Medical Visual Question Answering at Image CLEF 2019- VQA Med

Mohit Bansal, Tanmay Gadgil, Rishi Shah and Parag Verma

Pricewaterhouse Coopers US Advisory, Mumbai, India
mohit.b.bansal@pwc.com,tanmay.p.gadgil@pwc.com,
rishi.s.shah@pwc.com, parag.verma@pwc.com[1]

Abstract. This paper describes the submission created by PwC US-Advisory for the Medical Domain Visual Question Answering (VQA-Med) Task of Image CLEF 2019. The goal of the challenge was to create a Visual Question Answering System which uses medical images as context to generate answers. The VQA pipeline classifies the questions into two groups, the first group of questions involves giving answers from a fixed pool of predefined answer categories and the second group of questions involves generating answers based on the abnormality seen in the image. The first model uses question embeddings from the Universal Sentence Encoder and Image Embeddings from ResNet which are fed into an attention-based classifier to generate answers. The second model uses the same ResNet image embedding along with word embeddings from a Word2Vec model pre-trained on PubMed data which is used as an input to a sequence to sequence model which generates descriptions of abnormalities. This methodology helped us achieve reasonable results with a strict accuracy of 48% and a BLEU score of 53% on the challenge's test data.

Keywords: Visual Question Answering, Sequence to Sequence Model, Image CLEF 2019, Attention

1 Introduction

Technical advancements in the healthcare sector have helped in storing large number of health records electronically and this provides opportunities for the use of artificial intelligence (AI) in supporting clinical decision making and improving patient engagement. Visual Question answering (VQA) is one way in which the technical innovation in Artificial Intelligence can be used for the benefit of the healthcare sector. VQA is a relatively new approach that uses a combination of computer vision and natural language processing for designing

the model. Given the image and question input, VQA technique processes both the visual and textual inputs to give relevant answer to the input question.

Some interesting approaches have been published over the years on VQA, but for this challenge we have come up with a relatively simple architecture to best address the Image-CLEF 2019 problem. In this paper, we aim to explain the details about different models used for getting the image embeddings, question embeddings and answer embeddings and for finally training the model. In subsequent parts of this paper, we go on to discuss the possible applications of the VQA model in the industry and possible changes in model design for enhancing the accuracy.

2 Literature Review

In the last few years, there have been tremendous advancements in the field of AI. From neural network to computer vision, the field has transformed through leaps and bounds [1]. Among other things, its ability to impact the value chain has been felt unequivocally [2].

Healthcare as a domain is very fragmented and the components in the value chain don't adhere to accepted standards [3]. Such discretization provides immense opportunity for consolidation and optimization in the Provider Market [4]. A physician scans scores of documents and images before arriving at a conclusion. Without having a heuristic process in place, he or she may have to often spend substantial amount of time sieving through the scans and vitals before arriving at a conclusion. Visual Question Answering (VQA) is a recent advancement in AI which tries to answer a set of questions about an image [5]. Tasks such as feature extraction, understanding question and generating answers are an important part of the VQA. Complex systems built to extract information out of images and scans need to incorporate VQA along with natural language processing to leverage the huge data generated at the patient and provider interface [6]. These will not only supplement the information at hand with the clinician to make better decisions but also enable better knowledge management.

Earlier work in VQA was concentrated mostly towards image captioning [7]. In most of them, deep learning methods such as CNN, LSTM, DMN had been used. The proposed model aims to find a link between the semantics of the text description and the features extracted from an image.

3 Task Description and Dataset

As part of the Image-CLEF 2019 VQA-Med Task [12] [13], the participants were given medical images accompanied by few clinically relevant questions, and were supposed to answer the questions based on the visual image content. The training dataset consisted of 3200 medical images and 12,792 question-answer pairs, while the validation set had 500 medical images and 2000 question-answer pairs. Finally, after developing the model framework, the participants were required to predict answers for a test set of 500 medical images with 500 questions

4 Pipeline overview / Model Structure

In this section we discuss our model architecture, which primarily consists of 3 main components: Question Classifier Model, VQA Classifier Model, VQA Seq-2-Seq Model as shown in Fig. 1. Due to the complex nature of the problem, we decided to break the problem into simpler tasks which could be handled independently. We sorted the questions into its respective categories namely Modality, Plane, Organ system & Abnormality using the Question Classifier Model. The first group consisting of all questions on Modality, Plane & Organ system categories along with their image pairs was fed to VQA Classifier Model to implement a multi-class classification model. The second group consisting of all the questions on Abnormality category along with their image pair was fed to the VQA Seq-2-Seq Model to generate answer sequences for predictions.

