STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LXVIII**, Number 1, 2023
DOI: 10.24193/subbi.2023.1.04

DOMAS: DATA ORIENTED MEDICAL VISUAL QUESTION ANSWERING USING SWIN TRANSFORMER

TEODORA-ALEXANDRA TOADER

Abstract. The Medical Visual Question Answering problem is a joined Computer Vision and Natural Language Processing task that aims to obtain answers in natural language to a question, posed in natural language as well, regarding an image. Both the image and question are of a medical nature. In this paper we introduce DOMAS, a deep learning model that solves this task on the Med-VQA 2019 dataset. The method is based on dividing the task into smaller classification problems by using a BERT-based question classification and a unique approach that makes use of dataset information for selecting the suited model. For the image classification problems, transfer learning using a pre-trained Swin Transform based architecture is used. DOMAS uses a question classifier and seven image classifiers along with the image classifier selection strategy and achieves 0.616 strict accuracy and 0.654 BLUE score. The results are competitive with other state-of-the-art models, proving that our approach is effective in solving the presented task.

1. Introduction

Visual Question Answering (VQA) is a task that combines both the Natural Language Processing (NLP) Field and Computer Vision (CV). The inputs of a VQA model are an image and a question addressed in natural language, question that can be answered from the given image. The output is of course the answer returned in natural language as well. Medical Visual Question Answering (MVQA) is a task that evolved from the VQA task by constraining the domain of the image and question to be the medical domain. Therefore, the images can take the form of pictures obtained using medical imaging, such as X-rays, MRIs, CT scans, as will be in our case while the questions can enquire about different aspects associated with the image. Using intelligent

Received by the editors: 14 June 2023.

2010 *Mathematics Subject Classification.* 68T45, 68T50.

1998 *CR Categories and Descriptors.* I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *Language parsing and understanding*; I.2.7 [**Artificial Intelligence**]: Applications and Expert Systems – *Medicine and science*.

Key words and phrases. Medical Visual Question Answering, Swin Transformer.

algorithms to solve the MVQA tasks could benefit the medical field immensely as such a model could provide a second opinion to medical professionals and could also make medical investigations more accessible.

One big challenge of MVQA is the limited amount of data that is available compared to the general VQA task that has been more widely explored. The lack of data for such an extensive task can lead models to overfit and not provide enough generalization. One recent approach that has been successfully used in the domain as a solution to the problems caused by small amounts of data is transfer learning, which focuses on using information gained from solving one task on a second related task. An architecture that proved to be very successful in association with transfer learning is the transformer architecture introduced by Vaswani et al. in [17]. Transformers are deep learning models based on the attention mechanism that proved to be very efficient in both NLP and CV, especially when pretrained on large amounts of data and then fine-tuned for specific tasks.

In this paper we introduce DOMAS, a deep learning model that solves the MVQA task on the Med-VQA 2019 dataset. The architecture is based on transforming the complex MVQA task into smaller image classification problems by selecting the image model using a model based on BERT architecture [6] applied on the questions and our dataset knowledge and then solving the image classifications using models based on an impressive computer vision backbone, introduced by Liu et al. called the Swin Transformer. [10]. Our approach achieves a 0.616 accuracy and 0.654 Bleu score on the VQA-Med-2019 test set which makes it comparable with other state-of-the-art models. The purpose of this paper is to answer the following research questions:

- Can the Swin architecture be an alternative to the more commonly used CNN based networks on this task?
- Does using information about the dataset improve the classification for modality models, especially for questions with affirmative or negative answers?

The structure of the paper will be as follows. Section 2 will present other approaches from the literature. Section 3 will provide a detailed description of the dataset. Our approach will be described in Section 4, while the experiments will be detailed in Section 5. The last section will provide the conclusions of the study as well as possible ideas for future work.

2. Related Work

The recent advancements in computer vision and natural language processing also led to advancement in the joined image and language task that is VQA. Most state-of-the-art models such as the ones of Chen at al. [4], Wang et al. [19] and Bao et al.[3] make use of the transformer architecture for vision-language pre-training and for other tasks such as [4] which uses the Vision Transformer for feature extraction as well. The general VQA task has the advantage of large datasets, performant models being pre-trained on millions of images and text samples which is not currently possible for MVQA. However, other methods and architectures can also be used for MVQA.

