ImageCLEF 2020: An approach for Visual Question Answering using VGG-LSTM for different datasets

Sheerin Sitara Noor Mohamed [0000-0003-1752-2107], Kavitha Srinivasan [0000-0003-3439-2383]

Department of CSE, SSN College of Engineering, Kalavakkam – 603110, India
sheerinsitaran@ssn.edu.in, kavithas@ssn.edu.in

Abstract. The recent advancement and digitalization in the medical domain requires an image based question answering system to support clinical decisions. This system also helps the patients to know about their present conditions rapidly with more information. As an effort to promote the development, ImageCLEF 2020 organizes third edition of the Visual Question Answering (VQA) Task. In this task, the abnormality related questions are to be answered for the given set of radiology images. In the proposed system, VGGNet based on transfer learning approach and LSTM are used to extract the image and text feature vectors respectively in the encoder stage. Then, both feature vectors are combined and given as input to the decoder for predicting the answer. The purpose of selecting VGGNet and LSTM are: (i). VGGNet is able to extract medical image features effectively in small dataset (ii). LSTM is capable to accommodate significant information of the text. Moreover, the proposed model is evaluated for three datasets namely original dataset (4500 samples), reduced dataset (4348 samples) and augmented reduced dataset (4626 samples). The proposed model resulted in an accuracy of 0.282 and a BLEU score of 0.330 for augmented reduced dataset, which is ranked ninth among all participating group in ImageCLEF 2020 VQA-MED task.

Keywords. VQA; VGGNet; LSTM; medical domain; augmented dataset; reduced dataset; ImageCLEF

1 Introduction

The amount of data generated and used in this era are increasing exponentially and medical domain is not an exception. Also, everyone wants to know the answer for everything they come across in the internet world. Multiple search engines are working towards satisfying their knowledge thirst, unfortunately very few image based search engines are available in the market. However, these search engines are generalized and not suitable for medical domain.

The medical domain is wide, and it needs prior knowledge and analysis to answer the questions. These issues can be addressed by improving the medical image based question answering system. To enhance this research further, ImageCLEF is conducting VQA task in medical domain since 2018 [1].

The Visual Question Answering (VQA) in medical domain helps people (especially partially sighted) in better understanding of their condition and supports clinical decision. The challenges of VQA in medical domain includes: (i). Parameter selection and feature extraction for medical dataset which differs from the real time and abstract dataset (ii). Specific VQA model which works for all medical category is in developing stage. For example in [2], different approaches are required to answer different medical questions. The pre-trained model followed by BERT model answers organ, plane and modality related questions whereas abnormality related questions are answered effectively by sequence-to-sequence model (iii). Single optimal model which detects all types of medical abnormalities in different region needs some attention and effort. But, abnormality detection with respect to the particular region are available. For example, in [3], bifurcated structure detects four gastrointestinal abnormalities and three dermoscopic lesions in WCE images and PH2 dataset respectively. Two abnormality categories are detected separately and attained an accuracy of 97.8% and 97.5% respectively. (iv). Memory and time constraints.

The remaining part of the paper spans across following subsections. In Sect. 2, literature survey related to automation in medical domain, inference from VQA task for real world dataset and its recent advancement in medical domain are discussed. Sect. 3 gives brief description about the ImageCLEF VQA-Med 2020 dataset and two other proposed datasets used for analysis and validation. In Sect. 4, the design of the proposed VQA model based on inference attained and its implementation are explained. A brief summary about the result and the respective evaluation of all five runs are given in Sect. 5 and conclusion is given at the end.

