Grenze International Journal of Engineering and Technology, Jan Issue

Simple Methods is All You Need for Medical VQA: An ImageCLEF's Med-VQA Task Methods Review

Jitender Singh[1] and Surender Singh[2]
[1-2]Department of Computer Science and Engineering, Chandigarh University, Chandigarh, India
Email: 20mai1035@cuchd.in, surender.e8126@cumail.in

***Abstract*—Visual Question Answering (VQA) is an area of AI where the model takes an image and a free-form natural language question related to the image and generates an answer as the output to the question in natural language. In the medical domain, VQA generates the natural language answer to the given question related to a medical scan such as X-Ray, CT, MRI etc. With the public availability of medical VQA datasets introduced by ImageCLEF in 2018, it has become an active area of research. In this paper, we will discuss all four years of ImageCLEF's Med-VQA competition datasets and methods used by the participants to understand which methods perform better on the medical VQA problem. Also, we describe the limitations of the ImageCLEF dataset and a simple VQAMixUp technique that can be used with simple architectures such as VGG16 and GRU and performs equivalent to the large ensemble models of the winners of the competition.**

***Index Terms*— deep learning, visual question answering, medical VQA, ImageCLEF, VGG and GRU.**

I. INTRODUCTION

Artificial intelligence is being utilised to develop intelligent algorithms for medical imaging pattern recognition to improve clinical decision-making and patient involvement (AI). Many researchers are interested in image retrieval and aided diagnosis applications of automated medical imaging analysis. Medical systems employed hard-coded computer vision techniques before deep learning and neural networks. These algorithms were complicated and time-consuming. Convolution Neural Networks (CNNs) can be trained from medical data and eliminates the need for step-by-step algorithms. However, CNNs require annotated datasets but annotating medical datasets is expensive and requires experts. Ample research has been done to design network architectures that use less data and performs similar or better. Textual data from doctor reports are also essential for predicting image problems. Images, text, and tabular data (such as age, gender, and weight) can be used simultaneously as input to the neural networks. Different input modalities improve model correctness and dependability which leads to better predictions. Visual Question Answering (VQA) uses text and images as input and the network needs to predict the answer to the question using the image's content. More on VQA in the next section. Public availability of datasets and methodologies have made significant progress in the field of VQA in recent years. ImageCLEF [1] is the first one to introduce a public medical VQA dataset through ImageCLEF's Med-VQA 2018 [2] competition. They subsequently held the Med-VQA competition for four years.
In this review paper, our objective is to summarize the model architectures and training methods used by the participants of ImageCLEF's four years (2018-2021) Med-VQA task. By reviewing the extensive variety of

Grenze ID: 01.GIJET.9.1.86

methods used to get the best accuracy score, we noticed that some of the researchers have used simple methods and achieved top scores in the competition. We discuss all the methods used in detail and listed the comparisons in Tables II-V. We also found some limitations in the ImageCLEF datasets. Finally, we describe Singh et al. [7]'s methodology, which provides SOTA results on the ImageCLEF 2020 and 2021's datasets, depicting how a specific combination of training methods and basic neural network architectures can provide similar or better results compared to the winners of the competition.

II. Visual Question Answering (VQA)

Visual Question Answering (VQA) involves answering questions about the image's content. Machine learning models need both vision and language understanding to effectively answer an image-related question. VQA with natural images is complex due to a vast variety of images and open-ended questions. Large datasets and crowdsourced datasets make efficient model training possible. Still, the free-form and open-ended questions can ask anything about the image. Therefore, the model needs to find every single object and detail of the image.