Many approaches have been proposed for the ImageCLEF 2019 Med-VQA dataset, some of them during the competition that proposed the dataset. The two highest-ranking teams at the ImageCLEF 2019 competition for the VQA-Med task combined features extracted from image and text using a fusion algorithm. Yan et al. [20] used a VGG-16 [14] inspired network combined with Global Average Pooling for image feature extraction and the basic BERT [6] model as the question encoder. The fusion of the two types of features extracted was achieved by using multi-modal factorized bilinear pooling with co-attention [21]. Minh Vu et al. [18] also use a CNN based network, namely ResNET-152 [7], to extract image features and BERT for question features. The features are fused using an attention mechanism and global image features are obtained, while the question features are also linearly transformed to obtain global question features. The global features are then further combined using a bilinear transformation. Some contestants also made use of the nature of the dataset and divided the problem into four different problems, one for each type of question. Zhou et al. [22] propose a different type of model for the plane, organ and modality questions, where a classifier is used to get the answer, and for abnormality questions where a generative method is used. The simple classifier consists of an Inception ResNet-V2 [16] for image feature extraction and BERT for question embeddings. The features are combined through a Multi-Layer Perceptron (MLP). The generator used for the abnormality questions consists of a sequence-to-sequence model. The encoder part is similar to the classifier while the decoder consists of a long short-term memory network (LSTM), and it continuously generates the probability distribution of the next word. Another interesting submission is the one of the JUST team [2] which creates an individual model for each type of question as well. They consider the questions to be repetitive and therefore each model is in fact an image classification model or a combination of image classification models. We can observe that all proposed models used pre-trained networks for feature extraction as the models benefit tremendously from transfer learning given the dimension of the dataset. Particularly, the proposed models also use CNN based networks for image feature extraction.

More recent approaches on the ImageCLEF 2019 VQA-Med dataset also make use of the transformer architecture and obtain improved results. Ren at al. [12] propose CGMVQA, a model that can switch between a classification and a generative mode, by changing only the loss function and the output layer, in order to better fit the approached problem. They divide the task into five subtasks depending on the type of question: yes/no questions, organ, plane, modality and abnormality. To obtain the image features they extract from different convolutional layers of a ResNet-152. The questions are tokenized, and token, segment and positional embeddings are used to obtain the final features. These two types of features are used in the classification mode, for the generative mode, masked answers are also added as the method used for the generation is masking position by position. To get the final outputs, the features are fed into a slightly changed Transformer network. Another method that makes greater use of transformer capabilities is proposed by Khare et al. [8] where the authors propose MMBERT (Multimodal Medical BERT), a BERT like architecture that is pre-trained using self-supervised learning. The model is pretrained on medical images and their corresponding captions using MLM. The image features are extracted as in [12] and the captions are modified by replacing medical terms with the [MSK] token and then the embeddings are obtained using BERT. The obtained embeddings are passed through a BERT-like encoder and then a classifier is used to predict the initially masked word.

3. Dataset Description

The dataset we are going to use in our experiments is the VQA-Med-2019 dataset, introduced at the ImageCLEF 2019 competition for the VQA task [1]. The dataset contains 4200 images selected from MedPix database and 15 292 corresponding questions divided in the following way. For training 3200 images were allocated as well as 12 792 question-answer pairs; for validation 500 images with 2000 question answers pairs and for the test dataset, 500 images and questions.

The questions were divided into four different categories: Organ, Plane, Modality and Abnormality. The plane category includes images in 16 different planes, namely Axial; Sagittal; Coronal; AP; Lateral; Frontal; PA; Transverse; Oblique; Longitudinal; Decubitus; 3D Reconstruction; Mammo-MLO; MammoCC; Mammo-Mag CC and Mammo-XCC. The organ category has the smallest number of classes, the possible answers to all the questions belonging to a set of ten organs and organ systems namely: Breast; Skull and Contents; Face, sinuses, and neck; Spine and contents; Musculoskeletal; Heart and great vessels; Lung, mediastinum, pleura; Gastrointestinal; Genitourinary; Vascular and lymphatic. The modality category is slightly more complex than the previous two. There are 36 modalities, and the question can refer to the type of modality used, either what or yes/no questions. There are also questions related to contrast/noncontract in the image, what type of contrast is used and specifics of MRIs (if the images are t1-weighted, t2-weighted or flared). In total, there are 44 possible answers for all modality questions. Therefore, the modality category includes yes/no questions, what question and other closed questions. The abnormality category includes both yes/no questions, that inquire about the state of the image; if it is normal/abnormal and what questions, that inquire about the abnormality shown in the picture. The abnormality category is the most complex, with 1485 possible answers in the training set. One concern that we will consider with this category is the large number of answers to the validation questions that are not found through the answers in training. This could be a possible issue when treating this problem as a classification since the model will not be able to learn the classes that are not present while training.