2 Related Works

The recent studies shows a tremendous advancement in the medical domain. One of the best advancement is that the medical data in structured, semi-structured and unstructured formats are digitized. From the last decades, Artificial intelligence (AI) utilizes the digitization advancement and enhances an automation in the medical domain. In [4], natural language text (medical history, physical examination result, result of X-ray, ultrasound or ECG) are collected, analysed and used to find the dependency between features to improve the healthcare quality in multidisciplinary paediatric centre using deep linguistic techniques. The advantage of digitization is also applicable for medical imaging applications like image classification [5], caption generation [6] and computing severity level [7]. The reliability of these applications are based on the features extracted from the images. At present, the pre-trained models like Convolutional Neural Network (CNN) or pre-trained models like VGGNet or ResNet are playing a vital role in feature extraction for VQA related applications. VQA on medical domain emerged based on the knowledge obtained from real world datasets like MSCOCO dataset, DAQUAR, VQA Dataset, FM-IQA and Visual7W. The inferences are (i). The detailed understanding of the image and complex reasoning are required to answer the visual questions because it selectively targets background details and/or underlying context [8]. (ii). Questions are arbitrary and it imposes many sub-problems in computer vision like object location, detection and/or counting [9]. (iii) Improvement in rare question type has negligible impact on overall performance [10]. (iv). Least contributing question types need to be victimized because it pulls down the overall performance [10]. (v). Appropriate parameter selection (activation function, large mini-batches, smart shuffling of training data and word embedding by Glove, google images, etc.,) has its own impact in performance of the model.

Table 1. Brief description of ImageCLEF MED-VQA task for last three years

Dataset	**Training set**		**Validation set**		**Test set**		**Category**	**Performance analysis**			
	Image	**QA pairs**	**Image**	**QA pairs**	**Image**	**QA pairs**		**Accuracy**	**BLEU**	**CBSS**	**WBSS**
[11]	2278	5413	324	500	264	500	Organ, plane, modality and abnormality	-	0.162	0.186	0.338
[12]	3200	12792	500	2000	500	500	Organ, plane, modality and abnormality	0.644	0.624	-	-
[13]	4000	4000	500	500	500	500	Abnormality	0.496	0.542	-	-

From 2018, ImageCLEF is conducting VQA task in medical domain. The VQA-Med 2018 and VQA-Med 2019 dataset contains organ, plane, modality and abnormality related visual question answer pairs. In these tasks, most of the researcher applied pre-trained models like VGGNet, ResNet, etc., to encode medical images and Recurrent

Neural Networks (RNN) to generate question encodings. Some of the researchers applied attention based mechanism to extract relevant image features to answer the questions. The highest BLEU, WBSS and CBSS scores obtained in 2018 tasks are 0.162, 0.186 and 0.338 respectively [11]. In 2019, along with the above approaches, different pooling strategies and transformer-based approaches are also used and attained a highest accuracy and BLEU score as 0.644 and 0.624 respectively [12]. The overall summary of the ImageCLEF VQA tasks are tabulated in Table 1.

From the overall inference, VGGNet and LSTM are selected for the implementation of given task on VQA 2020. The advantages of selecting VGGNet [14] for image feature extraction includes: (i). Built on ImageNet dataset but works for other datasets and tasks. (ii). Outperforms the complex recognition tasks involving less detailed images. (iii). Addresses the vanishing gradient and exploding gradient problem. (iv). Illustrates the importance of deepest model in visual representation.

The advantages of using LSTM [15] are (i). Developed for TIMIT dataset, but it can solve any complex sequence learning problem in handwriting recognition, speech recognition, polyphonic music modelling, etc., (ii). The role of hyper parameters with respect to performance in LSTM structure includes: (a). Coupling the inputs, removing forget gates simplifies LSTM structure, reduces the number of parameters and computational cost, without significantly decreasing performance. (b). Gaussian noise is moderately helpful for TIMIT dataset but it is harmful for other datasets. (c). Highest measured interaction between hyper parameters are quite small.

3 Dataset Description

In this section, three medical VQA dataset are discussed along with its description. The three datasets are: ImageCLEF VQA-Med 2020 dataset (Original Dataset (OD)) and two other datasets used with modification (Reduced Dataset (RD) and Augmented Reduced Dataset (ARD)). In Original dataset, (ImageCLEF VQA-Med 2020 dataset), the dataset is divided into three subsets namely training set, validation set and test set as 4000, 500 and 500 with equivalent number of question answer pairs. In addition the dataset consists of abnormality related visual questions for different organs (e.g. lung, skull, spine, gastrointestinal, musculoskeletal), planes (e.g. axial, sagittal and corona) and modalities (e.g. CT, X-ray, MRI). For better learning, the training set and validation set (as a total 4500 samples) are used for training.

