Statistics and Applications {ISSN 2454-7395 (online)}
Volume 20, No. 2, 2022 (New Series), pp 279–289

Improvisation of Dataset Efficiency in Visual Question Answering Domain

Sheerin Sitara Noor Mohamed and Kavitha Srinivasan

Department of Computer Science and Engineering,
Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam – 603110, India

Received: 14 August 2021; Revised: 02 December 2021; Accepted: 02 January 2022

Abstract

The technology revolution moves the world towards automation and most of the activities are performed with minimum human intervention. The medical domain is not an exception, few developments in the medical domain helps both the patient and physician to some extent. As a part of this advancement, Visual Question Answering (VQA) in the medical domain is evolved and which helps the physician and partially visually sighted people in clinical decision making and patient education. One of the main disadvantages in achieving this advancement is data limitation problem. In this paper, two methods for handling the data limitation problem are explained and validated using appropriate pre-trained models like VGGNet and ResNet. The methods namely label smoothing and mixup are used to reduce the hard samples and augmentation of the medical data. From the performance analysis, it has been inferred that the highest accuracy and BLEU score are obtained for improved dataset as 0.297 and 0.313 for ResNet with a significant improvement of 7.9% and 5.9% respectively.

Key words: Medical VQA; Data augmentation; Label smoothing; Mixup; VGGNet; ResNet.

1. Introduction

The VQA in medical domain is an emerging field during last few years. But it has many challenges like data limitation, computation time and requires expert radiologist knowledge. Among the challenges, the data limitation issue is chosen based on the suggestions given by few researchers. Saurrouti *et al.* (2021) stated that the new training samples generation reduced the data insufficiency and avoided overfitting for VQA-RAD dataset. He *et al.* (2020) observed that the pathology images are rarely available, and the involvement of well trained pathologists in supporting dataset creation and validation are significantly minimal in real time. Hence, they addressed this issue by extracting pathology images and question-answer pairs from textbooks using semi-automated pipeline. Both the researchers stated the dataset insufficiency problems and adapted different approaches to reduce the issue to some extent. In this section, the existing methods to address the data limitation problem related to medical VQA dataset and its techniques are discussed.

Corresponding Author: Sheerin Sitara Noor Mohamed
E-mail: sheerinsitaran@ssn.edu.in

Nguyen *et al.* (2019) addressed the data limitation problem by combining the denoising auto encoder and meta-learning for large scale unlabelled data, but the compatibility between questions and the visual contents are neglected. Gong *et al.* (2021) proposed the multi-task pre-trained framework, which learns the linguistic compatibility feature set and visual content using classification and segmentation on the external dataset for data limitation problem. Chen *et al.* (2020) suggested few techniques to overcome the data limitation problem such as mixup, label smoothing and adaptive curriculum learning. According to Zhang *et al.* (2020) mixup is a simple but effective augmentation technique based on data centric efficient training. Szegedy *et al.* (2017) concluded that the label smoothing method avoids the model bias on the data by stabilizing the training progress. Bengio *et al.* (2009) stated that the noise in training set is unavoidable and it can be rectified by curriculum learning, which automatically reduces the weights of the samples with higher loss value.

One of the researcher, Chen *et al.* (2020) stated that many solutions are available to overcome data limitation problem but the easiest way is to collect more samples from the available datasets (VQA MED competition) and updating as per the user interest. ImageCLEF is one of the VQA-MED competition forums which has been conducting tasks related to medical image captioning and medical Visual Question Answering since 2018 by providing dataset as open-source. The dataset description of these tasks are given in Table 1 and the importance of each task are as follows: (i). ImageCLEF VQA-MED 2018 task concentrates on VQA dataset related to different organ, plane, modality and abnormality because very few medical VQA dataset was available during that time. (ii). In the ImageCLEF VQA-MED 2019 task, the number of samples for each category is increased to generate a better model but the abnormality type VQA samples in the dataset degrades the overall performance. (iii). ImageCLEF VQA-MED 2020 task concentrates on abnormality type queries for different organ, plane and modality (iv). In the ImageCLEF VQA-MED 2021 task, the number of classes and the equivalent abnormality type samples are increased. The dataset obtained from ImageCLEF VQA MED 2020 and 2021 tasks are augmented and used in this research work for VQA model generation.

