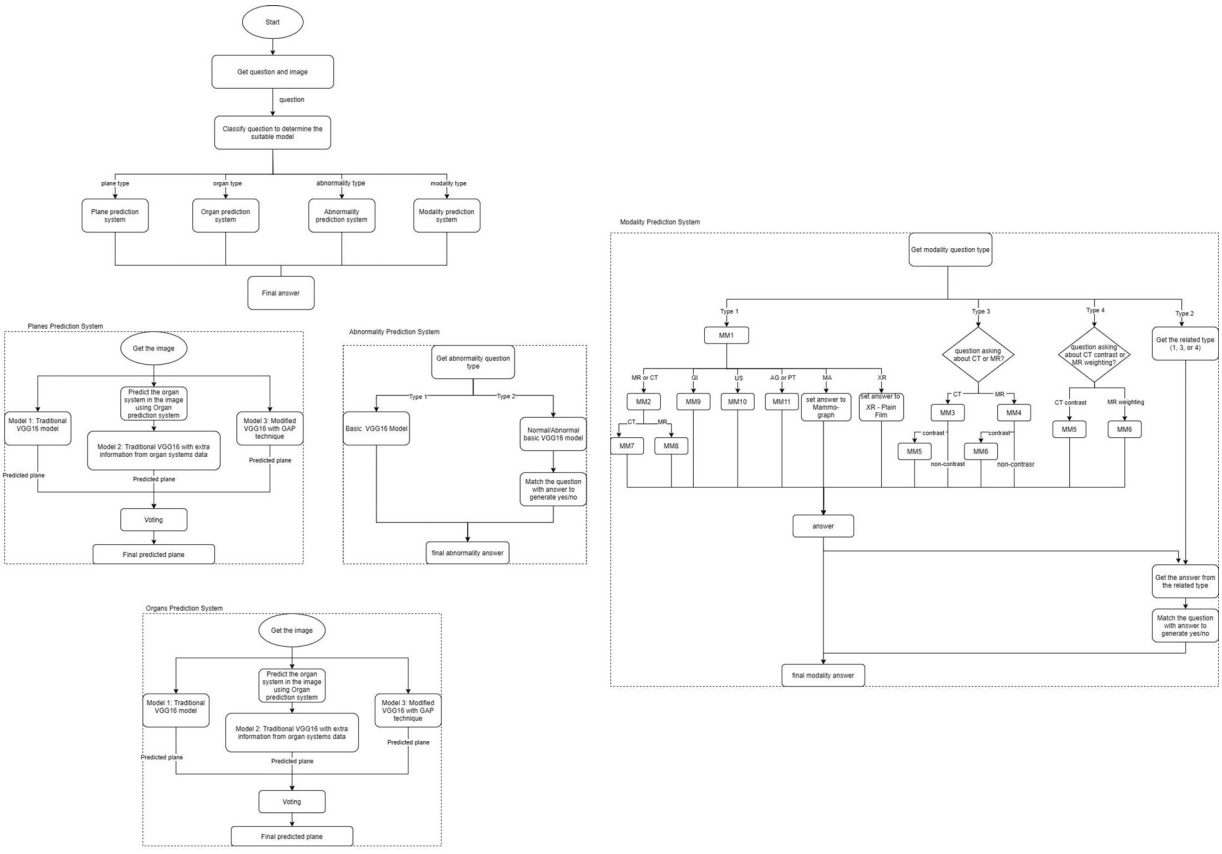

Fig. 3. Flowchart of the final VQA medical system.

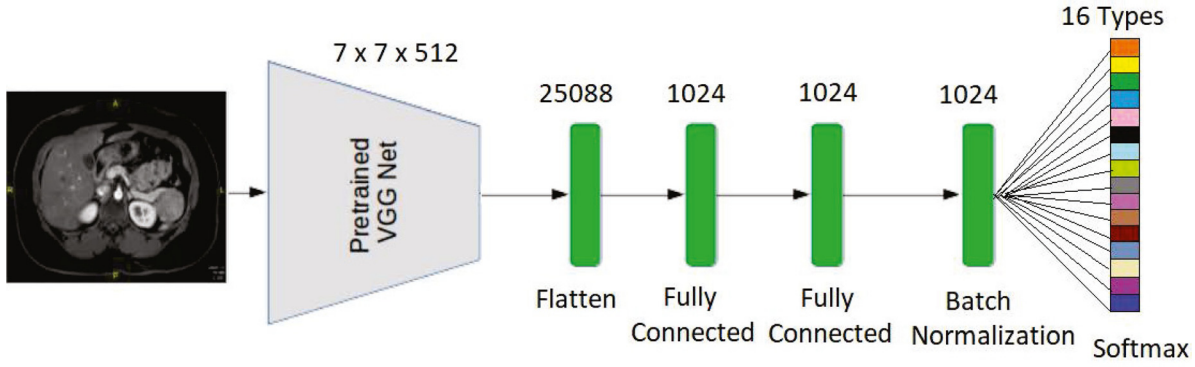

Fig. 4. Planes image classification model architecture.

in our case the data belongs to the medical domain, which is different from the ImageNet dataset domain. Hence, adjusting (retrain) the last four layers in the model to generate new weights according to our data is preferred. The output from this part is passed to two fully-connected layers with 1024 hidden nodes. Finally, the output is passed to a Softmax layer with 16 classes (the different 16 planes in the dataset). Since the data is imbalanced, we use class weights in order to give the classes with smaller numbers of images higher weights. Fig. 4 shows the plain model architecture in details.

4.1.2. Organ systems image classification model

We use the same model architecture for the plane model except that the last layer, which is the Softmax layer, has the ten organ systems classes.

4.1.3. Modality image classification models

As mentioned in the modality data description in the dataset section (Section 3), this category has different variations in question types and different main categories and subcategories. Thus, we create 11 models capable of answering every question type more accurately compared with what a general model can achieve.

Firstly, we explain the models we create, and, later, we explain how to combine them. “MM” is the abbreviation we use for modality models.

- MM1, the general model, for classifying image modality into eight main categories (XR, CT, MR, US, MA, GI, AG, and PT).
- MM2 model for distinguishing MR images from CT images.
- MM3 model for distinguishing contrast from noncontrast CT images.
- MM4 model for distinguishing contrast from noncontrast MR images.

- MM5 model for classifying CT contrast types (GI, IV, GI and IV).
- MM6 model for classifying MR weighting types (T1, T2 and Flair).
- MM7 model for classifying all CT subcategories.
- MM8 model for classifying all MR subcategories.
- MM9 model for classifying all GI subcategories.
- MM10 model for classifying all ultrasound subcategories.
- MM11 model for classifying PT images from AG images.

Note that no special models are created for PT and AG categories as the data for building them are insufficient. The available data for the AG category consists of 81 training images and 18 validation images. Moreover, 96% of the training images belong to only one class, and all the validation images are only for that class as well. The same applies for the PT category. The available data consists of 21 training images and a single validation image. About 85% of the training images belong to only one class, and the validation image for that class is zero. So, if the predicted main modality category is AG or PT, the subcategory answer will be the dominant class directly, which is AN-Angiogram for AG and NM-Nuclear Medicine for PT.

On the other hand, a special model (MM11) is created for classifying these two classes (AG and PT). The justification for creating this model comes from analyzing the errors of MM1 in classifying the eight main categories where most errors are in distinguishing these two classes. The training and validation data of the two classes are combined and re-split with 80% as the new training data and the remaining 20% as the new validation data to be used for this model exclusively.

Finally, it is worth mentioning that, due to the large variations in questions and answers of the modality category, it is not useful/practical to treat this part as a simple image classification task like what have been done for the plane and organ systems categories. Instead, as stated earlier, we approach this part in a hierarchical manner and build 11 models (MM1-MM11), where each model has its own data (i.e., the images and QA pairs) for training and validation purposes. It should be noted that relying on the data explicitly provided for each one of these models is not sufficient. Thus, we exploit the data provided for certain models to provide additional data for other models. Let us take Model MM6 as an example. This model is built for classifying MR weighting types (T1, T2, and Flair). The data for this model is gathered from questions such as the following.

- For question like “is this a t1 weighted, t2 weighted, or flair image” and “what is the mr weighting in this image”, the answer is one of three options: t1, t2 or flair. Thus, the image label is ready directly and the image is added to the training set of MM6.
- For a yes/no question like “is this a t1 weighted image”, if the answer is yes, then, we label the image with t1 add to the training set of MM6.
- For a question like “what imaging modality was used to take this image”, if the answer has the words “MR” along with any of the words “t1”, “t2” or “flair”, then, we use these words to create a new label for this image and add to the training set of MM6.