The Reduced Dataset (RD) consists of 4348 samples (from training and validation set) for training and 500 samples for testing. The reduced dataset is generated by two ways namely (i). Eliminate the least contributing samples (ii). Identify and reduce the number of samples of similar class, when the count deviates much from the remaining classes. These samples degrade the overall performance of the system and hence both approaches are applied.

The Augmented Reduced Dataset (ARD) consists of 4626 samples (from training and validation set) for training and 500 samples for testing. The dataset is augmented by collecting samples from VQA-Med 2018 and 2019. The collected samples are merged with RD to generate Augmented Reduced Dataset. Augmenting the training set improves the learning rate and as a result generates better model. The OD, RD and ARD contains 330, 316, 316 classes respectively.

4 System Design

In this VQA task, the VGGNet and LSTM techniques are used to answer the medical visual questions. The system design of the proposed model is shown in Fig. 1. In this, the feature information from the medical image and its question-answer pairs are extracted and concatenated by encoder. Then, the concatenated feature vector is decoded by timestamp to generate the answer, with post-processing at the end. The proposed model consists of five modules namely, (i). Pre-processing (ii). Encoder (iii). Decoder, (iv). Post-processing and (v). Answer prediction.

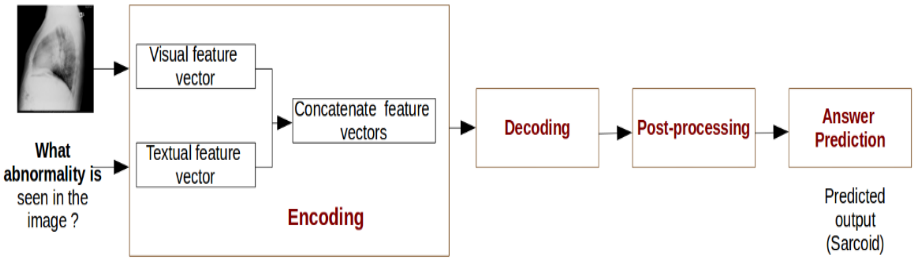
What abnormality is seen in the image ?
Visual feature vector
Textual feature vector
Concatenate feature vectors
Encoding
Decoding
Post-processing
Answer Prediction
Predicted output (Sarcoid)

Fig. 1. System design

4.1 Pre-processing

In the pre-processing stage, the input samples are converted to required format for effective image and text processing. As a first step, the images are reshaped to (229, 229) dimension (preferable input size of VGG16/VGG19 network). In text processing, comma is the best separator and hence the question-answer pairs are converted to comma separated file. The already existing comma within the field are converted to related special symbol (here we used semicolon). Otherwise, these commas within the field are encountered as separator, and end up with an imbalanced fields.

4.2 Encoder

Encoder transforms the feature vectors into the required format for the model to answer the questions. This transformation is required because the type and dimension of features extracted from image and its respective text are different. Hence a dimensionality mapping is required to bridge the gaps and then the features vectors are concatenated. To perform this, encoder has three sub-modules namely (i). Image processing (ii). Text processing (iii). Concatenation. The system architecture of the encoder is given in Fig.2

shows the each encoding stages along with the size of feature vector before and after concatenation.

Image Processing.

In the proposed system, the image features are extracted by VGG16/VGG19. The last layer of the VGGNet is frozen and the resulted model is used for image feature extraction as transfer learning approach. The last layer is frozen because VGGNet is trained for ImageNet dataset (1000 classes) but we required the output dimension to be 1024. For this reason, after the last before layer, the dense and fully connected layers are used to adjust the dimension of the image feature vector.

