JUST at ImageCLEF 2019 Visual Question Answering in the Medical Domain

Aisha Al-Sadi[1], Bashar Talafha[1], Mahmoud Al-Ayyoub[1], Yaser Jararweh[1], and Fumie Costen[2]

[1] Jordan University of Science and Technology, Jordan
asalsadi16@cit.just.edu.jo, talafha@live.com, {maalshbool, yijararweh}@just.edu.jo
[2] University of Manchester, UK
fumie.costen@manchester.ac.uk

Abstract. This paper describes our method for the Medical Domain Visual Question Answering (VQA-Med) Task of ImageCLEF 2019. The aim is to build a model that is able to answer questions about medical images. Our proposed model consists of sub-models, each specializing in answering a specific type of questions. Specifically, the sub-models we have are: "plane" model, "organ systems" model, "modality" models, and "abnormality" models. All of these models are basically image classification models based on pre-trained VGG16 network. We do not rely on the questions for the answers prediction since the questions on each type are repetitive. However, we do rely on them to determine the suitable model to be used for producing the answers and determine the suitable answer format. Our best model achieves 57% accuracy and 0.591 BLEU score.

Keywords: ImageCLEF 2019 · Visual Question Answering · Medical Image Interpretation · Medical Questions and Answers · VGG Network

1 Introduction

With the advances in the computer vision (CV) and natural language processing (NLP) fields, new challenging tasks emerge and one of them is Visual Question Answering (VQA), which grabbed the attention of both research communities. VQA is basically 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. Obviously, the difficulty level of the problem depends on the expected answer types, whether they are yes/no questions, multiple choice questions or open-ended questions.

Recently, VQA has been applied to different specific domains such as the medical domain. 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 of these challenges are massive and there are restrictions related to using them and integrating them into a single model. Thus, the Medical VQA is still at very early stages, but it is expected to improve over time [4].

This paper presents our participation in VQA-Med 2019 task [3], which is organized by ImageCLEF 2019 [6]. This is the second installment of this task[3] with the aim of answering questions about medical images. For that, we create sub-models, where each sub-model is specialized for answering a specific type of questions. All our models use the pre-trained convolutional neural networks (CNN), VGG16 [9], for visual features extractions.

The rest of paper is organized as follows. Section 2 presents the most relevant work, which includes the participants' models in the VQA-Med 2018 challenge [4]. Section 3 presents detailed analysis of the dataset, which we find useful in building our models. In Section 4, we present our proposed models for answering questions of each type. Validation results for all models and final test results are presented in Section 5. Finally, the paper is concluded in Section 6.

2 Related Works

The general VQA challenge[4] 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, number questions, and other questions. Different approaches were applied for the task and most solutions rely on deep learning techniques that combine the use of word embedding with different recurrent neural networks (RNN) for text embedding and features extraction, and CNN for visual features extraction supplemented with advanced techniques such as attention mechanisms.

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 challenge is organized for it. The first version of this competition is VQA-Med 2018 [4]. The dataset used in the 2018 version is different from the one used in the 2019 version. The 2018 version consists of 2,866 medical images and 6,413 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 questions types are mixed between asking

[3] The first VQA-Med task [4] was organized by ImageCLEF 2018.
[4] https://visualqa.org/index.html

about a "region" within the image, asking about what the image shows, yes/no questions, and other question types including asking about abnormalities shown in the image and image kind, etc. Five teams submitted their work, most of their approaches use deep learning techniques. They use pre-trained CNN models to extract image features such as VGG16 and ResNet [5]. There are many approaches based on 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, there are some teams that used advanced techniques in the task such as the stacked attention networks and multimodal compact bilinear (MCB) pooling.

JUST team [10] used VGG16 for image features extraction. They used an LSTM-based encoder-decoder model where they feed the question to the encoder and then concatenate the hidden state of the encoder with the image features to feed them to the decoder as the initial hidden states.[5]

The FSST team [2] dealt with the task as a multi-label classification problem. They extracted image features using VGG16 and word embedding of the question and feed it to a Bi-LSTM network to extract question features. Then concatenated question features and image features and fed them to a decision tree classifier.

TU team [11] provided two models. In the first model, which is basically the same architecture of [10], they used the pre-trained Inception-ResNet-v2 model to extract image features and Bi-LSTM instead of LSTM as [10]. In their second model, they computed the attention between the image features and the question features and concatenated it with the question features before feeding it to a Softmax layer for prediction.

