# HARENDRAKV at VQA-Med 2020: Sequential VQA with Attention for Medical Visual Question Answering

Harendra K. Verma[1] and Sindhu Ramachandran S.[2]

[1] Vadict Innovations Pvt. Ltd., Vadodara, Gujarat, India
hkv.aero@gmail.com
[2] Quest Global, Trivandram. Kerala, India
sindhu.ramachandran@quest-global.com

**Abstract.** This paper describes our approach for Medical Visual Question Answering (VQA-Med 2020) Task of ImageCLEF 2020. We used an encoder-decoder architecture for generating answers given the question and image. The encoder takes two inputs: the first is a feature vectors of image obtained from VGG16, and the second is a vector representation for each question using BERT. The question features are self-attended to get attention features. The question attention features, and image features are fused using multi-modal factorized bilinear pooling (MFB). The fused features are further self-attended to get fuse attention features. The thought vectors are obtained by concatenation of fuse attention and encoder LSTM hidden states. The decoder generates answer word by word for the input question and the image. The decoder consists of LSTM layer, Bahdanau Attention, and dense layer of answer vocabulary size with SoftMax. The answers are embedded using GLOVE word vectors before being passed to the decoder LSTM. Our best model achieves 37.8% accuracy and BLEU score of 0.439.

**Keywords:** Visual Question Answering, Sequential VQA Model, Attention, Radiology

## 1 Introduction

Visual question answering requires understanding of Computer Vision and NLP simultaneously. VQA has gained lot of attention from academicians and researchers due to introduction of new language models like BERT, new feature fusion techniques based on bilinear pooling and attention mechanisms. The recent advancements in artificial intelligence have encouraged the healthcare sector in storing large number of health records electronically. VQA can be used for automatic interpretation of radiology images, thereby helping make clinical decisions

In this paper, we present our method to build a deep learning model for ImageCLEF VQA-Med 2020 Task [1]. ImageCLEF conducts many tasks related to multimedia retrieval in many domains such as medicine, security, lifelogging, and nature [2]. Our model is based on encoder-decoder system where encoder takes question and radiology image as input and decoder generates the answer word by word. The Image and question features were extracted using pre-trained VGG16 and 12-Layer BERT Language model respectively. The extracted features for image and questions were fused using Multi-Modal Factorized Bilinear Pooling.

This paper is organized in the following manner: Section 2 provides a brief information regarding similar works done on VQA-Med which inspired our work. Section 3 presents a brief description of dataset used. Section 4 describes our model architecture based on encoder-decoder system. Section 5 describes our model evaluation and results achieved on test set. Section 6 presents conclusions and future work.

## 2 Dataset Description

The dataset used here is VQA-Med-VQA 2020 [1] used in ImageCLEF VQA-Med-VQA competition 2020.

- The training set: 4,000 radiology images with 4,000 associated Question-Answer (QA) pairs.
- The validation set: 500 radiology images with 500 QA pairs.
- The VQA test set: 500 radiology images with 500 associated questions.

Additional data for training was used from abnormality section of VQA-Med 2019 dataset which constitutes 3817 images and 3817 Question-Answer (QA) pairs.

Fig 1 shows few examples of questions and answers from dataset.


**Q.** what is most alarming about this x-ray?
**A.** c-spine fracture


**Q.** what abnormality is seen in the image?
**A.** carcinoma, small cell

Fig 1 Examples from VQA-Med-2020 dataset

## 3 Related Work

In past few years, many interesting approaches are reported on VQA and most of these approaches consider VQA as a classification problem. However, the generative models are the only better options if the answers are long. In particular, the best performing VQA models reported in VQA-Med 2019 task were classification based [3,4,5 and 6].

In this paper, we aim to explain our generative approach for VQA on the VQA-Med 2020 task. We have used VGG16 [7] for image feature extraction and 12-Layer BERT-Base [8] for question features extraction. The answers are decoded using 300d GLOVE [9] Word Embeddings. The glimpse attention mechanism is employed to get question image fuse self-attention while Bahdanau Attention [10] is used for decoder.