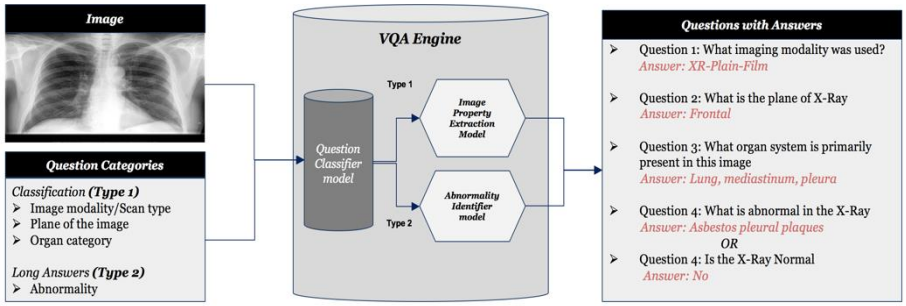

Fig. 1. Model Pipeline Overview

4.1 Question Classifier Model

To classify the questions, question embeddings were passed through a dense connected layer. This deep representation was then projected using a final dense layer followed by a softmax activation to form a distribution over all possible tags. Question embeddings were created using the Universal Sentence Encoder as it gives strong transfer performance on a number of NLP tasks and surpasses the performance of transfer learning using word level embeddings alone. Also, it handled variable input text length seamlessly to give a 512-dimensional vector output.

4.2 VQA Classifier Model

Once the questions were classified in the question classifier, questions and images from the category of Modality, Plane and Organ System were fed to the VQA classifier. The VQA classifier model used a Universal Sentence Encoder to create Question embeddings. To encode the information of the category to which the question belongs, one-hot-encoded vector of the Question Classifier was concatenated to the 512-dimensional vector to create a final 516-dimensional question embedding vector. A pre-trained convolutional neural network (CNN) model based on ResNet50 architecture [8] was used to embed the image.

To compute attention over image features, we concatenated tiled LSTM state with image features over the depth dimension and passed it through a 1×1 dimensional convolution layer of depth 512 followed by ReLU nonlinearity. The output feature was passed through another 1×1 convolution of depth C= 2 followed by softmax over spatial dimensions to compute attention distributions. We used these distributions to compute two image glimpses by computing the weighted average of image features. We further concatenated the image glimpses with the state of the LSTM and passed it through a fully connected layer of size 1024 with ReLU nonlinearity [10]. The output was then fed to a linear layer of size M = 66 followed by softmax to produce probabilities over the different sub-categorical answers in the top 3 categories. We used dropout of 0.5 on input features of all layers including the LSTM, convolutions, and fully connected layers and optimized the model with Adam optimizer. Fig 2 below shows the model architecture