4.1.4. Abnormality image classification models

For type 1 questions, which ask about the abnormality in the image, we create a special image classification model with the same architecture used in our previously discussed models except that the Softmax layer has to cover all 1.6K abnormalities in the data. As for abnormality type 2 questions which ask if the image is normal or abnormal, a special image classification model is created with the same architecture of plane model but with Softmax layer that predicts normal/abnormal labels.

4.2. Image classification with extra information

In these models, the same image classification model’s architecture in Section 4.1 is used, but with adding an extra layer to concatenate the extra information before the last Sotmax layer. This idea is applied to three data categories as follows.

- Feed the organ model with the image plane label.
- Feed the plane model with the image organ label.
- Feed the abnormality type 1 questions model with the image plane label and the image organ label.

The organ model with extra plane information architecture is illustrated in Fig. 5. The actual image labels are fed to the models to validate the idea. However, in the prediction phase of test data, the extra information, which is the image plane label, is predicted using our best models.

4.3. Image classification with data augmentation

Data Augmentation (DA) is a technique to enlarge datasets from existing data without new data, and it is becoming a very common technique in the CV field [38,39]. There are different DA methods proposed in the literature; however, the method must be selected carefully, due to the sensitivity of the medical images in this work. For that, augmentation will be applied to organ systems data only, which is within our ability as a non-medical expert to validate the quality of the generated images. Following are the techniques we use.

- Flipping horizontally.
- Add random noise.
- Shift left.
- Shift right.
- Blurring.
- Gamma correction.
- Change contrast.

One issue related to DA is dealing with data imbalance. However, balancing organ data in our case is not favorable since the variation in the number of images per class is very large. For instance, the class with the smallest number of images, which is the breast class, has only 65 images, whereas the class with the largest number of images, which is the skull and contents class, has 1216 images. Thus, different experiments have been conducted to determine the minimum number of images in each class after DA. We denote each such experiment with the term DAX, where X denotes the minimum number of images in each class after DA. I.e., each class with less than X images will be augmented with enough images to make the number of images (real and augmented) reach 200 images. We experiment with $X = 200, 300, 400$. For the DA200 experiment, 302 new images are added to four classes. For the DA300 experiment, 904 new images are added to seven classes. Finally, for the DA400 experiment, 1513 new images are added to seven classes.

4.4. Image classification model with global average pooling technique

In this model, the features extraction method provided by the first-place team [26] of the VQA-Med 2019 challenge is used. The authors built a modified version of the pre-trained model VGG16 by removing all fully connected layers and using a GAP layer after each set of convolutional layers in order to reduce overfitting. Then, they concatenated the results to produce a feature vector with size 1984 units as shown in Fig. 6. In this work, we use their model and add to it an addition Softmax layer for the classification step. This model is applied to organ systems data and plane data to validate the idea and compare its results with the traditional VGG16 image classification model.

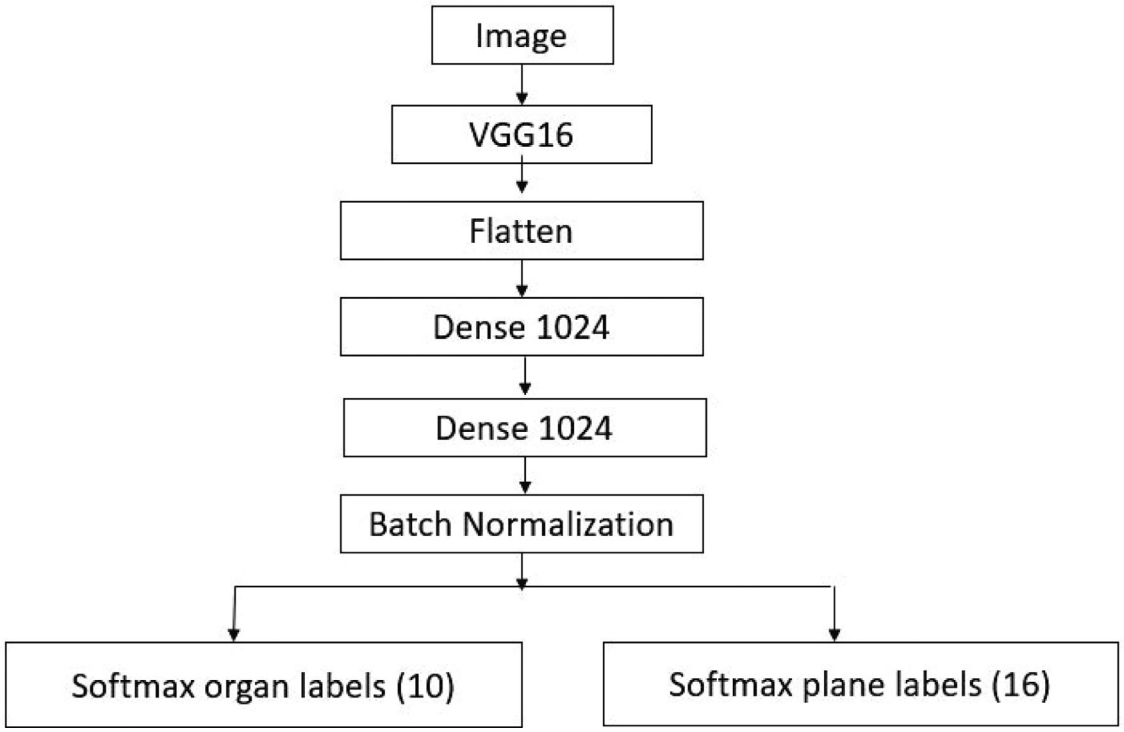

Fig. 5. Organ image classification model with extra plane information.

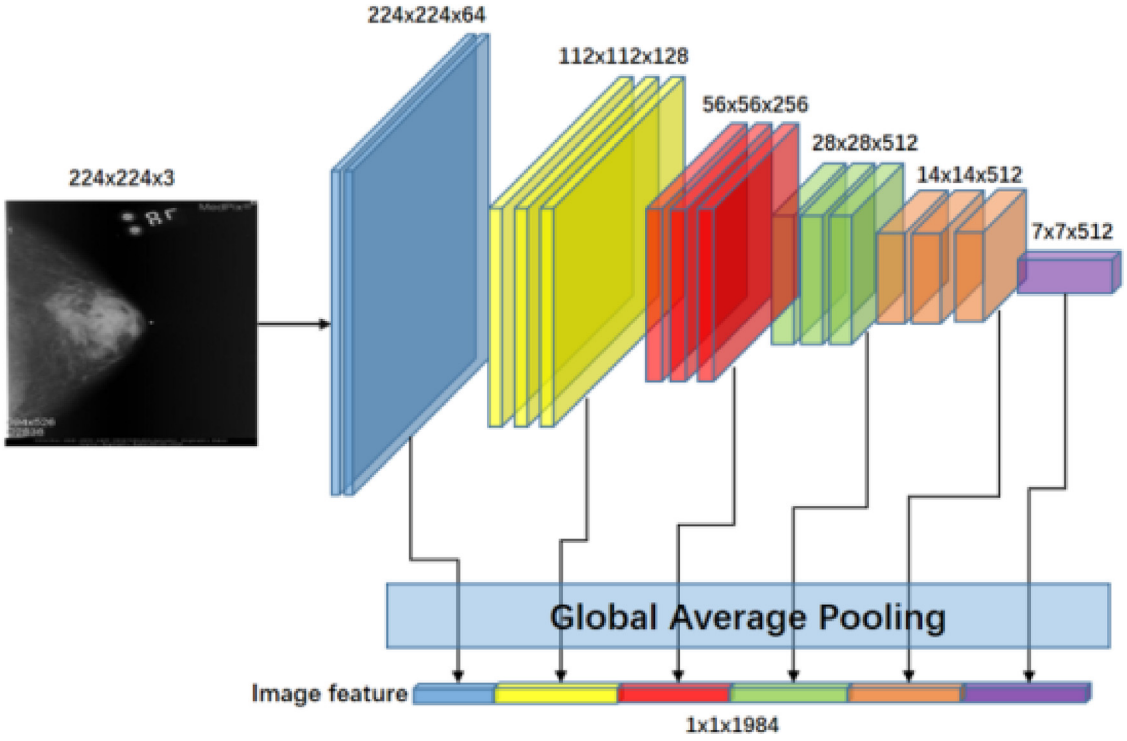

Fig. 6. Modified VGG16 model with GAP technique [26].

4.5. Ensemble models

In this section, we discuss how to combine some of the previously mentioned models in order to obtain better performance.