## 4 Model Description

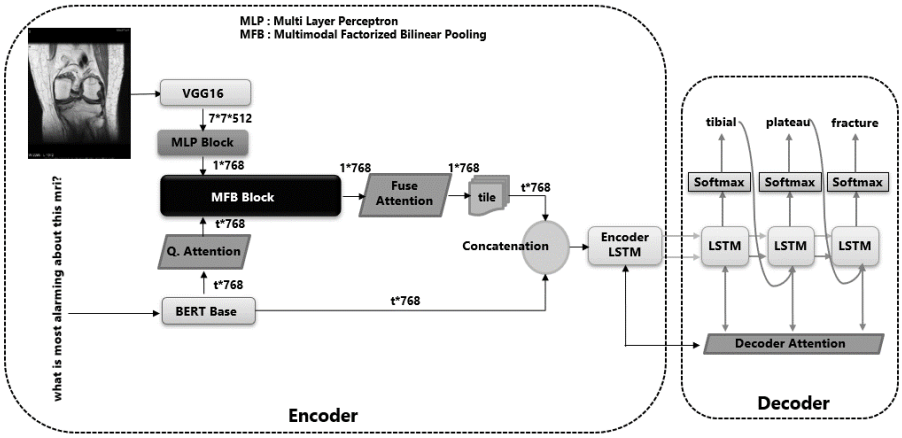


Fig 2 Model Structure

Fig 2 shows the proposed model based on encoder-decoder system. Different parts of this model are described in the following paragraphs.

The encoder has two main components. The first component is a Pretrained VGG16 [7] network which takes the radiology image as input and extracts feature vectors, while the second component is a Pretrained BERT [8] language model which encodes the question text into a vector representation. The output from encoder are thought vectors and encoder LSTM sequence. The input to the encoder LSTM is obtained by concatenation of question image fuse attention and question feature vectors from BERT.

The decoder consists of LSTM network and attention network that takes the thought vector as initial state and encoder LSTM sequence output and try to predict the answer word by word.

### 4.1 Encoder

The encoder takes image and the question as input and returns thought vectors and sequence output.

**Image Feature Extraction**

The image features are extracted using VGG16 Pretrained model from last pooling layer with size of *7*7*512*. This image feature vector is then passed to the MLP block having two fully connected layers with *2048* and *768* hidden nodes respectively along with dropout. The main purpose of this MLP block is to decrease the feature vector dimension to half of the LSTM output vectors.

**Question Feature Extraction**

The semantic meanings of the questions are extracted using *12*-layer, *768* hidden BERT Pretrained model. For each question, we will get a feature vector of size *t**768, where *t* is the max sequence length of the question.

**Question Image Fuse Attention**

It has been reported that the attention mechanism allows the VQA models to effectively learn which regions of the image are important for given question. It's always better to employ an effective attention mechanism to improve the performance of VQA models. Our model uses question image fuse attention based on glimpse attention networks. We have used two glimpses each for question and image as an optimal choice based on Fukui et. al. 2016 [11].

For question self-attention, the output from Pretrained BERT (*t*768*) is passed to question attention layer giving output attention feature vector of size *768*. The image features from MLP block (*1*768*) and question attention features are fused using multimodal bilinear factorize pooling (MFB) [12] giving an output vector of size *1*768*. Finally, the MFB output is passed to image attention layer with two glimpses giving an output of size *1*768*.

**Encoder LSTM**

Long Short-Term Memory networks or LSTM are a special type of RNN that are designed to avoid long range dependency problem of vanilla RNN networks. The LSTM cell carries an extra memory state (other than hidden state h) to store the context information. An LSTM has three different gates (i.e. Input gate, forget gate and output gate) which decide flow of information across time steps.

At each time step, LSTM cell has three inputs viz. current word ($x_t$), previous hidden state ($h_{t-1}$) and previous memory state ($c_{t-1}$), and three outputs viz. output hidden state ($h_t$), output memory state ($c_t$) and encoded sequence.

In our model, the input to encoder LSTM are obtained by concatenation of question features extracted from BERT and tiled fuse attention from MFB. The Encoder LSTM has hidden nodes of *768*. The outputs are hidden state *h* (*768*), memory state *c* (*768*) and encoded sequence of size *t*768*.

### 4.2 Decoder

The decoder generates answer word by word for the input question and the image. The decoder consists of LSTM layer, Bahdanau Attention Layer and dense layer of answer vocabulary size with SoftMax. The answers are embedded using GLOVE [9] word vectors before being passed to the decoder LSTM.