4. Proposed Approach

Following the lead of papers such as the ones proposed by Zhou et al. [22] and Al-Sadi et al.[2], we propose a model that divides the complex MVQA task into smaller and more manageable problems. By making use of the dataset knowledge, we can treat each individual problem as an image classification one and obtain comparable results with the current approaches from the literature. Unlike the presented approaches which use CNN based networks for image related tasks we choose to make use of an attention based state-of-the-art model for image classification, the Swin Transformer. We use pre-trained versions of Swin and finetune them to our specific classifications in order to obtain our results. We aim to see if the abilities of this model perform as well on these downstream tasks and moreover, if this transformer-based architecture can surpass the widely used CNN-based architectures. In order to apply the image classification models we need to divide the types of questions in a way that makes sense from the point of view of the created classes, therefore, we chose to create individual models for organ, as all answers are a type of organ or organ system, plane, as all answers in this category are planes in which the image is taken and abnormality. For the modality questions we created four models depending on the type of questions and possible answers that would create the classes. Therefore, we obtained a contrast model, used for questions that inquire about the way the image was taken, with contrast or noncontrast; a contrast type model, used for questions that inquire about the type of contrast used, possible answers being gi/iv/gi and iv; a type of weight model, for questions that examine whether the image is t1, t2 or flair weighted and finally the modalities model which predicts classes representing the type of image modalities.

4.1. **Model Overview.** DOMAS is therefore a model that joins two types of models: a question classification model and several image classification ones. An overview of the flow model can be observed in Figure 1.

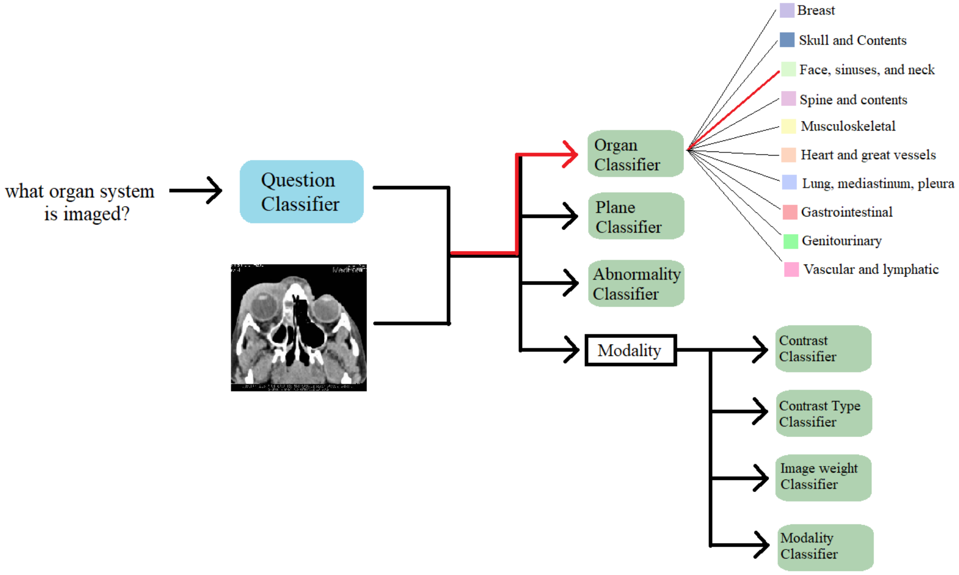


FIGURE 1. Model overview

As it can be seen, in Figure 1, the question is first passed through the question classifier which predicts which type of question it is organ, plane, modality or abnormality. Based on the predicted class an image classification model is selected. For example, in Figure 1 the question is classified as an organ question therefore the organ model is selected. For organ, plane and abnormality the corresponding model is selected. For modality, four models are available. To select the type of modality model we make use of dataset analysis. We observe that there is a limited number of questions for each of the four modality models in both train and validation that have as answer our classes. All questions that do not fit in one of these question lists and start with “is” or “was” are closed-ended questions, meaning the answer is either yes or no, and the remaining questions are image modality related questions. To handle yes or no questions, instead of treating them as classes and creating new models we further analyze them and make use of the initial four models.

We observe that each of the yes/no questions refers, in fact, to the information obtained using the previously mentioned models. Therefore, we create a function that extracts the type of modality model from the question and also the expected answer. For example, from the question "is this a t1 weighted image?" we extract the type of model, which is the weighting model, and also the expected class which is "t1". Therefore, we fed the image into the weighting classification model and if the class predicted by the model matches the expected one, we predict the answer "yes", and "no" otherwise. For modality type closed-ended questions we notice that the questions only enquire about "MRI" and "CT" scans, we return either "CT" or "MR" as the expected class. However, these answers are not classes for the modality model on their own. For that reason, we replace the perfect match for the yes answer with an inclusion. For example, if the expected class is "CT" and the predicted class is "CT - myelogram" we return the answer "yes" as "CT" is contained in the answer.