4.5.1. Ensemble model for organ systems data

So far, we have discussed two models for the organ systems data, which are the traditional VGG16 model and the modified VGG16 model with GAP technique. We have also discussed applying DA to these models. In this section, an ensemble model is built by following a hard voting scheme between these models as follows.

Table 5
Best validation results for each modality model along with the configuration that produces this result. For 'MM10', 'All' means that all optimizer/learning rates configurations under consideration produce the same result.

Subcategories Models	Validation Acc (%)	Best Parameters
MM1 (General model)	90	Adam (lr=0.0001)
MM2 (CT/MR model)	97.7	RMSprob (lr=0.0001)
MM3 (contrast/non-contrast CT)	79.3	RMSprob (lr=0.00001)
MM4 (contrast/non-contrast MR)	85.7	RMSprob (lr=0.0001)
MM5 (CT contrast types (GI/IV/GI and IV))	92.8	Nadam (lr=0.001)
MM6 (MR weighting types (T1/T2/Flair))	81	RMSprob (lr=0.00001)
MM7 (All CT subcategories)	73.1	RMSprob (lr=0.0001)
MM8 (All MR subcategories)	52.7	RMSprob (lr=0.00001)
MM9 (All GI subcategories)	76.2	Nadam (lr=0.0001)
MM10 (All ultrasound subcategories)	90	All
MM11 (AG/PT model)	100	RMSprob (lr=0.0001)
All models	**84.2**	-

- Model-1: Organ systems model with traditional VGG-16 and augmentation up to 200 images (Section 4.3)
- Model-2: Organ systems model with VGG16 and GAP technique without augmentation (Section 4.4)
- Model-3: Organ systems model with VGG16 and GAP technique with augmentation up to 200 images.

This configuration is found to produce the best results on the dataset under consideration.

4.5.2. Ensemble model for plane data

Similar to the organ systems data, we have discussed three models for the plane data as follows.

- Model-1: Plane model with traditional VGG-16.
- Model-2: Plane model with VGG16 and extra information from organ systems category.
- Model-3: Plane model with VGG16 and GAP technique.

Empirically, we find that combining them into a single ensemble model using a hard voting scheme yields better results.

Based on the final adopted models, Fig. 2 shows the steps taken to determine the required model in answer prediction. For simplicity of presentation, we omit some details such as how to handle modality question types 2–4 when different question forms are asking about specific things.

5. Results and analysis

The experiments to evaluate the proposed models and their results are discussed in this section. First, the evaluation results for each of the proposed models on the validation data are reported in order to find the best model for each of the four categories. After that, the results for the testing data are reported. The evaluation metrics are accuracy and cumulative 4-gram BLEU score. For all models, experiments are conducted using different optimizers [40] (which are RMSprop, Adagrad, Adadelta, Adam, Nadam, and Adamax) and different learning rates (which are 0.00001, 0.0001, 0.001, and 0.01): i.e., we conduct 24 experiments per model, each with 25 epochs. Then, the best results are taken.

5.1. Validation data results

In this section, we discuss the results for the proposed models on the validation data.

5.1.1. Basic image classification models results

Table 5 shows the best evaluation results on the validation data for each model belonging to the modality category along with the optimizer and the learning rate (lr) configuration that produces this result. The accuracy of MM10 is misleading since it predicts the dominant class all the time. We remind the reader that the overall modality validation accuracy (84.2%) is not the average accuracy for all models; it is the accuracy of the predicted answers for the modality questions in the validation set.

Table 6 shows the best evaluation results on the validation data for each model belonging to the abnormality category along with the optimizer and the learning rate (lr) configuration that produces this result. While Table 7 shows the best evaluation results on the validation data for each data category.

5.1.2. Image classification models with extra information results

In this experiment, extra information is fed to the image classification models of organ systems, planes, and abnormalities (type 2 questions). Table 8 shows the results of each model.

The accuracy achieved by the original abnormality classification model is 14.4%. With extra information, the accuracy is improved to 14.88%. However, it is worth mentioning that the extra information is the actual values of planes and organs, while, in test prediction, this extra information will be predicted. Thus, it is expected that accuracy decreases due to errors in predicting planes and organ systems, which is why we believe that the performance of the original abnormality classification model is more robust.

5.1.3. Image classification models with data augmentation results

As explained in the methodology section, DA is used for the organ systems data, with three experiments: DA200, DA300 and DA400. Table 9 shows the results of each experiment. These results coincide with recent research [38,39] about how DA should be used carefully that overusing it might negatively affect the outcomes.

5.1.4. Image classification model with global average pooling technique results

For this model, different experiments are made for organ systems; with the original data and with the augmented data. Due to the use of GAP, these experiments take 100 epochs to saturate unlike all our previously discussed experiments, which take 25 epochs. Table 10 shows the best results of these experiments. We also apply GAP for the plane data and the resulting accuracy is 78.8%.

5.1.5. Ensemble models

Tables 11 and 12 show the results of the three models discussed in Section 4.5.1 and the three models discussed in Section 4.5.2, respectively. The tables also show the results of ensembling each set of the models, which yields the best results.

Table 6
Best validation results for abnormality models along with the configuration that produces this result.

Model	Acc (%)	Best Parameters
Abnormality type 1 questions	14.7	RMSprob (lr=0.0001)
Abnormality type 2 questions (normal/abnormal)	77.7	Adam (lr=0.001)
All	**17.59**	-

Table 7
Best validation results for all data categories models along with the configuration that produces this result.

Model	Acc (%)	Best Parameters
Plane data questions	76.2	RMSprob (lr=0.00001)
Organ systems data questions	75.6	RMSprob (lr=0.00001)
Modality data questions	84.2	-
Abnormality data questions	17.59	-
All	**63.5**	-

Table 8
Image classification models with extra information validation results.

Model	Acc (%)	Best Parameters
Organ Model with Extra Plane Information	75	RMSprob (lr=0.0001)
Plane Model with Extra Organ Information	79.6	Nadam (lr=0.0001)
Abnormality (type 1 questions) Model with Extra Organ and Plane Information	14.88	RMSprop (lr=0.0001)

Table 9
Image classification models with DA validation results.

Experiment	Organ Acc (%)	Best Parameters
DA200	**77.2**	Adamax (lr=0.0001)
DA300	76.2	RMSprob (lr=0.00001)
DA400	74.2	RMSprob (lr=0.00001)

Table 10
Image classification model with GAP technique results.

Experiment	Organ Acc (%)	Best Parameters
GAP without DA	**77.8**	Adamax (lr=0.0001)
GAP with DA200	77.2	Adam (lr=0.0001)

Table 11
Ensemble model for organ systems results.

Model	Organ Acc (%)
Traditional VGG16 with DA200	77.2
Modified VGG16 with GAP without DA	77.8
Modified VGG16 with GAP and DA200	77.2
Ensemble model	**81.4**

Table 12
Ensemble model for plane data results.

Model	Plane Acc (%)
Traditional VGG16	76.2
Traditional VGG16 with extra information from organ systems data	79.6
Modified VGG16 with GAP	78.8
Ensemble model	**84.2**

5.2. Best validation data results and analysis

In this section, we discuss the best performing models on the validation data. These models are adopted for further experimentation on the test data. Based on the validation data results of all proposed models discussed in Section 5.1, Table 13 shows the adopted model for each data category; models that achieved the best results. To understand the results and explore the weaknesses of each of these models, we show the confusion matrices for each model.

Table 13
Best performing model for each category in the validation data.

Category	Best Acc (%)	Model
Organ Systems	81.4	The organ systems ensemble model
Plane	84.2	The planes ensemble model
Modality	84.2	Basic Image Classification Models
Abnormality	17.59	Basic Image Classification Models
Overall	**66.85**	-

For the organ systems part of the validation data, Fig. 7 shows the best model's confusion matrix. As the figure shows, the most frequent errors are in differentiating 'Face, Sinuses, and Neck' images from 'Skull and Contents' images and 'Gastrointestinal' images from 'Genitourinary' images. Consulting a physician and a radiologist leads to the realization that systems should not be penalized for such errors since images of one organ might show other organs. Specifically, the same image can show 'Face, Sinuses, and Neck' as well as 'Skull and Contents'. The doctors choose to focus on one based on case at hand. This issue only exists in the training and validation sets since they were automatically generated based on the captions associated with each image. Understandably, if the caption only mentions 'Face, Sinuses, and Neck' or any keyword associated with this organ, then, the answer would automatically be 'Face, Sinuses, and Neck' regardless of whether other answers such as 'Skull and Contents' are correct or not.

Another thing worth mentioning is that the number of false predictions of the 'Vascular and Lymphatic' organ (17 images), is larger than the number of correct predictions (4 images). One very likely reason for this is the errors in labeling training data for this organ. After consulting medical experts, it was noticed that a lot of images do not belong to this organ. Unfortunately, we were unable to correct all of them as mentioned in Section 3.5.