NLM team [1] also created two models. For the first model, they used Stacked Attention Network (SAN) with VGG16 for image features and LSTM for question features. As for the second model, they used Multimodal Compact Bilinear pooling (MCB) with ResNet-50 and ResNet-152 for image features and 2-layer LSTM question features. In SAN model, they compute the attention over the image, then combine the image features and question features for the second attention layer, then pass it to a Softmax layer as a classification problem. For 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.

UMass team [8] used ResNet-152 in extract image features, and a pre-trained word embedding on Wikipedia pages, PubMed articles and Pittsburgh clinical notes for text features. They created multiple attention maps using co-attention mechanism between image features and text features. Then, they generated answers using a sampling method as a classification task.

[5] https://github.com/bashartalafha/VQA-Med

3 Dataset Description

The dataset used in VQA-Med 2019 consists of 3,200 medical images with 12,792 Question-Answer (QA) pairs as training data, 500 medical images with 2,000 QA pairs as validation data, and 500 medical images with 500 questions as test data. The data is equally distributed over four categories based on the question types which are: plane category, organ category, modality category, and abnormality category.

We can determine the question category from the question words, i.e., if the word 'plane' appears in the question, then this is a plane question. While if the words 'organ' or 'part' appear in the question, then this is an organ question. If the words 'normal', 'abnormal', 'alarm' or 'wrong' appear in the question, then this is an abnormality question. Otherwise, this is a modality question. This is useful for test data questions since the category of the question is not given like in the training and validation questions.

3.1 Plane Category Data

Question on planes come in one of the following formats: "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. Figure 1 shows main planes and their distributions in training and validation data. As evidence in this figure, the data is unbalanced, with some planes being more frequent than the others. In fact, this imbalance is noticeable across all categories data.

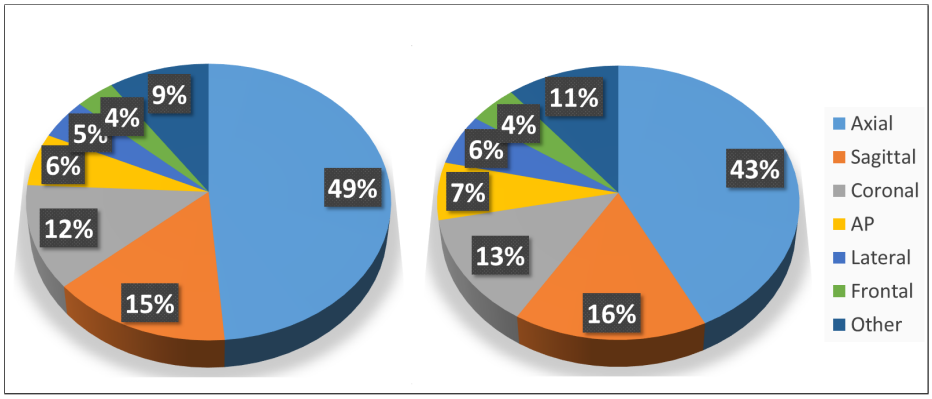

Fig. 1. Planes distribution in training data (left) and validation data (right).

3.2 Organ Systems Category Data

Question on organ systems come in one of the following formats: "what organ system is", "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", "what organ is this", etc. There are ten organ systems. Figure 2 shows all organ systems and their distribution in training and validation data.

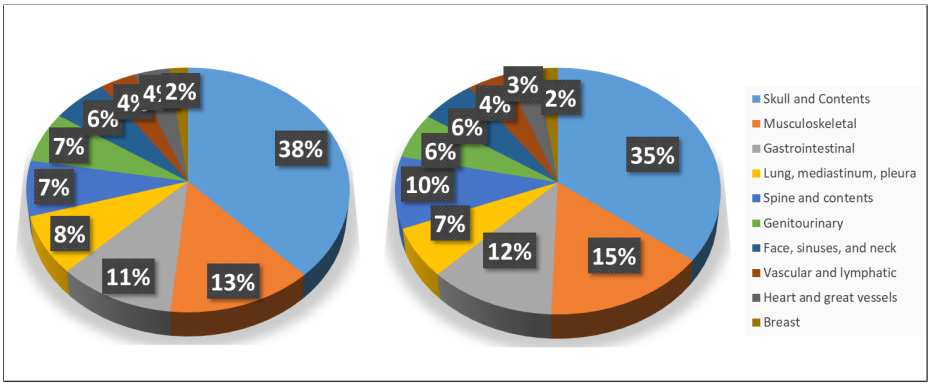

Fig. 2. Organ systems distribution in training data (left) and validation data (right).

3.3 Modality Category Data

There are eight main modality categories: XR, CT, MR, US, MA, GI, AG, and PT. Under each of these categories, there is a number of subcategories. 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.

