# SSN MLRG at VQA-MED 2021: An Approach for VQA to Solve Abnormality Related Queries using Improved Datasets

Noor Mohamed Sheerin Sitara [1] and Srinivasan Kavitha[2]

[1, 2] *Department of CSE, Sri Sivasubramaniya Nadar College of Engineering, Kalavakkam – 603110, India*

**Abstract**

The Visual Question Answering (VQA) in the medical domain attains tremendous advancement in last few years. To improvise the VQA research, ImageCLEF forum is organizing the fourth edition of VQA task in medical domain. This year, the abnormality related VQA queries are to be answered for the given set of radiology images. In the proposed system, VGGNet based on transfer learning approach and LSTM is used to extract image and text features respectively. The extracted three dimensional (embedding, image, text) feature vectors are concatenated into sequence of vectors by LSTM for predicting the answer. The purpose of selecting VGGNet and LSTM are: VGGNet, outperforms complex recognition tasks and also addresses vanishing gradient and exploding gradient problem and LSTM, solves complex sequence learning problems and overcomes long term dependency problems. In addition, the hyper parameters are chosen appropriately and four improved datasets are used to analyze the performance of the proposed model. These four datasets are build by collecting the samples from previous ImageCLEF VQA – MED tasks. The proposed model resulted in an accuracy of 0.196 and a BLEU score of 0.227 for one of the dataset, which is ranked tenth among all participating groups in ImageCLEF 2021 VQA-MED task.

**Keywords**

Visual Question Answering; VGGNet; Long Short Term Memory; medical domain; VQA dataset; augmented dataset; reduced dataset; ImageCLEF

## 1. Introduction

The recent studies of 2020 reveals that the 90% of data are unlabelled and 40 – 50 % of data is in the form of images [12]. Hence an Artificial Intelligent (AI) approach is required to analyze both image and text. Now-a-days, the advantage of AI approach is extended to different applications like text summarization, machine translation, sentiment analysis, image captioning, and Visual Question Answering (VQA). Among which, VQA comprises both image and text for real world dataset [1], abstract dataset [2] and medical dataset [3] are evolved in this decade. For medical dataset, ImageCLEF organizes medical related image captioning and VQA task since 2018 [3]. From 2020, ImageCLEF concentrates on solving abnormality related VQA questions [4].

The Visual Question Answering system of medical domain takes one or more abnormality related natural language questions with respective radiology images as input and predicts the appropriate answer as output. Some of the applications of medical VQA are: (i). Helps partially visually sighted people (ii). Helps in clinical support and decision. To answer the medical VQA queries, the visual information of the radiology image is extracted based on the significant textual content of the question. In other words, image features are extracted based on the text features and finally both feature vectors are concatenated to answer the respective questions. The different image processing techniques are, Convolutional Neural Network, pre-trained models like VGGNet, ResNet and DenseNet and, text

[1] CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania
EMAIL: sheerinsitaran@ssn.edu.in (A. 1); kavithas@ssn.edu.in (A. 2)
ORCID: 0000-0003-1752-2107 (A. 1); 0000-0003-3439-2383 (A. 2)

CEUR Workshop Proceedings (CEUR-WS.org)

processing techniques are Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU), Bidirectional Encoder Representation (BERT).

The overview of ImageCLEF VQA – MED tasks (2018, 2019 and 2020) are summarized and given in Table 1. From the results, the observations are: (i). ImageCLEF VQA – MED 2019 achieved better performance than ImageCLEF VQA – MED 2018, because of the increased number of samples for each class (ii). In ImageCLEF VQA – MED 2019 task, abnormality type VQA questions achieved less performance as compared with organ, plane and modality type questions (iii). Based on the ImageCLEF VQA – MED 2019 task outcome, ImageCLEF begin to concentrates on abnormality type questions since 2020 but the performance is reduced.

From the inference of previous tasks (especially ImageCLEF VQA – MED 2020 task) as tabulated in Table 1, VGGNet and LSTM (modification of RNN) are used in the proposed model for VQA system development. In addition, VGGNet and LSTM have some advantages, such as (i). VGGNet - Outperforms complex recognition tasks, addresses vanishing gradient and exploding gradient problem [5] (ii). LSTM – Solves complex sequence learning problems and overcomes long term dependency problems [6].