As for the planes part of the validation data, Fig. 8 shows the best model's confusion matrix. The first thing to note is that the validation data only covers 13 out of 16 possible planes; i.e., there are three planes without any image in the validation data. Another issue related to the imbalance of this dataset is that many errors are due to predicting the majority class, which is the 'Axial' plane. Finally, there are 15 errors in predicting 'Frontal' or 'Lateral' planes as 'AP' plane. According to the radiologist and the physician we consult, these are the same, especially the 'Frontal' plane and the 'AP' plane (which is an abbreviation for 'anterior-posterior') as they refer to imaging from the front side.

Regarding the modality part of the validation data, it is not easy to present and analyze the results since the number of possible answers is large and the results are produced from 11 different models organized in a hierarchical manner, which means that the errors are cumulative. For example, an error in predicting CT contrast image as MR T1 weighted is actually due to the error in initially predicting CT as MR from the general model.

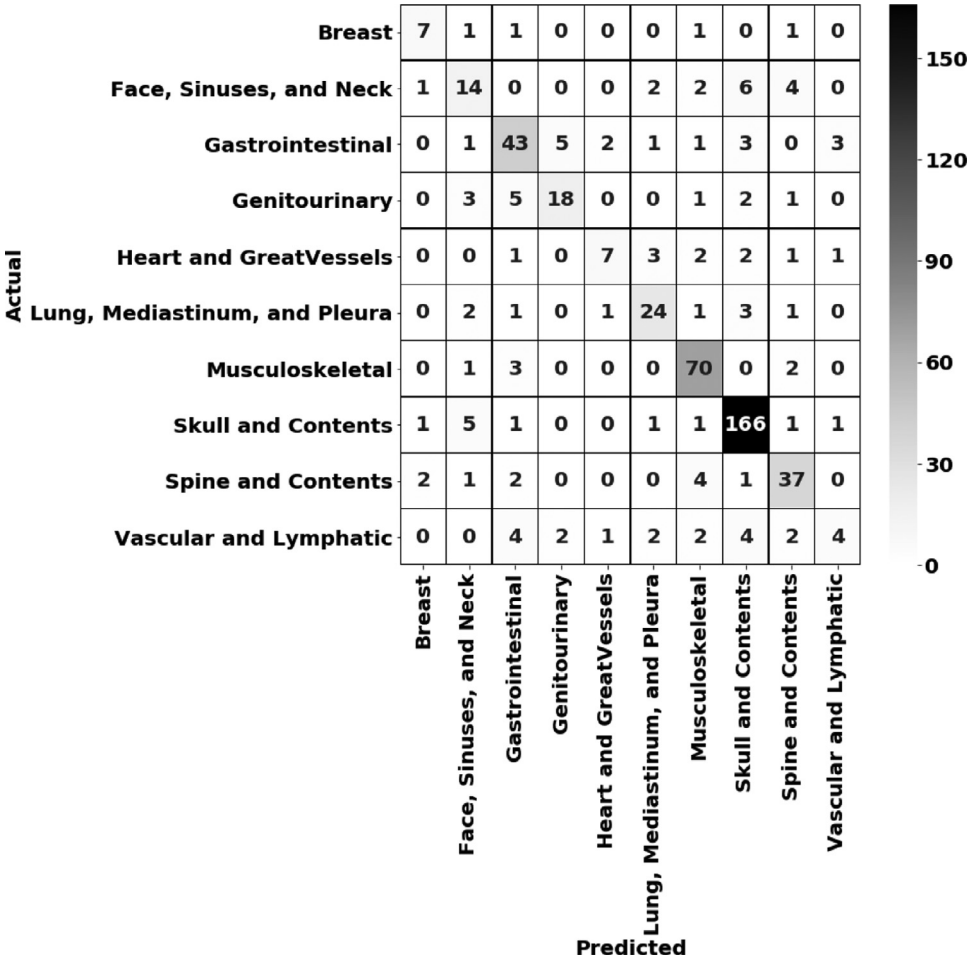

Fig. 7. Confusion matrix for organ systems validation data.

For this part of the validation data, there are 79 false predictions out of 500 possible ones. The most frequent errors we observe are as follows.

- There are 26 false predictions of the main modality made by the general model (MM1). It is worth mentioning that MM1 model accuracy, as reported in Table 5, is 90%; however, for the modality part of the validation data, it is only 88.65%. The difference is due to the fact that the data used in building the model is collected from all question types, not only from the questions asking about the main modality of the image.
- There are 21 false predictions for the T1/T2/Flair model (MM6), i.e., 77.9% accuracy. Again, these results are less than MM6's accuracy reported in Table 5 (which is 81%) because the data used to build the model is collected from different question types.
- There are 17 false predictions of the contrast/non-contrast CT.

The Abnormality model has the lowest accuracy among all models; however, this is expected for two reasons. Firstly, there is a very large number of labels (about 1.6K) for such a small dataset (about 3.7K). In fact, each label has an average of about two images only. Secondly, many labels have images in the training data and do not have any image in the validation data, and vice versa.

5.3. Test data results

The models with the best results on validation data are used for test data prediction. However, we remind the readers of the differences between the training/validation data and test data discussed in Section 3.5 such as the existence of errors and multiple correct answers for the same image. Another issue worth mentioning is about models that use extra information. In the validation data, we use the ground truth information, however, for the test data, we can only use the predicted extra information.

Table 14
Final test data results.

Category	Acc (%)
Organ Systems Data	75.2
Plane Data	77.6
Modality Data	72
Abnormality Data	18.4
Overall Data	**60.8**

Table 14 shows the test evaluation results for each data category and the overall evaluation. As can be seen in this table, the differences between our best results and the best results achieved by the top team of VQA-Med 2019 is rather small: 1.6 in accuracy and 1 BLEU score. Based on the detailed results for each team for each data category [4], our results for the abnormality part match those of the best team while outperforming their results in the organ systems category and the plane category. We are only outperformed in the modality category. This difference is expected due to the use of an advance technique in their model for question features extraction, which is the BERT model.