The questions of modality part are more diverse, we can classify them into four types:

- 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.
- 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 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 1 shows modality questions types distribution in training and validation data. Figure 3 shows the distribution of images of each main category from all questions types. Note that we are unable to determine the modality in some

cases, such as "is this an mri image" with "no" as the answer. With the variations in modality questions types and large number of subcategories for some categories, we prepare different data formats in order to focus on specific aim in each model.

Table 1. Modality questions distribution

	Training	Validation
Type 1	1,380 (43%)	229 (46%)
Type 2	1,184 (37%)	179 (36%)
Type 3	445 (14%)	73 (14%)
Type 4	191 (6%)	19 (4%)
Total	3,200	500

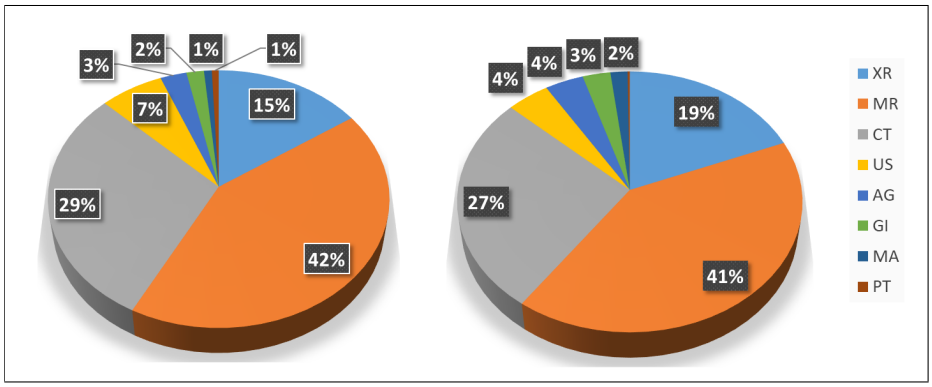

Fig. 3. Modality main categories distribution in training data (left) and validation data (right).

3.4 Abnormality Category Data

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

- Type 1: Questions asking about abnormality in the image. For example, "what is the abnormality/wrong/alarming in this image". This type represents 97% of abnormality training 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 types represents 3% of abnormality training questions and 5% of abnormality validation questions.

For Type 1 questions, there are 1,461 different abnormalities in the 3,082 training images, and 407 different abnormalities in the 477 validation images.

It is worth mentioning that the dataset has wrong answers for some images that might affect the model's accuracy. This is expected since the data was generated automatically. Even for non-medical people like ourselves, we are able to detect some errors, but it needs an expert to determine all wrong answers and correct them.

4 Methodology

Since we have different categories of questions, we create a special model for each category. Then, we combine them all in one model to be used for predicting answers. In order to use them correctly to answer a given question with a given image, we need to detect the suitable model to answer the question on the image and the question words. The following subsections describe the models we build for each subcategory before describing how to combine them.

4.1 Plane Model

The questions format on this category are repetitive and all questions have the same meaning even if they use different words. So, it is expected that the questions would not contribute anything in answer predictions and only the image can determine the plane answer. Hence, we deal with this part as an image classification task. We use the pre-trained model VGG16 with the last layer (the Softmax layer) removed and all layers (except the last four) frozen. The output from this part is fed into two fully-connected layers with 1024 hidden nodes followed by a Softmax layer with 16 plane classes. Figure 4 shows the plane model architecture in details. Since the data is unbalanced, we use class weights in order to give the classes with smaller numbers of images higher weights.

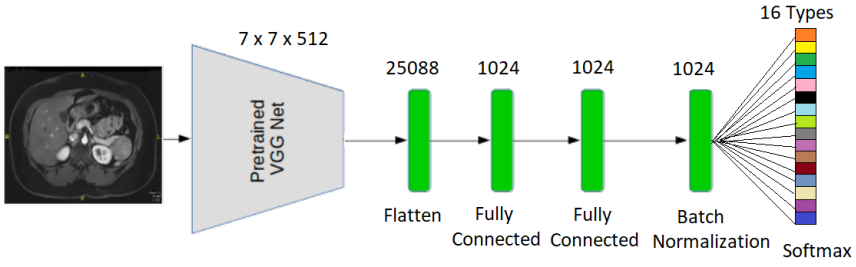

Fig. 4. Plane model architecture

4.2 Organ Model

The questions formats here are also repetitive and have the same meaning. So, we rely on the images only to get the organ system answer, i.e., as an image classification task. We use the same model architecture for plane model except that the last layer, which is the Softmax layer, has the ten organ systems classes.