**Table 1**
ImageCLEF VQA – MED task Overview

| Task | Widespread Techniques | | Remarkable Techniques | | Category | Remarkable Accuracy | Remarkable BLEU score |
|---|---|---|---|---|---|---|---|
| | Images | Texts | Images | Texts | | | |
| **ImageCLEF VQA – MED 2018 [7]** | Convolutional Neural Network (CNN) | Recurrent Neural Network (RNN) | ResNet | LSTM | Organ, plane, modality and abnormality | - | 0.162 |
| **ImageCLEF VQA – MED 2019 [8]** | VGGNet or ResNet | Bidirectional Encoder Representation (BERT) or RNN | CNN | BERT | Organ, plane, modality and abnormality | 0.624 | 0.644 |
| **ImageCLEF VQA – MED 2020 [9]** | CNN, VGGNet or ResNet | BERT or modification of RNN | DenseNet and ResNet | Skeleton based Sentence Mapping | Abnormality | 0.496 | 0.542 |

The research contributions of ImageCLEF VQA – MED 2021 task using the proposed model are: (i). For training the model, the dataset is augmented from ImageCLEF VQA-MED 2018, 2019 and 2020 (test set) datasets. From 2018 and 2019 datasets, 126 samples associated with abnormality related queries are collected and augmented. The ImageCLEF VQA-MED 2020 test set consists of 500 radiology images with respective 500 question-answer pairs are also used for augmenting the dataset. (ii). In terms of implementation, VGGNet followed by LSTM are used for answering the medical questions related to radiology images. (iii). For building the model, the hyper parameters like learning rate, number of epochs, batch size, momentum, dropout, etc., are selected and the values are fixed based on the performance measures.

The remaining part of the paper spans across following subsections. In Sect. 2, ImageCLEF VQA-MED 2021 task and its dataset are discussed and, compared with 2020 task. In Sect. 3, the design of the proposed VQA model and its implementation are explained. A brief summary about the results obtained and the performance evaluation are given in Sect. 4 with a conclusion at the end.

## 2. Task and Dataset Description

In this section, ImageCLEF VQA – MED 2021 task and given dataset are discussed with three types of improved datasets, which are build from the previous VQA datasets.

### 2.1. ImageCLEF VQA – MED 2021 task

ImageCLEF, a part of Conference and Labs of the Evaluation Forum is conducting tasks related to the medical domain since 2018. ImageCLEF VQA – MED 2021 task concentrates on abnormality type questions for different organs, planes and modalities. In this task, 33 participants were registered and 13 teams were participated with 75 successful runs.

### 2.2. ImageCLEF VQA – MED 2021 dataset

The ImageCLEF VQA-MED 2021 dataset [10] is given as four subsets namely, training set, validation set, new validation set and test set. The first two subsets are equivalent to ImageCLEF VQA-MED 2020 dataset and it is used for training. This set consists of 4500 radiology images and 4500 question-answer pairs, among which the validation set consists of 500 radiology images with respective 500 question-answer pairs. The new validation set consists of 500 question-answer pairs associated with 500 radiology images. Finally, the test set includes 500 radiology images and 500 questions about abnormality.

The datasets used for training the proposed model is given in Table 2. The acronyms, GD, GTD, AD and ARD represents Given Dataset, Given dataset along with Test dataset from ImageCLEF VQA-MED 2020, Augmented Dataset and Augmented Reduced Dataset. The Augmented Dataset consists of GTD along with the augmented samples from ImageCLEF VQA-MED 2018 and 2019. The Augmented Reduced Dataset is a modification of AD dataset, in which some of the samples are removed by two ways, (i). Least contributing samples, (ii). Identify and reduce the number of samples of similar cases where the count value deviates from the remaining classes.