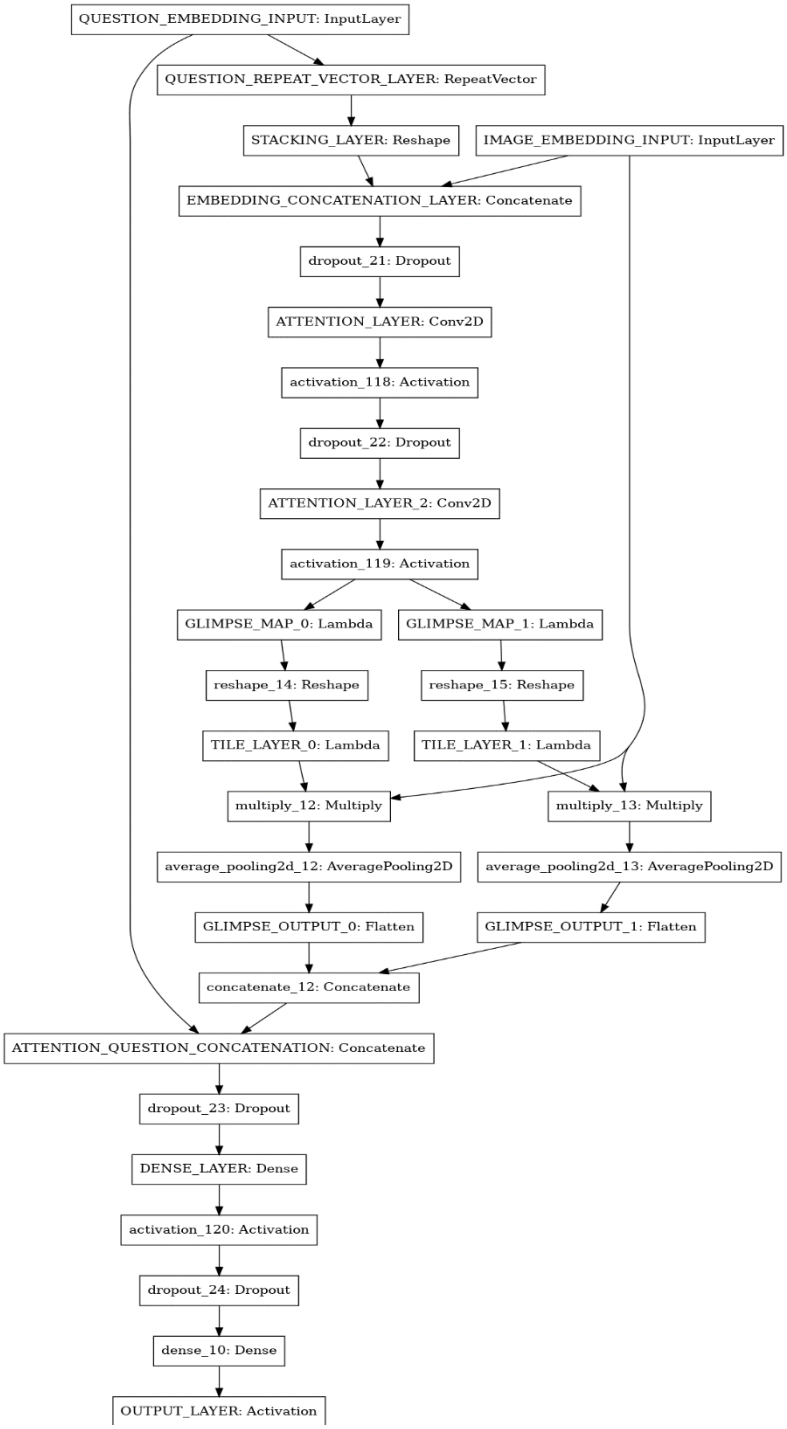

Fig. 2. VQA Classifier Architecture

4.3 VQA Seq-2-Seq Model

Questions and images from the Abnormality category were fed to the VQA Seq-2-Seq Model as shown in Fig 3. This is a custom encoder decoder architecture that was used to generate answer sequences. In the following section, this will be discussed in detail. The first component was a pre-trained convolutional neural network (CNN) model based on ResNet50 architecture that took the image as an input and extracted a vector representation for that image, while the second component was a word embedding layer that encoded the question into a vector representation which was passed through a LSTM network. The embeddings were created using pre-trained word2vec model on PubMed data. The decoder consisted of LSTM network that took the thought vector as initial state and 'Start of Sentence' <SOS> token as input in the first time step and tried to predict the answer tokens using softmax layer until 'End of Sentence' <EOS> was obtained [11].