After the correct model has been selected, the image is given as input to the model and the predicted class is obtained and transformed from its numeric representation used by the model into the textual one using the corresponding model dictionary completing the inference.

4.2. **Models Architectures.** The total of eight models, one for text and seven for images, have been trained individually. More details about the architectures and training process are available in this subsection.

The question classification model is used for differentiating between the four major types of questions. It is based on a pretrained BERT [6] model that we finetune for our classification. The question is pre-processed by applying the BERT tokenizer. The model consists of the pre-trained BERT, followed by a dropout layer and a linear layer produces the final prediction. Lastly, ReLU activation is applied. One concern regarding this model was that it might affect the overall accuracy of the model by misclassifying the questions which would result in an erroneous result from the start. Fortunately, the model achieves 100% accuracy on the task thus eliminating the concern and providing the model with a greater generalization power than a classification based on pattern matching.

For the image classification models we expand the dataset by using image augmentations. After testing with several augmentation types and excluding the ones that could alter the image in such way that the label would no longer be fitting, such as random rotations, horizontal and vertical flips, we decided to use random resized crop with a size of 224, which randomly crops the image and that resize it to the given size, as it was the version that yielded the best improvements.

For each type of image model we experimented with a Swin-based classifier. The architecture of the model can be observed in Figure 2.

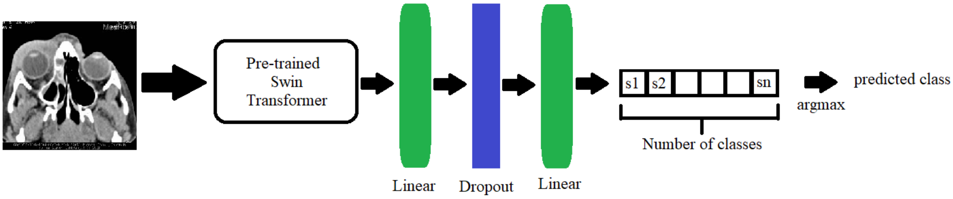


Figure 2. Image classification model architecture

As can be seen, the resized image enters the Swin model pretrained on the ImageNET dataset [5]. We experiment with different model versions such as tiny, small and base. The head of the model is modified in order to change the ImageNet classification task with our task. Therefore, the output of the Swin backbone is passed first through a linear layer. Next, ReLU activation and a dropout layer [15] are applied and finally the last linear layer obtains the final class. After analyzing the dataset, we observed that some classes are not present in the training dataset but appear in the validation dataset. We eliminated these classes, namely, "pet-CT fusion" from the modality split and "Mammo-XCC" from the plane classes. After this process the number of classes were 10 for organ classifier, 14 for plane, 1485 for abnormality and 44 for modality which were split into two for contrast mode, three for weighted model, three for contrast type model and 34 for modalities; the remaining two were the yes/no answers.

We chose Cross Entropy Loss since we performed multi-class classification and the Adam [9] and SGD optimizers depending on the case. Parameter settings and other implementation details will be further detailed in Section 5.

More details about hyperparameters settings as well as the results obtained by our model will be presented in Section 5

5. Experiments and Results

5.1. **Experimental Setup and Results.**

We trained and evaluated our models using a Google Colaboratory environment. The training was completed for each model using the integrated Nvidia T4 GPU. As mentioned before, we trained each model individually and selected the best models based on the classification accuracy. For evaluating the model we used two metrics namely the strict accuracy and Bleu score [11]. The parameter settings of the final models can be seen in Table 1. The parameters were chosen empirically. To select the Swin version for each model we performed experiments with the pre-trained tiny, small and base versions and selected the best performing one based on accuracy and F1-score. If different versions obtained the same results we selected the smallest model between them.

Model	**Swin Version**	**Optimizer**	**Linear Layer Output Dim**	**Activation Function**	**Dropout Rate**
Organ	Tiny	Adam	384	ReLU	0.5
Plane	Tiny	Adam	384	ReLU	0.5
Abnormality	Small	Adam	1536	ReLU	0.5
Modality Contrast	Small	SGD	384	ReLU	0.5
Modality Contrast Type	Tiny	Adam	128	ReLU	0.5
Modality Weighting	Base	Adam	384	ReLU	0.5
Modality Modalities	Small	Adam	384	ReLU	0.5

TABLE 1. Parameters Setting for the employed models

We used the models with the configurations presented in Table 1 in combination with the question classifier, which achieved 100% accuracy and obtained the final results which are presented in table 2