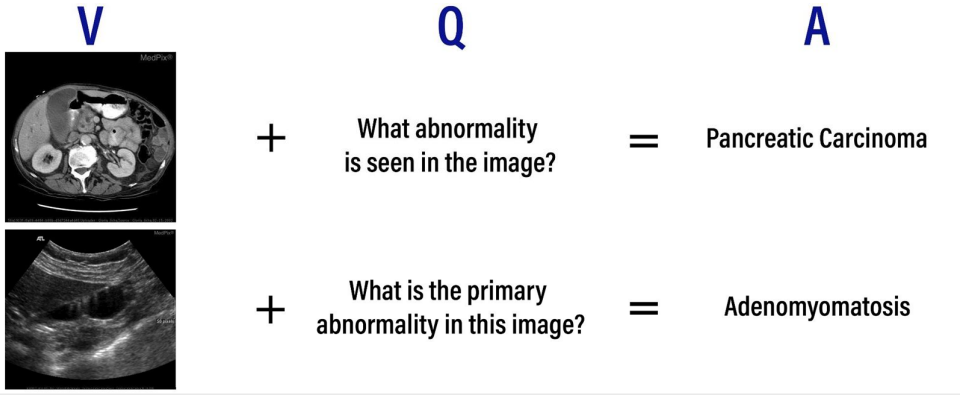

Figure 1. VQA Sample – CT Image with corresponding free-form question and answer pair

In contrast to the natural images VQA, the medical domain is constrained by the limited variety of scans. Figure 1 depicts a medical VQA sample. However, the number of questions one can ask is still large but smaller than the natural images. A major issue, as compared to natural images, is the expertise required to manually label the medical data. This leads to expensive data labelling. Hence, data is limited. Recently, Medical VQA became an active area of research with the induction of several publicly available benchmark datasets and the organization of challenges. Medical VQA systems can help overburdened and under-resourced healthcare systems worldwide. However, limited annotated data and domain-specific characteristics make training a reliable VQA model in medicine difficult. Therefore, rather than training complex models which require huge data, we can focus on simple methods and architecture. With a lower number of parameters, the model will be more generalized and less prone to overfitting. Due to the limited nature of questions, one can ask about the medical scan, the Med-VQA task is simpler than the natural images VQA. Therefore, the models for questions' features extraction can be basic models such as LSTM and GRU.

TABLE I. IMAGECLEF-VQA-MED DATASETS DETAILS

Year	Set Type	Images	Questions	Answers
2018	Train	2278	5413	4471
2018	Val	324	500	456
2018	Test	264	500	389
2019	Train	3200	12792	1552
2019	Val	500	2000	470
2019	Test	500	500	166
2020	Train	4000	38	332
2020	Val	500	26	232
2020	Test	500	40	NA
2021	Train	4500	40	332
2021	Val	500	16	236
2021	Test	500	24	NA

Unlike general-domain VQA, medical VQA has its unique set of concerns and challenges. Medical VQA datasets are difficult to create since expert annotation is costly because of its high need for professional expertise, and QA pairings can't be created directly from images. The capacity to handle medical terminology as well as non-medical terminology used by non-specialists is another issue that must be addressed. For example, dealing with medical images and extracting areas of interest that are specific to distinct body parts and diseases might not be feasible when dealing with a wide variety of images. These issues may need substantial resources, and there are several limits on their use and integration into one model.

III. IMAGECLEF MED-VQA DATASETS

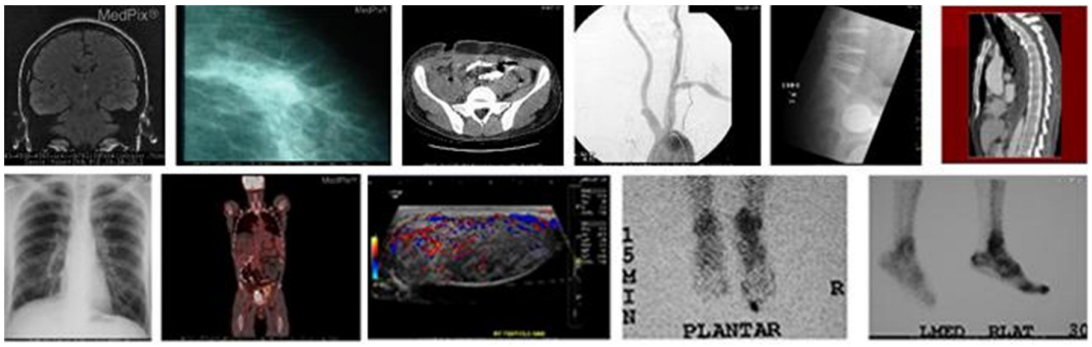

Figure 2. ImageCLEF sample images of different modalities in the training dataset. The modalities in the datasets are X-Ray, MRI, Ultrasound, CT, Angiogram, Mammogram, and PET

ImageCLEF [1] held Medical Visual Question Answering (Med-VQA) from 2018 to 2021. They introduced a different dataset every year. Some of the samples from these datasets are depicted in Figure 2. The four years datasets summary is given in Table I which shows the number of images, questions, and answers present in each year's training, validation, and test dataset. This section briefly describes the similarities and differences between them.