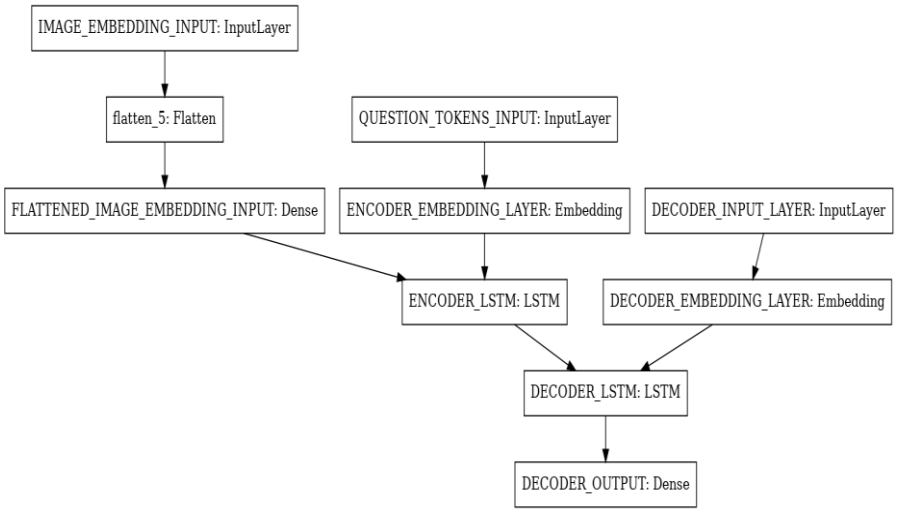

Fig. 3. VQA Seq-2-Seq Architecture

4.3.1 Encoder

The encoder layer consists of 2 main components: the first is used to obtain image embeddings from a CNN architecture and the second is a LSTM network with a pre-trained word embedding layer to encode questions. Fig 4 below describes the encoder architecture.

In the first component, the image embedding input layer takes the vectorized image from ResNet50. The image is flattened and passed through a fully connected dense layer followed by a RELU activation to create a deep 512-dimensional representation of the image. The main purpose of these two layers is to reshape and compress the feature vector dimension to match the hidden layer of the LSTM.

In the second component, the semantic meaning of the question is to be extracted with respect to the image. A 100-dimensional pre-trained word embedding layer is used to encode the word into a dense semantic space using a word2vec model trained on PubMed Data. This word representation is then fed to a LSTM network with 512 hidden nodes. LSTM is a special type of Recurrent Neural Network (RNN) that has been designed to solve the problem of vanishing gradient. The LSTM layer uses its memory cells to store the context information. LSTM has three gates (i.e. Input gate, forget gate and output gate) which will decide how the input will be handled. At any time step, inputs to the LSTM cell include the current word (x), previous hidden state (h-1) and previous memory state (c-1), and LSTM cell outputs are current hidden state (h) and current memory state (c). These states have 512 hidden nodes. At last time step in the sequence, the LSTM cell outputs its hidden state (h) and memory state (c). Both the hidden state and the memory state of the first LSTM layer is initialized with the image vector which helps LSTM layer learn the internal representations useful for extracting information relevant to answering the question.