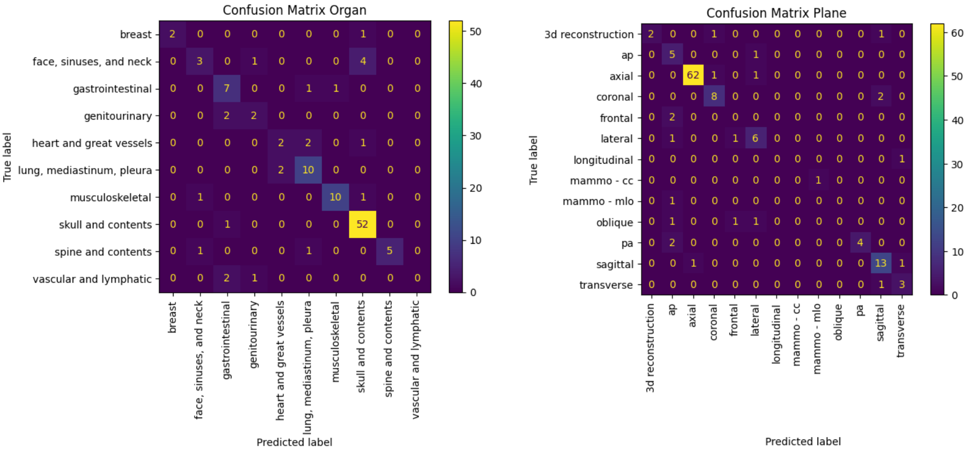
Metric	**Organ**	**Plane**	**Modality**	**Abnormality**	**Overall**
Strict Accuracy	0.744	0.824	0.824	0.072	0.616
BLEU Score	0.789	0.838	0.774	0.215	0.654

TABLE 2. Model Results

As we can see from the results, our model obtains a 61.6% score in strict accuracy and 65.4% BLEU score. The lowest performing model is as expected the abnormality model since the number of classes is indeed the largest. We can also observe that some models have high BLEU scores compared to accuracy which could mean that the model does not give a perfect answer but could give a close one.

5.2. **Discussion and comparison to related work.** Our model achieves promising results on the MVQA 2019 dataset, especially for the plane, organ and modality questions. For the abnormality model the high number of classes corresponding to the answers as well as the existence of many classes in both validation and test dataset which are not found in the train set lead to a lower performance. For the other models we will further discuss the results in order to better understand the strong points and the shortcomings of the models.

For the organ model, we found that some questions in the validation and test dataset have more than one organ system given as answers. More specifically, in the test dataset there are nine such answers out of the 125. This is not a case that we treated during training or as postprocessing, therefore the model cannot make a correct prediction from the perspective of strict accuracy. However, when analyzing these answers compared to the predictions, we found out that in seven out of the nine cases our model predicted an organ system that was part of the answer which partially increased the Bleu score. In order to analyze if there is a pattern in the misclassifications of the model, we constructed the confusion matrix, Figure 3 (left), from which we left out the answers that have more than one organ system. We can observe that the most frequent misclassifications are between face, sinuses and neck and skull and contents or gastrointestinal and genitourinary and vascular and lymphatic, but most classifications are indeed correct.



	breast	face, sinuses, and neck	gastrointestinal	genitourinary	heart and great vessels	lung, mediastinum, pleura	musculoskeletal	skull and contents	spine and contents	vascular and lymphatic
breast	2	0	0	0	0	0	0	1	0	0
face, sinuses, and neck	0	3	0	1	0	0	0	4	0	0
gastrointestinal	0	0	7	0	0	1	1	0	0	0
genitourinary	0	0	2	2	0	0	0	0	0	0
heart and great vessels	0	0	0	0	2	2	0	1	0	0
lung, mediastinum, pleura	0	0	0	0	2	10	0	0	0	0
musculoskeletal	0	1	0	0	0	0	10	1	0	0
skull and contents	0	0	1	0	0	0	0	52	0	0
spine and contents	0	1	0	0	0	1	0	0	5	0
vascular and lymphatic	0	0	2	1	0	0	0	0	0	0

	3d reconstruction	ap	axial	coronal	frontal	lateral	longitudinal	mammo - cc	mammo - mlo	oblique	pa	sagittal	transverse
3d reconstruction	2	0	0	1	0	0	0	0	0	0	0	1	0
ap	0	5	0	0	0	1	0	0	0	0	0	0	0
axial	0	0	62	1	0	1	0	0	0	0	0	0	0
coronal	0	0	0	8	0	0	0	0	0	0	0	2	0
frontal	0	2	0	0	0	0	0	0	0	0	0	0	0
lateral	0	1	0	0	1	6	0	0	0	0	0	0	0
longitudinal	0	0	0	0	0	0	0	0	0	0	0	0	1
mammo - cc	0	0	0	0	0	0	0	0	1	0	0	0	0
mammo - mlo	0	1	0	0	0	0	0	0	0	0	0	0	0
oblique	0	1	0	0	1	1	0	0	0	0	0	0	0
pa	0	2	0	0	0	0	0	0	0	0	4	0	0
sagittal	0	0	1	0	0	0	0	0	0	0	0	13	1
transverse	0	0	0	0	0	0	0	0	0	0	0	1	3