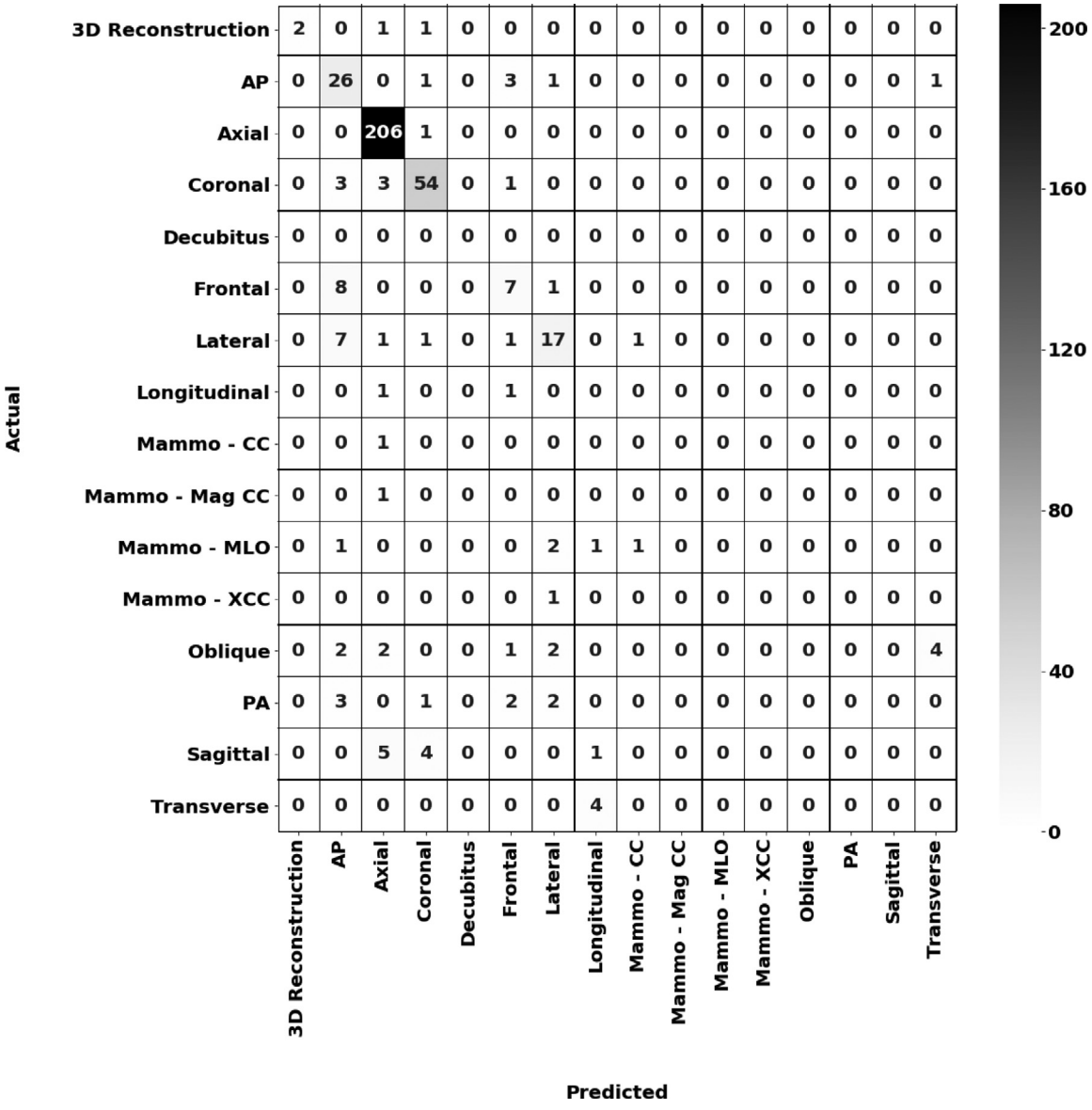

Fig. 8. Confusion matrix for planes validation data.

6. Conclusions and future work

In this paper, a medical VQA system is built based on the dataset provided by ImageCLEF VQA-Med 2019 challenge. The system consists of several sub-model. A special sub-model is built to classify the question category with 100% accuracy, and hence determine the appropriate model to answer it. The sub-models used for answering questions are based on the pre-trained VGG model. The best model overall accuracy is 60.8% with 63.4 BLEU score. Accuracy of plane, organ, and modality models are very good (72% to 77.6%), but the abnormality model accuracy is rather low (18.4%) due to the difficulty of this part of the task with such a small dataset. To achieve these results, different cutting-edge deep learning techniques are explored, such as pre-trained models, DA, GAP and ensemble models. In the future, we plan on addressing the shortcomings of the existing dataset by correcting the errors/issues we found in it and collecting more samples, especially for the abnormality part.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We would like to thank the Deanship of Research at the Jordan University of Science and Technology for supporting this work (Grant #20190180). We would also like to thank Dr. Asma' Al-Mnayyis, a Radiologist from the College of Medicine at Yarmouk University, Jordan, for her help with the medical concepts related to the dataset.

Appendix A. Data Categories

More details about the dataset categories are presented in the this section. The following list shows the 16 plane classes under consideration and Fig. A.9 shows samples of different planes.

- Axial.

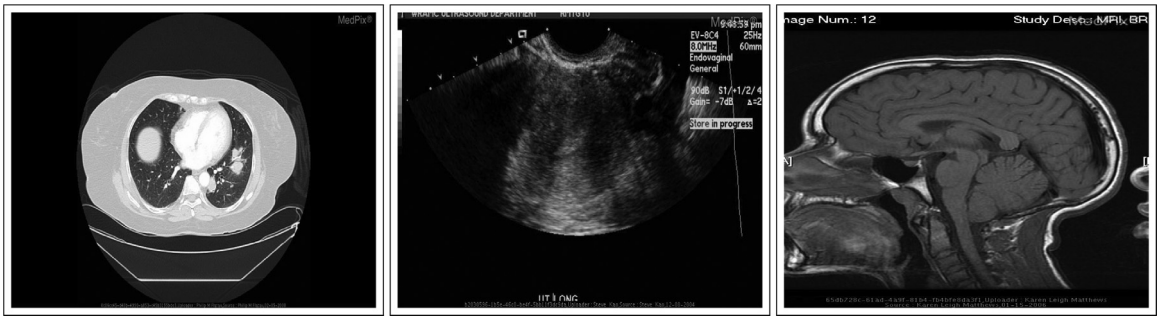

Fig. A.9. Sample images of different planes: Axial plane, Longitudinal plane, and Sagittal plane (from left to right) .

Fig. A.10. Sample images of different organs: Skull and contents organ, Spine and contents organ, and Musculoskeletal organ (from left to right).

Table A.15
Organ systems categories under consideration.

Skull and Contents	Face, sinuses, and neck
Spine and contents	Musculoskeletal
Heart and great vessels	Lung, mediastinum, pleura
Gastrointestinal	Genitourinary
Breast	Vascular and lymphatic

- Sagittal.
- Coronal.
- Anteroposterior (AP).
- Posteroanterior (PA).
- Lateral.
- Frontal.
- Transverse.
- Oblique.
- Longitudinal.
- Decubitus.
- 3D Reconstruction.
- Mammography with mediolateral oblique view (Mammo-MLO).
- Mammography with craniocaudal view (Mammo-CC).
- Mammo-CC with magnification (Mammo-Mag CC).
- Mammography with extended craniocaudal view (Mammo-XCC).

As for the ten organ systems under consideration, Table A.15 shows them while Fig. A.10 shows samples of different organs. Finally, the main eight modality categories, each with its subcategories are listed in Table A.16.

Appendix B. Other Approaches

In this section, we discuss the different deep learning approaches we use to solve the task at hand, but did not achieve the best results.

B1. Abnormality image classification based on data organ systems

As mentioned in Section 3, the number of unique answers in the abnormality data (type 1 questions) to be used in the classification task is around 1.6K, which is a large number for a classification model compared with the data size (about 3.7K images).

Table A.16
Modality categories under consideration: Main-modalities and Sub-modalities.

Main-modalities	Sub-modalities			
XR	XR - Plain Film			
MA	Mammograph			
US	US - Ultrasound		US-D - Doppler Ultrasound	
AG	AN - Angiogram		Venogram	
PT	NM - Nuclear Medicine		PET - Positron Emission	
GI	BAS - Barium Swallow	UGI - Upper GI	BE - Barium Enema	SBFT - Small Bowel
CT	CT - noncontrast	CT w/contrast (IV)	CT - GI & IV Contrast	CTA - CT Angiography
	CT - GI Contrast	CT - Myelogram	Tomography	
MR	MR-T1W w/Gadolinium	MR-T1W - noncontrast	MR-T2 weighted	MR-FLAIR
	MR-T1W w/Gd (fat suppressed)	MR-STIR	MR-DWI Diffusion Weighted	MR-Other Pulse Seq.
	MR-T2 FLAIR w/Contrast	MR-FIESTA	MR-ADC Map (App Diff Coeff)	MR-T1W SPGR
	MRA-MR Angiography/Venography	MR-FLAIR w/Gd	MR-PDW Proton Density	MR-T2* gradient GRE
	MR-T2*gradient,GRE,MPGR,SWAN,SWI			

Table B.17
Abnormality data distribution among organ systems and validation results.

Organ system	Abnormality Training Data		Abnormality Validation Data		Validation Acc (%)	Best Parameters
	# of questions	# of labels	# of questions	# of labels		
Skull and contents	1190	451	172	125	15.11	RMSprop(lr=0.0001)
Face, sinuses, and neck	190	108	29	29	10.34	Nadam (lr=0.001)
Spine and contents	226	123	46	42	17.39	RMSprop (lr=0.0001)
Musculoskeletal	418	263	69	62	4.35	Adamax (lr=0.0001)
Vascular and lymphatic	122	79	21	21	14.29	Nadam (lr=0.001)
Gastrointestinal	339	182	56	52	28.57	RMSprop (lr=0.0001)
Genitourinary	207	122	28	27	10.71	RMSprop (lr=0.0001)
Lung, mediastinum, pleura	234	147	31	30	12.90	RMSprop (lr=0.0001)
Heart and great vessels	104	65	15	15	6.67	Nadam (lr=0.0001)
Breast	52	34	10	10	20	Adam (lr=0.01)
Total	**3082**	-	**477**	-	**14.47**	-