As per the literature (Hasan *et al.*, 2018, Abacha *et al.*, 2019 and Abacha *et al.*, 2020), different techniques are used in medical VQA for visual and text feature extraction process. The visual feature extractions techniques are Convolutional Neural Network (CNN) or pre-trained models like VGGNet, ResNet or Inception – ResNet and, the text feature extraction techniques are Long Short Term Memory (LSTM), Bidirectional Long Short Term Memory (Bi-LSTM) or Bidirectional Encoder Representations from Transformers (BERT). Then the extracted features are encoded using Stacked Attention Networks (SAN), Bilinear Attention Network (BAN) or Multi-modal Factorized High-order pooling (MFH) for attention based feature fusion for training the model. One of the researcher, Aisha *et al.* (2020) proposed a VQA model for ImageCLEF VQA-MED 2020 dataset using VGG16, ResNet or DensNet where the last layer is equivalent to the number of classes of the dataset.

As of now, the data limitation problem are addressed in a few ways: (i). Extracting dataset from the textbook using semi-automated pipeline for pathology medical dataset (ii). Combining the denoising autoencoder and meta-learning approach for unlabelled large dataset (iii). By learning the visual content and linguistic compatibility in the feature set for the VQA-RAD dataset. The limitations of existing methods includes: (i). The generated dataset has limited set of queries (ii). The compatibility between the question and the visual content are neglected (iii). The cross model self-attention approach captures the long-range dependency between the question and visual content but the resulted number of samples are not increased. To overcome this specific issue, the number of samples are increased explicitly by mixup and label smoothing methods in the proposed work. Also, the modified dataset (improved) can be used directly in model creation, which in turn will reduce the overall computation time. For validating the model, different pre-defined techniques are available among which VGGNet and ResNet are significantly better than other techniques. Also, the datasets of ImageCLEF VQA-MED 2020 and 2021 are more suitable to analyse abnormality type questions as given in Table 1.

Table 1: ImageCLEF VQA – MED dataset description

ImageCLEF VQA-MED Dataset	Training Set		Validation Set		Test Set		Categories
	Images	QApairs	Images	QApairs	Images	QApairs	
Hasan *et al.* (2018)	2278	5413	324	500	264	500	Organ, plane, modality and abnormality
Abacha *et al.* (2019)	3200	12792	500	2000	500	500	Organ, plane, modality and abnormality
Abacha *et al.* (2020)	4000	4000	500	500	500	500	Abnormality
Abacha *et al.* (2021)	4500	4500	500	500	500	500	Abnormality

Source: ImageCLEF VQA-MED 2018 to 2021 tasks

The remaining sections of this paper are organised as follows. In Section 2, the dataset description and design of the proposed VQA model with dataset improvisation methods are briefly explained. In Section 3, the experimental setup and the results obtained are analysed with suitable quantitative metrics. Finally, summarized with conclusion and future work by stating the importance of dataset improvisation in medical domain.

2. Proposed Methodology

In the proposed system, the dataset is improvised by removing the hard samples and augmenting the data and, the pre-trained models are used to perform Visual Question Answering. For dataset improvisation, the methods label smoothing and mixup are used sequentially and vice-versa to improve the efficiency of the model. This can be further validated by the pre-trained models (VGGNet and ResNet) with the number of nodes in the last layer equivalent to the number of classes for the original and improved datasets. The overall system design of the proposed system is shown in Figure 1.