First publicly accessible VQA dataset in the medical domain: VQA-Med-2018 was offered in the ImageCLEF 2018 [2]. Semi-automatic methods were used to construct QA pairings. An automated question generation system used a combination of sentence reduction, response phrase recognition, and candidate question rating to create viable QA pairs. After that, all of the produced QA pairs were double-checked by two human annotators, one of whom was a clinical medicine specialist. Specifically, one step checks semantic accuracy, while the other validates the clinical relevance of the related medical imaging.

VQA-Med-2019 [3] contains modality, plane, organ system, and abnormality question types. The test set is manually checked by two medical experts whereas the training and validation datasets contain automated annotations. The abnormality category questions are more of a challenge because of the wide variety of answers than the previous three categories (modality, plane, and organ system), which may be dealt with as classification problems.

As compared to the previous years' datasets, the VQA-Med-2020 [4]’s dataset is focused on the abnormality questions. In contrast, previous datasets contained a wide variety of question categories. VQA-Med-2020 explicitly mentioned leveraging the 2019's dataset along with the 2020's dataset.

The VQA-Med-2021 in ImageCLEF 2021 [5] is developed using the same guidelines as the VQA-Med-2020, excluding the *Yes* and *No* abnormality questions. This leaves the dataset only with the abnormality questions. From the past year's participants' methods, it is clear that it is more relevant to consider this dataset as a classification task rather than sequence to sequence or text (answer) generation. Note that, the training set this year is the combination of the training and validation dataset of 2020 (4000 training + 500 validation). Only the test set is new and is manually verified by the experts.

A. ImageCLEF Med-VQA Dataset Limitations

Some of the issues with the ImageCLEF dataset are as follows

- The dataset contains seven different modalities i.e. X-Ray, MRI, ultrasound, CT, angiogram, mammography, and PET. Some of the questions do not mention the type of modality. However, the dataset is highly imbalanced. The training set contains around 4 PET samples whereas there are around 800 CT samples.

- The images in the dataset are noisy. For example, some X-Ray scans contain red background and some of them contain black background. Also, some scans are full body scans and some of them are of particular organs. Also, the orientation and plane of the samples have a wide variety. Figure 2 shows the different samples of the modalities present in the dataset.
- There are a few unique questions in the 2020-2021 datasets. Additionally, they are redundant as the top two questions hold 25% and 23.8% samples of the total questions in the training set.
- As most of the participants considered this dataset as classification and achieved better results than the participants who considered this task as a VQA task. Therefore, this dataset is more suitable as a classification task rather than VQA.

IV. METHODS

Several recent studies and participants of the ImageCLEF Med-VQA task have questioned the use of complex architectures and established the effectiveness of simpler models when trained systematically. These findings encouraged us to explore the methods and models used by the participants of the ImageCLEF Med-VQA competition to understand which methods and models perform better and how the participant handled the limited availability of the medical VQA dataset. Hence, we studied all the published papers in the ImageCLEF Med-VQA task from 2018 to 2021 to find the same. In this section, we will describe these methods in detail. Each year's ImageCLEF proceedings contain an overview paper that lays out the overall overview of the competition including the rank of each participant. We use these overview papers [2, 3, 4, 5] to collect the participants' literature to analyse the methodology they used. From the references, we were able to collect 5, 12, 6, and 7 articles from CEUR [6] 2018, 2019, 2020, and 2021, respectively, official ImageCLEF proceedings.

A. ImageCLEF Med-VQA 2018

TABLE II. IMAGECLEF MED-VQA 2018 OVERVIEW. NONE OF THE TEAMS USED ENSEMBLE MODELS. ALL MODELS USE IMAGENET PRE-TRAINED WEIGHTS FOR INITIALIZATION

Author	Considered Task	Vision Model	Text Model	Feature Fusion	Attention Type	BLEU Score	WBSS Score
UMass [8]	Multi-label classification	ResNet152	LSTM	MFB, MFH	co-attention	0.162	0.186
Zhou et al. [9]	Text generation	Inception ResNetV2	Bi-LSTM	concatenation	generic	0.135	0.174
NLM [10]	Multi-label classification	VGG16	LSTM	SAN	SAN-inbuilt	0.121	0.174
JUST [11]	Text generation	VGG16	LSTM	concatenation	-	0.061	0.12
Allaouzi et al. [12]	Multi-label classification	VGG16	LSTM	Decision Tree Classifier	-	0.054	0.10