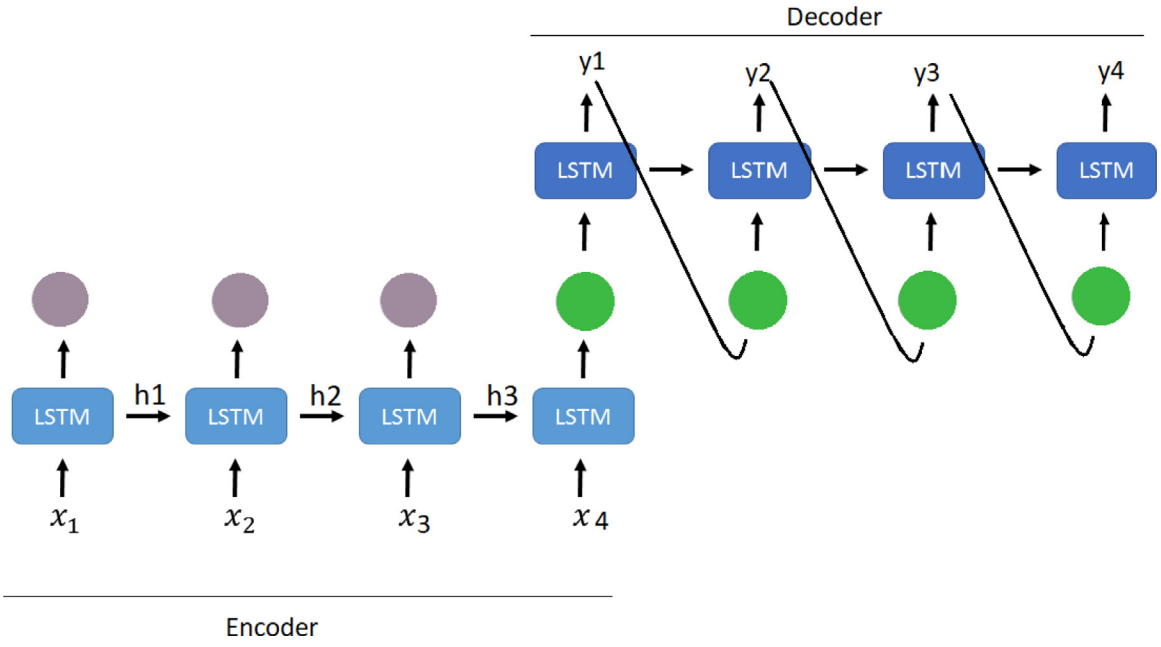

Fig. B.11. Encoder-Decoder model for Seq2Seq data.

Moreover, the distribution of these abnormalities across the organ systems is not balanced. To study this further, the abnormality data is divided into ten parts, where each part contains images and abnormalities for a specific organ system. Table B.17 shows the number of abnormality images that belong to each organ system.

A special model is created for each data part using the same image classification model used previously, but with a suitable Softmax layer on its labels. Then, for prediction, each image's organ system is predicted to determine the suitable abnormality model from the ten models aiming to predict the abnormality.

Table B.17 shows the best accuracy of each model, the optimizer, and the learning rate used to produce that accuracy. For many of these models, there is more than one combination of optimizer and learning rate that give the same best results; thus, one of them is mentioned randomly.

Here, the overall accuracy for the abnormality validation data (type 1 questions) is 14.47%, which is very close to the 14.4% accuracy achieved by our original abnormality classification model. However, it is worth mentioning that the organ system information here is the actual values, while, in test prediction, this information will be predicted. Thus, it is expected that accuracy drops due to the expected errors in predicting the organ systems.

B2. Abnormality as a sequence to sequence task

When the input and output data are both sequences of correlated elements, the problem is known as a sequence to sequence one. A large number of studies are dedicated for such type of problems due to its diverse applications, such as machine translation, text summarization, question answering, and image captioning. For the task at our hands, the abnormality questions of type 1 can be considered as a sequence since the answer can be a long free text, unlike the other categories. Thus, we treat this problem as a Seq2Seq problem and propose several models accordingly.

One of the most common architectures proposed for Seq2Seq problems is the encoder-decoder model [41,42], where the encoder and decoder are RNN networks such as LSTM [42] or Gated Recurrent Unit (GRU) [43]. In the simple version of the encoder-decoder, the encoder takes the input sequence and creates a context vector about it, which is sent to the decoder to be used besides the original input to produce the output sequence as shown in Fig. B.11. As the number of hidden units in the RNN increases, more information about the input is stored. However, the encoder cannot deal with items (words here) directly; a pre-trained word embedding technique must be used firstly, such as Word2Vec [10] or GloVe [11]. These techniques are used to convert words into vectors by capturing the relations between them.

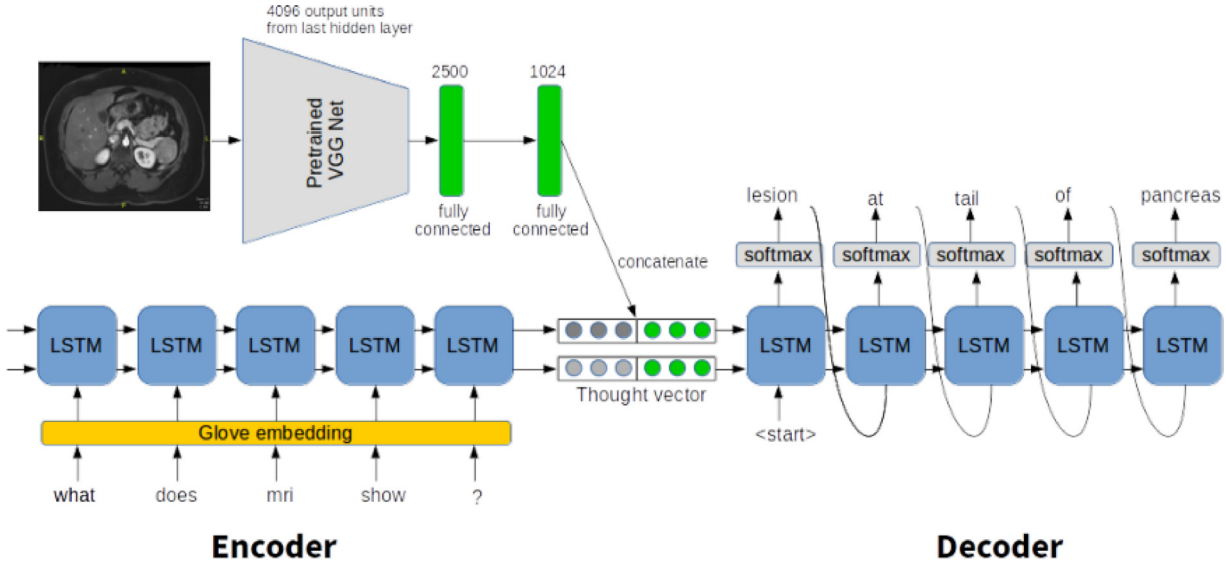

Fig. B.12. First Image-Question-Answer model [20].

	Xi		Yi	
i	Image feature vector	Partial Caption	Target word	
1	Image_1	startseq	the	data points corresponding to image 1 and its caption
2	Image_1	startseq the	black	
3	Image_1	startseq the black	cat	
4	Image_1	startseq the black cat	sat	
5	Image_1	startseq the black cat sat	on	
6	Image_1	startseq the black cat sat on	grass	
7	Image_1	startseq the black cat sat on grass	endseq	
8	Image_2	startseq	the	data points corresponding to image 2 and its caption
9	Image_2	startseq the	white	
10	Image_2	startseq the white	cat	
11	Image_2	startseq the white cat	is	
12	Image_2	startseq the white cat is	walking	
13	Image_2	startseq the white cat is walking	on	
14	Image_2	startseq the white cat is walking on	road	
15	Image_2	startseq the white cat is walking on road	endseq	

Fig. B.13. Image partial captioning approach example [44].

B2.1. Image-Question-Answer model 1

This model uses the simple encoder-decoder architecture described previously. Specifically, it uses GloVe embedding and LSTM cells, but the input of the LSTM encoder part here is the concatenation between the questions embedding and the image features as illustrated in Fig. B.12.

The encoder side consists of two parts: the image features and the question features. The image features are extracted using the VGG16 model from the last convolutional layer and the output is passed to two fully connected layers of sizes 2500 and 1024, respectively. Now for the question, firstly, the GloVe embedding is used to encode the question words with the pre-trained GloVe embedding and the output is passed to LSTM network. There are three outputs of the LSTM network: the main output and two hidden states. The main output of the encoder LSTM is passed as input to the decoder LSTM, while the two hidden states are concatenated with the extracted image features. The output, which we call the thought-vector, is passed to the decoder LSTM as its initial state. The decoder LSTM predicts the next possible word in the answer at each time step.

B2.2. Image-Question-Answer model 2

In this model, the question and the image features are concatenated to be fed later to the encoder LSTM. The encoder states output are the initial hidden states of the decoder LSTM. The difference between this work and the previous model is that the image features and the question features are concatenated in an earlier step, and the result of the concatenation is passed to the encoder LSTM, unlike the previous model, where the question features are passed to the encoder LSTM. Moreover, the resulting hidden states of the encoder LSTM are passed to the decoder as its initial hidden states.