4.3 Modality Models

As mentioned in the modality data description in the dataset section, this category has different variations in question types and different main categories and subcategories. For this part, we create many 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.

- M1, the general model, for classifying image modality into eight main categories (XR, CT, MR, US, MA, GI, AG, and PT).
- M2 model for distinguishing MR images from CT images.
- M3 model for distinguishing contrast from non-contrast CT images.
- M4 model for distinguishing contrast from non-contrast MR images.
- M5 model for classifying CT contrast types (GI/IV/GI and IV).
- M6 model for classifying MR weighting types (T1/T2/Flair).
- M7 model for classifying all CT subcategories.
- M8 model for classifying all MR subcategories.
- M9 model for classifying all GI subcategories.
- M10 model for classifying all ultrasound subcategories.

We did not create special models for the PT and AG categories as the data for building them are insufficient. The available data for the AG category is 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 are AN-Angiogram for AG and NM-Nuclear Medicine for PT.

4.4 Abnormality Models

For abnormality Type 2 questions, which ask if the image normal or abnormal, we create a special image classification model for that purpose with the same architecture of the plane model except that that the Softmax layer predicts normal/abnormal labels. While for Type 1 questions, which ask about the abnormality in the image, we experiment with different models since the task is quite challenging due to the given data being too small for the very large number of different abnormalities answers. The following are the main four methods with which we experiment.

- Method 1: We use an encoder-decoder architecture that takes an image as input and produces an answer as the output. The questions have the same meaning despite having different formats, hence, they are expected to not play an important role in producing the answers. We feed the image into an LSTM and use the hidden states of that LSTM as initial states of another LSTM for the answer sequence. We then add the encoder output and the decoder output, and pass the results into a dense layer followed by a Softmax layer.
- Method 2: In this method, we treat the problem as an image classification task using the same architecture of our previous models except that the Softmax layer has all unique abnormalities in the training data, which are 1,461 different abnormalities.
- Method 3: Firstly, we predict plane and organ classes of the test image. We then calculate the cosine similarity between the VGG16 features of the test image and all training images that have the same plane and organ of the test image. Finally, we get the most similar image and output its abnormality as the answer.
- Method 4: This is the same as Method 3, except that we take the two most similar images, and output the abnormality answer of the image which has the same abnormality question as that of the test image. If none of the two most similar image has the same question format, then we output the most similar image answer as in Method 3.

Algorithm 1 shows the steps taken to determine the required model for answer prediction. There are a lot of details that are not presented in the flowchart for simplicity, such as in modality questions Types 2-4, where different question formats are asking about specific things to determine the required model.

5 Evaluation and Results

In this section, we report the evaluation results of each of the previous models on the validation data separately. Then, we report our official results in the VQA-Med 2019 challenge. The evaluation metrics are accuracy and BLEU score [7]. For all models, we conduct several experiments for different optimizers and learning rates and report the best results. Table 2 shows the evaluation results for the validation data for each model belonging to the modality category.

Note that the accuracy of M10 is misleading since it predicts the dominant class all the time. It is worth mentiong that the overall modality validation accuracy, which is 75.4%, is not the average accuracy of all models. It is the accuracy of predicting modality validation questions using these models.

For the abnormality models, the accuracy of normal/abnormal model is 77.7%. While for other abnormality questions, Table 3 shows the validation accuracy and BLEU score for the different four methods. So, the best abnormality validation accuracy is 17.59% resulting from using Normal/Abnormal Model and Method 2 Abnormality Model.

Algorithm 1: Prediction Steps

Input: Image i and Question q
  if 'plane' word in q **then**
    Predict plane using Plane Model
  else if 'organ' or 'part' words in q **then**
    Predict organ plane using Organ Model
  else if 'normal', 'abnormal', 'alarm', or 'wrong' words in q **then**
    if q starts with "is this", "is there", "does this", "is the" or "are there" **then**
      Predict using Normal/Abnormal Model and answer yes/no based on that
    else
      if Method-1 **then**
        Predict using Abnormality Image Encoder-Decoder Model.
      else if Method-2 **then**
        Predict using Abnormality Image Classification Model.
      else if Method-3 **then**
        Predict using Abnormality Image Similarity Model.
      else
        Predict using Abnormality image Similarity with Question Format Model.
      end if
    end if
  else
    if Modality Type-1 Question **then**
      Predict main modality category
      Predict subcategory model based on the predicted main category from models M7-M10
    else
      Predict Answer using models M2-M6 based on what the question asks about
    end if
  end if
Output: Answer