**Table 2**
Dataset Description

| Datasets | Training Set | | Classes | Description |
|---|---|---|---|---|
| | Images | QA pairs | | |
| GD | 4500 | 4500 | 330 | Different abnormality related medical images along with associated question answer pairs |
| GTD | 5000 | 5000 | 366 | |
| AD | 5126 | 5126 | 366 | |
| ARD | 4848 | 4848 | 352 | |

The advancement in ImageCLEF VQA-MED 2021 task when compared to 2020 task are: (i). The number of VQA samples are increased (ii). Number of classes of abnormality type questions are increased

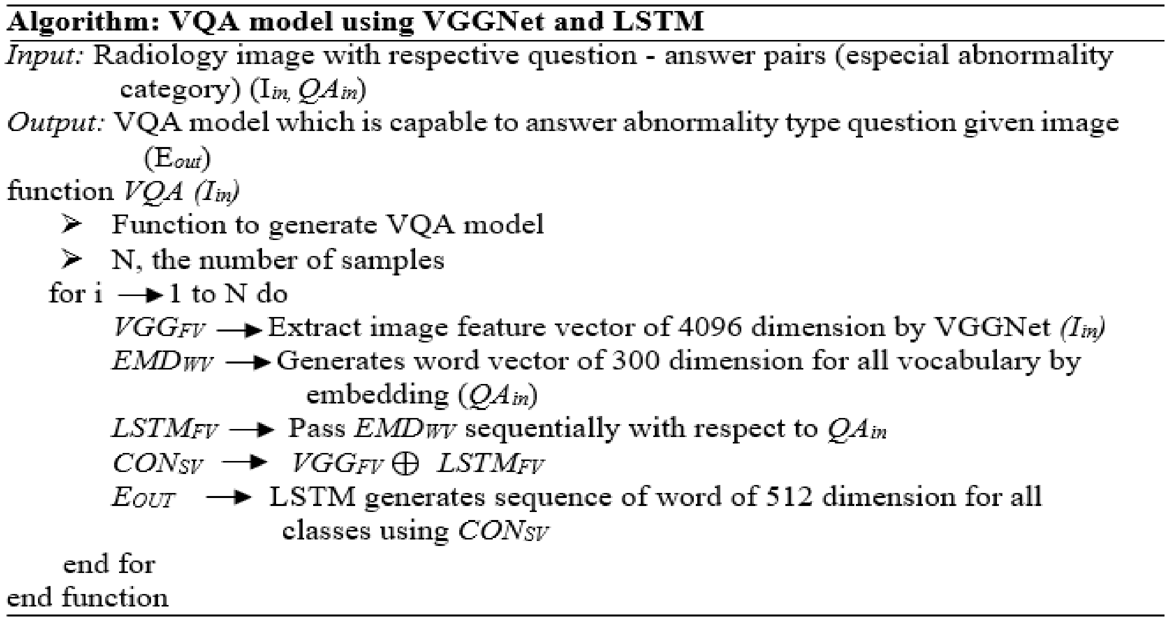
## 3. Proposed Methodology

The proposed VQA model comprises of VGGNet (used as Transfer Learning approach) and LSTM to answer the VQA queries related to radiology images. This VQA model is further tuned by hyperparameter selection as tabulated in Table 3 and supported by three improved VQA – MED dataset (as discussed in Section 2). VGGNet and LSTM are used to obtain the image features and text features respectively. These features are then combined using elementwise multiplication and used for model creation. The output of the model is the sequence of words for all possible answer classes.

**Table 3**
Hyper parameter selection and its respective values

| Hyper Parameters | Value |
|---|---|
| Number of epochs | 800 |
| Batch Size | 256 |
| Momentum | 0.9 |
| Dropout | 0.3 |
| Learning rate | 0.001 |