B2.3. Image-Question-Answer model 2 with attention

Attention gives the model the ability to focus on the important parts of the input at a specific moment instead of looking at the whole input. Mathematically, it is a vector of importance weights, which is obtained by calculating the attention value between each input item and output item. In this model, an attention layer is added between the encoder and the decoder. The attention weights are calculated for each hidden state at each time step in a way that gives important hidden states high weights.

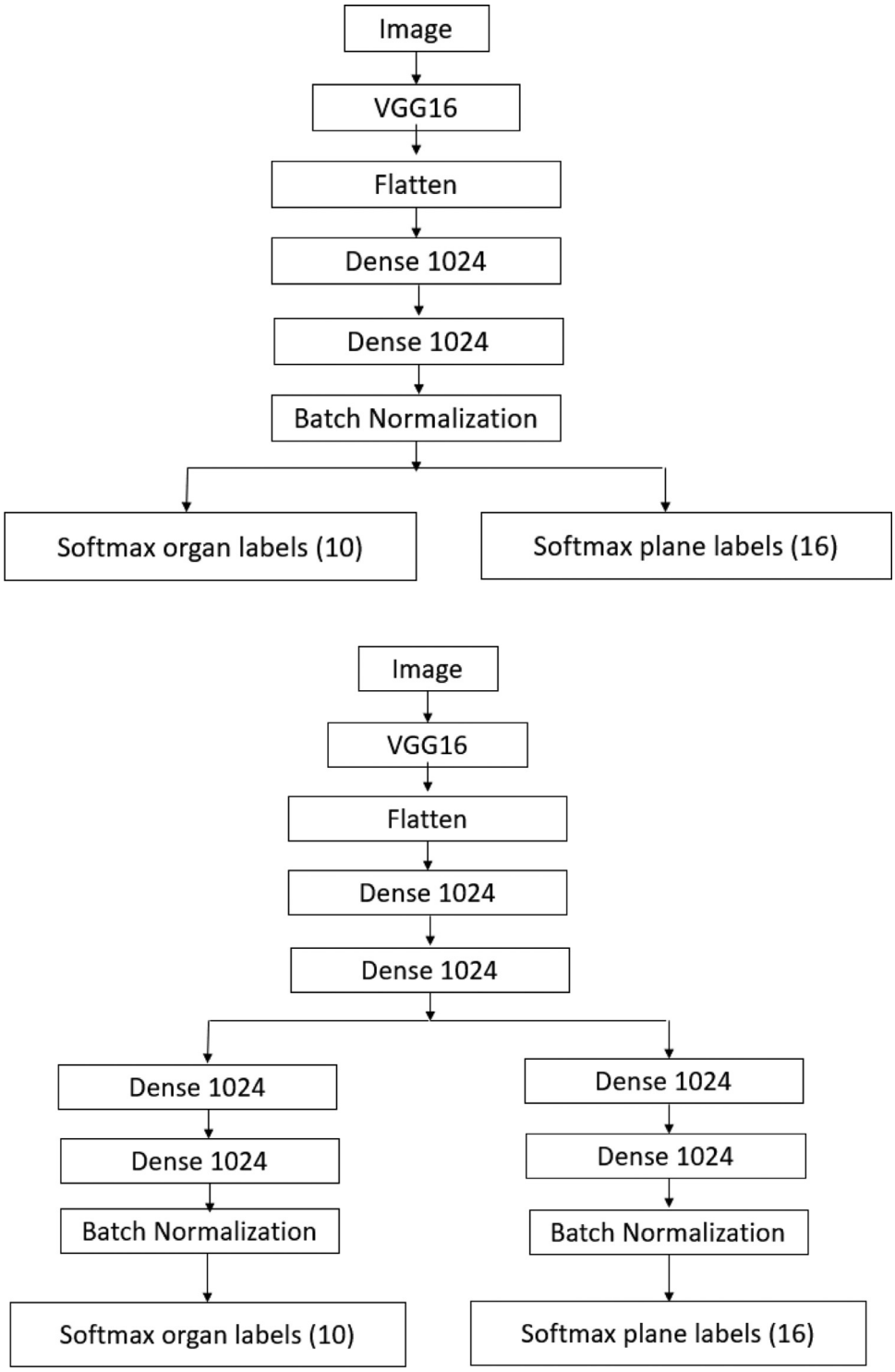

Fig. B.14. MTL architectures: Sharing all layers (top) and Sharing some layers (bottom).

B2.4. Image-Answer model

In the task at hand, questions are repetitive and, hence, they are not expected to play an important role in the answer prediction. Thus, we create an encoder-decoder model with image-answer only. this model is similar to the first Seq2Seq model, but, instead of question embedding, we use the image vector to be the input of the LSTM encoder.

B2.5. Image-Answer model with attention

This model is similar to the previous one, but it has an attention mechanism. The only modification on the previous encoder-decoder model is by adding a new layer, which is the attention layer between the encoder output and decoder output. This new layer output will be combined later and passed to the final output layer.

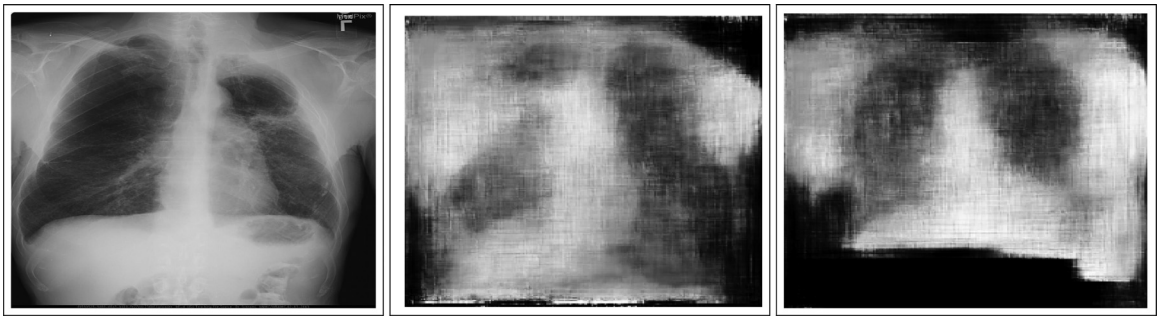

Fig. B.15. GAN generated images samples for 'Lung, Mediastinum, and Pleura' organ system with 'Frontal' plane: real image and generated images (from left to right).

Fig. B.16. GAN generated images samples for 'Lung, Mediastinum, and Pleura' organ system with 'Axial' plane: real image and generated images (from left to right).

Table B.18
Seq2Seq models validation results.

Model	BLEU Score
Image-Question-Answer Model 1	0.11
Image-Question-Answer Model 2	8
Image-Question-Answer Model 2 with Attention	9
Image-Answer	7.6
Image-Answer with Attention	0.8
Partial Image Captioning	1

B2.6. Partial image captioning model

In this model, the data is formulated in a new shape by attaching startSeq and endSeq tags to each answer, and then dividing the answer into words. After that, the image with the startSeq is used to predict the first word of the answer, and the image with the first word of the answer is used to predict the second word of the answer, and so on, as shown in Fig. B.13. In each time step, the concatenation of the image vector and the partial caption is passed to the encoder LSTM then to a dense layer followed by a Softmax layer to predict the next word.

Since this model deals with the data as Seq2Seq, it predicts the answer in a word-by-word manner; that is, it does not predict the whole answer directly as in classification tasks. Table B.18 shows the BLEU scores for each of the Seq2Seq models on the validation data. As expected from the very low BLEU scores, none of these models managed to predict any complete answer, which makes their accuracy zero%.

B3. Multi-Task learning image classification

Traditional machine and deep learning algorithms are dedicated to solving one problem at a time such as classification, segmentation, translation, etc. Inspired by human behavior, a new deep learning approach is proposed, which is multi-task learning (MTL). MTL is developing models that are able to learn multiple tasks simultaneously by sharing information between them. These tasks may belong to different domains. One of the benefits of MTL is the exploit of training time for training multiple problems instead of training one or two models. The same benefit is applicable for prediction time. There are many architectures for MTL varying in the size of the shared data, and the layers between the different tasks. For this task, image planes and organ systems data are chosen. Two MTL approaches are used, which are sharing all layers and sharing some layers as in Fig. B.14.

The best accuracy of planes classification for the validation images is 77.4%, which use RMSprop optimizer with a learning rate (lr) of 0.00001. For organ systems classification, the best accuracy is 73%, which use RMSprop optimizer with learning rate equals 0.0001.