Table II depicts the summary of 2018 articles. Three out of five participants used VGG16 for image features extraction whereas the other two used versions of ResNet architecture. All of the participants used LSTM for question features extraction. Only Zhou et al. used a Bidirectional LSTM. This is also the case between NLM and the JUST team. Both teams have used VGG16 for image features extraction and LSTM for question features extraction. However, team NLM used Stacked Attention Networks (SAN) which gave them the edge and secured third place. Once again, treating this dataset as a classification task also helped. Allaouzi et al. tried a little different approach by using a Decision Tree Classifier for feature fusion. But, this did not provide the expected results. The metrics used to evaluate the performance are Word-based Semantic Similarity (WBSS) and Bilingual Evaluation Understudy (BLEU) score. From the metrics, one can say that these metrics are used for the text generation task. However, with the classification, the answers will match exact words which, from our perspective, is an advantage over the text generation task which might miss some words. From the 2018 dataset, we can say that the ResNet architecture outperformed VGG16. The advantage of the first-place team was the co-attention mechanism with multi-modal factorized bilinear pooling (MFB) and Multi-modal Factorized High-order Pooling (MFH). We can also deduce that treating this VQA dataset as a classification problem can lead to better results.

B. ImageCLEF Med-VQA 2019

Table III depicts the overview of 2019's literature. This is the most participated in ImageCLEF's Med-VQA competition out of four. We found a total of 12 official published papers on CEUR 2019. Since this dataset is categorized, nine out of twelve participants considered this task as a classification task. The other three considered this a text generation problem. For image features, six participants used VGG (five VGG16 and one

VGG19), and four used ResNet architecture variants. Other than that, Allaouzi et al. used DenseNet121 and Liu et al used XceptionNet. For question features, the top three teams leveraged the BERT model. Seven of them used LSTM whose rank lies between fourth to twelfth. However, the JUST team did not use any textual model and Liu et al. used GRU. Six of them used features concatenation to fuse the features whereas others used MFB, MFH, and elementwise multiplication. Only five teams used attention mechanisms. Similar to 2018's winner, Zhejiang University was the only one to use MFB with a co-attention mechanism which provided them with the best results. We did not find a published paper from the second-place team. The third and fourth place teams used large ensemble models to boost their performance. Similar to 2018 VGG and ResNet architectures are the top performing architectures for image features. However, the BERT model outperformed LSTM on this dataset.

TABLE III. IMAGECLEF MED-VQA 2019 OVERVIEW. ALL MODELS USE IMAGENET PRE-TRAINED WEIGHTS FOR INITIALIZATION

Author	Considered Task	Vision Model	Text Model	Feature Fusion	Attention Type	No. of models in ensemble	BLEU Score	Accuracy
Zhejiang University [13]	Classification	VGG16 (GAP)	BERT	MFB	co-attention	-	0.644	0.624
Vu et al. [14]	Classification	ResNet152	BERT	attention	multi-glimpse attention	11	0.639	0.616
TUA1 [15]	Classification	Inception ResNetV2	BERT	concatenation	-	5	0.633	0.606
Shi et al. [16]	Classification	ResNet152	LSTM	MFH	co-attention	5	0.593	0.566
JUST [21]	Image classification	VGG16	-	-	-	1	0.591	0.57
Kornutta et al. [17]	Classification	VGG16	GloVe + LSTM	concatenation	-	1	0.582	0.558
Allaouzi et al. [20]	Text generation	DenseNet121	LSTM	concatenation	-	1	0.583	0.556
Turner et al. [18]	Classification	VGG16	LSTM	-	-	1	0.55	0.53
Bansal et al. [23]	Text generation	ResNet50	LSTM	concatenation	generic	2	0.53	0.48
Techno team [19]	Classification	VGG16	LSTM	concatenation	-	1	0.486	0.462
MIT Manipal [24]	Text generation	VGG19	LSTM (2 layers)	element-wise multiplication	-	1	0.462	-
Liu et al. [22]	Classification	XceptionNet + GAP	GRU	concatenation	generic	1	0.393	0.21