VGGNet, a pre-trained model, is used as a transfer learning approach. The transfer learning approach is adapted because of three factors namely, (i). Higher start – Model with transfer learning approach outperforms the model without transfer learning approach (ii). Higher slope – Performance rate gradually increases in the training phase (iii). Higher asymptote – Training rate converges smoothly.



```
Algorithm: VQA model using VGGNet and LSTM
Input: Radiology image with respective question - answer pairs (especial abnormality
        category) ($I_{in}$, $QA_{in}$)
Output: VQA model which is capable to answer abnormality type question given image
        ($E_{out}$)
function $VQA\ (I_{in})$
    ➢ Function to generate VQA model
    ➢ N, the number of samples
  for i ⟶ 1 to N do
      $VGG_{FV}$ ⟶ Extract image feature vector of 4096 dimension by VGGNet $(I_{in})$
      $EMD_{WV}$ ⟶ Generates word vector of 300 dimension for all vocabulary by
                    embedding $(QA_{in})$
      $LSTM_{FV}$ ⟶ Pass $EMD_{WV}$ sequentially with respect to $QA_{in}$
      $CON_{SV}$ ⟶ $VGG_{FV} \oplus LSTM_{FV}$
      $E_{OUT}$ ⟶ LSTM generates sequence of word of 512 dimension for all
                  classes using $CON_{SV}$
  end for
end function
```

## 4. Experiments and Results

The proposed model is executed on four datasets (as discussed in Section 2) and the performance is analyzed using five different runs as given in Table 4, are: (i). VGG16 concatenated with LSTM by excluding the last layer with less number of epochs for the given dataset (ii). Same as (i) but number of epochs is increased (iii). Similar to (ii) for Given dataset along with Test dataset from ImageCLEF VQA-MED 2020 (GTD). (iv). Same as (ii) for Augmented Dataset (v). Same as (ii) for Augmented Reduced Dataset.

**Table 4**
Brief Description about each run with performance score

| Run number | Dataset | Number of Epochs | Training Error | Validation Error | Test Set (Performance metrics) | |
|---|---|---|---|---|---|---|
| | | | | | Accuracy | BLEU Score |
| 1 | Given Dataset (GD) | 31 | 0.158 | 0.231 | 0.020 | 0.049 |
| 2 | Given Dataset (GD) | 401 | 0.130 | 0.219 | 0.172 | 0.213 |
| **3** | **Given dataset along with Test dataset from ImageCLEF VQA-MED 2020** (GTD) | **800** | **0.114** | **0.183** | **0.196** | **0.227** |
| 4 | Augmented Dataset (AD) | 800 | 0.134 | 0.225 | 0.172 | 0.211 |
| 5 | Augmented Reduced Dataset (ARD) | 800 | 0.132 | 0.221 | 0.170 | 0.208 |

The performance of the model depends on suitable hyper parameters and the appropriate values as given in Table 3. The result of the proposed model are analysed using suitable quantitative metrics for different runs. The quantitative metrics includes, mean square error for training and validation set and, accuracy and BLEU score for test set as given in Table 4. The overall inferences are: (i). Training error is lesser than validation error because most of the samples are learned in the training phase and followed early stopping (ii). For the third run, both training and validation error is minimum than other runs which leads to better prediction rate (iii). Among the five runs, the third run achieved a better accuracy of 0.196 and the BLEU score of 0.227 for GTD dataset.

**Table 5**
Top 10 ranking of ImageCLEF 2021 VQA-MED

| Rank | Team name | Accuracy | BLEU score | No. of runs submitted |
|---|---|---|---|---|
| 1 | duadua | 0.382 | 0.416 | 10 |
| 2 | Zhao_Ling_Ling_ | 0.362 | 0.402 | 10 |
| 3 | TeamS | 0.348 | 0.391 | 11 |
| 4 | Jeanbenoit_delbrouck | 0.348 | 0.384 | 13 |
| 5 | riven | 0.332 | 0.361 | 1 |
| 6 | Zhao_Shi_ | 0.316 | 0.352 | 4 |
| 7 | IALab_PUC | 0.236 | 0.276 | 7 |
| 8 | Li_Yong_ | 0.222 | 0.255 | 10 |
| 9 | silencec | 0.220 | 0.235 | 2 |
| **10** | **sheerin** | **0.196** | **0.227** | **5** |

The final result of the leaderboard is given in Table 5 where our team achieved 10th place in the listed ranks. Our proposed model achieved improved accuracy of 0.196 and BLEU score of 0.227 due to the usage of timestamps during the training phase. The timestamps play a major role in relearning the appropriate answers of the sample based on the previously predicted answer. It also helps the proposed model to learn temporal patterns from a sequence of question-answer pairs based on radiology images.