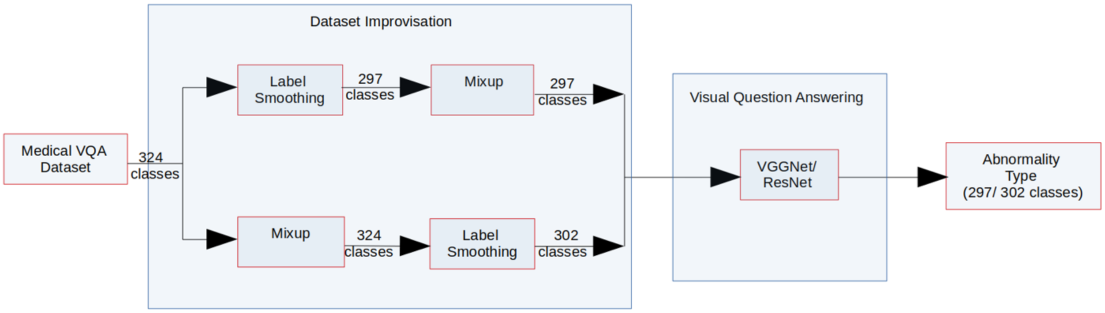

Figure 1: System design

2.1. Dataset description

The dataset used in this work is collected from ImageCLEF VQA-MED 2020 and 2021 for model generation and validation is mentioned in Table 2. The overall training set images are 5000 samples of radiology images with equivalent question-answer pairs. Similarly the validation set and test set comprises of 500 radiology images with its respective 500 question-answer pairs.

Table 2: Dataset description

ImageCLEF VQA-MED	**Training Set**		**Validation Set**		**Test Set**	
Dataset (Year)	**Images**	**QApairs**	**Images**	**QApairs**	**Images**	**QApairs**
2020	500	500	-	-	-	-
2021	4500	4500	500	500	500	500
Total	5000	5000	500	500	500	500

2.2. Label smoothing

The label smoothing methodology removes hard samples by adjusting the probability of target label is referred from (Gong *et al.*, 2021) as given in Equation (1),

$$p_a = f(x) = \begin{cases} 1-\varepsilon, & \text{if } a = b. \\ \frac{\varepsilon}{M-1}, & \text{otherwise.} \end{cases} \tag{1}$$

where ε is the small constant, M is the number of classes, and p_a denotes the probability of category a. This method is more suitable, when there are an elevated number of hard samples in the dataset and which affects the accuracy considerably. The hard samples are removed by grouping the representation of the samples from the same class into a tight cluster to improve the generalization ability. In general, the role of label smoothing in medical VQA dataset is to reduce the hard samples by adjusting the probability of target samples.

2.3. Mixup

The mixup methodology alleviates the data limitation problem by augmenting the dataset. Given the two samples (x_a and y_a) and (x_b and y_b), the new image $\hat{x}$ and $\hat{y}$ are created by linear interpolation by the equation is referred from (Gong *et al.*, 2021) as given in Equations (2) and (3),

$$\hat{x} = \lambda x_a + (1-\lambda)\, x_b \tag{2}$$

$$\hat{y} = \lambda y_a + (1-\lambda)\, y_b \tag{3}$$

$$p_a = f(x) = \begin{cases} Random(\beta(\alpha,\alpha)), & \text{if } \alpha > 0. \\ 1, & \text{otherwise.} \end{cases} \tag{4}$$

$$\beta(\alpha,\alpha) = \frac{\gamma(\alpha+\alpha))}{(\gamma(\alpha) * \gamma(\alpha))} \tag{5}$$

$$\gamma(\alpha) = (\alpha-1)! \tag{6}$$

where $\alpha\epsilon\Big[0,1\Big]$ is the shape parameter, γ is the factorial function to capture the continuous change, $\beta\epsilon\Big[0,1\Big]$ is the target probability distribution value modifier and $\lambda\epsilon\Big[0,1\Big]$ is a random value used to create new samples during the training process are given in Equations (6), (5) and (4), respectively.

The parameter α modifies the shape of the probability distribution and γ function is used to compute the range of probability distribution values. The α parameter in the beta distribution function controls the interpolation between feature-target pair using γ value. The beta distribution is chosen for two reasons, such as (i) to compute the probability distribution value from the range of alpha values (ii) the probability distribution function of beta distribution is approximately normal if $\gamma\Big(\alpha+\alpha\Big)$ is large. The role of mixup in medical VQA tasks is to augment the dataset by generating the new images from the existing images by linear interpolation.

2.4. VGGNet and ResNet

The medical VQA dataset improvised by label smoothing and mixup is given as input to the pre-trained models. In the pre-trained models, the last layer (fully connected layer) is frozen and the resultant model is used in the training process. The last layer is frozen because it is trained for the ImageNet dataset but the output dimension needs to be equivalent to the modified number of abnormality classes of the dataset to be validated. For this reason, the fully connected layer is frozen to predict the abnormality class types. The architecture of the pre-trained models, such as VGGNet and ResNet are referred from Simonyan *et al.* (2015) and He *et al.* (2016) for the implementation of proposed system.