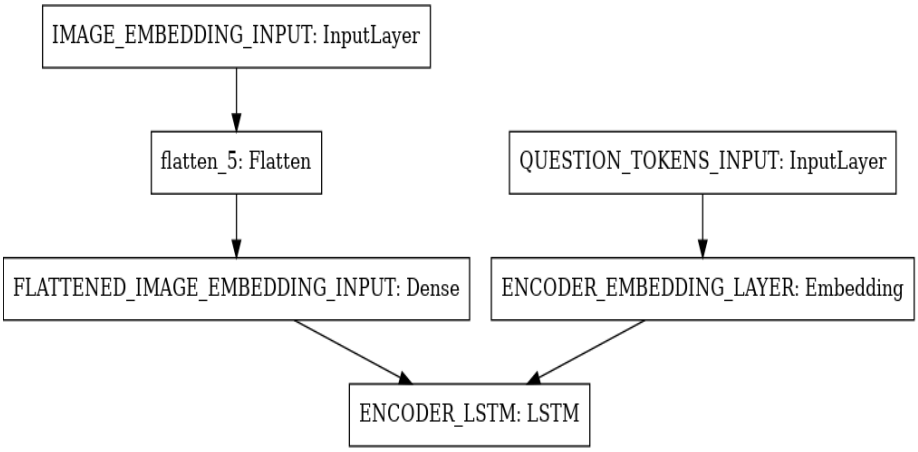

Fig. 4. Encoder Model Architecture

4.3.2 Decoder

The decoder model is responsible for generating the answer from the input image and question. The Decoder LSTM uses three inputs to generate a token which is mapped to a dictionary. These include the token of the previous word, the hidden state and the memory state of the previous LSTM layer as shown in Fig 5. At the first time step, LSTM cell takes 'Start of Sentence' <SOS> token, the hidden and memory state of the encoder model as the input and calculates the probability distribution of the target word using softmax layer. The word with the highest probability will be the first word of the answer; this word will be then passed to the second LSTM cell as input and predict the second word of the answer. The full answer will be generated by repeating this process until the model predicts 'End of Sentence' <EOS> token.

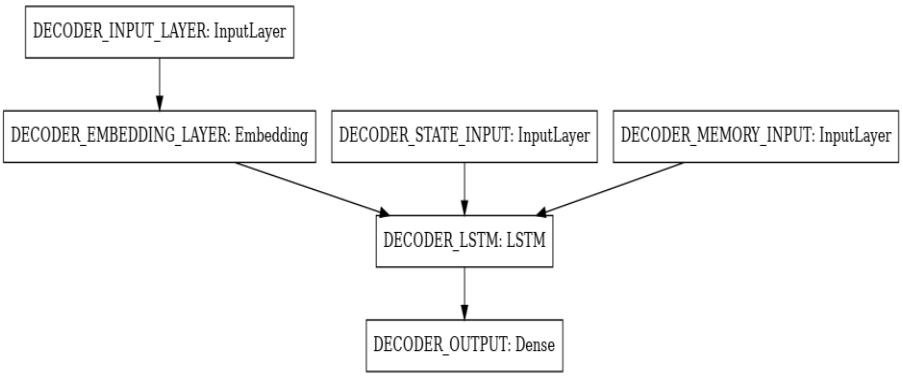

Fig. 5. Decoder Model Architecture

5 Results

Training data and validation data for the challenge comprised of 3,200 images with 12,792 Question-Answer (QA) pairs & 500 images with 2,000 QA pairs respectively. Test data contained 500 images and questions. We experimented with primarily 2 approaches for challenge submissions:

Approach 1: In the first experiment, we use the Question Classifier, followed by VQA Classifier on the top 3 categories and a second VQA Classifier model trained on Abnormality category train data only. We only consider top 657 most frequent answers in the abnormality classifier. Other answers are ignored and

do not contribute to the loss during training. This covers 74% of the answers in the validation set of abnormality category in VQA dataset [9]

Approach 2: In the second experiment, we use the proposed pipeline, Question Classifier, followed by VQA Classifier on the top 3 categories and VQA Seq-2-Seq Model on the Abnormality category questions

The question classifier model to classify the question category gave a 100% accuracy on the validation set. The VQA classifier trained on the first 3 categories (Modality, Plane and Organ Systems) delivered an overall 76.1% accuracy in predicting answers across 66 sub-categorical classes on validation data of the corresponding 3 categories. The second VQA classifier trained on the Abnormality category questions, used in Approach 1, delivered a 19.5% accuracy in predicting answers across 657 sub-categorical classes on the validation set of Abnormality category.