TABLE IV. IMAGECLEF MED-VQA 2020 OVERVIEW. ALL MODELS USE IMAGENET PRE-TRAINED WEIGHTS FOR INITIALIZATION

Author	Considered Task	Vision Model	Text Model	Feature Fusion	Attention Type	No. of models in ensemble	Extra Data	Additional Methods	BLEU Score	Accuracy
AIML [25]	Classification	ResNets, ResNexts, DenseNet, MobileNet	-	SSM	-	30+	2019	-	0.542	0.496
The Inception Team [26]	Classification	VGG16	-	-	-	2	-	ZCA Whitening	0.511	0.48
Jung et al. [27]	Classification	VGG16	BioBERT	MFH	co-attention	1	2019	GAP	0.502	0.466
HCP-MIC [28]	Classification	BBN-ResNeSt-50	BioBERT	-	-	1	2019	KL DIvergence	0.46	0.426
NLM [29]	Classification	ResNet-50	LSTM	Concatenation	-	1	2019	KL Divergence	0.441	0.40
Shengyan [30]	Sequence-to-Sequence	VGG16	GRU	Concatenation	generic	1	2019	-	0.412	0.376

C. ImageCLEF Med-VQA 2020

From the past two years' competition, participants anticipated that these ImageCLEF Med-VQA datasets are more suitable as a classification problem rather than a VQA task. Table IV depicts the summary of 2020's participating teams' methods. The ImageCLEF Med-VQA 2020's dataset is more inclined toward abnormality

questions only. The questions' diversity is also limited as compared to the 2019 and 2018 datasets. Consequently, five out of six 2020's participating teams considered this task as a classification task. As the questions seem redundant in the 2020's dataset, some of the teams omitted the questions' information and treated the VQA task as an image classification task. AIML, the winning team leveraged only the image models ensemble and the questions are used to categorize the images into six subtasks. They proposed a question's information extractor known as skeleton-based sentence mapping (SSM) which can categorize the images based on the questions into six sub-tasks. These sub-tasks are Fine imaging modality, coarse imaging modality, imaging plane, organ system, binary abnormality, and categorical abnormality. However, the ensemble they used contains more than 30 models with over 630.3M parameters. This model architecture is a multi-scale and multi-architecture ensemble with a wide range of input image resolutions. The ensemble consists of ResNets, ResNexts, DenseNet, and MobileNet architectures. They used 2019's dataset to extend the training dataset. Apart from this team, three teams used VGG16 and the other two used ResNet50 architecture for image features extraction. As per the text model, two teams used BioBERT whereas NLM used LSTM and Shengyan used the GRU model. The inception team used ZCA whitening only with the image model ensemble containing two models and secured second place. The winning method, MFB with co-attention, from 2018 and 2019 did not show up in 2020's competition. However, a similar method MFH with co-attention along with VGG16 and BioBERT is used by Jung et al. which secured third place in the competition. Surprisingly, the first place and second place winning teams only used questions to categorize the images into subtasks rather than using a text model to generate textual features. Both second and third-place teams used VGG16. Therefore, we can say that VGG16 performed better than ResNet architecture for image features on 2020's dataset.

TABLE V. IMAGECLEF MED-VQA 2021 OVERVIEW. ALL MODELS USE IMAGENET PRE-TRAINED WEIGHTS FOR INITIALIZATION. 2021'S TRAINING DATASET CONSISTS OF 2020'S TRAINING AND VALIDATION DATASET. ALL THE PARTICIPANTS CONSIDERED THIS TASK AS A CLASSIFICATION AND USED 2019'S DATASET TO EXTEND THE TRAINING DATASET

Author	Vision Model	Text Model	Feature Fusion	Attention Type	No. of models in ensemble	Additional Methods	BLEU Score	Accuracy
SYSU-HCP [31]	ResNet50, ResNeSt50, VGG	-	-	-	8	HGAP, label smoothing, SuperLoss, MixUp	0.416	0.382
Xiao et al. [32]	VGG16	BioBERT	MFH	co-attention	1	GAP	0.402	0.362
TeamS [33]	ResNeSt50	-	-	BBN attention	4	BBN	0.391	0.348
Lijie [34]	VGG8	BioBERT	MFH, MFB	co-attention	1	ZCA-whitening	0.352	0.316
PUC Chile [35]	DenseNet121	-	-	-	1	-	-	0.236
TAM [36]	ResNet34	LSTM	MFB	co-attention	1	-	0.255	0.222
SSN MLRG [37]	VGG16	LSTM	concatenation	-	1	-	0.227	0.196