FIGURE 3. Confusion Matrices for Organ and Plane

We constructed a confusion matrix for the plane model, Figure 3 (right), classification on the test set as well. As we can see the classes found while testing are fewer than the ones found in the train dataset. For this model there are fewer misclassifications, as expected after seeing the metrics. However, we can observe that the mistakes are more frequent for the less represented classes.

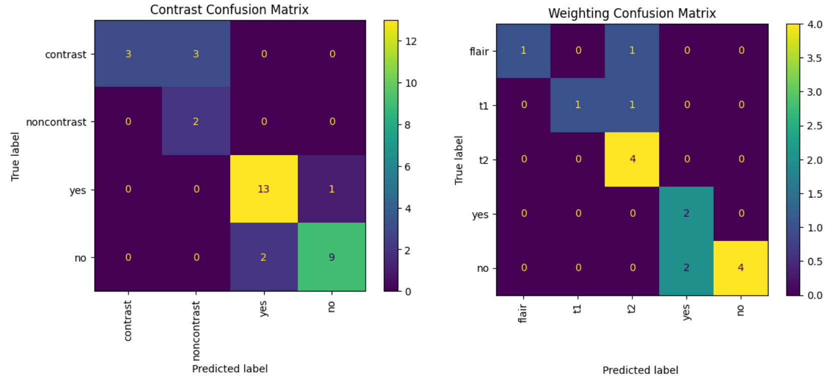
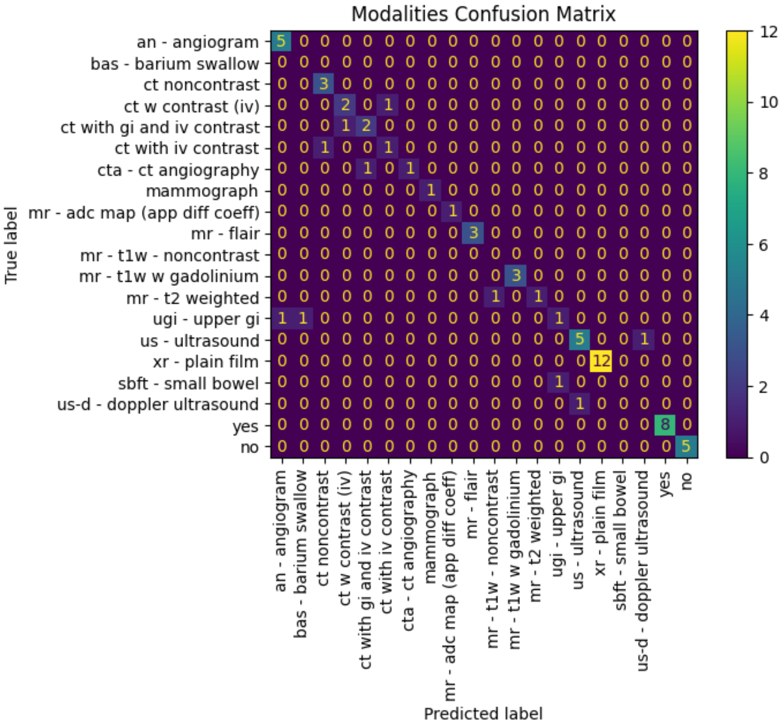


Figure 4. Confusion Matrices for Contrast and Weighting classifications

For the modality models we took a closer look at the results of the contrast, weighting and modality models. Even though our models do not treat yes and no as classes in the classification we constructed the matrices based on the final answer given by the modality model which was constructed to give an affirmative or negative answer as described in Section 4. For the contrast classification, one class that did not appear in training or validations set was found when testing therefore we removed the singular answer. We can see in Figure 4, that the model tends to predict the noncontrast class instead of contrast one rather than the other way around, which we also found to be true when analyzing more deeply the true meaning of the yes and no answers. For the weighting model we observed that there is a tendency to predict ‘t2’ type which is the best represented class out of the modality weighting types.

For the modalities confusion matrix, Figure 5, we used again the test classification data. We notice that many discrepancies between the predicted and true classes are generated by the different types of CT scans which explains the good prediction for yes or no answers since they only inquire whether the image is or not a CT scan or an MR image meaning that the prediction includes more classes which are usually only confused with one another. We can also see confusions between different types of ultrasounds or similar classes which explains why the Bleu score is higher that the strict accuracy.