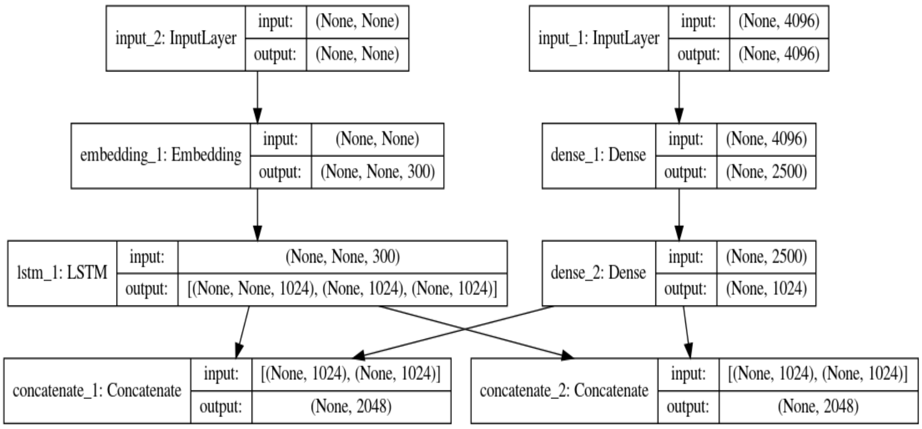

Fig. 2. Encoder

Text Processing.

Text processing, computes the dependency between the words and derives the information from the sequence of input words. The LSTM (an advanced type of RNN) is used to generate the text feature vector. The input text is tokenized into individual words and the minimum and maximum length of question and answer are computed. The LSTM computes the question embedding (using the Glove vector), timestamp by timestamp for the respective samples. This vector is given to the fully connected layer to project it to the same dimensional shape as image feature vector.

Concatenation.

The computed feature vectors (image and text feature vector) are combined using element wise multiplication and are later used by decoder for model creation.

4.3 Decoder

Both visual and textual features are merged into three dimensional vector (2048-dimensional space) which is a sequence of vectors. As both the image and textual features are represented as sequence of vectors (not as single vector), LSTM is required to feed the concatenated vector to the softmax layer. The system architecture of this sub-module, decoder is shown in Fig.3.

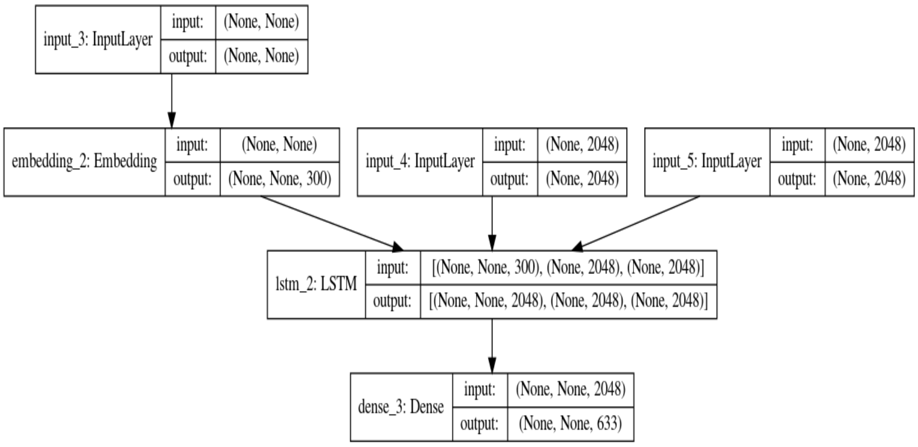

Fig. 3. Decoder

4.4 Post-processing

In post-processing, the generated answer needs to be converted to the required format as in training set. In this, semicolon in the generated answers are converted back to comma format.

4.5 Answer Prediction

In this stage, encoder-decoder model based on VGG-LSTM is generated. The answer for the test set can be predicted by the model. Further, the result can be analysed and evaluated using performance metrics like accuracy and BLEU score.

5 Experiments and Results

The proposed model is executed on three datasets (as discussed in Section 3) and analysed using five different combination of techniques, such as: (i). VGG16 (excluding last layer) followed by LSTM for original dataset (ii). Same as (i) for reduced dataset (iii). VGG16 (excluding last layer) followed by LSTM and post-processing at the end for Augmented Reduced Dataset (iv). Same as (iii), but the pre-trained model is VGG19 (v). Similar to first combination but post-processing is included at the end. From the results it is inferred that proposed model with post-processing is included at the end for Augmented Reduced Dataset gives better performance than the other combinations. In Table 2, OD, RD and ARD represents Original dataset, Reduced Dataset and Augmented Reduced Dataset respectively.