3. Experiments and Results

In this section, the implementation requirement and experimental setup are discussed for the proposed system. Then the significance of label smoothing and mixup methods are analysed from the results of the proposed model along with hyper parameters for two datasets (original and improved).

The implementation platform (hardware) for the proposed system are: (i). Intel x64 Processor (ii). 16 GB RAM (iii). 1TB Memory (including 50 GB disk space) (iv). SSD drive to support high speed Input/Output (v). Graphics Processing Unit. The software requirements includes: (i). Ubuntu 16.04 (ii). Python 3.6 (iii). Tensorflow library. The following paragraphs explains the significance of proposed system developed with this environmental setup.

In label smoothing, the number of hard samples are removed by adjusting the probability of target label using the parameters ϵ and M. The determination of appropriate class label based on target label probability is shown in Figure 2. In this, y-axis denotes the probability of the particular class and hence it ranges from 0.0 to 1.0 and x-axis denotes the comparison between target and predicted value and it achieves its peek value when both values are comparatively equal and it varies for each samples.

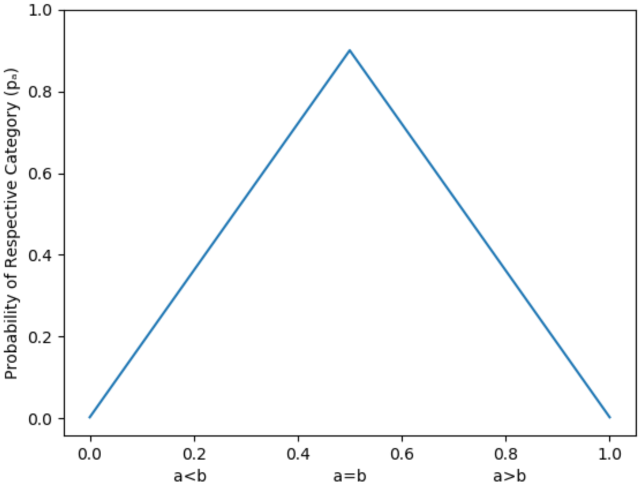

Figure 2: Appropriate class label with respect to target label probability

The hyper parameters and its values of label smoothing method is shown in Table 3 whereas common parameters are given in Table 5. Among the hyper parameters, MultiStepLR is used to modify the learning rate based on Stochastic Gradient Descent, which is updated whenever number of epochs reaches one of the two milestone (initially it starts with 0.1).

Table 3: Label smoothing - Hyper parameters for improved dataset

Hyper parameters	# value(s)
Learning rate	Starts with 0.1, update the value at 30^{th}, 60^{th} and 90^{th} epoches
Epoch	120
Pooling	1X1 (Adaptive Average Pooling)

In Mixup, the new image is generated using two images with appropriate parameters such as α, γ, β and λ. The choice of alpha value plays a significant role in linear interpolation of new image because it acts as a basic element for all required computation. The variation of beta value distribution with respect to α value is shown in Figure 3. The alpha value can be represented as $0 < \alpha \leq 1$. The value of α never be zero, because at this point beta distribution is undefined and hence the scale is 0.1 to 1.0 with an interval of 0.1 in x-axis. The resulted beta distribution ranges from 0.20 to 0.50 with an interval of 0.05 is the scale of y-axis.

The hyper parameters specific to mixup method is given in Table 4 and, the hyper-paramters and its value common to both mixup and label smoothing for improved dataset is given in Table 5. In Table 3, the learning rate is decreased by 10% after 100^{th} epoch then