B4. Data Augmentation using Generative Adversarial Networks

Generative adversarial networks (GANs) [45] is a deep learning technique that consists of two networks, the generator (G) and the discriminator (D). These two networks compete to outperform each other. The basic idea of GANs is that G generates fake data and D tries to classify this data and determine if its real or fake. As the quality of data generated by G increases, it becomes more difficult for D to differentiate real from fake data. The best point to reach is when G and D have very low loss, the accuracy of D is 50% (which means that it cannot distinguish between real and generated data), and at the same time, the quality of generated data is very good. The data type can be images, texts, audio signals, etc. Moreover, there are a lot of GANs variations, such as Deep Convolutional GAN (DCGAN) [33].

In the medical domain, using GANs started to gain popularity recently especially in generating new images to enlarge datasets such as in [46,47]. This work uses GANs in the same context, which is to generate new data as an augmentation method. To validate the data, basic GAN version and DCGAN are used to generate images for all organ systems, and for some organ systems each separately. The same DCGAN architecture in [47] is used. For different DCGAN experiments, the quality of the generated images is poor showing no details. Figs. B.15 and B.16 show samples from the real

images and the generated images for Lung, mediastinum, pleura organ system with Axial plane and Frontal plane.

In general, GANs are known to be difficult to be trained due to many reasons, such as the effect of each input parameter and the balancing between the two networks (D and G) [48,49]. However, in our case, the reason for the poor generated images is the small dataset size for each type of images. For example, the dataset size used in [46] and [47] is more than 6K images, all of which are for Lung, mediastinum, pleura organ system with Frontal plane. Another approach to benefit from their models is by using their model weights as a pre-trained to our model). Unfortunately, the weights are not available at the time of writing this paper. Moreover, we are unable reproduce the weights due to resources limitation. So, this will be an interesting future direction of this work.

References

[1] B. Ionescu, et al., ImageCLEF 2019: multimedia retrieval in medicine, lifelogging, security and nature, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, in: Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019), LNCS Lecture Notes in Computer Science, Springer, Lugano, Switzerland, 2019.

[2] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, 2002, pp. 311–318.

[3] S.A. Hasan, Y. Ling, O. Farri, J. Liu, M. Lungren, H. Müller, Overview of the ImageCLEF 2018 medical domain visual question answering task, in: CLEF2018 Working Notes, in: CEUR Workshop Proceedings, Avignon, France, 2018.

[4] A. Ben Abacha, S.A. Hasan, V.V. Datla, J. Liu, D. Demner-Fushman, H. Müller, Vqa-med: Overview of the medical visual question answering task at imageclef 2019, in: CLEF 2019 Working Notes, in: CEUR Workshop Proceedings, Lugano, Switzerland, 2019.

[5] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, A. van den Hengel, Visual question answering: a survey of methods and datasets, Comput. Vision Image Understanding 163 (2017) 21–40.

[6] A.K. Gupta, Survey of visual question answering: datasets and techniques, arXiv preprint arXiv:1705.03865 (2017).

[7] K. Kafle, C. Kanan, Answer-type prediction for visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4976–4984.

[8] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[9] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[10] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.

[11] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.

[12] M. Malinowski, M. Rohrbach, M. Fritz, Ask your neurons: A neural-based approach to answering questions about images, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1–9.

[13] M. Ren, R. Kiros, R. Zemel, Image question answering: a visual semantic embedding model and a new dataset, Proc. Advances in Neural Inf. Process. Syst 1 (2) (2015) 5.

[14] K.J. Shih, S. Singh, D. Hoiem, Where to look: Focus regions for visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4613–4621.

[15] Y. Zhu, O. Groth, M. Bernstein, L. Fei-Fei, Visual7w: Grounded question answering in images, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4995–5004.

[16] Z. Yang, X. He, J. Gao, L. Deng, A. Smola, Stacked attention networks for image question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 21–29.

[17] J. Lu, J. Yang, D. Batra, D. Parikh, Hierarchical question-image co-attention for visual question answering, in: Advances In Neural Information Processing Systems, 2016, pp. 289–297.

[18] Y. Peng, F. Liu, M.P. Rosen, Umass at imageclef medical visual question answering (med-vqa) 2018 task, in: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10–14, 2018., 2018.

[19] Y. Zhou, X. Kang, F. Ren, Employing inception-resnet-v2 and bi-lstm for medical domain visual question answering, in: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10–14, 2018., 2018.

[20] B. Talafha, M. Al-Ayyoub, Just at vqa-med: A vgg-seq2seq model, in: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10–14, 2018., 2018.

[21] A.B. Abacha, S. Gayen, J.J. Lau, S. Rajaraman, D. Demner-Fushman, Nlm at imageclef 2018 visual question answering in the medical domain, in: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10–14, 2018., 2018.

[22] I. Allaouzi, B. Benamrou, M. Benamrou, M.B. Ahmed, Deep neural networks and decision tree classifier for visual question answering in the medical domain, in: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10–14, 2018., 2018.

[23] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing systems, 2017, pp. 5998–6008.

[25] M. Lin, Q. Chen, S. Yan, Network in network, arXiv preprint arXiv:1312.4400 (2013).

[26] X. Yan, L. Li, C. Xie, J. Xiao1, L. Gu, Zhejiang university at imageclef 2019 visual question answering in the medical domain, in: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9–12, 2019., 2019.

[27] M.H. Vu, R. Sznitman, T. Nyholm, T. Lofstedt, Ensemble of streamlined bilinear visual question answering models for the imageclef 2019 challenge in the medical domain, in: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9–12, 2019., 2019.

[28] Y. Zhou, X. Kang, F. Ren, Tua1 at imageclef 2019 vqa-med: A classification and generation model based on transfer learning, in: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9–12, 2019., 2019.

[29] F. Ren, Y. Zhou, Cgmvqa: a new classification and generative model for medical visual question answering, IEEE Access 8 (2020) 50626–50636.

[30] L. Shi, F. Liu, M.P. Rosen, Deep multimodal learning for medical visual question answering, in: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9–12, 2019., 2019.

[31] H. Kim, Y. Sun, J. Hockenmaier, J. Han, Etm: Entity topic models for mining documents associated with entities, in: 2012 IEEE 12th International Conference on Data Mining, IEEE, 2012, pp. 349–358.

[32] A. Al-Sadi, B. Talafha, M. Al-Ayyoub, Y. Jararweh, F. Costen, Just at imageclef 2019 visual question answering in the medical domain, in: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9–12, 2019., 2019.

[33] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv preprint arXiv:1511.06434 (2015).

[34] T. Kornuta, D. Rajan, C. Shivade, A. Asseman, A.S. Ozcan, Leveraging medical visual question answering with supporting facts, in: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9–12, 2019., 2019.

[35] A. Lubna, S. Kalady, A. Lijiya, Mobvqa: A modality based medical image visual question answering system, in: TENCON 2019-2019 IEEE Region 10 Conference (TENCON), IEEE, 2019, pp. 727–732.

[36] W. Hussien, M. Al-Ayyoub, Y. Tashtoush, M. Al-Kabi, On the use of emojis to train emotion classifiers, arXiv preprint arXiv:1902.08906 (2019).

[37] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, Int J Comput Vis 115 (3) (2015) 211–252.

[38] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, J Big Data 6 (1) (2019) 60.

[39] M. Ebrahim, M. Alsmirat, M. Al-Ayyoub, Performance study of augmentation techniques for HEP2 CNN classification, in: 2018 9th International Conference on Information and Communication Systems (ICICS), IEEE, 2018, pp. 163–168.

[40] F. Chollet, et al., Keras, 2015, (https://keras.io).

[41] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using rnn encoder-decoder for statistical machine translation, arXiv preprint arXiv:1406.1078 (2014).

[42] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, in: Advances in Neural Information Processing Systems, 2014, pp. 3104–3112.

[43] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput 9 (8) (1997) 1735–1780.

[44] H. Lamba, Image captioning with keras, 2018,. [Accessed 1 Jul 2019], https://towardsdatascience.com/image-captioning-with-keras-teaching-computers-to-describe-pictures-c88a46a311b8.

[45] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.

[46] A. Madani, M. Moradi, A. Karargyris, T. Syeda-Mahmood, Semi-supervised learning with generative adversarial networks for chest x-ray classification with ability of data domain adaptation, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE, 2018, pp. 1038–1042.

[47] H. Salehinejad, S. Valaee, T. Dowdell, E. Colak, J. Barfett, Generalization of deep neural networks for chest pathology classification in x-rays using generative adversarial networks, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 990–994.

[48] A. Srivastava, L. Valkov, C. Russell, M.U. Gutmann, C. Sutton, Veegan: Reducing mode collapse in gans using implicit variational learning, in: Advances in Neural Information Processing Systems, 2017, pp. 3308–3318.

[49] Z. Chen, Y. Tong, Face super-resolution through wasserstein gans, arXiv preprint arXiv:1705.02438 (2017).