The overall experience from the VQA-MED 2021 task is based on the dataset only. It concentrates on abnormality type questions which can be answered more easily than the questions related to organ, plane and modality type. However, the accuracy is reduced by 11.4% as compared with previous year because of two reasons such as: large number of samples and the number of classes are also increased in VQA-MED 2021 task.

## 5. Conclusion

This paper describes an approach to solve Visual Question Answering on medical domain for ImageCLEF VQA - MED 2021 dataset. The ImageCLEF concentrates on abnormality related VQA dataset from previous year onwards. When compared with previous year, the number of samples, abnormality type and difficulty level are increased**.** For the VQA dataset, image features and text features are extracted using VGGNet and LSTM, finally both features are concatenated using LSTM to predict the answer. In this VQA model, word embedding is used which allows the model to focus on the part of the image which is relevant to both the image and the keyword in the question. As irrelevant parts of the radiology image are not taken into consideration and thus the classification accuracy is improved by reducing the chances of predicting wrong answers. To validate the model, four datasets namely, Given Dataset (GD), Given dataset along with Test dataset from ImageCLEF VQA-MED 2020 (GTD), Augmented Dataset (AD) and Augmented Reduced Dataset (ARD) are used in five different runs. Among the five runs of the proposed model the better result is achieved for Given dataset along with Test dataset from ImageCLEF VQA-MED 2020 (GTD) with an accuracy score of 0.196 and BLEU score of 0.227. Even though the 2021 dataset is complex, the appropriate parameter selection and improved datasets helps to maintain the performance of the proposed VQA system.

## 6. 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 challenge successfully.

## 7. References

[1] Agrawal, A., Lu, J., Antol. S., Mitchell, M., Zitnick, C.L., Parikh, D., Batra, D.: VQA: Visual Question Answering, International Journal of Computer Vision, 123 (1), pp. 4 – 31 (2017).

[2] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D. (2015). VQA: Visual question answering, Proceedings of the IEEE International Conference on Computer Vision, pp. 2425–2433 (2015).

[3] Ionescu, B., Muller, H., Peteri, R., Ben Abacha, A., Sarrouti, M., Demner-Fushman, D., Hasan, S.A., Kovalev, V., Kozlovski, S., Liauchuk, V., Dicente, Y., Pelka, O., Garcia Secode Herrera, A., Jacutprakart, J., Friedrich, C.M., Berari, R., Tauteanu, A., Fichou, D., Brie, P., Dogariu, M., Daniel Stefan, L., Gabriel Constantin, M., Chamberlain, J., Campello, A., Clark, A., Oliver, T.A., Moustahfid, H., Popescu, A., Deshayes-Chossart, J.: Overview of the ImageCLEF 2021: Multimedia Retrieval in Medical, Nature, Internet and Social Media Applications. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 12th International Conference of the CLEF Association (CLEF 2021), Romania, September 21-24. Lecture Notes in Computer Science, Springer (2021).

[4] Sheerin Sitara, N., Kavitha, S.: ImageCLEF 2020: An approach for Visual Question Answering using VGG-LSTM for different datasets. In: CLEF 2020 Working Notes, CEUR Workshop Proceedings, Greece, September 22-25 (2020).

[5] Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. In: International Conference on Learning Representations, Canada, pp. 1 - 14 (2014).

[6] 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).

[7] 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).

[8] Ben Abacha, A., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Miller, H.: VQAMed: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019. In: CLEF 2019 Working Notes. CEUR Workshop Proceedings, Switzerland (2019).

[9] 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).

[10] Ben Abacha, A., Sarrouti, M., Demner-Fushman, D., Hasan, S.A., Muller, H.: 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, Romania, September 21-24 (2021).

[11] 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).

[12] https://medium.com/pythoneers/vgg-16-architecture-implementation-and-practical-use-e0fef1d14557