After analyzing the results of the model overall as well as individually for each component we can affirm that using Swin as an image classification method for this dataset offers promising results especially for the organ, plane and individual modality models. Moreover, our approach to the models for



	an - angiogram	bas - barium swallow	ct noncontrast	ct w contrast (iv)	ct with gi and iv contrast	ct with iv contrast	cta - ct angiography	mammograph	mr - adc map (app diff coeff)	mr - flair	mr - t1w - noncontrast	mr - t1w w gadolinium	mr - t2 weighted	ugi - upper gi	us - ultrasound	xr - plain film	sbft - small bowel	us-d - doppler ultrasound	yes	no
an - angiogram	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bas - barium swallow	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ct noncontrast	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ct w contrast (iv)	0	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ct with gi and iv contrast	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ct with iv contrast	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
cta - ct angiography	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
mammograph	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
mr - adc map (app diff coeff)	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
mr - flair	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0
mr - t1w - noncontrast	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
mr - t1w w gadolinium	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0
mr - t2 weighted	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0
ugi - upper gi	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
us - ultrasound	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	1	0	0
xr - plain film	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0	0
sbft - small bowel	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
us-d - doppler ultrasound	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
yes	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0
no	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5

FIGURE 5. Confusion Matrix for Modalities classification

the modality questions makes these sub-tasks suitable for image classification by giving each model meaningful classes to discern from. By not using the yes and no answers as classes by themselves, but rather understanding from the question what the desired information is and constructing the inference accordingly we created models that can in fact predict the yes and no answers correctly, most of the time, without the need for extra classes or models.

In order to have the best understanding of where our results stand, we compared our model with other literature approaches, namely the first five teams of the Image-CLEF 2019 competition according to the dataset as well as the JUST [2] team since they also used an image classification approach and the two transformer-based approaches we mentioned in Section 2. The comparative results can be observed in Table 3, the results obtained by our proposed model being highlighted. As the test set was the same for all the papers, we did not replicate the experiments but rather got the corresponding results from each paper.

Model	**Organ**	**Plane**	**Modality**	**Abnormality**	**Overall Accuracy**	**Overall BLEU**
Henlin [22]	0.736	0.768	0.808	0.184	0.624	0.644
Yan [22]	0.736	0.768	0.808	0.168	0.62	0.640
Minhvu [18]	0.76	0.776	0.84	0.088	0.616	0.634
TUA1 [22]	0.792	0.816	0.744	0.072	0.606	0.633
UMMS [13]	0.736	0.76	0.672	0.096	0.566	0.593
JUST [2]	0.704	0.728	0.64	0.064	0.534	0.591
CGMVQA [12]	0.784	0.864	0.819	0.044	0.64	0.659
MMBERT [8]	0.768	0.864	0.833	0.14	0.672	0.69
DOMAS	**0.744**	**0.824**	**0.824**	**0.072**	**0.616**	**0.654**

Table 3. Results compared with literature approaches

Compared to the models submitted for the competition our model achieved the highest Bleu score and achieved the third score in accuracy. As for the individual models, it achieves the highest accuracy for the plane classification, ranks second for modality and third for plane. Compared to the JUST team which had an image classification approach as well but used VGG as a backbone for classification, our results rank higher in all the categories which proves that the Swin Transformer is a very suitable option for this task and could potentially be seen as a good alternative to the CNN based networks that are very popular choices for the MVQA task. Compared to the two transformer-based models, our model does not obtain better results, but the results are quite close. Our model surpasses the CGMVQA model for modality and abnormality results and it obtained a very close Bleu score. The MMBERT does perform better in all categories, which shows the great impact of extra training on more data.

Overall, our model obtains competitive results with the state-of-the-art models in all categories except the abnormality one where the high number of classes and small amount of data take a toll on the model's ability to effectively classify all the abnormality answers. We plan on further improving these results using various methods that are detailed in Section 6.

6. Conclusions and Future Work

In conclusion, we proposed DOMAS, a model that breaks down the complex MVQA task into multiple image classification tasks by processing the questions using a BERT-based architecture and knowledge we extract while performing exploratory data analysis on our dataset, Med-VQA 2019. Our approach proposes a Swin Transformer backbone for the image classification models as well as a unique way to select which models need to be developed based on the nature of the question in a way that all classes make sense from an image classification point of view. Our model achieves 0.616 score in strict accuracy and 0.654 BLEU score which ranks it the third in accuracy and first in BLEU among the participants in the ImageClef 2019 competition. The obtained results are also comparable with current state-of-the-art transformer-based model.