Table 2. Modality Validation Results

Subcategories Models	Validation Accuracy (%)
M1 (General model)	88.6
M2 (CT/MR model)	97.7
M3 (contrast/non-contrast CT)	74.7
M4 (contrast/non-contrast MR)	85.7
M5 (CT contrast types (GI/IV/GI and IV))	92.8
M6 (MR weighting types (T1/T2/Flair))	86.3
M7 (All CT subcategories)	66
M8 (All MR subcategories)	50.8
M9 (All GI subcategories)	76.2
M10 (All ultrasound subcategories)	90
All modality	75.4

Table 3. Validation Abnormality Results

	Validation Accuracy (%)	Validation BLEU Score
Method 1	0	0.046
Method 2	14.7	0.175
Method 3	14	0.193
Method 4	13	0.189

For the whole model, here are the results. For plane questions, the validation accuracy is 76.2%, and, for organ questions, it is 74.2%. While the final modality model accuracy is 75.4% and the best abnormality model accuracy is 17.59%. So, the final validation accuracy is 60.85%.

For the VQA-Med 2019 challenge, we submit four runs of test data predictions. The four runs have the same predictions of plane, organ, and modality questions and the difference between them is only in the abnormality part. In each run, we use different method (Methods 1-4 as described in the abnormality model section). Table 4 shows our submissions results, Run-2 which deals with abnormality questions as an image classification has the best accuracy score and best BLEU score among our submissions.

Table 4. Our Results in VQA-Med 2019 Results

	Accuracy (%)	BLEU Score
Run 1	0.528	0.553
Run 2	0.534	0.591
Run 3	0.528	0.55
Run 4	0.528	0.55

After the competition was finished and the test answers were published publicly, we compared our predicted answers with the correct answers. We found that the plane part has accuracy 72.8%, organ systems 70.4%, modality part 64%, and abnormality part 8%. Moreover, we discovered that, for the abnormality part, we submitted our predicted answers without stop words. So, some of our correct answers were considered as false ones. Correcting this part increase the accuracy of this part to 18.4%. Another technicality that caused some of the correct answers produced by our systems to be considered false is the fact that some modality questions have single correct answers in the testing set where they should have multiple correct answers accounting for the different formats in which the correct answer may appear. For example, if the actual answer is “ct w contrast iv”, then, our system’s predicted answer “ct with iv contrast” should be considered correct. So, by taking into consideration the two previous notes, the actual overall accuracy of our model reaches 57%.

6 Conclusion

In this paper, we described our participation in the ImageCLEF VQA-Med 2019 task. The proposed model consists of sub-models based on the pre-trained VGG16 model. Our model's overall accuracy is 57% with 0.591 BLEU score. Accuracy of the plane, organ, and modality models are good (ranging between 65% and 72%), however, the abnormality model's accuracy is rather low (18%), due to the difficulty of the task especially with the small dataset available. In the future, we plan on seeking the help of a medical expert in order to correct wrong answers and collect new data for the abnormality part.

References

1. Abacha, A.B., Gayen, S., Lau, J.J., Rajaraman, S., Demner-Fushman, D.: 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)
2. Allaouzi, I., Benamrou, B., Benamrou, M., Ahmed, M.B.: 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)
3. Ben Abacha, A., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Müller, H.: Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In: CLEF 2019 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Lugano, Switzerland (September 9-12 2019)
4. Hasan, S.A., Ling, Y., Farri, O., Liu, J., Lungren, M., Müller, H.: Overview of the ImageCLEF 2018 medical domain visual question answering task (September 10-14 2018)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
6. Ionescu, B., Müller, H., Péteri, R., Cid, Y.D., Liauchuk, V., Kovalev, V., Klimuk, D., Tarasau, A., Abacha, A.B., Hasan, S.A., Datla, V., Liu, J., Demner-Fushman, D., Dang-Nguyen, D.T., Piras, L., Riegler, M., Tran, M.T., Lux, M., Gurrin, C., Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Garcia, N., Kavallieratou, E., del Blanco, C.R., Rodríguez, C.C., Vasillopoulos, N., Karampidis, K., Chamberlain, J., Clark, A., Campello, A.: ImageCLEF 2019: Multimedia retrieval in medicine, lifelogging, security and nature. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019), LNCS Lecture Notes in Computer Science, Springer, Lugano, Switzerland (September 9-12 2019)
7. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002)
8. Peng, Y., Liu, F., Rosen, M.P.: 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)

9. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
10. Talafha, B., Al-Ayyoub, M.: 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)
11. Zhou, Y., Kang, X., Ren, F.: 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)