Table 2. Brief description about each run

Run number	Dataset	Techniques	Accuracy	BLEU score
1	OD	VGG16 and LSTM	0.274	0.321
2	ARD	VGG16 and LSTM	0.268	0.320
3	ARD	VGG16 and LSTM (Post processing)	**0.282**	**0.330**
4	RD	VGG19 and LSTM (Post processing)	0.248	0.292
5	OD	VGG16 and LSTM (Post processing)	0.276	0.323

The performance of the model depends on appropriate parameter selection also. In this model, RMSPROP optimizer is used with a learning rate of 0.001 and the batch size, epoch and dropout are set to 256, 400 and 0.2 respectively. For training the model using these hyper parameters, each run took approximately 180 minutes in GPU. Among the five runs, third run achieved a better accuracy score of 0.282 and the BLEU score of 0.330. The final result of the leaderboard is given in Table 3 where our team achieved 9th place in the listed ranks.

Table 3. Top 10 ranking of ImageCLEF 2020 VQA-MED

Rank	Team name	Accuracy	BLEU score	No. of runs submitted
1	z_liao	0.496	0.542	5
2	TheInceptionTeam	0.480	0.511	5
3	bumjun_jung	0.466	0.502	5
4	going	0.426	0.462	5
5	NLM	0.400	0.441	5
6	harendrakv	0.378	0.439	7
7	shengyan	0.376	0.412	5
8	kdevqa	0.314	0.350	4
9	**sheerin**	**0.282**	**0.330**	**5**
10	umassmednlp	0.220	0.340	4

6 Conclusion and Future Work

In this paper, an approach for Visual Question Answering (VQA) on medical domain is implemented for ImageCLEF VQA-Med 2020 dataset and further analysed using two different types of proposed datasets namely: Reduced Dataset (RD) and Augmented Reduced Dataset (ARD). The proposed model has five stages namely: (i). Pre-processing (ii). Encoding (iii). Decoding (iv). Post-processing and (v). Answer prediction. In pre-processing, the dataset has been converted to the specific input format as required for VGGNet and LSTM. Then the image and text features are extracted and concatenated. The concatenated feature vector is decoded for next level. In post-processing, the answer is converted to the format as in the training dataset. Finally, the generated model predicts the answer for the test set. Among the five runs of the proposed model the better result is achieved for augmented reduced dataset with an accuracy score of 0.282 and BLEU score of 0.330.

In medical VQA domain, large amount of information needs to be extracted and hence it has more memory constraint. This can be addressed with the help of GPU and the selection of optimal hyper parameters. In future, the proposed VQA model can be improvised by developing a design of Convolutional Neural Network (CNN) for medical images and fixing the appropriated hyper parameters with visualization of layers. In addition, the advanced text processing approach like BERT, which represent each sentence in 768-d question feature vector can be included.

7 Acknowledgement

Our profound gratitude to SSN College of Engineering, Department of CSE, for allowing us to utilize the High Performance Computing Laboratory and GPU Server for the execution of this challenge successfully.