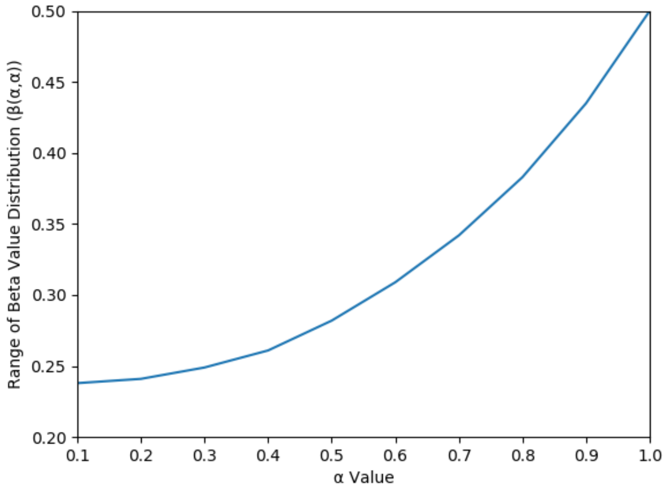

Figure 3: Range of beta value distribution with respect to α value

again 10% after 150^{th} epoch because smaller learning rate allows the model to learn more optimal set of weights but takes significantly longer time to train the model.

Table 4: Mixup - Hyper parameters for improved dataset

Hyper parameters	**# value(s)**
Learning rate	0.1, 0.01 and 0.001
Epoch	200
$WeightDecay$	0.0001
Pooling	4X4 (Average Pooling)

As a result of these two methods, the dataset comprises of 5000 VQA-MED samples with 324 classes is updated. The modifications in the number of samples and classes for label smoothing followed by mixup and vice versa are given in Table 6. In mixup followed by label smoothing method, the number of samples are augmented and then removed and hence few of the least contributing samples with higher loss values are also augmented.

The importance of improved dataset generated from label smoothing and mixup are validated using pre-trained models. The common hyperparameters used for VGGNet and ResNet for validating the model is shown in Table 7.

The results are analysed using the quantitative metrics namely accuracy and BLEU score for three cases such as, without dataset improvisation, with dataset improvisation (Label smoothing followed by mixup, mixup followed by label smoothing) as mentioned in Table 8.

From the overall results given in Tables 6 and 8 , some of the interesting inferences are: (i) The label smoothing followed by Mixup gives comparatively better results even though the number of samples are reduced (ii) Improvised dataset gives better results for both cases

Table 5: Hyper parameters common to label smoothing and Mixup for improved dataset

Hyper parameters	**# value**
WeightDecay	0.0001
Momentum	0.9
Normalization	64 (Batch Normalization)
Kernel size	3
Stride	1
Padding	1
Batch size	64
Type of Optimizer	Stochastic Gradient Descent
Type of Activation function	Rectified Linear Unit

Table 6: Improved dataset description

Label Smoothing followed by Mixup			**Mixup followed by Label Smoothing**		
Execution Sequence	**Number of Samples**	**Number of Classes**	**Execution Sequence**	**Number of Samples**	**Number of Classes**
Label Smoothing	4294	297	*Mixup*	5134	324
Mixup	4513	297	*Label Smoothing*	4700	302

(VGGNet and ResNet) (iii) For improved dataset, the overall accuracy is increased by 3.8% and 7.9% for VGGNet and ResNet respectively and (iv) For augmented dataset the accuracy and BLEU score are increased by 7.9% and 5.9% respectively for ResNet. In addition, the results of two metrics is graphically represented in Figure 4, for Datasets Vs ResNet only.

Table 7: VGGNet and ResNet - Hyper parameters

Hyper parameters	**# value**
Batchsize	128
Epoch	100
Dropout	0.2
Learning rate	0.001
Type of Optimizer	RMSPROP

In Figure 4, x-axis denotes the dataset and y-axis denotes the performance value achieved for three datasets which ranges between 0.20 and 0.32 with an interval of 0.02. The WoDI , LS_MU and MU_LS used in the graph represents Without Dataset Improvisation, Label Smoothing followed by Mixup and Mixup followed by Label Smoothing respectively. From Figure 4, it is clear that Label smoothing followed by mixup data improvisation achieved better accuracy and BLEU score using ResNet.