At first time step, decoder takes four inputs viz. the start of sequence <SOS>, hidden state of the encoder, memory state of encoder and encoder sequence output. The output will be the first word of the answer which is obtained with highest probability using SoftMax layer. This word will be the input to LSTM at second time step to predict the next word of the answer. This process keeps on going until the decoder predicts end of sequence <EOS> token.

## 5 Experiments and Results

A total of 5 runs were submitted to ImageCLEF VQA-Med-VQA 2020. The training was done on NVIDIA GeForce 940 MX GPU Device. The CUDA implementation of LSTM was employed from Tensorflow 2.0. Evaluation of the model was conducted using two different metrics: BLEU and Strict Accuracy, based on VQA-Med-VQA 2020 competition [1]. There are few pre-processing steps applied on each answer before running the evaluation metrics: lower-case, remove all punctuations and remove stopwords using NLTK. Table 1 shows model description and their accuracy scores of all the runs submitted.

Table 1 Results Comparison

| Run | Model (VGG+BERT+MFB+GLOVE) Parameters | Accuracy/BLEU |
|---|---|---|
| 1 | Model1 {encoder_hidden=1024, 34M, Adam, drop=0.3} | 0.34/0.398 |
| 2 | Model2 {encoder_hidden=768, 26M, RMSprop, drop=0.3} | 0.366/0.426 |
| 3 | Model2 {encoder_hidden=768, 26M, RMSprop, drop=0.4} | 0.34/0.408 |
| 4 | Model2 {encoder_hidden=768, 26M, RMSprop, drop=0.3, Train on Validation} | 0.378/0.438 |
| 5 | Model2 {encoder_hidden=768, 26M, RMSprop, drop=0.3, Added dropout on question features} | 0.36/0.423 |

There are two main models: model1 has hidden size of 1024 with approx. size of 36M and model2 with hidden size of 768 and approx. size of 26M. Model1 is trained

with Adam optimizer while model2 is trained with RMSprop optimizer. Experiments were also done with different dropouts.

## 6 Conclusion

In this paper we describe the model we submitted in ImageCLEF 2020 VQA-Med-VQA task. Our proposed model VGG+BERT+MFB+GLOVE employs an encoder-decoder architecture with an advanced feature fusion technique MFB with question image fuse attention. The image and question features were extracted using VGG16 and BERT Base networks respectively. The answers were encoded using GLOVE word embeddings. Our model achieved the accuracy of 37.8% and BLEU score of 0.439 on the test set.

## 7 References


1. Ben Abacha, S.A., Datla, A., Hasan, 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, CEURWS.org <http://ceur-ws.org>, Thessaloniki, Greece (September 22-25 2020)
2. Ionescu, B., Muller, H., Peteri, R., Abacha, A.B., 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., Ninh, V. T., Tu-Khiem Le, Zhou, L, Piras, L., Riegler, M., Halvorsen, P. I, Tran, M. T., Lux, M., Gurrin, C., Dang-Nguyen, D.T., Chamberlain, J., Clark, A., Campello, A., Fichou, D., Berari, R., Brie, P., Dogariu, M., Stephan, 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), LNCS Lecture Notes in Computer Science, Springer, Thessaloniki, Greece (September 22-25 2020).
3. Yan, X., Li, L., Xie, C., Xiao, J., Gu, L.: Zhejiang university at imageclef 2019 visual question answering in the medical domain. In: Working Notes of CLEF 2019 (2019)
4. Vu, M., Sznitman, R., Nyholm, T., Lfstedt, T.: Ensemble of streamlined bilinear visual question answering models for the imageclef 2019 challenge in the medical domain. In: Working Notes of CLEF 2019 (2019)
5. Zhou, Y., Kang, X., Ren, F.: Tua1 at imageclef 2019 vqa-med: A classication and generation model based on transfer learning. In: Working Notes of CLEF 2019 (2019)
6. Shi, L., Liu, F., Rosen, M.P.: Deep multimodal learning for medical visual question answering. In: Working Notes of CLEF 2019 (2019)
7. Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations (ICLR)
8. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. neural information processing systems, pages 5998{6008, 2017.
9. J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In EMNLP, volume 14, pages 1532–1543, 2014. 7

10. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In ICLR, 2015.
11. A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847, 2016. 2, 3, 4, 5, 6, 7, 8
12. Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 1839{1848, 2017.