The VQA Seq-2-Seq model, used in Approach 2, generated answer sequences with an accuracy of 23.4% on validation set of Abnormality category. Evaluation of test data was conducted based on the following metrics:

1. Accuracy (Strict)- an adapted version of the accuracy metric from the general domain VQA task that considers exact matching of a participant provided answer and the ground truth answer.;
2. BLEU metric [1] to capture the similarity between a system-generated answer and the ground truth answer.;

With different model hyper parameters, 5 valid submissions were given by our team; 2 from Approach 1 and 3 from Approach 2. The best results from both approaches were as follows:

Table 1. Performance statistics of model on test data

	Accuracy (Strict)	BLEU Score
Approach 1	0.484	0.531
Approach 2	0.488	0.534

It is worth mentioning that both the VQA Classifier and VQA Seq-2-Seq models are not very expensive to train. It took 14 seconds and 12 seconds

per epoch for the VQA-Classifier and VQA-Seq-2-Seq respectively on a Virtual Machine (VM) equipped with NVIDIA Tesla P4 GPU card with 8GB of RAM. The VM had Ubuntu OS with CUDA 10.1. For the implementation, we use Keras with TensorFlow 1.13.1 backend.

6 Conclusion

VQA has great potential with respect to the automation of information gathering from images. The medical domain can be greatly benefited in several ways by these techniques, like providing the coherence required to simplify the diagnosis process, providing clinicians with a tool to check the validity of their diagnosis and providing patients with details which may otherwise be overlooked during a consultation. It can empower the clinician to ascertain that enough information is at hand before arriving at any conclusion. Specific models relevant for a particular disease can also be made by training the algorithm on specific data sets, and this can create a better scanning process. Supplementing the models with more advanced techniques like attention and Multitask learning can greatly enhance their accuracy. It can also create standards in terms of models used by different providers. Moreover, the cost benefits incurred from such simplification can be propagated through the entire value chain. Clinicians, patients and insurance providers can benefit greatly from the value created by VQA based diagnosis systems.

References

1. Koushal, K., Gour, S.: Advanced Applications of Neural Networks and Artificial Intelligence: A Review, (2012).
2. Soroush, A., Nakhai, I., Bahreininejad, A.: Review on Application of Artificial Neural Networks in Supply Chain Management and its Future, (2009).
3. Lawton, R.: The Health Care Value Chain: Producers, Purchasers, and Providers, (2002).
4. Martin, G.: Examining the Impact of Health Care Consolidation, (2018).
5. Aishwarya, A., Jiasen, L., Stanislaw, A., Margaret, M., Lawrence, C., Dhruv, B., Devi, P.: VQA-Visual Question Answering, (2016).
6. Sonit, S.: Pushing the Limits of Radiology with Joint Modeling of Visual and Textual Information, (208).

7. Shuhui, Q.: Visual Question Answering Using Various Methods,California,94305
8. Alfredo, C., Eugenio, C., Adam, P.: An Analysis of deep neural network models for practical application, (2017).
9. Antol, S., Agrawal, A., Lu, M., Mitchell, D., Batra, C., Zitnick, L., Parikh, D.: VQA: Visual question answering. In International Journal of Computer Vision, (2015).
10. Vahid, K., Ali, E.: Show, Ask, Attend, and Answer: A Strong Baseline For Visual Question Answering. CoRR, abs/1704.03162, (2017).
11. Ilya, S., Oriol, V., Quoc, V.: Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems (NIPS) 27, pages 3104–3112, (2014).
12. Asma, B., Sadid, A., Vivek, V., Joey, L., Dina, D., Henning, M.: VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019 In: CLEF 2019 Working Notes. CEUR Workshop Proceedings (CEUR- WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-2380/, Lugano, Switzerland, (September 09-12, 2019)
13. Bogdan, I., Henning, M., Vivek, V., Renaud, P., Yashin, D., Vitali, L., Vassili, K., Dzmitri, K., Aleh, T., Asma, B., Sadid, A., Vivek, D., Joey, L., Dina, D., Duc-Tien, D. , Luca, P., Michael, R., Minh-Triet, T., Mathias, L., Cathal, G., Obioma, P., Christoph, M. , Alba, G., Narciso, G. , Ergina, K., Carlos, R., Carlos, C., Nikos, V., Konstantinos, K., Jon, C., Adrian, C., Antonio, C.: ImageCLEF 2019: Multimedia Retrieval in Medicine, Lifelogging, Security and Nature In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019), Lugano, Switzerland, LNCS Lecture Notes in Computer Science, Springer, (September 9-12, 2019).