Table 8: Performance analysis

Performance Metrics	Label Smoothing followed by Mixup		Mixup followed by Label Smoothing		Without improvisation of dataset	
	VGGNet	ResNet	VGGNet	ResNet	VGGNet	ResNet
Accuracy	0.234	0.297	0.230	0.284	0.196	0.218
BLEU Score	0.257	0.313	0.248	0.299	0.227	0.254

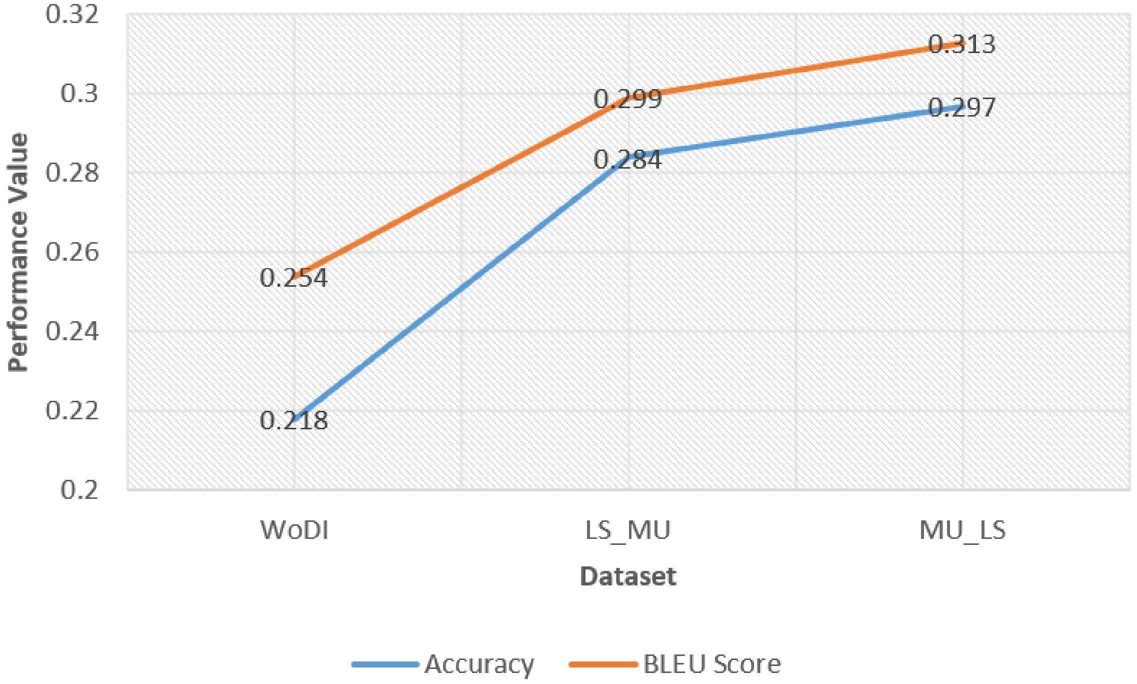

Figure 4: Dataset Vs performance analysis for ResNet

WoDI: Without Dataset Improvisation; LS_MU: Label Smoothing followed by Mixup; MU_LS: Mixup followed by Label Smoothing

4. Conclusion and Future Work

In this research, to strengthen the dataset of medical VQA, two methods namely label smoothing and mixup are chosen and its parameters are analysed and modified to improve the dataset. In label smoothing method, the hard samples are removed by adjusting the probability of target samples and mixup method augmented the new samples from the existing samples by linear interpolation. These methods improvises the efficiency of the dataset and overcomes the data limitation problem in the medical domain to some extent. The importance of dataset improvisation is validated using the pre-trained models (VGGNet and ResNet) with appropriate hyperparameters. The accuracy and BLEU score is improved by 3.8% and 3.0% for VGGNet, and 7.9% and 5.9% for ResNet respectively using the improved dataset. From the results, it has been inferred that the removal of hard samples and data augmentation improved the performance of the model significantly.

The important future direction is the creation of larger and varied dataset by increasing the number of samples in each category of medical domain with enhanced quality. Using this dataset, an improved VQA system can be developed to answer all medical queries. The VQA system development can be enhanced by selecting suitable hyper parameters to increase the efficiency and reliability of the system.

Acknowledgements

Our profound gratitude to Sri Sivasubramaniya Nadar College of Engineering, Department of CSE, for allowing us to utilize the High Performance Computing Laboratory and GPU Server for the execution of this research work successfully.

References

Abacha, A. B., Datla, V. V., Sadid A. Hasan, S. A., Demner-Fushman, D. and Muller, H. (2020). Overview of the VQA-Med Task at ImageCLEF 2020: Visual Question Answering and Generation in the Medical Domain. *CLEF 2020 Working Notes, CEUR Workshop Proceedings, Greece*, 1–9.