D. ImageCLEF Med-VQA 2021

2021's ImageCLEF Med-VQA dataset is bound to abnormality questions only. Therefore, all the participants considered the Med-VQA task as a classification task. Additionally, the training dataset is just the combination of 2020's training and validation set. Only the validation and test sets are new in 2021. All the participants used 2019's abnormality subset as an extension to the training set. Table V depicts the summary of 2021's participating teams' methods. SYSU-HCP team used a multi-architecture ensemble that contains 8 models with total parameters of 422.8M. The models in the ensemble are ResNet50, ResNeSt50, VGG16, VGG19, ResNet50-HAGAP, ResNeSt50-HAGAP, VGG16-HAGAP, and VGG19-HAGAP. Where the HAGAP is Hierarchical Adaptive Global Average Pooling (GAP) (1984 dimensional latent feature vector) pooling. They omitted the questions' information due to the redundant information since all the questions are about abnormality. The training process includes *MixUp* for data augmentation, label smoothing (LS) and SuperLoss curriculum learning (SL). Apart from the first place team, three teams used VGG, two of them used ResNet, and PUC Chile used DenseNet121 architecture for image features extraction. Three out of seven participants didn't use the text model whereas two of them used BioBERT and two of them used LSTM. Most of the teams that used text models used MFH and MFB feature fusion methods along with a co-attention mechanism. Only SSN MLRG used simple feature concatenation for feature fusion.

E. Simplified method for ImageCLEF 2020-2021 datasets

MixUp is a data augmentation technique used where the dataset has a low sample size. It helps the model to be more generalized and avoids overfitting. However, it is limited to the images only and cannot use the questions' information in data augmentation. To overcome this issue, Singh et al. [7] proposed *VQAMixUp* as an extension

of the mixup for the VQA task which can incorporate the questions information in data augmentation. Using this method along with other performance booster methods, they outperformed the winning teams of 2020 and 2021 on the validation datasets as their main focus is the abnormality subset. There were four teams, two for 2020 and two for 2021, that published the results on the validation datasets. They stated that they were able to achieve these results without ensemble models and dramatically reduced model parameters. They described that the participants have used complex methods and huge ensemble models but they have achieved similar or better results just by using simple model architectures such as VGG16 for images and GRU for questions features along with a few data augmentation and training methods. Compared to the AIML team, the winner of 2020's Med-VQA. They achieved an accuracy of 60.4% on the validation dataset with a single VGG16 and GRU model architecture whereas the AIML team achieved 59.6% using an ensemble of over 30 models. The winner of 2021's Med-VQA challenge, the SYSU-HCP team, achieved 69.4% accuracy on the 2021's validation dataset with an ensemble of 8 models whereas Singh et al. used the same model architecture containing VGG16 and GRU to train the model on 2021's dataset and has achieved 70.2% accuracy. Hence, the ImageCLEF's abnormality task can be solved using the simple model with the right combination of data augmentation, model architecture, and training methods.

V. Conclusions

In this study, we explored 31 ImageCLEF papers that provided experimental results on the datasets and provided a comprehensive description of the datasets and methods. We also pointed out some limitations of the ImageCLEF datasets such as datasets being noisy and highly imbalanced. In conclusion, most of the top-performing teams used VGG and ResNet architectures for image features and BERT, LSTM, and GRU models for text features. However, 2020 and 2021's winning teams used question information as an image filtering method instead of a text model. MFH, MFB with co-attention mechanism works better than other feature fusion methods. Additionally, the simple model architectures for images and text features along with a few smart training and data augmentation methods performed similar to or better than huge ensembles and complex training techniques.

References

[1] ImageCLEF, https://www.imageclef.org, last accessed 1 Aug, 2022.

[2] S. A. Hasan, Y. Ling, O. Farri, J. Liu, H. Müller, and M. P. Lungren, "Overview of ImageCLEF 2018 medical domain visual question answering task," in: CLEF (Working Notes), 2018.