References

1. Ionescu, B., Muller, H.,Peteri, R., Ben Abacha, A., Datla, V., Hasan, S. A., Demner-Fushman, D., Kozlovski, S., Liauchuk, V., Cid, Y.D., Kovalev, V., Pelka, O., Friedrich, C.M., Herrera, A. G. S. D., Ninh, V., Le, T., Zhou, l., Piras, l., Riegler, M., Halvorsen, P., Tran, M., Lux, M., Gurrin, C., Dang-Nguyen, D., Chamberlain, J., Clark, A., Campello, A., Fichou, D., Berari, R., Brie, P.,Dogariu, M., Stefan, L.D., Constantin, M. G.: Overview of the ImageCLEF 2020: Multimedia Retrieval in Lifelogging, Medical, Nature and Internet Applications. In: Experimental IR Meets Multilinguality, Multimodality and Interaction, Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020), Greece, September 22-25. LNCS Lecture Notes in Computer Science, Springer (2020).
2. Zhou, Y., Kang, X., Ren, F.: TUA1 a ImageCLEF 2019 VQA-Med: A Classification and Generation Model based on Transfer Learning. In: CLEF 2019 Working Notes. CEUR Workshop Proceedings, Switzerland (2019).
3. Baranov, A.A., Namazova-Baranova, L.S., Smirnov, I.V., Devyatkin, D.A., Shelmanov, A.O., Vishneva, E.A., Antonova, E.V., Smirnov, V.I.: Technologies for Complex Intelligent

Clinical Data Analysis. In: Annals of the Russian Academy of Medical Sciences, 71(2), pp. 160 - 171 (2016).

4. Hajabdollahi, M., Esfandiarpoor, R., Sabeti, E., Karimi, N., Soroushmehr, S.M.R., Samavi, S.: Multiple Abnormality Detection for Automatic Medical Image Diagnosis using Bifurcated Convolutional Neural Network. In: Biomedical Signal Processing and Control, 57, pp.101792 - 101802 (2020)
5. Liauchuk, V., Tarasau, A., Snezhko, E., Kovalev, V.: ImageCLEF 2018: Lesion-based TB-Descriptor for CT Image Analysis. In: CLEF 2018 Working Notes, CEUR Workshop Proceedings, Belarus (2018).
6. Herrera, A.G.S.D., Eickhoff, C., Andrearczyk, V., Miller, H.: Overview of the ImageCLEF 2018 Caption Prediction Tasks. In: CLEF 2018 Working Notes, CEUR Workshop Proceedings, China (2018).
7. Kavitha, S., Nandhinee, P.R., Harshana, S., Srividya, J.S., Harrinei, K.: ImageCLEF 2019: A 2D Convolutional Neural Network Approach for Severity Scoring of Lung Tuberculosis using CT Images. In: CLEF 2019 Working Notes. CEUR Workshop Proceedings, Switzerland (2019).
8. Antol, S., Agrawal. A., Lu, J., Antol, S., Mitchell, M., Zitnick, L., Batra, D., Parikh, D.: VQA: Visual Question Answering. In: International Conference on Computer Vision, pp. 2425 - 2433 (2015).
9. Kafle. K., Kanan, C.: Visual Question Answering: Datasets, Algorithms and Future Challenges. In: Computer Vision and Image Understanding, 163, pp.3 - 20 (2016).
10. Teney, D., Anderson, P., He, X., Hengel, A.V.D.: Tips and tricks for Visual Question Answering: Learning from the 2017 Challenge. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4223 - 4232 (2018).
11. Hasan, S.A., Ling, Y., Farri, O., Liu, J., Miller, H, Lungren, M.: Overview of ImageCLEF 2018 Medical Domain Visual Question Answering Task. In: CLEF 2018 Working Notes, *CEUR* Workshop Proceedings, Switzerland *(*2018).
12. Ben Abacha, A., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Miller, H.: VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019. In: CLEF 2019 Working Notes. CEUR Workshop Proceedings, Switzerland (2019).
13. Ben Abacha, A., Datla, V.V., Sadid A. Hasan, S.A., Demner-Fushman, D., Muller, H.: 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, Greece (2020).
14. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. In: International Conference on Learning Representations, Canada, pp. 1 - 14 (2014).
15. Greff. K., Srivastava, R.K., Koutnik, J., Steunebrink. B.R., Schmidhuber. J.: LSTM: A Search Space Odyssey. In: IEEE Transcations on Neural Networks and Learning Systems, 28(10), pp. 2222 - 2232 (2017).
16. Nguyen, .D., Do, T.T, Nguyen, B.X., Do, T., Tjiputra, E., Tran, Q.D.: Overcoming Data Limitation in Medical Visual Question Answering. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp.522 - 530 (2019)