Abacha, A. B., Hasan, S. A., Datla, V. V., Liu, J., Demner-Fushman, D. and Miller, H. (2019). VQAMed: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019. *CLEF 2019 Working Notes, CEUR Workshop Proceedings, Switzerland*, 1–11.

Abacha, A. B., Sarrouti, M., Demner-Fushman, D., Hasan, S.A. and Muller, H. (2021). Overview of the VQA-Med Task at ImageCLEF 2021: Visual Question Answering and Generation in the Medical Domain. *CLEF 2021 Working Notes, CEUR Workshop Proceedings, Romania*, 1–8.

Bengio, Y., Louradour, J., Collobert, R. and Weston, J. (2009). Curriculum Learning. *International Conference on Machine Learning, Canada*, 41–48.

Chen, G., Gong, H. and Li, G. (2020). HCP-MIC at VQA-Med 2020: Effective Visual Representation for Medical Visual Question Answering. *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Greece*, 1–11.

Gong, H., Chen, G., Liu, S., Yu, Y. and Li, G. (2021). Cross-Modal Self-Attention with Multi-task Pretraining for Medical Visual Question Answering. *ACM International Conference on Multimedia Retrieval (ICMR), Taiwan*, 456–460.

Gong, H., Huang, R., Chen, G. and Li, G. (2021). SYSU-HCP at VQA-Med 2021: A Data-Centric Model with Efficient Training Methodology for Medical Visual Question Answering. *CLEF 2021 Working Notes, CEUR Workshop Proceedings, Romania*, 1–11.

Hasan, S. A., Ling, Y., Farri, O., Liu, J., Miller, H. and Lungren, M. (2018). Overview of ImageCLEF 2018 Medical Domain Visual Question Answering Task. *CLEF 2018 Working Notes, CEUR Workshop Proceedings, Switzerland*, 1–8.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, 770 –778.

He, X., Zhang, Y., Mou, L., Xing, E. and Xie, P. (2020). PathVQA: 3000+ Questions for medical question answering. *arXiv preprint arXiv:2003.10286*, 1–12.

Muller, R., Kornblith, S. and Hinton, G. (2020). What does label smoothing help?. *Advances in Neural Information Processing System*, **32**, 1–10.

Nguyen, B. D., Do, T. T., Nguyen, B. X., Do, T., Tjiputra, E. and Tran, Q. D. (2019). Overcoming Data Limitation in Medical Visual Question Answering. *International Conference on Medical Image Computing and Computer-Assisted Intervention, Turkey*, 522–530.

Sarrouti, M., Abacha, A. B. and Demner-Fushman, D. (2021). Goal-driven visual question generation from radiology images. *Information*, **12(8)**, 1–16.

Sheerin, S. N. M. and Kavitha, S. (2020). ImageCLEF 2020: An approach for Visual Question Answering using VGG-LSTM for different datasets. *CLEF 2020 Working Notes, CEUR Workshop Proceedings, Greece*, 1–10.

Sheerin, S. N. M. and Kavitha, S. (2021). SSN MLRG at VQA-MED 2021: An Approach for VQA to Solve Abnormality Related Queries using Improved Datasets. *CLEF 2021 Working Notes, CEUR Workshop Proceedings, Romania*, 1–10.

Simonyan, K. and Zisserman, A. (2015). Very Deep Convolutional Networks for Large-scale Image Recognition. *International Conference on Learning Expectation, Canada*, 1–14.

Szegedy, C., Ioffe, S., Vanhoucke, V. and Alemi, A. A. (2017). Inception-v4, Inception-Resnet and the Impact of Residual Connections on Learning. *AAAI Conference on Artificial Intelligence, California*, 4278–4284.

Zhang, H., Cisse, M., Dauphin, Y. N. and Lopez-Paz, D. (2018). Mixup: Beyond Empirical Risk Minimization. *International Conference on Learning Representations, Canada*, 1–13

Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Zhang, Z., Lin, H., Sun, Y., He, T., Mueller, J., Manmatha, R., Li, M. and Smola, A. J. (2020). Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*,1–12.