[3] A. B. Abacha, S. A. Hasan, V. Datla, J. Liu, D. Fushman, and M. Henning "VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019," Lecture Notes in Computer Science, 2019.

[4] A. B. Abacha, V. V. Datla, S. A. Hasan, D. Demner-Fushman, and H. Muller "Overview of the vqa-med task at imageclef 2020: Visual question answering and generation in the medical domain," In CLEF 2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 22-25 2020).

[5] A. B. Abacha, M. Sarrouti D. Demner-Fushman, S. A. Hasan, and H. Müller "Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain," In CLEF 2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania, 2021.

[6] CEUR Workshop Proceedings (CEUR-WS.org), "Free Open-Access Proceedings for Computer Science Workshops," http://ceur-ws.org/, last accessed 30 Jul, 2022.

[7] J. Singh, D. Mahapatra, and D. R. Bathula "Medical VQA: MixUp Helps Keep it Simple," http://dx.doi.org/10.13140/RG.2.2.29705.11362, 2022.

[8] Y. Peng , F. Liu , and M. P. Rosen "UMass at ImageCLEF Medical Visual Question Answering(Med-VQA) 2018 Task," In CLEF 2018 Working Notes, CEUR Workshop Proceedings, 2018.

[9] Y. Zhou, X. Kang, and F. Ren "Employing Inception-Resnet-v2 and Bi-LSTM for Medical Domain Visual Question Answering," In CLEF 2018 Working Notes, CEUR Workshop Proceedings, 2018.

[10] A. B. Abacha, S. Gayen, J. J. Lau, S. Rajaraman, and D. Demner-Fushman "NLM at ImageCLEF 2018 Visual Question Answering in the Medical Domain," In CLEF 2018 Working Notes, CEUR Workshop Proceedings, 2018.

[11] B. Talafha and M. Al-Ayyoub "JUST at VQA-Med: A VGG-Seq2Seq Model," In CLEF 2018 Working Notes, CEUR Workshop Proceedings, 2018.

[12] I. Allaouzi, B. Benamrou, M. Benamrou, and M. B. Ahmed "Deep Neural Networks and Decision Tree classifier for Visual Question Answering in the medical domain," In CLEF 2018 Working Notes, CEUR Workshop Proceedings, 2018.

[13] X. Yan, L. Li, C. Xie, J. Xiao, and L. Gu "Zhejiang University at ImageCLEF 2019 Visual Question Answering in the Medical Domain," In CLEF 2019 Working Notes, CEUR Workshop Proceedings, 2019.