Our method shows that using the Swin Transformer architecture for working with images is beneficial in this task and could be seen as a viable alternative for the more popular CNN based networks which answers our first research question. Our model performs well for the organ, plane and modality models and we observe that our original approach of splitting the modality questions into four subcategories and obtaining the yes and no answers as a postprocessing of the model's output based on dataset knowledge rather than treating the answers as classes drastically improves the results in this category, making the response to our second research question an affirmative one. Moreover, using a BERT-based classifier, as opposed to a simpler pattern matching, for the questions also provides a certain generalizing power to the model even though the questions provided by the dataset are limited and quite redundant. However, the model also has some shortcomings such as lower level of robustness since we do use dataset specific knowledge and also treat the problem as a classification which makes the answer dependent on the training classes. We can observe this issue in the abnormality model where one cause of the lower performance could be the large number of answers in the test and validation datasets that are not present while training.

Future work plans include addressing some of these drawbacks. We plan on replacing the pattern matching methods used in the modality model classification with an intelligent approach that would predict the correct model as in the case of the question classification. Moreover, for yes and no questions we would aim to obtain the model as well as the expected class that would later be compared with the prediction in order to obtain the affirmative or negative answer. We also believe that the organ model could be improved by treating the case of multiple organ systems given as answer. As there is no information in the questions that could indicate that the expected answer is a compound one, we believe that one viable approach would be to return all answers which confidence exceeds a certain threshold. Finally, to improve the abnormality model, we would like to explore the possibility of using a generative model instead of a classification one or to use extra data for training this current classification model.

References

[1] Abacha, A. B., Hasan, S. A., Datla, V. V., Liu, J., Demner-Fushman, D., and Müller, H. VQA-Med: Overview of the medical visual question answering task at ImageCLEF 2019. *CLEF (working notes) 2*, 6 (2019).

[2] Al-Sadi, A., Talafha, B., Al-Ayyoub, M., Jararweh, Y., and Costen, F. JUST at ImageCLEF 2019 Visual Question Answering in the Medical Domain. In *CLEF (working notes)* (2019).

[3] Bao, H., Wang, W., Dong, L., Liu, Q., Mohammed, O. K., Aggarwal, K., Som, S., and Wei, F. VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts, 2022.

[4] Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., et al. PaLI: A Jointly-Scaled Multilingual Language-Image Model. *arXiv preprint arXiv:2209.06794* (2022).

[5] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009), pp. 248–255.

[6] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).

[7] He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition, 2015.

[8] Khare, Y., Bagal, V., Mathew, M., Devi, A., Priyakumar, U. D., and Jawahar, C. MMBERT: Multimodal BERT Pretraining for Improved Medical VQA. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)* (2021), IEEE, pp. 1033–1036.

[9] Kingma, D. P., and Ba, J. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980* (2014).

[10] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (2021), pp. 10012–10022.

[11] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (2002), pp. 311–318.

[12] Ren, F., and Zhou, Y. CGMVQA: A New Classification and Generative Model for Medical Visual Question Answering. *IEEE Access 8* (2020), 50626–50636.

[13] Shi, L., Liu, F., and Rosen, M. P. Deep Multimodal Learning for Medical Visual Question Answering. In *CLEF (working notes)* (2019).

[14] Simonyan, K., and Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition, 2015.

[15] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *The journal of machine learning research 15*, 1 (2014), 1929–1958.

[16] Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In *Proceedings of the AAAI conference on artificial intelligence* (2017), vol. 31.

[17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention Is All You Need. *Advances in neural information processing systems 30* (2017).

[18] Vu, M., Sznitman, R., Nyholm, T., and Löfstedt, T. Ensemble of Streamlined Bilinear Visual Question Answering Models for the ImageCLEF 2019 Challenge in the Medical Domain. In *CLEF 2019-Conference and Labs of the Evaluation Forum, Lugano, Switzerland, Sept 9-12, 2019* (2019), vol. 2380.

[19] Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., et al. Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks. *arXiv preprint arXiv:2208.10442* (2022).

[20] Yan, X., Li, L., Xie, C., Xiao, J., and Gu, L. ImageCLEF 2019 Visual Question Answering in the Medical Domain. *Zhejiang University* (2019).

[21] Yu, Z., Yu, J., Fan, J., and Tao, D. Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering. In *Proceedings of the IEEE international conference on computer vision* (2017), pp. 1821–1830.

[22] Zhou, Y., Kang, X., and Ren, F. TUA1 at ImageCLEF 2019 VQA-Med: a Classification and Generation Model based on Transfer Learning. In *CLEF (Working Notes)* (2019).

Department of Computer-Science, Faculty of Mathematics and Computer Science, Babeș-Bolyai University, Cluj-Napoca, Romania

Email address: teodora.toader@stud.ubbcluj.ro