[14] M. H. Vu, R. Sznitman, T. Nyholm, and T. Löfstedt "Ensemble of Streamlined Bilinear Visual Question Answering Models for the ImageCLEF 2019 Challenge in the Medical Domain," In CLEF 2019 Working Notes, CEUR Workshop Proceedings, 2019.
[15] Y. Zhou, X. Kang, and F. Ren "TUA1 at ImageCLEF 2019 VQA-Med: A classification and generation model based on transfer learning," In CLEF 2019 Working Notes, CEUR Workshop Proceedings, 2019.
[16] L. Shi, F. Liu, and M. P. Rosen "Deep Multimodal Learning for Medical Visual Question Answering," In CLEF 2019 Working Notes, CEUR Workshop Proceedings, 2019.
[17] T. Kornuta, D. Rajan, C. Shivade, A. Asseman, and A. S. Ozcan "Leveraging Medical Visual Question Answering with Supporting Facts," In CLEF 2019 Working Notes, CEUR Workshop Proceedings, 2019.
[18] A. Turner and A. B. Spanier "LSTM in VQA-Med, is it really needed? JCE study on the ImageCLEF 2019 dataset," In CLEF 2019 Working Notes, CEUR Workshop Proceedings, 2019.
[19] R. Bounaama and M. El A. Abderrahim "Tlemcen University at ImageCLEF 2019 Visual Question Answering Task," In CLEF 2019 Working Notes, CEUR Workshop Proceedings, 2019.
[20] I. Allaouzi, M. B. Ahmed, B. Benamrou "An Encoder-Decoder model for visual question answering in the medical domain," In CLEF 2019 Working Notes, CEUR Workshop Proceedings, 2019.
[21] A. Al-Sadi, B. Talafha, M. Al-Ayyoub, Y. Jararweh, and F. Costen "JUST at ImageCLEF 2019 Visual Question Answering in the Medical Domain," In CLEF 2019 Working Notes, CEUR Workshop Proceedings, 2019.
[22] S. Liu, X. Ou, J. Che, X. Zhou and H. Ding "An Xception-GRU Model for Visual Question Answering in the Medical Domain," In CLEF 2019 Working Notes, CEUR Workshop Proceedings, 2019.
[23] M. Bansal, T. Gadgil, R. Shah and P. Verma "Medical Visual Question Answering at Image CLEF 2019-VQA Med," In CLEF 2019 Working Notes, CEUR Workshop Proceedings, 2019.
[24] A. Thanki and K. Makkithaya "MIT Manipal at ImageCLEF 2019 Visual Question Answering in Medical Domain," In CLEF 2019 Working Notes, CEUR Workshop Proceedings, 2019.
[25] Z. Liao, Q. Wu, C. Shen, A. V. D. Henge, and J. Verjans "AIML at VQA-Med 2020: Knowledge Inference via a Skeleton-based Sentence Mapping Approach for Medical Domain Visual Question Answering," In CLEF 2020 Working Notes, CEUR Workshop Proceedings, 2020.
[26] A. Al-Sadi, H. Al-Theiabat, and M. Al-Ayyoub "The Inception Team at VQA-Med 2020: Pretrained VGG with Data Augmentation for Medical VQA and VQG," In CLEF 2020 Working Notes, CEUR Workshop Proceedings, 2020.
[27] B. Jung, L. Gu, and T. Harada "bumjun jung at VQA-Med 2020: VQA model based on feature extraction and multi-modal feature fusion," In CLEF 2020 Working Notes, CEUR Workshop Proceedings, 2020.
[28] G. Chen, H. Gong, and G. Li "HCP-MIC at VQA-Med 2020: Effective Visual Representation for Medical Visual Question Answering," In CLEF 2020 Working Notes, CEUR Workshop Proceedings, 2020.
[29] M. Sarrouti "NLM at VQA-Med 2020: Visual Question Answering and Generation in the Medical Domain," In CLEF 2020 Working Notes, CEUR Workshop Proceedings, 2020.
[30] S. Liu, H. Ding, and X. Zhou "Shengyan at VQA-Med 2020: An Encoder-Decoder Model for Medical Domain Visual Question Answering Task" In CLEF 2020 Working Notes, CEUR Workshop Proceedings, 2020.
[31] H. Gong, R. Huang, G. Chen, and G. Li "SYSU-HCP at VQA-Med 2021: A Data-centric Model with Efficient Training Methodology for Medical Visual Question Answering," In CLEF 2021 Working Notes, CEUR Workshop Proceedings, 2021.
[32] Q. Xiao, X. Zhou, Y. Xiao, and K. Zhao "Yunnan University at VQA-Med 2021: Pretrained BioBERT for Medical Domain Visual Question Answering," In CLEF 2021 Working Notes, CEUR Workshop Proceedings, 2021.
[33] S. Eslami, G. de Melo, and C. Meinel "TeamS at VQA-Med 2021: BBN-Orchestra for Long-tailed Medical Visual Question Answering," In CLEF 2021 Working Notes, CEUR Workshop Proceedings, 2021.
[34] J. Li and S. Liu "Lijie at ImageCLEFmed VQA-Med 2021: Attention Model-based Efficient Interaction between Multimodality," In CLEF 2021 Working Notes, CEUR Workshop Proceedings, 2021.
[35] R. Schilling, P. Messina, D. Parra, and H. Löbel "PUC Chile team at VQA-Med 2021: approaching VQA as a classification task via fine-tuning a pretrained CNN," In CLEF 2021 Working Notes, CEUR Workshop Proceedings, 2021.
[36] Y. Li, Z. Yang, and T. Hao "TAM at VQA-Med 2021: A Hybrid Model with Feature Extraction and Fusion for Medical Visual Question Answering," In CLEF 2021 Working Notes, CEUR Workshop Proceedings, 2021.
[37] N. M. S. Sitara and S. Kavitha "SSN MLRG at VQA-MED 2021: An Approach for VQA to Solve Abnormality Related Queries using Improved Datasets," In CLEF 2021 Working Notes, CEUR Workshop Proceedings, 2021.